LLM-based Related Work Section Generation Framework Incorporating Perspectives Researchers Value

Anonymous ACL submission

Abstract

This paper proposes a Large Language Model (LLM)-based framework to generate paper's related work section, incorporating perspectives valued by researchers. While LLMs excel at summarization, ambiguous instructions limit the clarity of a generated related work section for researchers. Through the surveys, we identified the preferred perspectives for a related work section: "categorization", "comparison", and "pointing out problems". We incorporate these perspectives into a prompt with few-shot examples. Furthermore, to provide the framework with explainability and aid in the fact-checking, we have the LLM select salient sentences from cited papers to extract evidences. Experimental results with human evaluation demonstrate that the generated related work section tends to be preferred over human-written ones and has fewer hallucinations. Our codes and the dataset we collected are available at https://anonymous.4open.science/r/anony_rwg/.

1 Introduction

Scholarly papers serve as one of the most essential cornerstones in the development of science and technology (Doumont et al., 2014). Papers clearly convey new discoveries and ideas, being crucial means for the accumulation of human knowledge. Within sections of papers, the "Related Work" plays a pivotal role for it. A related work section not only presents a list of existing research but also provides the context for the current work. For authors, the related work section entails extensive reading, sorting, and analyzing numerous publications, making it a laborious and time-consuming task.

To alleviate this situation, recent studies focus on the task of "related work section generation" (Li and Ouyang, 2022). The field of this research is pioneered by Hoang and Kan (2010), and researchers utilize the natural language processing (NLP) techniques to generate related work sections. From the viewpoint of how a related work section is crafted, existing studies are roughly classified into the extractive and abstractive ones. The extractive methods identify the key sentence of cited papers based on importance scores and generate the related work section by concatenating sentences (Deng et al., 2021; Hu and Wan, 2014). In the abstractive methods, authors mainly utilize Transformer (Vaswani et al., 2017)-based architectures and try to summarize the contents of cited papers (Liu et al., 2023; Chen et al., 2022, 2021). While these models can explain crucial aspects of the methodology, the generated style of sentences reflects the average of training data. In typical research papers, the related work section exhibits variations, including straightforward enumerations, and scattered claims with similarities. Thus, the output sentence style may not be optimal for readers (researchers). Since scholarly papers are meant for humans so far, we believe that explicitly capturing the writing style preferred by the researchers is crucial.

Large language models (LLMs) like ChatGPT (OpenAI, 2023a) shed light on this perspective. By teaching its role, LLMs can change its behavior depending on the prompts. According to the report, the text summarization capability of GPT-4 is on per with human-level performance (Pu et al., 2023). However, as mentioned in Section 4.3, GPT-4 tends to just enumerate methodologies when we simply instruct it to output a related work section, which is not achieving human-satisfactory level. Creating a prompt explicitly tailored for this task is required.

In this paper, we propose a LLM-based framework to generate related work section, which incorporate perspectives valued by researchers. Figure 1 illustrates the main idea of the proposed framework. As shown in this figure, some important perspectives for a related work section is instructed to LLM via a prompt. To identify the perspective researchers value, we conduct two surveys.

First one is a questionnaire based survey. We asked researchers to itemize what they are careful about when writing a related work section using a free-response format. As a result, we identify five perspectives "Quality", "Freshness", "Categorization", "Comparison", and "Problem". In the second survey, we investigate papers published in the top conference to verify the above result. As we expected, these five perspectives are covered at a high rate in many papers. In particular, we focus on three perspectives –categorization, comparison, and problem– that can be explicitly instructed to LLMs. We incorporate them into a prompt with few-shot examples. Additionally, we concentrate on the hallucination problem, wherein the output of the LLM includes incorrect sentences (Ji et al., 2023; Bang et al., 2023; Cao et al., 2018; Azaria and Mitchell, 2023). To assist users in fact-checking, we adopt a mechanism into a prompt to extract evidence from cited papers, providing the framework with explainability. Finally, through the experimental results with human evaluation, we demonstrate that the generated related work section tends to be preferred over human-written sections and has fewer hallucinations.

The contributions of this paper are as follows:

- We identify the perspectives needed in a related work section via surveys. The results are useful for not only researchers engaging in generating related work sections, but also researchers who would like to write a good related work section.
- Based on findings of surveys, we propose LLM-based framework that can generate a related work section for given cited papers. To the best of our knowledge, this is the first paper which demonstrates the possibility that the generated related work section outperforms human-written one through the human evaluation.
- For the development of this research field, we make our codes and the collected dataset publicly available.

The rest of this paper is organized as follows. We describe the preliminaries in Section 2. The proposed framework is presented in Section 3. Section 4 demonstrates the experimental results of human evaluations. Section 5 is the related work section, which is composed of the output of the proposed framework. In Section 6, we discuss the output related work section. Finally, we conclude the paper in Section 7.

2 Preliminaries

In our framework, we use contents of existing papers to generate a related work section. To clarify the paper we focus on, we use two terms, *target paper* and *cited papers* by following Chen et al. (2022). A target is the paper in which a related work section is generated. Cited papers are referred in the related work section of a target paper.

As the contents of papers, we use each paper's introduction. Since introductions generally include the essential information of papers, we believe that an effective related work section can be generated from them. Considering usability, we adopt an in-context learning approach. In this manner, we do not fine-tune the model and opt for GPT-4-turbo (OpenAI, 2023b) as a backbone LLM, leveraging few-shot prompting (Brown et al., 2020).

Formal definition of our work is as follows. Let n_c represent the number of the cited papers. Given the set of the information (title, author names, and introduction) of cited papers $C = \{c_j^{\text{info}} \mid 1 \leq j \leq n_c\}$ and that of a target paper t^{info}, our goal is to find a prompt p such that a generated related work section sentence $\hat{R}_{\text{target}} = \text{LLM}(C, t^{\text{info}} \mid p)$ is well preferred by researchers. To generate $\hat{R}_{\text{target}}$, existing works often use the set of actual related work sections $R_{\text{train}} = \{R_k \mid 1 \leq k \leq n_{\text{train}}\}$ as the ground truth data for training, where n_{train} is the large number of training samples. On the other hand, we use few-shot examples of related work sections instead of using R_{train}.

Note that there is a research field dedicated to efficiently seeking relevant studies in a particular area of research (van Dinter et al., 2021). We assume that the cited papers are given by authors of the target paper, and seeking them is out of the scope of our work.

3 Proposed Framework

3.1 Modes of Related Work Section

To gather insights for designing an effective framework for generating a related work section, we conducted the following two surveys on how researchers typically write an related work section.

3.1.1 Questionnaire: Important Points?

The purpose of this survey is to investigate what researchers are consciously considering when writing

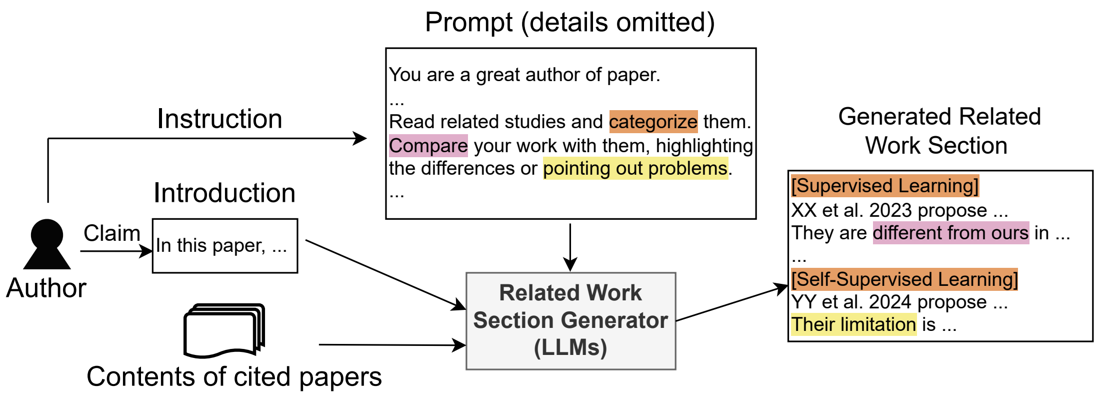


Figure 1: The overview of the proposed framework. The author inputs his/her introduction and contents of papers to the LLM with the proposed prompt. In the prompt, perspectives researchers value (identified by our surveys) are emphasized. The output is designed to include the perspectives of categorization, comparison, and problem.

a related work section of their papers. The respondents are 30 researchers in universities and enterprises including students. To explore the differences of consciousness based on the respondents' experiences, we prepared two questions:

- "How many peer-reviewed papers (including conferences proceedings) have been accepted for publication as the first author?"
- "Itemize what you are careful about when writing a related work section (What are important things for a good related work section?)"

For the second one, we opted for free-response format instead of providing choices in order to avoid biases. We then organized the collected answers into several perspectives. See Appendix A for the screenshot of this questionnaire.

The six perspectives we extracted from answers are as follows: **Quality**: Cited papers include papers of top conferences/journals. **Freshness**: Cited papers include papers published in a few years. **Categorization**: Cited papers are categorized into several categories. **Comparison**: Their proposals are explicitly compared with cited papers. **Problem**: Authors should point out the problems/limitations of cited papers. **Others** Other perspectives from above.

Although the comparison perspective generally includes the problem perspective, we separate them because of the broad scope of the comparison perspective. For each answer, we check if it includes each perspective and report the average inclusion rate. Figure 2 shows the survey results on important points in writing a related work section. As depicted in this figure, respondents with substantial experience tend to consciously consider various perspectives when writing a related work section. While the class of more than 5 tends to pay attention to quality and freshness, respondents with less experiences do not exhibit this tendency. This is likely because experienced researchers are aware that quality and freshness are often mentioned in the reviewing process. Thus, to generate a good related work section, cited papers input to the framework should include famous and newer ones.

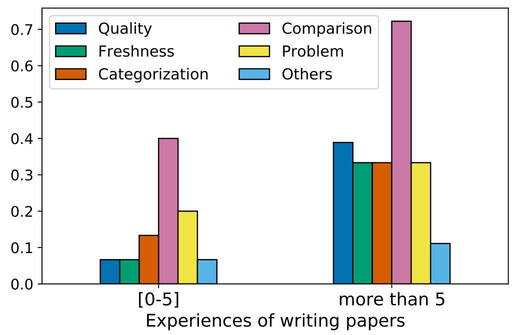


Figure 2: The survey results on important points in writing related work section.

As for other perspectives, we can see that whole respondents believe comparison is important, and thus the comparison perspective should be emphasized. When emphasizing comparison perspective, we believe both categorization and problem perspectives become important. Note that "others" include opinions such as "narrativity", "length" and "avoidance of excessive self-citation".

3.1.2 How Are Papers in A Top Conference?

To verify the survey results in Section 3.1.1, we investigated papers of ACL2023 (top conference in

Table 1: Covered perspectives rate in ACL2023 papers.

Perspective	Rate
Quality	1.00
Freshness	0.92
Categorization	0.82
Comparison	0.84
Problem	0.62

the field of NLP). Specifically, we randomly choose 50 papers including "Related Work" section. If a paper cites more than three top conference papers (A* or A rank at ICORE Ranking [1]), we regard it as satisfying quality perspective. Also, a paper that cites more than three papers published after 2021 is regarded as fresh. Categorization, comparison, and problem are the same criteria in Section 3.1.1.

Table 1 shows the rate of covered perspectives in ACL2023 papers. As we expected, all perspectives are covered at a high rate in many papers. This result indicates that all perspectives are reasonable and supported by leading researchers. There are some gaps between questionnaire result and this result, which reflect phenomena that researchers actually do but not consciously carry out. Note that the rate of the problem perspective in Table 1 is lower than that of the others. As an evidence, we find that some papers avoid pointing out problems or limitations by clarifying and emphasizing what authors do. This observation also highlights the significance of the comparison perspective.

From two survey results above, we observe that all perspectives play a crucial role in a related work section. As for categorization, comparison, and problem perspectives, we can explicitly instruct them to a LLM. Thus, we propose a framework centered around these three perspectives.

3.2 Methodology

By incorporating the identified perspectives into the prompt, we mimic the process of sophisticated researchers in writing a related work section. The goal is to generate a related work section with a writing style preferred by researchers. Figure 3 shows the main part of the proposed prompt. As shown in this figure, the prompt is divided into two parts: the *role playing part* and the *examples part*.

In the role playing part, we instruct that the role of the LLM is behaving like an author of a great paper and writing a great related work section. Furthermore, the way to structure a related work section is described by steps. By Step-1, we have the LLM recognize the main claim of the target paper. In Step-2, we include a mechanism to extract evidences of an output by instructing to select the salient sentences of cited papers. These are used for user to fact-check the output. Step-3 incorporates the "categorization" perspective and instructs to categorize the related studies based on given contents. The instruction to make subsections based on the established categories is also provided. Step-4 incorporates the two perspectives "comparison" and "problem". Note that we give the LLM options to emphasize comparison perspective or problem perspective. This is because all authors do not necessarily point out the problems of cited papers, as shown in Table 1.

In the examples part, we give the LLM a great example and a bad example. These examples are used to emphasize the perspectives and define the output style. We select the great example which satisfies all the perspectives from ACL2023 papers [2]. The bad example is crafted by removing the perspectives from the great example. By using Feedback, we teach the reason why each sample is good or bad. The Feedback includes the things we would like the LLM to follow. Thus, we do not mention the problem perspective here by the same reason as the case of Step-4. The output format of the evidence extraction is also defined in the examples part (Results of Step-2 in a great case). Salient sentence examples are also manually selected from the introduction of the great example.

Following this prompt, a title and an introduction of a target paper are provided by adhering the format defined in the great case. Similarly, the titles, author names, and introductions of cited papers are concatenated and given to the LLM.

4 Experiments

We evaluate the proposed framework to answer two research questions (RQ). The first one is **RQ1: How effective are related work sections generated by the proposed framework for humans?** Additionally, we focus on the hallucination problem (Bang et al., 2023; Cao et al., 2018). Even if the answer to RQ1 is satisfactory, an output containing numerous hallucinations would be rendered mean-

[1] https://www.core.edu.au/icore-portal

[2] To create these examples, we utilize the related work and introduction sections of a great work by Gao et al. (2023), while adhering to the CC-BY 4.0 license.

% ——— **Role playing part** ———
You are the author of the great paper.
Write a "Related Work" section based on the "Introduction" you have already written and the information about the related studies provided. Note that you must cite all of the listed related studies. Do not cite any papers that are not listed with their titles and introductions. % This part is also used in a baseline (Pure-GPT).

The authors of an excellent paper structures a "Related Work" section as follows.
Step-1: The authors confirm the authors' introduction to clarify what and how the authors have solved in the paper.
Step-2: The authors collect information on related studies. Then the authors carefully read each study's introduction and select the salient sentences.
Step-3: The authors categorize the related studies based on the selected salient sentences and their own introduction in order to write the Related Work section. Subsequently, the authors create subsections aligned with the established categories and assign concise and clear names to each subsection.
Step-4: Within each subsection, a comparison is made between related studies and the authors' work, focusing on what needs to be addressed and highlighting the differences or pointing out problems. Note that they ensure that differences or problems do not overlap across subsections. If there is any duplication, re-categorize accordingly.

% ——— **Examples part** ———
Below are a great case and a bad case as examples. In the examples, some of them are omitted, but you must not omit them.
=====
<Great case>
[Your Title: **(actual title of a paper that satisfies all the perspectives.)**]
[Your Introduction: **(actual introduction of a paper that satisfies all the perspectives.)**]
[Information about Related Studies: (omitted)] % this omitted is care for the 2nd sentence.

Results of Step-2:
Selected salient sentences from (X et al., 2019):
"**(manually selected sentences like:)** [...] In this work, we introduce [...] The main challenge to [...] The key insight [...]"

Selected salient sentences from (Y et al., 2020):
"**(manually selected sentences like above)**"

Related Work Section:
(actual contents in related work section that satisfies <Great case>. Subsections are represented by #### like:)
Category name
[...] In comparison, we use [...] the problem of building effective [...]

[**Feedback**: This related work section is very good. The reasons are:
- Authors categorize the cited papers by subsections.
- Authors pointed out the difference between their paper and existing papers.]
====
<Bad case>
[Your Title: **(the same title as <Great case>)**]
[Your Introduction: **(the same introduction as <Great case>)**]
[Information about Related Studies: (omitted)]

Related Work Section:
(The sentences of the great case in which good points are manually removed.)
[**Feedback**: This related work section is not good. The reasons are:
- Authors do not categorize papers. They just enumerate existing papers.
- Authors do not mention the relationship between their paper and existing papers.]
====

Figure 3: The main part of the proposed prompt. In this prompt, perspectives researchers value (identified by our surveys) are colored. Note that **(bold)** indicates the sentences are omitted here. Please see an actual prompt on Anonymous Github. The sentence after % is our comment. In the role playing part, we instruct that the role of LLM is behaving like a great author and three important perspectives are incorporated in Step-3 and 4. In the examples part, we give the LLM two examples (a great case and a bad case) to assist output style.

ingless. Thus, we must address **RQ2: To what extent does the proposed framework exhibit hallucination?** All experiments are conducted in English. As for the GPT-4-turbo hyperparameters, temperature is set to 0, and other parameters are set to the default values.

4.1 Evaluation Methodology

Experiment 1: To answer RQ1, we conduct a human evaluation. The participants are experienced researchers in the field of artificial intelligence who are colleagues of this paper's authors. The participants compare three related work sections: human-written (**Original**), the output of the proposed framework (**Proposal**), and that of **Pure-GPT**. Pure-GPT is a baseline that uses a simple prompt based on the *italic part* in Figure 3. We randomly present these three related work sections to participants, anonymizing them as A, B, and C. The participants are requested to judge which ones are preferred in a pairwise comparison manner. That is, they check the pairs (A, B), (A, C), and (B, C) by choosing options from: "X is better than Y", "Y is better than X", and "X and Y are of equivalent quality". In addition, for each related work section, they answer the following three questions with yes or no: "Does it properly categorize related studies?", "Does it compare the author's work with related studies?", and "Does it mention the challenge/limitations of related studies?". Note that the hallucination issue is addressed in the next experiment. Hence, participants assume each cited paper's description is factual and are asked to review and select the options. The details can be found in Appendix A.

Experiment 2: To answer RQ2, we read the output sentences and assess whether descriptions are correct. We separate the output into three parts: descriptions of cited papers, extracted evidences, and descriptions of the target. For the description of cited papers, we assign scores of 0 (incorrect), 0.5 (not incorrect but less confidence), and 1 (correct). If a given cited paper is not cited by the LLM, we skip score assigning process. Alternatively, we report the ignored citation rate. For extracted evidences, we check if they include hallucinations or not. Besides, to evaluate the effectiveness of extracted evidences, we define the hit rate. This is the rate of descriptions for cited papers that can be labeled as correct solely based on the extracted evidence. For the description regarding target, we check if each of them includes hallucination or not.

4.2 Dataset

In the experiments, we use 10 human-written (target) papers randomly collected from ACL2023 long papers. This is because the the common research area of the researchers participating in this experiment is NLP. For each target paper, we manually compiled contents of all cited papers into JSON format [3]. As this process is labor-intensive, we make the collected data publicly available [4] to contribute to the activation of the research community. Note that we checked each paper's license as mentioned in Ethical Consideration.

4.3 Results

RQ1: Effectiveness for Humans: Table 2 shows the experimental results on effectiveness of each method for humans. As we can see from this table, the win rate of Proposal is greater than 0.5 in both cases of vs Original and vs Pure-GPT. The task of generating the related work section includes elements of the summarization task. In this sense, this result shows that the proposed prompt successfully leverages the powerful summarization capabilities of GPT-4-turbo to generate refined related work sections preferred by researchers. Given that the papers are collected from the top conference, this result is unexpected for us. Careful readers might notice that proper categorization rate at Original is lower than the statistics shown in Table 1. While we regard the papers using subsections or paragraphs to group existing works as satisfying category perspective in Section 3.1.2, the proper categorization is required in this experiment. Actually, 7 out of 10 papers used in this experiment are using subsections or paragraphs to group existing works. This means that the participants at least judge the categorization by the proposed framework to be more correct than the ones by humans. Comparing Proposal with Pure-GPT, Proposal outperforms Pure-GPT with a large margin. As can be seen from the perspective satisfied, Pure-GPT tends to just enumerate descriptions of each method without comparing or categorizing. This result also indicates that the identified perspectives are crucial in writing a related work section.

RQ2: Hallucination Issue: Table 3 shows the correctness of the generated sentences by the pro-

[3]To automatically collect the dataset, we attempted to use some tools that parse PDFs of research papers. However, this attempt failed due to issues such as inaccurate parsing occurred, resulting in manual collection.

[4]This will be available when this paper is published.

Table 2: Effectiveness of each method for humans.

	Win rate		Perspective satisfied		
	vs Original	vs Pure-GPT	Categorization	Comparison	Problem
Proposal	**0.56**	**0.90**	0.90	0.90	0.40
Original	-	0.80	0.40	0.60	0.50
Pure-GPT	-	-	0.20	0.20	0.10

posed framework. As we can see from this table, the description correctness of cited papers is nearly 1.0, meaning that generated sentences regarding cited papers are almost correctly written. While there are no cited papers' description judged as 0 (incorrect), some papers are not included in the output, meaning that GPT-4-turbo ignores them. This issue occurs in 50% of target papers and the ratio of ignored cited papers per target is 9%. We observe that this issue tends to happen in the cases: (1) The surname of the author matches the surname of another paper's author. (2) The context can be established without citing it (e.g. papers with different properties, such as papers describing old background technique).

The extracted evidences also do not include hallucinations. On the other hand, some sentences vary in summarized forms despite the instruction to select salient sentences of cited papers. In any case, its usefulness is clear as 70% of the extracted evidences are directly used for fact-checking. Appendix B covers an example of the extracted evidence.

Unfortunately, the part of target paper's description includes a small amount of hallucinations. We find two target papers' descriptions include hallucinations in one or two sentences. In both cases, we observe that GPT-4-turbo attempts to forcibly compare methodologies and mentions groundless stuffs. Note that after experiments, we observe that the hallucinations disappear by repeating generation. The proposed framework is a powerful tool for generating a draft of the related work section with a refined structure, incorporating perspectives valued by researchers. Authors should carefully review the generated draft and address the hallucination issue, from an ethical standpoint.

5 Related Work

The sentences below are the generated related work section where this paper is the target. Although we performed the generation several times and adopted the best one, we do not manually change any single word of the output. Discussion regarding this generated section is done in Section 6.

Table 3: The correctness of the generated sentences by the proposed framework.

Description correctness of cited papers	0.96 ↑
Ignored papers inclusion rate (/target)	0.50 ↓
The ratio of ignored cited papers (/target)	0.09 ↓
Hallucination rate in extracted evidences	0.00 ↓
Hit rate of extracted evidences	0.70 ↑
Hallucination rate in target paper's part	0.20 ↓

Foundations of Related Work Generation: The task of generating related work sections has been recognized as a significant challenge within the automatic summarization community (Hoang and Kan, 2010). Early efforts in this domain focused on understanding the argumentative patterns in related work sections and exploring summarization tactics for their creation. Hu and Wan (2014) advanced this field by proposing an optimization approach to automatically generate related work sections, selecting sentences from both the target and reference papers to form a coherent narrative. This extractive approach laid the groundwork for subsequent research in the area.

Abstractive Approaches and Causal Interventions: Moving beyond extractive methods, recent studies have introduced abstractive techniques to generate more sophisticated summaries. Liu et al. (2023) introduced a Causal Intervention Module for Related Work Generation (CaM), aiming to mitigate the impact of spurious correlations in the generation process. Similarly, Chen et al. (2021, 2022) developed models that not only abstract content from multiple papers but also capture the relationships between them, with the Relation-aware Related work Generator (RRG) and the target-aware related work generator (TAG), respectively. These models represent a shift towards generating sections that are not only informative but also contextually aware of the target paper's contributions.

Extractive Methods and Sentence Reordering: Despite the advancements in abstractive methods, extractive approaches remain relevant. Deng et al. (2021) proposed a method for generating descriptive related work sections by extracting salient sentences and reordering them logically. This method emphasizes the importance of maintaining the accuracy and objectivity of the original texts while presenting them in a structured manner.

Meta Studies and Frameworks: Li and Ouyang (2022) provided a meta-study of automatic related work generation, critically reviewing the state-of-the-art from various perspectives, including problem formulation and methodological approaches. This comprehensive analysis underscores the complexity of the task and the need for a holistic understanding of the various components involved in generating related work sections.

Novel Models and Techniques: Wang et al. (2019) introduced a novel Bayesian model that probabilistically links the target paper with reference papers, capturing the relevance between them. This approach highlights the importance of modeling the relationships between papers in generating related work sections. Wang et al. (2021) and Ge et al. (2021) both proposed frameworks that integrate background knowledge and content information, with AutoCite focusing on multi-modal representation fusion for contextual citation generation and BACO emphasizing the generation of citing sentences using background knowledge from citation networks.

Contributions of the Current Work: Our work builds upon these foundations by proposing an LLM-based framework that incorporates perspectives valued by researchers, such as categorization, comparison, and problem identification. Unlike previous methods, our framework is designed to generate related work sections that are not only informative and contextually aware but also aligned with the writing style preferred by researchers. By conducting surveys to identify the perspectives researchers value most, we have tailored our LLM prompts to produce related work sections that are preferred over human-written ones and exhibit fewer hallucinations. This approach represents a significant step forward in the automatic generation of related work sections, combining the strengths of both extractive and abstractive methods with the nuanced understanding of researcher preferences.

6 Discussion

The generated related work section is sufficiently organized for our draft. We fact-check the generated contents and there are no hallucinations. As can be seen, while the categorization and comparison perspectives are satisfied, the problem perspective is not included in the generated sentences, corresponding to the case mentioned in Section 3.2. We believe this is because we do not explicitly point out the problem of each method in our introduction. The case where the problem perspective is satisfied is shown in Appendix B. For the categorization perspective, although we would prefer to merge the categories of "Extractive Methods and Sentence Reordering" and "Meta Studies and Frameworks" to the first category, the categorization itself seems to be correct. As concerns, there are vague category names such as "Novel Models and Techniques". Also, the proposed framework sometimes makes an independent category for the target paper. From the viewpoint of using the proposed framework as a drafting tool, these are acceptable since renaming or removing them is a painless work. For the comparison perspective, although the claim of the proposal is relatively long, the different points from existing works are emphasized. The proposed framework is a practically effective tool since shortening claims is also not tough work for authors.

Note that we believe this paper is the more difficult case than that of papers in other fields. While the meaning of "related work" is generally "similar research", it indicates "section" in our context. In addition, each paper includes many words of "paper", such as the target and cited papers. Considering most of papers state the claims after "in this paper", capturing crucial parts becomes more difficult. Thus, we consider these uncommon features may complicate the interpretation of the context.

7 Conclusion

In this paper, we have proposed the framework to generate paper's related work section based on LLM. Through surveys, we identified the perspectives researchers value in writing related work section. The perspectives "comparison", "categorization", and "pointing out problems" are incorporated into the proposed prompt. Through the experiments using top conference papers, we demonstrate the possibility that the generated related work section by the proposed framework tends to be preferred over human-written ones and that of straightforward prompt-based method.

Limitation & Ethical Consideration

While the proposed framework generates the related work section based on text, it does not take into consideration the figures and tables mentioned within each paper. By leveraging multimodal models, capable of handling both text and images, there is an expectation for the development of frameworks that can consider these elements as well (Yin et al., 2023; Fu et al., 2023). Furthermore, if high-quality datasets containing structured figures, tables, and full text of papers were available, we might generate not only a related work section but also other sections such as an introduction section.

As for the dataset we provide, it includes papers that have been made publicly available as open access under the CC-BY 4.0 license [5].

As mentioned in Section 4.3, we reported the presence of hallucinations to some extent. The occurrence of hallucinations has decreased with the performance improvement of LLMs, which still poses a significant concern (OpenAI, 2023b; Zhang et al., 2023). Although the proposed framework can be a powerful tool to generate drafts for a good related work section, it does not eliminate the need for thorough review and appropriate revision by authors. From an ethical standpoint, authors should bear the responsibility of verifying all generated content before it is published, regardless of the presence of hallucinations.

Moreover, one potential issue associated with the development of automated writing methods like ours may hinder the growth of researchers' research skills. For researchers, writing papers plays a crucial role not only in disseminating scientific findings but also in enhancing their skills. Through the writing process, they may gain a deeper understanding of the related research, improve their presentation abilities, and clarify the direction of their research.

Finally, there is a concern about the misuse of these automated writing methods for creating fake scientific papers, posing ethical issues that need to be appropriately addressed.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of EMNLP*, pages 967–976, Singapore.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of IJCNLP and AACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*, volume 33, pages 1877–1901.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of AAAI*, volume 32.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Target-aware abstractive related work generation with contrastive learning. In *Proceedings of SIGIR*, pages 373–383.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of ACL and IJCNLP*, pages 6068–6077, Online.

Zekun Deng, Zixin Zeng, Weiye Gu, Jiawen Ji, and Bolin Hua. 2021. Automatic related work section generation by sentence extraction and reordering. In *Proceedings of AII@ iConference*, pages 101–110.

Jean-Luc Doumont, Laura Grossenbacher, Christina Matta, and Jorge Cham. 2014. English communication for scientists.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of ACL*, pages 1762–1777, Toronto, Canada.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A background knowledge- and content-based framework for citing sentence generation. In *Proceedings of ACL and IJCNLP*, pages 1466–1478, Online.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Proceedings of Coling (Posters)*, pages 427–435, Beijing, China.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of EMNLP*, pages 1624–1633, Doha, Qatar.

[5] https://creativecommons.org/licenses/by/4.0/deed.en

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Xiangci Li and Jessica Ouyang. 2022. Automatic related work generation: A meta study.

Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, Liang Hu, and Ivor Tsang. 2023. Causal intervention for abstractive related work generation. In *Findings of EMNLP*, pages 2148–2159, Singapore.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. New models and developer products announced at devday. Accessed on Feb. 15, 2024.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, volume 30.

Pancheng Wang, Shasha Li, Haifang Zhou, Jintao Tang, and Ting Wang. 2019. Toc-rwg: explore the combination of topic model and citation information for automatic related work generation. *IEEE Access*, 8:13043–13055.

Qingqin Wang, Yun Xiong, Yao Zhang, Jiawei Zhang, and Yangyong Zhu. 2021. Autocite: Multi-modal representation fusion for contextual citation generation. In *Proceedings of WSDM*, pages 788–796.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models.

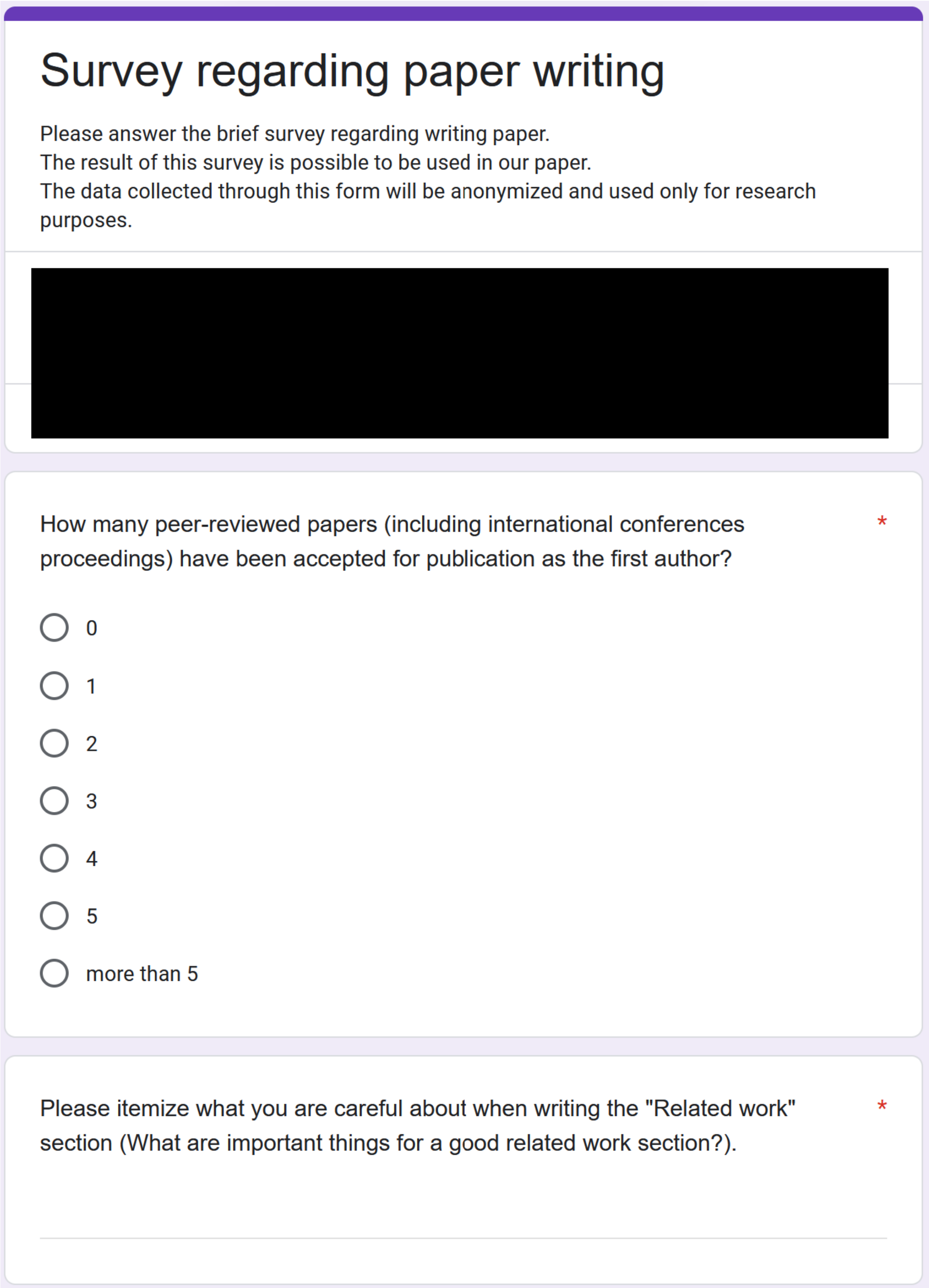
Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball.

A Survey and Human Evaluation

We conduct a survey to explore important perspectives for the related work section generation and perform a human evaluation experiment to assess the generated related work sections. Figure 4 shows the instructions and response form for the perspective survey. Regarding the evaluation of the generated related work sections, as indicated in Figure 5, participants are instructed to assess them using only their human abilities (without tools like ChatGPT). After reading and agreeing to these instructions, participants evaluate each related work section, as illustrated in Figure 5, and then compare the related work sections as shown in Figure 6.

B Output Example

Figure 7 shows an example of output salient sentences. In this example, the LLM can correctly extract the parts which describe the cited paper. Figure 8 shows the part of the output related work section by the proposed framework. As shown in this figure, the output includes the problem perspective.



Survey regarding paper writing

Please answer the brief survey regarding writing paper.
The result of this survey is possible to be used in our paper.
The data collected through this form will be anonymized and used only for research purposes.

[redacted]

How many peer-reviewed papers (including international conferences proceedings) have been accepted for publication as the first author? *

- 0
- 1
- 2
- 3
- 4
- 5
- more than 5

Please itemize what you are careful about when writing the "Related work" section (What are important things for a good related work section?). *

Figure 4: A screenshot of the questionnaire for the perspective survey. (The blacked-out area is for concealing Google account information.)

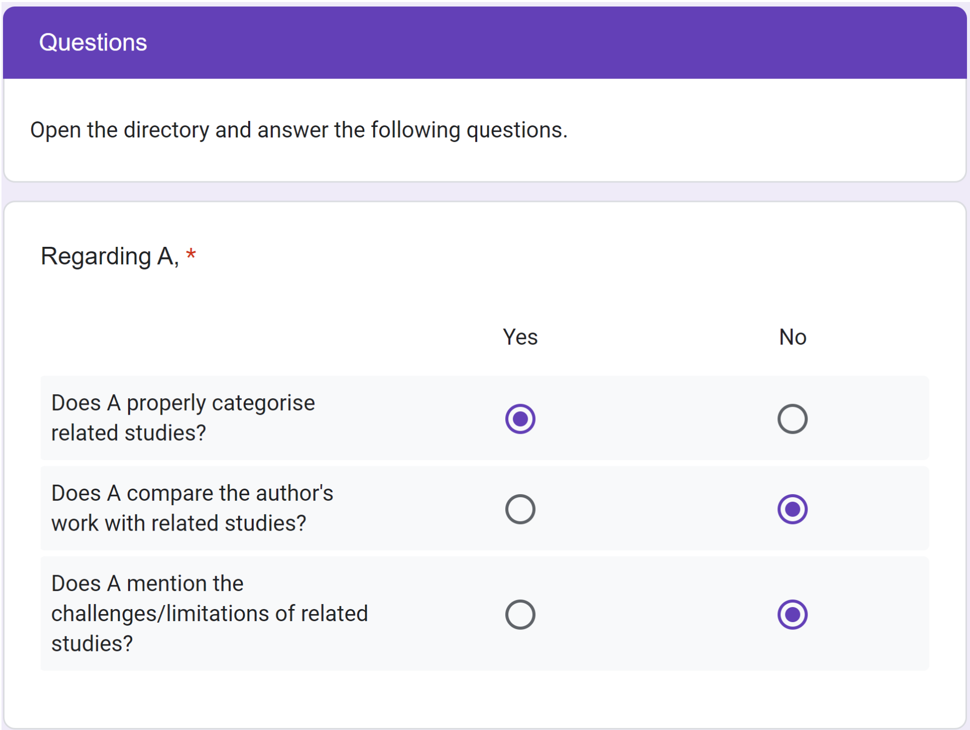
Evaluation of Related Work Sections in Research Papers

Thank you for your participation.
Your task involves reviewing the content of three Related Work sections for each paper and providing an assessment.
The data has been pre-shared in a zip file, and its name is based on the data number (e.g., 22). Inside the directory, you will find three text files for the Related Work sections (A.txt, B.txt, C.txt) and an Introduction file (introduction.txt).
Flow:

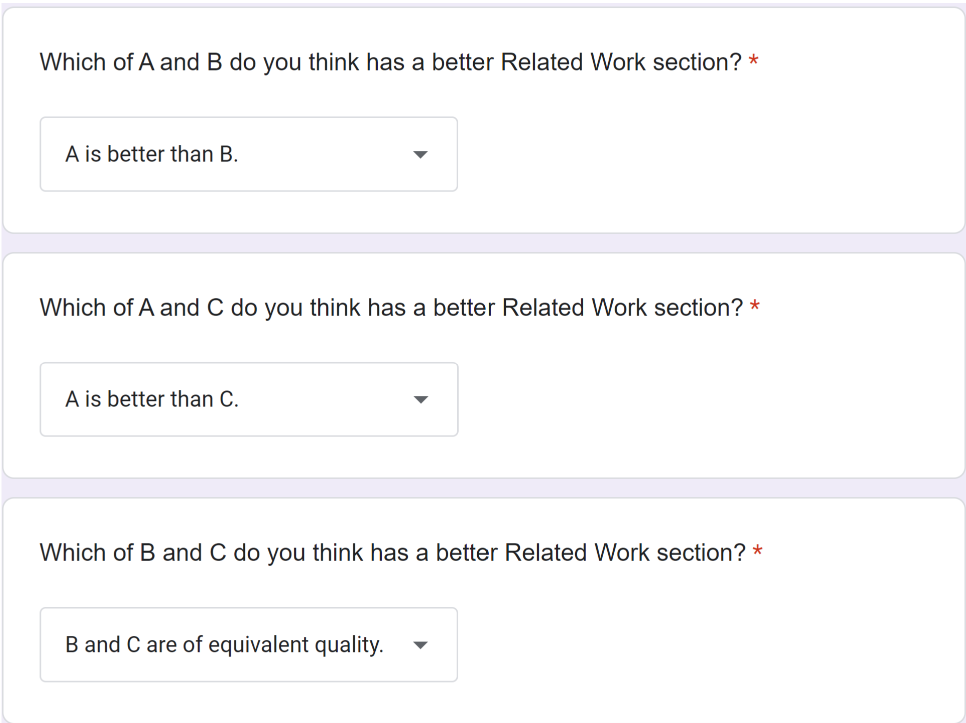
1. **Read Introduction:**
   - Open the file named introduction.txt and familiarize yourself with the content.
   - Understand the context of the paper based on the introduction.
2. **Evaluate Related Work Sections (A.txt, B.txt, C.txt):**
   - Open each Related Work section file (A.txt, B.txt, C.txt).
   - Read them carefully so that you can answer the questionnaire items.
   - The evaluation includes a comparative format, comparing two Related Work sections.

Important Notes:
Please read and evaluate without relying on tools such as ChatGPT, solely using human (your) capabilities. Please evaluate the content from cited papers assuming it is factual.
The data collected by this form will only be used for research purposes.

Figure 5: A screenshot of the instructions given to participants for the evaluation of the generated related work sections.



(a) Participants evaluate related work sections anonymized as A, B, and C, respectively.



(b) Participants conduct comparative evaluations for each pair.

Figure 6: Screenshots of the questionnaire used for the evaluation of the generated related work sections.

Selected salient sentences from (Han et al., 2021):
"[...] we propose prompt tuning with rules (PTR) for many-class classification tasks. [...] PTR can apply logic rules to encode prior knowledge about tasks and classes into prompt tuning."

Figure 7: The output example of the evidence extraction.

Hate Speech Detection and Datasets
The detection of hate speech is a critical and challenging task that has garnered significant attention within the NLP community (Poletto et al., 2020; Schmidt et al., 2017). While traditional machine learning techniques have been employed to predict hate speech from social media (Al-Makhadmeh et al., 2020), the complexity of hate speech and its multifaceted nature make it difficult for both humans and automated systems to recognize effectively (Poletto et al., 2020). The need for large, clearly annotated datasets is paramount for the development of robust and nuanced classification systems (Vidgen et al., 2021). However, existing models often struggle with generalization and lack explanations for their decisions (Mathew et al., 2022), highlighting the importance of interpretability in hate speech detection systems.

Figure 8: The output example by the proposed method (Problem perspective is included).