XtraGPT : LLMs FOR HUMAN-AI COLLABORATION ON CONTROLLABLE SCIENTIFIC PAPER REFINEMENT

This work does not advocate using LLMs to replace human creativity or ethical standards in research.

Anonymous authors
Paper under double-blind review

ABSTRACT

The increasing volume of scientific publications highlights the growing need for high-quality academic writing. However, while groundbreaking ideas are often present, many papers fail to meet academic writing standards. Unlike open-ended applications of large language models (LLMs) in research, which delegate creative tasks to AI, we emphasize a human-centered approach where researchers provide ideas and drafts while LLMs strictly follow user instructions for refinement. The XtraGPT training and evaluation processes, and models will be **open-sourced**.

We propose **XtraGPT**, LLMs designed to assist authors by delivering instruction-driven, context-aware revisions that (1) adhere to user instructions, (2) align with general academic writing standards, and (3) are consistent with the whole paper. Leveraging a dataset of 7,040 ICLR 24 papers and 140,080 question-answer pairs, XtraGPT enhances specific sections without compromising the paper's integrity. Experimental results show XtraGPT-7B surpass similar size models and is competitive with GPT-4o-mini in providing high-quality, context-aware refinements. We also found that scaling up model parameters provides limited improvement for the difficulty of paper scoring. Modifying six sections with XtraGPT can improve the paper's rating according to the predictor.

By prioritizing controllability in the task of paper refinement, XtraGPT empowers researchers to focus on innovation while relying on the system to handle the demands of academic writing with context understanding and adherence to academic standards and user instructions.

1 INTRODUCTION

Hercules, the hero who achieved great deeds through the persecution of Hera, took on the twelve labors commanded by Eurystheus. Each task was not only a challenge to his courage and wisdom, but also a journey of growth and self-overcoming, much like the process of constantly refining and improving a paragraph in the creation of a paper.

The rapid growth of scientific publications has created an increasing demand for high-quality academic writing tools. While many papers present groundbreaking ideas, their overall clarity, coherence, and writing quality often fall short of meeting academic standards. Large language models (LLMs) have shown remarkable capabilities in general-purpose text generation and question-answering (Dubey et al., 2024; Liu et al., 2024a; Achiam et al., 2023; Bai et al., 2023), but their potential in assisting fine-grained and controllable paper refinement remains underexplored.

Existing research in applying LLMs to academic writing focuses on four main areas: (1) *full-paper generation without user intervention*, which lacks fine-grained refinement or user-instruction alignment (Shao et al., 2024; Jiang et al., 2024; Anonymous, 2024a; Asai et al., 2024; Schmidgall et al., 2025); (2) *idea generation*, where LLMs propose research ideas directly, raising ethical concerns over authorship and creative responsibility (Baek et al., 2024; Ghafarollahi & Buehler, 2024; Li et al., 2024a; Si et al., 2024; Gu et al., 2024); (3) *paper review and domain-specific question-answering*, with little emphasis on directly improving the overall quality of writing (D'Arcy et al., 2024; Liang et al., 2024; Lu et al., 2024; Anonymous, 2024b; Asai et al., 2024; Chen et al., 2024b; Lála et al., 2023; Song et al., 2024; Lin et al., 2024); and (4) *polishing tools*, such as AI-assisted writing apps,

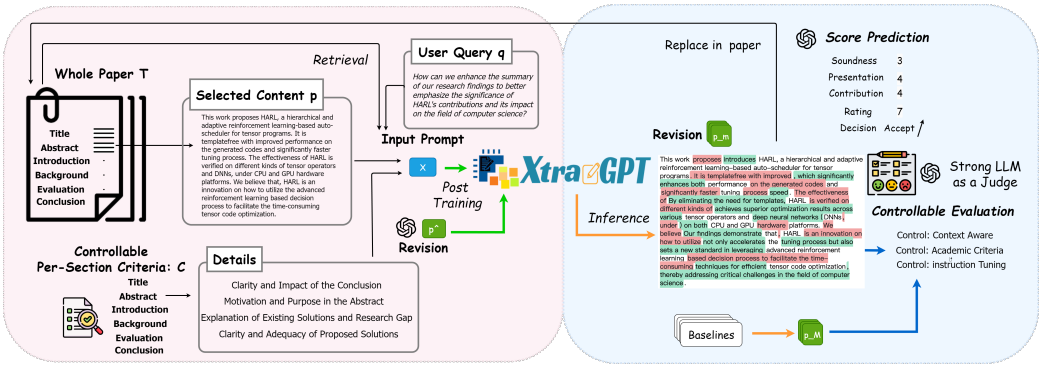


Figure 1: The schematic process of XtraGPT. The post training and evaluation processes ensure controllable section-level fine-grained paper refinement.

which focus on superficial improvements without understanding the context of the paper or the academic writing standards (CoWriter, 2025; van Zeeland, 2023).

The controllable generation of large models is emphasized (Ge et al., 2025). Despite advancements in leveraging llms for academic writing, none of the existing approaches address the necessity of a fine-grained, controllable paper refinement process. Scientific writing requires more than polished language — it demands a deep understanding of a paper's ideas and general academic writing standards to provide meaningful revisions. Human authors must retain control over the creative process, generating ideas and drafts while using tools that enhance their work in targeted and specific ways. Our approach, akin to a powerful code editor (Cursor, 2024), empowers authors to select specific sections of their paper for refinement and receive reliable, context-aware suggested revisions that align with their intent while preserving the core ideas of the work.

Developing such a paper refinement framework faces significant challenges: 1) *The lack of paired training data for instruction and refinement.* Most available datasets contain completed papers, offering little insight into pre- and post-improvement versions, making it difficult to model the refinement process effectively. 2) *The limited capability of LLMs to refine based on the context of an entire paper.* Current models struggle to deeply understand a paper's global structure, interconnected ideas, and nuanced context, which are essential for meaningful revisions. *The absence of a comprehensive summarization of general academic writing standards.* While academic writing relies on clarity, coherence, sound argumentation, and adherence to specific formats, an LLM-understandable and representative summarization of these qualities is lacking, complicating the evaluation and refinement of papers.

To address these challenges, we propose **XtraGPT** (Figure 1), a framework for controllable, fine-grained paper refinement that bridges the gap between human creativity and AI-assisted writing. XtraGPT enables authors to improve their drafts with minimal writing overhead by understanding the structure and context of scientific papers, providing section-level revisions tailored to user instructions, and maintaining the core ideas while enhancing clarity, coherence, and adherence to academic standards. By addressing the three major challenges mentioned above, XtraGPT sets a new direction in scientific writing tools.

Our key contributions include:

(1) **XtraQA**: a dataset of 7,040 research papers enriched with over 140,000 question-answer pairs for section-grained paper refinement by extracting high-quality data tailored for academic papers;

(2) **XtraGPT**: the first LLMs explicitly designed for fine-grained, controllable paper refinement, with its controllability demonstrated across three dimensions: contextual refinement, section-level fine-grained standards, and instruction-following ability;

(3) We qualify the effect of controllable paper refinement through a testset of 7000 question-answer pairs. Through detailed experiments, we provide several insights on paper refinement and scoring.

(4) XtraGPT adheres to the principle that human creatively generates ideas, while AI minimizes the mechanical burden of writing.

Table 1: Comparison of current full-paper AI generators on quality issues, full-paper In-Context Learning (ICL), Retrieval-Augmented Generation (RAG) or not, evaluation, controllability and whether include Human-Computer Interaction (HCI) or generate paper from scratch. Controllability refers to a generative system's ability to adapt to user needs, provide fine-grained control over content, and allow dynamic interaction and adjustment during the generation process.

Full-Paper AI Generator	Quality Issues	ICL	Evaluation	Control	HCI
PaperRobot (Wang et al., 2019)	Not LLM based, bad QA quality	✔	Human	✔	✔
August et al.(August et al., 2022)	Only definition	✗	Human	✗	✗
STORM (Shao et al., 2024)	Biased & Factual Hallucination	✔	Automatic & Human	✗	✗
CO-STORM (Jiang et al., 2024)	Lack of Consistency	✔	Automatic & Human	✗	✔
CycleResearcher (Anonymous, 2024a)	Reward Hacking & Outdated	✗	Automatic & Human	✗	✗
OpenScholar (Asai et al., 2024)	Disorganized Logic & Overlength	✔	Automatic & Human	✗	✗
Agent Lab (Schmidgall et al., 2025)	Structure Rigidity	✗	Automatic & Human	✗	✗
(Ifargan et al., 2024)	from scratch	✗	Automatic	✔	✔
AI Scientist (Lu et al., 2024)	no control idea	✗	Automatic	✗	✗
XtraGPT	**Controllable Refinement** 😎	✔	Automatic & Human	✔	✔

Experiments demonstrate that XtraGPT delivers context-aware, high-quality revisions that strictly follow user instructions, with comparable results to GPT-4o-mini (OpenAI et al., 2024) but using only 7 billion parameters. Additionally, we found that LLMs struggle with paper scoring even with scaling, and it is hard to achieve a rating MAE below 1.5. Moreover, modifying six sections with XtraGPT can enhance the paper's rating according to the predictor.

Our philosophy is that when the core idea of paper is strong enough, we assist authors in producing smooth and polished writing, turning the writing process into a minimal overhead task.

2 Background and Motivation

2.1 Limitations of Current AI Paper Generation Methods

Why Can't Existing LLMs Excel in Paper Generation? Table 1 provides a comparative analysis of existing AI paper generators, which are designed to generate entire papers. These systems struggle to simultaneously ensure comprehensive retrieval, fine-grained control, and effective human-computer interaction. Additionally, they often exhibit various quality issues, making it challenging to achieve **Controllable AI Paper Refinement**.

Why do we need Section-Level Fine-Grained Control? The success of o1 (OpenAI, 2024) and r1 (DeepSeek-AI et al., 2025) models lies in their ability to explore problems from multiple perspectives with fine-grained reasoning.

Academic Papers are inherently complex and sparse, making them difficult for models and even human experts to learn and evaluate effectively. As demonstrated in our experiments in Section 5.2, even models with substantial capacity find it challenging to directly learn and comprehend entire papers. Fortunately, the success of o1(OpenAI, 2024) and r1 (DeepSeek-AI et al., 2025) models lies in their ability to explore problems from multiple perspectives with fine-grained reasoning. To address this, we target on the paper into 6 **sections** (title, abstract, introduction, background, evaluation, conclusion) and establish **fine-grained criteria** for selected content across six key paragraphs. This approach is akin to *Hercules* completing 12 meticulous tasks. However, a significant challenge remains: **the lack of labeled data** or paired examples showing pre- and post-improvement versions of papers, leaving us with only the final versions for evaluation.

What criteria influence the overall evaluation of a research paper? According to the review form provided in NeurIPS 2024, **full-paper level** evalaution of paper contains soundness, presentation and contribution, which is positively correlated with the acceptance rate. However, to effectively evaluate the improvements made to a paper, we need to move beyond **section-level** assessments.

Therefore, we collaborated with experts in the AI field to develop fine-grained principles which is specifically tailored for AI papers. The criteria are detailed in Figure 10, 11, 13, 14, 15, 12, and 16.

2.2 Motivation and Philosophy

The motivation of this paper is to assist researchers in improving the quality of their AI-generated academic writing while ensuring that the necessary academic standards and language precision are maintained. The goal is to address the issues highlighted in Table 1, where existing full-paper AI generators struggle with quality control and limitations in adaptability.

We argue that revising a paper is a meticulous process, akin to *the heroism of Hercules*, overcoming numerous obstacles and challenges. Paper writing should not be done without **proper quality control**, and that directly generating a full paper without refinement is not the best approach. We believe that when the core idea of a paper is solid, using the controllable refinement capabilities of XtraGPT can help authors quickly **revise the writing**, turning the rewriting process into a task that involves minimal effort. This allows authors to leverage AI's capabilities for rapid revisions while maintaining the integrity of their work.

3 Data Collection

As outlined in Section 2.1, a significant challenge in implementing HCI paper refinement is the lack of high-quality QA data. To address this, we introduce **XtraQA**, the first dataset designed to assist authors in improving their paragraphs. XtraQA comprises 140,800 QA pairs, with 133,800 pairs allocated for training.

We initially collected all 6,994 PDFs (after excluding 64 excessively long PDFs from a total of 7,042) in ICLR 2024 and converted them into parsable markdown format. For each article, we generated 20 criteria-based questions for user-selected paragraphs, resulting in 140,800 QA pairs. Subsequently, we employed GPT-4o-mini (only 1.7% hallucination rate from (Hong et al., 2024)) to generate improved versions of these paragraphs, denoted as $\hat{p}$. As analyzed in Table 9, human annotators confirmed that the dataset is sufficiently robust to compete with GPT-o1-mini.

To evaluate the paragraph improvement capabilities of LLMs, we randomly sampled 5% of the dataset (350 papers, comprising 7,000 QA pairs) to create the QA benchmark. We used length-controlled win rate (Dubois et al., 2024) to establish an LLM arena, with XtraGPT as the anchor, avoiding the widely criticized ROUGE and BLEU metrics for direct answer evaluation. In the QA benchmark, the distribution of QAs across six sections—title, abstract, introduction, background, evaluation, and conclusion—is 2:4:6:2:3:2, corresponding to 700, 1,400, 2,100, 700, 1,050, and 700 QAs, respectively. Throughout the data collection process, we maintained stringent quality control measures to ensure the reliability of the dataset.

3.1 Submission Data Analysis

We analyzed all ICLR 2024 submissions, finding that 64.71% received replies, with 82.4% of those reaching a final decision and a 36.3% acceptance rate among the filtered and parsed PDFs. The full-paper score distribution is shown in Figure 2, and paper length distribution is shown in Section C, with a maximum of around 16,384 tokens.

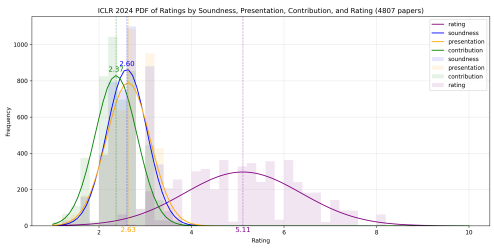


Figure 2: ICLR 2024 PDF ratings of full-paper criterias

3.2 XtraQA Data Generation

The XtraQA dataset was constructed using parsed text T from ICLR 2024 submissions. Queries q were generated based on predefined criteria c. Leveraging the full text T, the GPT-4o-mini model was employed to produce the revised paragraph $\hat{p}$.

The dataset for supervised fine-tuning (SFT) is defined as:

$$D_{SFT} = \{(q, T, p, \hat{p})\}$$

3.2.1 CONTROLLABILITY ASSURANCE OF $q, p, \hat{p}$

The queries q were guided by section-level improvement criteria (Table 2), ensuring the enhancement of selected paragraph quality, with detailed criteria illustrated in Figures 10 through 12. The improved paragraphs $\hat{p}$ were generated using a carefully designed prompt (6). The quality of both the supervised fine-tuning dataset D_{SFT} and the enhanced paragraphs $\hat{p}$ was rigorously validated by human annotators, as listed in Table 9 and detailed in Table 8.

Domain Style-Invariant Assumption To address whether varying writing styles and development speeds across different domains affect the overall evaluation of articles, we engaged three human evaluators from the fields of inference speedup, graph and FPGA. They used a specialized interface (Figure 17,18) to annotate the data on criteria definition 16 with different colors. Our findings suggest that our models do not need to be designed separately for different domains and perform consistently well across them.

Table 2: Evaluation Criteria for Title, Abstract, Introduction, Background, Evaluation, and Conclusion

Aspect	Comments
Title	Consistency and Alignment of Title with Content Conciseness and Clarity of Title
Abstract	Clarity and Impact of the Conclusion Motivation and Purpose in the Abstract Explanation of Existing Solutions and Research Gap Clarity and Adequacy of Proposed Solutions
Introduction	Strength and Clarity of Motivation in the Introduction Review of Existing Approaches in Introduction Audience Alignment and Appropriateness Clarity and Visibility of Contributions Clarity and Specificity of Problem Definition Integration of State-of-the-Art in Problem Framing
Background	Contextual Relevance and Clarity of Background Coverage of Key Preliminary Concepts Clarity and Consistency of Terminology
Evaluation	Experimental Setup Clarity and Reproducibility Depth and Clarity of Figures and Tables Analysis Experimental Support for Main Innovations
Conclusion	Broader Impact and Future Directions Clarity and Impact of Key Innovations and Findings

Table 3: Human evaluation on improvement acceptance rates before and after paragraph. We asked 3 human evaluators based on 5, 3, 5 papers, with about 100, 60, 100 questions scored from 1-5. The **Aggregated** column averages the results of the 3 human evaluators.

QA Controllability Assurance	**Judge 1**	**Judge 2**	**Judge 3**	**Aggregated**
GPT-4o-Mini				
-Instruction Following	76.6	74.3	77.4	76.1
-Criteria Following	73.6	74.7	76.6	74.9
-In-Context Ability	59.4	66.3	72.8	66.2
-Agree revision?	49.2	61.7	71.6	60.8
—				
GPT-o1-mini				
-Instruction Following	76.4	79.0	74.8	76.7
-Criteria Following	74.0	77.0	74.2	75.1
-In-Context Ability	62.0	68.0	73.6	67.9
-Agree revision?	56.0	66.3	72.8	65.0

3.2.2 QUALITY ASSURANCE OF T

We analyzed 6,994 after-filtered PDFs from ICLR 2024 using the deep learning-based academic paper parser nougat (Blecher et al., 2023), which converts PDFs into tokenizable markdown text T. To ensure the quality of T, we chose nougat, as its performance outperforms rule-based tools like pymupdf (PyMuPDF, 2024), and Marker (Paruchuri, 2024) according to (Li et al., 2024d), which were used in the ICLR analyses by Lu et al. (2024), Anonymous (2024b), and Anonymous (2024a) (using MagicDoc (Magic-Doc, 2024)). Afterward, we will perform post-processing, keeping only the content before the service and removing the acknowledge information, so that T can remain length within 16384.

4 XtraGPT

4.1 EXPERIMENT SETTINGS

We post train D_{SFT} on Qwen-2.5-1.8B-Instruct and Qwen-2.5-7B-Instruct to get XtraGPT using the LLaMA-Factory (Zheng et al., 2024) framework on a setup consisting of 4 NVIDIA H100 GPUs, with 80 GB of memory and inference on XtraQABench using the vLLM (Kwon et al., 2023) framework on a setup consisting of 1 NVIDIA A100 GPUs, with 80 GB of memory. The computing environment was configured with CUDA 12.2 and cuDNN 9.1 for optimized deep learning performance. Detailed parameters are listed in Table 10.

4.2 CONTROLLABLE INSTRUCTION POST TRAINING

We train on the XtraQA training set $D_{SFT} = \{(q, T, p, \hat{p})\}$, which consists of 133,800 QA pairs. This fine-tuning process enhances the base model's **controllability** to follow instructions, ensure criteria compliance, and maintain contextual understanding.

4.3 HOW TO EVALUATE THE MODEL'S QUALITY ON CONTROLLABLE PAPER REFINEMENT?

Previous papers including (Anonymous, 2024b) have used simple metrics like ROUGE to evaluate the full-text generation capabilities in the AI research process. However, such metrics only ensure adherence to raw text-level answers and fail to provide controllability over specific capabilities. To address this, we adopt the concept of Length Controlled Win Rate (Dubois et al., 2024) against XtraGPT as anchor and utilize alpaca_eval_gpt4_turbo_fn as a judge (Zheng et al., 2023), which reaches 68.1% human agreement according to (Tatsu-lab, 2023), with a slight modification focused on evaluating the controllability of outputs using the instruction 8, 7. Length Controlled Win Rate calculates how many times XtraGPT (m) can win against baseline models (M).

Why LLM as a controllable paper revision judger? Previous work demonstrates high alignment with automated reviewers (Lu et al., 2024), while (Schmidgall et al., 2025) say still needs both.

In this study, we chose GPT-4o-mini to generate data instead of Openai o1 or Deepseek R1 (DeepSeek-AI et al., 2025) because our task does not rely on complex reasoning or deep thinking, planning but rather focuses on the ability to handle long-context understanding. GPT-4o-mini excels in this area, effectively understanding and generating coherent paragraphs. For sequence-level tasks like paragraph rewriting, the evaluation criteria are often subjective. Using LLM as an evaluator of the generated content provides consistent quality feedback, a method proven effective in the development of InstructGPT and ChatGPT. Therefore, LLM as a judge is well-suited for quality evaluation in our scenario, avoiding the high cost of manual annotation while providing efficient feedback.

The reliability of Instruction and the bias of LLM paper revision We identified several issues with LLM-based paper revisions: overuse of certain GPT-style words like "comprehensive" to exaggerate the paper's impact, making superficial changes, and a tendency to generate excessively long revision segments. To address these issues, we meticulously designed 6 to avoid such problems during generation, along with 8 and 7 to emphasize these concerns during evaluation.

While win rate effectively reflects the relative performance of our model compared to others in paragraph rewriting tasks, it becomes unreliable due to length bias, as shown in Table 11. This issue has also been noted in other studies. To mitigate this, we employ length-controlled win rates (Dubois et al., 2024), which adjust for the bias introduced by varying lengths of generated content, ensuring a fairer evaluation, supported by methods from AlpacaEval (Tatsu-lab, 2023).

Definition of LC win rate Let b represent the baseline model and xtra represent our model. Let θ denote the prediction value. The length-controlled win rate is defined as:

$$q_{\theta,\phi,\psi}(y = m \mid z_m, z_M, x) := \text{logistic}\left(\text{model} + \text{length}\right)$$

where the model term is $\theta_m - \theta_M$ and the length term is $\phi_{M,b} \cdot \tanh\left(\frac{\text{len}(z_m) - \text{len}(z_M)}{\text{std}(\text{len}(z_m) - \text{len}(z_M))}\right)$

We omit the instruction difficulty term as we focus solely on the improvement effect. The length-controlled win rate is then calculated as:

$$\text{winrate}^{LC}(m, M) = 100 \cdot \mathbb{E}_x \left[q_{\theta,\phi,\psi}(y = m \mid z_m, z_M, x)\right]$$

When lengths are inconsistent, the length term adjusts the final estimated value to account for this bias. This approach ensures a fair comparison by controlling for length variations.

Table 4: Length-controlled (LC) win rates of various models against XtraGPT (*anchor*) across different evaluation categories. Models are ranked in descending order based on their weighted LC win rates. The judge is modified alpaca_eval_gpt4_turbo_fn (Prompt 7).

Models	Title (2)	Abstract (4)	Introduction (6)	Background (3)	Evaluation (3)	Conclusion (2)	Overall↑
Qwen2-72B-Instruct	35.93	63.43	67.63	71.18	77.26	64.77	65.31
Deepseek-v3-671B	52.36	57.33	62.08	56.26	73.23	50.00	59.75
GPT-4o-Mini	50.97	50.00	52.29	58.49	51.35	45.96	51.86
XtraGPT (*anchor*↑)							
Qwen-2.5-7B-Instruct	50.41	47.11	43.71	46.56	46.05	49.79	46.44
Qwen-QWQ-32B-Preview	37.83	34.57	32.13	40.58	30.04	32.91	34.22
Llama-3.1-8B-Instruct	34.78	30.64	35.31	41.60	40.29	18.36	33.51
Qwen2.5-1.5B-Instruct	36.07	30.87	25.80	21.34	24.18	24.27	26.80
GPT-3.5-Turbo	25.73	23.99	21.52	23.16	30.97	17.39	24.24
Llama-3.2-3B-Instruct	19.93	6.45	9.35	3.80	8.26	4.64	8.73

5 Analasis

5.1 Q1: How about the win rate of baselines against XtraGPT?

Based on the data in Table 4, the XtraGPT model demonstrates superior performance compared to several baseline models, especially in categories like Introduction, Abstract, and Background, surpassing many open-source 7B models.

While Qwen2-72B-Instruct leads in some categories, such as Introduction and Evaluation, XtraGPT remains highly competitive across all dimensions, showing reliability and strength in various tasks. Compared to Deepseek-v3-671B (59.75%) and GPT-4o-Mini (51.86%), XtraGPT's overall win rate of 65.31% surpasses both, highlighting its advantage in comprehensive performance. Moreover, XtraGPT significantly outperforms smaller models like Qwen-2.5-7B-Instruct (46.44%) and Llama-3.1-8B-Instruct (33.51%), demonstrating its consistent strength across multiple evaluation criteria.

In conclusion, XtraGPT not only leads among open-source 7B models but also shows strong competitive capabilities against larger models like GPT-4o-Mini in paper revision tasks.

The table 5 shows the quality ratings of XtraGPT by humans as judges, and combined with Table 4's LLM as a judge, it highlights XtraGPT's outstanding performance.

Table 5: Expert evaluation of XtraGPT results.

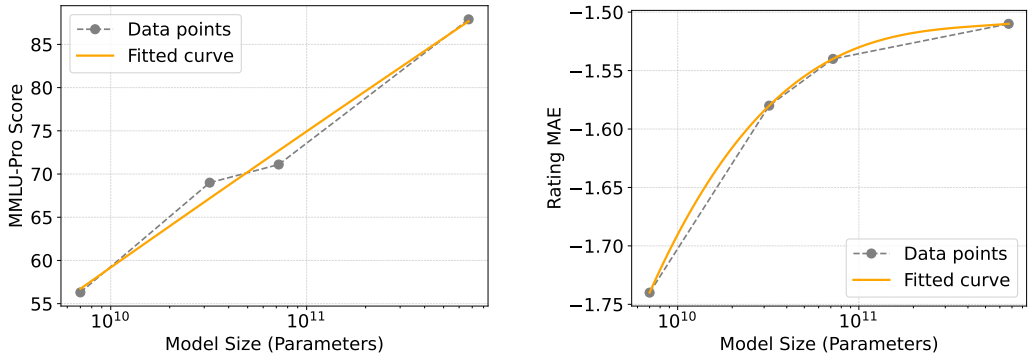
XtraGPT	Judge 1	Judge 2	Judge 3	Aggregated
Instruction Following	65.0	79.7	81.8	75.5
Criteria Following	66.8	74.0	81.8	74.2
In-Context Ability	55.8	68.0	81.2	68.3
Agree revision?	49.2	64.5	80.2	64.6

5.2 Q2: Can LLMs Score Full Papers? Scaling Laws of LLMs as Review Judges

In the context of academic paper evaluation, the only available human expert review labels at full-paper granularity come from OpenReview. Unfortunately, due to the high cost and inherent biases of human reviews—evidenced by a standard deviation of 1.26 in reviewer ratings for each paper in 2024— it is impractical to invite expert reviewers for every benchmarking scenario that requires full-paper scoring.

To address this limitation, several studies (Lu et al., 2024; Anonymous, 2024a;b) have explored the use of LLMs for predicting full-paper scores. A key question remains: are LLMs suited to be a reliable reviewer? To investigate this, we follow the approach of Lu et al. (2024), applying NeurIPS review guidelines and few-shot examples to assess our test set.

As shown in Figure 3, scaling up model parameters is significantly more challenging for paper scoring compared to MMLU-Pro. We can infer that the bottleneck in the paper scoring task **cannot be simply**



(a) MMLU-Pro scores on multi-task understanding across different # of model parameters

(b) MAE scores on paper scoring across different # of model parameters

Figure 3: Scaling trends of Qwen-2.5-7B/32B/72B/Max-Instruct performance. (a) MMLU-Pro scores stably improve with model size. Scaling is effective on multi-task understanding. (b) In the paper scoring task, the rating MAE struggles to go below 1.5. As the model size increases, the reduction in MAE becomes smaller, indicating that scaling offers limited performance improvement.

solved by scaling the model. LLMs struggle with paper scoring, which is already quite challenging even for human experts (1.16 rating MAE per paper according to Anonymous (2024a)).

Table 6: Different LLMs as scorer on the evaluation of 404 ICLR2024 papers. MAE measures the Mean Absolute Error of the avg rating of human and llm reviewers.

Criteria	Soundness			Presentation			Contribution			Rating ↑				Accept Rate
	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	MAE ↓	Acc.
Human Reviewer on ICLR 2024 original papers														
Human Reviewer	1.00	4.00	2.60±0.43	1.00	4.00	2.63±0.48	1.00	4.00	2.37±0.45	1.00	9.00	5.11±1.26	-	36.3%
Human R. - Rejected paper	1.00	3.75	2.46±0.40	1.00	4.00	2.50±0.46	1.00	3.50	2.20±0.39	1.00	7.50	4.53±0.97	-	-
Deepseek-V3 Reviewer as Scorer on ICLR 2024 original papers														
Deepseek-V3-64k	2.00	6.00	3.38±0.70	2.00	6.00	3.42±0.72	2.00	6.00	3.71±0.62	4.00	8.00	6.49±0.95	1.45	75.0%
Qwen (Above 7B) as Scorer on ICLR 2024 original papers														
Qwen-2.5-7B-Instruct	2.00	6.00	3.00±0.37	2.00	6.00	2.78±0.59	2.00	6.00	3.03±0.36	6.00	7.00	6.92±0.27	1.74	95.2%
Qwen-2.5-32B-Instruct	2.00	4.00	2.90±0.33	2.00	4.00	2.69±0.49	2.00	4.00	2.97±0.34	4.00	8.00	6.73±0.63	1.58	81.9%
Qwen-2.5-72B-Instruct	2.00	6.00	3.00±0.32	2.00	6.00	2.66±0.58	2.00	6.00	3.28±0.53	3.00	8.00	6.67±0.73	1.54	78.4%
Qwen-2.5-Max-LongContext	2.00	4.00	3.03±0.29	2.00	4.00	2.70±0.59	2.00	4.00	3.09±0.31	3.00	8.00	6.68±0.73	1.51	74.3%
GPT-4o as Scorer on ICLR 2024 original papers														
GPT-4o	1.00	4.00	3.05±0.53	2.00	4.00	3.13±0.61	2.00	4.00	3.43±0.54	3.00	8.00	6.79±0.85	1.60	88.0%

We list different LLMs as paper scorer in Table 6. We can derive that LLMs as reviewers tend to give higher accept rate than human. The rating MAE of Deepseek-V3 is competitive against other models, which reaches near the 1.16 human bias of a specific data (Anonymous (2024a)). Based on these findings, and thanks to the success of DeepSeek (Liu et al., 2024a; DeepSeek-AI et al., 2025), we adopt DeepSeek-V3 (Liu et al., 2024a) as our scoring model to evaluate the quality of our own models.

5.3 Q3: Does XtraLLM revision impact the full paper?

To evaluate the quality of papers at the full paper level, we randomly sampled a passage from each section (6 passages total) and brought it back for evaluation. Our findings show that even modifying just a single passage per section leads to an increase in the overall score. Additionally, the acceptance rate after revision also improved. This highlights the effectiveness of XtraGPT in enhancing paper quality.

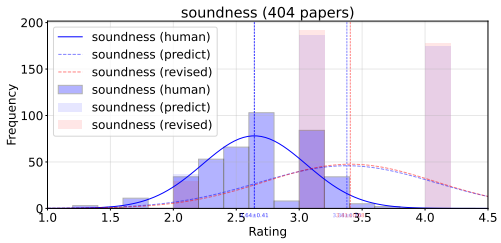
We test on 404 paper which have rating in the QA benchmark. From the data presented in Table 7, it can be observed that after replacement, the AcceptRate improved from 75.0% to 75.8% (same LLM Deepseek-V3 as reviewer). Additionally, the average scores for soundness, presentation, and contribution all saw increases of 0.03, 0.02, and 0.02 respectively. The overall rating improved by

0.02. These results demonstrate that modifying just six sections of the paper can significantly enhance the quality of the full paper, showcasing the effectiveness of XtraGPT in improving the overall paper.

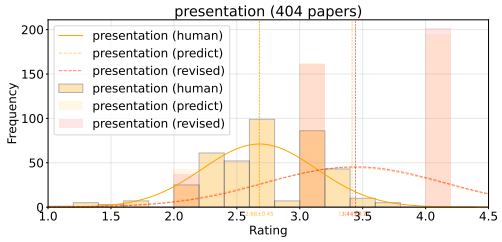
According to the overrating bias from LLM scorer in Section 5.2, we calculate the bias from Deepseek-V3 against human by the sum of the differences in ratings (without absolute values). The average bias of Deepseek-V3 is 1.03 across all 404 papers. After minusing the 1.03 bias caused by Deepseek-V3 as the scorer, the revision from XtraGPT against origin paper is 0.03 (before minus 1.05). It means after revision, the overall rating improves 0.02. Detailed results of different paper-level criteria are shown in Figure 4.

Table 7: Evaluation of 404 XtraGPT improved papers. We replace the revised paragraph back to the paper to re-evaluate the paper score by paper score classifier. The predictor is DeepSeek-V3.

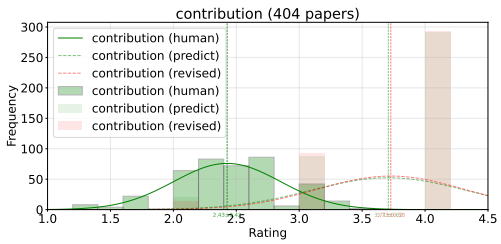
Criteria	Soundness			Presentation			Contribution			Rating ↑				Accept Rate
	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	Min.	Max.	Avg.	MAE ↓	Acc. R.↑
Human Reviewer on ICLR 2024 original papers														
Human Reviewer	1.00	4.00	2.60±0.43	1.00	4.00	2.63±0.48	1.00	4.00	2.37±0.45	1.00	9.00	5.11±1.26	-	36.3%
Human R. - Rejected paper	1.00	3.75	2.46±0.40	1.00	4.00	2.50±0.46	1.00	3.50	2.20±0.39	1.00	7.50	4.53±0.97	-	-
Deepseek-V3 Reviewer as Scorer on ICLR 2024 original papers														
Deepseek-V3-64k	2.00	6.00	3.38±0.70	2.00	6.00	3.42±0.72	2.00	6.00	3.71±0.62	4.00	8.00	6.49±0.95	**1.45**	75.0%
Deepseek-V3 Reviewer as Scorer on XtraGPT improved papers														
XtraGPT (ours)	2.00	6.00	**3.41±0.68**	2.00	6.00	**3.44±0.71**	2.00	6.00	**3.73±0.58**	4.00	8.00	6.51±0.91	1.47	**75.8%**



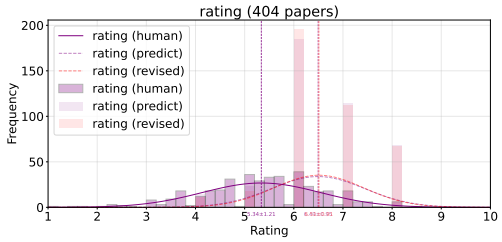
(a) Soundness score predict



(b) Presentation score predict



(c) Contribution score predict



(d) Rating predict

Figure 4: Comparison of human-assigned, predicted, and revised ratings for four evaluation criteria across 404 papers. (a) Soundness, (b) Presentation, (c) Contribution, and (d) Overall Rating. Each subfigure shows the distribution of ratings along with fitted density curves

6 Conclusion

In this work, we introduce XtraGPT, a series of LLM designed to help researchers refine scientific papers through fine-grained, controllable revisions. Leveraging a dataset of 7,040 ICLR 2024 papers and over 140,000 question-answer pairs, XtraGPT provides context-aware, instruction-driven revisions that improve clarity, coherence, and adherence to academic standards while preserving the integrity of the original work. Our experiments show that XtraGPT-7B outperforms similarly sized models and competes with larger models like GPT-4o-mini in delivering high-quality refinements. We also find that scaling model parameters beyond 100 billion is necessary for LLMs to achieve human-level paper scoring capabilities.

XtraGPT's ability to enhance sections such as the introduction, abstract, and conclusion positively impacts paper quality and acceptance rates. By enabling human-AI collaboration, XtraGPT allows researchers to maintain creative control while reducing the mechanical burden of writing, ensuring high academic standards without deviating from core ideas. We believe XtraGPT offers a significant step forward, providing researchers with a practical solution to produce high-quality papers with minimal effort.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yiming Ai, Zhiwei He, Ziyin Zhang, Wenhong Zhu, Hongkun Hao, Kai Yu, Lingjun Chen, and Rui Wang. Is cognition and action consistent or not: Investigating large language model's personality. *arXiv preprint arXiv:2402.14679*, 2024.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. Litsearch: A retrieval benchmark for scientific literature search. *arXiv preprint arXiv:2407.18940*, 2024.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, pp. 413–425, 2024.

Anonymous. Cycleresearcher: Improving automated research via automated review. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=bjcsVLoHYs. under review.

Anonymous. Peer review as a multi-turn and long-context dialogue with role-based interactions: Benchmarking large language models. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=uV3Gdoq2ez. under review.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Openscholar: Synthesizing scientific literature with retrieval-augmented lms, 2024. URL https://arxiv.org/abs/2411.14199.

Joshua Ashkinaze, Julia Mendelsohn, Li Qiwei, Ceren Budak, and Eric Gilbert. How ai ideas affect the creativity, diversity, and evolution of human ideas: Evidence from a large, dynamic experiment. *arXiv preprint arXiv:2401.13481*, 2024.

Tal August, Katharina Reinecke, and Noah A. Smith. Generating scientific definitions with controllable complexity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8298–8317, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.569. URL https://aclanthology.org/2022.acl-long.569/.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents, 2023. URL https://arxiv.org/abs/2308.13418.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089*, 2024a.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Mingchen Zhuge, Jürgen Schmidhuber, Xin Gao, and Xiangliang Zhang. Scholarchemqa: Unveiling the power of language models in chemical research question answering. *arXiv preprint arXiv:2407.16931*, 2024b.

CoWriter. Cowriter - your ai platform for speeding up creative writing, 2025. URL https://cowriter.org/login.

Cursor. Cursor - the ai code editor, 2024. URL https://www.cursor.com/.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2024. URL https://arxiv.org/abs/2404.04475.

Yubin Ge, Neeraja Kirtane, Hao Peng, and Dilek Hakkani-Tür. Llms are vulnerable to malicious prompts disguised as scientific language, 2025. URL https://arxiv.org/abs/2501.14073.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer,

Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. Llms can realize combinatorial creativity: generating creative ideas via llms for scientific research, 2024. URL https://arxiv.org/abs/2412.14141.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904, 2024. doi: 10.48550/ARXIV.2404.05904. URL https://doi.org/10.48550/arXiv.2404.05904.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. *arXiv preprint arXiv:2402.10811*, 2024.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous llm-driven research from data to human-verifiable research papers, 2024. URL https://arxiv.org/abs/2404.17605.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J. Semnani, and Monica S. Lam. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations, 2024. URL https://arxiv.org/abs/2408.15232.

Hao Kang and Chenyan Xiong. Researcharena: Benchmarking llms' ability to collect and organize information as research agents. *arXiv preprint arXiv:2406.10291*, 2024.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. PlatoLM: Teaching LLMs in multi-round dialogue via a user simulator. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7841–7863, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.424. URL https://aclanthology.org/2024.acl-long.424/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Bruce W Lee, Yeongheon Lee, and Hyunsoo Cho. Language models show stable value orientations across diverse role-plays. *arXiv preprint arXiv:2408.09049*, 2024.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*, 2024a.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. Scilitllm: How to adapt llms for scientific literature understanding. *arXiv preprint arXiv:2408.15545*, 2024b.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. CryptoTrade: A reflective LLM-based agent to guide zero-shot cryptocurrency trading. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1094–1106, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.63. URL https://aclanthology.org/2024.emnlp-main.63/.

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, and Le Sun. Readoc: A unified benchmark for realistic document structured extraction, 2024d. URL https://arxiv.org/abs/2409.05137.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024.

Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. Agentsims: An open-source sandbox for large language model evaluation. *arXiv preprint arXiv:2308.04026*, 2023.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science. *arXiv preprint arXiv:2407.00466*, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–25, 2024b.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pp. 1–11, 2024.

Magic-Doc. Magic-doc: A toolkit that converts multiple file types to markdown. `https://github.com/InternLM/magic-doc`, 2024.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

OpenAI. Introducing openai o1, 2024. URL `https://openai.com/index/introducing-openai-o1-preview/`. Accessed: 2025-01-30.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng,

Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

Vik Paruchuri. Marker: Convert pdf to markdown + json quickly with high accuracy, 2024. URL https://github.com/VikParuchuri/marker.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. Citeme: Can language models accurately cite scientific claims? *arXiv preprint arXiv:2407.12861*, 2024.

PyMuPDF. Pymupdf. https://github.com/pymupdf/PyMuPDF, 2024.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. Assisting in writing wikipedia-like articles from scratch with large language models, 2024. URL https://arxiv.org/abs/2402.14207.

Zhengliang Shi, Shen Gao, Zhen Zhang, Xiuying Chen, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. Towards a unified framework for reference retrieval and related work generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5785–5799, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.385. URL https://aclanthology.org/2023.findings-emnlp.385/.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Xiaoshuai Song, Muxi Diao, Guanting Dong, Zhengyang Wang, Yujia Fu, Runqi Qiao, Zhexu Wang, Dayuan Fu, Huangxuan Wu, Bin Liang, et al. Cs-bench: A comprehensive benchmark for large language models towards computer science mastery. *arXiv preprint arXiv:2406.08587*, 2024.

Weiwei Sun, Zhengliang Shi, Wu Jiu Long, Lingyong Yan, Xinyu Ma, Yiding Liu, Min Cao, Dawei Yin, and Zhaochun Ren. MAIR: A massive benchmark for evaluating instructed retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14044–14067, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.778. URL https://aclanthology.org/2024.emnlp-main.778/.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pp. 2024–11, 2024.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*, 2024.

Tatsu-lab. Alpacaeval: An automatic evaluator for instruction-following language models, 2023. URL https://github.com/tatsu-lab/alpaca_eval.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024. URL https://qwenlm.github.io/blog/qwq-32b-preview/.

Hilde van Zeeland. Texgpt: Harness the power of chatgpt in overleaf, 2023. URL https://blog.writefull.com/texgpt-harness-the-power-of-chatgpt-in-overleaf/.

Qian Wang, Tianyu Wang, Qinbin Li, Jingsheng Liang, and Bingsheng He. Megaagent: A practical framework for autonomous cooperation in large-scale llm agent systems, 2024. URL https://arxiv.org/abs/2408.09955.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. PaperRobot: Incremental draft generation of scientific ideas. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1980–1991, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1191. URL https://aclanthology.org/P19-1191/.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Jinghui Zhang, Dandan Qiao, Mochen Yang, and Qiang Wei. Regurgitative training: The value of real data in training large language models. *arXiv preprint arXiv:2407.12835*, 2024.

Zining Zhang, Bingsheng He, and Zhenjie Zhang. Harl: Hierarchical adaptive reinforcement learning based auto scheduler for neural networks. In *Proceedings of the 51st International Conference on Parallel Processing*, ICPP '22, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450397339. doi: 10.1145/3545008.3545020. URL https://doi-org.libproxy1.nus.edu.sg/10.1145/3545008.3545020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models, 2024. URL https://arxiv.org/abs/2403.13372.

A Related Work

LLMs Assist Academic Writing Research on LLMs for academic writing falls into four primary categories. First, *automated paper generation* attempts to produce complete papers but often lacks user control and academic rigor (Shao et al., 2024; Jiang et al., 2024; Anonymous, 2024a; Asai et al., 2024; Schmidgall et al., 2025). Second, *research ideation* employs LLMs to propose novel ideas and methodologies, though concerns regarding authorship and originality persist (Baek et al., 2024; Ghafarollahi & Buehler, 2024; Li et al., 2024a; Si et al., 2024). Third, thanks to the success of retrieval by instruction (Sun et al., 2024), *automated reviewing and research question answering* assist in literature searches and manuscript evaluations but do not directly refine writing quality (D'Arcy et al., 2024; Liang et al., 2024; Lu et al., 2024; Anonymous, 2024b; Asai et al., 2024; Chen et al., 2024b; Lála et al., 2023; Song et al., 2024; Lin et al., 2024). Lastly, *LLM-assisted writing tools* enhance grammar and style and Shi et al. (2023) improves a small paragraph of paper, they lack deep contextual awareness necessary for high-quality academic discourse (CoWriter, 2025; van Zeeland, 2023).

LLMs Assist Research Beyond writing, LLMs are increasingly utilized in autonomous research. (Swanson et al., 2024) introduced LLM agents functioning as research assistants, integrating human feedback into scientific workflows. ChemCrow (M. Bran et al., 2024) and Coscientist (Boiko et al., 2023) highlight LLM-led ideation and experimentation in chemistry, while ResearchAgent (Baek et al., 2024) automates research generation, iterative refinement, and review. AI Scientist (Lu et al., 2024) extends automation to coding, experimentation, and manuscript review. Despite these advancements, studies caution that LLMs require human oversight to ensure reproducibility and scientific rigor (Si et al., 2024).

Gaps and Contributions LLMs also contribute to research tasks such as code generation (Chen et al., 2021; Nijkamp et al., 2022), literature search (Ajith et al., 2024; Kang & Xiong, 2024; Press et al., 2024; Li et al., 2024b), and automated paper reviewing (D'Arcy et al., 2024; Liang et al., 2024; Lu et al., 2024; Weng et al., 2024). While they support ideation (Si et al., 2024), concerns about reduced creativity and homogenization persist (Chakrabarty et al., 2024; Anderson et al., 2024). Hybrid human-LLM approaches are seen as the most effective way to enhance research workflows (Ashkinaze et al., 2024; Liu et al., 2024b; Padmakumar & He, 2024).

Recently, the controllable generation of LLMs have been emphasized (Ge et al., 2025). While much work has focused on using LLMs for idea generation, review, and automation, little research directly addresses refining research papers to enhance coherence, clarity, and adherence to academic standards. Our work bridges this gap by leveraging LLMs specifically for structured refinement, allowing researchers to focus on deeper reasoning tasks while ensuring scholarly rigor.

LLM simulation Researchers have increasingly utilized Large LLMs to construct simulations, treating LLM agents as proxies for humans to perform actions and interactions (Park et al., 2023; Lin et al., 2023; Kong et al., 2024; Wang et al., 2024). These simulations have shown promise in diverse fields such as society, economics, policy, and psychology (Park et al., 2023; Li et al., 2024c; Chen et al., 2024a), while also serving as data generators and evaluators for LLM training (Tang et al., 2024; Zhang et al., 2024). However, LLMs face significant limitations in simulation tasks. Studies (Ai et al., 2024; Petrov et al., 2024; Hu & Collier, 2024; Lee et al., 2024) highlight their inability to maintain contextual consistency and produce fine-grained outputs. For example, Lee et al. (2024) found that LLMs exhibit consistent values and preferences even when role-playing diverse personas, underscoring their lack of adaptability and nuanced understanding.

B Prompts

Figure 5 shows the prompt for QA.

C ICLR 2024 markdown Token Distribution

ICLR 2024 markdown Token Distribution showed in Figure 9.

The Prompt for QA

Act as an expert model for improving articles **PAPER_CONTENT**.
<SELECTED_CONTENT>
User Selected
</SELECTED_CONTENT>
<QUESTION>
<User Question>
</QUESTION>

Figure 5: Prompts for QA

D HUMAN LABEL DETAILS

QA Controllability Assurance	**Judge 1**	**Judge 2**	**Judge 3**
GPT-4o-Mini -Instruction Following	(78+77+72+78+78)/500	(76+68+79)/300	(75+81+80+78+73)/500
-Criteria Following	(79+74+63+77+75)/500	(77+68+79)/300	(76+81+77+74+75)/500
-In-Context Ability	(73+53+48+62+61)/500	(67+57+75)/300	(69+76+74+73+72)/500
-Agree revision?	(48+48+44+53+53)/500	(65+56+64)/300	(67+74+74+72+71)/500
—-			
GPT-o1-mini -Instruction Following	(79+71+76+76+80)/500	(79+81+77)/300	(75+80+78+77+64)/500
-Criteria Following	(72+70+74+74+80)/500	(79+77+75)/300	(74+80+78+76+63)/500
-In-Context Ability	(74+53+58+65+60)/500	(68+68+68)/300	(73+75+81+75+64)/500
-Agree revision?	(58+50+53+59+60)/500	(66+66+67)/300	(72+76+78+76+62)/500

Table 8: Human evaluation on improvement acceptance rates before and after paragraph. we ask 3 human evaluators based on 5,3,5 paper, about 100,60,100 questions in score 1-5. The **Aggregated** column aggregates the results of 3 human evaluators.

XtraGPT	**Judge 1**	**Judge 2**	**Judge 3**
Instruction Following	(62+61+72+76+74)/500	(80+80+79)/300	(86+80+83+82+78)/500
Criteria Following	(60+60+69+72+73)/500	(74+74)/300	(82+82+82+81+82)/500
In-Context Ability	(58+51+48+61+61)500	(67+69)/300	(85+80+82+79+80)/500
Agree revision?	(50+45+44+55+52)/500	(65+64)/300	(83+79+82+80+77)/500

Table 9: our model human evaluation.

E SECTION-LEVEL CRITERIA DETAILS

Section-level criterias are detailed in Table 10,11,13,14,15,12.

The Prompt for Generating QA pairs

You are an advanced language model designed to assist users in improving their articles. Users will provide an article in LaTeX or Markdown format and specify a **section** along with **criteria** for improvement. Your task is to identify a specific selected content from the provided section, align it with the given criteria, and offer actionable feedback to improve the content.
Instructions:
1. **Role 1**: We have a paper improvement task with a specific criteria 'criteria['prompt']'. Now play a role as an author of the provided paper content. Select a specific content from the section 'section' (or equivalent), and ask a chatbot assistant to help you improve that selected content.
- **The selected paper content must be a worth-improving paragraph(s)** that might not achieve the standards of the criteria 'criteria['prompt']', and that content should come from the section 'section'. The selected content will be labeled as **BEFORE IMPROVEMENT**.
- Provide a concise, conversational improvement-related question labeled as **QUESTIONS**. These questions should not explicitly tell what rules or standards to follow or what the specific goal should be. Instead, offer a high-level instruction that may hint at the criteria without stating them directly. The aim is to allow for creativity and subtle alignment with the criteria.
- Keep the question short and conversational.
2. **Role 2**: Act as an expert model for improving articles.
The revised version of the selected content should be labeled as AFTER IMPROVEMENT and specifically address the QUESTIONS on BEFORE IMPROVEMENT above. Avoid adding unnecessary length, unrelated details, overclaims, or vague statements. Focus on clear, concise, and evidence-based improvements that align with the overall context of the paper.
Provide a detailed explanation of the changes made, labeled as EXPLANATION, with clear references to the paper's content. Ensure the explanation demonstrates how the revisions align with the context and criteria of the paper.
— PAPER CONTEXT STARTS
paper_latex
— PAPER CONTEXT ENDS
Response Format (must be strictly followed):
— BEFORE IMPROVEMENT STARTS
<Selected content>
— BEFORE IMPROVEMENT ENDS
— QUESTIONS START
<Concise, improvement-related question based on the criteria 'criteria['prompt']'>
— QUESTIONS END
— AFTER IMPROVEMENT STARTS
<Revised version of the selected content to answer the **Questions** above> — AFTER IMPROVEMENT ENDS
— EXPLANATION STARTS
<An explanation of the changes made, showing how they align with the context of the article and address the criteria. Include references from the paper context where relevant.>
— EXPLANATION ENDS

Figure 6: Prompts for Generate XtraQA

The Prompt for Judging

You are a highly efficient assistant, who evaluates and rank large language models (LLMs) based on the quality of their responses to given prompts. This process will create a leaderboard reflecting the most accurate and human-preferred answers.
I require a leaderboard for various large language models. I'll provide you with prompts given to these models and their corresponding responses. Your task is to assess these responses, ranking the models in order of preference from a human perspective. Once ranked, please output the results in a structured JSON format for the make_partial_leaderboard function.
Prompt

```
{
    "instruction": "{instruction}",
}
```

Model Outputs
Here are the unordered outputs from the models. Each output is associated with a specific model, identified by a unique model identifier.

```
{
    {
        "model": "m",
        "output": "{output\_1}"
    },
    {
        "model": "M",
        "output": "{output\_2}"
    }
}
```

Task
Evaluate based on the quality and relevance to the instructions. The following is the definition of the quality on the section `<section>`: `<criteria["prompt"]>`. If the model's output refers to information beyond `<Selected content>`, it receives a slightly higher score.

Figure 7: Prompts for Judging (modified from alpaca_eval_gpt4_turbo_fn).

F HYPERPARAMS

Hyperparameter	**value**
Batch Size	{1,2}
Cut-off Len	16384
max_new_tokens	512
Epoch	{10,20}
Learning Rate	{1e-5,2e-5}
Details	
Weight Update Per	{4 Step, 6 Step}

Table 10: Hyperparameters

G WINRATE

Table 11 shows the win rate without length control, which is unreasonable compared to Table 4.

The Prompt for ranking

Human: I want you to create a leaderboard of different large-language models. To do so, I will give you the instructions (prompts) given to the models, and the responses of two models. Please rank the models based on which responses would be preferred by humans. All inputs and outputs should be Python dictionaries.
Here is the prompt:

```
{
    "instruction": "{instruction}",
}
```

Here are the outputs of the models:

```
{
    "model": "model_1",
    "answer": "{output_1}"
},
{
    "model": "model_2",
    "answer": "{output_2}"
}
```

Now please rank the models by the quality of their answers, so that the model with rank 1 has the best output. Then return a list of the model names and ranks, i.e., produce the following output:

```
[
    {'model': \texttt{<model-name>},
    'rank': \texttt{<model-rank>}},
    {'model': \texttt{<model-name>},
    'rank': \texttt{<model-rank>}}
]
```

Your response must be a valid Python dictionary and should contain nothing else because we will directly execute it in Python. Please provide the ranking that the majority of humans would give.

Figure 8: Prompts for Scoring.

Models	**Title**	**Abstract**	**Introduction**	**Background**	**Evaluation**	**Conclusion**	**Average↑**
qwen2-72B-Instruct	53.57	70.93	77.52	86.76	91.90	73.71	75.73
GPT-4o-Mini	65.57	59.71	70.05	67.81	70.86	62.14	66.02
Qwen-QWQ-32B-Preview	62.97	66.42	61.33	73.24	72.48	74.29	69.88
Deepseek-V3-671B Liu et al. (2024a)	63.79	59.29	66.19	61.24	88.95	58.57	66.33
Qwen-2.5-7B-Instruct	60.79	70.93	60.52	56.48	74.48	70.43	65.60
PaperCursor (base Qwen-2.5-7b-instruct) (*anchor*↑)							
Llama-3.1-8B-Instruct	47.41	39.64	41.24	55.24	55.71	30.29	44.92
Qwen2.5-1.5B-Instruct	34.36	32.39	26.14	21.24	26.48	31.29	28.65
GPT-3.5-Turbo	28.57	20.79	19.38	20.95	23.05	11.43	20.70
Llama-3.2-3B-Instruct	27.43	9.29	10.90	9.71	14.67	6.43	13.07

Table 11: Win rates of various models against XtraGPT (*anchor*) across different evaluation categories. Models are ranked in descending order based on their averaged win rates.

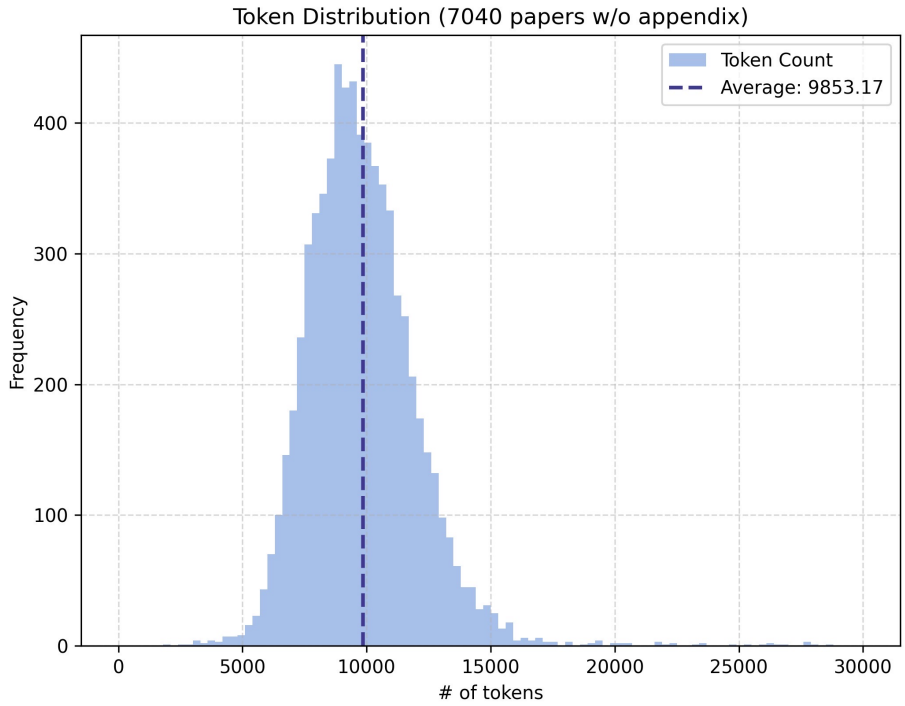


Figure 9: ICLR 2024 Paper Token Distribution (without Appendix)

Criteria Details of Section Title

1. Consistency and Alignment of Title with Paper's Content:
Evaluate the degree to which the paper's title accurately captures its principal topics, arguments, or findings. Does the title reflect the scope and focus of the paper, and is it consistent with the main concepts and keywords presented in the abstract and introduction? Identify any discrepancies or misalignment between the title and the content.

2. Conciseness and Clarity of Title:
Evaluate the paper's title for redundancy. Are there repeated words or concepts that could be removed without changing the core meaning? Does the final title remain succinct, clear, and accurately convey the paper's main focus or contribution?

Figure 10: Criteria Details of Section Title

H Annotators for Controllable Quality Assurance

I Case Study

We chose HARL Zhang et al. (2023) in Figure 19 as a case study to demonstrate the application of XtraGPT in human-AI collaboration. XtraGPT helps the author of MegaAgent refine the paper in a controllable manner.

Criteria Details of Section Abstract

1. Clarity and Impact of the Conclusion:
Evaluate the clarity and impact of the conclusion in the abstract. Does it clearly summarize the research steps, highlight key outcomes, and explain the significance of these outcomes for the field of computer science? Are the primary technical advancements and their contributions presented in a concise and unambiguous manner?

2. Motivation and Purpose in the Abstract:
Evaluate how well the abstract communicates the research's motivation. Does it clearly articulate the broader issue, concept, or problem in Computer Science that the work addresses? Does it explicitly state the specific research problem being solved and why it is important?

3. Explanation of Existing Solutions and Research Gap:
Assess how well the abstract explains the shortcomings of current solutions and highlights the corresponding research gap. Does it clearly articulate why existing methods are insufficient and how the proposed approach addresses these limitations? Is the explanation comprehensible to a wide audience, from domain experts to non-specialists?

4. Clarity and Adequacy of Proposed Solutions:
Assess how effectively the abstract communicates the proposed solutions. Does it clearly identify the research gap or problem being addressed, and explain how the proposed solution tackles this gap? Does it highlight the novelty or contribution of the solution, demonstrating its relevance or improvement over existing work? Rate the clarity, completeness, and significance of the explanation provided in the abstract.

Figure 11: Criteria Details of Section Abstract

Criteria Details of Section Conclusion

1. Broader Impact and Future Directions:
Assess the thoroughness of the paper's conclusion or discussion sections in addressing the broader impact of the research. Does the paper provide specific and clear avenues for future work?

2. Clarity and Impact of Key Innovations and Findings:
Evaluate whether the conclusion effectively highlights the paper's key innovations.

Figure 12: Criteria Details of Section Conclusion

J Baseline Model Details

K Controllability Annotation Criterias and Interface

To ensure our data and model quality, We invited three AI experts specializing in inference speedup, graph neural networks (GNN), and Field Programmable Gate Arrays (FPGA) to annotate 5, 3, and 5 papers, respectively. Each paper includes 20 question-answer pairs per model, focusing on section-level improvements. These pairs are distributed across different sections of the paper as follows: 2 for the title, 4 for the abstract, 6 for the introduction, 3 for the background, 3 for the evaluation, and 2 for the conclusion. The controllable criteria used for evaluation are presented in Figure 16. The annotators' operating interface and the interface of XtraGPT are listed in Figure 17,18.

Criteria Details of Section Introduction

1. Strength and Clarity of Motivation in the Introduction:
Evaluate whether the motivation in the Introduction is specific and convincing. Does the paper avoid over-generalization and clearly articulate the significance of the issue? Are concrete examples, statistics, or contextual details used to establish why the problem matters?

2. Review of Existing Approaches in Introduction:
Assess the thoroughness and clarity of the literature review within the introduction. Does the paper cite and critique relevant prior works, highlighting both their methodologies and limitations? Does the introduction establish how the proposed work builds upon or differentiates itself from existing research, and is there sufficient context provided to demonstrate the significance of the current study? Are any quantitative or qualitative comparisons made where appropriate?

3. Audience Alignment and Appropriateness:
Evaluate whether the introduction is effectively tailored to its target audience. Is the complexity, depth, and choice of terminology suitable for the presumed background knowledge of the readership? Does the introduction provide sufficient context without oversimplifying or overwhelming the intended audience?

4. Clarity and Visibility of Contributions:
Assess the clarity and visibility of the paper's contributions. Are the core contributions explicitly stated in a dedicated paragraph or section toward the end of the introduction? Are they understandable to a broad scientific audience, presented succinctly, and positioned logically following the problem statement and background information?

5. Clarity and Specificity of Problem Definition:
Evaluate the paper's problem definition in terms of four key elements: current situation, ideal situation, the gap between them, and how the research aims to address this gap. Are these components clearly stated, distinct, and directly tied to the research objectives? Does the definition provide sufficient clarity and focus for the research?

6. Integration of State-of-the-Art in Problem Framing:
Evaluate how effectively the introduction incorporates the State-of-the-Art (SOTA) to frame the research problem. Does it include explicit references to key works, methodologies, or findings that highlight relevant gaps or limitations in the field? Is there a clear logical link between the SOTA discussion and the stated research objectives, demonstrating how the proposed work builds upon or extends existing research?

Figure 13: Criteria Details of Section Introduction

Criteria Details of Section Background

1. Contextual Relevance and Clarity of Background:
Assess how effectively the background section establishes context for the research. Does it provide a clear overview of the broader field in computer science, then narrow down to the specific problem? Does the paper avoid making unwarranted assumptions about the reader's prior knowledge? Finally, does it clarify why addressing the problem is important to the field?

2. Coverage of Key Preliminary Concepts:
Evaluate the thoroughness and clarity of the paper's background or preliminary section. Does it introduce and define all the critical concepts, algorithms, or theorems necessary to understand the technical contributions? Are these concepts clearly explained, logically organized, and accessible to readers who are not experts in the field? Does the paper use consistent terminology and adequately explain symbols, abbreviations, or specialized terms before their first usage?

3. Clarity and Consistency of Terminology:
Assess the clarity and consistency of the key terms introduced in the background section. Are all critical terminologies defined at their first occurrence and used consistently throughout the paper? Does the paper avoid undefined shifts or redefinitions of terms, and does it align terminology with standard conventions in the field?

Figure 14: Criteria Details of Section Background

Criteria Details of Section Evaluation

1. Experimental Setup Clarity and Reproducibility:
Evaluate how clearly and thoroughly the experimental setup is described. Does the paper provide all necessary information on hardware/software configurations, parameter settings, data preprocessing steps, and any contingency procedures, such that others could replicate the experiments with the same resources?

2. Depth and Clarity of Figures and Tables Analysis:
Evaluate the thoroughness and clarity of the paper's analysis of figures and tables. Are the data clearly explained and linked to the research objectives or hypotheses? Do the authors discuss trends, patterns, or anomalies, and interpret quantitative metrics in a way that highlights their significance? Is there a clear comparison to baselines or related work, demonstrating how the results fit into or advance the field? Do the authors emphasize key takeaways and practical or theoretical implications arising from the findings?

3. Experimental Support for Main Innovations:
Evaluate how thoroughly the paper's main innovations or contributions are backed by experimental evidence. Does the paper provide direct tests or comparisons to validate each innovation? Are quantitative or qualitative results clearly linked to the claims made, with appropriate metrics and comparisons against baselines or existing methods? Are ablation studies or sensitivity analyses included to demonstrate the significance of each component? If certain claims are not experimentally supported, have the authors either provided additional experiments or adjusted their claims accordingly?

Figure 15: Criteria Details of Section Evaluation

Criteria

Each QA pair is evaluated based on four metrics, each scored from 1 to 5:
Evaluation Metrics (1-5 Scoring Criteria)
1. **Instruction Following:** Evaluate whether the answer correctly follows the given instruction.
1 – The answer completely ignores or contradicts the instruction.
2 – The answer only partially follows the instruction, with major missing elements.
3 – The answer follows the instruction but lacks completeness or clarity.
4 – The answer mostly follows the instruction with minor inconsistencies.
5 – The answer strictly follows and fully satisfies the instruction.
2. **Criteria Following:** Evaluate whether the revised text improves the original content based on predefined criteria.
1 – The revision does not follow any criteria and worsens the content.
2 – The revision attempts to follow the criteria but makes the content unclear.
3 – The revision follows the criteria but does not provide a significant improvement.
4 – The revision improves clarity and correctness while adhering to the criteria.
5 – The revision strictly follows the criteria and significantly improves the original content.
3. **In-Context Ability:** Evaluate whether the model's output appropriately references information within Selected Content.
1 – The output ignores Selected Content and adds irrelevant external information.
2 – The output relies on external information without justification.
3 – The output primarily references Selected Content but includes minor unrelated details.
4 – The output correctly refers to Selected Content with minimal external additions.
5 – The output strictly remains within Selected Content while providing a relevant and precise response.
4. **Agree Revision:** Evaluate whether the revision is convincing enough for the user to adopt it as a replacement.
1 – The revision is clearly worse than the original text.
2 – The revision is slightly better but has major flaws, making it unlikely to be adopted.
3 – The revision is neutral or slightly better, but adoption is uncertain.
4 – The revision is clearly better, and most users would likely adopt it.
5 – The revision is significantly better, and users would confidently adopt it.

Figure 16: The criteria for human instructors.

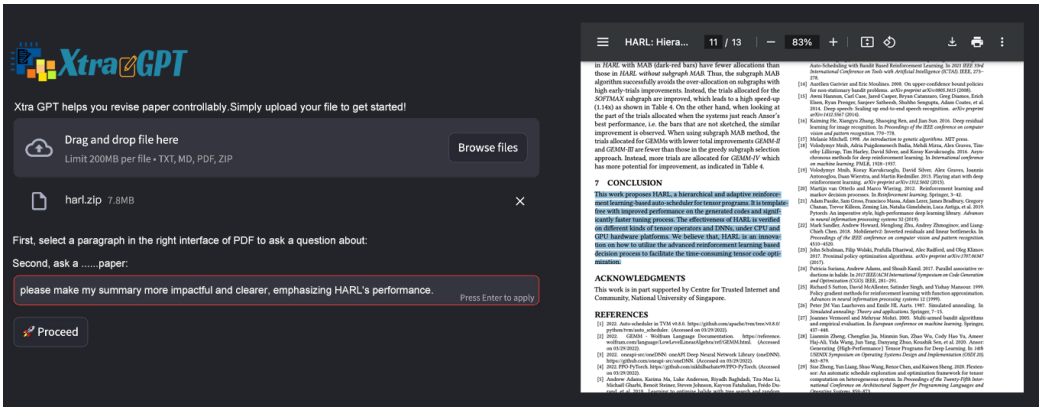


Figure 17: Demo of XtraGPT.

```
This work proposes HARL, a hierarchical and adaptive
reinforcement learning-based auto-scheduler for tensor
programs. It is templatefree with improved performance
on the generated codes and significantly faster tuning
process. The effectiveness of HARL is verified
on different kinds of tensor operators and DNNs, under
CPU and GPU hardware platforms. We believe that, HARL
is an innovation on how to utilize the advanced
reinforcement learning based
decision process to facilitate the time-consuming
tensor code optimization.
```

```
In this paper, we introduce HARL, a hierarchical and
adaptive reinforcement learning-based auto-scheduler
that significantly enhances tensor program exploration.
By employing a hierarchical RL architecture, HARL
enables learning-based decisions at various levels of
search granularity, leading to a $22\%$ improvement in
tensor operator performance and a $4.3\mathrm{x}$
increase in search speed compared to the state-of-the-
art auto-scheduler, Ansor. Our experiments reveal that
HARL not only outperforms Ansor in both tuning speed
and final program performance but also achieves up to
$8\%$ improvement in inference performance on end-to-
end neural networks. These results underscore the
effectiveness of HARL in addressing the challenges of
neural network optimization, providing a robust
framework for future research in this domain.
```

Figure 18: A comparison of the paragraph before and after revision.

Selected Content: This work proposes HARL, a hierarchical and adaptive reinforcement learning-based auto-scheduler for tensor programs. It is templatefree with improved performance on the generated codes and significantly faster tuning process. The effectiveness of HARL is verified on different kinds of tensor operators and DNNs, under CPU and GPU hardware platforms. We believe that, HARL is an innovation on how to utilize the advanced reinforcement learning based decision process to facilitate the time-consuming tensor code optimization.
User Interaction: How can we effectively summarize the key innovations of HARL and their contributions to the field of neural network optimization? What specific results and comparisons to existing methods should we include to demonstrate the significance of our findings?

Responses: In this paper, we introduce HARL, a hierarchical and adaptive reinforcement learning-based auto-scheduler that significantly enhances tensor program exploration. By employing a hierarchical RL architecture, HARL enables learning-based decisions at various levels of search granularity, leading to a 22% improvement in tensor operator performance and a $4.3\mathrm{x}$ increase in search speed compared to the state-of-the-art auto-scheduler, Ansor. Our experiments reveal that HARL not only outperforms Ansor in both tuning speed and final program performance but also achieves up to 8% improvement in inference performance on end-to-end neural networks. These results underscore the effectiveness of HARL in addressing the challenges of neural network optimization, providing a robust framework for future research in this domain.
Explanations: The revised conclusion emphasizes the key innovations of HARL, specifically highlighting the 22% performance improvement and $4.3\mathrm{x}$ speed increase over Ansor. By including quantitative results and comparisons to existing methods, we demonstrate the significance of our findings in the context of neural network optimization. The hierarchical RL architecture is clearly linked to these improvements, showcasing how it facilitates more efficient exploration of tensor programs. This alignment with the criteria of effectively summarizing key innovations and providing concrete results is crucial for advancing the field.

Figure 19: A use case on XtraGPT.

Models	Backbone	GitHub Stars	Huggingface Download
	Closed-Source		
GPT-4-Turbo (OpenAI et al., 2024)		/	
GPT-4o-mini (OpenAI et al., 2024)		/	
GPT-3.5-Turbo (Brown et al., 2020)		/	
	Open-Source (> 8B)		
Deepseek-R1 (DeepSeek-AI et al., 2025)	Llama	50K	285K
Deepseek-V3-671B (Liu et al., 2024a)	DeepSeek-V3-Base	63K	374K
Deepseek-V3-32B (Liu et al., 2024a)	DeepSeek-V3-Base	63K	374K
Qwen-2-72B-Instruct (Yang et al., 2024)	Qwen-2-72B-Instruct	45.3K	374K
QwQ-32B-Preview (Team, 2024)	Qwen2.5-32B-Instruct	15K	198K
Phi-4 (14B) (Abdin et al., 2024)	-	-	557K
	Open-Source (≤8B)		
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Llama-3.1-8B-Instruct	28.1K	5.75M
Qwen-2.5-7B-Instruct (Yang et al., 2024)	Qwen-2.5-7B	12.6K	1.27M
Llama-3.2-3B-Instruct (Grattafiori et al., 2024)	Llama-3.2-3B	28.1K	1.48M
Qwen-2.5-1.5B-Instruct (Yang et al., 2024)	Qwen-2.5-1.5B	12.6K	551K

Table 12: Details information of baseline models. Data collected at 30.1.2025. The "/" indicates that the model uses a private download link or that its download statistics on HuggingFace are not disclosed.