NORTHWESTERN
UNIVERSITY

Computer Science Department

Technical Report
Number: NU-CS-2023-19

December, 2023

Large Language Models for Automatic Peer Review and Revision in Scientific Documents

Mike D'Arcy

Abstract

In this dissertation, we investigate the capacity of large language models (LLMs) to assist with writing scientific papers, both by revising papers in response to feedback and by generating feedback on paper drafts. Our investigation aims to both understand the limits of LLMs in this highly technical setting and to advance the development of tools to accelerate the scientific process. We study revision and feedback generation separately, focusing first on the relatively simpler task of writing or identifying a relevant paper edit given a human-written critique and then on the task of identifying critiques and writing feedback comments.

Our findings suggest that while LLMs do show potential for generating feedback comments and edits for papers, they still suffer from significant limitations when attempting to comprehend or produce nuanced and technical text, often exhibiting surface-level reasoning and producing generic outputs. However, we show that by using multiple LLM instances that engage in internal discussion, the quality and specificity of outputs can be substantially improved.

Keywords

Machine Learning, Natural Language Processing, Peer Review, Writing Assistance, Multi-agent system

NORTHWESTERN UNIVERSITY

Large Language Models for Automatic Peer Review and Revision in Scientific Documents

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Mike D'Arcy

EVANSTON, ILLINOIS

March 2024

ABSTRACT

Large Language Models for Automatic Peer Review and Revision in Scientific Documents

Mike D'Arcy

In this dissertation, we seek to evaluate LLM capabilities for reviewing and revising scientific documents and to develop new methods to improve them. The capabilities of large language models (LLMs) have advanced dramatically in recent years, performing on par with humans in some tasks. However, the ability of models to comprehend and produce long, highly technical text—such as that of scientific papers—remains under-explored.

We construct ARIES, a dataset of scientific paper drafts, their associated peer reviews, and the new drafts after reviews, and we link individual feedback comments to specific edits that address them. Using ARIES, we study the ability of LLMs to edit scientific papers in response to feedback and to generate feedback comments.

Our findings suggest that LLMs do show potential for generating feedback comments and edits for papers, but still suffer from significant limitations when attempting to comprehend or produce nuanced and technical text, often exhibiting surface-level reasoning and producing generic outputs. When revising a document in response to feedback, LLMs often write edits by quoting or paraphrasing the given feedback (48% of the time, compared to 4% for humans) and

tend to include less technical detail (38% of model edits vs 53% of human edits had technical details). Similarly, when generating feedback comments for papers, baseline methods using GPT-4 were rated by users as producing generic or very generic comments more than half the time, and only 1.5 comments per paper were rated as good overall in the best baseline.

We explore ways to mitigate these shortcomings and develop MARG-S, an approach for generating paper feedback using multiple specialized LLM instances that engage in internal discussion. We show that MARG-S substantially improves the ability of GPT-4 to generate specific and helpful feedback, reducing the rate of generic comments from 51% to 17% and generating 4.2 good comments per paper (a 2.8$\times$ improvement).

Acknowledgements

First and foremost, I would like to thank my family for their love and support throughout my life and my academic career. I'd also like to thank Chris Zorman for introducing me to the world of academic research many years ago (and Andrew Barnes, who mentored me at the time) and for providing wisdom at several points along my journey; those experiences had a tremendous impact on my research mindset and on my choice to pursue a PhD in the first place.

I would like to thank my committee members, Tom Hope, Larry Birnbaum, Bryan Pardo, and Doug Downey for their feedback and advice. I am also grateful to everyone I worked with at AI2; Sergey Feldman, Arman Cohan, Amanpreet Singh, Erin Bransom, Bailey Kuehl, and many others. In particular, I am thankful to my collaborators Alexis Ross and Jonathan Bragg for many insights and discussions on this dissertation.

Many thanks to my friends and colleagues in the NLP group at Northwestern for their support and camaraderie. Mohammed Alam, David Demeter, and Zheng Yuan, thank you for your mentorship and support throughout my early years of graduate school.

Finally, I would like to thank my advisor, Doug Downey, for his guidance and support throughout my time at Northwestern. Doug's excitement for science and insightful perspectives on research have been inspirational throughout my time here, and I am grateful that he has always made time to provide mentorship and advice when I needed it most.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Writing high-quality scientific papers is a challenging task, requiring authors to not only carefully review related work and consider nuanced details in the design of methods and evaluation but also to find effective ways to organize and communicate their findings with a broader audience. As it is easy to overlook potential pitfalls or miss valuable insights, authors often solicit feedback on their drafts, and most publication venues require papers to undergo peer review as a standard practice. Peer review serves both to vet papers for publication and to provide new perspectives and suggestions to authors that help them to improve the work.

While peer review feedback can provide valuable insights and identify mistakes, it suffers from several drawbacks in practice. Reviewers may be biased, unreasonable, or disinterested, leading to sparse or overly-harsh feedback that is not particularly helpful for authors [36, 62]. The process also places a high burden on reviewers, who are typically professional scientists with busy schedules, and it can take weeks or months for authors to finally receive feedback.

In recent years, the capabilities of large language models (LLMs) have advanced dramatically, resulting in modern models such as GPT-4 that can perform comparably to humans in some tasks [50]. These advancements provide hope that LLMs may be able to assist human researchers with their writing; however, most modern LLMs can only consume limited amounts of text and are primarily trained on non-technical text such as news articles and websites. The ability of models to comprehend and produce long, highly technical text—such as that of scientific papers—remains under-explored. LLMs have shown promise in tasks such as summarizing or

extracting information from a text, but the tasks of critiquing and revising scientific documents pose unique challenges in that they require reasoning about highly technical and specialized subjects, carefully attending to small details, and understanding not only what is written in the text but also what is missing.

For example, consider the following feedback comment from a reviewer of a paper about a fabric physics modeling technique [19]:

> *"Experiments have been focused on simulated woven cloth. Yet, the models are heavy-handedly designed. This casts some doubt regarding the generalizability of the proposed method."*

This comment may appear straightforward at a glance, but writing such a comment requires both careful reasoning and background experience. It requires not only understanding the details of the proposed method and evaluation setting, but also inferring information about the design process and recognizing how this limits the conclusions that can be drawn from the evaluation. Notice that the comment would likely not be relevant if the paper had included very comprehensive and realistic experiments or provided a strong justification of the generalizability (e.g., if the design decisions of the model were made completely independently of the evaluation setting); such nuances must be accounted for when writing feedback.

Revising the paper based on the comment carries similar reasoning challenges. Authors must understand the intent of the comment—which does not make a direct request—and then determine how to modify the paper to address it. This may involve choosing among many possibilities—such as adding additional experiments, providing a theoretical argument for generalizability, or adding a clarification that the design is based on past work—and then finding an eloquent way to integrate that information into the paper.

In this dissertation, we investigate the capacity of large language models (LLMs) to assist with writing scientific papers, both by revising papers in response to feedback and by generating feedback on paper drafts. Our investigation aims to both understand the limits of LLMs in this highly technical setting and to advance the development of tools to accelerate the scientific process. We study revision and feedback generation separately, focusing first on the relatively simpler task of writing or identifying a relevant paper edit given a human-written critique and then on the task of identifying critiques and writing feedback comments.

Our findings suggest that while LLMs do show potential for generating feedback comments and edits for papers, they still suffer from significant limitations when attempting to comprehend or produce nuanced and technical text, often exhibiting surface-level reasoning and producing generic outputs. However, we show that by using multiple LLM instances that engage in internal discussion, the quality and specificity of outputs can be substantially improved.

1.1. Modeling the relationship between feedback comments and edits:

In chapter 2, we investigate the relationship between reviewer feedback and the edits made by authors in response. Previous work on edit modeling either focuses on stylistic and grammatical edits [30, 34, 46, 47, 67] or incorporates no feedback [14, 27, 48] or very different kinds of feedback—such as explicit instructions [25, 44, 55, 58, 77] or descriptions of edits created post-hoc [17, 57, 60]. In contrast, we investigate contentful editing in a highly technical domain—scientific papers—and the review comments these edits are conditioned on are much more complex than previous kinds of feedback, as illustrated in Figure 2.2.

As no prior resources exist for studying the review-revision relations we wish to explore, we construct ARIES, a dataset of scientific paper drafts, their associated peer reviews, and the

new drafts after reviews. By manually examining the reviews and revisions, we obtain a set of individual feedback comments linked to specific paper edits.

We apply large language models (LLMs) to the task of aligning feedback comments to edits; that is, given the set of all edits that authors made to their paper after receiving reviews, the model must determine which specific edits (if any) address each reviewer comment (e.g., the comment *"is your dataset public?"* aligns with the edit "we study performance on a [+private+] dataset..."). We find that the alignment task is challenging even for GPT-4 [50], a state-of-the-art model. Further analysis reveals that LLMs often fail to see past the surface-level wording of a comment or edit to grasp the nuanced semantics. Comments are often erroneously aligned to edits that are topically similar, and real comment-edit pairs are often missed when the comment is worded in an indirect way or when an edit is written to rebut the comment rather than strictly obeying the suggestion.

In addition to the alignment task, we investigate whether GPT-4 can generate good edits when given feedback comments. We find that GPT-4 generally produces edits that are fluent and relevant to the topic of the comments. However, as with the alignment task, it fails to model the underlying intent; whereas real authors sometimes respond to feedback by adding clarifications that suggest the feedback itself is mistaken, GPT-4 almost always obeys the feedback. In addition, the generated edits often borrow wording from the feedback itself rather than tightly integrating edits into the context of the paper, and tend to include less technical detail.

1.2. Generating peer review comments:

In chapter 3, we investigate the ability of GPT-4 to generate peer review comments for scientific papers and propose multi-agent review generation (MARG-S), a method of prompting GPT-4 to generate comments.

Past work on automatic review generation primarily does so using (relatively) small models that cannot consume the full text of a paper [65, 78], whereas the GPT-4 model we study is substantially more powerful and can consume larger amounts of text (full papers when combined with MARG). Previous work that does use GPT-4 [41] generates reviews with a single pass through the model, which we find results in generic and vague comments—similar to the issues we observed in edit generation—and typically makes it impossible to consume the entire text of a paper due to input length limitations. To mitigate these issues, MARG-S structures the review generation task as an interaction between several instances of GPT called *agents*. To handle long papers beyond the token limit of the model, MARG-S splits papers into chunks and gives each chunk to a separate agent, and to improve comment quality we introduce "expert" agents that handle specific sub-tasks and engage in internal discussion to identify weaknesses of papers.

The idea of applying multiple instances of a model to a task is not new, but is typically seen in the context of multi-agent reinforcement learning on games and robotics tasks [51, 80]. Recently, there has also been work on multi-persona modeling with LLMs to simulate artificial societies [39, 52] and to improve reasoning abilities [15, 68]. Unlike those works, we explore the use of multi-agent modeling to scale input size limits and investigate its potential for the highly technical task of scientific review generation, and for MARG-S we design specialized agents and sub-tasks to promote diverse and high-quality review comments.

We evaluate MARG-S using both an automated evaluation and a user study. The automated evaluation aligned generated comments with those extracted from real reviews, and we find that MARG-S achieves a recall of 15.8% compared with 9.7% for a method from contemporaneous work (LiZCa). Note that while the numbers may appear small, it is expected that many comments will be missed due to the diversity of human comments, conservative matching of generated and real comments, and lack of visual information such as figures for the models. In fact, because human typically write fewer comments than MARG-S, they typically achieve only 9.4% recall against other reviewers.

In the user study, we find that MARG-S generates more total good comments and a higher proportion of good comments compared to LiZCa. Of note, we find that MARG-S's comments are rated as much more specific on average. Overall, 83% of MARG-S's comments are rated as specific, compared with only 49% for LiZCa. This is crucial, as it is relatively easy to generate generic comments like "add more experiments" that may technically be valid but offer little real insight.

Our findings provide several insights towards improved understanding of technical text with LLMs. The improvements in recall on the automated evaluation and the high quality scores in the user study suggest that splitting a long paper among multiple agents is an effective way to scale LLM systems beyond the limits of the base model. In addition, while the low specificity of baselines supports the general conclusion that GPT has difficulty producing detailed and technical text, the much higher specificity of MARG-S provides a path towards mitigating this limitation through multi-agent modeling and careful structuring of the task. Nonetheless, when we investigated the accuracy of comments we found that MARG-S produces just as many

highly-inaccurate comments as other methods, indicating that more work is needed to reduce hallucinations and logical errors.

CHAPTER 2

ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews

2.1. Introduction

In this chapter we focus on a task that encapsulates multiple challenges in reasoning about scientific text: revising papers in response to peer review feedback. This task provides a testbed for evaluating NLP systems on important and understudied capabilities needed for effective scientific assistants—performing the task requires a deep understanding of the full text of a scientific paper, and the ability to infer the intent behind technical human feedback and act upon it (revise the paper).

Feedback on paper drafts, whether from co-authors, readers, or reviewers, can be challenging to interpret and address because it often includes complex critiques of a paper's substance and can be phrased in an indirect way. For example, consider a reviewer who wants authors to use a more realistic dataset in their evaluation. This could be expressed in a variety of ways; it could be stated as a direct request (*"Apply the method to a realistic dataset"*), or more indirectly as a criticism (*"The evaluation is only on a synthetic dataset"*) or as a question (*"Is the current dataset truly representative of the real-world?"*). Similarly, an author editing the manuscript in response has several options: they could simply comply with the request, or they could clarify that no realistic datasets are publicly available, or they might even argue that the reviewer is mistaken and add a justification of their dataset's realism.

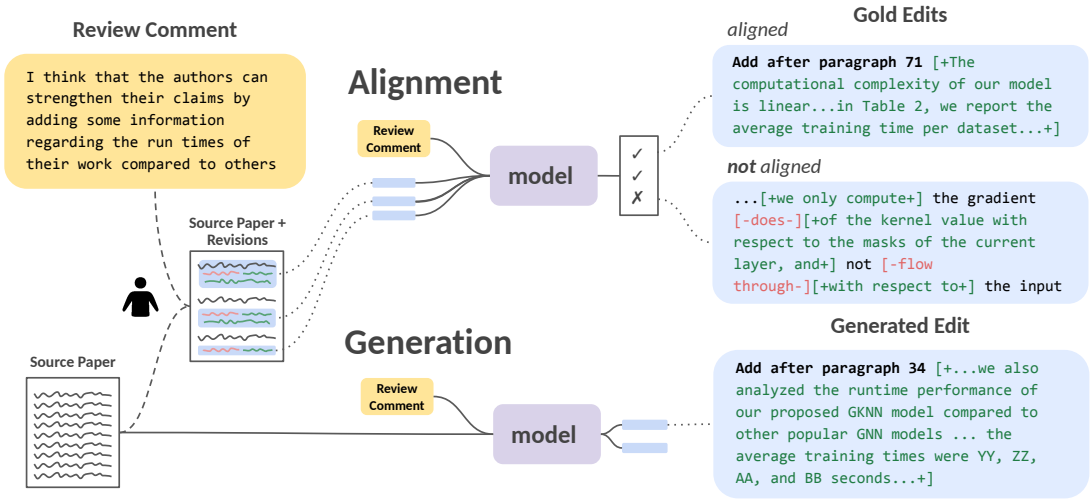

Figure 2.1. Overview of our tasks. In comment-edit alignment, a model is given a review comment and set of candidate edits derived from a source paper and a revised target paper, and it must align the comment to the edit(s) that are associated with it. In edit generation, a model is given a review comment and a source paper and must generate an edit that addresses the comment, possibly using placeholders for missing information.

In this work, we evaluate whether large language models (LLMs) possess the reasoning abilities required to model the relationship between feedback and edits. We release **ARIES** (**A**ligned, **R**eview-**I**nformed **E**dits of **S**cientific Papers), a real-world dataset of computer science paper drafts, the corresponding reviewer feedback, and the author responses and revisions that address the feedback.[1]

Using this dataset, we formulate two novel tasks, shown in Figure 2.1: **comment-edit alignment**, in which a model must determine which review comments made about a paper correspond to each of the edits made after the feedback, and **edit generation**, in which a model must generate edits directly from a given reviewer comment and paper text.

[1]The dataset and code are available at: `https://github.com/allenai/aries`

In addition to serving as challenging testbeds for LLM evaluation, these tasks have the potential to advance assisted reading and writing applications. Automatic alignment could enable tools that allow readers to quickly find parts of a document that address particular questions or comments [12, 23] or that help authors, reviewers, and area chairs more easily track revisions. Edit generation could power collaborative writing tools that allow authors to rapidly iterate on their manuscripts in response to feedback.

We evaluate ten baseline methods and find that the alignment task is challenging for existing models, including even large models such as GPT-4, and that comments and edits with indirect relationships are especially difficult. For the generation task, we find that GPT-4 does produce edits that are coherent and on-topic on a surface level, but fails to model the underlying intent; unlike real authors, it almost never makes edits that suggest the feedback is mistaken, often paraphrases the feedback rather than tightly integrating edits into the context of the paper, and tends to include less technical detail.

In summary, our contributions are as follows:

- We propose the novel tasks of (1) aligning high-level draft feedback to specific edits and (2) generating revisions for scientific papers given reviewer feedback (section 2.3).
- We construct ARIES, a real-world dataset containing 196 human-labeled review comments matched to their corresponding paper edits, as well as 3.9K reviewer comments automatically matched to edits using author responses from OpenReview, with 92% precision (section 2.4).
- We evaluate a wide range of baseline methods on our comment-edit alignment task, finding that it is challenging even for modern LLMs. The best model (GPT-4) achieves only 27.0 micro-F1 compared to human performance of 70.7 (section 2.5).

- We conduct a thorough analysis of edit generation with GPT-4, detailing several systemic differences between generated and real edits, and suggest future work directions (section 2.6).

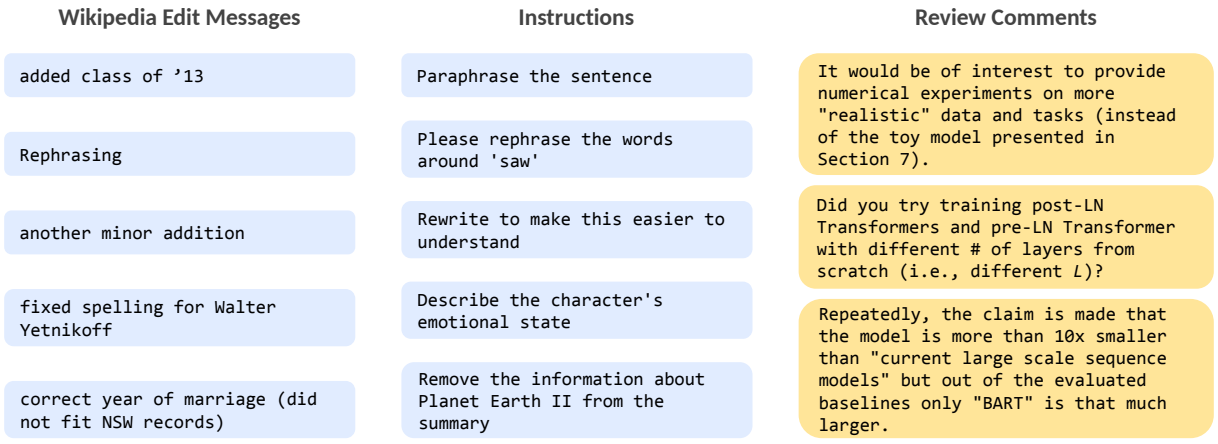

Figure 2.2. Representative examples of the kinds of conditioning information used to guide edits in our work (review comments) compared to previous work which considered Wikipedia edits [17] and author-provided instructions [25, 44, 55, 77]. Review comments are longer and less direct, requiring more knowledge and reasoning to interpret.

2.2. Related work

To our knowledge, our work is the first to study contentful edits conditioned on complex feedback in a highly technical domain (scientific papers). Previous work on edit modeling either focuses on stylistic and grammatical edits or incorporates no feedback or very different kinds of feedback—such as explicit instructions or descriptions of edits created post-hoc. Those settings don’t present the same challenging reasoning requirements as our tasks. Figure 2.2 illustrates how the content and linguistic complexity of review comments differs substantially from that of the conditioning information used in past work.

Style and Grammar Edits. Early work on edit modeling focused on grammatical error correction (GEC), which aims to identify and correct grammatically incorrect or misspelled text, and work in this area dates back several decades [34, 67]. With the increase in language modeling capabilities in recent years, there has been progress in making more sophisticated edits such as rewriting a sentence to improve clarity, style, or structure [30, 46, 47]. However, these areas of research do not target the kinds of substantive revisions often made to papers in response to reviews, such as adding an entire sentence or paragraph to discuss a result or justify a design choice.

Assisted Writing Systems. Several works develop writing assistants that incorporate human input to guide the edits. In some cases the human input is restricted to specific actions, such as marking words that the system should omit [20] or selecting proposed edits to apply [13, 38], while in other cases the user can provide a natural language instruction [25, 44, 55, 58, 77]. However, the kinds of instructions found in these works are different from the draft feedback we investigate in that they are written by humans who know they are interacting with an automated system, resulting in more direct and specific instructions than the open-ended feedback that authors often receive for a draft.

Much of the previous research on edit modeling focuses on Wikipedia, using Wikipedia edit messages as a proxy for instructions when generating edits [17, 57, 60]. Wikipedia edit messages are generally written post-hoc and provide varying levels of information about the content of the edit, often giving only a very vague summary like "add reference". In contrast, review comments generally provide enough information for a human to identify the content of the necessary edit, as in many cases their purpose is in part to guide the authors' revisions.

Lee and Webster [37] create a corpus of essays by English-as-a-second-language students with sentences aligned to feedback from teachers and the corresponding revisions. Their task has a similar structure to ours, but in practice the vast majority of the feedback in their data is focused on simple word-level grammatical issues. ArgRewrite [29, 79] is also a dataset of student essay revisions with teacher feedback, and contains some contentful comments, but the essays are very short (~500 words) compared to scientific papers (~5000 words) and the comments are not aligned to specific edits.

Scientific Edits. Some work does explore scientific-domain edits, but these don't associate edits with reviewer comments and often focus on classification rather than generation. Jiang, Xu, and Stevens [27] and Du et al. [14] analyze and tag edit intentions on ArXiv papers but do not use feedback. Du et al. [13] develop a system for human-in-the-loop editing in several domains, including Wikipedia and Arxiv, but the feedback is limited to accepting/rejecting suggested edits, and the focus is on fluency and style edits. Mita et al. [48] construct a dataset and evaluation framework for scientific document revision, and they do consider some document-level revisions such as reordering sentences. Nonetheless, the aim of the revisions is to improve writing quality rather than to alter the semantics of the text, and peer review comments are not used.

Finally, Kuznetsov et al. [35] identify edits between paper versions and separately align reviewer comments to referenced text in the source paper, but do not explore the connection between feedback and edits. We note that linking comments to source text is insufficient to study feedback-based editing due to both spurious edits and our finding in subsection A.1.2 that most feedback-based edits add a new paragraph or section instead of modifying existing text.

2.3. Task Definitions

As shown in Figure 2.1, we consider two versions of the task of determining how a document should be edited to address a given piece of feedback: comment-edit alignment and edit generation. Both tasks express the differences between an original (source) document and revised (target) document as a list of *edits*, where each edit represents a specific change from source text at some location in the paper into new text in the target paper. Specifically, an edit consists of a paragraph in the source and its corresponding revised paragraph in the target, where either paragraph (but not both) can be null in the case of deletions or additions.

In the **comment-edit alignment** task, the goal is to identify the edit(s) that correspond to a given review comment. The input is a comment and a list of edits, which include both original and revised text. In our evaluation, we derive the list of input edits by using a paper's gold revisions, but they could consist of any candidate revisions. The output is a set of binary classifications over the list of edits, indicating whether each edit addresses the comment. Note that this results in a many-to-many mapping; one comment may result in several edits to the paper, and (less commonly in our data) multiple comments may be addressed by one edit.

In the **edit generation** task, the objective is to generate appropriate edits to a paper based on feedback. The input for this task consists of a comment and the original paper text. The output is the generated edit, which should address the reviewer's feedback and be coherent within the context of the paper.

2.4. Dataset Construction

Both the comment-edit alignment and edit generation tasks require a dataset with paper edits aligned to specific feedback comments. In this section, we describe our approach for collecting

and annotating ARIES, a corpus of computer science paper revisions and reviews with both manual and synthetic annotations of comment-edit alignments.

At a high level, the construction process is as follows: First, we obtain a corpus of paper draft PDFs, their peer reviews, and revised drafts from OpenReview (subsection 2.4.1). Next, we manually identify spans in reviews that represent actionable comments (subsection 2.4.2). Then, we manually identify the edits that correspond to each review comment to obtain a small but high-quality dataset for evaluating models (subsection 2.4.3). Finally, we develop a synthetic labeling approach to automatically extract comments and align them to edits using author responses (subsection 2.4.4). This approach results in edits with high precision (but low recall), and with it we create a much larger dataset suitable for training models. Statistics of our final dataset are in Table 2.1.

2.4.1. Collecting papers and reviews

We obtain papers, reviews, and author responses from computer science conferences on OpenReview.[2] For each paper, we use the latest PDF that was uploaded before the first review as the original version and the latest available PDF as the revised version. We omit papers that do not have a revised version uploaded after reviews were posted, resulting in a set of 6,501 paper records. We use Grobid [21] and S2ORC [45] to parse the paper PDFs.

We identify edits between the source and target papers by finding pairs of paragraphs with high bigram overlap. More details can be found in section A.3.

[2] https://openreview.net

On average, a paper revision typically has 40% of its paragraphs unchanged, 14% "minor" edits (with less than 10 tokens changed, usually fixing typos or grammar), 14% "major" edits, 8% fully deleted paragraphs, and 23% fully new paragraphs.

Statistic	**Manual**	**Synthetic**
Papers	42	1678
Comments	196	3892
Aligned Edits	131	3184

Table 2.1. Statistics for manually- and synthetically-labeled data. Papers, reviews, and aligned edits are counted only when they correspond to included comments. Edits are counted only once, even if they correspond to multiple comments.

2.4.2. Identifying Actionable Feedback

To create our manually-annotated evaluation data (196 instances), we first extract sentences from reviews which constitute actionable feedback. We define *actionable feedback* as feedback that states or implies a specific request that could be addressed by an edit to the paper. Reviews generally consist of a summary of the paper in question, some comments on the strengths of the work, the weaknesses of the work (which may include some specific suggestions for improvement), and an overall opinion of whether the paper should be accepted or rejected. In this work we care primarily about the weaknesses and suggestions, although actionable feedback can sometimes appear elsewhere. Actionable feedback can be phrased in a wide variety of ways, including as questions or as implicitly negative remarks. However, a positive comment (*"The paper is sound and of certain interest"*) or one that simply summarizes the paper is *not* considered actionable for our purposes.

Two annotators manually annotated 42 reviews to extract the token spans corresponding to actionable feedback (details in section A.4), ultimately resulting in 196 comments. In some cases, a comment might only make sense in the context of some other sentence from the review. For example, in *"The paper is missing several things: (1) a definition of L, (2) ImageNet baseline, (3) ..."*, the phrase "ImageNet baseline" is only interpretable in the context of the top-level comment. Where this occurs (9% of comments), we annotate both the context and comment spans and concatenate them into a single comment.

Inter-annotator agreement was measured on a set of 10 reviews that were annotated by both annotators, with a total of 60 non-overlapping spans between the two annotators. We find that 88% of spans overlap between annotators, but due to differences in amounts of included context the token-level Jaccard overlap is 65%. In subsection A.1.1, we conduct further analysis on the types of actionable review comments in our extracted data.

2.4.3. Aligning Comments to Edits

The extracted actionable comments (subsection 2.4.2) were mapped to their corresponding edits in the paper by an expert annotator (the author of this dissertation). For each comment, the annotator was given the original and revised paper PDFs and the list of edits and asked to identify which edits were made in response to the comment. As additional context, the annotator was given the responses authors made to the reviewers on the OpenReview forum to assist with finding all of the intended edits, as authors often state in their response where they made edits to address each point made by the reviewer. Agreement was calculated against a second annotator on a sample of 25 comments, obtaining a Cohen's κ of 0.8.

In total, 78% of comments were addressed by the authors. However, 28% were addressed only in the author response and not with edits to the paper, and 7% were addressed in the paper but not visible in the parsed text (either because of a parsing error, or because the edit was purely visual, such as changing a figure), leaving 43% (85 comments) aligned to textual edits (the comments without edits are still included as challenging examples for our comment-edit alignment task). The aligned comments each correspond to 2.1 edits on average.

2.4.4. Creating Synthetic Data

To produce a large training set with high-quality comment-edit alignments, manual annotation is not feasible; each review takes approximately 30 minutes to fully process and requires annotators with extensive domain expertise, and our corpus contains 24k reviews. Thus, we automatically generate a large silver dataset of comment-edit alignments by leveraging the fact that authors often quote reviewer comments directly in author responses, and the edits that correspond to a comment are often highly similar to the author response text discussing the comment.

We automatically identify the quoted review comments in author responses by searching for lines with a small edit distance to a contiguous span of review text (with a minimum length of 40 characters, to eliminate spurious matches). The corresponding response text for each comment is matched to edits with high textual overlap; we informally observe that edits with at least 25% bigram overlap to the response text almost always correspond to the quoted comment. Using this threshold, we link responses and edits to obtain a set of 3892 high-precision alignments from the training corpus.

Unlike the manually-annotated data, the synthetic data has low recall; applying the synthetic labeling algorithm to our hand-labeled data identifies only 2% of the matches. However, they

have high precision: We manually checked 50 sampled alignments and found that 46 were correct. Furthermore, we find that the synthetically-aligned data has similar statistics to the manually-annotated data; see subsection A.1.3 for details.

2.5. Comment-Edit Alignment

In this section, we evaluate models on the comment-edit alignment task using our constructed dataset. As described in section 2.3, the comment-edit alignment task is a binary classification task where the input is review comment and a list of candidate edits, and the output is binary for each comment-edit pair, specifying whether the comment and edit are aligned. In model inputs, edits are textually represented using a "diff" format with additions and deletions enclosed in `[+ +]` and `[- -]` brackets, respectively.

For manually-annotated data, for a given comment, we consider all edits for the corresponding paper as candidate edits, labeled as positive if the edit was annotated as addressing the comment and negative otherwise. Given the low recall of the synthetic data (discussed in subsection 2.4.4), we can only use the synthetic labels to produce positive comment-edit alignment pairs; thus, we pair comments with edits sampled from other documents as negative candidates. Additional details are provided in section A.5.

2.5.1. Models

We consider four kinds of model architectures, detailed below. For all models that produce similarity scores or probability outputs, we tune a decision threshold on the dev set to maximize micro-F1. In addition, we use a version of BM25 tuned for high recall (>90%) on the dev set as

a first-pass candidate filter for the GPT-4 based methods, which increases evaluation speed and reduces GPT-4 API costs.

Bi-encoder: The model separately consumes each review comment and edit to create an embedding for each, with a goal that embeddings for corresponding comments and edits are closer to each other than those for non-corresponding pairs are. We prefix the comments with "review comment:" and the edits with "edit:" to allow the model to treat the two text types differently. For fine-tuning, we use a triplet loss; given a triplet consisting of a comment c, a positive edit x_+, and a negative edit x_-, the loss is

$$\mathcal{L} = \max(0, \text{sim}(c, x_-) - \text{sim}(c, x_+) + 0.5)$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity.[3]

The bi-encoder models we use are DeBERTaV3-large [22] and SPECTER2 (base) [61]. For SPECTER2, we also include a non-finetuned variant, as the pretrained weights are already designed to produce good scientific text representations.

Pairwise cross-encoder: The model consumes a comment-edit pair separated by a `[SEP]` token and outputs a score representing the likelihood of a positive label. DeBERTaV3-large [22], LinkBERT [76], and GPT-4 [50] models are used with this format. For GPT-4, we try both a zero-shot setting where only instructions are given and a (2-way) one-shot setting where one positive and one negative example are given in the prompt.

Multi-edit cross-encoder: The model consumes all edits for a paper at once, including unchanged paragraphs as "edits" for context; in essence, this is a full "diff" of the paper with an edit ID number attached to each paragraph. We additionally feed all comments for a paper at

[3]This loss is similar to the one used to train the SPECTER2 base model we use in our experiments, although we found cosine similarity to work slightly better than Euclidean distance in our preliminary experiments.

once, each with a unique ID. The output is formatted as a list of JSON objects, each containing a comment ID and a list of edit IDs. In practice, a diff of the full paper is often too long to fit model length limitations, and in these cases we split the paper into 2-3 chunks and merge the output lists. We use GPT-4 [50] for this variant, with a maximum input size of 7,500 tokens (the maximum total length is 8,192, and we allow roughly 700 tokens for the response).[4]

Bag of words: We try a simple BM25 ranker [59] that scores a comment against the post-revision text of an edit. As an additional baseline, we apply BM25 using generated edits from GPT-4 (discussed in section 2.6) and refer to this as BM25-generated. As we show in section 2.6, GPT-generated edits are competitive with human edits in terms of the overall comprehensiveness with which they address comments, but they also differ substantially from human edits in style and content. The BM25-generated baseline serves as a way to empirically probe the similarity of the two kinds of edits.

Human: As a strong baseline, we evaluate how well an expert human annotator can perform on this task given the same inputs as the models. That is, the human is shown a comment and a full diff of the parsed source and target papers, but—unlike the annotators who labeled the task data—does not have access to author responses with which to identify unintuitive responses or to the PDFs with which to identify parsing errors.

	Micro				Macro			
Model	**AO-F1**	**P**	**R**	**F1**	**AO-F1**	**P**	**R**	**F1**
BM25	13.3	12.2	10.5	11.3	77.1	73.8	62.4	43.8
BM25-generated	14.7	4.6	40.3	8.3	50.7	7.6	80.3	9.6
Specter2 (no finetuning)	14.0	8.1	14.4	10.3	68.6	63.0	62.8	39.9
Specter2 bi-encoder	19.6	17.0	29.3	21.5	67.8	55.5	70.5	38.5
DeBERTa bi-encoder	3.1	9.9	12.2	10.8	72.6	47.5	61.8	31.9
LinkBERT cross-encoder	2.8	10.1	28.4	14.4	71.3	39.2	70.8	26.8
DeBERTa cross-encoder	8.5	7.4	25.6	10.0	70.9	30.2	71.5	22.5
GPT-4 cross-encoder 0-shot	38.7	-	-	-	70.6	-	-	-
GPT-4 cross-encoder 1-shot	42.1	-	-	-	74.8	-	-	-
GPT-4 multi-edit	36.2	24.2	30.4	27.0	74.6	62.0	70.4	46.2
Human	70.6	65.6	76.8	70.7	89.2	92.7	86.2	82.7

Table 2.2. Precision (P), Recall (R), and F1 of comment-edit alignment on test data (the manually-annotated set). The micro-average is over all comment-edit pairs, while the macro-average is grouped by comment. Addition-Only F1 (AO-F1) is the F1 score when only addition-only edits are considered; due to budget constraints, this is the only feasible setting for pairwise cross-encoder GPT. Overall, GPT-4 methods are all much better than the smaller locally-trained models, but none reach human performance.

2.5.2. Results

Table 2.2 reports precision, recall, and F1 scores for models. The micro- scores are computed over all comment-edit pairs, while the macro- scores are macro-averaged by comment[5] to down-weight cases where a model incorrectly predicts many edits for one difficult comment. In addition to results over the full dataset, we also run experiments on just edits that add a full paragraph as addition-only F1 (AO-F1); this setting is easier because it does not require models to understand which tokens have been added, removed, or unchanged, and is a better fit for BM25, which

[4]OpenAI has indicated plans for a 32k-sized model, but that has not been released as of this work.

[5]Implementation note: F1 is considered 100 for comments where the model correctly predicts that there are no corresponding edits.

cannot represent the differences between these tokens. Results are averaged over three trials with different random seeds for training. The prompt templates used for GPT-4 can be found in section A.2.

We find that task is challenging, with none of the models reaching human-level performance. GPT-4 methods are best, but interestingly it appears that giving GPT-4 a full chunk of the document at once (GPT-4 multi-edit) results in slightly worse performance than the pairwise approach, aside from an improvement in efficiency.

For LinkBERT and DeBERTa, we surprisingly find poor micro-AO-F1 performance; it appears that the models sometimes assign similar scores to several instances, making it likely that the decision threshold on the dev set will be suboptimal. Nonetheless, the models can still obtain good macro-AO-F1 scores, and this issue is far less prevalent on the full dataset results.

For DeBERTa, we find that the cross-encoder and bi-encoder variants have similar performance. However, the Specter-based bi-encoder substantially outperforms both DeBERTa and LinkBERT cross-encoders, which is especially notable because Specter has only about a quarter of the parameters of those models. We conjecture that Specter's pretraining makes it an especially good fit for this task; the citation prediction objective it pretrains on, which constrains papers that cite each other to have similar embeddings, is similar to the comment-edit alignment task in that two texts may be "similar" for purposes of the task even if they are semantically and syntactically very different.

The results of BM25-generated indicate that using generated edits as inputs provides only a slight improvement to micro-AO-F1, and actually worsens macro-AO-F1 (although the harm to macro-F1 may be amplified by the fact that the decision threshold is tuned on micro-F1). This suggests that the differences in style and content between GPT-4 and human generated edits are

large enough to prevent effective alignment despite GPT's edits appearing plausible in many cases. We discuss the differences in more detail in section 2.6.

Across all methods, including human performance, we observe that macro-F1 is substantially higher than micro-F1, suggesting that some comments are especially error-prone. For example, 55% of GPT-4 multi-edit's errors correspond to just 20% of the comments. Nuanced comments on documents with many edits may lead to several incorrect predictions—*e.g.,* if they involve many small changes to technical details and equations—whereas other instances may be more straightforward. In the next section, we analyze specific failure modes that we observe.

2.5.3. False Positives

We manually examined 50 randomly-sampled false positives of the best-performing model, GPT-4 multi-edit, and identified four common categories of mistakes that it makes. The categories and their frequencies are described in the following paragraphs. Note that the categories are partially overlapping, so the total exceeds 100%, and 10% of the errors did not fit clearly into any category.

Too-Topical (40%). In some cases, the model assigns a positive label to an edit that contains some words that are topically or lexically similar to the words in the review comment, but do not actually address the comment. In many cases, this happens even when the words are only part of the original text and were not added or deleted in the edit.

Diff-ignorance (28%). In some cases, a comment asks about something that is already present in the paper in some form—*e.g.,"add CIFAR10 experiments"* when there is already one CIFAR10 experiment in the paper, or asking to remove a misleading claim. The model sometimes aligns

these comments to edits of paragraphs with preexisting (or deleted) content that is relevant to the comment, failing to correctly account for the add/delete markup.

Over-Generation (28%). This failure mode is unique to the multi-edit task format, in which models attempt to generate a full list of all comment-edit alignments for a paper in one pass. We observe some cases where GPT-4 outputs more consecutive edits in a list than it should; for example, if edits 17 and 18 are relevant to some comment, the model might add 19, 20, 21, 22 and so on. In rare cases, the list extends beyond the highest edit id in the input. Although it is difficult to precisely determine the factors that influence GPT-4's output, we hypothesize that GPT-4 may be suffering in part from exposure bias, and as it begins to generate a consecutive sequence it gets stuck in a loop and fails to stop at the correct place. This phenomenon has previously been studied in smaller models [10], and may be occurring to a much lesser degree with GPT-4 as well.

Bad Parsing (12%). Some errors are simply the result of the PDF parser extracting text differently for different versions of a paper, causing text to appear edited when it was not. In some of these cases, the "edits" in question do look as though they partially address the comment, similar to the errors in the "diff-ignorance" category, and the model erroneously (albeit reasonably) aligns to those edits without realizing they were already in the original paper.

2.5.4. False Negatives

Similarly to how many false positives arise when an edit uses terms similar to the ones the reviewer used in their comment, we observe that false negatives often occur when there is *low* overlap between the language of the comment and the edit. For example, a comment may ask how a method was implemented and the corresponding edit adds a link to a code release, or

a comment asks for a proof and the corresponding edit adds an equation. In such cases the model must understand e.g. that adding a link to code is a way of addressing a request for implementation details.

We attempt to quantify how the explicitness of the relationship between a comment and edit affects alignment performance. We leverage two metrics: The first is a measure of **edit compliance**: Specifically, we annotate how directly an edit obeys a given comment on a 1-3 scale (1 being least compliant, 3 being most compliant). More details on the metric and compliance annotations are in section 2.6. The second is a measure of **comment directness**: how "direct" or "indirect" the comments are. A direct comment is one that indicates a clear action; this could be phrased in the negative, but still explicitly specifies what needs to be done (*"It is unfortunate that you didn't [do experiments on imagenet]"*). An indirect comment does not state the specific request, and is usually a statement of fact or observation that requires an understanding of linguistic and scientific norms to interpret (*"Only one dataset was used"*).

We measure the performance impact of indirectness and compliance on the multi-edit GPT-4 method in Table 2.3, and we find that both factors result in a substantial difference. GPT's micro-F1 is 30% lower on indirect comments compared to direct ones, and 24% lower when edits are non-compliant. These results suggest that GPT-4 struggles to understand complex comment-edit interactions and performs better on those with simple, direct relationships.

2.6. Edit Generation

In this section, we explore the edit generation task introduced in section 2.4.

	GPT-4	Human
Direct comment	40.4	78.6
Indirect comment	28.2	61.3
Compliance = 3	39.5	71.5
Compliance < 3	30.1	77.3

Table 2.3. Alignment micro-F1 for GPT and humans on direct/indirect comments and compliant/non-compliant edits. Note that the values are higher than in Table 2.2 because comments with no corresponding edits were not annotated. GPT and humans both do much worse with indirectly-phrased comments than direct ones. GPT also struggles to match to non-compliant edits, whereas humans are unaffected.

2.6.1. Experimental Setup

Our goal is to understand the differences in style and content between the kinds of edits human authors write and those that models generate, which will provide insight into model behavior and point to directions for future improvements. However, we note that evaluating the *correctness* of generated edits is beyond the scope of our analysis, as it is both difficult to do fairly (models may lack information such as lab notes and raw data) and difficult to do correctly (robust judgements require a very deep expertise in a given paper). Nonetheless, in our preliminary analysis we observed that almost all model-generated edits would appear plausible to a reader with only cursory knowledge of the paper (the title and abstract).

We generate edits with GPT-4, which was the best model for comment-edit alignment and is known to be a powerful general-purpose generation model [50]. Note that this model cannot be fine-tuned as of this work, so we use prompting to instruct the model to do the task. The prompt template is provided in subsection A.2.4.

2.6.2. Manual Analysis

We explore the differences between GPT-written and author-written edits more deeply with an analysis by two expert judges (with multiple CS/ML publications) on 85 comments. The comments were divided between the two judges, except for 10 instances that were annotated by both in order to measure agreement. Each instance includes the original paper, the review comment, and both GPT's generated edits and the set of real edits that were made to the paper in response to the comment. The judges are aware of which edits are model-generated and which are real, as it would be impossible to conceal the stylistic differences; however, we do not believe this impacts our goal of understanding the trends between the two edit types, as the judges scored edits using several specific factors described in the following rubric. Examples of these factors can be found in Table 2.4:

- **Compliance (1-3):** The edit might argue that the comment is irrelevant or infeasible to address (1), address the spirit of the comment but not specifically what was asked (2), or directly comply with the reviewer's advice (3).
- **Promises (true/false):** The edit promises to address part of the comment in future work or a future revision; we include cases where the model says it provides information elsewhere (e.g., in its Appendix) but does not give the corresponding edit for that section.
- **Paraphrases (true/false):** The edit reuses the wording from the comment itself.
- **Technical details (true/false):** The edit contains specific details or placeholders for details such as citations, mathematical expressions, or numerical results.

Factor	Comment	Edit
Compliance=1	... Isn't this percentage too much? Can't we use, e.g., 5% of all nodes for training?	[+... our split of 80% -10% -10% is a standard split+]
Compliance=2	... there is a hyperprameter in the radius decay, how it will affect the performance is crucial ...	[+... this learnable radius is not effective the in terms of an classification performance compared to that the predefined radius decay+]
Compliance=3	the experimental setup requires significantly more details on the hardware ...	[+We conducted our experiments using NVIDIA Tesla V100 GPUs ...+]*
Promises	it would be interesting to know how the proposed method would work, for instance, for node classification (e.g., Cora, Citeseer)	[+... the performance of our method on node classification tasks is beyond the scope of this paper and is left as an interesting direction for future work.+]*
Paraphrases	... it should be investigated ... with respect to more natural perturbations, e.g. noisy input, blurring, ...	[+... we also investigate their performance with respect to more natural perturbations, such as noisy input, blurring, ...+]*
Technical details	... This does put into question whether the full closed loop model is actually useful in practice	[+... we evaluated the performance of a closed-loop N-CODE model ... Here, the control parameters are a matrix of dynamic weights, $\theta(t) \in \mathbb{R}^{m \times m}$ ...+]

Table 2.4. Examples of comment-edit pairs exhibiting each scored factor in the edit generation analysis (subsection 2.6.2). Edits marked with an asterisk (*) are generated by GPT, while the others are real. Text is ellipsized for brevity.

We note that the edit generation task is made technically impossible by the fact that some edits may require information that the model does not have, such as the results of additional experiments. We mitigate this by instructing the model to use placeholders or to hallucinate technical details that it does not know (details in section A.2). In addition, for each comment we measure **answerability**: whether it can be addressed *without* placeholders or halluncinations. In other words, a perfect model should be able to address answerable comments using just the original paper and background knowledge.

Additionally, for each (GPT, real) edit pair, we evaluate which has greater **comprehensiveness** in addressing the reviewer's comment, as there are many cases where one edit is more thorough or goes beyond what the reviewer asked, even though both have the same compliance. This is not the same as correctness; instead, comprehensiveness measures how thoroughly an edit *attempts* to address a comment, possibly using placeholders or hallucinating unavailable information.

2.6.3. Results

	Ans.	**Non-ans.**	**All**
GPT better	31%	19%	25%
Real better	19%	40%	29%
Same	50%	42%	46%
Frequency	51%	49%	100%

Table 2.5. Fraction of the time that a given model's generated edits were deemed more comprehensive (but not necessarily correct), broken down by answerability. The Frequency is the fraction of comments that fall into each category. Overall, GPT generations are comparable to real edits, with GPT being better for comments that don't require additional data and real edits being better for those that do.

From an initial inspection of GPT's generated edits, we find that the model almost always produces coherent and on-topic edits that respond to the given review comments. Table 2.5 shows that GPT-generated edits are competitive with human-authored edits in comprehensiveness, often being rated as more comprehensively addressing the given comment when sufficient information is available but doing worse for comments that require additional data to address. On average, GPT almost matches real edits in this regard.

However, we observe in Table 2.6 that the kinds of edits generated by GPT-4 are very different than those produced by real edits. The most striking difference we observe is the tendency for GPT-4 to paraphrase the comment when writing its edit (48% for GPT-4 vs. 4% for human edits). Qualitatively, we notice that GPT-4's edits are often written as though they are meant to be a standalone response, whereas the real edits are more tightly integrated into the context of the paper. In addition, real edits are more likely to use specific technical details as opposed to a high-level response, an effect which is understated in Table 2.6 due to the cases where both edits contain some technical details but one contains substantially more. To account for these cases, we additionally record relative technicality judgements for each (GPT, real) edit pair and find that the difference grows: the real edits are more technical in 38% of cases compared to only 12% for GPT (p=10^{-3}). Overall, the reduced level of technicality and the tendency to paraphrase may make GPT-4's edits preferable for those who simply want clear responses to their questions and feedback, but they also make edits less informative for the most engaged readers who care about technical details.

	GPT	**Real**	κ	p
Compliance	2.9	2.6	0.6	10^{-4}
Promises	21%	6%	1.0	10^{-2}
Paraphrases	48%	4%	0.7	10^{-11}
Technical details	38%	53%	0.7	0.06

Table 2.6. Edit generation analysis. We report average Compliance and fraction of examples that include each of the other factors. We report Cohen's κ for all factors on 10 instances and report p-values using Wilcoxon's signed-rank test for Compliance and Fisher's exact test for others. GPT is more compliant, often paraphrases the comment directly in its edits, and tends to include fewer technical details than real edits.

We also note that while most edits from both GPT-4 and humans follow the reviewer's specific instructions, human edits deviate from the reviewer's request more often: 94% of GPT-4 edits are highly compliant (compliance = 3), while only 68% of human edits are. The actual discrepancy in this factor may be even higher, as real authors often choose not to make an edit at all when they disagree with a comment, opting instead to discuss it on the OpenReview forum.

The high compliance of the model is not especially surprising given that GPT-4 is trained to follow instructions, but it does have implications for GPT-4's suitability as an editing assistant. Often, the proper edit requires thinking critically about the reviewer's critique rather than simply following it, and GPT-4's output is less suitable in those cases.

2.7. Conclusion and Future work

In this work, we have introduced the novel tasks of comment-edit alignment and edit generation for scientific paper revisions based on high-level draft feedback from reviewers. We have constructed and released a dataset containing pairs of computer science paper drafts with edits aligned at the paragraph level, along with their corresponding reviews and author responses. We hope the dataset will enable research on assisted writing technologies and serve as a challenging testbed for large language models.

It is interesting that models (including GPT-4) do so poorly on the comment-edit alignment task despite GPT being able to generate plausible edits in the generation task. As our analysis shows, the kinds of edits produced by GPT can be very different from the real edits authors make to their papers, and the fact that GPT fails to recognize many of the real comment-edit pairs suggests that it may have gaps in its reasoning that would be interesting to explore further in

future work. We hope that the insights from our analyses can help motivate and guide future studies.

A shortcoming of the generated GPT edits is their relative lack of technical details. However, this may be caused in part by their lack of access to information about detailed experimental results, code, and lab notes for the paper, which the authors have when doing their revisions. As a long-term goal, we believe that an ideal writing assistant would observe the entire research process and consume all relevant information when writing an edit; in some cases, this might even include suggesting additional experiments for humans to run. However, this requires further work both to create applications that can collect this information and to develop efficient methods to provide this information to large language models, which are currently limited in input size and expensive to run.

2.8. Limitations

Our study is limited to scientific papers in English from the field of AI, and future work should consider a broader set of scientific disciplines and languages. Our evaluations are limited to measuring the correctness and types of alignments and generations produced by today's large language models (LLMs); future work should apply the techniques within real assisted writing applications and evaluate their impact on users. We use proprietary LLMs like GPT-4 in certain experiments, and those results may be difficult reproduce if changes to the proprietary services occur.

CHAPTER 3

MARG: Multi-Agent Review Generation for Scientific Papers

3.1. Introduction

In the previous chapter, we constructed a peer review and revision dataset and studied the task of editing papers given peer-review feedback. While it is a challenging and useful task in its own right, the editing task is only one piece of the review-revision process; in particular, it depends on having a human reviewer to carefully read the paper and produce feedback to condition the edits on.

In this chapter, we study the task of generating peer-review feedback for a scientific paper automatically. This task comprises several reasoning challenges: a reviewer must understand the intent and significance of a work, the technical details of the methodology, and the nuances of how an experiment or proof can be claimed to support a particular conclusion. They must then identify the ways in a which a paper does or does not fall short and articulate suggestions for improvement.

Modern large language models (LLMs) face a technical challenge in addition to the reasoning challenges involved in generating reviews: namely, they are limited in the total number of tokens they can effectively reason over at once. As scientific papers can be quite long (thousands or tens of thousands of tokens, in our case), there are many cases in which it is not even possible to provide the whole paper in the model's input. Even for models that technically support large inputs, they often cannot use the full capacity effectively in practice [42, 54].

We propose multi-agent review generation (MARG), a method for generating peer-review feedback by prompting an LLM (GPT-4). We find that by using multiple instances of GPT (hereinafter referred to as "agents"), giving each a portion of the paper, and allowing the agents to communicate with each other, it is possible to generate feedback across the whole paper. We additionally find that by including aspect-specific "expert" GPT agents to separately assist with generating comments on experiments, clarity, and impact, the method can perform significantly better than when having a lone agent attempt to generate all types of feedback at once.

In a user study, MARG generated 4.2 "good" comments per paper (rated by users), whereas a simple baseline of having a single agent generate all comments generated only 1.5 good comments, and a recently proposed method [41] produced only 0.5. In addition, we found that while users perceived the majority of the comments generated by the baselines as being generic, the vast majority (83%) of MARG's comments were rated as specific.

In summary, our contributions are as follows:

- We propose a novel method (MARG) that can generate high-quality peer-review feedback even for papers longer than the context size of the base model.
- We evaluate the quality of our generated feedback against two baselines, using both automatic metrics and a user study. We find that our method outperforms the strongest baseline by 6.1 recall points in the automated evaluation and generates 2.8x as many helpful comments per review in the user study.
- We conduct a thorough analysis of the generated feedback, finding that our proposed method is particularly good at generating specific comments, but offers little improvement in accuracy.

3.2. Related work

3.2.1. Review generation

There has been a variety of work that aims to score or improve papers in specific aspects, such as checking statistical tests [49], plagiarism detection [28], citation recommendation [1], and review score prediction [5, 9], among others [33]. While these are useful tools, they are limited in scope compared to the breadth of feedback authors receive from a real review; our work aims to produce free-form textual review comments across a variety of aspects.

Past work on automatic review generation primarily does so using (relatively) small models that cannot consume the full text of a paper [78] or use template-filling instead of generating nuanced free-form comments [65]. More recent work has explored using GPT-4 [50] to verify author checklists [43], but this limits the variety in generated comment types.

Impressona [8] is an editor that allows writers to create AI personas (via GPT-4) to write comments on their work; this is valuable for personalization of feedback, but doesn't focus on finding good techniques and prompts for scientific review generation, and doesn't explore LM-LM interactions; as we show, a simple prompt (akin to what a user might try initially) does poorly on our task compared to our method.

Contemporaneously with our work, Liang et al. [41] conducted a large user study of review generation using GPT-4, finding that GPT-4 could generate helpful review comments. However, that work simply truncated long papers and did not attempt to address the input size limitations of GPT-4. In addition, they used a single prompt rather than attempting to construct specialized prompts and "experts" for different comment types, as we do. We compare our proposed method

to that of Liang et al. [41] and find that while their approach is more efficient, ours produces more helpful comments.

3.2.2. Multi-agent modeling

In games and robotics tasks, where there are often distinct roles being performed or multiple physical agents operating in the same environment, various problem-solving algorithms and reinforcement learning techniques have been studied to enable cooperation between agents [51, 80]. Not all of these use communication for cooperation, and those that do typically exchange symbols or vectors rather than natural-language messages.

Recent work has explored multi-persona interaction with prompted LLMs to simulate artificial societies [39, 52] and to improve reasoning abilities [15, 68], but this work does not explore the use of multi-agent modeling to scale input size limits and does not investigate their potential for highly technical tasks like scientific review generation.

Contemporaneously with our work, Hong et al. [24] and Wu et al. [72] have proposed general frameworks for multi-agent modeling with large language models such as GPT. Wang et al. [68] has also proposed multi-persona collaboration as a way to improve LLM creativity, although they do not investigate the ability of multi-agent modeling to scale input size limits. However, none of these works explore review generation applications.

3.2.3. LLM context management

One effect of multi-agent modeling is to circumvent the input-size limitations of LLMs, which are often prohibitive for long documents. A variety of other techniques have been investigated in prior work.

Several works have proposed methods for modifying LLM architectures in order to increase the effective input size by using alternative attention formulations [7, 26, 32, 66] or incorporating memory retrieval [73]. However, architecture changes often cannot be applied without retraining models from scratch, and powerful LLMs such as GPT are sometimes available only through a fixed API that does not allow low-level model modifications. This motivates us to explore techniques that can be applied without changing the underlying model.

Recently, there has been work exploring context management in LLMs by having models summarize a large input one chunk at a time and then operate on the concatenation of the summaries [71], recursively summarize their input/output history to compress it [64], or incorporate retrieval [4, 74]. These strategies are effective when only part of the input is needed or when it is clear in advance what details will be important; however, in our review generation task, a paper's shortcomings may involve nuanced details that would be lost with extraction or summarization techniques, so we divide the input among multiple agents that collectively retain the full text throughout the task.

3.3. Task definition

We formulate our task as follows: given a scientific paper, generate a list of *actionable* feedback comments that could help authors to improve the paper. Actionable feedback is defined the same way as in subsection 2.4.2; that is, we focus on suggestions and criticism rather than positive feedback. In addition, we focus on substantive comments rather than simple grammatical or stylistic errors.

In both our multi-agent approach and our simple baseline, a paper is split into chunks of text so that each chunk can fit into the model's input. The splits are made on paragraph boundaries to

avoid breaking sentences, and when presenting the text to the model we annotate each paragraph with its position in the paper (paragraph 1, 2, 3, etc) and the name of the section it appears in.

We note that the input format we use does not include figures or tables (as GPT-4 is a pure language model,[1] it cannot consume this information), and many equations are garbled or incomplete due to parsing limitations. Nonetheless, we expect that many comments can be identified from the text alone, as the main conclusions from tables and figures are often stated in text.

3.4. Multi-agent review generation

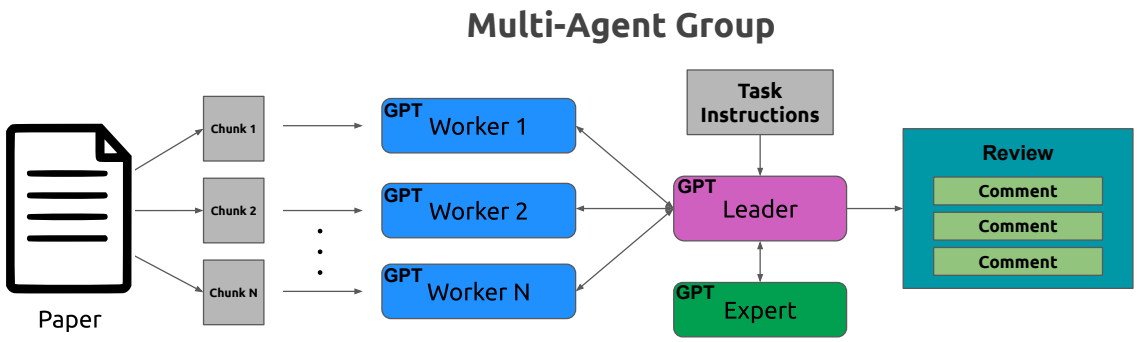

Figure 3.1. Overview of our multi-agent architecture.

In this section, we describe our proposed multi-agent method for generating peer-review feedback, which we call MARG-S (**M**ulti-**A**gent **R**eview **G**eneration with **S**pecialized Agents). At a high level, our multi-agent architecture is formulated as follows: We define an *agent* as one instance of a chat-based LLM (ChatGPT, in our case); each agent has its own chat history and prompt(s). We initialize a set of agents, including three distinct types: (1) a **leader** agent, which is in charge of coordinating the task and the communication among agents, (2) one or more

[1]OpenAI has announced that a vision-enabled version of GPT-4 is being privately tested, but this was not available to us at the time of this work.

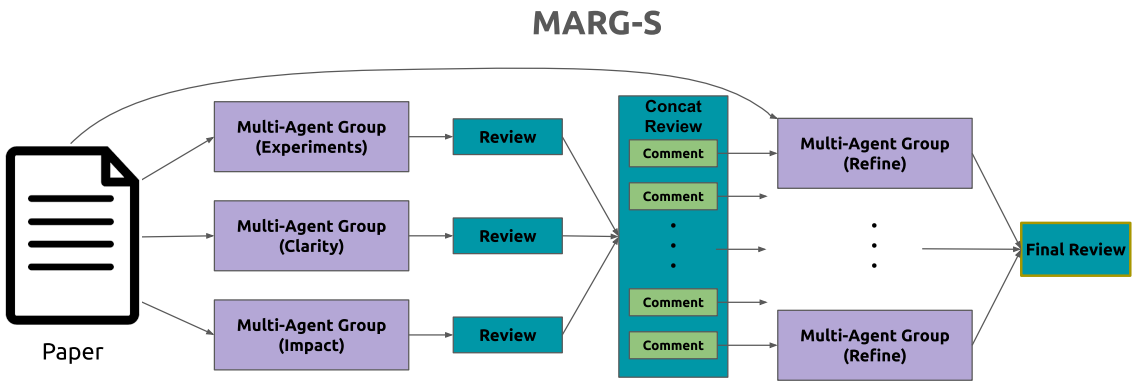

Figure 3.2. Overview of MARG-S, which consists of several specialized multi-agent groups. The comments from each group are concatenated to produce the overall review, and each comment is refined (and potentially pruned) by an additional multi-agent group to produce the final review.

worker agents, which each receive a chunk of the task data (the paper), and (3) zero or more **expert** agents, which are prompted to specialize in some sub-task that serves to assist the leader agent in performing the task effectively. The leader agent is given a protocol with which it can broadcast a message to all other agents and receive responses. Finally, the leader agent is given the task instructions, and must send messages to other agents in order to obtain information and delegate sub-tasks in order to produce the final output.

3.4.1. Agents

Chat-based LLMs, such as ChatGPT, take an input that consists of a list of messages. Each message consists of the message content and the "role" that the message is from, which in ChatGPT's case can be either the "system", the "user", or the "assistant" (i.e., generated by the LLM). Typically, an input to the model starts with a "system" message that describes general instructions that it must follow (e.g., "always give concise and helpful answers"), then the "user" writes a message ("summarize the following passage: ..."), and the generated response is treated

as an "assistant" message. The message history serves as a form of context management; with it, agents can use information from previous interactions in the conversation when formulating future responses.

We use the "system" message at the start of a message history to give unique instructions to each agent type. For example, the "leader" agent is told that it is the leader, that it must coordinate other agents to complete the user's requests, and that it can communicate by using a special `"SEND MESSAGE"` command to broadcast messages to other agents. It is also given some guidelines to improve its reasoning; for example, it is instructed to create a high-level plan from its task instructions before it begins communicating and performing sub-tasks. The "worker" agents are told that they must obey instructions from the leader agent, and "experts" are given special instructions depending on the sub-task they need to perform.

Despite their name, expert agents do not actually have more information or expertise than any other agent. Rather, they are given a special prompt that is designed to encourage them to specialize in a particular sub-task. For example, an expert agent that is asked to focus on experiments and evaluation is given a prompt that encourages it to think about the kinds of experiments that it would expect to see in order to support a particular claim, and then compare those hypothesized experiments to the real experiments in the paper. We found in preliminary testing that giving such instructions to the leader agent tends to work poorly and often ignores details of the instructions, as though the model is overloaded by the number of instructions it is trying to follow, while refactoring the subtask to the separate expert model produces a much higher-quality result.

All agents are given some information about the agent group; they are told how many agents are in the group and the IDs of the agents (while the IDs are not directly used in the

communication protocol, they are useful for internal chain-of-thought; for example, the leader might note that it needs to follow up with a particular agent). In our setting there is always exactly one leader agent, N worker agents for a paper with N chunks, and zero or more expert agents.

3.4.2. Communication

The leader agent is able to send messages to other agents by outputting a special string (`"SEND MESSAGE:"`) followed by the message content. The message is then broadcast to all other agents in the group. When an agent receives a message, it is appended to the history as a "system" message with the header "Message from <agent id>:" preceding the message. The LLM is then run to generate a response to the received message, and this response is always treated as a reply to the leader agent. Replies from all agents are added to the message history of the leader agent before generating the next output from the leader.

When the leader agent generates an output that does not send a message (and thus does not seek any additional information), the task is complete and we prompt the agent to return the final answer.

Error correction. We attempt to correct a few common errors that occur in practice when agents try to use the communication protocol. In some cases, agents get stuck in a loop, often when the task is complete. For example, the leader agent might send a message saying "Thank you all for your feedback and cooperation.", the workers respond with "You're welcome, Agent 0.", the leader says "Thank you all for your responses.", and this loop of unending pleasantries continues. Such loops typically devolve into exactly the same messages being sent repeatedly, so

we check if a message is ever duplicated and if so, we interject with a user message indicating that the message has been duplicated and that it should not be sent again.

We also observe from preliminary experiments that the leader agent does not always remember to follow the protocol for sending a message and simply writes the message body without the necessary header, especially as the conversation grows longer. This is mitigated by including a short reminder every time the leader agent receives messages, reminding it that it must use the appropriate protocol if it wants to respond.

Finally in some cases the leader agent explicitly addresses a message to one agent (e.g., the expert), but that agent does not recognize the message as being addressed to them. To identify such cases and speed up inference, we add a prompt instruction with a specific string an agent should output if they wish to not respond to a message. We detect the presence of any agent ID in a sent message, and if the agent in question outputs the no-response string, we inject a follow-up message reminding them that their name is in the message and encouraging them to respond.

3.4.3. Context management

For documents that result in long exchanges between agents, it is possible for the message history to eventually exceed the input token limit of the LLM. To mitigate this, we prune old messages from the history on each round of communication. The pruning strategy is different depending on the agent type.

History length is most limited for the worker agents, which each have a paper chunk occupying most of their token limit, so the histories for workers were trimmed to the initial prompts plus the three most recent messages. For the leader agent, we observe that (1) a long history is sometimes necessary for in-depth discussions, (2) the majority of tokens in the history arise

from all the messages it receives from (potentially many) other agents, and (3) as the leader relays information between other agents, it generally summarizes any important information from messages it receives. We therefore prune the past messages received from other agents, but keep the full history of outgoing messages. Finally, for expert agents we never observed issues with the token limit, so no pruning was applied.

3.4.4. Review generation

To tune prompts for review generation, we performed several hundred rounds of manual iteration on a small set of papers from ARIES. As the review generation task is somewhat subjective and there are a large number of potential shortcomings with different levels of severity, it is not always straightforward to determine whether a model has made a clear error or if it simply has a difference of opinion with respect to what the most important comments are. We found it helpful to manually alter some of the papers to create severe and obvious errors that we could expect the model to identify; for example, removing an entire section or adding an unfounded claim (e.g., *"the proposed method achieves artificial general intelligence"*). Surprisingly, these "obvious" errors were often not trivial for the system to recognize, making the altered papers useful for finding and mitigating blind-spots. The final prompts are shown in section B.1, and an outline of our system structure is described in the following paragraphs and shown in Figure 3.2.

We use three independent multi-agent groups to generate different kinds of review comments. The task prompt given to the leader agent is different for each comment type, and each group has one expert. The comment types are based loosely on points in the ICLR reviewer guidelines.[2] In particular, it asks *"[...] is the submission clear, technically correct, experimentally rigorous,*

[2] `https://iclr.cc/Conferences/2023/ReviewerGuide`

reproducible, does it present novel findings (e.g. theoretically, algorithmically, etc.)?" We group and slightly reframe these points to arrive at the following comment types:

Experiments and evaluation: The leader is instructed to focus on verifying that the experiments and theoretical proofs are correct and adequately support the paper's claims. The expert in this group is told to "design high-quality experiments" given the main claims made in the paper, inspired in part by the fact that making predictions is an effective active reading strategy to improve comprehension in humans [16, 18]. In preliminary experiments without the expert, the model could identify some bad experiments and give generic comments, but struggled to realize when an experiment was missing. Explicitly designing experiments provides a baseline with which to compare the experiments in the paper, allowing the model to recognize missing or incomplete experiments.

Clarity and reproducibility: The leader is instructed to focus on ensuring that the paper clearly explains key concepts and proposed methods, and that it provides all necessary details to implement any proposed methods and reproduce experiments. The expert in this case is instructed to be "highly curious" and to ask questions of the leader agent in order to learn more about the paper. This process aids in identifying any questions that *can't* be answered based on the paper, which become comments.

Novelty and impact: The leader is instructed to focus on the novelty and impact of the paper. However, we note that for our study the task of accurately retrieving related work is out of scope, so this comment type is limited to identifying errors in the paper's own explanations. Specifically, the model is instructed to verify that the paper clearly states and justifies its motivations, goals, and key findings, and that it thoroughly discusses how it fits into the existing literature. The

expert in this case is instructed to be skeptical of the paper and ask questions to determine if it actually makes a significant contribution to its field.

3.4.5. Refinement

After generating a review, we find that it is very helpful to include a "refinement" stage, in which the model is given the review comments and asked to improve (or remove) them. Various errors can arise during the initial comment generation, and we observe that models tend to be poor at self-reflection and correction during that stage. Including refinement as a separate stage can resolve many of the errors introduced during the initial generation.

To refine comments, we initialize a new multi-agent group with no expert agent. For each comment, we provide the comment to the leader agent with a prompt instructing it to ensure that the comment is clear, that it is specific, and that it is valid (i.e., does not suggest something that is already done in the paper). The model outputs a list; usually this list contains one element (the newly-refined comment), but may contain more (if the original comment mixed two different suggestions) or be null (if the comment was invalid). The comments are processed independently (i.e., by separate multi-agent groups).

3.5. Baseline methods

In this section, we will describe the baseline methods that we compare against our multi-agent approach. We consider three baselines: a simple baseline that treats chunks independently and uses a one-line prompt, a baseline that treats paper chunks independently but uses a more sophisticated prompt, and a recently proposed method for generating peer-review feedback [41]. Prompts for these methods can be found in section B.1.

3.5.1. Single-Agent Review Generation with Basic prompt (SARG-B)

This baseline is designed to emulate a simple approach that a ChatGPT user might use to get feedback on their paper if they did not want to do any prompt tuning. We use a single agent to generate all comments for the paper. The paper is split into the same chunks as for the multi-agent baseline, but the chunks are processed independently using a very simple prompt:

```
Write feedback comments in the style of a scientific paper review for the
following portion of a scientific paper.  You can skip minor grammar
comments.
```

After applying the model to each chunk, the resulting comment lists are combined by a similarly simple prompt:

```
Here are some lists of review comments that were made about different
portions of the paper: <comment lists>
Merge these lists into a final list of review comments.  Any comments that
are duplicates (saying essentially the same thing as other comments) should
be merged or deleted.
```

3.5.2. Single-Agent Review Generation with Tuned Prompt (SARG-TP)

This baseline is designed to emulate a more sophisticated approach that a ChatGPT user might use to get feedback on their paper if they were willing to do some prompt tuning. We use a single agent to generate all comments for the paper, but we use a more sophisticated prompt (subsection B.1.3) that is designed to encourage the model to generate more specific and actionable comments. As with the other simple baseline, we generate comments independently for each paper chunk and then merge the resulting lists with GPT.

Similarly to our multi-agent method, we include a refinement step in this baseline. For each paper chunk, we give the model the chunk and the final list of comments, and ask it to output a new, refined list of comments. This provides an opportunity to remove incorrect comments

that arise from the independent processing of each chunk. For example, if one chunk contains the introduction but not the experiments, the model might initially write a comment that claims the experiments are missing, but in the refinement stage will be able to prune it when it sees the chunk that does contain experiments.

3.5.3. Multi-Agent Review Generation with Tuned Prompt (MARG-TP)

This baseline is designed to provide a direct comparison with the prompt-tuned single-agent baseline and explore the benefits of multi-agent modeling. Whereas our full MARG-S approach leverages several advantages of multi-agent that would be difficult to directly compare in a single-agent setting (e.g., the use of expert agents), this multi-agent baseline uses a prompt designed to be as similar as possible to the prompt-tuned single-agent baseline. Of course, we still must include some instructions that explain the communication protocol and instruct the agents to work together, but the task prompt includes all the same language as in the single-agent setting. Similarly, we use a refinement prompt that is as similar as possible to the single-agent setting, although the refinement stage still differs in that we do not manually apply it on each chunk (as this would defeat the point of using multiple agents).

3.5.4. Multi-Agent Review Generation with Specialized Agents (MARG-S)

Our full MARG-S approach is described in section 3.4, and uses three independent multi-agent groups to generate different kinds of review comments. MARG-S outputs the concatenation of the three mini-reviews generated by those groups. In addition to the full approach, we evaluate each of the three mini-reviews separately. We refer to these as MARG-S (experiments), MARG-S

(clarity), and MARG-S (impact). In addition, we include a "no refinement" baseline that skips the refinement stage.

3.5.5. Liang et al. [41] baseline (LiZCa)

We also compare against a recently proposed method for generating peer-review feedback [41], which we refer to as "LiZCa" (from the names of the lead authors of that paper; the method was not given a name in that work). Unlike our methods, this method simply truncates the paper rather than applying to multiple chunks. In addition, it includes the captions of figures and tables in the input.

The prompt used in Liang et al. [41] instructs the model to generate an "outline" style review, and includes non-actionable positive comments. Fortunately, when comparing their method's comments with real reviews, they developed a prompt to extract and merge the parts of an outline that focus on "criticisms" and to ignore minor grammar comments. This roughly matches the type of comments we target, so we use that prompt to produce the final list of comments that we use in this baseline.

We note that Liang et al. [41] used a different PDF parsing library (pikepdf) than ours (Grobid), but for consistency with our other baselines we run it with Grobid.

3.6. Automated evaluation

To automatically evaluate the quality of generated reviews, we measure their overlap with real reviews from papers in the ARIES corpus described in chapter 2. That is, we attempt to match the generated comments to comments extracted from real (human-written) reviews. Because ARIES only has comment annotations for a small set of reviews, we use GPT to extract

comments from all reviews for a subset of 30 papers and treat this as our test set. To match our intended type of feedback, GPT is instructed to focus only on actionable feedback comments and to ignore minor comments on style and grammar.

We note that this form of evaluation is imperfect in that real reviewers do not always identify every reasonable critique of a paper, and in some cases they may make critiques that are unreasonable. Thus, the generated review could contain good comments that happen to be different from ones the real reviewers made, or it could miss comments that are actually invalid. Thus, the measured overlap should be treated as a lower bound for the fraction of good-quality comments. In addition, the nuanced nature of the matching task makes it impossible to fully capture the similarities and differences between real and generated comments using binary alignments, and this could lead to biases. We nonetheless use automated evaluation as an inexpensive but rough approximation of the relative quality of different methods, and separately conduct a user study in section 3.7 to obtain a more realistic evaluation.

The matching procedure and results are outlined in the following subsections.

3.6.1. Measuring overlap

Given a set of generated review comments C_{gen} and the set of ground truth real-reviewer comments C_{real} for a paper, we automatically align individual comments between the reviews that have the same meaning. That is, we ultimately obtain a binary label for every comment pair (C^i_{gen}, C^j_{real}) indicating whether the two comments are making the same request. To do this, we begin with a "many-many" matching stage that efficiently compares the full set of comments in both reviews and identifies possibly-matching pairs, followed by a more accurate (but more expensive) pairwise stage that examines the candidate pairs to produce a final list.

Real-reviewer comment	Generated comment
The experimental methodology used in the paper is not well detailed, making it difficult to reproduce the reported results.	More details about the experiments conducted would be beneficial. This should **include information about the datasets used, the training process, and the evaluation process**. To ensure the reproducibility of the results, **consider providing the code used to implement the model, the specific parameters used, and any other necessary information**. This will allow other researchers to replicate your work and further validate your findings. **[high relatedness, more specific]**
The paper does not include enough baselines for Fair Federated Learning to compare against. Even if some methods do not satisfy privacy considerations, they should still be included for the reader to understand how the proposed method compares against such methods, especially given that the results are not promising. **Some baselines to consider include Cui et al or Tran et al.**	The authors should **consider including a comparison of their proposed method with existing methods** in the experimental results section. This would help to highlight the advantages and improvements of their proposed method. **[high relatedness, less specific]**
The datasets used in the study are not representative due to their simplicity and experimental nature.	The evaluation of the proposed method may not be comprehensive enough. The authors could **include more datasets** in their evaluation to demonstrate the robustness of their method. The paper could benefit from a more detailed discussion on the limitations of the proposed method. **[medium relatedness, more specific]**

Table 3.1. Aligned pairs of comments with corresponding relatedness and relative specificity scores from the alignment model; the bold is added to emphasize key differences. Notice that in the third row with "medium" relatedness, the reviewer comment is suggesting that the datasets need to be more representative (but a larger number of datasets is not necessarily needed) whereas the generated comment only asks for more datasets (not identifying the issue with the current datasets). In the two "high" relatedness cases, one comment fully subsumes the other (high relatedness) but includes much more specific details and rationales (less/more relative specificity).

In the many-many matching stage, we feed all comments from both reviews into GPT-4 and prompt it to output a list of all matching comments. As GPT has somewhat inconsistent performance, we do five such passes, randomly permuting both the order of comments within each review and the order in which reviews are presented. The final output of this stage is the list of comment pairs that were produced by at least two of the five runs—a ratio we heuristically found to work well in preliminary experiments.

In the pairwise stage, we give one comment pair at a time to GPT and prompt it to produce two scores: one of four levels of relatedness ("none", "weak", "medium", or "high"), and a "relative specificity" ("less", "same", "more") indicating how specific the generated comment is relative to the real review comment. To be considered a match, a comment pair must have "medium" or "high" relatedness, and the generated comment must have "same" or "more" specificity compared to the human comment. An example of an aligned pair of comments can be found in Table 3.1.

The final output is a list of alignment edges between the lists of generated and real-reviewer comments. We note that this may result in a many-many mapping; one generated comment might match multiple reviewer comments, and one reviewer comment might match multiple generated comments. This can happen when there are similar comments within one list or if, for example, a reviewer makes a broad suggestion like "Evaluate on more datasets" and the generated review contains several comments, each with a different specific dataset recommendation.

3.6.2. Metrics

Using the alignments between C_{gen} and C_{real}, we evaluate several metrics, described below. However, we note that the many-many nature of the mapping between the comments indicates

that these are not proper sets, and traditional set-based metrics such as the union and intersection are not well-defined. For our purposes, we define directional intersection operators $\overleftarrow{\cap}$ and $\overrightarrow{\cap}$ representing the set of aligned elements in the left or right operand, respectively. For example, $C_{gen} \overleftarrow{\cap} C_{real}$ is the set of elements of C_{gen} that align to any element in C_{real}.

- **Recall:** $\frac{|C_{gen} \overrightarrow{\cap} C_{real}|}{|C_{real}|}$, the fraction of real-reviewer comments that are aligned to any generated comment.
- **Precision:** $\frac{|C_{gen} \overleftarrow{\cap} C_{real}|}{|C_{gen}|}$, the fraction of generated comments that are aligned to any real-reviewer comment.
- **(Pseudo-)Jaccard:** The Jaccard index is a commonly-used measure of set overlap. Let $intersection = \frac{|C_{gen} \overleftarrow{\cap} C_{real}| + |C_{gen} \overrightarrow{\cap} C_{real}|}{2}$; then the Jaccard index is $\frac{intersection}{|C_{gen}| + |C_{real}| - intersection}$.

To compute these metrics over a set of papers, we macro-average on the level of reviews. That is, given a set of papers in our test set, we generate a review for each, measure the aforementioned metrics between each generated review and each corresponding real review, and then average all of the results to obtain a single value for each metric.

3.6.3. Results

We include a selection of example generated comments in Table 3.3. Results of the automated evaluation are shown in Table 3.2. We additionally include a human-review baseline, which is the average of the metrics computed between each real review and each other real review for the same paper (i.e., $\frac{1}{n} \sum_{i=1}^{n} \text{metric}(human_i, \{human_k | k \neq i\})$. Note that while this is theoretically unbiased for recall, it may result in lower precision and Jaccard scores for human reviewers.

Method	Recall	Precision	Jaccard	# comments
SARG-B	7.43	1.40	1.25	19.7
SARG-TP	10.62	4.61	3.46	11.6
MARG-TP	8.49	5.34	3.52	8.5
LiZCa	9.67	**9.96**	**5.58**	4.0
MARG-S	**15.84**	4.41	3.53	19.8
no refinement	11.92	3.32	2.70	18.3
experiments-only	4.36	4.83	2.23	4.1
clarity-only	3.25	2.65	1.46	6.9
impact-only	8.88	4.75	3.32	8.8
Human	9.42	12.00	5.45	4.7

Table 3.2. Automated evaluation results with recall, precision, and Jaccard values, in addition to the average number of comments generated by each method. The proposed MARG-S method outperforms all baselines in terms of recall, but generates more comments than other baselines and thus has lower precision and Jaccard scores.

We find that our proposed MARG-S method outperforms all baselines in terms of recall, but generates more comments than other baselines and thus has lower precision and Jaccard scores. With that said, we believe that recall is the most important metric in this evaluation. While higher precision and Jaccard should be preferred at similar levels of recall, it is relatively easy for a human to recognize and ignore bad comments; thus, it is more important for the system to maximize the number of good comments than to minimize the number of bad ones.

The simple baseline (SARG-B) performs poorly on all metrics; despite being tied with MARG-S for the highest number of generated comments, it has the lowest recall of all methods. This is not unexpected, but highlights the importance of careful prompting with GPT-4.

Interestingly, we find that between SARG-TP and MARG-TP (which use essentially the same task prompt), SARG-TP generates more comments and has better recall. This suggests that simply applying a multi-agent approach does not always result in a performance improvement;

Method	Example comment
SARG-B	The paper could benefit from a more detailed discussion of the results, including the implications of the findings and how they contribute to the existing body of knowledge.
LiZCa	The experimental evaluation could be more comprehensive. The authors should consider including more diverse tasks and environments in their experiments to demonstrate the robustness of their method. The paper could benefit from a more detailed analysis of the experimental results, including a discussion on why the proposed method outperforms the baselines.
MARG-S (experiments)	The authors have compared their method with several baselines, including DeepMDP, HiP-BMDP-nobisim, Distral, PCGrad, GradNorm, and PEARL. However, it would be beneficial to include comparisons with other state-of-the-art methods in multi-task and Meta-RL setups to further validate the effectiveness of the proposed method. This would help ensure that the results are not specific to the current set of comparisons and can generalize across different settings. Additionally, providing a detailed discussion on why the proposed method outperforms each baseline could offer more insights into the strengths and weaknesses of the proposed method.
MARG-S (clarity)	The paper mentions an encoder that maps observations from state space to a learned, latent representation, but it does not provide specific details about the type of encoder used or the process of how it learns the latent representation. These details are crucial for understanding how the model works and how it achieves its performance. Therefore, I recommend that the authors include this information in the paper.
MARG-S (impact)	The authors provide a theoretical proof for the 'Transfer bound' formula, which is a significant contribution. This formula is crucial for measuring the transferability of a policy learned on one task to another, taking into account the error from the learned representation. However, to ensure its robustness and applicability in real-world scenarios, it would be beneficial if the authors could empirically test this formula in reinforcement learning environments. For instance, the agent could be trained on one game and then tested on a different game with similar mechanics. This would provide empirical evidence supporting the theoretical proof and demonstrate the practical utility of the formula.

Table 3.3. Example comments generated by each method (SARG-TP and MARG-TP omitted for brevity) for the same paper. Qualitatively, we find that MARG-S writes relatively long and specific comments, whereas other methods tend to write shorter and more generic comments.

instead, the use of multiple agents enables the design of richer internal problem-solving structures via expert agents. Indeed, we see that the specialized MARG-S (impact) is able to approximately match the performance of MARG-TP despite focusing on only one type of comment.

We notice that the human baseline actually has a lower recall than some of the LLM baselines, although it has the highest precision. This is consistent with the results of Liang et al. [41], which found that Human-Human agreement was slightly lower than LiZCa-Human agreement.[3] Humans generate fewer comments than other approaches, which offers a partial explanation for the low recall, but it is nonetheless interesting to observe that human reviewers can have very different perspectives of the same work, and highlights the challenge of the review generation task (and the potential weaknesses of alignment-based evaluation).

MARG-S ablations: Among the sub-reviewers of MARG-S, the impact-focused model tends to produce the best results. The experiment-focused model does well considering the small number of comments it produces, but as it produces half as many comments as the impact model it also has half the recall. Finally, the clarity-focused model struggles compared to the other two. The poor performance of the clarity model may be due in part to the subjective nature of clarity judgements and the fact that language models do not necessarily perceive text in the same way that humans do (e.g., humans prefer that terms be defined before they are used, but a model that consumes a full document at once might not see a problem if terms are defined later). In addition, we note that due to the fact that the input does not capture visual information such as figures, tables, and the arrangement of symbols in equations, there are many resulting clarity issues that

[3]Note that while the relative differences are similar to those reported in Liang et al. [41], our absolute recall scores are lower. We conjecture that this is primarily due to differences in the alignment step; in particular, the pairwise filtering makes our approach more conservative.

are not present in the full paper, and getting the model to identify the "real" issues from among the large number of parsing- and input-format-related issues is challenging.

We observe that without the refinement stage, MARG-S's performance is reduced on all metrics, but it still obtains reasonable results; recall remains the second-highest of all methods. Interestingly, the number of generated comments is slightly lower than with the refinement stage, indicating that the refinement stage splits one comment into multiple comments more often than it prunes comments.

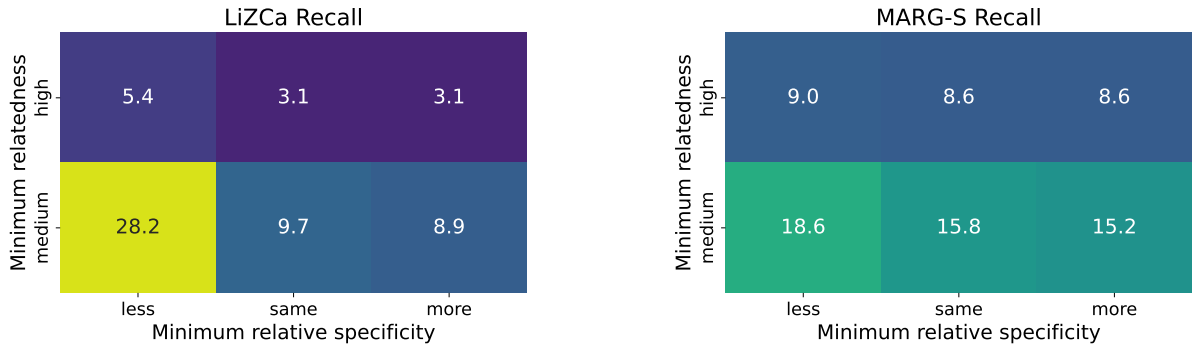

Figure 3.3. Recall of MARG-S and LiZCa for different alignment cutoff levels of relatedness and relative specificity. The ("medium", "same") cell corresponds to our default setting. LiZCa obtains very high recall in the most lenient setting, but rapidly drops for stricter settings that prevent vague comments from being counted as matches. MARG-S obtains relatively consistent results for all levels of specificity (as most of its comments are considered "more" specific) but still experiences a decline when requiring highly-related matches.

Effect of the matching thresholds. We qualitatively observe that several of the methods we evaluate produce many generic or vague comments. Many of these are not counted towards recall due to our constraint that a generated comment must be equally or more specific compared to the real comment it matches with. In addition, some aligned pairs of comments are questionable, especially for pairs that have only "medium" relatedness; for example, a comment asking for a

Method	Input tokens	Generated tokens
SARG-B	15,532	3,249
SARG-TP	54,914	6,853
MARG-TP	134,860	5,264
LiZCa	7,398	869
MARG-S	1,236,344	51,255

Table 3.4. Average number of input and generated tokens per paper for each method. This includes tokens used for internal discussion in multi-agent methods, but not tokens used outside of the method (e.g., for measuring the alignment metric). MARG-S generates substantially more tokens than other methods, and thus is more expensive to run.

"more thorough comparison" to baselines is considered a match for one that indicates that the proposed method underperforms the baselines in some cases.

To evaluate the impact of threshold choices, we select our method and the LiZCa baseline and evaluate all combinations of thresholds for relative specificity and "medium" or "high" relatedness. The results are shown in Figure 3.3.

The difference between thresholds is striking in the figure. LiZCa generates a large number of comments that broadly match to a real comment, but are much more vague (and thus less useful) and often do not have exactly the same meaning. When either the relatedness or the specificity thresholds are increased even by one step, the recall drops sharply. In contrast, MARG-S loses relatively little recall at higher specificity thresholds, as it almost always outputs specific comments. However, MARG-S still has a drop when requiring "high" relatedness, indicating that it has many matches in which the generated comment does not express exactly the same sentiment as the real one.

Cost. In Table 3.4 we report the average number of tokens generated by each method during the alignment-based evaluation. LiZCa generates the fewest tokens and has the best cost to recall

ratio overall, making it an attractive choice in budget-constrained settings. While MARG-S has the best recall, it also generates roughly an order of magnitude more tokens than other methods, suggesting that it takes on diminishing returns in efficiency to obtain the recall improvement.

The extra tokens used by MARG-S result in it taking roughly an hour longer than other methods to generate reviews. This may serve as an inconvenience in practice, and it would be beneficial to explore ways to reduce it. For example, it may be possible to dynamically switch to cheaper LLMs to handle simpler messages or develop methods to route communications more effectively (reducing the number of redundant messages). We also note that our implementation performs only one inference at a time for simplicity, but in theory, it is highly parallelizable (due to having three separate groups for different comment types, separate groups for the refinement stage, and several agents communicating at once in each group), and the time needed to generate a review could likely be reduced by 2-10x depending on the document size.

3.7. User study

We conduct a user study to obtain a more reliable (but more expensive) evaluation compared to the automated metrics. To reduce burden on participants, we only evaluate a subset of methods in the user study: MARG-S (our best method on the automated metrics), LiZCa (baseline from prior work), and SARG-B (the simplest baseline).

3.7.1. Study design

Participants. We recruit 6 volunteers[4] from a large research organization to participate in the study. All participants are researchers in the fields of natural language processing and human-computer interaction.

Figure 3.4. The survey interface. Participants were asked to rate the specificity, accuracy, and overall helpfulness of each comment, and to rate the overall review.

Survey. The study was conducted using a web interface in which participants could upload a paper PDF. We then ran each review generation method to produce a set of reviews, where each review was a list of comments. When all reviews were generated, participants would receive

[4]While this is a small number of participants, we note that the number of rated comments is much higher (each participant rates many comments per method), and we obtain statistically significant conclusions from mixed-effect analyses in which we control for participant bias as a random effect.

an email notification with a link to page with reviews and a set of survey questions, depicted in Figure 3.4. The survey page did not describe the review generation methods or give any indication of which method generated a given review, and the generated reviews were displayed in a random order to reduce bias (the order of comments within reviews was not randomized, however).

For each comment, participants were asked to rate its specificity, accuracy, and to provide an overall rating. The following guidelines for these ratings were provided at the start of the survey:

- **Specificity:** Does the comment make a suggestion specific to the paper, or is it generic (could apply to many papers)? Please note that a comment may be verbose without being specific, or vice versa.
- **Accuracy:** Does the comment display an accurate understanding of the paper and make a valid critique? For example, suppose a comment says the paper is missing statistical significance tests and should include them. If the paper doesn't have significance tests and could potentially benefit from including them, please rate the comment as "accurate" (even if the importance of those tests is questionable). If the paper has tests on one or two results but not all, and the comment doesn't mention this, the comment would have a "minor" inaccuracy. If the paper already has extensive significance tests or provides substantial justification for not including them, the comment would have a "major" inaccuracy.
- **Overall rating:** How helpful is the comment overall? Is the comment one that you would want to see in a review (Good), one that you might not mind seeing but don't care much about (Neutral), or one that is useless or invalid (Bad)?

Method	Bad	Neutral	Good	Total
SARG-B	9.3	4.8	1.5	15.7
LiZCa	2.3	1.0	0.5	3.8
MARG-S	7.7	3.7	**4.2**	15.5

Table 3.5. Average number of each comment rating per review for each method. MARG-S generates the most good comments. LiZCa generates substantially fewer comments than the other methods, and therefore has the fewest bad comments per review but also the fewest good comments.

In addition, participants were asked questions at the end of each review. Specifically, they were asked to rate whether the review was too long or too short on a 5-point scale and to provide an overall rating for the review on a 5-point scale.

Finally, we asked participants about their research and reviewing experience, and about their authorship of the submitted paper.

3.7.2. Total good comments

As in the alignment-based evaluation, we argue that bad comments have relatively small cost compared to the value of good comments. As there is no straightforward way to adjust the total number of generated comments (unlike in a classification task, where the decision threshold could be adjusted continuously), the total number of good comments is the most appropriate metric with which to compare methods.

Table 3.5 shows the average number of each comment rating per review for each method. We find that MARG-S generates the most good comments by a wide margin compared to SARG-B (p=0.08, related-sample t-test) and LiZCa (p=0.02). LiZCa generates substantially fewer comments than the other methods, and therefore has the fewest bad comments per review but also the fewest good comments.

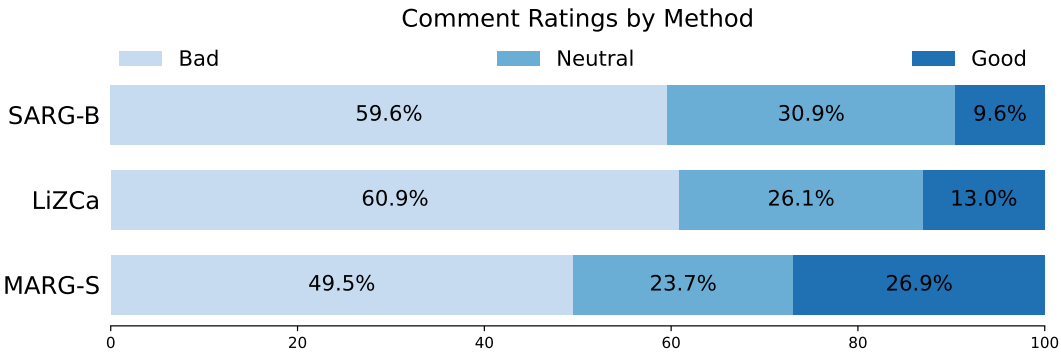

Figure 3.5. Average quality ratings for each method. LiZCa and SARG-B are rated similarly, while MARG-S has over twice the fraction of "good" comments compared to the other two methods.

Review length. MARG-S generates the most good comments, but does this come at the cost of generating overly-long reviews? It seems that in general, MARG-S reviews do tend to be longer than authors would like, while LiZCa reviews are too short. Specifically, MARG-S was rated as "way too long" by 4 of the 6 participants (and "just right" by the other two), while LiZCa was rated as "too short" by 3, "way too short" by 2, and "just right" by 1 of the participants. SARG-B occupied a middle ground, rated as "too short" by 1, "too long" by 3, and "just right" by 2 of the participants. Although SARG-B generates a similar number of comments as MARG-S, the comments it generates are much shorter, which is likely why its length is perceived as being more reasonable than MARG-S.

3.7.3. Average comment ratings

The distribution of user ratings of comment quality, accuracy, and specificity are shown in Figure 3.5, Figure 3.6, and Figure 3.7, respectively.

We find that MARG-S has the highest proportion of "good" comments, and is significantly better than SARG-B ($p=10^{-2}$ for per-comment Barnard's exact test, p=0.08 for per-user related

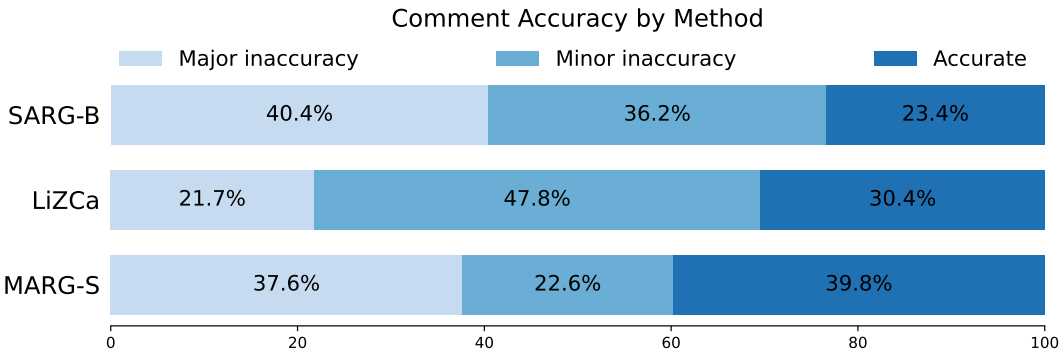

Figure 3.6. Average accuracy ratings for each method. MARG-S has the most fully accurate comments, but its inaccurate comments are more likely to have "major" inaccuracies compared to LiZCa, which typically has only "minor" inaccuracies. SARG-B is less accurate than both other methods.

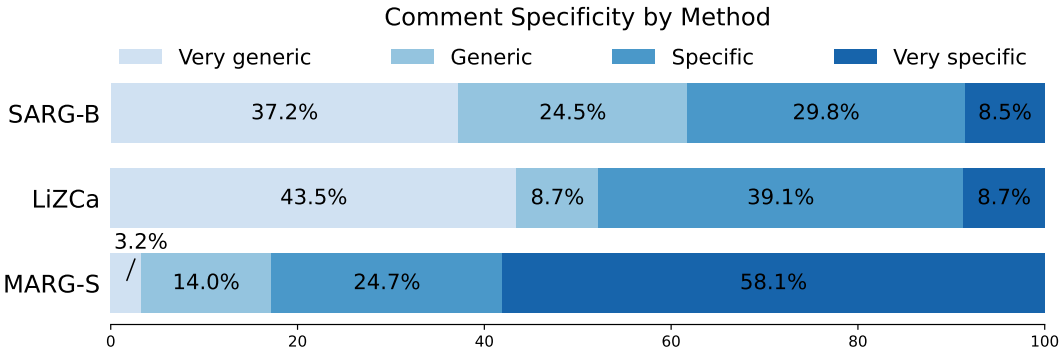

Figure 3.7. Average specificity ratings for each method. LiZCa and SARG-B have similar proportions of the "very" specific or generic comments, but LiZCa has substantially more somewhat specific comments. MARG-S is extremely specific compared to the other two methods; over 83% of MARG-S comments are rated specific or very specific, compared to only 49% for LiZCa.

sample t-test), although the difference between MARG-S and LiZCa is not significant (p=0.18 per-comment, p=0.37 per-user). When asked about the overall helpfulness of the reviews, 4 of the 6 participants rated MARG-S substantially better than LiZCa (2 points higher on the 5-point scale), one rated them the same, and one rated LiZCa as slightly better.

The accuracy ratings in Figure 3.6 show a similar trend as the comment quality ratings. MARG-S has the highest proportion of fully accurate comments, but the differences are not significant. The accuracy and quality ratings are also similar in that MARG-S has a somewhat bimodal distribution, with a greater proportion of ratings at either extreme than in the middle. This may be related to the specificity differences between methods.

The most striking difference between the methods is in specificity. MARG-S has "very specific" comments at more than 6 times the rate of the other two methods, a significant increase ($p=0.002$, per-user related-sample t-test). Overall, 83% of its comments are rated as "specific" or "very specific", compared to only 49% for LiZCa ($p=0.08$).

Finally, we observe that MARG-S has a high proportion of "good" comments rated by users despite having a relatively low precision in the automated evaluation (Table 3.2). The difference suggests that it may generate many comments which are helpful but also different than the kinds of suggestions a real reviewer would tend to make. This could be a promising sign indicating that MARG-S can serve as a useful source of novel inspiration for authors—even when the paper has already been reviewed by humans—and that it may be a source of inspiration for reviewers as well.[5]

3.7.4. Relationships between factors

The bimodal distributions of accuracy and quality ratings for MARG-S suggest that there may be a relationship between specificity and accuracy/quality. In particular, more specific comments may be easier to make strong judgements about, whereas comments that are generic or vague

[5]Of course, it is also possible that the kinds of novel suggestions MARG-S makes only *appear* useful to authors and actually do not improve the paper in ways that reviewers (or readers) care about. Measuring the extent to which this is the case would require a much more sophisticated study of how these comments affect the long-term impact of papers, and we leave this to future work.

	Coef	Std. err	z-value	Pr(>|z|)
MARG-S	-0.12	0.43	-0.27	0.79
LiZCa	-0.74	0.57	-1.29	0.20
Inaccuracy (minor)	-1.14	0.42	-2.72	10^{-2}
Inaccuracy (major)	-6.33	1.10	-5.76	10^{-8}
Specificity (specific)	1.98	0.43	4.56	10^{-5}
Random effect std. dev		$\sigma = 1.13$		

Table 3.6. Cumulative link fixed effects for specificity, accuracy, and method on the overall rating of a comment. Specificity is positively associated ratings, as is accuracy (inaccuracies have a negative effect). The review generation method has a relatively small independent effect compared to the other factors, suggesting that specificity and accuracy capture a large portion of the aspects that contribute to perceived comment quality.

	Rating=Bad		Rating=Neutral		Rating=Good	
Factor	Coef	p-value	Coef	p-value	Coef	p-value
(Intercept)	0.02	0.98	-1.54	0.02	-2.00	0.02
MARG-S	0.36	0.53	-0.47	0.33	0.38	0.52
LiZCa	0.97	0.17	-0.71	0.28	-0.22	0.80
Inaccuracy (minor)	0.70	0.22	1.20	10^{-2}	-2.44	10^{-3}
Inaccuracy (major)	6.39	10^{-6}	-3.27	10^{-2}	-19.46	0.99
Specificity (specific)	-2.71	10^{-4}	1.03	0.03	2.07	10^{-2}
Random effect std. dev	$\sigma = 2.06$		$\sigma = 0.71$		$\sigma = 1.49$	

Table 3.7. Mixed-effects logistic regression coefficients and p-values for the effect of specificity, accuracy, and method on the probability of a comment receiving a given overall rating. Specificity is positively associated with neutral and good ratings, while major inaccuracies are strongly predictive of bad ratings.

may be hard to clearly classify. To investigate this, we fit logistic regression mixed-effects models to find the effect of specificity on the classification probabilities of the overall rating and on the accuracy while controlling for the generation method. In addition, we analyze the

Factor	Inaccuracy=major Coef	p-value	Inaccuracy=minor Coef	p-value	Inaccuracy=none Coef	p-value
(Intercept)	-0.05	0.89	-0.89	0.08	-1.48	10^{-4}
Specific	-0.79	0.03	0.04	0.91	0.68	0.07
MARG-S	0.03	0.93	-0.45	0.27	0.45	0.22
LiZCa	-1.02	0.08	0.76	0.15	0.30	0.57
Random effect std. dev	$\sigma = 0.57$		$\sigma = 1.00$		$\sigma = 0.16$	

Table 3.8. Mixed-effects logistic regression coefficients and p-values for the effect of specificity on accuracy.

tendency of both specificity and accuracy to result in higher ratings using a cumulative link mixed-effects model.

We binarize specificity in these analyses by grouping "specific" and "very specific" judgements together as well as "generic" and "very generic" ones. The logistic regression and cumulative link models are implemented in R, using the `lme4.glmer` [6] and `ordinal.clmm` [11] functions, respectively. We treat the submission ID as a group variable (random effect).

Results of the logistic regression analysis are shown in Table 3.7 (predicting overall rating) and Table 3.8 (predicting accuracy given specificity). Surprisingly, we find that specificity has a positive association with neutral ratings, contradicting our original hypothesis that the high specificity of MARG-S contributes to its bimodal rating distribution. Higher specificity does not appear to produce the bimodal accuracy distribution either, and instead seems to weakly correspond with higher accuracy. It is unclear why specificity would influence accuracy in this way, but we speculate on three possibilities:

- **Calibration:** There is evidence that humans tend to give more precise answers when they are more confident [69]. The model may mimic this tendency and write more specific comments when it has greater confidence.

- **GPT-4 mode switching:** GPT-4 may have an intrinsic tendency to write comments that are either good in both specificity and accuracy or bad in both. It has been rumored that GPT-4 uses a mixture-of-experts architecture,[6] in which case the correlated behavior may be related to expert routing.
- **Human bias:** Humans may have a tendency to perceive comments as more specific when they are more accurate, even if the specificity is not actually relevant to the accuracy. For example, *"There is only one baseline for comparison. You should add more."*, is very generic, and this is easy to see when it is inaccurate. However, if there really is only one baseline and adding more would be useful, it may be perceived as more specific because it appears to demonstrate an understanding of the paper.

The analysis in Table 3.6 shows that accuracy is highly predictive of overall rating, particularly for major inaccuracies. In fact, we find that 99% of all comments with a major inaccuracy are rated as bad, as opposed to 32% for minor inaccuracies and 27% for accurate comments. Specificity plays a larger role among comments without a major inaccuracy; within this group, only 11% of non-specific comments were rated as good, while 49% of specific comments were. Still, specificity and accuracy are not perfect predictors of comment quality; even among comments that were rated as both fully accurate and very specific, only 54% were rated as good.

3.7.5. Complements and ratings

We observe qualitatively that some generated comments include complements or flattering remarks; for example, a comment might say *"While the authors have done a commendable job in [...], the paper could benefit from [...]"*. To test whether these complements might bias the

[6] OpenAI has not publicly released architecture details; the mixture-of-experts claim was made by an AI researcher on a podcast and is consistent with the speed and cost of the model.

user ratings, we use GPT-4 to detect the presence of such remarks in all generated comments, using the following prompt:

```
Determine whether following review comment for a scientific paper includes a
 complement or flattering remark about the paper.  Output a JSON object with
 the key "has_complement" set to true or false.  Output only JSON with no
additional commentary.

Comment: {comment}
```

We fit a cumulative link mixed-effect model with accuracy, specificity, method, and "has_complement" as fixed effects and submission id as a random effect. We find that "has_complement" has a coefficient of 0.12 (p=0.80), which is small relative to the coefficients of other factors we observed in Table 3.6 and smaller than the random effect standard deviation ($\sigma = 1.13$), and we cannot reject the null hypothesis that the coefficient is 0. Thus, it does not appear that flattery causes a meaningful bias. Of course, we note that detecting complements is somewhat subjective and can be a matter of degree, so it is still possible that there are more subtle biases in user ratings; we leave further analyses to future work.

3.8. Failure analysis

While MARG-S does well relative to other methods, there are still a large number of comments rated as "bad", and the precision and recall in the automated evaluation are still rather low in absolute terms. In this section, we qualitatively analyze the conversation message logs of the multi-agent system and identify several common classes of errors in the communication. The analysis was carried out by an author of this work with several publications in the field of machine learning and natural language processing, and the papers being analyzed were broadly related to the topic of machine learning.

3.8.1. Scope

There are two main stages of the multi-agent system: (1) the "main" stage, in which the model comes up with a list of comments, and (2) the refinement stage, in which the comments from the main stage are refined and potentially pruned if they are redundant. For 10 papers from the automated evaluation, we analyze the main stage for all three sub-reviewers (experiments, clarity, impact), for a total of 30 conversations. We additionally analyze the refinement stage for one randomly-selected comment from each of the 30 papers in the automated evaluation test set.

Checking each message against the paper for factual inconsistencies is expensive and error-prone, especially given the number of claims and comments that can be generated in the main stage, so for the main stage we only consider errors that are apparent from the conversations themselves. For the refinement stage, we do refer to the paper to check whether the models missed basic facts; however, it is important to note that only a limited amount of time (approximately 5-15 minutes) was spent to check comments against each paper, and due to the highly technical nature of these works it is possible that some factual errors were overlooked. Nonetheless, the fraction of invalid comments identified in this analysis is similar to the fraction of bad-rated comments found in the user study, so we believe the findings are reasonably accurate.

3.8.2. Main stage

Below, we describe the error categories we identified for the main stage, along with the percentage of conversations that contain the error type. If the same error type appears multiple times in the same message log, we only count it once. It is worth noting that not all errors ultimately result in erroneous comments, as it is possible for agents to point out each others' errors and address them.

Overall, 70% of conversations contain at least one of these error types:

- **Missing context (MC) (53%):** The leader agent fails to include key context in a message to another agent. In general, this tends to happen when it messages an expert agent and fails to include some information about the paper that the expert needs to proceed.
- **Missing context - misplaced `SEND MESSAGE` (MC-MSM) (47%):** A subtype of MC, this error occurs when the leader agent does include the necessary context in its generated output, but places the `SEND MESSAGE` marker after it instead of before.
- **Fails to Identify Error (FIE) (17%):** When the leader makes one of the aforementioned errors, worker or expert agents should point this out and ask the leader to try again, but they sometimes fail to do this.
- **Ignores Relevant Information (IRI) (10%):** An agent ignores part of a message that it should have responded to.
- **Failure to Respond (FR) (7%):** An agent does not recognize a message as being relevant and gives an empty or vapid response.
- **Skipping Steps (SS) (7%):** The leader moves to a later step too early. For example, writing the final review comments before the expert's questions are resolved, or skipping the initial step where it is supposed to get a summary of the paper.
- **Message loop (LOOP) (7%):** The agents enter a loop of similar messages, triggering the duplicate-message detector described in section 3.4.2.
- **Exceeds input token limit (EITL) (7%):** The conversation exceeds the input token limit for the underlying model. These cases occur when the expert asks too many

questions, which can happen when the expert repeatedly asks for slightly more details each time it gets an answer to a question.

Qualitatively, we noticed that there is a very common pattern for missing-context errors. Specifically, when the leader first addresses the expert, it tries to include a summary of the paper to give context for the expert, but it misplaces the `SEND MESSAGE` indicator.

This error occurs in 33% of conversations, but in 80% of those cases the expert points out the error and the leader corrects it. Interestingly, in many instances of the error, the leader tries to use a placeholder ("[insert summary here]") despite never being instructed to do so; for example (magenta text verbatim, black text is edited):

> **Agent 0 (leader):** Summary: <omitted for brevity>
>
> Step 3: Share the summary with Agent 3 and ask for their input, specifically focusing on potential shortcomings of the paper's assumptions.
>
> SEND MESSAGE: Agent 3, here is a summary of the paper: **[insert summary here]**. Could you please provide your input on potential shortcomings of the paper's assumptions?

Also interesting is the fact that in all cases when the leader fails to include the summary, there are no additional missing-context errors in the remainder of the message log. We conjecture that the early failure (and the following correction) may serve as a form of one-shot example that encourages the model to avoid such errors later in the discussion.

3.8.3. Refinement stage

Because the refinement stage works with one comment at a time, we do check the comment against the paper to determine if it is relevant. While this is somewhat subjective, we attempt to

give the model the benefit of the doubt; if the comment is factually consistent and does raise a potentially valid suggestion (even if minor or difficult to address), we consider the comment as valid in the sense that it is fine for the system not to prune it.

Below, we describe the error categories we identified for the refinement stage, along with the percentage of conversations that contain the error type. As with the main stage, if the same error type appears multiple times in the same conversation, we only count it once.

- **Failure to prune a comment (47%):** The system fails to prune a comment that is invalid. This can happen for several reasons:
  - **Ignored information (17%):** The comment is already addressed in the parsed paper text or contradicts information in the text, but the model did not recognize it.
  - **Unavailable information (13%):** The comment is already addressed in the paper or contradicts information in the paper, but that information is not available in the parsed text (either due to parsing errors, or because it is in a figure or table).
  - **Irrelevant (17%):** The comment asks for something that is trivial or does not make sense in the context of the paper (e.g., requesting an experiment to confirm a claim that the paper does not make or that is trivially true by definition).
- **Revising instead of pruning (30%):** The original comment for refinement is invalid, but the system revises it rather than pruning it. This is not always a failure, as the revision may result in a good comment, but there are also many cases where the revised comment has questionable relevance.
- **Failure to respond (17%):** An agent does not recognize a message as being relevant and gives an empty or vapid response.

- **Message loop (10%):** The agents enter a loop of similar messages, triggering the duplicate-message detector described in section 3.4.2.
- **Pruned (3%):** The system marks the comment as invalid and prunes it. This only occurred for one comment in our sample, and the decision to prune was correct in that case.

Our analysis finds that the system fails to prune invalid comments in 47% of cases, which is close to the 49% of comments that were rated as bad in the user study. Most of these cases appear to occur because the system lacks some information from the paper, either because the information isn't extracted by the PDF parser or because the model simply ignores it. The relatively high fraction of cases where models ignore information is quite interesting and suggests that GPT-4 can still fail in very basic ways, despite the strong results that have been seen in other studies of its reasoning ability. Our result is supported by some recent work, however, which suggests that large models do not always use their full input effectively, and have a particularly hard time attending to tokens near the middle of the input [42].

Another common trend we find is that even when presented with information that would render the original comment invalid (e.g., the paper already covers it), the GPT-4 leader agent strongly prefers to revise the comment to extend the request rather than pruning the comment. For example, consider the following comment:

> *"The paper does not provide specific details about the experiment setup using the MiniImagenet benchmark. The authors should provide more detail about this setup, including the specific settings and how the results were measured."*

The paper actually does include details, which the leader agent summarizes:

> **Agent 0 (leader):** Summary of findings:
> - Agent 1 and Agent 2 did not find any information about the MiniImagenet benchmark in their parts of the paper.

> - Agent 3 found a section that discusses the setup for the MiniImagenet benchmark, including details about data efficiency and computational efficiency, and the tuning of hyper-parameters.
> - Agent 4 did not find any information about the MiniImagenet benchmark in their part of the paper.
> - Agent 5 found a section titled "D.2 FEW-SHOT MINIIMAGENET" which provides details about the experiment setup using the MiniImagenet benchmark, including the structure of the dataset, the protocol for meta-training, and the comparison of the BMG meta-objective against the MG meta-objective.
> - Agent 6 found information about the model used in the MiniImagenet benchmark, the training setup, and the behavior of BMG and MG with different data and time budgets. ...

However, while it recognizes the original comment as invalid, it refines the comment to ask for the information to be presented more clearly instead of pruning it:

> **Agent 0 (leader):** ... The comment is not valid and should be removed. However, it may be helpful to suggest that the authors consolidate the information about the MiniImagenet benchmark into one section for clarity.
>
> Revised comment: "The paper provides detailed information about the experiment setup using the MiniImagenet benchmark, including the specific settings and how the results were measured. However, this information is spread across different sections of the paper, which may make it difficult for readers to find and understand. The authors could improve the paper by consolidating this information into one section."

In this case, the revised comment is a valid comment. However, this is not always the case; the leader often doesn't check the validity of the new revised comment with the other agents, so if the refinement introduces an invalid request it typically will not catch the error. It may be possible to improve accuracy by repeating the refinement stage several times, although it would be expensive to do so for every comment.

3.9. Conclusion

In this work, we have introduced MARG, a novel method for review generation, which uses a network of LLM agents that communicate to share information across different parts of a paper

and to engage in internal discussion to write better comments. We evaluated MARG against both our own simple baselines and a contemporaneously-published GPT-4 baseline and found that MARG produces more good comments in both an alignment-based evaluation and a user study. The user study found that MARG is especially strong in terms of specificity and tends to generate very detailed comments compared to other methods. However, a majority of comments across all methods (including MARG) are rated as bad, and 20-40% are rated as highly inaccurate, suggesting that substantial work is still needed.

MARG is substantially more expensive to use compared to other methods (in terms of both time and API cost), and exploring ways to reduce this, such as dynamically switching to faster and cheaper models for simpler parts of the task, could be a promising avenue for future work. In addition, future work could extend the method to incorporate background literature, which would enable more informed critiques of related work and baseline choices. Finally, while splitting the paper into chunks allows MARG to consume papers beyond the base model's input size limits, it is still limited in that very large inputs can result in a large number of messages on each round of communication (one per chunk) which overflow the input; it would be interesting to explore ways to compress or prune messages to further increase the system's effective input capacity.

CHAPTER 4

Conclusions and Future Work

In this dissertation, we have presented our investigations into the automation of two aspects of peer review: (1) how edits can be aligned to—or generated from—feedback comments for a scientific paper, and (2) how feedback comments can be generated given a paper. Our results revealed fundamental weaknesses of LLMs for these tasks—such as a tendency to produce generic and non-technical outputs and a bias towards reasoning based on word- and topic-level characteristics rather than deeper semantics—and also identified strategies to mitigate these weaknesses for review generation.

In chapter 2, we constructed ARIES, a dataset of peer review comments aligned to specific paper edits, and investigated the ability of LLMs to produce such alignments. We found that even state-of-the-art LLMs perform poorly compared to humans when tasked with aligning feedback comments to corresponding edits. Performance is higher in simpler cases (where the comment makes a direct request and the corresponding edit complies with the request) than in the more nuanced cases where edits might address the underlying intent of a comment but not the explicit details of the request. Similarly, we found that when generating edits, GPT-4 can generally produce an edit that addresses a given comment on a surface level, but generally misses the technical details and sometimes non-compliant aspects of human-written comments.

In chapter 3, we proposed MARG, an approach for generating feedback comments for papers using GPT-4. MARG structures the feedback generation process as an interaction between independent agents, allowing it to scale to papers beyond the base model's input capacity (by

assigning different chunks of the paper to different agents) and to engage in internal discussion with "expert" models that assist with specific subtasks to identify better comments. We compare MARG to simple baselines and to the previous state-of-the-art for GPT-4 review generation, and find that while the baselines have a high tendency to generate generic comments, MARG generates comments that are substantially more specific and more helpful overall. However, we find that all methods are still prone to generating many inaccurate comments, and none acheive high coverage of real reviewer comments.

4.1. Future work

LLM reasoning

Taken together, our findings from these studies suggest that LLMs do show promise for generating feedback comments and edits for papers. However, there are still significant limitations when attempting to comprehend or produce nuanced and technical text. The results of MARG show that it is possible to obtain detailed generations from GPT-4 with careful prompting and structuring of the problem, but even then, the generated details are not always correct. More work is needed to develop systems that produce detailed outputs but also maintain logical consistency.

We also note that many of the findings in our work are based on experiments with GPT-4. While GPT-4 is one of the more popular and powerful LLMs available today, many others have been developed [2, 63, 81], and GPT in particular is trained to have "guard-rails" that make it less likely to generate harmful or hallucinatory outputs. It would be interesting to explore how differences in the base LLM can affect performance on the tasks we investigate.

Multi-agent modeling

There are numerous design decisions involved in our multi-agent review generation approach, and many paths remain unexplored. For example, we used the same LLM (GPT-4) for all agents at all rounds of communication, but it would be interesting to explore the possibility of using different LLMs—possibly fine-tuned to specialize in different tasks—for different agents or even for the same agent at different times depending on the message it is responding to (e.g., if a worker is simply asked to extract a numerical result from the paper, the power of GPT-4 may not be necessary). This may help to reduce costs and improve reasoning.

Another potential area of improvement is in the context management of individual agents. In our work, we used a simple pruning-based approach to prevent the message history from becoming too large, but it would be interesting to consider more sophisticated approaches such as history summarization [64]. Given recent results that show LLMs are not always able to use their full capacity effectively [42, 54], future work could also explore whether using smaller paper chunks distributed across more agents could affect results.

Finally, we note that our multi-agent approach could in theory be applied to other tasks such as multi-document summarization or multi-hop question answering, and it would be interesting to investigate whether it could have advantages over existing approaches in those settings. The internal dialogue generated by multi-agent modeling may also improve the explainability of outputs, and in future work it would be valuable to study whether users find it helpful to receive excerpts of the internal conversation in addition to the final output to better calibrate their confidence in the result.

Extending applications

In this dissertation, we separately investigated the tasks of generated feedback and generating edits. A natural extension of these investigations would be to combine the systems into a single assistant that can autonomously improve papers by generating comments and the revising the paper based on those comments. The comment generation would make the system more explainable than a system that attempts to generate edits for a paper text directly, and may also serve as a useful chain-of-thought step to improve the quality of edits.

In addition, while our MARG-S approach is designed and evaluated for scientific paper reviewing in this work, it is theoretically adaptable to a wide range of other tasks. In general, any task that would benefit from input data larger than the base model's capacity could benefit from information-sharing between multiple agents, and any task that can be broken into different sub-tasks could potentially benefit from expert agents and group specialization.

In particular, critiquing tasks in other domains are a promising target for extending MARG. For example, code review involves many similar aspects to scientific papers (highly technical, large inputs, specialized aspects such as speed optimizations, readability, and stability), as do document-reviewing tasks in other technical settings such as legal documents and medical diagnoses/prognoses. However, MARG may be applicable to tasks beyond critiquing as well; for example, to multi-document summarization or financial portfolio generation.

Broadening the input domain

Our studies used only the textual content of papers and reviews as input to LLMs. However, figures and tables often provide a substantial amount of information that is not summarized in the text. Several multimodal LLMs have been developed that can consume visual information

along with text [3, 40, 75]. In addition, it is not clear how best to represent the structure of mathematical equations for a language model, and methods for better spatial and mathematical reasoning could be highly beneficial for review and revision tasks.

Beyond alternative modalities, another key influence in real scientific papers and reviews is the related work in the field. While LLMs like GPT-4 have a large amount of world knowledge baked into the model weights via pretraining, they do not stay up to date with the latest publications (which is crucial when working on the frontiers of knowledge) and may not even be trained on all past publications. Retrieving related papers as input could improve both edit and comment generation; this is especially interesting to try in the context of multi-agent modeling, which provides a natural way to add those papers without overflowing the model's input limit.

Finally, our long-term vision for this area is to expand scientific assistants beyond simple paper feedback and design a system that consumes all information related to a given study. That is, the system might consume lab notes, code, meeting recordings, and so on, such that it has the same (or more) information as the human researcher. The system would then be capable of interjecting throughout the research process to provide feedback before a paper is ever drafted, and eventually may even carry out experiments or write the paper based on the results.

Bibliography

[1] Zafar Ali, Pavlos Kefalas, Khan Muhammad, Bahadar Ali, and Muhammad Imran. "Deep learning in citation recommendation models survey". In: *Expert Systems with Applications* 162 (Dec. 2020), p. 113790. ISSN: 09574174. DOI: 10.1016/j.eswa.2020.113790. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417420306126 (visited on 10/21/2021).

[2] Rohan Anil et al. *PaLM 2 Technical Report*. en. Sept. 2023. URL: http://arxiv.org/abs/2305.10403 (visited on 11/02/2023).

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*. Oct. 2023. DOI: 10.48550/arXiv.2308.12966. URL: http://arxiv.org/abs/2308.12966 (visited on 11/02/2023).

[4] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. *LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding*. Aug. 2023. DOI: 10.48550/arXiv.2308.14508. URL: http://arxiv.org/abs/2308.14508 (visited on 10/26/2023).

[5] Setio Basuki and Masatoshi Tsuchiya. "The Quality Assist: A Technology-Assisted Peer Review Based on Citation Functions to Predict the Paper Quality". In: *IEEE Access* 10 (2022), pp. 126815–126831. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3225871. URL: https://ieeexplore.ieee.org/document/9968010 (visited on 10/26/2023).

[6] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, Gabor Grothendieck, Peter Green, et al. "Package 'lme4'". In: *URL http://lme4. r-forge. r-project. org* (2009).

[7] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. 2020. DOI: 10.48550/ARXIV.2004.05150. URL: https://arxiv.org/abs/2004.05150.

[8] Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. *Writer-Defined AI Personas for On-Demand Feedback Generation*. en. Sept. 2023. URL: http://arxiv.org/abs/2309.10433 (visited on 09/21/2023).

[9] Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agarwal, and Asif Ekbal. "PEERRec: An AI-based approach to automatically generate recommendations and predict decisions in peer review". en. In: *International Journal on Digital Libraries* (July 2023). ISSN: 1432-1300. DOI: 10.1007/s00799-023-00375-0. URL: https://doi.org/10.1007/s00799-023-00375-0 (visited on 10/27/2023).

[10] Ting-Rui Chiang and Yun-Nung Chen. "Relating Neural Text Degeneration to Exposure Bias". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 228–239. DOI: 10.18653/v1/2021.blackboxnlp-1.16. URL: https://aclanthology.org/2021.blackboxnlp-1.16 (visited on 05/25/2023).

[11] Rune Haubo Bojesen Christensen. "Package 'ordinal'". In: *Stand* 19.2016 (2015).

[12] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. "A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 4599–4610. DOI: 10.18653/v1/2021.naacl-main.365. URL: https://aclanthology.org/2021.naacl-main.365 (visited on 05/25/2023).

[13] Wanyu Du, Zae Myung Kim, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. "Read, Revise, Repeat: A System Demonstration for Human-in-the-loop Iterative Text Revision". In: *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 96–108. DOI: 10.18653/v1/2022.in2writing-1.14. URL: https://aclanthology.org/2022.in2writing-1.14 (visited on 05/22/2023).

[14] Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. "Understanding Iterative Revision from Human-Written Text". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3573–3590. DOI: 10.18653/v1/2022.acl-long.250. URL: https://aclanthology.org/2022.acl-long.250.

[15] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. May 2023. DOI: 10.48550/arXiv.2305.14325. URL: http://arxiv.org/abs/2305.14325 (visited on 11/03/2023).

[16] Nell K Duke and P David Pearson. "Effective practices for developing reading comprehension". In: *Journal of education* 189.1-2 (2009), pp. 107–122.

[17] Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. "Text Editing by Command". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 5259–5274. DOI: 10.18653/v1/2021.naacl-main.414. URL: https://aclanthology.org/2021.naacl-main.414.

[18] Linda G. Fielding and And Others. *How Discussion Questions Influence Children's Story Understanding. Technical Report No. 490*. en. Tech. rep. Jan. 1990. URL: https://eric.ed.gov/?id=ED314724 (visited on 12/14/2023).

[19] Deshan Gong, Zhanxing Zhu, Andy Bulpitt, and He Wang. "Fine-grained Differentiable Physics: A Yarn-level Model for Fabrics". In: *Proceedings of the International Confernece on Learning Representations*. 2022.

[20] David Grangier and Michael Auli. "QuickEdit: Editing Text & Translations by Crossing Words Out". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 272–282. DOI: 10.18653/v1/N18-1025. URL: https://aclanthology.org/N18-1025 (visited on 05/26/2023).

[21] *GROBID*. https://github.com/kermitt2/grobid. 2008–2023. swh: 1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

[22] Pengcheng He, Jianfeng Gao, and Weizhu Chen. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. 2021.

[23] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. "Conversations with Documents: An Exploration of Document-Centered Assistance". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. CHIIR '20. New York, NY, USA: Association for Computing Machinery,

Mar. 2020, pp. 43–52. ISBN: 978-1-4503-6892-6. DOI: 10.1145/3343413.3377971. URL: https://dl.acm.org/doi/10.1145/3343413.3377971 (visited on 05/25/2023).

[24] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. *MetaGPT: Meta Programming for Multi-Agent Collaborative Framework*. en. Aug. 2023. URL: http://arxiv.org/abs/2308.00352 (visited on 09/25/2023).

[25] Takumi Ito, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, and Kentaro Inui. "Langsmith: An Interactive Academic Text Revision System". en. In: *arXiv:2010.04332 [cs]* (Oct. 2020). URL: http://arxiv.org/abs/2010.04332 (visited on 11/23/2021).

[26] Maor Ivgi, Uri Shaham, and Jonathan Berant. "Efficient Long-Text Understanding with Short-Text Models". In: *Transactions of the Association for Computational Linguistics* 11 (Mar. 2023), pp. 284–299. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00547. URL: https://doi.org/10.1162/tacl_a_00547 (visited on 10/26/2023).

[27] Chao Jiang, Wei Xu, and Samuel Stevens. *arXivEdits: Understanding the Human Revision Process in Scientific Writing*. en. Oct. 2022. URL: http://arxiv.org/abs/2210.15067 (visited on 12/01/2022).

[28] A. U. Kalnins, K. Halm, and M. Castillo. "Screening for Self-Plagiarism in a Subspecialty-versus-General Imaging Journal Using iThenticate". en. In: *American Journal of Neuroradiology* 36.6 (June 2015), pp. 1034–1038. ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A4234. URL: https://www.ajnr.org/content/36/6/1034 (visited on 10/27/2023).

[29] Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. "ArgRewrite V.2: an Annotated Argumentative Revisions Corpus". en. In: *Language Resources and Evaluation* 56.3 (Sept. 2022), pp. 881–915. ISSN: 1574-020X, 1574-0218. DOI: 10.1007/s10579-021-09567-z. URL: http://arxiv.org/abs/2206.01677 (visited on 04/10/2023).

[30] Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. "Improving Iterative Text Revision by Learning Where to Edit from Other Revision Tasks". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9986–9999. URL: https://aclanthology.org/2022.emnlp-main.678 (visited on 05/18/2023).

[31] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.

[32] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. "Reformer: The Efficient Transformer". en. In: 2020. URL: https://openreview.net/forum?id=rkgNKkHtvB (visited on 10/28/2023).

[33] Kayvan Kousha and Mike Thelwall. "Artificial intelligence to support publishing and peer review: A summary and review". en. In: *Learned Publishing* n/a.n/a (Aug. 2023). ISSN: 1741-4857. DOI: 10.1002/leap.1570. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/leap.1570 (visited on 10/26/2023).

[34] Karen Kukich. "Techniques for automatically correcting words in text". en. In: *ACM Computing Surveys* 24.4 (Dec. 1992), pp. 377–439. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/146370.146380. URL: https://dl.acm.org/doi/10.1145/146370.146380 (visited on 05/19/2023).

[35] Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. "Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review". In: *Computational Linguistics* 48.4 (Dec. 2022), pp. 949–986. ISSN: 0891-2017. DOI: 10.1162/coli_a_00455. URL: https://doi.org/10.1162/coli_a_00455.

[36] Carole J Lee, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. "Bias in peer review". In: *Journal of the American Society for information Science and Technology* 64.1 (2013), pp. 2–17.

[37] John Lee and Jonathan Webster. "A Corpus of Textual Revisions in Second Language Writing". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 248–252. URL: https://aclanthology.org/P12-2049.

[38] Mina Lee, Percy Liang, and Qian Yang. "CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities". en. In: *CHI Conference on Human Factors in Computing Systems*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–19. ISBN: 978-1-4503-9157-3. DOI: 10.1145/3491102.3502030. URL: https://dl.acm.org/doi/10.1145/3491102.3502030 (visited on 05/18/2023).

[39] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. *CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society*. Mar. 2023. DOI: 10.48550/arXiv.2303.17760. URL: http://arxiv.org/abs/2303.17760 (visited on 10/27/2023).

[40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". In: *ICML*. 2023.

[41] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. *Can large language models provide useful feedback on research papers? A large-scale empirical analysis*. en. Oct. 2023. URL: http://arxiv.org/abs/2310.01783 (visited on 10/25/2023).

[42] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. *Lost in the Middle: How Language Models Use Long Contexts*. en. July 2023. URL: http://arxiv.org/abs/2307.03172 (visited on 10/26/2023).

[43] Ryan Liu and Nihar B. Shah. *ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing*. June 2023. DOI: 10.48550/arXiv.2306.00622. URL: http://arxiv.org/abs/2306.00622 (visited on 10/26/2023).

[44] Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed H. Awadallah. *On Improving Summarization Factual Consistency from Natural Language Feedback*. en. Dec. 2022. URL: http://arxiv.org/abs/2212.09968 (visited on 02/21/2023).

[45] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. "S2ORC: The Semantic Scholar Open Research Corpus". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 4969–4983. DOI: 10.18653/v1/2020.acl-main.447. URL: https://aclanthology.org/2020.acl-main.447.

[46] Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. "EdiT5: Semi-Autoregressive Text Editing with T5 Warm-Start". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2126–2138. URL: https://aclanthology.org/2022.findings-emnlp.156 (visited on 05/19/2023).

[47] Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. "Encode, Tag, Realize: High-Precision Text Editing". en. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 5053–5064. DOI: 10.18653/v1/D19-1510. URL: https://www.aclweb.org/anthology/D19-1510 (visited on 05/19/2023).

[48] Masato Mita, Keisuke Sakaguchi, Masato Hagiwara, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. *Towards Automated Document Revision: Grammatical Error Correction, Fluency Edits, and Beyond*. en. May 2022. URL: http://arxiv.org/abs/2205.11484 (visited on 02/21/2023).

[49] Michèle B. Nuijten and Joshua R. Polanin. ""statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses". en. In: *Research Synthesis Methods* 11.5 (2020), pp. 574–579. ISSN: 1759-2887. DOI: 10.1002/jrsm.1408. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1408 (visited on 10/27/2023).

[50] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[51] Afshin Oroojlooy and Davood Hajinezhad. "A review of cooperative multi-agent deep reinforcement learning". In: *Applied Intelligence* 53.11 (Oct. 2022), pp. 13677–13722. ISSN: 0924-669X. DOI: 10.1007/s10489-022-04105-y. URL: https://doi.org/10.1007/s10489-022-04105-y (visited on 10/28/2023).

[52] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. *Generative Agents: Interactive Simulacra of Human Behavior*. en. Aug. 2023. URL: http://arxiv.org/abs/2304.03442 (visited on 10/27/2023).

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[54] Guanghui Qin, Yukun Feng, and Benjamin Van Durme. "The NLP Task Effectiveness of Long-Range Transformers". en. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 3774–3790. DOI: 10.18653/v1/2023.eacl-main.273. URL: https://aclanthology.org/2023.eacl-main.273 (visited on 10/26/2023).

[55] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. *CoEdIT: Text Editing by Task-Specific Instruction Tuning*. en. May 2023. URL: http://arxiv.org/abs/2305.09857 (visited on 05/18/2023).

[56] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[57] Machel Reid and Graham Neubig. "Learning to Model Editing Processes". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3822–3832. URL: https://aclanthology.org/2022.findings-emnlp.280.

[58] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. "A Recipe for Arbitrary Text Style Transfer with Large Language Models". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 837–848. DOI: 10.18653/v1/2022.acl-short.94. URL: https://aclanthology.org/2022.acl-short.94 (visited on 05/18/2023).

[59] Stephen Robertson and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". In: *Foundations and Trends in Information Retrieval* 3.4 (2009), pp. 333–389.

[60] Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. *PEER: A Collaborative Language Model*. Aug. 2022. DOI: 10.48550/arXiv.2208.11663. URL: http://arxiv.org/abs/2208.11663 (visited on 01/14/2023).

[61] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. "SciRepEval: A Multi-Format Benchmark for Scientific Document Representations". In: *ArXiv* abs/2211.13308 (2022).

[62] Richard Smith. “Peer Review: A Flawed Process at the Heart of Science and Journals”. In: *Journal of the Royal Society of Medicine* 99.4 (2006), pp. 178–182. DOI: 10.1177/014107680609900414. URL: https://doi.org/10.1177/014107680609900414.

[63] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. en. July 2023. URL: http://arxiv.org/abs/2307.09288 (visited on 11/02/2023).

[64] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. *Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models*. en. Aug. 2023. URL: http://arxiv.org/abs/2308.15022 (visited on 10/26/2023).

[65] Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. “ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis”. en. In: *arXiv:2010.06119 [cs]* (Dec. 2020). URL: http://arxiv.org/abs/2010.06119 (visited on 11/23/2021).

[66] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. *Linformer: Self-Attention with Linear Complexity*. en. June 2020. URL: http://arxiv.org/abs/2006.04768 (visited on 10/26/2023).

[67] Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. “A Comprehensive Survey of Grammatical Error Correction”. In: *ACM Transactions on Intelligent Systems and Technology* 12.5 (Dec. 2021), 65:1–65:51. ISSN: 2157-6904. DOI: 10.1145/3474840. URL: https://dl.acm.org/doi/10.1145/3474840 (visited on 05/19/2023).

[68] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. *Unleashing Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration*. en. July 2023. URL: http://arxiv.org/abs/2307.05300 (visited on 10/26/2023).

[69] Matthew Welsh, Daniel Navarro, and Steve Begg. “Number Preference, Precision and Implicit Confidence”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 33. 2011.

[70] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. *HuggingFace’s Transformers: State-of-the-art Natural Language Processing*. 2020.

[71] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. *Recursively Summarizing Books with Human Feedback*. en. Sept. 2021. URL: http://arxiv.org/abs/2109.10862 (visited on 10/26/2023).

[72] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework*. en. Aug. 2023. URL: http://arxiv.org/abs/2308.08155 (visited on 09/25/2023).

[73] Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. "Memorizing Transformers". en. In: 2022. URL: https://openreview.net/forum?id=TrjbxzRcnf-&continueFlag=ab9fd4c6147dec86186d6bb2eed056b5 (visited on 10/26/2023).

[74] Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. *Retrieval meets Long Context Large Language Models*. en. Oct. 2023. URL: http://arxiv.org/abs/2310.03025 (visited on 10/26/2023).

[75] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. *The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)*. Oct. 2023. DOI: 10.48550/arXiv.2309.17421. URL: http://arxiv.org/abs/2309.17421 (visited on 11/02/2023).

[76] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. "LinkBERT: Pretraining Language Models with Document Links". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8003–8016. DOI: 10.18653/v1/2022.acl-long.551. URL: https://aclanthology.org/2022.acl-long.551 (visited on 04/20/2023).

[77] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. "Wordcraft: Story Writing With Large Language Models". In: *27th International Conference on Intelligent User Interfaces*. IUI '22. New York, NY, USA: Association for Computing Machinery, Mar. 2022, pp. 841–852. ISBN: 978-1-4503-9144-3. DOI: 10.1145/3490099.3511105. URL: https://dl.acm.org/doi/10.1145/3490099.3511105 (visited on 05/18/2023).

[78] Weizhe Yuan and Pengfei Liu. "KID-Review: Knowledge-Guided Scientific Review Generation with Oracle Pre-Training". en. In: *Proceedings of the First MiniCon Conference*. Feb. 2022. (Visited on 05/25/2022).

[79] Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. “A Corpus of Annotated Revisions for Studying Argumentative Writing”. en. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 1568–1578. DOI: 10.18653/v1/P17-1144. URL: http://aclweb.org/anthology/P17-1144 (visited on 04/20/2023).

[80] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. “Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms”. en. In: *Handbook of Reinforcement Learning and Control*. Ed. by Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever. Studies in Systems, Decision and Control. Cham: Springer International Publishing, 2021, pp. 321–384. ISBN: 978-3-030-60990-0. DOI: 10.1007/978-3-030-60990-0_12. URL: https://doi.org/10.1007/978-3-030-60990-0_12 (visited on 10/28/2023).

[81] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. *A Survey of Large Language Models*. en. Sept. 2023. URL: http://arxiv.org/abs/2303.18223 (visited on 11/02/2023).

APPENDIX A

Appendices for chapter 2

A.1. Data Analysis

In this section, we discuss observations about the kinds of edits and review comments we find in our dataset.

A.1.1. Comments

To explore the kinds of comments found in reviews, we asked annotators to categorize extracted comments in the manually-annotated data partition according to what kind of action the comment request from the author, using the following action classes:

- **Compare** a proposed method or resource to a baseline from prior work
- **Apply** a proposed method or theory to a different task or dataset
- **Use** a method from prior work to improve a proposed method
- **Define** a term or symbol
- **Discuss** a related paper
- **Report** an additional metric or analysis for an existing experiment or observation
- **Explain** a detail about the proposed method or finding
- **Remove** something, such as a confusing or misleading claim

The results of our analysis are summarized in Table A.1; 7% of comments did not clearly fit any category and were omitted from this analysis, leaving 182 comments. We observe that

Action	Occurred	Addressed
Explain	42%	39%
Compare	14%	32%
Report	10%	39%
Remove	8%	50%
Apply	8%	33%
Use	8%	33%
Define	7%	47%
Discuss	7%	67%

Table A.1. Rates at which different action classes occurred in comments and the frequency with which they were actually addressed by authors in their revisions.

comments asking to compare with a new baseline, use a new component in a proposed method, or apply the same method to an additional dataset or setting were the least likely to be addressed in revisions. This is likely because those kinds of requests require (potentially substantial) additional experimental work to be done. Requests to define terms or discuss related work were the most commonly addressed, although the small number of comments in those categories means those estimates are likely to be high-variance.

A.1.2. Edits

Most (71%) edits made in response to review comments consist of solely adding a contiguous span of text. Many edits both add and delete text (34%), and very few (2%) consist of only deletions. On average, an edit adds 89 tokens and deletes 12.

A.1.3. Comparison of Synthetic vs. Manually-Annotated Data

We ask whether the synthetically aligned data (subsection 2.4.4) results in different kinds of comments and edits than the manually-annotated data. Surprisingly, we find that the two sets of

data have similar statistics for most of the properties we measure. The synthetic data has the same ratio of edits that add a full new paragraph as opposed to altering an existing paragraph (64% vs 65% new paragraphs for manual and synthetic data).

The main source of potential bias is the number of added tokens in the edits, which trends much higher for synthetic data than manual data (158 vs 89 tokens for manual vs synthetic). This is expected to some degree, due to the minimum matching length used in our synthetic labeling algorithm. This length bias leads to an increased average unigram overlap between review comments and edits, which is higher for synthetic data (8.3%) than for manually-annotated data (6.6%). However, when we control for length by binning comment-edit pairs based on the geometric mean of comment and edit length, the average difference between bins is only 0.4%.

A.2. GPT-4 Prompts

Here we provide the prompts used for the GPT experiments. Because it was only feasible to evaluate the pairwise GPT methods on fully-additive edits, note that we designed the prompts accordingly.

The prompts were created with 10-20 iterations of manual adjustment on a small handful of instances from the training set. We also found in preliminary experiments that the 1-shot GPT pairwise setting did better when the negative example was given first and the positive example second; we suspect that the model is biased towards an alternating `[yes, no, yes, no]` sequence of examples, and because the majority of candidates are negatives it is better to set up the sequence to bias in favor of a "no".

A.2.1. GPT-pairwise (0-shot)

Consider the following review comment for a scientific paper: **<comment>**

Consider the following paragraph, which was added to the paper after the review:

<edit>

Is the new paragraph likely to have been added for the purpose of addressing this review comment? Answer with "yes" or "no".

A.2.2. GPT-pairwise (1-shot)

You need to determine which edits correspond to a given reviewer comment for a scientific paper. Given a comment and a paper edit (where changes are enclosed by brackets with +/- to indicate additions/deletions), you must determine whether the edit was likely added to the paper to address the comment. Here are some examples:

<examples, formatted identically to the main query below, followed by "Answer: yes|no">

Consider the following review comment for a scientific paper: **<comment>**

Consider the following paragraph, which was added to the paper after the review:

<edit>

Is the new paragraph likely to have been added for the purpose of addressing this review comment? Answer with "yes" or "no".

A.2.3. GPT multi-edit

Consider the following comments that a reviewer made about a scientific paper (each followed by a unique comment id):

<comment>

comment id: **<comment id>**

<repeat for all comments>

Below is a partial diff of the original paper text and the paper text after the authors made revisions in response to various reviews. Changes are indicated with brackets "[]" with a "+" for additions and a "-" for deletions. Below each paragraph is a unique "edit id". Determine which edits were meant to address the given reviewer comments above.

—BEGIN PAPER DIFF—

<edit>

section: **<section name>**

edit id: **<edit id>**

<repeat above until out of edits or reached model token limit>

—END PAPER DIFF—

Which edit ids correspond to each of the reviewer's comments? The relationship is many-to-many; one comment could correspond to several edits, and several comment could correspond to the same edit. There could also be comments that the authors didn't address at all or edits that were not made in response to any particular comment.

Write the answer as JSON lines with the format {"comment_id": <comment id>, "edit_ids": [<edit ids>], "notes": ""} where each record has a comment id and the list of edit ids that correspond to it. The "notes" field is optional and can contain any notes about edits you weren't sure about or reasons for including/omitting certain edits.

While the "notes" field in the prompt was included in the hope that it might provide insight into the model's reasoning and assist in diagnosing its errors, in practice we found that the notes were usually not very informative (e.g., *"this comment was not addressed by the edits"*), and therefore we did not do a formal analysis of the model's notes.

A.2.4. GPT edit generation

Consider the following excerpt of a scientific paper which is under review for a conference:

— START —

Abstract: **<abstract>**

Body: **<sequence of body paragraphs>**

— END —

A reviewer made the following comment about the paper: **<comment>**

Write a response to the reviewer and an edit (or edits) that could be added somewhere in the paper (or Appendix) to resolve the reviewer's comment. Above an edit, write the location in the paper where it should be made. The edit should not explicitly say that it is written in response to a reviewer comment; it just needs to improve the paper such that a future reviewer would be unlikely to make the same comment. If addressing the comment requires additional experiments or information that you do not have access to, you can use placeholders or fill in reasonable guesses for that information. An edit may be a new sentence, paragraph, or section, depending on the comment.

For ease of parsing, write "Response:" before the reviewer response, "Location:" before the edit location(s), and "Edit:" before the edit(s).

The above prompt asks for an author response, which we did not discuss in the main paper. During preliminary prompt tuning, we found that the model had a tendency to phrase its edits as though it was writing a response directly to the reviewer (often including a phrase along the lines

of, "as the reviewer suggests, we ..."). Encouraging the model to write a direct response separate from the paper edit appeared to mitigate this tendency and improve the quality of the edits.

A.3. Edit extraction details

In this section we provide additional details on how we align text between drafts to construct a list of edits.

To perform initial alignment of drafts, we create a mapping $m(j) \rightarrow i$ from a paragraph t_j in the target document to paragraph s_i in the source document. For each pair, we score the similarity as

$$\text{sim}(i, j) = \text{bigram}(s_i, t_j) - \frac{|m(j-1) + 1 - i|}{|S|}$$

where $|S|$ is the number of paragraphs in the source document and bigram() is the bigram overlap. The term $\frac{|m(j-1)+1-i|}{|S|}$ is used to encourage matching to a paragraph close to where the previous paragraph was matched to. We take the most similar match if $\text{sim}(i, j) > 10\%$, otherwise we consider it a new paragraph. Unmapped source paragraphs are considered to be deleted.

In some cases, PDF parsing errors cause a paragraph to be split differently in different drafts. E.g., the paragraph may be broken across a page boundary differently, so the parser thinks it is two paragraphs. To prevent this from resulting in spurious "edits", we post-process all matches to check if the similarity would become higher by merging s_i or t_j with an adjacent paragraph. If so, we merge them.

A.4. Additional annotation information

To extract actionable comments from reviews, annotators were shown the review text in a web-based annotation interface where they could highlight spans corresponding to comments.

The annotators were instructed to select comments based on the definition in subsection 2.4.2. That is, any comments which imply some action should be performed to improve the paper are included, but non-actionable comments such as summaries, positive comments, or comments too vague and fundamental to be addressable (*"Lacks novelty."*) are excluded. As a rule of thumb for unclear cases, a comment was included if the annotators could imagine rewriting it as an imperative to-do list item.

The interface allowed for selecting arbitrary token spans, but in practice almost all spans aligned roughly to sentence boundaries. The majority (78%) of extracted comments were one sentence long and some (19%) were two sentences long, with only 4% being more than two sentences.

A.5. Implementation details

For training models on the comment-edit alignment task, we sample 20 negative edits for each comment. The negative edits are sampled from the pool of training documents excluding the one to which the comment applies, to mitigate the low recall of synthetic data.

For all methods, we exclude edited paragraphs with fewer than 100 characters, as shorter ones are often badly parsed equations or text fragments that appear erroneously as edits.

All neural models are trained on NVIDIA Quadro RTX 8000 and RTX A6000 GPUs. We use the Adam optimizer [31] with a learning rate of 2e-5, batch size of 16, and β of [0.9, 0.999], running for a maximum of 8192 steps and selecting the best model on the dev set. The experiments are implemented using Pytorch 1.10 [53] and Huggingface Transformers 4.21 [70] for transformer models and Gensim 4.3 [56] for BM25. For GPT-4, we use the `gpt-4-0314` model and use a temperature of zero in all experiments.

A.6. Comment-source alignment

While comment-edit alignment maps a comment to a corresponding edit made to the paper, *comment-source alignment* maps a comment to the corresponding source paragraph where an edit should be made. Good comment-source alignment would make it possible to break down the edit generation task into separate location-finding and edit-generation stages, which could improve efficiency and accuracy.

Comment-source alignment is more challenging than comment-edit alignment because it doesn't provide information about the content of the edit. It is also more ambiguous: there are sometimes multiple places where an edit could potentially be made to address a comment, and we only use the location of the real edit as a correct answer. Nonetheless, in this section we apply our baseline models to the task to quantify its difficulty.

We use the same baseline models as in section 2.5, excluding GPT-4 due to its high cost. To reduce the degree of ambiguity in the task, we reduce our dataset to include only comment-edit pairs that correspond to specific source paragraphs (excluding newly-added paragraphs).

A.6.1. Results

Results are reported in Table A.2. While results are not directly comparable to comment-edit alignment due to the different dataset split, the micro-f1 results are all substantially lower; macro scores are higher, but this is likely the result of models being biased towards predicting negatives and therefore getting more perfect scores on comments that have no aligned edits.

	Micro			Macro		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
BM25	5.0	3.5	4.1	81.9	75.1	61.1
Bi-encoder						
Specter no-finetune	3.5	22.8	6.1	51.8	80.6	43.1
Specter	3.0	13.5	4.8	60.0	77.5	44.6
DeBERTa	0.4	0.6	0.5	68.0	74.2	50.7
Cross-encoder						
LinkBERT	1.4	0.6	0.8	72.8	74.3	53.8
DeBERTa	2.3	7.6	3.0	75.7	41.6	40.9

Table A.2. Precision (P), Recall (R), and F1 of comment-source alignment on test data. The micro-average is over all comment-edit pairs, while the macro-average is grouped by comment.

APPENDIX B

Appendices for chapter 3

B.1. Prompts

B.1.1. MARG-S

MARG-S: Leader agent system prompt

```
You are part of a group that needs to perform tasks that involve a scientific paper.
However, the paper is very long, so each agent has only been given part of it.  You are
the leader in charge of interacting with the user and coordinating the group to
accomplish tasks.  You will need to collaborate with other agents by asking questions or
 giving instructions, as they are the ones who have the paper text.

Communication protocol:
To broadcast a message other agents, write "SEND MESSAGE: " and then your message;
alternatively, if you forget to include it until the end of your message, you can write
"SEND FULL MESSAGE" and everything you just wrote will be sent.  This will be a common
failure, so if other agents remark that you didn't include some information, check that
you used the right version of SEND MESSAGE, and consider using SEND FULL MESSAGE instead
.

Additional instructions:
When you are given a task, your first step should be to draft a high-level plan with a
list of steps, concisely describing how you will approach the task and your strategy for
 communicating with other agents.  Then, execute the plan.  When executing the plan,
write the current step you are working on each time you move to the next step, to remind
 yourself where you are.  You are allowed to create a sub-plan for a step if it is
complicated to do in one pass.

You should continue to pay attention to details in the original task instructions even
after you draft your plan.  Optionally, it may be helpful to share a plan with other
agents to help guide them in some cases.

Other agents do not know anything about the task being performed, so it is your
responsibility to convey any information about the task that is necessary for them to
provide helpful responses.  You should make this part of your high-level plan.
Depending on the task, you may need to do multiple rounds of communication to exchange
all the necessary information; you should follow up with other agents if they provide a
bad response or seem to have misunderstood the task.  In addition, because other agents
can only communicate with you but not each other, you may need to help relay information
 between agents.

Because each agent has a different piece of the paper, communication is key for
performing tasks that require understanding the full paper.  In addition, depending on
the responses you receive, you may need to ask follow-up questions, clarify your
requests, or engage in additional discussion to fully reason about the task.

To reduce communication errors, after you send a message you should write a short
description of what you expect the response to look like.  If the response you get doesn
't match your expectation, you should review it and potentially ask follow-up questions
to check if any mistakes or miscommunications have occurred.  It could be the case that
an agent (including yourself) has misread something or made a logic error.
```

MARG-S: Worker system prompt

```
You are part of a group that needs to perform tasks that involve a scientific paper.
However, the paper is very long, so each agent has only been given part of it.  The
leader of the group is Agent 0, who will coordinate with the user and convey questions
or task instructions to you.

Sometimes you will need more information in order to understand a question or task or to
 interpret your portion of the paper; in these cases, you should send a message to
request this information from other agents.  For example, if there are key terms that
you don't know the definitions for or parts of the paper chunk that you are missing
important context for, you might need to ask for more information in order to understand
 it.  In addition, if a message or request you receive is unclear or does not seem
relevant to you, you should explain your confusion and request any additional
clarification needed.

Communication protocol:
To send a message to the group leader, write "SEND MESSAGE: " and then your message.
Include all necessary information, but be concise; do not include any extra greetings or
 commentary.

To reduce communication errors, after you send a message you should write a short
description of what you expect the response to look like.  If the response you get doesn
't match your expectation, it is not necessarily wrong, but you should review it and
potentially ask follow-up questions to ensure that no mistakes or miscommunications have
 occurred.

Because the leader always broadcasts messages to all agents, you might sometimes get
messages that aren't relevant to you; in this case, just respond with "This doesn't seem
 relevant to me, so I will stand by for further instructions.".  However, if the message
 contains information that contradicts information in your part of the paper, you should
 respond and mention the issue, even if the message wasn't directed at you.  In addition
, you should be aware that sometimes the leader accidentally leaves some information out
 from its messages, so if a message looks like it might be directed at you but is simply
 incomplete, you should ask follow-up questions to confirm.
```

MARG-S: Worker chunk prompt

```
Your paper chunk is shown below:
--- START PAPER CHUNK ---
{paper_chunk}
--- END PAPER CHUNK ---

Information about agents: There are {num_agents} agents in the group, including yourself
.  You are {agent_name}.  The other agent(s) are: {other_agent_names}.

Write "Ready" if you have understood the assignment.  You will then receive messages.
```

MARG-S (experiments): Leader task prompt

```
Task: Write a list of feedback comments, similar to the suggestions a reviewer might make.  In addition, focus on major comments rather than minor comments; major comments are important things that affect the overall impact of the paper, whereas minor comments are small things like style/grammar or small details that don't matter much for whether the paper should be accepted to a venue.

Be specific in your suggestions, including details about method or resource names and any particular steps the authors should follow. However, don't suggest things that have already been included or addressed in the paper.  Remember that you can collaborate if necessary, but also remember that other agents can't see anything you write prior to "SEND MESSAGE", so you may need to repeat information so that they are aware of it.  For example, if you write some comments and ask for additional ones, you may want to provide your original comments so that the agent knows what they are.

Your review comments should be specific and express an appropriate level of importance.  For example, suppose a paper is missing some important details needed to understand a proposed method.  A comment like "The authors could add more details about the proposed method, such as XYZ." is bad because it is too generic; even for a paper with a good method description it is always possible to add more details, so it isn't clear if there is actually a significant problem with the current paper.  Instead, in this scenario it is much better to leave a comment like "The description of the proposed method is unclear because it is missing some key details such as XYZ.  Without these details it is hard to know whether ___.".  Make sure your high-level plan mentions this instruction.

Some comments are a matter of degree.  For example, maybe the paper includes one baseline but no others; you would need to determine whether or not that is acceptable for meeting the goals of the paper and supporting its claims, and decide whether it is important enough to leave a comment about.  You can discuss with other agents as needed to help determine this.

You will need to communicate with other agents to understand the paper and learn what has already been addressed and what is still missing from the paper.

The main type of feedback you should focus on the thoroughness of the experiments and consistency of claims.  You should ensure that information is consistent across the paper and that claims are appropriately supported by evidence.  Your high-level plan should be roughly as follows:
1. Identify the main goals, contributions, and claims of the paper.  What questions is the paper trying to answer, and why are those questions important or interesting?  What findings does it contribute to the field?
    a. Go through the paper paragraph by paragraph and write down anything that looks like it might be part of the main goals or contributions, and ask other agents to do the same.
    b. Put all the information together, filtering out anything that turned out to be unimportant and merging similar points.  This should result in a concise list of summarized claims.
2. Identify expectations for fulfilling the goals and claims.  For this part, you should collaborate closely with the experiment design expert.  Give them information about the paper's topic and the claims and goals you summarized in the previous step, and explain the task so that they can help you.  Remember to put the information after SEND MESSAGE so that it gets sent correctly.  Note that other agents will see your message and may try to respond despite not being the expert; you should make it clear that you only want to communicate with the expert, and only respond to the true expert's messages.  During this step, you must obey all of the expert's instructions and answer all of their questions.  The expert is {expert_2}.
    a. Come up with a clear description of experiments, analyses, and ablations that you would use to verify the paper's claims if you were doing the study yourself.  Be specific and detailed in your description; what experiments should be conducted, how should they be set up, and why are they helpful for verifying the claims?
3. Check whether the paper matches your expectations
    a. Go through the actual evaluations and experiments in the paper and identify the similarities and differences between them and your experiment description.  Make sure to pay careful attention to details.  This will require communication with other agents to collect all the necessary information.  If agents do not provide all the needed information or if something is ambiguous, you must send additional messages to resolve the communication issues.
    b. For each way the paper's experiments don't match your expectations, determine if this constitutes a shortcoming of the paper, or if the paper's experiments still fulfill the goals and claims of the paper.  It may be helpful to share your thoughts, the claims, the expected experiments, and the real experiments with other agents and get their opinions on whether the paper's experiments fall short.
    b. If the paper's experiments are suboptimal or inadequate, write a feedback comment explaining the shortcoming and what the authors should do to resolve the issue.  Be detailed and specific in your feedback to make it clear what the authors should do and why the suggestion is important.
```

MARG-S (experiments): Expert prompt

```
You are part of a group of agents that must perform tasks involving a scientific paper.
 You are an expert scientist that designs high-quality experiments, ablations, and
analyses for scientific papers.  When the leader sends a message to you to ask for
assistance in coming up with experiments to include in a paper or judging the quality of
 experiments that are in a paper, you should help.

You should ensure that you fully understand the claims and goals of the paper before
giving suggestions.  You can send messages back to the leader to ask questions about the
 paper's claims, goals, methods, and so on.  It is crucial to understand what the paper
is attempting to investigate in order to design experiments to support the investigation
.  Obtain any information you need in order to design good experiments, and ask follow
up questions if needed.

Be detailed and specific in the experimental suggestions you give.  What should the
setup be?  What settings or methods should be compared?  What metrics or measurement
techniques should be used?  How should the results be analyzed?  Make it clear which
specific details are important and why (e.g., particular choices of settings, baselines,
 metrics, environments, procedures, and so on), and which details are unimportant.

If you are asked to check the quality of an existing experimental procedure, one useful
approach is to come up with how you would have conducted the experiments and compare the
 given approach to that in order to generate potential areas for improvement.  If you
find a shortcoming, explain the issue clearly: why is the existing experiment misleading
 or why does it fail to fulfill the goals of the investigation?

Finally, note that you may receive messages from the group leader that are not relevant
to you.  This is because the group leader always broadcasts all messages to all agents.
 If you get an irrelevant message, simply respond by saying "I do not believe the
request is relevant to me, as I do not have a paper chunk.  I will stand by for further
instructions.".
```

MARG-S (impact): Leader task prompt

```
Task: Write a list of feedback comments, similar to the suggestions a reviewer might
make.  The main type of feedback you should focus on is the novelty and significance of
the work.  The motivations, goals, and key findings of the paper need to be clearly
explained, and the paper needs to explain how it fits into the related literature in the
 field and how it builds and expands on this work in a meaningful way.  If any of those
things are unclear or missing from the paper, you should comment on them.

Once you have established what the motivations, goals, and key findings of the paper are
, you should carefully scrutinize whether they are reasonable and well-justified or if
they need to be improved.  For example, if a paper proposes a new method that is
motivated by real-world use cases, but requires unrealistic assumptions to operate, the
paper needs to justify that somehow.

Important: {expert_1} doesn't have a paper chunk, but they are good at coming up with
questions and potential shortcomings of the paper's assumptions.  Explain the paper to {
expert_1} and answer any questions they have until they say they are finished.  You will
 likely need to pass their questions and comments along to the other agents that have
the paper, and pass the answers back to the expert.  Write feedback based on any points
{expert_1} indicates are in need of improvement.

Think carefully in a logical, step-by-step way.  Ask questions or give instructions to
other agents to help you accomplish the task, including follow-up questions or requests
as needed.  Write potential feedback comments as you come up with them so that you can
keep them in mind; you can always remove or revise them later for the final list.
```

MARG-S (impact): Expert prompt

You are part of a group of agents working with a scientific paper. You are highly curious and skeptical of papers, and your job is to help ensure that the paper has clearly explained its motivations, goals, and key findings and determine whether the paper actually makes a significant contribution to its field. The group leader will give you a summary of the paper, and you should ask questions to fully understand the paper's motivations, goals, and key findings. This includes asking follow-up questions as needed.

Scrutinize the paper heavily, identifying any hidden assumptions or potential issues that could undermine the paper's claimed goals and motivations. For example, suppose a paper proposes a robot navigation algorithm that implicitly works only with omnidirectional instantly-accelerating robots; a questionable hidden assumption in this case would be that real-world robots can effectively be treated as omnidirectional, which is often untrue. It would be important for the authors to provide some kind of justification for the assumption in this case (for example, that there exist robots that can turn in place and accelerate quickly enough to be treated as omnidirectional in practice). Keep in mind that the issues might not be so obvious in practice, so you should think carefully and explore multiple perspectives and possibilities.

Think of the kinds of questions a scientific paper reviewer might ask, or what they might suggest is confusing or poorly justified in the paper.

Always make sure that you understand the terms and concepts used in the paper. If you are unsure about the definition of a term or how it is meant to be interpreted in a particular context, you should ask about it, as it is important for the paper to explain such things.

You will communicate with the group leader, who in turn will handle communications with other agents who have the paper itself. Because the leader always broadcasts messages to all agents, you might sometimes get messages that aren't relevant to you; in this case, just respond with "This doesn't seem relevant to me, so I will stand by for further instructions.". However, if you have asked questions and it doesn't seem like the leader is responding or trying to get information from other agents so that it can respond to you, you should interject and tell the leader that they need to answer you.

When you are done talking with the group leader, tell them that you are done with your review, and give them a summary list of any missing information, poorly justified points, or other suggestions that you identified.

MARG-S (clarity): Leader task prompt

```
Task: Write a list of feedback comments, similar to the suggestions a reviewer might
make.  The main type of feedback you should focus on is the clarity and reproducibility
of the work.  The methods, experimental settings, and key concepts of the paper need to
be clearly explained, and the paper needs to provide enough context and background
information for anyone with general experience in the field to understand it.  If any of
 those things are unclear or missing from the paper, you should comment on them.

Once you have established what the methods, experiments, and key concepts of the paper
are, you should carefully scrutinize whether they are clearly explained and detailed or
if they need to be improved.

Important: {expert_1} doesn't have a paper chunk, but they are good at coming up with
questions that test the paper's clarity.  Explain the paper to {expert_1} and answer any
 questions they have until they say they are finished.  You will likely need to pass
their questions and comments along to the other agents that have the paper, and pass the
 answers back to the expert.  Write feedback based on any points {expert_1} indicates
are in need of improvement.

Think carefully in a logical, step-by-step way.  Ask questions or give instructions to
other agents to help you accomplish the task, including follow-up questions or requests
as needed.  Write potential feedback comments as you come up with them so that you can
keep them in mind; you can always remove or revise them later for the final list.
```

MARG-S (clarity): Expert prompt

```
You are part of a group of agents working with a scientific paper.  You are highly
curious and have incredible attention to detail, and your job is to help ensure that the
 paper has clearly explained its methods, experimental settings, and key concepts and
determine whether the paper is well-organized and can be easily understood and
reproduced.  The group leader will give you a summary of the paper, and you should ask
questions to fully understand the paper's methods, experimental settings, and key
concepts.  This includes asking follow-up questions as needed.

Scrutinize the paper heavily, identifying any missing details or potential issues that
could make it ambiguous or hard to understand.  Keep in mind that the issues might not
be so obvious in practice, so you should think carefully and explore multiple
perspectives and possibilities.  In particular, make sure the paper provides all
information necessary to implement any proposed methods, including any information on
any background concepts needed to understand how the methods work.  Also ensure that the
 paper provides enough information to replicate the experimental settings, including any
 hyperparameters, equipment and material specifications, or other implementation details
.

Think of the kinds of questions a scientific paper reviewer might ask, or what they
might suggest is confusing or poorly explained in the paper.

Always make sure that you understand the terms and concepts used in the paper.  If you
are unsure about the definition of a term or how it is meant to be interpreted in a
particular context, you should ask about it, as it is important for the paper to explain
 such things.

You will communicate with the group leader, who in turn will handle communications with
other agents who have the paper itself.  Because the leader always broadcasts messages
to all agents, you might sometimes get messages that aren't relevant to you; in this
case, just respond with "This doesn't seem relevant to me, so I will stand by for
further instructions.".  However, if you have asked questions and it doesn't seem like
the leader is responding or trying to get information from other agents so that it can
respond to you, you should interject and tell the leader that they need to answer you.

When you are done talking with the group leader, tell them that you are done with your
review, and give them a summary list of any missing or misleading information, ambiguous
 statements, poorly organized points, or other suggestions that you identified.
```

MARG-S: Refinement prompt

```
Refine and improve the following review comment that was written about a scientific
paper.  The goal is for the comment to be detailed and helpful, similar to a comment
that a scientific paper reviewer might write.  The comment should not ask for things
that are already in the paper, it should include enough detail for an author to know
clearly how to improve their paper, the purpose and value of the suggestion should be
clearly justified, and so on.  Remove the comment if it is bad (i.e., if it fails to
meet those criteria).  You may need to incorporate additional information in the paper
to refine the comment.  You should focus on "major" comments that are important and have
 a significant impact on the paper's quality, as opposed to minor comments about things
like writing style or grammar.  If the comment you are given is minor, express this fact
 as part of the revised comment.

Your revised review comment should be specific and express an appropriate level of
importance.  For example, suppose a paper is missing some important details needed to
understand a proposed method.  A comment like "The authors could add more details about
the proposed method, such as XYZ." is bad because it is too generic; even for a paper
with a good method description it is always possible to add more details, so it isn't
clear if there is actually a significant problem with the current paper.  Instead, in
this scenario it is much better to leave a comment like "The description of the proposed
 method is unclear because it is missing some key details such as XYZ.  Without these
details it is hard to know whether ___.".  Make sure your high-level plan references
this instruction.

Note that only you are being given the comment; you will need to share it with other
agents if you want them to have context.  When receiving responses, it may be helpful to
 first summarize the findings from all agents before applying the information to the
review comment.

Some comments are a matter of degree.  For example, maybe the paper includes one
baseline but no others; you would need to determine whether or not that is acceptable
for meeting the goals of the paper and supporting its claims, and decide whether it is
important enough to leave a comment about.  You can discuss with other agents as needed
to help determine this.

It may be helpful to work step-by-step examining one aspect of the comment at a time and
 considering what information is needed to verify that it is valid and important as well
 as what kind of clarification and rewording could help to make it clearer and more
specific.

Here is the comment:
{review_comments}
```

B.1.2. SARG-B

SARG-B: System prompt

```
You are ReviewGPT, an expert scientific paper reviewer.
```

SARG-B: Task prompt

```
Write feedback comments in the style of a scientific paper review for the following
portion of a scientific paper.  You can skip minor grammar comments.
--- START PAPER CHUNK ---
{paper_chunk}
--- END PAPER CHUNK ---
```

B.1.3. SARG-TP

SARG-TP: System prompt

```
You need to perform tasks that involve a scientific paper.  When you are given a task,
your first step should be to draft a high-level plan, concisely describing how you will
approach the task.  Then execute that plan.
```

SARG-TP: Chunk prompt

```
A chunk of text from a scientific paper is shown below:
--- START PAPER CHUNK ---
{paper_chunk}
--- END PAPER CHUNK ---

Write "Ready" if you have understood the assignment.  You will then be given tasks.
```

SARG-TP: Task prompt

Task: Write a list of feedback comments, similar to the suggestions a reviewer might make. Focus on major comments rather than minor comments; major comments are important things that affect the overall impact of the paper, whereas minor comments are small things like style/grammar or small details that don't matter much for whether the paper should be accepted to a venue.

Be specific in your suggestions, including details about method or resource names and any particular steps the authors should follow. However, don't suggest things that have already been included or addressed in the paper.

Your review comments should have a clear purpose; obviously, it is always possible to simply say the authors should include more details or do more experiments, but in practice the authors have limited space to write and limited time to work, so each comment needs to have a clear purpose.

B.1.4. MARG-TP

MARG-TP: Leader system prompt

```
You are part of a group that needs to perform tasks that involve a scientific paper.
However, the paper is very long, so each agent has only been given part of it.  You are
the leader in charge of interacting with the user and coordinating the group to
accomplish tasks.  You will need to collaborate with other agents by asking questions or
 giving instructions, as they are the ones who have the paper text.

Communication protocol:
To broadcast a message other agents, write "SEND MESSAGE: " and then your message;
alternatively, if you forget to include it until the end of your message, you can write
"SEND FULL MESSAGE" and everything you just wrote will be sent.  This will be a common
failure, so if other agents remark that you didn't include some information, check that
you used the right version of SEND MESSAGE, and consider using SEND FULL MESSAGE instead
.

Additional instructions:
When you are given a task, your first step should be to draft a high-level plan with a
list of steps, concisely describing how you will approach the task and your strategy for
 communicating with other agents.  Then, execute the plan.  When executing the plan,
write the current step you are working on each time you move to the next step, to remind
 yourself where you are.  You are allowed to create a sub-plan for a step if it is
complicated to do in one pass.

You should continue to pay attention to details in the original task instructions even
after you draft your plan.  Optionally, it may be helpful to share a plan with other
agents to help guide them in some cases.

Other agents do not know anything about the task being performed, so it is your
responsibility to convey any information about the task that is necessary for them to
provide helpful responses.  You should make this part of your high-level plan.
Depending on the task, you may need to do multiple rounds of communication to exchange
all the necessary information; you should follow up with other agents if they provide a
bad response or seem to have misunderstood the task.  In addition, because other agents
can only communicate with you but not each other, you may need to help relay information
 between agents.

Because each agent has a different piece of the paper, communication is key for
performing tasks that require understanding the full paper.  In addition, depending on
the responses you receive, you may need to ask follow-up questions, clarify your
requests, or engage in additional discussion to fully reason about the task.

To reduce communication errors, after you send a message you should write a short
description of what you expect the response to look like.  If the response you get doesn
't match your expectation, you should review it and potentially ask follow-up questions
to check if any mistakes or miscommunications have occurred.  It could be the case that
an agent (including yourself) has misread something or made a logic error.
```

MARG-TP: Worker system prompt

```
You are part of a group that needs to perform tasks that involve a scientific paper.
However, the paper is very long, so each agent has only been given part of it.  The
leader of the group is Agent 0, who will coordinate with the user and convey questions
or task instructions to you.

Sometimes you will need more information in order to understand a question or task or to
 interpret your portion of the paper; in these cases, you should send a message to
request this information from other agents.  For example, if there are key terms that
you don't know the definitions for or parts of the paper chunk that you are missing
important context for, you might need to ask for more information in order to understand
 it.  In addition, if a message or request you receive is unclear or does not seem
relevant to you, you should explain your confusion and request any additional
clarification needed.

Communication protocol:
To send a message to the group leader, write "SEND MESSAGE: " and then your message.
Include all necessary information, but be concise; do not include any extra greetings or
 commentary.

To reduce communication errors, after you send a message you should write a short
description of what you expect the response to look like.  If the response you get doesn
't match your expectation, it is not necessarily wrong, but you should review it and
potentially ask follow-up questions to ensure that no mistakes or miscommunications have
 occurred.
```

MARG-TP: Worker chuk prompt

```
Your paper chunk is shown below:
--- START PAPER CHUNK ---
{paper_chunk}
--- END PAPER CHUNK ---

Information about agents: There are {num_agents} agents in the group, including yourself
.  You are {agent_name}.  The other agent(s) are: {other_agent_names}.

Write "Ready" if you have understood the assignment.  You will then receive messages.
```

MARG-TP: Task prompt

```
Task: Write a list of feedback comments, similar to the suggestions a reviewer might
make.  Focus on major comments rather than minor comments; major comments are important
things that affect the overall impact of the paper, whereas minor comments are small
things like style/grammar or small details that don't matter much for whether the paper
should be accepted to a venue.

Be specific in your suggestions, including details about method or resource names and
any particular steps the authors should follow.  However, don't suggest things that have
 already been included or addressed in the paper.

Your review comments should have a clear purpose; obviously, it is always possible to
simply say the authors should include more details or do more experiments, but in
practice the authors have limited space to write and limited time to work, so each
comment needs to have a clear purpose.
```

B.2. Example multi-agent interaction

In this section, we show example conversation logs from MARG-S (experiments), MARG-S (clarity), MARG-S (impact), and the refinement stage. Note that for the leader agent, the full output is shown, but only the portion following the `SEND MESSAGE` marker is actually sent to other agents. Messages consisting of the no-response indicator have been pruned.

B.2.1. MARG-S (experiments)

Agent 0 (leader): High-Level Plan:
1. Identify the main goals, contributions, and claims of the paper.
a. Ask each agent to summarize the main goals, contributions, and claims from their section of the paper.
b. Compile and filter the responses to create a concise list of the paper's main goals, contributions, and claims.
2. Identify expectations for fulfilling the goals and claims.
a. Communicate with Agent 4 (the experiment design expert) to develop a clear description of the experiments, analyses, and ablations that would be necessary to verify the paper's claims.
3. Check whether the paper matches the expectations.
a. Ask each agent to describe the evaluations and experiments conducted in their section of the paper.
b. Compare the actual experiments with the expected experiments and identify any discrepancies.

c. Discuss any discrepancies with the agents to determine if they constitute shortcomings of the paper.
d. Write feedback comments for any identified shortcomings, explaining the issue and suggesting how the authors could resolve it.

Now, let's execute the plan.

Step 1a: Ask each agent to summarize the main goals, contributions, and claims from their section of the paper.

SEND MESSAGE: Please provide a summary of the main goals, contributions, and claims from your section of the paper. Be as specific as possible.

Expected response: Each agent should provide a summary of the main goals, contributions, and claims from their section of the paper.

Agent 1 (worker): The main goal of the paper is to introduce a new Graph Neural Network (GNN) framework, DIGRAC, for node clustering in directed networks. The method is based on maximizing flow imbalance between clusters, which is a novel approach compared to traditional methods that focus on edge density. The paper claims that this approach can reveal clusters even when there is no density difference between them. The main contributions of the paper include the introduction of a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA). The paper also claims that DIGRAC can naturally incorporate node features, unlike existing spectral methods, and does not require label supervision, unlike existing GNN methods. Experimental results are presented to demonstrate the effectiveness of the method.

Expected response: Confirmation from Agent 0 that the summary is accurate and complete, or requests for additional information or clarification.

Agent 2 (worker): The main goal of my section of the paper is to describe the method of Directed Graph Clustering (DIGRAC) and its implementation, as well as the experiments conducted to test its efficacy. The paper introduces a self-supervised loss for clustering and a directed mixed path aggregation scheme (DIMPA) to implement DIGRAC. The method aims to find a partition that maximizes the imbalance flow to capture clusters in the meta-graph. The paper also introduces a global probabilistic imbalance score and discusses different variants for selecting pairs of clusters. The DIMPA scheme is used to build node embeddings and obtain the probability assignment matrix. The paper also discusses the experiments conducted to test the efficacy of DIGRAC, using both synthetic and real-world data sets. The performance of DIGRAC is measured using the Adjusted Rand Index (ARI) and accuracy for node clustering, node classification, and link direction prediction. The paper claims that DIGRAC outperforms its competitors, especially when there exist ambient nodes.

EXPECTED RESPONSE: Confirmation of receipt and understanding of the summary, and potentially

further instructions or questions.

Agent 3 (worker): The main goals of my section of the paper are to present the experimental results of the DIGRAC model and to discuss future work and potential improvements. The experimental results show that the DIGRAC model performs well in node clustering tasks, especially when using the "sort" loss function and when additional label information is available. The future work section discusses potential improvements to the model, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications. The paper also suggests that the model could be improved by considering more pairwise imbalance scores and by using regularization in sparse settings.

Expected response: Confirmation from Agent 0 that my summary is accurate and complete, or requests for additional information or clarification.

Agent 4 (expert): I do not believe the request is relevant to me, as I do not have a paper chunk. I will stand by for further instructions.

Agent 0 (leader): Step 1b: Compile and filter the responses to create a concise list of the paper's main goals, contributions, and claims.

Based on the responses from the agents, the main goals, contributions, and claims of the paper are:

1. The paper introduces a new Graph Neural Network (GNN) framework, DIGRAC, for node clustering in directed networks.
2. The method is based on maximizing flow imbalance between clusters, which is a novel approach.
3. The paper introduces a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA).
4. DIGRAC can naturally incorporate node features and does not require label supervision.
5. The paper presents experimental results to demonstrate the effectiveness of the method.
6. The paper discusses potential improvements to the model, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications.

Step 2a: Communicate with Agent 4 (the experiment design expert) to develop a clear description of the experiments, analyses, and ablations that would be necessary to verify the paper's claims.

SEND MESSAGE: Agent 4, based on the main goals, contributions, and claims of the paper, could you help us develop a clear description of the experiments, analyses, and ablations that would be necessary to verify the paper's claims? Here are the main points:

1. The paper introduces a new Graph Neural Network (GNN) framework, DIGRAC, for node clustering in directed networks.
2. The method is based on maximizing flow imbalance between clusters, which is a novel approach.

3. The paper introduces a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA).
4. DIGRAC can naturally incorporate node features and does not require label supervision.
5. The paper presents experimental results to demonstrate the effectiveness of the method.
6. The paper discusses potential improvements to the model, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications.

Expected response: Agent 4 should provide a detailed description of the experiments, analyses, and ablations that would be necessary to verify the paper's claims.

Agent 4 (expert): Based on the main goals, contributions, and claims of the paper, here are the experiments, analyses, and ablations that would be necessary to verify the paper's claims:

1. **Baseline Comparison**: To validate the effectiveness of the DIGRAC framework, it should be compared with existing state-of-the-art methods for node clustering in directed networks. This comparison should include both traditional methods and other GNN-based methods. The comparison should be done on multiple datasets with varying characteristics to ensure the robustness of the results.

2. **Ablation Study**: To understand the contribution of each component of the DIGRAC framework, an ablation study should be conducted. This involves removing or modifying one component at a time (e.g., the probabilistic imbalance loss, the DIMPA scheme) and observing the impact on the performance. This will help to understand the importance of each component and how they contribute to the overall performance.

3. **Feature Importance Analysis**: Since DIGRAC can naturally incorporate node features, an analysis should be conducted to understand the importance of different features. This can be done by using feature importance methods such as permutation importance or SHAP values.

4. **Scalability Analysis**: To validate the claim that DIGRAC can be adapted for large scale applications, experiments should be conducted on large datasets. The scalability can be evaluated in terms of both computational time and memory usage.

5. **Semi-supervised Learning Experiment**: To validate the potential improvement of semi-supervised learning, experiments should be conducted where a small amount of labeled data is used in addition to the unlabeled data. The performance of the semi-supervised learning approach can be compared with the unsupervised learning approach.

6. **Automatic Cluster Detection Experiment**: To validate the potential improvement of automatic detection of the number of clusters, experiments should be conducted where the number of clusters is not provided as input. The performance of the automatic cluster detection approach can be compared with the approach where the number of clusters is provided.

The performance of the methods can be evaluated using standard clustering metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and F1 score. The statistical significance of the results should be tested using appropriate statistical tests.

Agent 0 (leader): Step 3a: Ask each agent to describe the evaluations and experiments conducted in their section of the paper.

SEND MESSAGE: Please describe the evaluations and experiments conducted in your section of the paper. Be as specific as possible.

Expected response: Each agent should provide a detailed description of the evaluations and experiments conducted in their section of the paper.

Agent 1 (worker): My section of the paper does not contain specific details about the evaluations and experiments conducted. It mainly discusses the problem definition, the proposed DIGRAC method, and the self-supervised loss for clustering. However, it does mention that experimental results on synthetic data and real-world data at different scales demonstrate that the method can achieve state-of-the-art performance for a wide range of network densities and topologies.

Expected response: Agent 0 should acknowledge the information provided and may ask for further details or clarification if needed.

Agent 2 (worker): In my section of the paper, the authors conduct several experiments to evaluate the effectiveness of their proposed method, DIGRAC.

Firstly, they use synthetic data generated from Directed Stochastic Block Models (DSBM) with different parameters and structures. They vary the number of clusters, the ratio between the largest and smallest cluster size, the number of nodes, and the direction flip probability. They also consider different meta-graph structures, including "cycle", "path", "complete", and "star", with or without ambient nodes.

Secondly, they test DIGRAC on five real-world digraph datasets: Telegram, Blog, Migration, WikiTalk, and Lead-Lag. The number of clusters and values of β are set differently for each dataset.

In terms of performance measures, they use the Adjusted Rand Index (ARI) for node clustering, and accuracy for node classification and link direction prediction. They also introduce normalized imbalance scores to evaluate clustering performance based on flow imbalance.

Finally, they compare DIGRAC against 11 recent methods from the literature for clustering digraphs, including InfoMap, DGCN, and three variants of DiGCN.

The authors also discuss their training setup and hyperparameter selection process. They use 10% of all nodes from each cluster as test nodes, 10% as validation nodes to select the model, and the remaining 80% as training nodes. Hyperparameters are selected via a greedy search.

Agent 3 (worker): In my section of the paper, the authors conducted experiments on real-world data sets to evaluate the performance of the DIGRAC model. They used the largest K eigenvectors of the Hermitian matrix to construct an input feature matrix. They tested the model on a DSBM("cycle", T, n = 1000, K = 5, p = 0.02, ρ = 1) model and compared the test ARI performance using three variants of loss functions on the same digraph. They found that the "sort" choice performed best among these variants. They also experimented on adding seeds, with the seed ratio defined as the ratio of the number of seed nodes to the number of training nodes. They found that seed nodes with a supervised loss function enhance performance, and inferred that the model can further boost its performance when additional label information is available.

In the conclusion and future work section, they mentioned that further work will include additional experiments in the semi-supervised setting, when there exist seed nodes with known cluster labels, or when additional information is available in the form of must-link and cannot-link constraints. They also discussed extending the framework to detect the number of clusters, instead of specifying it a-priori. They also mentioned the need for additional preliminary analysis on how many pairwise imbalance scores to consider. They also discussed the need to build a more powerful framework that can automatically detect the value β used in the current model, to select the subset of influential pairs of imbalances.

Expected response: Agent 0 should confirm receipt of the message and may ask for additional details or clarifications.

Agent 4 (expert): I do not believe the request is relevant to me, as I do not have a paper chunk. I will stand by for further instructions.

Agent 0 (leader): Step 3b: Compare the actual experiments with the expected experiments and identify any discrepancies.

Based on the responses from the agents and the expectations from Agent 4, the paper seems to have conducted a thorough set of experiments. They have used both synthetic and real-world datasets, varied parameters, and compared their method with several recent methods. They have also considered different meta-graph structures and used multiple performance measures.

However, there are a few discrepancies:

1. The paper does not seem to have conducted ablation studies to evaluate the individual contributions of the novel components of their method, such as the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA).

2. The paper does not seem to have tested the method's performance when additional label information is available, despite mentioning that the method can naturally incorporate node features and does not require label supervision.
3. The paper does not seem to have conducted experiments to evaluate the potential improvements they discussed, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications.

Step 3c: Discuss any discrepancies with the agents to determine if they constitute shortcomings of the paper.

SEND MESSAGE: Based on the responses and the expectations, I have identified a few discrepancies:

1. The paper does not seem to have conducted ablation studies to evaluate the individual contributions of the novel components of their method, such as the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA).
2. The paper does not seem to have tested the method's performance when additional label information is available, despite mentioning that the method can naturally incorporate node features and does not require label supervision.
3. The paper does not seem to have conducted experiments to evaluate the potential improvements they discussed, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications.

Do you agree that these are shortcomings of the paper? If so, how should the authors address these issues?

Expected response: Each agent should provide their opinion on whether the identified discrepancies constitute shortcomings of the paper and suggest how the authors could address these issues.

Agent 1 (worker): I agree that these are potential shortcomings of the paper. To address these issues, the authors could conduct additional experiments to evaluate the individual contributions of the novel components of their method. This could involve running ablation studies where they remove or modify one component at a time and observe the impact on the results. They could also test the method's performance when additional label information is available to see if this improves the results. Finally, they could conduct experiments to evaluate the potential improvements they discussed, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications.

Expected response: Agent 0 should acknowledge the suggestions and may provide further instructions or ask for additional input.

Agent 2 (worker): I agree with the identified discrepancies. The authors could have conducted ablation studies to evaluate the individual contributions of the novel components of their method. This would help

to understand the impact of each component on the overall performance.

Regarding the use of additional label information, the authors could have conducted experiments to evaluate the performance of their method when such information is available. This would provide a more comprehensive evaluation of their method.

As for the potential improvements discussed in the paper, the authors could have conducted experiments to evaluate these improvements. This would provide evidence to support their claims and demonstrate the potential of their method.

To address these issues, the authors could conduct additional experiments to evaluate the individual contributions of the novel components of their method, the performance of their method when additional label information is available, and the potential improvements they discussed. They could also provide a more detailed discussion on these aspects in their paper.

Agent 3 (worker): I agree with the identified discrepancies. The authors could have conducted ablation studies to evaluate the individual contributions of the novel components of their method. This would have provided a clearer understanding of the impact of each component on the overall performance of the method.

Regarding the second point, the authors did mention that they experimented with adding seed nodes with a supervised loss function and found that it enhanced performance. However, they could have provided more details on these experiments and perhaps conducted additional experiments to evaluate the method's performance when more label information is available.

As for the potential improvements discussed, it would have been beneficial if the authors had conducted experiments to evaluate these. For instance, they could have conducted experiments in a semi-supervised setting, tested the automatic detection of the number of clusters, and evaluated the method's performance in large scale applications. These experiments would have provided more concrete evidence of the method's potential and areas for future work.

Expected response: Agent 0 should confirm receipt of the message and may ask for additional details or clarifications.

Agent 4 (expert): I agree with the identified discrepancies and consider them as shortcomings of the paper. Here's how the authors could address these issues:

1. **Ablation Studies**: The authors should conduct ablation studies to evaluate the individual contributions of the novel components of their method. This involves removing or modifying one component at a time (e.g., the probabilistic imbalance loss, the DIMPA scheme) and observing the impact on the performance. This will help to understand the importance of each component and how they contribute to

the overall performance.

2. **Incorporation of Label Information**: The authors should test the method's performance when additional label information is available. This can be done by conducting experiments where a small amount of labeled data is used in addition to the unlabeled data. The performance of the semi-supervised learning approach can be compared with the unsupervised learning approach.

3. **Potential Improvements**: The authors should conduct experiments to evaluate the potential improvements they discussed. This includes experiments for semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications. The performance of these improved approaches can be compared with the original approach.

By addressing these issues, the authors can provide a more comprehensive evaluation of their method and its potential improvements.

Agent 0 (leader): Step 3d: Write feedback comments for any identified shortcomings, explaining the issue and suggesting how the authors could resolve it.

Based on the discussion with the agents, here are the feedback comments:

1. **Ablation Studies**: The paper would benefit from ablation studies to evaluate the individual contributions of the novel components of the method, such as the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA). This could involve running experiments where one component is removed or modified at a time, and the impact on the results is observed. This would help to understand the importance of each component and how they contribute to the overall performance.

2. **Incorporation of Label Information**: The paper mentions that the method can naturally incorporate node features and does not require label supervision. However, it would be beneficial to test the method's performance when additional label information is available. This could involve conducting experiments where a small amount of labeled data is used in addition to the unlabeled data, and the performance of the semi-supervised learning approach is compared with the unsupervised learning approach.

3. **Potential Improvements**: The paper discusses potential improvements to the model, such as semi-supervised learning, automatic detection of the number of clusters, and adaptations for large scale applications. However, it would be beneficial to conduct experiments to evaluate these improvements. This would provide more concrete evidence of the method's potential and areas for future work.

These comments should be addressed to provide a more comprehensive evaluation of the method and its potential improvements.

B.2.2. MARG-S (clarity)

Agent 0 (leader): High-level plan:

1. Ask each agent (Agent 1, Agent 2, Agent 3) to summarize their part of the paper, focusing on the methods, experimental settings, and key concepts.
2. Compile the summaries into a coherent explanation of the paper.
3. Explain the paper to Agent 4 and answer any questions they have.
4. Pass any questions or comments from Agent 4 to the relevant agent(s) (Agent 1, Agent 2, Agent 3) for clarification.
5. Pass the answers back to Agent 4.
6. Write feedback based on any points Agent 4 indicates are in need of improvement.
7. Review and finalize the feedback comments.

Now, let's execute the plan.

Step 1: Ask each agent to summarize their part of the paper, focusing on the methods, experimental settings, and key concepts.

SEND MESSAGE: Please provide a summary of your part of the paper, focusing on the methods, experimental settings, and key concepts.

Expected response: Summaries from Agent 1, Agent 2, and Agent 3, each focusing on the methods, experimental settings, and key concepts in their part of the paper.

Agent 1 (worker): My part of the paper introduces a new Graph Neural Network (GNN) framework called DIGRAC, which is designed for clustering directed networks. Unlike existing methods that focus on edge density, DIGRAC uses directionality as the main signal for clustering. It introduces a novel probabilistic imbalance loss and optimizes directed flow imbalance for clustering without requiring label supervision. The paper also introduces a directed mixed path aggregation scheme, DIMPA, for the implementation of the framework. The method is tested on synthetic and real-world data, demonstrating superior performance in terms of the Adjusted Rand Index (ARI) and imbalance scores. The paper also discusses the applicability of DIGRAC in various fields such as time series analysis, ranking, and anomaly detection.

Expected response: Confirmation from Agent 0 that the summary is clear and accurate, or requests for additional information or clarification.

Agent 2 (worker): My part of the paper discusses a method for clustering in directed graphs, specifically focusing on the self-supervised loss for clustering and the directed mixed path aggregation (DIMPA) scheme. The method aims to find a partition that maximizes the imbalance flow to represent clusters in a

meta-graph. The paper also introduces a global probabilistic imbalance score and discusses how to select pairs of clusters for this score. The DIMPA scheme is used to obtain the probability assignment matrix and feed it to the loss function. This scheme involves building node embeddings by capturing local network information and taking a weighted average of information from neighbors within a certain number of hops. The paper also discusses experiments conducted to test the method, using both synthetic and real-world data sets. The performance of the method is measured using the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction. The paper also introduces normalized imbalance scores to evaluate clustering performance based on flow imbalance.

Expected response: Confirmation from Agent 0 that the summary is accurate and complete, or requests for additional information or clarification.

Agent 3 (worker): My part of the paper discusses the experimental results and future work of a model called DIGRAC. In the experimental results, the paper discusses how DIGRAC uses the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. The paper also discusses the use of different loss functions and the impact of adding seed nodes with a supervised loss function. In the future work section, the paper discusses the potential for DIGRAC to be used in a semi-supervised setting, the possibility of extending the framework to detect the number of clusters, and the need for preliminary analysis to determine how many pairwise imbalance scores to consider. The paper also mentions the potential for further research in the sparse regime and the adaptation of the pipeline for directed clustering in extremely large networks.

EXPECTED RESPONSE: Acknowledgement from Agent 0 and possibly further questions or tasks related to the content of my part of the paper.

Agent 0 (leader): Step 2: Compile the summaries into a coherent explanation of the paper.

The paper introduces a new Graph Neural Network (GNN) framework called DIGRAC, designed for clustering directed networks. Unlike existing methods that focus on edge density, DIGRAC uses directionality as the main signal for clustering. It introduces a novel probabilistic imbalance loss and optimizes directed flow imbalance for clustering without requiring label supervision. The paper also introduces a directed mixed path aggregation scheme, DIMPA, for the implementation of the framework. The method is tested on synthetic and real-world data, demonstrating superior performance in terms of the Adjusted Rand Index (ARI) and imbalance scores. The paper also discusses the applicability of DIGRAC in various fields such as time series analysis, ranking, and anomaly detection.

The method aims to find a partition that maximizes the imbalance flow to represent clusters in a meta-graph. The paper also introduces a global probabilistic imbalance score and discusses how to select pairs of clusters for this score. The DIMPA scheme is used to obtain the probability assignment matrix and feed it to the loss function. This scheme involves building node embeddings by capturing local network information and taking a weighted average of information from neighbors within a certain number of

hops. The performance of the method is measured using the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction. The paper also introduces normalized imbalance scores to evaluate clustering performance based on flow imbalance.

In the experimental results, the paper discusses how DIGRAC uses the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. The paper also discusses the use of different loss functions and the impact of adding seed nodes with a supervised loss function. In the future work section, the paper discusses the potential for DIGRAC to be used in a semi-supervised setting, the possibility of extending the framework to detect the number of clusters, and the need for preliminary analysis to determine how many pairwise imbalance scores to consider. The paper also mentions the potential for further research in the sparse regime and the adaptation of the pipeline for directed clustering in extremely large networks.

Step 3: Explain the paper to Agent 4 and answer any questions they have.

SEND MESSAGE: I have summarized the paper based on the information provided by the other agents. The paper introduces a new Graph Neural Network (GNN) framework called DIGRAC, designed for clustering directed networks. It introduces a novel probabilistic imbalance loss and optimizes directed flow imbalance for clustering without requiring label supervision. The paper also introduces a directed mixed path aggregation scheme, DIMPA, for the implementation of the framework. The method is tested on synthetic and real-world data, demonstrating superior performance in terms of the Adjusted Rand Index (ARI) and imbalance scores. The paper also discusses the applicability of DIGRAC in various fields such as time series analysis, ranking, and anomaly detection. The method aims to find a partition that maximizes the imbalance flow to represent clusters in a meta-graph. The paper also introduces a global probabilistic imbalance score and discusses how to select pairs of clusters for this score. The DIMPA scheme is used to obtain the probability assignment matrix and feed it to the loss function. This scheme involves building node embeddings by capturing local network information and taking a weighted average of information from neighbors within a certain number of hops. The performance of the method is measured using the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction. The paper also introduces normalized imbalance scores to evaluate clustering performance based on flow imbalance. In the experimental results, the paper discusses how DIGRAC uses the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. The paper also discusses the use of different loss functions and the impact of adding seed nodes with a supervised loss function. In the future work section, the paper discusses the potential for DIGRAC to be used in a semi-supervised setting, the possibility of extending the framework to detect the number of clusters, and the need for preliminary analysis to determine how many pairwise imbalance scores to consider. The paper also mentions the potential for further research in the sparse regime and the adaptation of the pipeline for directed clustering in extremely large networks. Do you have any questions or comments about the paper?

Expected response: Questions or comments from Agent 4 about the paper.

Agent 4 (expert): Thank you for the summary, Agent 0. I have several questions and comments about the paper:

1. Could you provide more details about the novel probabilistic imbalance loss introduced in the paper? How is it calculated and what makes it different from other loss functions used in similar studies?

2. The paper introduces a directed mixed path aggregation scheme, DIMPA. Could you elaborate on how this scheme works? What is the significance of taking a weighted average of information from neighbors within a certain number of hops?

3. How does the paper define and calculate the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction?

4. The paper discusses the use of the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. Could you provide more details on this process? How is the value of K determined?

5. The paper mentions the impact of adding seed nodes with a supervised loss function. Could you elaborate on this? How are these seed nodes selected and what is their role in the process?

6. In the future work section, the paper discusses the potential for DIGRAC to be used in a semi-supervised setting. Could you provide more details on how this might be implemented?

7. The paper also mentions the potential for further research in the sparse regime. Could you explain what is meant by the "sparse regime"?

8. Lastly, could you provide more information on the experimental settings? What kind of synthetic and real-world data was used? What were the hyperparameters, equipment, and material specifications used in the experiments?

These details are crucial for understanding the paper's methods, experimental settings, and key concepts, and for determining whether the paper is well-organized and can be easily understood and reproduced.

Agent 0 (leader): Step 4: Pass any questions or comments from Agent 4 to the relevant agent(s) for clarification.

SEND MESSAGE: Agent 4 has several questions about the paper. I would appreciate if you could provide answers to the following:

1. Could you provide more details about the novel probabilistic imbalance loss introduced in the paper? How is it calculated and what makes it different from other loss functions used in similar studies?

2. The paper introduces a directed mixed path aggregation scheme, DIMPA. Could you elaborate on how this scheme works? What is the significance of taking a weighted average of information from neighbors within a certain number of hops?

3. How does the paper define and calculate the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction?

4. The paper discusses the use of the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. Could you provide more details on this process? How is the value of K determined?

5. The paper mentions the impact of adding seed nodes with a supervised loss function. Could you elaborate on this? How are these seed nodes selected and what is their role in the process?

6. In the future work section, the paper discusses the potential for DIGRAC to be used in a semi-supervised setting. Could you provide more details on how this might be implemented?

7. The paper also mentions the potential for further research in the sparse regime. Could you explain what is meant by the "sparse regime"?

8. Lastly, could you provide more information on the experimental settings? What kind of synthetic and real-world data was used? What were the hyperparameters, equipment, and material specifications used in the experiments?

Expected response: Detailed answers to Agent 4's questions from Agent 1, Agent 2, and Agent 3.

Agent 1 (worker): In response to Agent 4's questions:

1. The novel probabilistic imbalance loss introduced in the paper is a self-supervised loss function inspired by Cucuringu et al. (2020a). It aims to cluster the nodes by maximizing a normalized form of cut imbalance across clusters. The imbalance flows are normalized to obtain an imbalance score with values in [0, 1] for interpretability and ease of comparison.

2. The directed mixed path aggregation scheme, DIMPA, is a method devised for the implementation of the DIGRAC framework. Unfortunately, my part of the paper does not provide detailed information on how this scheme works or the significance of taking a weighted average of information from neighbors within a certain number of hops.

3. The Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are mentioned as measures of performance in the paper. However, my part of the paper does not provide detailed information on how these are defined and calculated.

4. The paper discusses the use of the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. Unfortunately, my part of the paper does not provide detailed information on this process or how the value of K is determined.

5. The paper mentions the impact of adding seed nodes with a supervised loss function. However, my part of the paper does not provide detailed information on how these seed nodes are selected or their role in the process.

6. In the future work section, the paper discusses the potential for DIGRAC to be used in a semi-supervised setting. However, my part of the paper does not provide detailed information on how this might be implemented.

7. The term "sparse regime" is not explicitly defined in my part of the paper. However, in the context of graph theory, a sparse graph is one in which the number of edges is close to the minimal number of edges needed for the graph to be connected.

8. My part of the paper does not provide detailed information on the experimental settings, the kind of synthetic and real-world data used, or the hyperparameters, equipment, and material specifications used in the experiments.

I hope this helps. If you need more detailed information on any of these points, I would recommend reaching out to the other agents who may have access to different parts of the paper.

Expected response: Confirmation from Agent 0 that the message has been received and understood, or follow-up questions for further clarification.

Agent 2 (worker): In response to Agent 4's questions:

1. The novel probabilistic imbalance loss introduced in the paper is designed to optimize the directed flow imbalance for clustering. It is calculated based on the CI vol_sum from a specific equation mentioned in the paper. This loss function is different from others as it specifically targets the imbalance in the flow of information in the network, which is a unique aspect of directed networks.

2. The Directed Mixed Path Aggregation (DIMPA) scheme is a method to obtain the probability assignment matrix and feed it to the loss function. It involves building node embeddings by capturing local network information and taking a weighted average of information from neighbors within a certain number of hops. This is significant as it allows the model to consider not just the immediate neighbors of

a node, but also nodes that are further away, thus capturing more global information about the network.

3. The Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are standard metrics used in clustering and classification tasks. The paper does not provide specific details on how these are calculated.

4. When node attributes are not available, the paper suggests using the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix. The value of K is not specified in my section of the paper.

5. The paper does mention the impact of adding seed nodes with a supervised loss function, but does not provide specific details on how these seed nodes are selected or their role in the process.

6. The paper suggests that DIGRAC could potentially be used in a semi-supervised setting in future work, but does not provide specific details on how this might be implemented.

7. The term "sparse regime" typically refers to situations where the data or network is sparse, i.e., there are fewer connections or interactions between nodes. The paper suggests that further research could be done in this area, but does not provide specific details.

8. The paper uses both synthetic and real-world data for experiments. The synthetic data is generated using Directed Stochastic Block Models (DSBM), with various parameters such as the number of clusters, the ratio between the largest and smallest cluster size, and the number of nodes. The real-world data includes five datasets: Telegram, Blog, Migration, WikiTalk, and Lead-Lag. The paper does not provide specific details on the hyperparameters, equipment, or material specifications used in the experiments.

I hope this helps answer Agent 4's questions. If there are any further queries, feel free to ask.

Expected response: Confirmation from Agent 0 that the answers are satisfactory or further questions from Agent 4.

Agent 3 (worker): I can provide answers to questions 4 and 5.

4. The paper uses the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. This is a common technique in spectral graph theory, where the eigenvectors of a matrix associated with a graph (in this case, a Hermitian matrix) are used to extract features that capture the structure of the graph. The value of K is not specified in my section of the paper, and it may be determined based on the specific requirements of the application or through a process of experimentation.

5. The paper discusses the impact of adding seed nodes with a supervised loss function. These seed nodes are nodes for which the cluster labels are known a priori, and they are used to guide the clustering process. A supervised loss function is applied to these seed nodes, which helps to improve the performance of the clustering algorithm. The specific method for selecting these seed nodes is not detailed in my section of the paper.

I hope this helps. Please let me know if you need further clarification on these points.

Expected response: Confirmation from Agent 0 that my answers have been received and are satisfactory, or follow-up questions for further clarification.

Agent 0 (leader): Step 5: Pass the answers back to Agent 4.

SEND MESSAGE: Here are the answers to your questions:

1. The novel probabilistic imbalance loss introduced in the paper is a self-supervised loss function inspired by Cucuringu et al. (2020a). It aims to cluster the nodes by maximizing a normalized form of cut imbalance across clusters. The imbalance flows are normalized to obtain an imbalance score with values in [0, 1] for interpretability and ease of comparison. This loss function is designed to optimize the directed flow imbalance for clustering. It is calculated based on the CI vol_sum from a specific equation mentioned in the paper. This loss function is different from others as it specifically targets the imbalance in the flow of information in the network, which is a unique aspect of directed networks.

2. The Directed Mixed Path Aggregation (DIMPA) scheme is a method to obtain the probability assignment matrix and feed it to the loss function. It involves building node embeddings by capturing local network information and taking a weighted average of information from neighbors within a certain number of hops. This is significant as it allows the model to consider not just the immediate neighbors of a node, but also nodes that are further away, thus capturing more global information about the network.

3. The Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are standard metrics used in clustering and classification tasks. The paper does not provide specific details on how these are calculated.

4. When node attributes are not available, the paper suggests using the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix. This is a common technique in spectral graph theory, where the eigenvectors of a matrix associated with a graph (in this case, a Hermitian matrix) are used to extract features that capture the structure of the graph. The value of K is not specified in the paper, and it may be determined based on the specific requirements of the application or through a process of experimentation.

5. The paper discusses the impact of adding seed nodes with a supervised loss function. These seed nodes

are nodes for which the cluster labels are known a priori, and they are used to guide the clustering process. A supervised loss function is applied to these seed nodes, which helps to improve the performance of the clustering algorithm. The specific method for selecting these seed nodes is not detailed in the paper.

6. The paper suggests that DIGRAC could potentially be used in a semi-supervised setting in future work, but does not provide specific details on how this might be implemented.

7. The term "sparse regime" typically refers to situations where the data or network is sparse, i.e., there are fewer connections or interactions between nodes. The paper suggests that further research could be done in this area, but does not provide specific details.

8. The paper uses both synthetic and real-world data for experiments. The synthetic data is generated using Directed Stochastic Block Models (DSBM), with various parameters such as the number of clusters, the ratio between the largest and smallest cluster size, and the number of nodes. The real-world data includes five datasets: Telegram, Blog, Migration, WikiTalk, and Lead-Lag. The paper does not provide specific details on the hyperparameters, equipment, or material specifications used in the experiments.

Do you have any further questions or comments about the paper?

Expected response: Further questions or comments from Agent 4 about the paper, or confirmation that they are finished.

Agent 4 (expert): Thank you for the detailed responses, Agent 0. I have a few follow-up questions and comments:

1. You mentioned that the novel probabilistic imbalance loss is calculated based on the CI vol_sum from a specific equation mentioned in the paper. Could you provide more details about this equation? How is the CI vol_sum calculated?

2. Regarding the Directed Mixed Path Aggregation (DIMPA) scheme, you mentioned that it involves taking a weighted average of information from neighbors within a certain number of hops. How is this weighted average calculated? What factors determine the weight assigned to each neighbor?

3. You mentioned that the paper does not provide specific details on how the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are calculated. This could be a potential issue as these metrics are crucial for evaluating the performance of the method. Could the authors provide more details on how these metrics are calculated?

4. Regarding the use of the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix, you mentioned that the value of K is not specified in the paper. This could be a potential issue

as the choice of K could significantly impact the performance of the method. Could the authors provide more guidance on how to choose the value of K?

5. You mentioned that the paper does not provide specific details on how seed nodes are selected for the supervised loss function. This could be a potential issue as the selection of seed nodes could significantly impact the performance of the method. Could the authors provide more guidance on how to select these seed nodes?

6. You mentioned that the paper does not provide specific details on how DIGRAC could potentially be used in a semi-supervised setting. This could be a potential issue as it leaves the reader with unanswered questions about the future direction of the research. Could the authors provide more details on this?

7. You mentioned that the paper does not provide specific details on the hyperparameters, equipment, or material specifications used in the experiments. This could be a potential issue as it makes it difficult for other researchers to reproduce the experiments. Could the authors provide more details on these aspects?

These details are crucial for understanding the paper's methods, experimental settings, and key concepts, and for determining whether the paper is well-organized and can be easily understood and reproduced.

Agent 0 (leader): Step 4 (Continued): Pass any follow-up questions or comments from Agent 4 to the relevant agent(s) for clarification.

SEND MESSAGE: Agent 4 has several follow-up questions and comments about the paper. I would appreciate if you could provide further clarification on the following:

1. Could you provide more details about the equation used to calculate the CI vol_sum for the novel probabilistic imbalance loss? How is the CI vol_sum calculated?

2. Regarding the Directed Mixed Path Aggregation (DIMPA) scheme, how is the weighted average of information from neighbors within a certain number of hops calculated? What factors determine the weight assigned to each neighbor?

3. Agent 4 noted that the paper does not provide specific details on how the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are calculated. Could you provide more details on how these metrics are calculated?

4. Regarding the use of the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix, Agent 4 noted that the value of K is not specified in the paper. Could you provide more guidance on how to choose the value of K?

5. Agent 4 noted that the paper does not provide specific details on how seed nodes are selected

for the supervised loss function. Could you provide more guidance on how to select these seed nodes?

6. Agent 4 noted that the paper does not provide specific details on how DIGRAC could potentially be used in a semi-supervised setting. Could you provide more details on this?

7. Agent 4 noted that the paper does not provide specific details on the hyperparameters, equipment, or material specifications used in the experiments. Could you provide more details on these aspects?

Expected response: Detailed answers to Agent 4's follow-up questions from Agent 1, Agent 2, and Agent 3.

Agent 2 (worker): Regarding the Directed Mixed Path Aggregation (DIMPA) scheme, the weighted average of information from neighbors within a certain number of hops is calculated by row-normalizing the adjacency matrix to obtain A s. This is then used to capture local network information. The weight assigned to each neighbor is determined by the adjacency matrix and the number of hops. The paper does not provide specific details on the calculation of these weights.

As for the use of the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available, the value of K is not specified in the paper. It may be determined based on the specific requirements of the application or through a process of experimentation.

Expected response: Confirmation of the information provided or further questions on the DIMPA scheme and the use of eigenvectors.

Agent 3 (worker): In response to Agent 4's questions:

1. The CI vol_sum is not explicitly defined in my section of the paper. It seems to be related to the loss function used in the DIGRAC model, but without further context or information, I cannot provide a detailed explanation of how it is calculated.

2. The Directed Mixed Path Aggregation (DIMPA) scheme is not detailed in my section. I would need more information to provide a detailed explanation of how the weighted average of information from neighbors is calculated.

3. The Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are not explained in my section. These are standard metrics in machine learning and their calculation would be based on standard definitions.

4. The paper suggests using the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix when node attributes are not available. The value of K is not specified in my section of the paper. It may be determined based on the specific requirements of the application or through a process of

experimentation.

5. The paper discusses the impact of adding seed nodes with a supervised loss function. These seed nodes are nodes for which the cluster labels are known a priori, and they are used to guide the clustering process. The specific method for selecting these seed nodes is not detailed in my section of the paper.

6. The paper suggests that DIGRAC could potentially be used in a semi-supervised setting in future work, but does not provide specific details on how this might be implemented in my section.

7. The paper uses both synthetic and real-world data for experiments. The specific details on the hyperparameters, equipment, or material specifications used in the experiments are not provided in my section of the paper.

I expect the response to contain further questions or clarifications from Agent 0 or Agent 4, or instructions for the next steps.

Agent 0 (leader): Step 5 (Continued): Pass the answers back to Agent 4.

SEND MESSAGE: Here are the further clarifications to your follow-up questions:

1. The CI vol_sum is related to the loss function used in the DIGRAC model, but the agents do not have further context or information to provide a detailed explanation of how it is calculated.

2. The weighted average of information from neighbors within a certain number of hops in the Directed Mixed Path Aggregation (DIMPA) scheme is calculated by row-normalizing the adjacency matrix to obtain A s. This is then used to capture local network information. The weight assigned to each neighbor is determined by the adjacency matrix and the number of hops. The paper does not provide specific details on the calculation of these weights.

3. The Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are standard metrics in machine learning and their calculation would be based on standard definitions. The paper does not provide specific details on how these are calculated.

4. When node attributes are not available, the paper suggests using the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix. The value of K is not specified in the paper. It may be determined based on the specific requirements of the application or through a process of experimentation.

5. The paper discusses the impact of adding seed nodes with a supervised loss function. These seed nodes are nodes for which the cluster labels are known a priori, and they are used to guide the clustering process. The specific method for selecting these seed nodes is not detailed in the paper.

6. The paper suggests that DIGRAC could potentially be used in a semi-supervised setting in future work, but does not provide specific details on how this might be implemented.

7. The paper uses both synthetic and real-world data for experiments. The specific details on the hyperparameters, equipment, or material specifications used in the experiments are not provided in the paper.

Do you have any further questions or comments about the paper?

Expected response: Further questions or comments from Agent 4 about the paper, or confirmation that they are finished.

Agent 4 (expert): Thank you for your responses, Agent 0. I have no further questions at this time. However, I would like to summarize the potential issues and suggestions for the paper:

1. The paper should provide more details about the novel probabilistic imbalance loss, specifically how the CI vol_sum is calculated in the equation mentioned.

2. The paper should provide more details about the Directed Mixed Path Aggregation (DIMPA) scheme, specifically how the weighted average of information from neighbors is calculated and what factors determine the weight assigned to each neighbor.

3. The paper should provide more details on how the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are calculated.

4. The paper should provide more guidance on how to choose the value of K when using the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix.

5. The paper should provide more guidance on how to select seed nodes for the supervised loss function.

6. The paper should provide more details on how DIGRAC could potentially be used in a semi-supervised setting.

7. The paper should provide more details on the hyperparameters, equipment, and material specifications used in the experiments to make it easier for other researchers to reproduce the experiments.

These suggestions are aimed at improving the clarity and reproducibility of the paper. I am now done with my review.

Agent 0 (leader): Step 6: Write feedback based on any points Agent 4 indicates are in need of improvement.

Based on Agent 4's comments, here is the list of feedback comments for the paper:

1. The paper should provide more details about the novel probabilistic imbalance loss, specifically how the CI vol_sum is calculated in the equation mentioned.

2. The paper should provide more details about the Directed Mixed Path Aggregation (DIMPA) scheme, specifically how the weighted average of information from neighbors is calculated and what factors determine the weight assigned to each neighbor.

3. The paper should provide more details on how the Adjusted Rand Index (ARI) and accuracy for node classification and link direction prediction are calculated.

4. The paper should provide more guidance on how to choose the value of K when using the largest K eigenvectors of a Hermitian matrix to construct an input feature matrix.

5. The paper should provide more guidance on how to select seed nodes for the supervised loss function.

6. The paper should provide more details on how DIGRAC could potentially be used in a semi-supervised setting.

7. The paper should provide more details on the hyperparameters, equipment, and material specifications used in the experiments to make it easier for other researchers to reproduce the experiments.

Step 7: Review and finalize the feedback comments.

The feedback comments provided above are based on the comments from Agent 4 and are aimed at improving the clarity and reproducibility of the paper. These comments are now finalized.

B.2.3. MARG-S (impact)

Agent 0 (leader): High-level plan:

1. Ask each agent (1, 2, 3) to summarize their part of the paper, focusing on the motivations, goals, and key findings.
2. Compile the summaries into a coherent explanation of the paper.
3. Share the explanation with Agent 4 and ask for their questions and potential shortcomings.

4. Relay Agent 4's questions to the relevant agents (1, 2, 3) and get their responses.
5. Share the responses with Agent 4 and continue the discussion until Agent 4 is finished.
6. Based on the discussion with Agent 4 and the information from the other agents, write a list of feedback comments.

Now, let's execute the plan.

Step 1: Ask each agent (1, 2, 3) to summarize their part of the paper, focusing on the motivations, goals, and key findings.

SEND MESSAGE: Please provide a summary of your part of the paper, focusing on the motivations, goals, and key findings.

Expected response: Summaries from each agent that include the motivations, goals, and key findings of their part of the paper.

Agent 1 (worker): The part of the paper I have focuses on introducing a new method for node clustering in directed networks, called DIGRAC. The motivation behind this method is the observation that directionality in networks can reveal latent properties of network flows and can be a vital signal for directed clustering, even in the absence of any edge density differences. The goal of DIGRAC is to recover clusters with strong and imbalanced flow among them. The key findings presented in this part of the paper are that DIGRAC, which optimizes directed flow imbalance for clustering without requiring label supervision, attains state-of-the-art results on directed graph clustering for a wide range of noise and sparsity levels and graph structures and topologies. The paper also discusses how DIGRAC complements existing approaches and extends its applicability beyond settings where the input data is a digraph.

Expected response: Confirmation from Agent 0 that the summary is accurate and complete, or requests for additional information or clarification.

Agent 2 (worker): My part of the paper primarily discusses the methodology and experimental results of a novel method for clustering in directed graphs, called DIGRAC. The method uses a self-supervised loss function for clustering and a directed mixed path aggregation scheme to build node embeddings. The paper also introduces a new measure, the imbalance score, to evaluate clustering performance based on flow imbalance. The experiments section compares DIGRAC against other methods from the literature and shows that DIGRAC performs well across a range of network densities and noise levels, and is particularly effective when there are ambient nodes present.

Expected response: Confirmation that the summary is clear and accurate, or requests for additional information or clarification.

Agent 3 (worker): My part of the paper primarily covers the experimental results and conclusions of a study on DIGRAC, a model for node clustering. The experimental results section discusses the performance of DIGRAC on real-world data sets, with a focus on the impact of different loss functions and the addition of seed nodes. The conclusion and future work sections outline the current capabilities of DIGRAC, such as creating node embeddings and performing directed clustering, and suggest future improvements, including semi-supervised settings, automatic detection of the number of clusters, and adaptations for large scale industrial applications.

Expected response: Confirmation from Agent 0 that my summary is clear and accurate, or requests for additional details or clarifications.

Agent 0 (leader): Step 2: Compile the summaries into a coherent explanation of the paper.

The paper introduces a new method for node clustering in directed networks, called DIGRAC. The motivation behind this method is the observation that directionality in networks can reveal latent properties of network flows and can be a vital signal for directed clustering, even in the absence of any edge density differences. The goal of DIGRAC is to recover clusters with strong and imbalanced flow among them. The method uses a self-supervised loss function for clustering and a directed mixed path aggregation scheme to build node embeddings. The paper also introduces a new measure, the imbalance score, to evaluate clustering performance based on flow imbalance. The key findings presented in the paper are that DIGRAC, which optimizes directed flow imbalance for clustering without requiring label supervision, attains state-of-the-art results on directed graph clustering for a wide range of noise and sparsity levels and graph structures and topologies. The paper also discusses how DIGRAC complements existing approaches and extends its applicability beyond settings where the input data is a digraph. The experimental results section compares DIGRAC against other methods from the literature and shows that DIGRAC performs well across a range of network densities and noise levels, and is particularly effective when there are ambient nodes present. The conclusion and future work sections outline the current capabilities of DIGRAC, such as creating node embeddings and performing directed clustering, and suggest future improvements, including semi-supervised settings, automatic detection of the number of clusters, and adaptations for large scale industrial applications.

Step 3: Share the explanation with Agent 4 and ask for their questions and potential shortcomings.

SEND MESSAGE: Here is a summary of the paper: [insert summary here]. Do you have any questions or potential shortcomings to point out?

Expected response: Questions and potential shortcomings from Agent 4.

Agent 4 (expert): I'm sorry, but there seems to be a mistake. The summary of the paper was not included in your message. Could you please provide the summary so I can start my review?

Agent 0 (leader): Apologies for the oversight. Let's correct that.

SEND MESSAGE: Here is a summary of the paper: The paper introduces a new method for node clustering in directed networks, called DIGRAC. The motivation behind this method is the observation that directionality in networks can reveal latent properties of network flows and can be a vital signal for directed clustering, even in the absence of any edge density differences. The goal of DIGRAC is to recover clusters with strong and imbalanced flow among them. The method uses a self-supervised loss function for clustering and a directed mixed path aggregation scheme to build node embeddings. The paper also introduces a new measure, the imbalance score, to evaluate clustering performance based on flow imbalance. The key findings presented in the paper are that DIGRAC, which optimizes directed flow imbalance for clustering without requiring label supervision, attains state-of-the-art results on directed graph clustering for a wide range of noise and sparsity levels and graph structures and topologies. The paper also discusses how DIGRAC complements existing approaches and extends its applicability beyond settings where the input data is a digraph. The experimental results section compares DIGRAC against other methods from the literature and shows that DIGRAC performs well across a range of network densities and noise levels, and is particularly effective when there are ambient nodes present. The conclusion and future work sections outline the current capabilities of DIGRAC, such as creating node embeddings and performing directed clustering, and suggest future improvements, including semi-supervised settings, automatic detection of the number of clusters, and adaptations for large scale industrial applications. Do you have any questions or potential shortcomings to point out?

Expected response: Questions and potential shortcomings from Agent 4.

Agent 3 (worker): I have a question regarding the "imbalance score" mentioned in the summary. My part of the paper does not provide a definition or explanation for this term. Could you please provide more information about what the imbalance score is and how it is used in the DIGRAC method?

I expect the response to provide a detailed explanation of the imbalance score, including its definition and its role in the DIGRAC method.

Agent 4 (expert): Thank you for the summary. I have several questions and potential shortcomings to point out:

1. The paper introduces a new measure, the imbalance score, to evaluate clustering performance based on flow imbalance. How is this imbalance score calculated? Is it a standard measure in the field or a novel contribution of this paper? If it's a novel contribution, how do the authors justify its use and effectiveness?

2. The paper claims that DIGRAC attains state-of-the-art results on directed graph clustering for a wide range of noise and sparsity levels and graph structures and topologies. What specific benchmarks or datasets were used to validate this claim? How does DIGRAC compare to existing methods in terms of computational efficiency?

3. The paper mentions that DIGRAC is particularly effective when there are ambient nodes present. Could you clarify what is meant by "ambient nodes"?

4. The future work section suggests adaptations for large scale industrial applications. Does the paper provide any preliminary results or discussions on how DIGRAC might perform in such settings?

5. The paper discusses how DIGRAC complements existing approaches and extends its applicability beyond settings where the input data is a digraph. Could you provide more details on how this is achieved?

6. The paper introduces a self-supervised loss function for clustering and a directed mixed path aggregation scheme to build node embeddings. Are there any assumptions or limitations associated with these methods that the authors should address?

7. Lastly, the paper suggests future improvements, including semi-supervised settings, automatic detection of the number of clusters. Are there any existing methods that already incorporate these features? If so, how does DIGRAC plan to improve upon them?

Agent 0 (leader): Step 4: Relay Agent 4's questions to the relevant agents (1, 2, 3) and get their responses.

SEND MESSAGE: Agent 4 has several questions about the paper. I'll list them below. Please provide answers to any questions that pertain to your section of the paper.

1. How is the imbalance score calculated? Is it a standard measure in the field or a novel contribution of this paper? If it's a novel contribution, how do the authors justify its use and effectiveness?
2. What specific benchmarks or datasets were used to validate the claim that DIGRAC attains state-of-the-art results on directed graph clustering for a wide range of noise and sparsity levels and graph structures and topologies? How does DIGRAC compare to existing methods in terms of computational efficiency?
3. Could you clarify what is meant by "ambient nodes"?
4. Does the paper provide any preliminary results or discussions on how DIGRAC might perform in large scale industrial applications?
5. Could you provide more details on how DIGRAC complements existing approaches and extends its applicability beyond settings where the input data is a digraph?
6. Are there any assumptions or limitations associated with the self-supervised loss function for clustering and the directed mixed path aggregation scheme to build node embeddings?
7. Are there any existing methods that already incorporate semi-supervised settings and automatic detection of the number of clusters? If so, how does DIGRAC plan to improve upon them?

Expected response: Answers to Agent 4's questions from the relevant agents.

Agent 1 (worker): In response to Agent 4's questions:

1. The imbalance score is a novel contribution of this paper. It is calculated as a normalized form of cut imbalance across clusters. The authors justify its use and effectiveness through experimental results on synthetic data and real-world data at different scales.

2. The paper does not provide specific benchmarks or datasets in my section. However, it mentions that DIGRAC attains state-of-the-art results on directed graph clustering for a wide range of noise and sparsity levels and graph structures and topologies.

3. The term "ambient nodes" is not mentioned in my section of the paper.

4. The paper does not provide any preliminary results or discussions on how DIGRAC might perform in large scale industrial applications in my section.

5. DIGRAC complements existing approaches in various aspects. It can detect alternative patterns in the data, such as meta-graph structures, which are otherwise not detectable by existing methods. It can include exogenous information, in the form of node-level features or labels, thus borrowing their strength. It also introduces an imbalance-based objective.

6. The self-supervised loss function for clustering is inspired by Cucuringu et al. (2020a), aiming to cluster the nodes by maximizing a normalized form of cut imbalance across clusters. The directed mixed path aggregation scheme, abbreviated as DIMPA, is devised for the implementation of the framework. It is described as simple and effective.

7. The paper does not provide information on existing methods that already incorporate semi-supervised settings and automatic detection of the number of clusters in my section.

I expect Agent 0 to confirm the accuracy of my responses or provide additional information if necessary.

Agent 2 (worker): In response to question 1, the imbalance score is calculated based on the concept of "imbalance flow" which is introduced in this paper. It is a measure of how much the flow of information in a network deviates from a balanced state. The imbalance score is a novel contribution of this paper and its effectiveness is justified through the experimental results presented in the paper.

In response to question 2, the paper uses both synthetic and real-world datasets to validate the performance of DIGRAC. The synthetic datasets are generated using Directed Stochastic Block Models with varying parameters to simulate different network densities and noise levels. The real-world datasets

include Telegram, Blog, Migration, WikiTalk, and Lead-Lag. The paper does not provide a direct comparison of computational efficiency between DIGRAC and existing methods.

In response to question 3, "ambient nodes" refer to nodes that are not part of a meaningful cluster in the network. They are not part of the main directional signal planted in the graph and are considered as background noise.

In response to question 6, the self-supervised loss function for clustering and the directed mixed path aggregation scheme to build node embeddings are novel contributions of this paper. The loss function is designed to maximize the imbalance flow in the network, which is a unique approach compared to traditional clustering methods that aim to maximize intra-cluster similarity and inter-cluster dissimilarity. The directed mixed path aggregation scheme is used to capture local network information by taking a weighted average of information from neighbors within a certain number of hops. The paper does not discuss any specific assumptions or limitations associated with these methods.

EXPECTED RESPONSE: Confirmation from Agent 0 that the answers provided are satisfactory or further questions from Agent 4.

Agent 3 (worker): I can provide some information regarding questions 2 and 4.

For question 2, my section of the paper discusses the experimental results of DIGRAC on real-world data sets. The paper mentions that the node in-and outdegrees may not be identical across clusters in these data sets. The authors used the largest K eigenvectors of the Hermitian matrix to construct an input feature matrix. They experimented with different loss functions and found that the "sort" variant performed best. However, the specific benchmarks or datasets used for these experiments are not mentioned in my section.

For question 4, the conclusion of my section mentions that adapting DIGRAC for directed clustering in extremely large networks, possibly combined with sampling methods or mini-batch, is a future direction worth exploring. This suggests that the authors are considering the application of DIGRAC in large scale industrial applications, but no preliminary results or discussions are provided in my section.

I expect the response to acknowledge the information I provided and possibly provide additional details if available from other sections of the paper.

Agent 0 (leader): Step 5: Share the responses with Agent 4 and continue the discussion until Agent 4 is finished.

SEND MESSAGE: Here are the responses to your questions:

1. The imbalance score is a novel contribution of this paper. It is calculated as a normalized form of cut imbalance across clusters. The authors justify its use and effectiveness through experimental results

on synthetic data and real-world data at different scales.
2. The paper uses both synthetic and real-world datasets to validate the performance of DIGRAC. The synthetic datasets are generated using Directed Stochastic Block Models with varying parameters to simulate different network densities and noise levels. The real-world datasets include Telegram, Blog, Migration, WikiTalk, and Lead-Lag. The paper does not provide a direct comparison of computational efficiency between DIGRAC and existing methods.
3. "Ambient nodes" refer to nodes that are not part of a meaningful cluster in the network. They are not part of the main directional signal planted in the graph and are considered as background noise.
4. The conclusion of the paper mentions that adapting DIGRAC for directed clustering in extremely large networks, possibly combined with sampling methods or mini-batch, is a future direction worth exploring. This suggests that the authors are considering the application of DIGRAC in large scale industrial applications, but no preliminary results or discussions are provided.
5. DIGRAC complements existing approaches in various aspects. It can detect alternative patterns in the data, such as meta-graph structures, which are otherwise not detectable by existing methods. It can include exogenous information, in the form of node-level features or labels, thus borrowing their strength. It also introduces an imbalance-based objective.
6. The self-supervised loss function for clustering and the directed mixed path aggregation scheme to build node embeddings are novel contributions of this paper. The loss function is designed to maximize the imbalance flow in the network, which is a unique approach compared to traditional clustering methods that aim to maximize intra-cluster similarity and inter-cluster dissimilarity. The directed mixed path aggregation scheme is used to capture local network information by taking a weighted average of information from neighbors within a certain number of hops. The paper does not discuss any specific assumptions or limitations associated with these methods.
7. The paper does not provide information on existing methods that already incorporate semi-supervised settings and automatic detection of the number of clusters.

Do you have any further questions or comments?

Expected response: Further questions or comments from Agent 4, or a confirmation that they are finished.

Agent 4 (expert): Thank you for the detailed responses. I have a few follow-up questions and comments:

1. Regarding the imbalance score, it's good to know that it's a novel contribution of this paper. However, since it's a new measure, it would be beneficial if the authors could compare it with other existing measures in the field. This would help to validate its effectiveness and provide a more comprehensive evaluation of DIGRAC's performance.

2. It's concerning that the paper does not provide a direct comparison of computational efficiency between DIGRAC and existing methods. This is an important aspect to consider, especially when discussing the potential for large scale industrial applications. The authors should provide this information to

give a complete picture of DIGRAC's performance.

3. The explanation of "ambient nodes" is clear. However, it would be interesting to know how DIGRAC handles these nodes. Are they simply ignored or does the algorithm have a specific way of dealing with them?

4. The authors should provide more details on how DIGRAC can detect alternative patterns in the data, such as meta-graph structures, and how it can include exogenous information. These are important features that could set DIGRAC apart from existing methods, so they should be explained more thoroughly.

5. The self-supervised loss function and the directed mixed path aggregation scheme sound interesting. However, the authors should discuss any potential limitations or assumptions associated with these methods. This would help to identify any potential shortcomings and areas for future improvement.

6. The lack of information on existing methods that incorporate semi-supervised settings and automatic detection of the number of clusters is a missed opportunity. The authors should provide a comparison with these methods to show how DIGRAC could potentially improve upon them.

In summary, while the paper introduces some novel concepts and methods, there are several areas where more information or justification is needed. The authors should provide more details on the imbalance score, computational efficiency, handling of ambient nodes, detection of alternative patterns, inclusion of exogenous information, and potential limitations of the proposed methods. They should also compare DIGRAC with existing methods that incorporate semi-supervised settings and automatic detection of the number of clusters.

Agent 0 (leader): Step 6: Based on the discussion with Agent 4 and the information from the other agents, write a list of feedback comments.

Feedback comments:

1. The imbalance score is a novel contribution of this paper. However, since it's a new measure, it would be beneficial if the authors could compare it with other existing measures in the field. This would help to validate its effectiveness and provide a more comprehensive evaluation of DIGRAC's performance.

2. The paper does not provide a direct comparison of computational efficiency between DIGRAC and existing methods. This is an important aspect to consider, especially when discussing the potential for large scale industrial applications. The authors should provide this information to give a complete picture of DIGRAC's performance.

3. It would be interesting to know how DIGRAC handles "ambient nodes". Are they simply ignored or does the algorithm have a specific way of dealing with them?

4. The authors should provide more details on how DIGRAC can detect alternative patterns in the data, such as meta-graph structures, and how it can include exogenous information. These are important features that could set DIGRAC apart from existing methods, so they should be explained more thoroughly.

5. The authors should discuss any potential limitations or assumptions associated with the self-supervised loss function and the directed mixed path aggregation scheme. This would help to identify any potential shortcomings and areas for future improvement.

6. The lack of information on existing methods that incorporate semi-supervised settings and automatic detection of the number of clusters is a missed opportunity. The authors should provide a comparison with these methods to show how DIGRAC could potentially improve upon them.

In summary, while the paper introduces some novel concepts and methods, there are several areas where more information or justification is needed. The authors should provide more details on the imbalance score, computational efficiency, handling of ambient nodes, detection of alternative patterns, inclusion of exogenous information, and potential limitations of the proposed methods. They should also compare DIGRAC with existing methods that incorporate semi-supervised settings and automatic detection of the number of clusters.

B.2.4. Refinement Stage

The comment being refined is as follows:

> *"Ablation Studies: The paper would benefit from ablation studies to evaluate the individual contributions of the novel components of the method, such as the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA). This could involve running experiments where one component is removed or modified at a time, and the impact on the results is observed. This would help to understand the importance of each component and how they contribute to the overall performance."*

Agent 0 (leader): High-level plan:

1. Understand the comment: The comment is about the lack of ablation studies in the paper. Ablation studies are used to understand the contribution of each component of a method by removing or modifying them one at a time. The comment suggests that the paper would benefit from such studies, particularly for the novel components like the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA).

2. Share the comment with other agents: I will share the comment with the other agents and ask them to check their parts of the paper to see if there are any ablation studies mentioned.

3. Gather information: I will ask the agents to provide information about the novel components mentioned in the comment (probabilistic imbalance loss and DIMPA), their role in the method, and any results or discussions related to them in the paper.

4. Evaluate the comment: Based on the information gathered, I will evaluate whether the comment is valid and important. If the paper already includes ablation studies or if the novel components are sufficiently explained and justified, the comment may not be necessary.

5. Refine the comment: If the comment is valid, I will refine it to be more specific and helpful. This could involve suggesting specific experiments for the ablation studies or providing more context about why these studies are important.

6. Finalize the comment: I will finalize the comment and ensure it meets the criteria of being detailed, helpful, and important.

Now, let's execute the plan.

Step 1: Understanding the comment
The comment is suggesting that the paper lacks ablation studies, particularly for the novel components like the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA). These studies could help understand the importance of each component and how they contribute to the overall performance.

Step 2: Sharing the comment with other agents
SEND MESSAGE: We have received a review comment suggesting that the paper lacks ablation studies for the novel components of the method, such as the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA). Please check your parts of the paper to see if there are any ablation studies mentioned. Also, provide information about these novel components, their role in the method, and any results or discussions related to them.

Expected response: The agents should provide information about whether there are any ablation studies in their parts of the paper and details about the novel components mentioned in the comment.

Agent 1 (worker): In my part of the paper, the authors introduce a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA) as part of their proposed method, DIGRAC. The probabilistic imbalance loss is used in a self-supervised manner to act on the digraph induced by all training nodes. The global imbalance score, one minus whom is the self-supervised loss function, is

aggregated from pairwise normalized cut imbalances. The method is end-to-end in combining embedding generation and clustering without an intermediate step. The directed mixed path aggregation scheme (DIMPA) is part of the implementation of the framework. However, there is no explicit mention of ablation studies for these components in my part of the paper.

Expected response: Confirmation from Agent 0 about the information provided and possibly further instructions.

Agent 2 (worker): In my part of the paper, the authors discuss the novel components of the method, namely the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA).

The probabilistic imbalance loss is used to find a partition that maximizes the imbalance flow, capturing groups of nodes that represent clusters in the meta-graph. The normalization by volumes is carried out to penalize partitions that put most nodes into a single cluster. The authors also discuss different variants of this loss function.

The DIMPA scheme is used to implement DIGRAC. It is a digraph extension of the KernelGCN method and is used to build node embeddings by capturing local network information. The authors describe how they row-normalize the adjacency matrix and add a weighted self-loop to each node. They also discuss how they define two feature mapping functions for source and target embeddings.

However, there is no explicit mention of ablation studies for these components in my part of the paper. The authors do discuss the efficacy of their method in the experiments section, but they do not isolate the impact of these individual components.

EXPECTED RESPONSE: Agent 0 should confirm receipt of this information and may provide further instructions or questions. They may also ask other agents for additional information if needed.

Agent 3 (worker): In my part of the paper, there is no explicit mention of ablation studies for the probabilistic imbalance loss or the directed mixed path aggregation scheme (DIMPA). However, there is a discussion about the performance of different loss functions, including a supervised loss function and a variant called L sort vol_sum. The paper suggests that adding L sort vol_sum to the loss could be beneficial and that the current choice "sort" performs best among these variants. There is no mention of a probabilistic imbalance loss or DIMPA in my section.

Expected response: Agent 0 should confirm receipt of my message and may provide further instructions or questions based on the information I provided. They may also share information from other agents regarding the ablation studies and the novel components of the method.

Agent 0 (leader): Step 3: Gathering information
The agents have confirmed that there are no explicit ablation studies mentioned in their parts of the paper.

They have also provided detailed information about the novel components of the method, the probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA).

Step 4: Evaluating the comment
Based on the information provided by the agents, it seems that the comment is valid. The paper introduces novel components but does not conduct ablation studies to evaluate their individual contributions. This could make it difficult for readers to understand the importance of each component and how they contribute to the overall performance.

Step 5: Refining the comment
Now, I will refine the comment to make it more specific and helpful. I will suggest specific experiments for the ablation studies and provide more context about why these studies are important.

SEND MESSAGE: The paper introduces a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA) as part of the proposed method. However, there are no ablation studies to evaluate their individual contributions. It would be beneficial to conduct experiments where the probabilistic imbalance loss and DIMPA are removed or modified one at a time, and the impact on the results is observed. This would help to understand the importance of each component and how they contribute to the overall performance. For example, an ablation study could involve comparing the performance of the method with and without the probabilistic imbalance loss, or with different variants of this loss function. Similarly, the impact of the DIMPA scheme could be evaluated by comparing the performance with other path aggregation schemes.

Expected response: The agents should confirm receipt of the refined comment and provide any additional feedback or suggestions.

Step 6: Finalizing the comment
Once the agents confirm the refined comment, I will finalize it. The final comment should be detailed, helpful, and important, providing clear guidance for the authors on how to improve their paper.

Agent 0 (leader): Apologies for the oversight. Let's correct that.

SEND MESSAGE TO ALL AGENTS: The paper introduces a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA) as part of the proposed method. However, there are no ablation studies to evaluate their individual contributions. It would be beneficial to conduct experiments where the probabilistic imbalance loss and DIMPA are removed or modified one at a time, and the impact on the results is observed. This would help to understand the importance of each component and how they contribute to the overall performance. For example, an ablation study could involve comparing the performance of the method with and without the probabilistic imbalance loss, or with different variants of this loss function. Similarly, the impact of the DIMPA scheme could be evaluated by comparing the performance with other path aggregation schemes. Please confirm receipt of this refined comment and

provide any additional feedback or suggestions.

Expected response: The agents should confirm receipt of the refined comment and provide any additional feedback or suggestions.

Agent 1 (worker): I confirm receipt of the refined comment. I agree that an ablation study would be beneficial to understand the individual contributions of the novel probabilistic imbalance loss and the DIMPA scheme. However, my part of the paper does not provide enough information to suggest specific alternatives for these components.

I expect a response that acknowledges my confirmation and feedback.

Agent 2 (worker): I confirm receipt of the refined comment. I agree with the suggestion of conducting ablation studies to evaluate the individual contributions of the novel probabilistic imbalance loss and the directed mixed path aggregation scheme (DIMPA). This would indeed provide valuable insights into the importance of each component and their contribution to the overall performance of the proposed method.

I expect a response confirming the receipt of my message and possibly further instructions or questions.

Agent 3 (worker): I confirm receipt of the refined comment. However, my part of the paper does not contain information about a probabilistic imbalance loss or a directed mixed path aggregation scheme (DIMPA). My section discusses the use of different loss functions, including a supervised loss function and L sort vol_sum, and their impact on the performance of the model. It also mentions the use of seed nodes to enhance performance. In the future work section, it discusses the possibility of extending the framework to detect the number of clusters and the performance in the sparse regime.

Expected response: Agent 0 should clarify if the probabilistic imbalance loss or DIMPA are mentioned in other parts of the paper or if they are related to the loss functions and methods discussed in my section.

Agent 0 (leader): Step 6: Finalizing the comment
All agents have confirmed receipt of the refined comment. Agent 3 mentioned that their part of the paper discusses different loss functions and the use of seed nodes, but does not mention the probabilistic imbalance loss or DIMPA. This information is consistent with what we know from the other agents and does not change the validity of the comment.

The final refined comment is:

"The paper introduces a novel probabilistic imbalance loss and a directed mixed path aggregation scheme

(DIMPA) as part of the proposed method. However, there are no ablation studies to evaluate their individual contributions. It would be beneficial to conduct experiments where the probabilistic imbalance loss and DIMPA are removed or modified one at a time, and the impact on the results is observed. This would help to understand the importance of each component and how they contribute to the overall performance. For example, an ablation study could involve comparing the performance of the method with and without the probabilistic imbalance loss, or with different variants of this loss function. Similarly, the impact of the DIMPA scheme could be evaluated by comparing the performance with other path aggregation schemes."

SEND MESSAGE TO ALL AGENTS: Thank you for your feedback. The final refined comment has been formulated. No further action is required at this time.

Expected response: The agents should confirm receipt of the final message.

Agent 1 (worker): I confirm receipt of the final message.

I expect a response that acknowledges my confirmation.

Agent 2 (worker): I confirm receipt of the final message. No further action is required from my side at this time.

Agent 0 (leader): SEND MESSAGE TO AGENT 3: Please confirm receipt of the final message.

Expected response: Agent 3 should confirm receipt of the final message.

Agent 3 (worker): Confirmation of receipt of the final message.

I expect a response acknowledging my confirmation.

Agent 0 (leader): SEND MESSAGE TO AGENT 3: Thank you for your confirmation. No further action is required at this time.

Agent 0 (leader): Apologies for the oversight. Let's correct that.

SEND MESSAGE TO AGENT 3: Thank you for your confirmation. No further action is required at this time.

Expected response: Agent 3 should confirm receipt of the final message.

Agent 0 (leader): Apologies for the oversight. Let's correct that.

SEND MESSAGE TO AGENT 3: Thank you for your confirmation. No further action is required

at this time.

Expected response: Agent 3 should confirm receipt of the final message.

Agent 0 (leader): My apologies for the confusion. The task has been completed successfully and no further communication is required.

Agent 0 (leader): {
"revised_comment": "The paper introduces a novel probabilistic imbalance loss and a directed mixed path aggregation scheme (DIMPA) as part of the proposed method. However, there are no ablation studies to evaluate their individual contributions. It would be beneficial to conduct experiments where the probabilistic imbalance loss and DIMPA are removed or modified one at a time, and the impact on the results is observed. This would help to understand the importance of each component and how they contribute to the overall performance. For example, an ablation study could involve comparing the performance of the method with and without the probabilistic imbalance loss, or with different variants of this loss function. Similarly, the impact of the DIMPA scheme could be evaluated by comparing the performance with other path aggregation schemes."
}