
000
001 REDHAT : REDUCING HALLUCINATION IN ESSAY
002 CRITQUES FROM LARGE LANGUAGE MODELS
003
004

005 **Anonymous authors**

006 Paper under double-blind review
007
008
009

010 ABSTRACT
011

012 Essay critiques refer to the textual assessment of an essay, serving as the basis for
013 the grading of the essay, and are also crucial for the improvements of the essay.
014 Essay critique generation has received increasing attention after the blooming of
015 large language models (LLMs), which show promising potential in writing and
016 critiquing essays. However, current LLMs suffer from hallucinations when gener-
017 ating essay critiques (e.g., baseless criticism), which are still under-explored in the
018 community. To facilitate research in reliable essay critique generation, we first de-
019 fine this task with a unified input-output format as well as clear evaluation criteria.
020 To minimize hallucinations in critique generation, we introduce RedHat, a novel
021 approach that embeds the key information from an essay directly into the gener-
022 ation process through document-level question-answering, ensuring critiques stay
023 firmly anchored to the evaluated content. We collected a large-scale, high-quality
024 essay critique dataset called `EssayC`, annotated by human experts over multiple
025 LLM-generated critiques, from an undergraduate essay writing course. We ex-
026 perimented RedHat backboned by commercial and open-sourced LLMs. Results
027 showed that critiques generated by RedHat are preferred by auto-judger and hu-
028 man experts over baseline in around 20% of cases on `EssayC` in ambiguity and
029 informativeness, with a decrement around 10% on hallucinations under our eval-
030 uation criteria.

031 1 INTRODUCTION
032

033 Essay critiques are pivotal for grading writings (Triawan et al., 2023; Suresh et al., 2023; Wang
034 et al., 2018), providing constructive feedback (Abbas & Herdi, 2018) and improving writing
035 skills (Noroozi et al., 2023). With the advancement of large language models (LLMs) (Ouyang
036 et al., 2022; Rafailov et al., 2024; Ethayarajh et al., 2024), LLM-as-a-judge (Zheng et al., 2024a)
037 based critique models have shown promising results in providing explainable and informative cri-
038 tiques in instruction following tasks (Ke et al., 2023; OpenAI, 2024a). Although applying LLMs in
039 essay assessment seems promising, our study found that LLMs are plagued by hallucinations when
040 generating essay critiques and therefore not suitable for direct application.

041 Hallucination in LLMs refers to the phenomenon that the generated content is not grounded on
042 factual or correct information (Rawte et al., 2023). Figure 1 presents hallucinations from GPT-
043 4o (OpenAI, 2024b) generated essay critiques. It exhibits two typical types of hallucination in this
044 task: (1) providing advice that is not appropriate nor does not match the essay content, and (2)
045 proposing fallacies that do not exist in the assessed essay. These hallucinated critiques significantly
046 hinder the usability of LLM in essay critique generation.

047 Existing research focuses on instructing LLMs to automated essay scoring (AES) (Kundu & Bar-
048 bosa, 2024), yet improving critique quality is still under-explored. Lack of consensus on how to
049 evaluate an essay in detail leads to such negligence in critique improvement. First, the essay is a
050 form of open-ended generation (Brahman et al., 2022), ranging from narrative to argumentative, each
051 with distinct purposes. Detailed requirements differ between writing an analysis part and a conclu-
052 sion part. This complicates the detailing of assessment criteria in the evaluation prompt, leading to
053 the fact that type I hallucination in Figure 1 often happens. Unfortunately, human expert evaluation
is extremely costly and inefficient (10 seconds for LLMs versus half an hour for human) for assess-

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

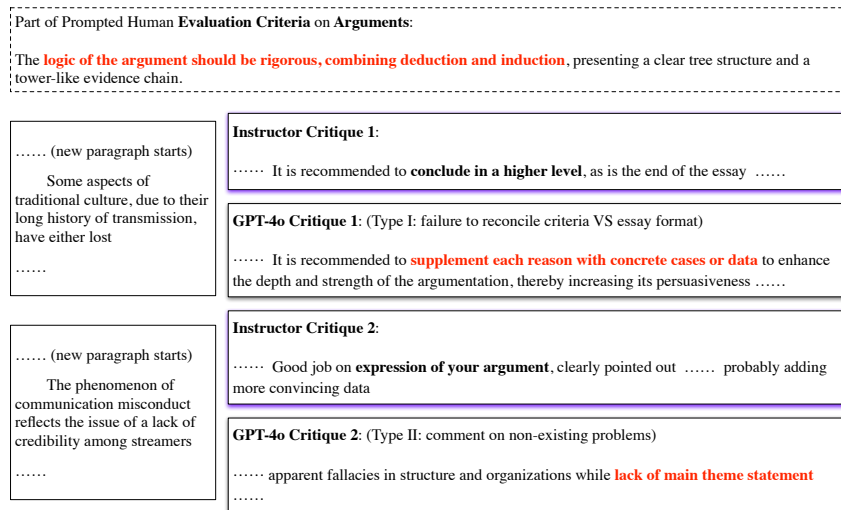


Figure 1: **Essay critique hallucination explained.** Here we listed two typical hallucinations caused by LLM’s overly following evaluation criteria in the prompt of the whole essay when using GPT-4o-mini to generate essay critique. In the case of **Type I** hallucination, we find that GPT-4o-mini is overly criticizing a conclusion part using descriptions from the criteria. In the case of **Type II** hallucination, GPT-4o-mini does not capture the author’s argument expressed in complex format and presented in the beginning. GPT-4o-mini is fed with the whole essay and criteria. The prompt for generating critique is listed in Appendix I.

ment both for essays and critiques, causing a lack of research resources, especially for the detection of Type II hallucination in Figure 1. These factors hinder the understanding and de-hallucination of LLM-generated essay critiques.

In response to the lack of resource challenges, we start with an undergraduate writing training course. We regard such scenarios as a generalizable scene of essay critiquing, with abundantly available expert annotations. In our preliminary study of the application of LLMs in such a course, we found that the LLM hallucination behavior in essay critiques is closely related to two factors: (1) misinterpretation of the essay, and (2) over-exaggeration of a certain standard in the criteria. These factors plague the LLM with hallucinations in following the critiquing prompt.

Based on the findings above, we propose RedHat (**Reduce HallucinaTion**), to reduce the loss of credibility in LLM-generated essay critiques caused by hallucination. We noticed a widely adopted assessment technique for humans (Marton & Säljö, 1976; Chung et al., 2023), that embodies the abstract concept of understanding an essay into the concrete practice of answering questions based on it. We consulted essay experts to identify crucial questions that often need to be clarified for a thorough understanding of an essay. The answers to these questions were beneficial to facilitate the LLM’s comprehension of the essay’s structure and the author’s arguments. This approach ensures that the model recognizes the function of each paragraph within the overall context of the essay, reducing the risk of generating hallucinations caused by overly following evaluation instructions. These question-answer pairs were embedded into the evaluation prompt, providing additional grounding information for the essay during the generation process

We compare accessible alignment techniques including post-pretraining on long contexts and supervised finetuning with RedHat. We suggest the generalizability of RedHat across different base-LLMs, languages and writing genres. We show that alignment would cause more hallucinations on synthetic training data constructed out of human experts’ critiques. This indicates the source of such new hallucinations. In our evaluation setting, RedHat augmented LLM is consistently preferred by human annotators compared to baselines. We utilize the optimized critiques as guidance for essay improvement. In our machine-aided refinement setting, the polished content is generally preferred by human annotators. These showed the potential of our method in relieving hallucination in critiques, thus providing essays with informative and practicable help.

Works	Granularity	Target	Content Len	Generation Format	Generation Len	Open sourced
Ours	Paragraph	Argumentative Writing	5K	Critique	~100	✓
(Tyser et al., 2024)	Whole	CS Conference Paper	>10K	Review	Unlimited	✗
(Liu & Shah, 2023)	Whole	CS Conference Paper	>10K	Review	Unlimited	✓
(Tang et al., 2024)	Whole	ASAP-AES ¹	150-550	Score	Integer	✓
(Noroozi et al., 2023)	Sentence	Argumentative Writing	<800	Feedback	30-50	✗

Table 1: **A Brief Comparison with Previous Work.** We conclude between the scope that AI feedback covers (**Granularity**), assessment content (**Target**), content length (**Writing Len**), AI feedback format, length and whether the works’ dataset, method and evaluation results are publicly available (**Open Sourced**). Our work integrates a fine-grained perspective towards this field.

2 FINE-GRAINED ESSAY CRITIQUE GENERATION

2.1 CREATION OF ESSAYC

Our task is to leverage Large Language Models (LLMs) to automate the generation of critiques for the drafts of undergraduates’ argumentative essays. The goal is to provide students with meaningful feedback that aligns with the detailed feedback provided by the instructors, so as to help the students improve their writing before final submission (**critiquing**). Most previous works studies whether LLM feedback for essay **scoring** (Tang et al., 2024), distinguish the quality of paired abstract (Liu & Shah, 2023) or **trait scoring** in conference review (Tyser et al., 2024). Our evaluation criteria including topic, literature, arguments and structure, language, and norms can be inspected in Appendix C. As summarized in Table 1, this work distinguishes itself from previous work mainly from types and granularity from the LLMs, the assessed target and more importantly, the shift from scoring to critiquing.

In order to provide a unified, even and reproducible test-bench for such task, we curated `ESSAYC` addressing such task and concerns. `ESSAYC` randomly collects undergraduate essays whose topics cover *Environment Science, Biological Science, Software Engineering, Game Industry, Earth, Social Science, Journalism and Communications, Economics, Humanities, Literature Comments*, and so on. Most science, engineering and humanity and social science are covered. Human comments may be incomplete in grammar and organization. We used GLM4-130B to refine and complete their grammar and structure based on the human comments. As a result, 36 essays are randomly picked out of the above process under each field topics. The leftovers are beneficial as training data for supervised fine-tuning.

To enhance quality, we asked the annotators to read through the teacher’s critiques in the paragraph and filter out unqualified ones, such as those with only punctuation marks or subjective comments expressing unrelated feelings. Then we devised a raw critique-quality classifier on GLM-4-9B to auto-filter the leftovers in the train data section. Critique numbers drop from 675 to 395 in the test set, and from 51238 to 31694 in the training set after filtration. The statistics can be found in Table 2 column 2.

2.2 ESSAY CRITIQUE GENERATION TASK DESCRIPTION

Paragraph-level feedback is an effective granularity for improving written content since it can effectively help authors localize the problem while maintaining most contextual information. In contrast, sentence-level feedback can sometimes be less meaningful because not every sentence in a text

	EssayC	English
Essays	36	10
Avg Len	5204.7	42087.3
Critiqued Paras	395	100
Avg Para Len	278.2	1278.4
Avg Tea Cri Len	76.78	/
Pointwise Annotations	5530	200
Annotation Dims	4	4
Pairwise Annotations	1580	100
Avg Cri len	98.53	89.65

Table 2: **Statistics about `ESSAYC`.** **Avg** is short for average. **Para** is short for paragraphs. **Tea** is short for teachers. **Cri** is short for critique. **English** stands for the English subset of conference papers used in experiment.

requires revision, nor are writing problems typically confined to a single sentence. We focus on essay assessment for pedagogical purposes, mainly helping novice writers improve their writing with detailed and informative critiques, and we formulate the task as follows.

Task Formalization: Given an essay \mathcal{E} , a set of instructor evaluation criteria Γ , and the paragraph \mathcal{P} to be critiqued, a model f (e.g., an LLM) is required to generate a critique c for that paragraph:

$$c = f(\mathcal{E}, \Gamma, \mathcal{P}) \quad (1)$$

Objective: The goal of this task is to generate critiques that meet three essential criteria. First, the critique should be free from **hallucination**, and accurately interpret the author’s viewpoints and the factual evidence in the text without introducing inaccuracies. Second, it must be **detailed**, demonstrating a thorough understanding of the paragraph under evaluation, rather than providing vague or superficial feedback. Finally, the critique should be **informative**, offering meaningful insights that assist authors in improving their writing. To maintain clarity and readability, we stipulate that comments must be limited to a maximum of 100 words in our study. Formally, the generation of critique c should maximize the **informativeness** $\mathcal{U}(c)$ and **Ambiguity** $\mathcal{A}(c)$ while minimizing **hallucination** $\mathcal{H}(c)$, subject to the length constraint:

$$\max_c \mathcal{U}(c) - \mathcal{A}(c) - \mathcal{H}(c), \quad \text{subject to } \text{Len}(c) \leq 100 \quad (2)$$

This constrained problem reflects the trade-off between reducing hallucinations and increasing detail, with the ultimate goal of optimizing the informativeness of the feedback provided to the writer.

3 HALLUCINATION IN ESSAY ASSESSMENT

We conducted an empirical study using students’ feedback on LLM-generated critiques. Students give textual judgments over randomly presented critiques to their essays generated by LLMs including ChatGLM3-6b, GLM-4 Plus API (Du et al., 2021; GLM, 2024), and ChatGLM3 finetuned on the instructors’ comments. We found that most prominent issue is hallucination in critiques, as reported to be “the generation of plausible looking yet factually incorrect statements” from (Bang et al., 2023).

As (Maynez et al., 2020) defined *Extrinsic Hallucination* as “ignoring the source material altogether” and *Intrinsic Hallucination* as “misrepresenting information from the document” in summarization task, we found the hallucination in generated essay critiques can be divided mainly into two types as follows:

- **Type I:** Criticizing writing fallacies that do not exist in the essay. As the cases in Table 3 show, LLM emphasize some baseless errors. This type shares commons with the above *Extrinsic Hallucination*.
- **Type II:** Overemphasizing details and reversal of priorities in argumentation structures. The primary concern lies in the tendency to recommend inclusion of excessive details, which consequently undermines the clarity and conciseness of the argument. This diverts from the actual intent proposed in the criteria and the essay. This type is partly related to *Intrinsic Hallucination*.

Under the two main types of hallucinations, we discuss the specific manifestation of them. As listed in Table 3, ignoring the context information, or misunderstanding authors’ perspective originate from **Type I** hallucination. Overreaction and Over-elaboration originate from **Type II** hallucination. These consist of the major aspects for human judgment of critique quality in experiments, such as Table 4 and Appendix H.

We also observed that the occurrence of hallucination varies depends on the position of the critiqued content within the essay. The conclusion part of the essay exhibits the least amount of hallucination, whereas the body sections exhibit the highest incidence. Figure 3 illustrates the human scoring of critique quality, primarily based on the extent of hallucination. Hallucinations are most pronounced in the essay sections ranging from positions between 0% to 30% and around 80% with respect to the total essay length, indicating that LLMs struggle particularly in these areas.

Hallucination	Description	Example Cases	Type
Ignoring context info	Overlooked the contextual information, failing to notice the perspectives and evidence the author has already provided in the surrounding text	(critiquing conclusion part) This section provides background information on "carnivalization"; however, it is somewhat lacking in argumentation and support for the viewpoint.	Type I
Overcorrection at the word or sentence level	Incorrect correction of words or phrases, or overcorrection	In addition, the argument lacks detailed support, and terms such as "universality" and "social attributes" are not thoroughly explained. (no need for explanation)	Type II
Misunderstand the author's perspective	Failed to understand the author's perspective in the evaluated paragraph and its connection to the article	The evaluated paragraph is fairly clear in terms of structure, laying the foundation for subsequent analysis by explaining the 4C marketing theory ... (which is not the author's intention)	Type I
Over-elaboration of non-essential information	Overemphasizing details, reversal of priorities in structures	((already presented evidence) ... further specific evidence is needed to support its conclusion, particularly in clarifying how these strategies hindered technology sharing.	Type II
Citation-related error	Incorrect identification of citations or mistaking the citation for the object of evaluation	The evaluated paragraph has logical issues in its argumentation. The author rejects the definition of health based on "bodily integrity ... (which is the citation part view)	Type I & II
Vague assessment	Copying words from evaluation criteria, with no in-depth revision advice	The argument in this paragraph is relatively clear. However, the supporting evidence appears somewhat limited. And ...	Type II

Table 3: **Hallucination in LLM essay critiques:** the red background texts are the hallucination part and the blue) comments are explanations.

4 REDHAT REDUCES HALLUCINATION IN CRITIQUES

4.1 BACKGROUND

As we have discussed above, the hallucination in essay critique generation mainly originates from LLMs' un-faithfulness to the essay. LLM is not fully aware of the essay contents, leading to the generation problems when it tries to align with the evaluation criteria.

To bridge the gap between LLM's faithfulness of the essay and the following of assessing prompt, we propose RedHat. We noticed the phenomenon in education and psychology (Marton & Säljö, 1976), that breaking the understanding task into question-answering task is able to speed up human's comprehension of long documents. There is an opportunity to ease the evaluation instruction by switching it into series of questions. Then by finding answers from the essay, LLM can reduce its hallucination by more factual information.

4.2 CRITERIA EMBODIMENT

Following the idea above, we embody the evaluation criteria Γ into a list of questions. To ensure the questions' relevance, we prompted GPT-4 to propose questions conditioned on Γ and the essay content. The questions shall cover the essence of Γ , the above process has to be repetitive to be exhaustive. Formally, denote \mathcal{E} as the essay, \mathcal{P} as the critiquing paragraph, p_{question} as prompt for this task, questions are produced in the following iterative process:

$$q_n = \text{Question}_\theta(\Gamma, \mathcal{E}, p_{\text{question}}, q_{1:n-1}) \quad (3)$$

The number of questions n is a hyper-parameter. The above process is not economic in reality, with repetition for each new essay exhibiting redundancy on common questions. We repeat the experiments with different essays and pick a list of common questions as the general solution. The questions are reviewed by human writing experts, listed in the Appendix E.

Another important part of criteria decomposition is seeking answers to those questions in the essay. Fortunately, current LLM techniques all showed compelling performance on DocQA and long

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

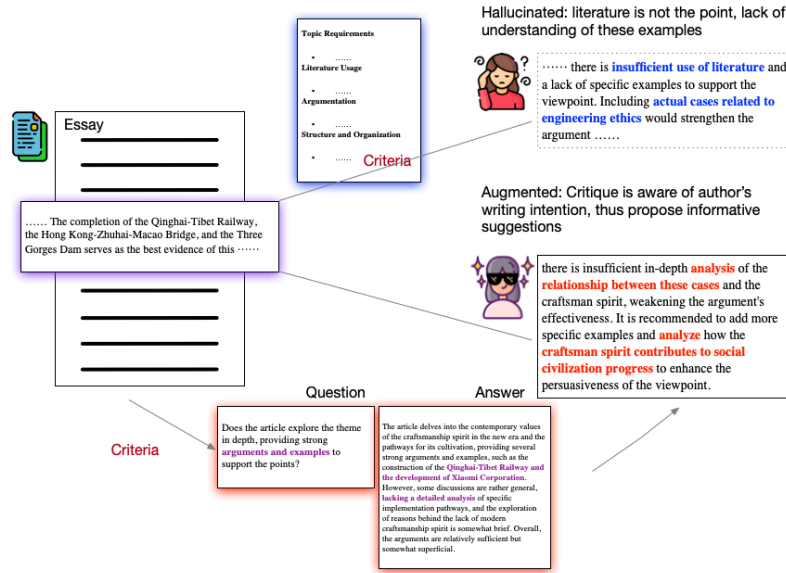


Figure 2: **RedHat Explained.** Converting essay evaluation criteria into a question checklist is beneficial for critique generation. Directly following criteria would ignore the understanding of the essay. RedHats designed to reduce hallucination and ambiguity, and improve critique informativeness. RedHat engages necessary information for understanding the essay in question-answering pairs into the critique generation prompt.

context retrieval (Lewis et al., 2020). The answering process can be streamlined into a separate document question-answering process as formalized below:

$$a_n = \text{DocQA}(q_n, \mathcal{E}) \quad (4)$$

4.3 REORGANIZING THE CRITIQUING PROCESS

We state that the direct insert of the question-answering result pairs into the critiquing process is enough to mediate LLM hallucination in critiquing. One of the most practical methods for combining QA results like RAG (Lewis et al., 2020) calls for a domain-specific retriever for such a situation. Though shared with a promising goal to reduce hallucination (Shuster et al., 2021), our task differs from RAG in contents. RedHat tries to rectify LLM’s inability to grasp and understand the whole essay, while an essay is presented in the original inputs. RAG supplements external knowledge to relieve the ignorance of LLM knowledge, and usually the knowledge does not explicitly exist in original inputs.

We ground the question-answering results into the original critique generation prompt, as the below formula reveals:

$$c_n = \text{LLM}_C(\Gamma, \mathcal{E}, \mathcal{P}, \{q_i, a_i\}_{i=1}^n, p_{\text{critique}}) \quad (5)$$

5 EXPERIMENTS

5.1 EXPERIMENT SETUP

Dataset: We mainly experiment RedHat on EssayC discussed in section 3. To validate RedHat’s effectiveness, we additionally picked a subset from artificial intelligence conference papers as previous works with English-dominated LLMs did. We intentionally chose those papers containing less formulas and illustrations, and more importantly, ensuring the paper authors’ are accessible so that they could judge the quality over the generated critiques. We pick 10 paragraphs with longer word counts from each paper to be critiqued. The statistics of the English subset is listed in Table 2.

	Overall(↑)	Hallu%(↓)	Ambig%(↓)	Info%(↑)
Human Critiques	3.387	47.34	11.65	30.63
Qwen-2-7b-Instruct	3.187	62.53	14.68	11.90
+ 5-shots	3.178	61.01	11.14	12.66
+ RedHat	<u>3.267</u>	62.03	7.59	<u>15.70</u>
+ RedHat-weak	3.323	58.73	<u>8.10</u>	18.73
+ PT	2.777	71.65	24.30	-9.11
+ SFT	2.615	74.43	27.59	-19.24
+ SFT+RedHat	2.636	77.72	22.28	-10.63
glm-4-9b-chat	3.190	65.99	14.97	6.60
+ RedHat	3.327	63.96	9.64	13.20
+ RedHat-weak	<u>3.246</u>	<u>65.99</u>	<u>13.45</u>	6.35
+ PT	3.053	69.04	23.86	-3.04
+ SFT	2.503	79.44	16.50	-31.72
+ SFT+RedHat	2.574	79.44	14.47	-27.92
ChatGPT-4o	2.448	76.92	11.99	-9.99
+ RedHat	3.549	42.96	8.99	23.98

Table 4: **Main experiment** on EssayC (GLM-4 and Qwen-2) and English subset (ChatGPT-4o). All results are judged by human experts. **Hallu** is short for hallucination (0-100%). **Ambig** is short for ambiguity(0-100%). **Info** is short for informativeness(-100-100%). Due to the discriminating ability of human, the three dimensions are evaluated in human detection of fallacies or goodness. Beside the three dimensions, an **Overall score** is given mainly based on hallucination based on the number of deficits detected.

Base LLMs: To validate RedHat’s effectiveness in more LLMs, we select GLM-4-9B-chat (GLM, 2024) and Qwen-2-7B-Instruct (Qwen, 2024) to be studied on EssayC. We select ChatGPT-4o to study the English conference paper subset.

Baselines: Since there are plenty of human written critiques in EssayC construction, supervised-finetuning (SFT) is a direct baseline method. **SFT** tries to show whether it is applicable to avoid hallucinating from direct learning from teachers’ critiques. **Post-pretraining (PT)** tries to clarify our doubt about whether hallucination originates from alienness to long document form reading. **Few-shot** tries to explore the feasibility of bypassing hallucination with in-context examples. Details for few-shot, training and data preparation can be found in Appendix F. Additionally, we also apply RedHat to the supervised finetuned model, to investigate its further application. We are also interested in the quality of answers to the RedHat questions, therefore we compared the LLM self-generated answers in inference (**Weak**) and GPT-4 generated answers.

Metrics: Each of the critique is evaluated with four dimensions: hallucination, ambiguity, informativeness and overall. **Hallucination** (↓ 0 ~ 100%) is evaluated by the true or false detection rate. If one falls to fit the 6 hallucination types mentioned in Table 3, it is marked as true in hallucination. **Ambiguity** (↓ 0 ~ 100%) is calculated whether the critique is ambiguous or not. **Informativeness** (↑ -100 ~ 100%) is calculated whether the critique provided useful improvement advice for polishing. They scored three levels of informativeness: of positive help, of no help, of negative help. **Overall score** (↑ 0 ~ 5) models the task target in Formula 1, and is calculated through: (1) minus 2 per hallucination found; (2) minus 1 for ambiguous; (3) minus 1 for being of negative help or plus 1 for being of positive help (4) truncate into interval 0 to 5.

Evaluator: We mainly refer to trained human graduate teaching assistant scores as evaluation results. The details of our human annotations are listed in the Appendix H. We also conducted pairwise preference annotation with base-LLM and RedHat generated critiques. In this scene, human ranks two critiques into which is better or both is good or bad. Each generated comment is annotated by two graduate teaching assistants. In case of discrepancies, a third graduate teaching assistant makes the final decision. Our overall Inter Annotator Agreement is 0.71 in GLM-4 and Qwen-2 as a whole, ensuring annotation consistency and reducing random interference. We also utilized GPT-4o as auto evaluation method to explore the accessibility of automatic evaluators.

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

5.2 MAIN RESULTS

We showed the results in Table 4, with Qwen-2-7B, GLM-4-9B, ChatGPT-4o. Statistics from the RedHat (Orange background) showed increments in all dimensions compared to base-LLMs. Few-shot benefits the base Qwen2 but is less evident compared to RedHat. However, SFT and PT cause decrement in all dimensions, indicating that direct adjust LLM parameters in the aim of fitting MLE loss are not solutions to hallucination reduction in essay critique generation. Additionally, the reduction of hallucination usually correlates to the reduction of ambiguity and the increment of informativeness. Last but not least, considering RedHat and RedHat-weak, answers provided by GPT (regarded as an DocQA oracle for its high accuracy) or LLM itself all contributed to hallucination reduction and overall improvement.

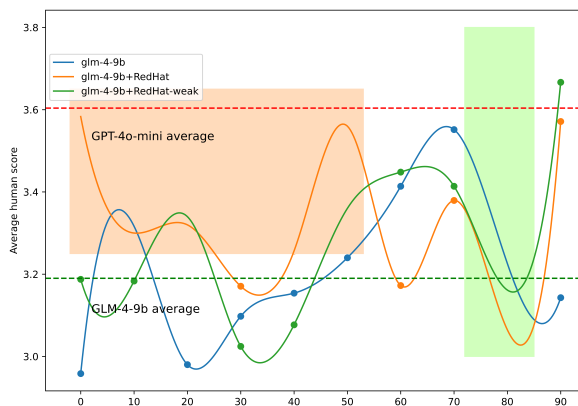


Figure 3: **Distribution** of overall scores with the **position** in the article. The x-axis shows the relative length of the annotated text to the essay. The y-axis shows the average overall score by a human.

Figure 3 depicts a dynamic relation between the critiqued piece and its position in the essay. In the Orange box, RedHat mainly relieved the hallucination in this part. At the 80% point of the article, we observe a notable decline in performance across all methods, as the Green box highlights. We hypothesize that this is where the author begins to conclude their argument, rather than continuing to elaborate further. At this stage, the model tends to overextend by providing more detailed explanations than necessary.

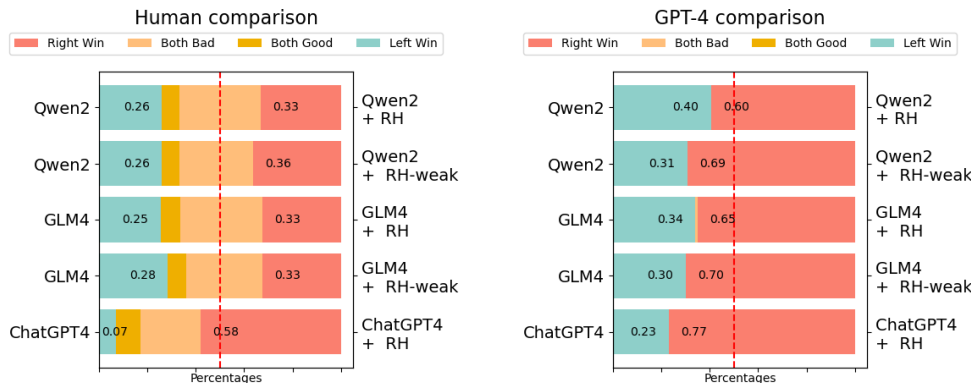


Figure 4: **Results from comparison of critiques generated by baseline with our methods.** Both human experts and GPT-4o judgements are plotted. **RH** is short for RedHat.

The comparison between baseline-LLM and RedHat are shown in Figure 4. In the figure, human are more preferred to critiques generated by RedHat by $\Delta 7.11\%$ in GLM, 10.36% in Qwen. On the one hand, the high tie rates in human judgments result from the number of hallucination types. If one of the six hallucination types is detected from each of the critiques, the pair would be graded as *both is bad*. On the other hand, GPT-4o as comparison evaluator showed low *tie rate*, indicating its potential bias or unawareness of hallucination. Appendix G contains a detailed discussion of them. In conclusion, the overall trend of GPT-4o judgments matches with human judgments and shows the improvements from .

	Question				Answer			
	R-L	B-1	BLEURT	BERTScore	R-L	B-1	BLEURT	BERTScore
Qwen-2-7B-Instruct	10.64	15.62	27.45	71.09	11.16	6.69	21.39	74.50
+RedHat	11.13	18.68	25.24	72.39	12.17	8.40	31.00	76.16
+RedHat-weak	10.92	18.95	26.24	72.50	12.21	8.69	29.41	76.41
+SFT	8.41	4.60	41.21	65.11	6.40	1.88	23.42	66.63
+SFT+RedHat	8.73	5.10	44.56	65.50	6.75	2.03	21.55	67.22
+PT	8.20	4.07	40.82	64.81	6.05	1.67	27.49	66.02
GLM-4-9B-chat	8.89	5.69	42.00	66.09	9.88	3.55	55.28	68.00
+RedHat	9.88	7.51	44.26	67.31	10.18	4.43	55.51	69.72
+RedHat-weak	9.96	7.54	44.11	67.17	10.52	4.60	55.71	69.90
+SFT	7.90	3.75	37.45	64.25	6.75	2.16	50.36	65.04
+SFT+RedHat	8.28	4.28	40.08	64.85	7.25	2.42	51.32	65.96
+PT	8.73	5.77	41.10	65.90	8.91	3.54	54.07	67.89

Table 5: **Overlaps between generated critiques and questions.** **R-L** is short for Rouge score calculated with longest common substrings. **B-1** is short for BLEU score calculated with unigrams.

5.3 HOW QAS HELP REDUCE HALLUCINATION?

To explore how QA results assist in comment generation, we designed the following analytical experiments to investigate the impact of QA accuracy on outcomes and the overlap between the generated critiques and the QA.

Question-Answer Quality: We evaluate the validity of questions by analyzing the similarities between critiques, questions and answers. We calculated word-level overlapping with **ROUGE** (Lin, 2004) and **BLEU** (Papineni et al., 2002), and semantic similarity with **BLEURT** (Sellam et al., 2020), **BERTScore** (Zhang et al., 2019), between the generated critiques and the corresponding questions list, answers list, as shown in Table 5.

We can observe several findings from the results in Table 5. First, with RedHat, similarities between generated critiques and questions do not significantly increase, indicating the questions are not leaking the desired contents to the LLM. Second, similarities gain with answers is observed, especially with Qwen + RedHat, showing that detailed information about the essay is conveyed in the answers by RedHat.

Answer Accuracy Influence on Performance:

In our methodology, we assume that the responses to the questions are correct, which are generated by a perfect long-document question-answering model. We invited human essay evaluation experts to score the correctness of answers for the questions on different essays. GLM-4 show a 14.4% error rate, followed by 7.8% from Qwen-2 and 4.4% ChatGPT-4o. The decrease in error rates corresponds to the gain in point-wise scoring (RedHat-weak rows) of hallucination and pairwise comparison. However, the overall influence of RedHat still outperform baseline-LLMs, suggesting the robustness of our method.

	GLM-4 v.s. GLM-4-RedHat			
	Win	Tie	Lose	Δ
Human	45.74	5.32	48.94	3.20
GPT-4o	19.56	58.67	21.78	2.22

Table 6: **Critique Effect on Essay Polish.** Preference picking between through human and GPT-4o-0815.

5.4 HOW UPDATED CRITIQUES HELP WITH ESSAY POLISH?

Comment generation needs to serve the educational context, meaning it should assist students in improving their essay quality. We designed an experiment to verify that the comments generated by the RedHat can better help improve text quality.

We selected 100 essay samples from the test set and paired the comments generated by glm-4-9b-chat and glm-4-9b-chat-RedHat. We used GPT-4o as a professional text enhancer, prompting it to

	glm-4-9b-chat	glm-4-9b-chat+RedHat
486		
487		
488	Paragraph	The renovation of Juer Hutong involves two key aspects: one is addressing planning issues from the perspective of a historically and culturally significant city, and the other is dealing with residential design from the standpoint of urban design. In terms of historical and cultural city planning, there are two main issues: first, the facade is inconsistent. The upper floors of No. 41 Juer Hutong are white, which appears out of place against the primarily gray hutong landscape. Second, multi-story buildings still disrupt the traditional hutong atmosphere to some extent. Although the three-story buildings in No. 41 maintain the traditional courtyard-style rooftops, preserving the urban texture from an aerial view, from within the hutong, the height of the buildings blocks the view, offering an experience similar to that of high-rise residential complexes.
489		
490		
491		
492		
493		
494		
495	Critique	This paragraph reveals shortcomings in structural organization. It analyzes the issues of the Juer Hutong renovation from the perspective of a historically and culturally significant city but abruptly shifts to the analysis of a specific case (No. 41 Juer Hutong) without proper transition or explanation. It is recommended to briefly outline the main argument after mentioning the planning issues of the historical and culturally significant city, then proceed to the detailed analysis with references, in order to enhance the internal logic and overall coherence.
496		
497		
498		
499		
500		
501		
502		
503		
504		
505		
506	Overall	2.0
507	Reasons	Over-correction at the word or sentence level; of no use
508		
509		
510		
511		
512		
513		
514		
515		
516		
517		
518		
519		
520		
521		
522		
523		
524		
525		
526		
527		
528		
529		
530		
531		
532		
533		
534		
535		
536		
537		
538		
539		

Table 7: Example between the quality of two critiques by glm-4-9b-chat and glm-4-9b-chat + RedHat. Already translated into English from the original Chinese texts.

revise the annotated text based on the comments. Finally, we invited master and doctoral students, as well as teachers from the humanities, to compare the quality of the revised texts. The results are shown in the Table 6. From the results, it is evident that our method, using GPT-4o as the enhancer, can effectively generate in-depth comments and improve text quality. Also, we find that there is a huge gap between human evaluation and GPT-based automatic evaluation. Which is another hint on the unexposed bias inside LLMs.

5.5 CASE STUDY

We pick one case from a certain student essay titled “Why was the renovation of Ju’er Hutong highly praised but not widely promoted”. In the specific paragraph presented, the author made an abrupt twist in conveying from “the two aspects of renovation” to “two main issues of historical and cultural city planning”. Although glm-4-9b-chat points out the structural issues in it, it mistakenly focuses on the revision of the terms. On the other hand, glm-4-9b-chat with RedHat recognized the issue and provided suggestions to outline the main argument. This benefits the coherence both in the paragraph and the essay, showing the augmentation of the LLM’s ability against hallucination caused by local phrases.

6 CONCLUSIONS

In this work, we proposed RedHat, an effective method for reducing hallucination in LLM-generated critiques in essay assessment. RedHat enhanced GLM-9b-chat, Qwen-2-7B-Instruct and ChatGPT-4o by adding an essay-level digest in a question-answering format for the LLM. In our pedagogical application setting, results showed that our method reduced hallucination, ambiguity and improved their informativeness. On the other hand, our generated critiques also greatly helped polish the original essay content. The method is both effective in reducing the hallucination both with EssayC and with the English conference papers.

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

REFERENCES

- M Fadhly Farhy Abbas and Herdi Herdi. Solving the students' problems in writing argumentative essay through collaborative writing strategy. *English Review: Journal of English Education*, 7(1): 105–114, 2018.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Love-
nia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multi-
task, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interac-
tivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and
Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural
Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for
Computational Linguistics (Volume 1: Long Papers)*, pp. 675–718, Nusa Dua, Bali, November
2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.45. URL
<https://aclanthology.org/2023.ijcnlp-main.45>.
- Faeze Brahman, Baolin Peng, Michel Galley, Sudha Rao, Bill Dolan, Snigdha Chaturvedi, and
Jianfeng Gao. Grounded keys-to-text generation: Towards factual open-ended generation. *arXiv
preprint arXiv:2212.01956*, 2022.
- Shuyang Cao and Lu Wang. Cliff: Contrastive learning for improving faithfulness and factuality in
abstractive summarization, 2021. URL <https://arxiv.org/abs/2109.09209>.
- HE Chung, HS Song, YR Cho, YR Oh, AY Kim, HJ Joo, and D Yim. Story comprehension skills of
school-aged children by passage type and question type according to story conditions. *J. Speech-
Lang. Hear. Disord*, 32:49–60, 2023.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard
of wikipedia: Knowledge-powered conversational agents, 2019. URL [https://arxiv.org/
abs/1811.01241](https://arxiv.org/abs/1811.01241).
- Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. Multi-
fact correction in abstractive text summarization, 2020. URL [https://arxiv.org/abs/
2010.02443](https://arxiv.org/abs/2010.02443).
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang.
Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint
arXiv:2103.10360*, 2021.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model
alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.
- Team GLM. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv
preprint arXiv:2406.12793*, 2024.
- Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu.
Iflyea: A chinese essay assessment system with automated rating, review generation, and rec-
ommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational
Linguistics and the 11th International Joint Conference on Natural Language Processing: System
Demonstrations*, pp. 240–248, 2021.
- Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. A probabilistic framework for llm
hallucination detection via belief tree propagation, 2024. URL [https://arxiv.org/abs/
2406.06950](https://arxiv.org/abs/2406.06950).
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large
language models: Principles, taxonomy, challenges, and open questions, 2023. URL <https://arxiv.org/abs/2311.05232>.
- Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization
with semantic-driven cloze reward, 2020. URL <https://arxiv.org/abs/2005.01159>.

594 You-Jin Jong, Yong-Jin Kim, and Ok-Chol Ri. Review of feedback in automated essay scoring.
595 *arXiv preprint arXiv:2307.05553*, 2023.
596

597 Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan
598 Zeng, Yuxiao Dong, Hongning Wang, et al. Critiquellm: Scaling llm-as-critic for effective and
599 explainable evaluation of large language model generation. *arXiv preprint arXiv:2311.18702*,
600 2023.

601 Asheesh Kumar, Tirthankar Ghosal, and Asif Ekbal. A deep neural architecture for decision-aware
602 meta-review generation. pp. 222–225, 2021.
603

604 Anindita Kundu and Denilson Barbosa. Are large language models good essay graders? *arXiv*
605 *preprint arXiv:2409.13120*, 2024.

606 Paraskevas Lagakis and Stavros Demetriadis. Automated essay feedback generation in the learning
607 of writing: A review of the field. pp. 443–453. Springer, 2021.
608

609 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
610 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented genera-
611 tion for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:
612 9459–9474, 2020.

613 Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Knowledge-based
614 review generation by coherence enhanced text planning. pp. 183–192, 2021.

615 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization*
616 *branches out*, pp. 74–81, 2004.
617

618 Ryan Liu and Nihar B. Shah. Reviewergpt? an exploratory study on using large language models
619 for paper reviewing, 2023. URL <https://arxiv.org/abs/2306.00622>.

620 Ference Marton and Roger Säljö. On qualitative differences in learning—ii outcome as a function
621 of the learner’s conception of the task. *british Journal of educational Psychology*, 46(2):115–127,
622 1976.
623

624 Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality
625 in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault
626 (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,
627 pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/
628 2020.acl-main.173. URL <https://aclanthology.org/2020.acl-main.173>.

629 Omid Noroozi, Seyyed Kazem Banihashem, Harm JA Biemans, Mattijs Smits, Mariëtte TW Ver-
630 voort, and Caro-Lynn Verbaan. Design, implementation, and evaluation of an online supported
631 peer feedback module to enhance students’ argumentative essay quality. *Education and Informa-
632 tion Technologies*, 28(10):12757–12784, 2023.

633 OpenAI. Criticgpt, 2024a. URL <https://criticgpt.org/>.

634 OpenAI. GPT-4o, 2024b. URL <https://openai.com/index/hello-gpt-4o/>.

635 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
636 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-
637 low instructions with human feedback. *Advances in neural information processing systems*, 35:
638 27730–27744, 2022.

639 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic
640 evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association*
641 *for Computational Linguistics*, pp. 311–318, 2002.
642

643 Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. Review-
644 llm: Harnessing large language models for personalized review generation. *arXiv preprint*
645 *arXiv:2407.07487*, 2024.
646

647 Team Qwen. Qwen2 technical report, 2024. URL <https://arxiv.org/abs/2407.10671>.

-
- 648 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
649 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*
650 *in Neural Information Processing Systems*, 36, 2024.
- 651 Md Mizanur Rahman, Harold Jan Terano, Md Nafizur Rahman, Aidin Salamzadeh, and Md Saidur
652 Rahaman. Chatgpt and academic research: A review and recommendations based on practical
653 examples. *Rahman, M., Terano, HJR, Rahman, N., Salamzadeh, A., Rahaman, S.(2023). ChatGPT*
654 *and Academic Research: A Review and Recommendations Based on Practical Examples. Journal*
655 *of Education, Management and Development Studies*, 3(1):1–12, 2023.
- 656 Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models.
657 *arXiv preprint arXiv:2309.05922*, 2023.
- 658 Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text gener-
659 ation. *arXiv preprint arXiv:2004.04696*, 2020.
- 660 Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bow-
661 man, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Tim-
662 othy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan,
663 Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023.
664 URL <https://arxiv.org/abs/2310.13548>.
- 665 Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation
666 reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- 667 V Suresh, R Agasthiya, J Ajay, A Amrith Gold, and D Chandru. Ai based automated essay grading
668 system using nlp. In *2023 7th International Conference on Intelligent Computing and Control*
669 *Systems (ICICCS)*, pp. 547–552. IEEE, 2023.
- 670 Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. Harnessing llms for multi-dimensional
671 writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14), 2024.
- 672 Farid Triawan, Hideki Mima, Jeffrey Scott Cross, et al. Automated essay grading of constructive
673 response test responses for mechanical engineering students. In *2023 IEEE International Confer-*
674 *ence on Teaching, Assessment and Learning for Engineering (TALE)*, pp. 1–4. IEEE, 2023.
- 675 Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg,
676 Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te’eni, and Iddo Drori. Ai-driven review
677 systems: Evaluating llms in scalable and bias-aware academic reviews, 2024. URL <https://arxiv.org/abs/2408.10365>.
- 678 Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural
679 machine translation, 2020. URL <https://arxiv.org/abs/2005.03642>.
- 680 Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuan-Jing Huang. Automatic essay scoring in-
681 corporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on*
682 *empirical methods in natural language processing*, pp. 791–797, 2018.
- 683 Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu.
684 Measuring and reducing llm hallucination without gold-standard answers, 2024. URL <https://arxiv.org/abs/2402.10412>.
- 685 Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model
686 hallucinations can snowball, 2023. URL <https://arxiv.org/abs/2305.13534>.
- 687 Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston.
688 Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and
689 Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Com-*
690 *putational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July
691 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL <https://aclanthology.org/P18-1205>.
- 692 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluat-
693 ing text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

702 Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. On large language models'
703 selection bias in multi-choice questions. *arXiv preprint arXiv:2309.03882*, 2023.
704

705 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
706 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
707 chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024a.

708 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyuan Luo. Llamafactory: Unified
709 efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024b.
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

A EXTENDED DISCUSSION OF RELATED WORK

A.1 HALLUCINATION DATASOURCES

The hallucination in natural language generation task is generally defined as the phenomenon that model generated contents contain information that contradicts or is unfaithful to user instructions, additional background context, and/or factual knowledge. Various previous studies have attempted to mitigate hallucination problem in enormous traditional NLG tasks. Due to their discrepancy in task formats, nevertheless, these works define hallucination in rather different ways and aspects and design methodologies tailored to solving these problems in concern. In conversation tasks, (Zhang et al., 2018) proposed PersonaChat dataset aiming to relieve the problem of self-consistency violation in chit-chat. (Dinan et al., 2019) attempts to incorporate external knowledge corpus for more factual knowledge-based dialogue generation. In abstractive summarization tasks (a most related domain of our task), efforts have been paid to alleviate the hallucination problems embodied as generating spans not entailed by the source text. Early works explores methods to improve factuality from source content understanding (Huang et al., 2020), training process (Cao & Wang, 2021) and post-training phase (Dong et al., 2020).

A.2 HALLUCINATION DETECTION

Such method needs external knowledge sources, or reference answer for judging. There are also pioneers who invented reference free methods. FEWL(Wei et al., 2024) weights multiple LLMs answers as proxy of golden answers, which theoretically provided plausibility for judgment. (Hou et al., 2024) utilizes the belief of LLM to check their hallucination problem via decomposing statements into child statements to check in a hierarchical way.

Essay evaluation is both reference and knowledge sparse task, making it hard for quantification on judging. Our method inherits the above ideology by embodying the concept of faithfulness to essay as correctly performs the docOA task.

A.3 HALLUCINATION CAUSES

The causes of hallucination on knowledge-intensive tasks are various. Previous works have focused on those arising from deficiencies in data collection and preprocessing, training, and inferencing phases. In terms of data sources, the emergence of hallucination could be attributed to incorrect or bi-ased data, absence of real-time or proprietary knowledge, or wrong utilization of knowledge (Huang et al., 2023). In the training phase, (Sharma et al., 2023) shows that the training process of RLHF may wrongly lead LLM to produce content that flatter users but disobeys facts. In the inference phase, (Wang & Sennrich, 2020) claims that the discrepancy between the training and inferencing pattern of the AR model could lead to hallucination. (Zhang et al., 2023) finds that hallucinations already generated can mislead LLM to continue producing error statements.

In our work concerning hallucination in essay evaluation tasks, hallucination could be caused by more complex factors. Due to blurred or even seemingly contradictory criteria of judgment, evaluators could generate outputs not consistent with previous contents, likewise tending to generate tangential evaluations.

A.4 ESSAY CRITIQUE GENERATION

Utilizing LLMs to judge and refine human writing has become a buoyant application of recent LLM systems. Several systematic evaluations have been conducted on the capability of LLM to generate critique content for human writings in various scenarios (Tang et al., 2024; Rahman et al., 2023; Lagakis & Demetriadis, 2021; Jong et al., 2023; Lagakis & Demetriadis, 2021). There are also emerging systems built for providing critique generation (Tyser et al., 2024; Peng et al., 2024; Gong et al., 2021; Kumar et al., 2021; Li et al., 2021), manifesting remarkable performance. The primary difference between their work and ours is that their system focuses on generating evaluative comments, whereas we prioritize minimizing hallucinations in the feedback to help writers improve text quality. Also, there lacks of an agreement on a unified testbench.

810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863

B LIMITATIONS

There are two limitations of this work. First, the development of automated hallucination detection techniques for essay critique generation is necessary but requires extensive data labeling, which was constrained by practical budget limitations; thus, we believe it is important to explore synthetic data for the purpose as a focus for future research. Second, exploring how LLM-generated critiques influence LLM-generated essays could deepen our understanding of LLM-based automatic reviews. If successful, it will greatly improve the potential of LLMs for enhancing human-written texts.

C CRITERIA FOR ESSAYS

C.1 CHINESE ARGUMENTATIVE ESSAYS

The essay content studied in our work exhibits four structural and content characteristics. **Topic** is the background and the author’s core perspective to be delivered in the essay. An essay must have a well-defined topic to discuss. Students need to choose a focused, valuable question from a clearly identified discipline that allows for in-depth discussion. **Literature** is the bridge between the essay and the information outside the essay. It is essential to engage in a thorough discussion about existing literature to clearly understand the issue at hand and cite sources appropriately throughout. **Arguments and Structure** refer to the chain of thoughts that depict how arguments are articulated. When presenting arguments, the structure should follow the “problem-argument-reason-evidence” structure to ensure persuasiveness. Arguments should be clear, well-supported, and employ proper logical reasoning, often utilizing both deductive and inductive methods. **Language and Norms:** First-person pronouns should be avoided, and the arguments must be original. When referring to others’ opinions, it is crucial to paraphrase appropriately and refrain from plagiarism.

Student essays are evaluated with respect to the standards uniformly above. We believe such criteria are beneficial for narrowing down possible variances stemming from different assessors’ subjectivity. When evaluating the model’s generated critiques, human labelers can then better focus on hallucinations in the critiques, conditioned on the above criteria.

C.2 ENGLISH CONFERENCE PAPERS

Generally, we refer to ICLR 2025 review instructions for details (<https://iclr.cc/Conferences/2025/ReviewerGuide>). We applied the ICLR reviewer guidelines as evaluation criteria. Since ICLR reviewer guidelines have already contained more than 10 questions in it, we replace the guideline questions with description of the expectation for a good conference paper on those questions. The following is the evaluation criteria version without questions we used.

1. **Thoroughly Read the Paper:** The paper should be read carefully in its entirety. Related works and citations must be reviewed to ensure a comprehensive evaluation. Sufficient time should be allocated for this process.

2. **Key Considerations While Reading:**

2.1 **Objective of the Work:** The paper should have a clear goal, such as addressing a known problem or application, highlighting a new issue, or presenting new theoretical findings. Different objectives should be assessed based on their potential value and impact.

2.2 **Strong Points:** The submission should be clear, technically correct, experimentally rigorous, reproducible, and present novel findings in areas such as theory or algorithms.

2.3 **Weak Points:** Any shortcomings in clarity, technical correctness, rigor, reproducibility, or novelty should be noted.

864 2.4 Open-Mindedness: The value of the paper should be considered
865 from the perspective of the entire ICLR community, even if it may
866 not seem immediately relevant or interesting to individual
867 reviewers.
868

869 3. Evaluating Core Aspects for Recommendation:
870

871 3.1 Problem Definition: The paper should tackle a specific
872 question or problem with clarity.
873

874 3.2 Motivation and Context: The approach should be well-motivated
875 and appropriately contextualized within the literature.
876

877 3.3 Support for Claims: The paper should provide rigorous evidence
878 to support its claims, ensuring results are both correct and
879 scientifically valid.
880

881 3.4 Significance: The work should contribute new, valuable
882 knowledge to the community, whether empirical, theoretical, or
883 practical, regardless of whether it achieves state-of-the-art
884 results.
885

886 4. Initial Review Structure:
887

888 4.1 Summary: Clearly summarize the paper's contributions in a
889 positive and constructive manner.
890

891 4.2 Strengths and Weaknesses: Identify the paper's strong and weak
892 points comprehensively.
893

894 4.3 Initial Recommendation: Provide an initial recommendation
895 (accept or reject) with a clear rationale.
896

897 4.4 Supporting Arguments: Present evidence and arguments that
898 support the recommendation.
899

900 4.5 Clarifying Questions: Include questions for the authors to
901 address ambiguities and provide additional evidence for the
902 assessment.
903

904 4.6 Improvement Suggestions: Offer constructive feedback aimed at
905 improving the paper. Clarify that these suggestions are for
906 improvement and not necessarily decision-critical.
907

908 5. Complete the CoE report:
909

910 5.1 Familiarize yourself with the ICLR Code of Ethics before
911 starting reviews.
912

913 5.2 Assess whether the paper has potential CoE violations and
914 provide explanations if applicable. The CoE report will involve
915 answering these questions as part of the review process.
916

917 6. Active Participation in Discussions:
918

919 Actively engage in the asynchronous discussion phase, where
920 reviewers, authors, and area chairs exchange feedback. Be open to
921 revising your initial recommendation based on new insights or
922 updates to the submission.

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

7. Borderline Paper Discussions:

Participate in virtual meetings organized by Area Chairs (ACs) to discuss borderline cases. Familiarize yourself with feedback from other reviewers to contribute meaningfully to the discussions. Reviewers who fail to attend without emergencies will have their absence noted.

8. Final Recommendation:

Update your review to reflect any new information or revisions during the discussion phase. Clearly articulate the reasoning behind your final recommendation, including what influenced any changes to your assessment.

With the above criteria, the prompts for English conference paper critiquing is structured as follows.

Suppose you are a professional essay polisher for international conference in learning representation. Based on the following review criteria, provide suggestions to improve the appointed paragraph.

```
[review criteria begins]
{criteria}
[review criteria ends]
```

```
[paper begins]
{paper}
[paper ends]
```

```
[paragraph begins]
{paragraph}
[paragraph ends]
```

Now begin your suggestions within 100 words. Your suggestions should aim at pointing out the weaknesses and providing constructive feedback.

D DATA PREPARATION

We collect over 6,000 student essays from our course archives from Fall 2019 to Spring 2024, and randomly select 50 essays to serve as the test set for our evaluation.

Below are our considerations for picking:

Diversity of Topics: The selected 50 essays cover a broad spectrum of topics, including literature, cultural criticism, gaming industry reviews, electric vehicles, technology, and artificial intelligence. These topics were categorized into distinct thematic groups to ensure a diverse representation of subject matter for our testing.

Content and Instructor Feedback: All essays were initial drafts submitted by students for one-on-one feedback from their course instructors. The instructors provided paragraph-level comments, primarily focusing on the writing issues and offering suggestions for improvements.

Ethics and Privacy Considerations: To ensure the ethical use of student data, we obtained approval from the course teaching team for the use of these essays. Additionally, all essays were anonymized

by removing personal identifiers such as student names, IDs, locations, and any other sensitive information. We applied standard anonymization techniques to ensure privacy and a manual review was conducted to confirm that no personal information remained in the dataset.

De-noising: We apply format-revision and correction to the the essays. We also filter out very casual teacher comments like single punctuation like '?', or commenting on unrecognizable pieces.

E ESSAY READING QUESTION LIST

E.1 QUESTIONS FOR CHINESE ARGUMENTATIVE ESSAYS

We list the questions in Table 8 that we collected from the essay writing experts. They are crucial questions in understanding an essay. The picking threshold is the agreement over 15 TAs and instructors.

Setting	Prompt
Question 1	文章是否有一个明确的主题或中心思想? Does the article have a clear theme or central idea?
Question 2	作者在文章的开头是否清晰地提出了主要观点或论点? Does the author clearly present the main point or argument at the beginning of the article?
Question 3	作者是否清晰地表达了他们的观点, 且这些观点在文章各部分中得到一致的支持和阐述? Has the author articulated their views clearly, with consistent support and elaboration throughout the various sections of the article?
Question 4	这些观点是否贯穿全文, 有没有与主题无关的内容? Are these viewpoints consistently maintained throughout the text, or is there unrelated content?
Question 5	文章是否深入探讨了主题, 提供了有力的论据和例子来支持观点? Does the article delve deeply into the subject, providing strong evidence and examples to support its arguments?
Question 6	作者是否展示了对题目有深刻的理解和分析, 还是仅仅停留在表面? Does the author demonstrate a profound understanding and analysis of the topic, or do they merely scratch the surface?
Question 7	作者在文章中是否深入分析了主题, 提供了充分的论据、例子和细节来支持他们的观点? Has the author thoroughly analyzed the theme within the article, offering ample evidence, examples, and details to back up their points?
Question 8	有没有考虑到不同的视角或反驳意见, 并且对这些进行了回应? Have different perspectives or counterarguments been considered, and have these been adequately addressed?
Question 9	文章中的语言是否清晰、准确且具有表现力? Are the statements in the article clear, accurate, and expressive?
Question 10	语言风格是否与文章的目的和受众相匹配? Does the writing style align with the article's purpose and audience?
Question 11	文章是否有明显的语法、拼写或标点错误? 这些错误是否会干扰读者的理解或降低文章的专业性和可信度? Are there noticeable grammatical, spelling, or punctuation errors in the article? Do these errors hinder the reader's understanding or diminish the professionalism and credibility of the piece?
Question 12	文章是否提出了独特的见解或创新的观点, 或者只是重复了常见的观点? Does the article present unique insights or innovative viewpoints, or does it merely reiterate common ideas?
Question 13	有没有引入新颖的例子或视角来讨论主题, 从而使文章在众多类似文章中脱颖而出? Has the author introduced novel examples or perspectives to discuss the theme, allowing the article to stand out among similar works?
Question 14	文章的结构是否合理? Is the structure of the article logical?
Question 15	段落之间的衔接是否流畅? Is there a smooth transition between paragraphs?
Question 16	作者是否按照一个清晰的逻辑顺序来组织他们的论点和证据? Does the author organize their arguments and evidence in a clear logical sequence?
Question 17	每一段是否都有一个明确的中心思想, 并且与前后的段落自然衔接? Does each paragraph have a distinct central idea that connects naturally with the preceding and following paragraphs?
Question 18	段落之间是否有过渡句来帮助读者理解文章的整体结构? Are there transitional sentences between paragraphs to assist the reader in understanding the overall structure of the article?

Table 8: Crucial questions list for EssayC.

E.2 QUESTIONS FOR ENGLISH ARTIFICIAL INTELLIGENCE CONFERENCE PAPERS

We select the question list from the ICLR guideline and list them in Table 9.

Setting	Prompt
Question 1	What is the goal of the paper?
Question 2	Is it to better address a known application or problem, draw attention to a new application or problem, or to introduce and/or explain a new theoretical finding? A combination of these?
Question 3	Is the submission clear, technically correct, experimentally rigorous, reproducible, does it present novel findings (e.g. theoretically, algorithmically, etc.)?
Question 4	What is the specific question and/or problem tackled by the paper?
Question 5	Is the approach well motivated, including being well-placed in the literature?
Question 6	Does the paper support the claims?
Question 7	Are results, whether theoretical or empirical, correct and scientifically rigorous?
Question 8	What is the significance of the work?
Question 9	Does it contribute new knowledge and sufficient value to the community?
Question 10	Does the paper convincingly demonstrate new, relevant, impactful knowledge (including empirical, theoretical, for practitioners, etc.)?
Question 11	What questions would you like answered by the authors to help you clarify your understanding of the paper and provide the additional evidence you need to be confident in your assessment?
Question 12	Is there a potential violation of the Code of Ethics (CoE)?
Question 13	If there is a potential violation, why might there be a potential violation?

Table 9: Crucial questions list For English artificial intelligence conference papers.

F DETAILS FOR TRAINING AND IMPLEMENTATIONS

F.1 SUPERVISED FINETUNING

F.1.1 DATA PREPARING

We conducted our supervised finetuning over augmentation of teachers’ original comments from historical archives apart from the test set. We found original teachers’ comments are informal and fragmented, and directly finetuning on them causes damage to the LLM’s performance. Therefore, we extracted teachers’ comments and deployed a GLM-4-130B for augmentation. The aim of augmentation is to rewrite the semantically low-quality comments into fluent ones, easing for LLM to fit on. The prompt for augmentation can be found in Appendix I.

As a result, we adopt 31,694 polished human paragraph-level critiques as training data, excluded from the `EssayC` testset split mentioned in Section 2. The format of the data is arranged into (evaluation prompt, essay, and target paragraph) as input, and polished paragraph as output. The train and valid set are split based on essays to avoid potential leakage.

F.1.2 TRAINING DETAILS

We split the data into training and validation sets with a 0.95:0.05 ratio. The training epoch is set as 1.15, for from empirical observation, the lowest loss on the validation set falls around epoch 1.1 to 1.2. We adjust learning rate from $\{1e-5, 2e-5, 3e-5, 5e-5, 1e-4\}$, weight decay rate $\{1e-3, 1e-2\}$, betas for Adam $\{[0.9, 0.999], [0.9, 0.9]\}$, scheduler between $\{linear, cosine\}$. Finally, we pick the following config for the least evaluation loss. The training is implemented with LLaMA-Facotory (Zheng et al., 2024b).

- `per_device_train_batch_size`: 1
- `gradient_accumulation_steps`: 2
- `learning_rate`: 1.0e-5
- `weight_decay`: 0.01

-
- 1080 • adam_beta1: 0.9
 - 1081 • adam_beta2: 0.999
 - 1082 • max_grad_norm: 1.0
 - 1083 • num_train_epochs: 1.15
 - 1084 • lr_scheduler_type: cosine
 - 1085 • warmup_ratio: 0.1

1088 F.2 POST TRAINING

1089 F.2.1 DATA PREPARING

1090 As for post-pretraining, we follow two steps: (1) pre-training on Chinese academic papers in the
1091 field of literature, social science and humanities and (2) followed by SFT on the previous data to
1092 ensure the alignment of the critiquing task.

1093 We crawled 128,321 academic papers from the **Chinese National Social Science Base**. The papers
1094 mainly come from journals, such as *Exploration and Free Views*, *Fiction Monthly Shanghai Literature*,
1095 *Beijing Literature Novella Month*, *Science Technology Critiques*, *Tanzhen Technology Review*
1096 and so on. We use OCR with doc2x API (<https://v2.doc2x.noedgeai.com>) and applied the follow-up
1097 data filter and typo fixing with GPT-4 and GLM-4. The above process produces 27,430 pure text
1098 papers of an average around 30,000 Chinese characters. The whole tokens surpassed 1.5 billion.

1101 F.2.2 TRAINING DETAILS

1102 We split the data into training and validation sets with a 0.95:0.05 ratio. The training epoch is set as
1103 6.0, for from empirical observation, the lowest loss on the validation set falls around epoch 5.0 to
1104 7.0.

1105 We adjust learning rate from {1e-5, 2e-5, 3e-5, 5e-5, 1e-4}, weight decay rate {1e-3, 1e-2}, betas
1106 for Adam {[0.9, 0.999], [0.9, 0.9]}, {linear, cosine}. Finally, we pick the following config for the
1107 least evaluation loss. The training is implemented with LLaMA-Facotory (Zheng et al., 2024b).

- 1110 • per_device_train_batch_size: 1
- 1111 • gradient_accumulation_steps: 1
- 1112 • learning_rate: 3.0e-5
- 1113 • weight_decay: 0.01
- 1114 • adam_beta1: 0.9
- 1115 • adam_beta2: 0.999
- 1116 • max_grad_norm: 1.0
- 1117 • lr_scheduler_type: cosine
- 1118 • warmup_ratio: 0.1
- 1119 • bf16: true

1123 F.3 FEW-SHOT IMPLEMENTATION

1124 In our experiment, we experimented with 5-shot structure to test its feasibility to handle the task.
1125 The structure of 5-shot is listed as follows. Note that the beginning of the prompts and the ending of
1126 the prompts remain the same as prompts for baseline-LLM inference in Table 10. The only change
1127 is the insertion of the five examples.

1128 [Evaluation Prompt begins and ends]
1129
1130 Explanation of the criteria.
1131
1132
1133

1134 [Evaluation Criteria begins and ends]
1135
1136 Explanation of the essay.
1137
1138 [Target essay begins and ends]
1139
1140 Explanation of paragraph.
1141
1142 [Target Paragraph begins and ends]
1143
1144 There are five examples for your critiques. You can refer to them
1145 or mimic.
1146
1147 [Example 1 begins]
1148 Target Essay 1
1149 Paragraph 1
1150 Critique 1
1151 [Example 1 ends]
1152
1153 [Example 2 begins]
1154 Target Essay 2
1155 Paragraph 2
1156 Critique 2
1157 [Example 2 ends]
1158
1159 [Example 3 begins]
1160 Target Essay 3
1161 Paragraph 3
1162 Critique 3
1163 [Example 3 ends]
1164
1165 [Example 4 begins]
1166 Target Essay 4
1167 Paragraph 4
1168 Critique 4
1169 [Example 4 ends]
1170
1171 [Example 5 begins]
1172 Target Essay 5
1173 Paragraph 5
1174 Critique 5
1175 [Example 5 ends]
1176
1177 Now, please provide your evaluation. Note that although five
1178 aspects are listed in the evaluation criteria, you only need to
1179 evaluate one dimension based on the
1180 most prominent feature in the paragraph. In your evaluation,
1181 please integrate your notes to grasp the overall framework,
1182 thought process, and logic of the article. Your feedback should
1183 help the student improve the quality of the paragraph. If there
1184 are issues, please point them out and offer suggestions for
1185 improvement. Please respond with your feedback directly without
1186 using formalities, and your evaluation should not exceed 100 word.
1187

1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241

	Mode	Win	Tie	Lose
Pair A	Forward	0.16	0.01	0.82
	Reverse	0.19	0.02	0.80
	Average	0.18	0.02	0.81
Pair B	Forward	0.24	0.01	0.76
	Reverse	0.45	0.01	0.54
	Average	0.34	0.01	0.65
Pair C	Forward	0.68	0.00	0.32
	Reverse	0.80	0.00	0.20
	Average	0.74	0.00	0.26
Pair D	Forward	0.46	0.02	0.52
	Reverse	0.63	0.01	0.36
	Average	0.55	0.01	0.44

Figure 5: **Position Bias** by GPT evaluator. **Forward** shows that critique A is posited far from the end of the prompt while **Reverse** is the opposite case. The scores we reported are the algorithmic average of the two modes.

G POSITION BIAS OF THE EVALUATOR

We observe significant position bias on the pairwise scoring of GPT-4o-mini. As we find in Table 5. We compared four settings from top to down:

- GPT-4o-mini-0718 V.S. GPT-4o-mini-0718-RedHat
- glm-4-9b-chat V.S. glm-4-9b-chat-RedHat
- glm-4-9b-chat V.S. glm-4-9b-chat-RedHat-weak
- glm-4-9b-chat-sft V.S. glm-4-9b-chat-sft-RedHat

As the table showed, GPT showed a significant preference on the item that is near to the end of the prompt (**Reverse**). Previous works in multiple choices (Zheng et al., 2023) also discussed such a phenomenon.

H HUMAN ANNOTATION

H.1 WRITING EXPERT INFORMATION

We hired 15 writing experts for the human annotation stage. They are serving as teaching assistants in the undergraduate writing course. The group primarily consists of graduate students and advanced undergraduates (juniors and seniors), representing a diverse range of academic departments. This interdisciplinary composition ensures the accessibility and relevance of articles across various disciplines and research topics.

H.2 ANNOTATION GUIDELINE TRANSLATED

The following verbatim is our annotation document for human expert annotators. The original document is in Chinese and we translate it into English.

Evaluation Scoring and Annotation Guidelines Document (For critique quality evaluation)

I. Task Description & Objectives

1242 The model is tasked with evaluating human-written paragraphs.
1243 However, due to limitations in the model's capabilities, the
1244 evaluation may produce instances of hallucination and other
1245 issues. The core objective of this task is to assess the overall
1246 quality of the model's comments based on specific dimensions and
1247 to conduct preference scoring and comparison.
1248
1249 In a given essay, multiple comments are provided for a particular
1250 paragraph. Our tasks are as follows:
1251
1252 1. Scoring { Evaluate the quality of comments based on three
1253 dimensions (hallucination, detail, and informativeness) and assign
1254 scores accordingly.
1255
1256 2. Subjective Ranking { Subjectively rank the quality of selected
1257 pairs of comments.
1258
1259 II. Data Field Description
1260
1261 Fixed Fields
1262
1263 - Original Text: The original essay is in document format, which
1264 can be accessed for viewing (annotations from the instructor can
1265 be seen after downloading).
1266
1267 - Original Paragraph: The paragraph being evaluated by the model,
1268 sourced from a specific section of the paper.
1269
1270 - Comments A | H | C | G | I | D | E | F: Eight different model
1271 comments on the original paragraph, including opinions on
1272 structure, content, and format.
1273
1274 Annotation Fields
1275
1276 - Scores for Comments A | H | C | G | I | D | E | F: Score +
1277 corresponding deduction reasons (drop-down list) + 4 sets of
1278 preference comparisons, totaling 20 points.
1279
1280 1. Comment Scoring (8 scores + corresponding multiple-choice
1281 reason boxes):
1282
1283 - Each comment is scored out of a maximum of 5 points, with
1284 deductions made based on error types; specific rules can be
1285 found in STEP 3.
1286
1287 2. Preference Selection (4 single choices):
1288
1289 - A & H Comment Comparison: Preference comparison between
1290 comments A and H.
1291
1292 - C & G Comment Comparison: Preference comparison between
1293 comments C and G.
1294
1295 - C & I Comment Comparison: Preference comparison between
1296 comments C and I.
1297
1298 - D & E Comment Comparison: Preference comparison between
1299 comments D and E.
1300

1296 Preliminary Notes:
1297
1298 The order in the multi-dimensional table from left to right will
1299 follow the sequence:
1300
1301 A, H, C, G, I, D, E, F. Reading from left to right generally does
1302 not require looking back. Note that preference comparisons will be
1303 interspersed throughout.
1304
1305 Scoring is supported by objective dimensions, but these dimensions
1306 may not always correspond directly to the actual quality of the
1307 comments. Preference selection can include subjective factors,
1308 allowing evaluators to choose the most helpful comment between two
1309 options.

1310 III. Specific Scoring Rules (Deduction System)
1311
1312 Scores will be assigned based on the following three dimensions,
1313 with a total score of 5 points, deducting down to 0 points. If the
1314 final score is 5 (full score) and there are no other deduction
1315 points, please check the box for constructive feedback (add 1
1316 point) to provide a reason for the full score.

1317 Dimension 1: Hallucination
1318
1319 - A single hallucination error results in a deduction of 2 points,
1320 two errors lead to a 3-point deduction, and more than two errors
1321 lead to a 4-point deduction. The following rules were previously
1322 detailed in the pre-annotation documentation regarding
1323 hallucination classification:
1324

1325 1. Ignoring Context and Multimodal Information
1326
1327 - Explanation: While the entire paper may not provide this
1328 information, it can be inferred from the feedback given by
1329 human authors whether the model's comments overlook contextual
1330 text information or multimodal information (such as images or
1331 links).
1332
1333 - Typical Context Issues: The author may have presented a
1334 viewpoint or concept in the surrounding context that the model
1335 fails to recognize. This is easily identified with human
1336 feedback, but without it, relevant contextual information must
1337 be judged.
1338
1339 - Multimodal: When the model evaluates articles that combine
1340 text and images, it may fail to effectively parse and integrate
1341 the meanings of the illustrations within the text, leading to
1342 deviations or errors in assessing the relationships between
1343 text and images.

1343 2. Vocabulary, Grammar, and Punctuation Correction Hallucinations
1344 (Overcorrection, Errors)
1345
1346 - The model may provide unnecessary overcorrections regarding
1347 ordinary vocabulary and grammar in the paper|for example,
1348 demanding an explanation for a simple word and providing
1349 examples.

1350 - Corrections made to punctuation and grammar may be incorrect.
1351
1352 - Sentences that lack fluency should be categorized in this
1353 group.
1354

1355 3. Misunderstanding Concepts, Viewpoints, and Logical Structures
1356
1357 - Failure to recognize or understand the main viewpoints,
1358 concepts, and logical structures expressed by the author in the
1359 paragraph, yet proceeding to make corrections.

1360 4. Content Structure - Overcorrection of Non-Key Information
1361
1362 - Requires thorough reading and understanding of the original
1363 paragraph's theme and arguments, assessing whether the model
1364 displays the following issues:
1365

1366 1. Failure to correctly identify the main argument of the
1367 paragraph, resulting in corrections that do not align with
1368 the actual situation.

1369 2. Proposing expansions or corrections that focus on
1370 non-essential information.
1371

1372 3. Errors in summarizing the author's viewpoint.
1373

1374 4. Misunderstanding of the inter-paragraph relationships at
1375 the chapter level.
1376

1377 5. Proposing additions or expansions due to a failure to
1378 differentiate between the author's argumentation logic and
1379 specific concepts.

1380 5. Citation-Related Errors|Content Formatting Comments
1381
1382 - The model may encounter the following hallucinatory issues
1383 regarding citations in the paper:
1384

1385 1. Incorrectly treating a citation as an evaluation target.
1386

1387 2. Failing to recognize or incorrectly identifying citation
1388 information.
1389

1390 3. Guiding errors in citation formatting.
1391

1392 4. Incorrectly assuming that there is citation information
1393 when the original text does not provide any.

1394 Dimension 2: Detail Level
1395
1396 - Deductions of 1 point will be applied for vague evaluations.
1397
1398 - Vague evaluations:
1399
1400 - Comments provided by the model are overly generic and lack
1401 substantial content, making them applicable in any context.
1402
1403 Dimension 3: Constructiveness

-
- 1404 - Constructive feedback adds 1 point; lack of substantial help
1405 results in a deduction of 1 point.
1406
- 1407 - Note: If a comment has no issues and is constructive, it can
1408 still receive a score of 5.
1409
- 1410 - Evaluation lacking helpfulness:
1411
- 1412 - The model's comments do not offer constructive suggestions
1413 that would aid in improving the paper, resulting in a deduction
1414 of 1 point.
- 1415 - It is important to distinguish the constructiveness dimension
1416 from the hallucination dimension: having hallucinations does not
1417 automatically warrant a deduction for constructiveness. If the
1418 AI provides helpful suggestions for improving the paper, then no
1419 deduction is necessary; however, if the AI misleads the reader,
1420 then a deduction should be applied.
1421
- 1422 - Care should be taken to avoid double deductions stemming from
1423 hallucination issues that lead to a lack of helpfulness.
- 1424 - If the comments provided by the model are highly beneficial for
1425 the improvement of the paper, an additional point can be awarded
1426 based on this dimension.
1427
1428
1429

1430 The following is the document for preference picking on polished essays.
1431

1432 I. Task Description & Objectives 1433

1434 People can polish articles of varying quality by following
1435 different types of comments. The core of this task is to score
1436 preferences of the polished text according to specific dimensions
1437 based on the comments.

1438 II. Data Field Description 1439

1440 Fixed Fields 1441

1442 - Original Text: The original essay is in document format, which
1443 can be accessed for viewing (annotations from the instructor can
1444 be seen after downloading).
1445

1446 - Original Paragraph: The paragraph being evaluated by the model,
1447 sourced from a specific section of the paper.
1448

1449 - Polishing A | H | C | G | I | D | E | F: The polished based on
1450 the original text and original paragraph, which are to be
1451 evaluated.

1452 Annotation Fields 1453

1454 1. Preference Selection (3 single choices, win / lose / good tie /
1455 bad tie):
1456

1457 - A & H Polishing Comparison: Preference comparison between
polishings A and H.

-
- 1458
1459 - C & G Polishing Comparison: Preference comparison between
1460 polishings C and G.
1461
1462 - A & J Polishing Comparison: Preference comparison between
1463 polishings A and J.
1464
1465 2. Selection Reasons for Preference (choose from 1-5. Please refer
1466 to Section III for detailed information.)
1467
1468 III. Criteria for Preference Selection
1469
1470 The following describes the characteristics of high-quality
1471 polishing:
1472
1473 1. Adaptability to the Original Text (Original Structure):
1474 - When the polished paragraph is inserted into the article,
1475 does it align with the main flow of the original text, without
1476 deviating in the logical chain?
1477 - The viewpoint of the polished paragraph should not
1478 contradict any content already present in the original text.
1479
1480 2. Language Characteristics:
1481 - Does it comply with the writing norms taught in our writing
1482 courses?
1483
1484 3. Argumentation Process:
1485 - Whether the development of the polished paragraph follows
1486 the required "tree structure", problem, viewpoint, reasons,
1487 and evidence.
1488 - Regardless of the complexity of the viewpoint, whether the
1489 viewpoint information is effectively conveyed to the reader?
1490
1491 4. Literature and Examples:
1492 - Avoid irresponsible citations, incorrect citations,
1493 counterfactual references, or irrelevant citations.
1494
1495 5. Cannot Discern Quality Difference
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

I PROMPTS FOR ALL EXPERIMENT SETTINGS

Setting	Prompt
1512	
1513	
1514	Chinese prompt
1515	你是一位专业的写作老师，你正在教授一位同学论述性写作，同学提交了他的论文草稿，请你根据你制定的以下标准，对论文草稿中的一段话进行点评。
1516	[评价标准开始]
1517	选题要求
1518	* 选题基于明确的研究空白；
1519	* 需具备学理深度、新颖性和研究价值；
1520	* 研究对象和视角应聚焦明确。
1521	文献使用
1522	* 文献检索应充分且符合CRAAP原则（时效性、相关性、权威性、准确性、无利益冲突）；
1523	* 根据具体研究问题平衡使用前沿文献和经典文献；
1524	* 与文献进行充分对话，深入理解选题及方法，合理运用文献进行观点论证。
1525	观点论证
1526	* 观点明确，论据充分；
1527	* 论证逻辑严密，结合演绎与归纳，呈现清晰的树形结构和塔式积木式的证据链。
1528	结构组织
1529	* 内容有清晰的主线，条理分明；
1530	* 概念、框架应前后一致，文内合理呼应；
1531	* 通过标题、段首词句衔接，实现流畅过渡；
1532	* 论点前置，吸引读者，回答“为什么写、写了什么、怎么写”；
1533	* 结尾自然，无不必要的评论或总结。
1534	规范与语言
1535	* 遵守学术规范，论证应为原创，合理引用而非照搬；
1536	* 排版整洁，符合模板要求，引用符合标准格式；
1537	* 语言准确、简洁、理性，避免使用个性化表达。
1538	[评价标准结束]
1539	以下是学生的作文，你需要先阅读并理解其内容。
1540	[学生作文开始]
1541	{essay}
1542	[学生作文结束]
1543	下面是需要你评价的局部段落，请在评价的时候定位其在文章中的位置。
1544	[待评价段落开始]
1545	{paragraph}
1546	[待评价段落结束]
1547	现在请开始你的评价。请注意，评价标准中列出了5点要求，但是你只需要根据待评价文段中最明显的特征，在选题要求、文献使用、观点论证、结构组织、规范与语言中选取一个维度进行评价即可。你的评价旨在帮助学生提升待评价段落的质量，如果待评价段落中存在问题，请将其指出，并且提供改进建议。请直接回复你的评价，不要套话。你的评价不要超过100字。
1548	
1549	English translation
1550	You are a professional writing instructor teaching a student argumentative writing. The student has submitted a draft of their essay. Based on the following criteria, please provide feedback on a specific paragraph from the draft.
1551	[Evaluation Criteria Begins]
1552	Topic Selection Requirements
1553	- The topic should be based on a clear research gap.
1554	- It should have academic depth, novelty, and research value.
1555	- The research object and perspective should be focused and specific.
1556	Use of Literature
1557	- The literature search should be thorough and meet the CRAAP principles (Currency, Relevance, Authority, Accuracy, and Purpose).
1558	- Use a balanced mix of cutting-edge and classic literature, depending on the research question.
1559	- Engage deeply with the literature to understand the topic and methodology, and use it appropriately to support arguments.
1560	Argumentation
1561	- The argument should be clear, with sufficient evidence.
1562	- The logic should be rigorous, combining deduction and induction, presenting a clear tree structure and a block-by-block evidence chain.
1563	- Structure and Organization
1564	The content should follow a clear main line, well-structured.
1565	- Concepts and frameworks should be consistent and logically referenced throughout the essay.
1566	- Smooth transitions should be achieved through appropriate use of headings and introductory phrases.
1567	- The thesis should be upfront, engaging the reader, answering "why write, what is written, how it is written."
1568	- The conclusion should be natural, without unnecessary commentary or summary.
1569	Academic Norms and Language
1570	- Follow academic standards; arguments should be original, with proper citation instead of paraphrasing or copying.
1571	- The formatting should be neat, adhering to template requirements, and citations should follow the correct format.
1572	- The language should be accurate, concise, and objective, avoiding personal expressions.
1573	[Evaluation Criteria Ends]
1574	Here is the student's essay; please read and understand its content first.
1575	[Student Essay Begins]
1576	{essay}
1577	[Student Essay Ends]
1578	Below is the specific paragraph to be evaluated. When providing feedback, please identify its position in the essay.
1579	[Paragraph to be Evaluated Begins]
1580	{paragraph}
1581	[Paragraph to be Evaluated Ends]
1582	Now, please provide your evaluation. Note that although five aspects are listed in the evaluation criteria, you only need to evaluate one dimension based on the most prominent feature in the paragraph. Your feedback should help the student improve the quality of the paragraph. If there are issues, please point them out and offer suggestions for improvement. Please respond with your feedback directly without using formalities, and your evaluation should not exceed 100 words.

Table 10: Prompt for critiquing essays directly based on essays and paragraphs with **zero-shot base LLMs** in Chinese.

Setting	Prompt
1566	
1567	
1568	Chinese prompt
1569	你是一位专业的写作老师，你正在教授一位同学论述性写作，同学提交了他的论文草稿，请你根据你制定的以下标准，对论文草稿中的一段话进行点评。
1570	[评价标准开始]
1571	选题要求
1572	* 选题基于明确的研究空白；
1573	* 需具备学理深度、新颖性和研究价值；
1574	* 研究对象和视角应聚焦明确。
1575	文献使用
1576	* 文献检索应充分且符合CRAAP原则（时效性、相关性、权威性、准确性、无利益冲突）；
1577	* 根据具体研究问题平衡使用前沿文献和经典文献；
1578	* 与文献进行充分对话，深入理解选题及方法，合理运用文献进行观点论证。
1579	观点论证
1580	* 观点明确，论据充分；
1581	* 论证逻辑严密，结合演绎与归纳，呈现清晰的树形结构和塔式积木式的证据链。
1582	结构组织
1583	* 内容有清晰的主线，条理分明；
1584	* 概念、框架应前后一致，文内合理呼应；
1585	* 通过标题、段首词句衔接，实现流畅过渡；
1586	* 论点前置，吸引读者，回答“为什么写、写了什么、怎么写”；
1587	* 结尾自然，无不必要的评论或总结。规范与语言
1588	* 遵守学术规范，论证应为原创，合理引用而非照搬；
1589	* 排版整洁，符合模板要求，引用符合标准格式；
1590	* 语言准确、简洁、理性，避免使用个人化表达。
1591	[评价标准结束]
1592	以下是学生的作文，你需要先阅读并理解其内容。
1593	[学生作文开始]
1594	{essay}
1595	[学生作文结束]
1596	为了更好地理解这篇文章的内容，你带着几个主要问题阅读文章，并且得到了对文章的总体认识。下面是你的问题和相应回答：
1597	[你的笔记开始]
1598	{qa_notes}
1599	[你的笔记结束]
1600	下面是需要你评价的局部段落，请在评价的时候定位其在文章中的位置。
1601	[待评价段落开始]
1602	{paragraph}
1603	[待评价段落结束]
1604	
1605	现在请开始你的评价。请注意，评价标准中列出了5点要求，但是你只需要根据待评价文段中最明显的特征，在选题要求、文献使用、观点论证、结构组织、规范与语言中选取一个维度进行评价即可。在你的评价过程中，请你结合你的笔记，把握文章的整体框架、思路、逻辑。你的评价旨在帮助学生提升待评价段落的质量，如果待评价段落中存在问题，请将其指出，并且提供改进建议。请直接回复你的评价，不要套话。你的评价不要超过100字。
1606	
1607	
1608	
1609	
1610	
1611	English translation
1612	You are a professional writing instructor teaching a student argumentative writing. The student has submitted a draft of their essay. Based on the following criteria, please provide feedback on a specific paragraph from the draft.
1613	[Evaluation Criteria Begins]
1614	Topic Selection Requirements
1615	- The topic should be based on a clear research gap.
1616	- It should have academic depth, novelty, and research value.
1617	- The research object and perspective should be focused and specific.
1618	Use of Literature
1619	- The literature search should be thorough and meet the CRAAP principles (Currency, Relevance, Authority, Accuracy, and Purpose).
1620	- Use a balanced mix of cutting-edge and classic literature, depending on the research question.
1621	- Engage deeply with the literature to understand the topic and methodology, and use it appropriately to support arguments.
1622	Argumentation
1623	- The argument should be clear, with sufficient evidence.
1624	- The logic should be rigorous, combining deduction and induction, presenting a clear tree structure and a block-by-block evidence chain.
1625	- Structure and Organization
1626	The content should follow a clear main line, well-structured.
1627	- Concepts and frameworks should be consistent and logically referenced throughout the essay.
1628	- Smooth transitions should be achieved through appropriate use of headings and introductory phrases.
1629	- The thesis should be upfront, engaging the reader, answering “why write, what is written, how it is written.”
1630	- The conclusion should be natural, without unnecessary commentary or summary.
1631	Academic Norms and Language
1632	- Follow academic standards; arguments should be original, with proper citation instead of paraphrasing or copying.
1633	- The formatting should be neat, adhering to template requirements, and citations should follow the correct format.
1634	- The language should be accurate, concise, and objective, avoiding personal expressions.
1635	[Evaluation Criteria Ends]
1636	Here is the student’s essay; please read and understand its content first.
1637	[Student Essay Begins]
1638	{essay}
1639	[Student Essay Ends]
1640	To better understand the content of this article, you read it with several key questions in mind, gaining an overall insight into the work. Below are your questions and their corresponding answers:
1641	[Your notes begin]
1642	{qa_notes}
1643	[Your notes end]
1644	Below is the specific paragraph to be evaluated. When providing feedback, please identify its position in the essay.
1645	[Paragraph to be Evaluated Begins]
1646	{paragraph}
1647	[Paragraph to be Evaluated Ends]
1648	Now, please provide your evaluation. Note that although five aspects are listed in the evaluation criteria, you only need to evaluate one dimension based on the most prominent feature in the paragraph. In your evaluation, please integrate your notes to grasp the overall framework, thought process, and logic of the article. Your feedback should help the student improve the quality of the paragraph. If there are issues, please point them out and offer suggestions for improvement. Please respond with your feedback directly without using formalities, and your evaluation should not exceed 100 word.
1649	

Table 11: Prompt for critiquing essays using RedHat. It reserve a field ‘qa_notes’ for the question-answering results.

1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673

Setting	Prompt
Chinese prompt	<p>请扮演一位专业的论文评审专家，在读懂论文的基础上判断两条评语的质量。请先阅读以下的长文。</p> <p>[文章开始] essay [文章结束]</p> <p>下面是一对评语与评语对应的段落，请你判断哪一条评语质量更好。评语的质量好坏主要体现在：</p> <ol style="list-style-type: none"> 1. 评语是否理解了段落的内容，特别是在作者的写作意图基础上展开的； 2. 评语是否足够深入，特别是对改进段落质量有帮助 3. 评语是否避免了幻觉，例如事实错误、逻辑错误、过分解读、不理解文本本身等； <p>[段落开始] {paragraph} [段落结束]</p> <p>[评语1开始] {comment1} [评语1结束]</p> <p>[评语2开始] {comment2} [评语2结束]</p> <p>你需要给出四种判断之一：1更好；2更好；1和2一样好；1和2一样差。请以两对括号包括你的回答，例如“[[1更好]]”，或者“[[2更好]]”等。请直接给出你的判断。</p>
English translation	<p>Please act as a professional paper reviewer and assess the quality of two comments based on your understanding of the paper. First, read the following text.</p> <p>[Article begins] essay [Article ends]</p> <p>Below is a pair of comments along with the corresponding paragraph. Please determine which comment has better quality. The quality of the comments is primarily evaluated based on:</p> <ol style="list-style-type: none"> 1. Whether the comment accurately understands the content of the paragraph, especially in relation to the author's intent; 2. Whether the comment is sufficiently in-depth, particularly in its usefulness for improving the quality of the paragraph; 3. Whether the comment avoids misconceptions, such as factual errors, logical fallacies, over-interpretation, or misinterpretation of the text itself. <p>[Paragraph begins] {paragraph} [Paragraph ends]</p> <p>[Comment 1 begins] {comment1} [Comment 1 ends]</p> <p>[Comment 2 begins] {comment2} [Comment 2 ends]</p> <p>You need to provide one of four judgments: 1 is better; 2 is better; 1 and 2 are equally good; 1 and 2 are equally poor. Please enclose your answer in double brackets, such as "[[1 is better]]" or "[[2 is better]]". Please provide your judgment directly.</p>

Table 12: Prompt for GPT-4o-mini-0718 to **compare the critique quality** between the polished texts with different critiques.

1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

Setting	Prompt
Chinese prompt	请扮演一位专业的论文润色专家，在读懂论文的基础上，结合你的阅读笔记，以及一个评阅意见，对一个段落进行润色、优化。
	[文章开始] {essay} [文章结束]
	[阅读笔记开始] {notes} [阅读笔记结束]
	[段落开始] {paragraph} [段落结束]
	[评语开始] {critique} [评语结束]
	依据评语，请直接写出你改进后的段落，不需要其他说明。
English translation	Please act as a professional paper editing expert. After fully understanding the paper, and based on your reading notes as well as a critique, revise and optimize a given paragraph.
	[Start of Essay] {essay} [End of Essay]
	[Start of Reading Notes] {notes} [End of Reading Notes]
	[Start of Paragraph] {paragraph} [End of Paragraph]
	[Start of Critique] {critique} [End of Critique]
	Based on the critique, please directly write the improved paragraph without any further explanation.

Table 13: Prompts for instructing GPT-4o-0806 to **polish the original text** based on the critique.

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

Setting	Prompt
Chinese prompt	<p>请扮演一位专业的论文评审专家，读懂论文的基础上比较一段话不同润色结果的的质量。请先阅读以下的长文。</p> <p>[文章开始] {essay} [文章结束]</p> <p>下面是一对润色结果与原文段落，请你判断哪一条润色结果质量更好。润色结果的质量好坏主要体现在： 1. 放入原文中的位置是否通顺、合理，在文章片段上表意连贯、思路清晰； 2. 本身没有明显可见的事实错误、论述不当； 3. 使得文章结构更加完整，不明显偏离原文主线。</p> <p>[原文开始] {paragraph} [原文结束]</p> <p>[润色结果1开始] {polish1} [润色结果1结束]</p> <p>[润色结果2开始] {polish2} [润色结果2结束]</p> <p>你需要给出四种判断之一：1更好；2更好；1和2一样好；1和2一样差。请以两对括号包括你的回答，例如“[[1更好]]”，或者“[[2更好]]”等。请直接给出你的判断。</p>
English translation	<p>Please act as a professional paper reviewer and assess the quality of different revisions of a paragraph based on your understanding of the paper. First, read the following text.</p> <p>[Article begins] {essay} [Article ends]</p> <p>Below is a pair of revisions compared to the original paragraph. Please determine which revision has better quality. The quality of the revisions is primarily evaluated based on: 1. Whether the placement of the revisions within the original text is coherent and reasonable, maintaining a clear flow of ideas; 2. The absence of obvious factual errors or inappropriate arguments; 3. The enhancement of the overall structure of the paper without significantly deviating from the original main line.</p> <p>[Paragraph begins] {paragraph} [Paragraph ends]</p> <p>[Revision result 1 begins] {polish1} [Revision result 1 ends]</p> <p>[Revision result 2 begins] {polish2} [Revision result 2 ends]</p> <p>You need to provide one of four judgments: 1 is better; 2 is better; 1 and 2 are equally good; 1 and 2 are equally poor. Please enclose your answer in double brackets, such as “[[1 is better]]” or “[[2 is better]]”. Please provide your judgment directly.</p>

Table 14: Prompt for GPT-4o-0806 to **compare the quality between the polished texts** with different critiques.