ABGEN: Evaluating Large Language Models in Ablation Study Design and Evaluation for Scientific Research

Yilun Zhao[*Y] **Weiyuan Chen**[*Y] **Zhijian Xu**[Y] **Manasi Patwardhan**[T]
Yixin Liu[Y] **Chengye Wang**[Y] **Lovekesh Vig**[T] **Arman Cohan**[Y]
[Y] Yale NLP Lab [T] TCS Research

Abstract

We introduce **ABGEN**, the first benchmark designed to evaluate the capabilities of LLMs in designing ablation studies for scientific research. ABGEN consists of 1,500 expert-annotated examples derived from 807 NLP papers. In this benchmark, LLMs are tasked with generating detailed ablation study designs for a specified module or process based on the given research context. Our evaluation of leading LLMs, such as o4-mini and DeepSeek-R1, highlights a significant performance gap between these models and human experts in terms of the importance, faithfulness, and soundness of the ablation study designs. Moreover, we demonstrate that current automated evaluation methods are not reliable for our task, as they show a significant discrepancy when compared to human assessment. To better investigate this, we develop **ABGEN-EVAL**, a meta-evaluation benchmark designed to assess the reliability of commonly used automated evaluation systems in measuring LLM performance on our task. We investigate various LLM-as-Judge systems on ABGEN-EVAL, providing insights for future research on developing more effective and reliable LLM-based evaluation systems for complex scientific tasks.

Data yale-nlp/AbGen
Code yale-nlp/AbGen

1 Introduction

In empirical scientific fields, designing experiments and selecting the appropriate experimental settings often present considerable challenges and requires significant domain expertise. Oftentimes, scientists learn about the flaws in their experimental design and missing ablations after going through a peer review process, which involves domain experts carefully evaluating a scientific work. The complexity of tasks in experimental science underscores the need for innovative approaches to support researchers in optimizing their workflows. Meanwhile, LLMs have demonstrated remarkable capabilities across a range of tasks integral to scientific processes, such as reviewing manuscripts (D'Arcy et al., 2024; Du et al., 2024), scientific writing (Altmäe et al., 2023; Xu et al., 2024), scientific code generation (Liu et al., 2023; Yang et al., 2024b). This raises a compelling question: **Can LLMs be effectively leveraged to assist scientists in the process of experimental design?**

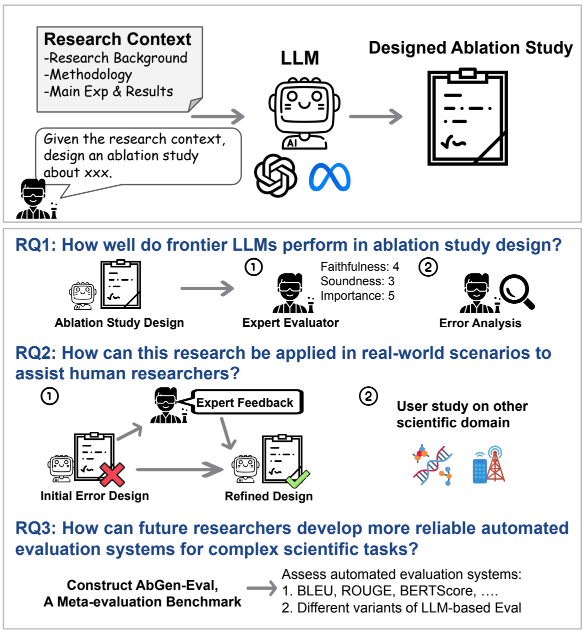

Figure 1: Overview of the research: the ablation study design task and three research questions investigated.

While addressing this question is inherently complex due to the diverse nature of scientific disciplines and difficulty of evaluation, our objective is to introduce the first comprehensive benchmark as well as an evaluation methodology to facilitate measuring progress on this task. We particularly introduce **ABGEN**, the first benchmark for evaluat-

[*] Equal Contributions. Correspondence: Yilun Zhao (yilun.zhao@yale.edu)

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12479–12491
July 27 - August 1, 2025

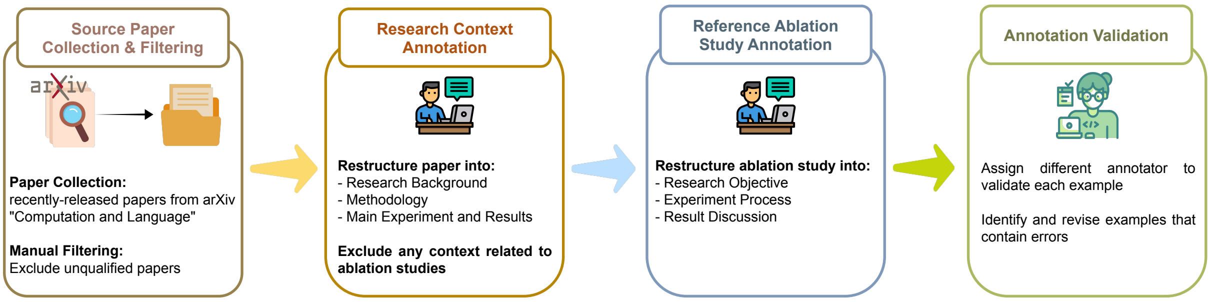

Figure 2: An overview of ABGEN construction pipeline.

ing LLMs in the context of designing ablation studies for scientific research. The dataset consists of 1,500 examples derived from 807 scientific papers in natural language processing (NLP). Each example is carefully annotated and validated by NLP experts and includes a comprehensive research context along with a reference ablation study, both restructured from the original research paper. The research context is divided into three sections: research background, methodology, and the main experiment setup and results. As illustrated in Figure 1, the LLMs are tasked with generating a detailed ablation study design for a specified module or process based on the provided research context.

As outlined in Figure 1, we investigate three research questions in this study. Our main contributions are summarized below:

- We propose ABGEN, the first benchmark designed to evaluate the capabilities of LLMs in ablation study designs for scientific research (§2).
- We design a comprehensive human and automated evaluation systems for ABGEN (§3).
- We conduct a systematic evaluation of leading LLMs, analyzing their strengths and limitations on our new task, and providing insights for future advancements (§4.2).
- Our user studies reveals the potential of LLMs in ablation study design by interaction with human researchers, and highlights the adaptability of this approach to other scientific domains (§4.3).
- We develop the meta-evaluation benchmark, ABGEN-EVAL, and investigate various LLM-based evaluation methods to provide insights for creating more reliable automated evaluation systems for complex scientific tasks (§5).

2 ABGEN Benchmark

To systematically study the capabilities and limitations of current LLMs and measuring progress in assisting scientists with the design of their experimental workflows, we introduce a new benchmark named ABGEN. The LLMs are tasked with generating detailed ablation study designs for a specified module or process based on the given research context. We focus on scientific research within the NLP domain, as the involved expert annotators primarily have expertise in NLP (*i.e.,* each has at least one publication in a top-tier NLP or AI venue as a leading author). Detailed biographies of the annotators participating in the ABGEN annotation and LLM performance evaluation process are provided in Table 6 in Appendix A.1. We believe that future research could extend our benchmark construction pipeline to extend to other scientific domains.

In the following subsections, we first provide a formal definition of the ABGEN task and then detail each step within the benchmark construction process. We present an overview of the ABGEN construction pipeline in Figure 2.

2.1 ABGEN Task Formulation

We formally define the task of ABGEN in the context of LLMs. Specifically, given:

- The **research context** $\boldsymbol{C}$, which is an expert-annotated context of a specific scientific study. This context is restructured from the original paper by expert annotators, including sections of research background, methodology, and main experiment setup and results (§2.3).
- The name of a specific essential module or process, denoted as $\boldsymbol{M}$, which is described in the *methodology* section within research context $\boldsymbol{C}$.

The LLM is tasked with generating the design for an ablation study, $\boldsymbol{A}$, aimed at evaluating the contribution and impact of $\boldsymbol{M}$ within the overall research framework:

$$\hat{\boldsymbol{A}} = \arg\max_{A} P_{\mathbf{LLM}}(\boldsymbol{A} \mid \boldsymbol{C}, \boldsymbol{M}) \quad (1)$$

The ablation study design should include a clear statement of the research objective, along with a detailed description of the experimental process.

2.2 Source Paper Collection and Filtering

Source Paper Collection. We collect scientific papers from arXiv under the "Computation and Language" category, targeting those first released between March 1, 2024 and August 30, 2024. For each paper, we adopt the tool[1] developed by Lo et al. (2020) to extract its content. Specifically, this tool parses LaTeX source files of papers into JSON format, extracting features including the paper title, abstract, main sections, and appendix. We convert tables within the papers into HTML format. Both recent works (Sui et al., 2024; Fang et al., 2024) and our preliminary studies reveal that the evaluated LLMs can comprehend such table format effectively. Next, we describe our approach and criteria for inclusion of the papers for annotation, as well as the details of the annotation process.

Research Paper Manual Filtering. For each collected NLP paper, the expert annotator first determines if they are familiar with the paper's topic. If not, we randomly assign the paper to another annotator. Papers whose topics are unfamiliar to both annotators are excluded. The annotators are then instructed to determine whether the paper qualifies for inclusion in our benchmark. Specifically, we exclude: (1) Papers that are not focused on experimental work (*e.g.,* surveys, position papers, dissertations), as they do not involve ablation study design; (2) Papers with fewer than two ablation studies, as these may not provide sufficient breadth of experimental evidence. Additionally, annotators may exclude papers they deem to be of low quality based on their expert judgment. After applying these filtering criteria, 807 papers remain for further annotation.

2.3 Research Context Annotation

After determining that a research paper qualifies for benchmark inclusion, annotators are instructed to restructure the original paper into research context that maintains the original meaning but exclude any content related to ablation studies. The research context contains the following three sections: (1) **Research Background**, which is restructured from the introduction and related work sections, describing the paper's motivation, research problem, and relevant prior work. (2) **Methodology**, which is restructured from the methodology sections, This section describes the proposed method or model, including key components and innovations. (3) **Main Experiment Setup and Results**, which is restructured from the experiment sections. This section details the primary experimental setup, including datasets, baselines, and evaluation metrics used in main experiments, as well as the main experimental results.

2.4 Reference Ablation Study Annotation

Annotators are then tasked with restructuring each ablation study in the research paper into a reference ablation study. It consists of the following three sections: (1) **Research Objective**, a one- or two-sentence description of the research problem and the goal of the ablation study. If this statement is not explicitly provided in the original ablation study, annotators are required to infer and summarize it. (2) **Experiment Process**, a detailed account of the experimental setup, including the experimental groups, datasets, procedures, and the evaluation tools and metrics used. Annotators are requried to ensure that the process is clearly understandable and replicable based on the provided description. (3) **Result Discussion**, an analysis of the outcomes, where annotators summarize the key findings and their implications. It's worth noting that we do not instruct LLMs to generate this part in our experiments, as our main focus is on evaluating their ability to design ablation studies rather than execute and analyze experiments. However, we believe these features could be valuable for future research, which we elaborate on in the limitations section.

2.5 Annotation Validation

For each annotated example, we assign an annotator to validate the annotated research context and reference ablation study based on the original research paper. They are required to identify and revise examples that contain errors. Out of the 1,500 annotated examples, 273 were identified as erroneous and were subsequently revised. We conducted a final human evaluation of data quality on 100 examples. As shown in Table 5 (Appendix A.1), for each validation metric, over 95% of the samples received a satisfaction rating of at least 4 out of 5. This result indicates the high quality of ABGEN.

[1] https://github.com/allenai/s2orc-doc2json

Property	Value (avg./max)
Research Context Word Length	1,847.8 / 6,253
Research Background	319.6 / 1,178
Methodology	904.4 / 4,685
Exp Setup & Results	623.7 / 2,174
Ref. Ablation Study Word Length	145.5 / 518
Research Objective	6.1 / 15
Experiment Process	72.5 / 264
Result Discussion	67.1 / 336
# NLP Research	807
# Ref. Ablation Study per Research	1.9 / 3
AbGen Size	1,500
Testmini Set	500
Test Set	1,000

Table 1: Data statistics of the AbGen benchmark.

2.6 Data Statistics

Table 1 illustrates the data statistics of the AbGen benchmark. We randomly split the dataset into two subsets: *testmini* and *test*. The *testmini* subset contains 500 examples and is intended for both method validation and human analysis and evaluation. The *test* subset comprises the remaining 1,000 examples and is designed for standard evaluation.

3 AbGen Evaluation

The automated evaluation of LLM generation for tasks relevant to scientific workflows remains an unsolved problem in the community. Recent benchmark work, such as SciMON (Wang et al., 2024a) for novel scientific direction generation and MARG (D'Arcy et al., 2024) for peer review generation, primarily rely on human evaluation to assess LLM-based system performance. In our study, we also employ human evaluation by expert annotators as the *primary* assessment method. Additionally, in Section 5, we investigate different variants of LLM-based evaluation methods, aiming to provide insights for future work to develop automated evaluation systems for a large-scale evaluation.

3.1 Evaluation Criteria

This section discusses the human and automated evaluation protocols developed for AbGen evaluation. We assess the following three dimensions for the generated ablation study design.

- **Importance**: The generated ablation study design will provide valuable insights into understanding the role of the specified module or process within the overall methodology.
- **Faithfulness**: The generated ablation study design aligns perfectly with the given research context. There are no contradictions between the generated content and the main experimental setup within the provided research context.
- **Soundness**: The generated ablation study design is logically self-consistent without ambiguious description. The human researchers would be able to clearly understand and replicate the ablation study based on the generated context.

To determine these three dimensions, we gathered feedback from three external senior NLP researchers, all of whom serve as area chairs for the ACL Rolling Review. Through iterative discussions, we identified these dimensions as critical for evaluating the quality and utility of generated ablation study designs. This feedback process also helped us in refining the assessment guidelines used for human evaluation (§3.2). We do not evaluate the *fluency* of the generated ablation study, as both recent works (D'Arcy et al., 2024; Zeng et al., 2024) and our preliminary findings find that leading LLMs consistently produce fluent text free of grammatical errors.

3.2 Human Evaluation Protocol

For human evaluation, we use Likert-scale scores ranging from 1 to 5 for each criterion (*i.e.,* importance, faithfulness, and soundness). Given the research context and an LLM-generated ablation study, human evaluators are asked to score the generated content for each criteria. Initially, the reference ablation study is not provided to the evaluator. This approach encourages evaluators to carefully review the generated content in light of the research context, reducing the likelihood of bias from comparing it to the reference. This is particularly important, as LLMs may generate ablation studies that, while reasonable, differ from the reference. After submitting their initial scores, evaluators are then given the reference ablation study and asked to adjust their scores if they identify any aspects they may have initially overlooked.

To assess inter-annotator agreement of our human evaluation, we sample 40 fixed LLM-generated outputs that are separately evaluated by all four expert annotators. They achieve inter-annotator agreement scores (*i.e.,* Cohen's Kappa) of 0.735, 0.782, and 0.710 for the criteria of importance, faithfulness, and soundness, respectively.

3.3 Automated Evaluation

While human evaluation is generally reliable, it is time-consuming and does not scale well. To address this, we also employ an LLM-as-a-judge system for automated evaluation. Specifically, we use GPT-4.1-mini as the base evaluator. For each model-generated response, the evaluator is provided with the research context and a reference ablation study. It then assigns a score from 1 to 5 for each evaluation criterion (*i.e.,* importance, faithfulness, and soundness). Before assigning the final score, the evaluator is required to generate an explanation outlining the rationale behind its scoring. To gain a deeper understanding of the reliability of LLM-as-Judge systems, we develop the meta-evaluation benchmark, ABGEN-EVAL, which is detailed in Section 5.

4 LLMs for Ablation Study Design

4.1 Experiment Setup

Evaluated Systems. We examine the performance of 18 frontier LLMs across two distinct categories on our benchmark: (1) **Proprietary LLMs**, including o4-mini (OpenAI, 2025a), GPT-4o (OpenAI, 2024), GPT-4.1 (OpenAI, 2025b), Gemini-2.5-Flash (Gemini, 2024); and **Open-source LLMs**, including Llama-3.1-70B, Llama-3.3-70B, Llama-4-Scout-17B and Llama-4-Maverick-17B (AI@Meta, 2024; Meta AI, 2025), Mistral-Large (Jiang et al., 2024), Deepseek-V3, DeepSeek-R1-0528-Qwen3-8B, and Deepseek-R1 (DeepSeek-AI, 2024, 2025), Phi-4 (Microsoft et al., 2025), Gemma-3-27b-it (Team et al., 2025) , Qwen2.5-32B, Qwen3-8B, Qwen3-32B and Qwen3-235B-A22B, (Yang et al., 2024a; Team, 2025). Table 7 in Appendix presents the details of these evaluated LLMs in ABGEN.

Measuring Performance of Real Paper and Expert. To provide an informative estimate of real paper and expert-level performance on ABGEN, we randomly sample 20 examples from 10 papers in the *testmini* set. We enlist two expert annotators (*i.e.,* Annotators 1 and 4, as described in Table 6 in Appendix A.1) to individually solve these examples. To ensure fairness, we mix these 20×2 expert-annotated data and corresponding 20 reference ablation study within the standard human evaluation process. The expert evaluators are not informed of the sources of these ablation study examples when evaluation. We report the evaluation results on Table 2.

> **Ablation Generation Prompt**
>
> `[System Input]:`
> Given the research context, design an ablation study for the specified module or process. Begin the design with a clear statement of the research objective, followed by a detailed description of the experiment setup. Do not include the discussion of results or conclusions in the response, as the focus is solely on the experimental design. The response should be within 300 words. Present the response in plain text format only.
>
> `[User Input]:`
> Research Context:{research context}
> Design an ablation study about {ablation module} based on the research context above.

Figure 3: Prompt for ablation study generation.

Implementation Details. For all the experiments, we set temperature as 1.0 and maximum output length as 1024 (as the maximum length of reference ablation study is 264 words as presented in Table 1). Figure 3 illustrates the default prompt used across all generation experiments. The model is tasked with generating the design for an ablation study, based on the provided annotated research context and the specified module or process name. Specifically, the LLMs are required to first generate a one-sentence description of the research objectives, followed by a detailed description of the experimental setup for the ablation study.

4.2 Results and Analysis

> 💡 **RQ1:** How well do frontier LLMs perform in designing ablation studies?

Table 2 illustrates the performance of the evaluated LLMs on ABGEN. The human evaluation results demonstrate that ABGEN poses significant challenges to current LLMs. Even the best-performing LLM, DeepSeek-R1-0528, performs much worse than human experts. This gap highlights the critical need for further advancements in LLMs, especially in applying them to complex scientific tasks. Moreover, we observe a significant disparity between automated evaluation systems and human assessments. For instance, despite receiving lower scores in LLM-based evaluations compared to Qwen3-235B-A22B, GPT-4.1 consistently outperforms it in every criterion according to human evaluation. These results suggest that current automated evalu-

System	LLM-based Eval (1-5)			Human Evaluation (1-5)			
	Import.	Faith.	Sound.	Import.	Faith.	Sound.	**Avg.**
Reference (orig)	–	–	–	4.70	4.90	4.70	4.77
Human Expert	4.82	4.84	4.33	4.65	4.93	4.83	4.80
DeepSeek-R1-0528	4.80	4.85	4.39	4.23	4.0	4.11	4.11
o4-mini	4.80	4.81	4.33	4.23	3.78	4.00	4.00
GPT-4.1	4.82	4.84	4.28	4.12	3.87	4.02	4.00
DeepSeek-V3	4.78	4.80	4.19	3.98	3.79	3.96	3.91
Qwen3-235B-A22B	4.83	4.76	4.31	4.26	3.43	4.00	3.90
Gemini-2.5-Flash	4.63	4.52	4.01	3.89	3.94	3.76	3.86
Gemma-3-27b-it	4.70	4.75	4.21	3.78	3.81	3.96	3.85
GPT-4o	4.81	4.75	4.15	3.88	3.67	3.91	3.82
Qwen3-32B	4.82	4.74	4.22	3.90	3.47	3.98	3.78
Qwen3-8B	4.77	4.69	4.16	3.86	3.46	3.89	3.74
Mistral-Small-3.1-24B	4.74	4.63	4.12	3.74	3.35	3.84	3.64
Phi-4	4.74	4.65	4.12	3.70	3.34	3.78	3.61
Llama-4-Maverick-17B	4.66	4.64	4.04	3.46	3.66	3.68	3.60
DeepSeek-R1-0528-Qwen3-8B	4.69	4.68	4.12	3.71	3.18	3.65	3.51
Qwen2.5-32B	4.73	4.64	4.08	3.53	3.17	3.72	3.47
Llama-4-Scout-17B	4.71	4.51	4.04	3.49	3.22	3.50	3.40
Llama-3.1-70B	4.68	4.46	4.05	3.58	2.91	3.55	3.35
Llama-3.3-70B	4.68	4.45	4.03	3.27	3.08	3.49	3.28

Table 2: Human and automated evaluation results of LLMs on ABGEN. For automated evaluation, we use GPT-4.1-mini as the base evaluator and report scores on the *test* subset. For human evaluation, we randomly sample 100 examples from the *testmini* subset. Each model output is assessed by an expert evaluator. The average human score is used as the primary metric for ranking model performance in this table.

ation systems may not be fully reliable for our task. To gain a deeper understanding of the reliability of current automated evaluation systems, we develop the meta-evaluation benchmark, ABGEN-EVAL, which is detailed in Section 5.

Error Analysis. We further conduct a comprehensive error analysis to better understand the capabilities and limitations of the top-performing LLMs on our task. This error analysis is based on 100 failure cases of GPT-4o from the *testmini* set, where the average human evaluation scores are below 3. We identify five common error types, and provide detailed explanations for each type in Table 3. These error cases demonstrate that generating constructive ablation study designs based on research context is still challenging for LLMs.

4.3 User Studies on Real-world Scenarios

> **RQ2:** How can this research be applied in real-world scenarios to assist human researchers in designing ablation studies?

To investigate this research question, we design and conduct following two user studies:

LLM-Researcher Interaction While LLMs currently lag behind human experts in designing ablation studies, they still hold value as tools to assist researchers. To explore this potential, we examine scenarios where researchers interact with LLMs, providing feedback to guide the refinement of their outputs. Specifically, we first sample 20 failure cases from *testmini* set—each with an average human score below 3—from both GPT-4o and Llama-3.1-70B. Two expert annotators are then tasked with reviewing these LLM-generated ablation study designs, identifying errors, and providing constructive feedback for improvement within a 50-word limit. We then feed the research context, initial ablation study design, and researcher feedback back into the same LLMs, instructing them to regenerate the ablation study design. Another expert evaluator is then assigned to assess the revised version, following the same human evaluation protocol in Section 3.2. As shown in Table 4, incorporating researcher feedback can significantly enhance LLM performance in refining their outputs.

Domain Generalization of Our Research. Our research primarily focuses on NLP domains. To explore the adaptability of our work across other

Error Type	Explanation
Misalignment with research context	This error arises when the generated experiment process contradicts with the baseline in the research context or introduces factual errors.
Ambiguity and Difficulty in Reproduction	This error arises when the generated experiment process contains ambiguous steps or lacks the necessary datasets or tools, for human researchers to replicate the ablation study.
Partial Ablation or Incomplete Experimentation	This error arises when the generated experiment process partially addresses the ablation module, such as only ablating a sub-module, or missing experimental groups.
Insignificant Ablation Module	This error arises when the generated research objective is focused on an insignificant ablation module in research context.
Inherent Logical Inconsistencies	This error arises when the generated experiment process contains inherent logical inconsistencies, such as gaps in implementation steps.

Table 3: A summary of GPT-4o's failure cases.

User Study	Import.	Faith.	Sound.
User Study 1: LLM-Researcher Interaction			
GPT-4o			
Initial Failure Case	3.9	2.1	2.0
Revision with Feedback	4.8 (+0.9)	4.2 (+2.1)	4.6 (+2.6)
Llama-3.1-70B			
Initial Failure Case	3.7	1.8	1.7
Revision with Feedback	4.5 (+0.8)	3.9 (+2.1)	4.1 (+2.4)
User Study 2: Domain Generalization			
GPT-4o			
NLP Domain (as Main Exp)	3.9	3.4	3.3
Biomedical Domain	3.7	3.4	3.1
Computer Network Domain	3.8	3.3	3.4
Llama-3.1-70B			
NLP Domain (as Main Exp)	3.3	2.8	2.8
Biomedical Domain	3.0	2.8	2.9
Computer Network Domain	3.1	2.9	3.0

Table 4: Human evaluation result from two user studies. The findings demonstrate (1) the potential of LLMs in designing ablation studies through interaction with human researchers, and (2) the adaptability of our research across different scientific domains.

scientific fields, we conducted user studies in the areas of biomedical sciences and computer networks. Specifically, we engage two experts—one in computer networking and one in biomedical research—to provide five research papers from their respective fields that were first published after May 1, 2024, and with which they are familiar. Following the same procedure as ABGEN annotation, they annotate the research context and reference ablation studies from five corresponding papers, resulting in a total of 27 examples over ten papers. We then provide them with LLM-generated ablation study designs and ask them to strictly follow our human assessment guidelines to evaluate the LLM outputs. As shown in Table 4, the human evaluation scores for GPT-4o and Llama-3.1-70B are consistent with the results observed in the NLP domain experiments. We believe that future work could extend our research framework to other scientific domains.

5 Investigating Automated Evaluation for Ablation Study Design

💡 **RQ3:** How can future researchers develop more reliable and effective automated evaluation systems for complex scientific tasks?

As discussed in Section 4.2, we observe a significant discrepancy between automated and human evaluation results when assessing LLM performance on ABGEN. To investigate this issue further, we conduct a systematic meta-evaluation of commonly used automated evaluation systems.

5.1 ABGEN-EVAL Benchmark

We construct the meta-evaluation benchmark, ABGEN-EVAL, based on the human assessments results collected in Section 4. ABGEN-EVAL comprises 18 LLM outputs $\times$ 100 human assessments $= 1,800$ examples. Each example includes an LLM-generated ablation study design and three human scores assessing the study's importance, faithfulness, and soundness, respectively (detailed in §3.2). In line with previous meta-evaluation studies (Fabbri et al., 2021; Chen et al., 2021; Liu et al., 2024), in ABGEN-EVAL, the human evaluation results on the system-generated ablation study is considered the gold standard. The performance of automated evaluation systems is measured by the **system-level** and **instance-level** correlation between scores of human evaluation and automated evaluation systems. Specifically, given n input scientific papers and m ablation study generation systems, the human evaluation and an automatic metric result in two n-row, m-column score matrices H, M respectively. The *system*-level correlation is calculated on the aggregated system scores:

$$r_{\text{sys}}(H, M) = \mathcal{C}(\bar{H}, \bar{M}), \quad (2)$$

where $\bar{H}$ and $\bar{M}$ contain m entries which are the average system scores across n data samples (e.g., $\bar{H}_0 = \sum_i H_{i,0}/n$), and $\mathcal{C}$ is a function calculating a correlation coefficient (e.g., the Pearson's correlation coefficient). In contrast, the *instance*-level correlation is an average of sample-wise correlations:

$$r_{\text{sum}}(H, M) = \frac{\sum_i \mathcal{C}(H_i, M_i)}{n}, \quad (3)$$

where H_i, M_i are the evaluation results on the i-th data sample.

5.2 Experiments

For the two LLM-based evaluation systems, we developed multiple variants to investigate how different factors influence their effectiveness. These factors include: (1) whether reference ablation studies are included in the prompt, (2) the use of CoT reasoning, (3) the choice of base LLMs, ranging from open-source to proprietary models, and (4) whether evaluation is based on specific criteria or overall scores.

As illustrated in **??**, the current automated evaluation systems show relatively low correlations, indicating that they are not reliable for assessing generated ablation study designs. The reference-based approach consistently outperforms the reference-free approach. Additionally, using CoT prompting can further improve the effectiveness of LLM-based evaluations.

6 Related Work

LLMs have been employed for different scientific tasks for enhancing researchers' scientific workflows, such as conducting literature reviews (Wang et al., 2024b; Agarwal et al., 2024), peer-review and meta-review generation (D'Arcy et al., 2024; Tan et al., 2024; Wu et al., 2022; Zhou et al., 2024a), question answering over scientific papers (Dasigi et al., 2021; Saikh et al., 2022; Lee et al., 2023), scientific paper writing (Xu et al., 2024) and research hypothesis generation (Wang et al., 2024a; Zhou et al., 2024b). However, the potential of LLMs to effectively assist scientists in the experimental design process remains largely unexplored. Additionally, the challenge of developing effective and reliable automated evaluation systems for complex scientific tasks is still an open question.

7 Conclusion

This paper introduces AbGen, the first benchmark designed to evaluate LLMs in generating ablation studies for scientific research. Through a comprehensive assessment, we highlight both the strengths and limitations of leading LLMs on AbGen, providing valuable insights for future advancements. Our findings offer practical guidance on how to apply this research in real-world scenarios, ultimately aiding human researchers. Additionally, we identify a discrepancy between automated evaluations and human assessments in our task. To investigate this, we also develop a meta-evaluation benchmark, providing insights into developing more reliable automated evaluation for complex scientific tasks.

Acknowledgments

This project is supported by Tata Sons Private Limited, Tata Consultancy Services Limited, and Titan. We are grateful to Nvidia Academic Grant Program for providing computing resources.

Limitations

In this work, we introduce AbGen and perform a comprehensive analysis of various LLMs' capabilities on the proposed new task. However, there are still some limitations:

First, this study does not explore advanced prompting techniques (Yao et al., 2023; Wang et al., 2024a) or LLM-Agent-based methods (D'Arcy et al., 2024; Majumder et al., 2024). Our focus is on assessing the fundamental capabilities of leading LLMs in ablation study design. The goal is to provide insights into their strengths and limitations, laying the groundwork for future advancements. We encourage researchers to build upon our benchmark and findings to develop more advanced approaches for this task.

Second, as shown in our results on AbGen-Eval, the reported automated evaluation scores are not yet perfect. To support further research, we will make all model outputs from Section 4 publicly available. This will enable other researchers to conduct different automated evaluations and ensure consistent rankings by re-running their assessments on our model outputs. Additionally, our human evaluation protocol is designed to minimize the need for repeated human evaluations by future researchers. By strictly adhering to our assessment guidelines, researchers can reliably assess and compare their methods with existing approaches in an independent and consistent manner.

Lastly, we do not evaluate LLMs specialized in scientific domains (Groeneveld et al., 2024; Wadden et al., 2024), as our preliminary experiments find that they struggle to follow the instructions to generate ablation studies. However, we believe that training LLMs on scientific data (Lo et al., 2020) could enhance their performance on our task.

References

Shubham Agarwal, Issam H. Laradji, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review.

AI@Meta. 2024. The llama 3 herd of models.

Signe Altmäe, Alberto Sola-Leyva, and Andres Salumets. 2023. Artificial intelligence in scientific writing: a friend or a foe? *Reproductive BioMedicine Online*, 47(1):3–9.

Yiran Chen, Pengfei Liu, and Xipeng Qiu. 2021. Are factuality checkers reliable? adversarial meta-evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2082–2095, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. *arXiv preprint arXiv:2406.16253*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee, and Moontae Lee. 2023. QASA: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052. PMLR.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Yuliang Liu, Xiangru Tang, Zefan Cai, Junjie Lu, Yichi Zhang, Yanjun Shao, Zexuan Deng, Helan Hu, Zengxian Yang, Kaikai An, et al. 2023. Ml-bench: Large language models leverage open-source libraries for machine learning tasks. *arXiv preprint arXiv:2311.09835*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. 2024. Data-driven discovery with large generative models.

Meta AI. 2025. Llama 4: Natively multimodal mixture-of-experts language model.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2025a. Addendum to openai o3 and o4-mini system card: Openai o3 operator.

OpenAI. 2025b. Introducing gpt-4.1 in the api.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23:289 – 301.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. Peer review as a multi-turn and long-context dialogue with role-based interactions.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Qwen Team. 2025. Qwen3 technical report.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2024. Sciriff: A resource to enhance language model instruction-following over scientific literature.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty.

Xintao Wang, Jiangjie Chen, Nianqi Li, Lida Chen, Xinfeng Yuan, Wei Shi, Xuyang Ge, Rui Xu, and Yanghua Xiao. 2024b. Surveyagent: A conversational system for personalized and efficient research survey. *arXiv preprint arXiv:2404.06364*.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. Kiwi: A dataset of knowledge-intensive writing instructions for answering research questions. *arXiv preprint arXiv:2403.03866*.

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. 2024a. Qwen2 technical report.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024b. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024a. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *International Conference on Language Resources and Evaluation*.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024b. Hypothesis generation with large language models.

A Appendix

A.1 ABGEN Benchmark

Annotation Quality	%S $\geq$ 4
Research Context	
Correctly structured	99.0
Excluding ablation-relevant content	96.5
Reference Ablation Study	
Correctly structured	98.5
Non-overlapping	96.0
Justifiable within research context	97.5

Table 5: Human evaluation over 200 samples of ABGEN. Three internal evaluators were asked to rate the samples on a scale of 1 to 5 individually. We report percent of samples that have an average score $\geq$ 4 to indicate the annotation quality of ABGEN.

A.2 Experiment Setup

User Study Prompt

[System Input]:

Revise or rewrite the initial generation based on research context and user feedback.

[User Input]:

Research context: {research context}
Initial generation: {initial generation}
User feedback: {user feedback}

Redesign an ablation study about the {ablation module}, according to user feedback ...

Figure 4: Prompt for LLM-researcher interaction.

ID	# NLP/AI Publication	Data Annotation	Data Validation	Human Evaluation	Human Performance
1	> 10	✓	✓		✓
2	> 10			✓	
3	> 10			✓	
4	5-10	✓	✓		✓
5	1-5	✓		✓	
6	1-5	✓	✓	✓	

Table 6: Details of annotators involved in dataset construction and LLM performance evaluation. ABGEN is annotated by experts in NLP domains, ensuring both the accuracy of the benchmark and the reliability of the human evaluation.

Organization	Model	Release	Version	Context Window
		Proprietary Models		
OpenAI	o4-mini	2025-4	`o4-mini-2025-04-16`	–
	GPT-4.1	2025-4	`gpt-4.1-2025-04-14`	–
	GPT-4o	2024-8	`gpt-4o-2024-08-06`	–
Google	Gemini-2.5-Flash	2024-5	`gemini-2.5-flash-preview-05-20`	–
		Open-source Multimodal Foundation Models		
Mistral AI	Mistral-Small-3.1	2025-3	`Mistral-Small-3.1-24B`	128k
Microsoft	Phi-4	2025-3	`Phi-4`	16k
Google	Gemma-3-27b-it	2025-3	`gemma-3-27b-it`	16k
DeepSeek	DeepSeekV3	2024-12	`DeepSeekV3`	160k
	DeepSeekR1	2025-5	`DeepSeek-R1-0528`	160k
	DeepSeek-R1-0528-Qwen3-8B,	2025-5	`DeepSeek-R1-0528-Qwen3-8B`	160k
Alibaba	Qwen2.5-32B	2025-1	`Qwen2.5-32B-Instruct`	32k
	Qwen3-8B	2025-5	`Qwen3-8B`	40k
	Qwen3-32B	2025-5	`Qwen3-32B`	40k
	Qwen3-235BA22B	2025-5	`Qwen3-235B-A22B`	32k
Meta	Llama-3.1-70B	2024-6	`Llama-3.1-70B-Instruct`	32k
	Llama-3.3-70B	2025-5	`Llama-3.3-70B-Instruct`	32k
	Llama-4-Scout-17B	2025-5	`Llama-4-Scout-17B-Instruct`	32k
	Llama-4-Maverick-17B	2025-5	`Llama-4-Maverick-17B-Instruct`	32k

Table 7: Details of the organization, release time, maximum context length, and model source (*i.e.,* url for proprietary models and Huggingface model name for open-source models) for the LLMs evaluated in ABGEN.