AutoSurvey: Large Language Models Can Automatically Write Surveys

Yidong Wang[1,2*], **Qi Guo**[2,3*],
Wenjin Yao[2], **Hongbo Zhang**[1], **Xin Zhang**[4], **Zhen Wu**[3],**Meishan Zhang**[4],
Xinyu Dai[3], **Min Zhang**[4], **Qingsong Wen**[5], **Wei Ye**[2†], **Shikun Zhang**[2†], **Yue Zhang**[1†]

[1]Westlake University, [2]Peking University,
[3]Nanjing University, [4]Harbin Institute of Technology, Shenzhen, [5]Squirrel AI

Abstract

This paper introduces AutoSurvey, a speedy and well-organized methodology for automating the creation of comprehensive literature surveys in rapidly evolving fields like artificial intelligence. Traditional survey paper creation faces challenges due to the vast volume and complexity of information, prompting the need for efficient survey methods. While large language models (LLMs) offer promise in automating this process, challenges such as context window limitations, parametric knowledge constraints, and the lack of evaluation benchmarks remain. AutoSurvey addresses these challenges through a systematic approach that involves initial retrieval and outline generation, subsection drafting by specialized LLMs, integration and refinement, and rigorous evaluation and iteration. Our contributions include a comprehensive solution to the survey problem, a reliable evaluation method, and experimental validation demonstrating AutoSurvey's effectiveness.

1 Introduction

Survey papers provide essential academic resources, offering comprehensive overviews of recent research developments, highlighting ongoing trends, and identifying future directions [1, 2, 3, 4]. However, crafting these surveys is increasingly challenging, especially in the fast-paced domain of Artificial Intelligence including large language models(LLMs) [5, 6, 7, 8]. Figure 1a illustrates a significant trend: in just the first four months of 2024 alone, over 4,000 papers containing the phrase "Large Language Model" in their titles or abstracts were submitted to arXiv. This surge highlights a critical academic issue: the rapid accumulation of new information often outpaces the capacity for comprehensive scholarly review and synthesis, emphasizing the growing need for more efficient methods to synthesize the expanding literature. Moreover, as depicted in Figure 1b, while the number of survey papers has rapidly increased, the growing difficulty of producing traditional human-authored survey papers—due to the sheer volume and complexity of data—remains a significant challenge. This challenge is evidenced by the lack of comprehensive surveys in many fields (Figure 1c), which hinders knowledge transfer and makes it difficult for new researchers to efficiently navigate the vast amount of available information.

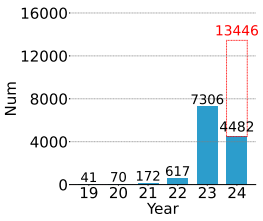
The advent of LLMs [7, 9] presents a promising avenue for addressing these challenges. These models, trained on extensive text corpora, demonstrate remarkable capabilities in understanding and generating human-like text, even in long-context scenarios [10, 11, 12]. Despite these advancements, the practical application of LLMs to survey generation is fraught with challenges. Firstly, **context**

*Equal contribution. yidongwang37@gmail.com, qguo@smail.nju.edu.cn; Yidong Wang did this work during his internship at Squirrel AI.

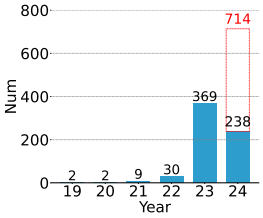
†Correspondence to: wye@pku.edu.cn, zhangsk@pku.edu.cn, zhangyue@westlake.edu.cn.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

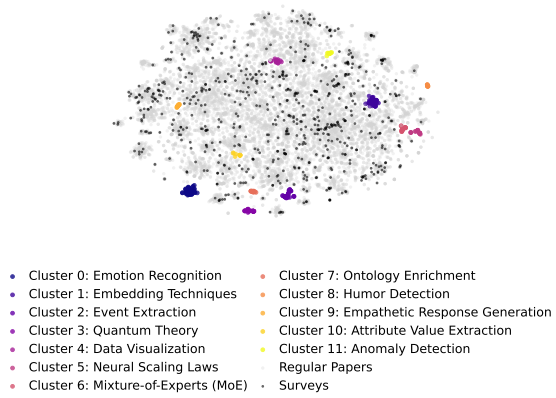
 https://doi.org/10.52202/079017-3655



(a) #Paper about LLM.



(b) #Survey about LLM.



(c) T-SNE visualization of surveys and papers about LLMs. Clusters represent groups of papers identified through clustering, which currently lack comprehensive survey coverage.

Figure 1: Depicting growth trends from 2019 to 2024 in the number of LLMs-related papers (a) and surveys (b) on arXiv, accompanied by a T-SNE visualization. The data for 2024 is up to April, with a red bar representing the forecasted numbers for the entire year. While the number of surveys is increasing rapidly, the visualization reveals areas where comprehensive surveys are still lacking, despite the overall growth in survey numbers. The research topics of the clusters in the T-SNE plot are generated using GPT-4 to describe their primary focus areas. **These clusters of research voids can be addressed using AutoSurvey at a cost of $1.2 (cost analysis in Appendix D) and 3 minutes per survey. An example survey focused on Emotion Recognition using LLMs is in Appendix F.**

window limitations: LLMs encounter inherent restrictions in output length due to limited processing windows [13, 14, 15, 16, 17]. While several advanced large models, including GPT-4 and Claude 3, support inputs exceeding 100k tokens, their output is still limited to fewer than 8k tokens (the output length of GPT-4 is 8k, and the output length of Claude 3 is 4k). Writing a comprehensive survey typically requires reading hundreds of papers, resulting in input sizes far beyond the capacity of even the most advanced models. Moreover, a well-written survey itself spans tens of thousands of tokens, making it highly challenging to generate such extensive content directly with large models. Secondly, **parametric knowledge constraints**: Sole reliance on an LLM's internal knowledge is insufficient for producing surveys that require comprehensive and accurate references [18, 19, 20]. LLMs may generate content based on inaccuracies or even non-existent "hallucinated" references. Moreover, these models cannot incorporate the latest studies not included in their training data, which limits the breadth and depth of the surveys they generate. Thirdly, **the lack of evaluation benchmark:** after production, reliable metrics to evaluate the quality of outputs from LLMs are lacking. Relying on human review for quality assessment is not only resource-intensive but also lacks scalability [21, 22, 23]. This presents a significant obstacle to the widespread adoption of LLMs for academic synthesis, where rigorous standards of accuracy and reliability are paramount.

In response to these challenges, we introduce AutoSurvey: a speedy and well-organized methodology for conducting comprehensive literature surveys. Specifically, AutoSurvey's primary innovations include: **logical parallel generation**: AutoSurvey employs a two-stage generation approach to parallelly generate survey content efficiently. Initially, multiple LLMs work concurrently to create detailed outlines. A final, comprehensive outline is then synthesized from these individual outlines, setting a clear framework for content development. Subsequently, each subsection of the survey is generated in parallel and guided by the outline, which significantly accelerates the process. To overcome potential transition and consistency issues due to segmented generation phases, AutoSurvey integrates a systematic revision phase. After the initial parallel generation, each section undergoes thorough revision and polishing, ensuring smooth transitions and enhanced overall document consistency. The sections are then seamlessly merged to produce a cohesive and well-organized final document. **Real-time knowledge update**: AutoSurvey incorporates a Real-time Knowledge Update mechanism using a Retrieval-Augmented Generation (RAG) approach [24, 25, 26]. This feature ensures that every aspect of the survey reflects the most current studies. When a survey topic is input by the user,

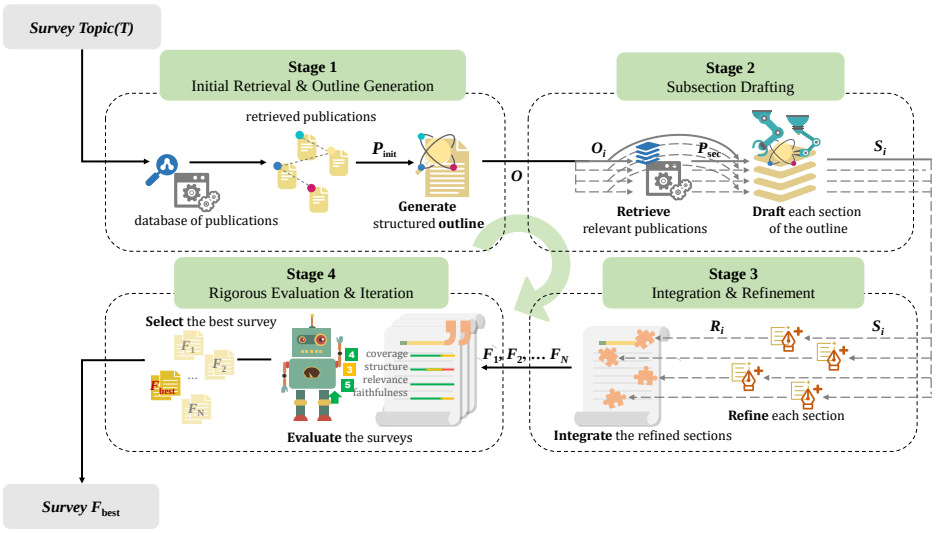


Figure 2: The AutoSurvey Pipeline for Generating Comprehensive Surveys.

AutoSurvey leverages the RAG system to retrieve the latest relevant papers, forming the basis for generating a structured and informed outline. During subsection writing, the system dynamically pulls in new research articles relevant to the specific content under development. This approach ensures that citations are current and the survey content is aligned with the latest developments in the field, significantly enhancing the accuracy and depth of the literature review. **Multi-LLM-as-judge evaluation**: AutoSurvey employs the Multi-LLM-as-Judge strategy, leveraging the LLM-as-Judge method for text evaluation [22, 21, 23]. This approach generates initial evaluation metrics using multiple large language models, which process a substantial corpus of high-quality surveys. These metrics are refined by human experts to ensure precision and adherence to academic standards. The Multi-LLM-as-Judge method assesses generated content across two main dimensions: (1) Citation Quality, verifying the accuracy and reliability of the information presented, with sub-indicators for Recall and Precision. (2) Content Quality, consisting of Coverage (assessing the extent of topic encapsulation), Structure (evaluating logical organization and coherence), and Relevance (ensuring alignment with the main topic). By utilizing multiple LLMs, this strategy minimizes bias and ensures a balanced and comprehensive assessment, upholding rigorous academic standards.

Extensive experimental results across different survey lengths (8k, 16k, 32k, and 64k tokens) demonstrate that AutoSurvey consistently achieves high citation and content quality scores. At 64k tokens, AutoSurvey achieves 82.25% recall and 77.41% precision in citation quality, outperforming naive RAG-based LLMs (68.79% recall and 61.97% precision) and approaching human performance (86.33% recall and 77.78% precision). In content quality at 64k tokens, AutoSurvey scores 4.73 in coverage, 4.33 in structure, and 4.86 in relevance, closely aligning with human performance (5.00, 4.66, and 5.00 respectively). At shorter lengths (8k, 16k, and 32k tokens), AutoSurvey also maintains strong performance across all metrics. Furthermore, the Spearman's rho values indicate a moderate positive correlation between the rankings provided by the LLMs and those given by human experts. The mixture of models achieves the highest correlation at 0.5429, indicating a strong alignment with human preferences. These results reinforce the effectiveness of our multi-LLM scoring mechanism, providing a reliable proxy for human judgment across varying survey lengths.

In conclusion, to the best of our knowledge, AutoSurvey is the first system to explore the potential of large model agents in writing extensive academic surveys. It proposes evaluation criteria for surveys that align with human preferences, providing a valuable reference for future related research.

2 Methodology

In this section, we describe the methodology employed by AutoSurvey to automate the creation of comprehensive literature surveys. Our approach systematically progresses through four distinct phases—Initial Retrieval and Outline Generation, Subsection Drafting, Integration and Refinement,

https://doi.org/10.52202/079017-3655

and Rigorous Evaluation and Iteration. Each phase is meticulously designed to address specific challenges associated with survey creation, thereby enhancing the efficiency and quality of the resulting survey document. The pseudo code of AutoSurvey can be found at Algorithm 1.

Algorithm 1 AUTOSURVEY: Automated Survey Creation Using LLMs.

Input: Survey topic T, publications database D
Output: Final refined and evaluated survey document F_{best}
for each survey generation trial $t = 1$ to N **do**
  Phase 1: Initial Retrieval and Outline Generation
  Retrieve initial pool of publications $P_{\text{init}} \leftarrow \text{Retrieve}(T, D)$
  Generate outline $O \leftarrow \text{Outline}(T, P_{\text{init}})$
  Phase 2: Subsection Drafting
  for each section O_i in O **in parallel do**
    Retrieve relevant publications $P_{\text{sec}} \leftarrow \text{Retrieve}(O_i, D)$
    Draft subsection $S_i \leftarrow \text{Draft}(O_i, P_{\text{sec}})$
  end for
  Phase 3: Integration and Refinement
  Refine the merged document to improve coherence $R_i \leftarrow \text{Refine}(S_i)$
  Merge subsection drafts into a single document $F_t \leftarrow \text{Merge}(R_1, R_2, \ldots, R_n)$
end for
Phase 4: Rigorous Evaluation and Iteration
Evaluate and select the best survey document $F_{\text{best}} \leftarrow \text{Evaluate}(F_1, F_2, \ldots, F_N)$
Return: Refined and evaluated survey F_{best}

Initial Retrieval and Outline Generation The process begins with the Initial Retrieval and Outline Generation phase. Utilizing an embedding-based retrieval technique, AutoSurvey scans a database of publications to identify papers most pertinent to the specified survey topic T. This phase is crucial for ensuring that the survey is grounded in the most relevant and recent research. The retrieved publications P_{init} are then used to generate a structured outline O, which ensures comprehensive coverage of the topic and logical structuring of the survey. To provide more detailed guidance for writing subsections, the outline generation includes not only titles for each subsection but also brief descriptions. These descriptions convey the main idea of each subsection, aiding in the overall clarity and direction of the survey. Given the extensive amount of relevant papers extracted during this stage, the total length of P_{init} often exceeds the context window size of the LLM. To address this, papers are randomly divided according to the LLM's context window size, resulting in the creation of multiple outlines. The model then consolidates these outlines to form the final comprehensive outline. Finally, the outline O of the entire survey is represented as $O = \text{Outline}(T, P_{\text{init}})$.

Subsection Drafting With the structured outline in place, the Subsection Drafting phase commences. During this phase, specialized LLMs draft each section of the outline in parallel. This method not only accelerates the drafting process but also ensures detailed and focused content generation for each survey section, adhering to the thematic boundaries established by the outline. When writing the content of each subsection, the sub-outline O_i of that subsection will be used to retrieve the necessary relevant reference papers P_{sec} to provide information that aligns more closely with the main idea of the subsection. During the writing process, the model is required to cite the provided reference papers to support the generated content. The references in the generated content will be extracted and mapped to the corresponding arXiv papers (see Appendix B for details). The i_{th} subsection S_i can be expressed as: $S_i = \text{Draft}(O_i, P_{\text{sec}})$.

Integration and Refinement Following the drafting phase, each section S_i is individually refined to enhance readability, eliminate redundancies, and ensure a seamless narrative. The refined sections R_i are then merged into a cohesive document F, which is essential for maintaining a logical flow and coherence throughout the survey. During the refinement process, the model needs to polish each subsection based on the local context (considering the previous and following subsections) to improve readability, eliminate redundancies, and enhance coherency. Additionally, the model is required to check the correctness of the cited references in the content and correct any errors in the citations. This procedure can be represented by: $F = \text{Merge}(R_1, R_2, \ldots, R_n)$, where $R_i = \text{Refine}(S_i)$.

Rigorous Evaluation and Iteration The final phase involves a rigorous evaluation and iteration process, where the survey document is assessed through a Multi-LLM-as-Judge strategy. This evaluation critically examines the survey in several aspects. The insights gained from this evaluation are used to guide further refinements, ensuring the survey meets the highest academic standards. The best survey is chosen from N candidates. The final output of AutoSurvey is $F_{\text{best}} = \text{Evaluate}(\{F_1, F_2, \ldots, F_N\})$.

The methodology outlined here—from initial data retrieval to sophisticated multi-faceted evaluation—ensures that AutoSurvey effectively addresses the complexities of survey creation in evolving research fields using advanced LLM technologies.

3 Experiments

Setup We conduct comprehensive experiments to evaluate the performance of AutoSurvey, comparing it against traditional methods for generating survey papers. For the drafting phase of AutoSurvey, we utilize Claude-3-Haiku, known for its speed and cost-effectiveness, capable of handling 200K tokens. For evaluations, we employ a combination of GPT-4, Claude-3-Haiku, and Gemini-1.5-Pro[3]. The evaluation covers the following key performance metrics:

- **Survey Creation Speed**: To estimate the time it takes for humans to write a document, we use a mathematical model with the following parameters: L (the length of the document), E (the number of experts), M (the writing speed of each expert), T_r (the preparation time for research and data collection), T_w (the actual writing time, $T_w = \frac{L}{E \times M}$), and T_e (the editing and revision time, $T_e = \frac{1}{2}T_w$). Assuming an ideal situation where $E = 10$, $M = 2000$ tokens/hour, $T_r = 5$ hours, and $T_e = \frac{1}{2}T_w$, the total time $Time$ is calculated as:

$$Time = T_r + T_w + T_e = T_r + \frac{L}{E \times M} + \frac{1}{2} \times \frac{L}{E \times M}. \tag{1}$$

  For Naive RAG and AutoSurvey, we count all the time of API calls. The speed is calculated as $Speed = \frac{1}{Time(\text{hours})}$.

- **Citation Quality**: Adopted from [27], this metric assesses the accuracy and relevance of citations in the survey. Assuming a set of claims $C = \{c_1, c_2, \ldots\}$ extracted from the survey, the metric utilizes an NLI model h to decide whether a claim c_i is supported by its references $\text{Ref}_i = \{r_{i_1}, r_{i_2}, \ldots\}$, where each r_{i_k} represents one paper cited. $h(c_i, \text{Ref}_i) = 1$ means that the references can support the claim, and $h(c_i, \text{Ref}_i) = 0$ otherwise. Refer to Appendix C for more details. Citation quality encompasses two sub-metrics:
  - **Citation Recall**: Measures whether all statements in the generated text are fully supported by the cited passages, which is calculated as

$$\text{Recall} = \frac{\sum_{i=1}^{|C|} h(c_i, \text{Ref}_i)}{|C|}. \tag{2}$$

  - **Citation Precision**: Identifies irrelevant citations, ensuring that the provided citations are pertinent and directly support the statements. Before listing the formula for precision, a function g is defined as $g(c_i, r_{i_k}) = (h(c_i, \{r_{i_k}\}) = 1) \cup (h(c_i, \text{Ref}_i \setminus \{r_{i_k}\}) = 0)$, which measures whether the paper r_{i_k} is related to the claim c_i. The precision is

$$\text{Precision} = \frac{\sum_{i=1}^{|C|} \sum_{k=1}^{|\text{Ref}_i|} h(c_i, \text{Ref}_i) \cap g(c_i, r_{i_k})}{\sum_{i=1}^{|C|} |\text{Ref}_i|}. \tag{3}$$

- **Content Quality**: An overarching metric evaluating the excellence of the written survey, encompassing three sub-indicators. Each sub-indicator is judged by LLMs according to a 5-point rubric, calibrated by human experts to meet academic standards. Note that the detailed scoring criteria are provided in Table 1.
  - **Coverage**: Assesses the extent to which the survey encapsulates all aspects of the topic.
  - **Structure**: Evaluates the logical organization and coherence of each section.
  - **Relevance**: Measures how well the content aligns with the research topic.

[3]Specifically, we use gpt-4-0125-preview, claude-3-haiku-20240307 and Gemini-1.5-pro-preview.

https://doi.org/10.52202/079017-3655

Table 1: Content Quality Criteria.

Criteria	Scores
Coverage	*Score 1*: The survey has very limited coverage, only touching on a small portion of the topic and lacking discussion on key areas. *Score 2*: The survey covers some parts of the topic but has noticeable omissions, with significant areas either underrepresented or missing. *Score 3*: The survey is generally comprehensive in coverage but still misses a few key points that are not fully discussed. *Score 4*: The survey covers most key areas of the topic comprehensively, with only very minor topics left out. *Score 5*: The survey comprehensively covers all key and peripheral topics, providing detailed discussions and extensive information.
Structure	*Score 1*: The survey lacks logic, with no clear connections between sections, making it difficult to understand the overall framework. *Score 2*: The survey has weak logical flow with some content arranged in a disordered or unreasonable manner. *Score 3*: The survey has a generally reasonable logical structure, with most content arranged orderly, though some links and transitions could be improved such as repeated subsections. *Score 4*: The survey has good logical consistency, with content well arranged and natural transitions, only slightly rigid in a few parts. *Score 5*: The survey is tightly structured and logically clear, with all sections and content arranged most reasonably, and transitions between adjacent sections smooth without redundancy.
Relevance	*Score 1*: The content is outdated or unrelated to the field it purports to review, offering no alignment with the topic. *Score 2*: The survey is somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to. *Score 3*: The survey is generally on topic, despite a few unrelated details. *Score 4*: The survey is mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions. *Score 5*: The survey is exceptionally focused and entirely on topic; the article is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.

Baselines We compare AutoSurvey with surveys authored by human experts (collected from Arxiv) and naive RAG across 20 different computer science topics across 20 different topics in the field of LLMs (see Table 7). For the naive RAG, we begin with a title and a survey length requirement, then iteratively prompt the model to write the content until completion. Note that we also provide the model with the same number of reference papers with AutoSurvey. To make a more comprehensive comparison, we additionally introduced two baselines: RAG+Reflection, which involves reflecting the generated survey content from RAG, and RAG+Query Rewriting, where the LLM reformulates the retrieval query based on the topic.

For AutoSurvey, we utilize a corpus of 530,000 computer science papers from arXiv as the retrieval database. During the initial drafting stage, we retrieve 1200 papers relevant to the given topic and split them into several chunks with a window size of 30,000 tokens. The model generates an outline for each chunk and merges these outlines into a final comprehensive outline, using only the abstracts of the papers at this stage. The outline predetermines the number of sections as 8. For subsection drafting, the models generate specific sections using the outline and 60 papers retrieved based on the subsection descriptions, focusing on the main body of each paper (up to the first 1,500 tokens). During the reflection and polishing stage, the same reference papers are provided to the model to ensure consistency and accuracy. The iteration number N is set to 2. Note that human writing surveys used for evaluation are excluded during the retrieval process. For more details of implementations, see Appendix B, and the prompts are presented in Appendix F.

Table 2: Results of Naive RAG, Human writing and AutoSurvey. Both of AutoSurvey and Naive RAG use Claude-haiku as the writer. **Note that human writing surveys used for evaluation are excluded during the retrieval process.**

Survey Length (#tokens)	Methods	Speed	Citation quality		Content Quality			Avg.
			Recall	Precision	Coverage	Structure	Relevance	
8k	Human writing	0.16	80.00	87.50	4.50	4.16	5.00	4.52
	Naive RAG	79.67	$78.14_{\pm 5.23}$	$71.92_{\pm 6.83}$	$4.40_{\pm 0.48}$	$3.86_{\pm 0.71}$	$4.86_{\pm 0.33}$	4.33
	Naive RAG+Reflection	-	$82.25_{\pm 2.89}$	$76.84_{\pm 2.59}$	$4.46_{\pm 0.37}$	$4.02_{\pm 0.49}$	$4.86_{\pm 0.41}$	4.42
	Naive RAG+Query Rewriting	-	$80.99_{\pm 3.43}$	$71.83_{\pm 3.59}$	$4.84_{\pm 0.23}$	$4.05_{\pm 0.57}$	$4.88_{\pm 0.19}$	4.56
	AutoSurvey	107.00	$82.48_{\pm 2.77}$	$77.42_{\pm 3.28}$	$4.60_{\pm 0.48}$	$4.46_{\pm 0.49}$	$4.8_{\pm 0.39}$	4.61
16k	Human writing	0.14	88.52	79.63	4.66	4.38	5.00	4.66
	Naive RAG	43.41	$71.48_{\pm 12.50}$	$65.31_{\pm 15.36}$	$4.46_{\pm 0.49}$	$3.66_{\pm 0.69}$	$4.73_{\pm 0.44}$	4.23
	Naive RAG+Reflection	-	$79.67_{\pm 2.94}$	$73.73_{\pm 2.32}$	$4.57_{\pm 0.45}$	$4.28_{\pm 0.59}$	$4.83_{\pm 0.23}$	4.55
	Naive RAG+Query Rewriting	-	$77.73_{\pm 3.86}$	$66.29_{\pm 4.56}$	$4.70_{\pm 0.31}$	$3.67_{\pm 0.63}$	$4.79_{\pm 0.37}$	4.32
	AutoSurvey	95.51	$81.34_{\pm 3.65}$	$76.94_{\pm 1.93}$	$4.66_{\pm 0.47}$	$4.33_{\pm 0.59}$	$4.86_{\pm 0.33}$	4.60
32k	Human writing	0.10	88.57	77.14	4.66	4.50	5.00	4.71
	Naive RAG	22.64	$79.88_{\pm 4.35}$	$65.03_{\pm 8.39}$	$4.41_{\pm 0.64}$	$3.75_{\pm 0.72}$	$4.66_{\pm 0.47}$	4.23
	Naive RAG+Reflection	-	$80.50_{\pm 3.66}$	$72.18_{\pm 3.31}$	$4.82_{\pm 0.17}$	$4.08_{\pm 0.61}$	$4.49_{\pm 0.44}$	4.44
	Naive RAG+Query Rewriting	-	$76.56_{\pm 4.38}$	$65.36_{\pm 4.92}$	$4.61_{\pm 0.33}$	$3.96_{\pm 0.65}$	$4.88_{\pm 0.21}$	4.45
	AutoSurvey	91.46	$83.14_{\pm 2.44}$	$78.04_{\pm 3.14}$	$4.73_{\pm 0.44}$	$4.26_{\pm 0.69}$	$4.8_{\pm 0.54}$	4.58
64k	Human writing	0.07	86.33	77.78	5.00	4.66	5.00	4.88
	Naive RAG	12.56	$68.79_{\pm 11.00}$	$61.97_{\pm 13.45}$	$4.4_{\pm 0.61}$	$3.66_{\pm 0.47}$	$4.66_{\pm 0.47}$	4.19
	Naive RAG+Reflection	-	$73.12_{\pm 2.49}$	$68.36_{\pm 3.65}$	$4.66_{\pm 0.31}$	$4.06_{\pm 0.53}$	$4.76_{\pm 0.21}$	4.47
	Naive RAG+Query Rewriting	-	$69.77_{\pm 5.24}$	$62.21_{\pm 6.73}$	$4.45_{\pm 0.33}$	$3.88_{\pm 0.57}$	$4.69_{\pm 0.29}$	4.31
	AutoSurvey	73.59	$82.25_{\pm 3.64}$	$77.41_{\pm 3.84}$	$4.73_{\pm 0.44}$	$4.33_{\pm 0.47}$	$4.86_{\pm 0.33}$	4.62

Main Results The results of our experiments comparing human writing, Naive RAG, and AutoSurvey for generating academic surveys are summarized in Table 2. The key findings are:

- *AutoSurvey significantly outperforms both human writing and Naive RAG in terms of speed.* For instance, AutoSurvey achieves a speed of 73.59 surveys per hour for a 64k-token survey, compared to 0.07 for human writing and 12.56 for Naive RAG, highlighting the larger gap in speed for longer context generation.
- *AutoSurvey demonstrates superior citation quality compared to other baselines, with performance close to human writing.* For an 8k-token survey, AutoSurvey achieves citation recall and precision scores of 82.48 and 77.42, respectively, surpassing Naive RAG (78.14 recall, 71.92 precision). While human writing achieves the best performance, ours is close to human's across different lengths. We also observe a significant decline in citation quality for other baselines as the survey length increased, whereas AutoSurvey maintains stable performance. We investigate this phenomenon in our ablation study.
- *AutoSurvey excels in content quality, scoring 4.60 on average for a 16k-token survey.* It achieves 4.66 for coverage, 4.33 for structure, and 4.86 for relevance, matching human writing and surpassing Naive RAG.

The experiments indicate that AutoSurvey provides a balanced trade-off between quality and efficiency. It achieves near-human levels of coverage, relevance, and citation quality while maintaining a significantly lower time cost. While human writing still leads in structure and overall quality, the efficiency and performance of AutoSurvey make it a compelling alternative for generating academic surveys. Naive RAG-based LLM, though effective, falls short in several key areas compared to both human writing and AutoSurvey, making it the least preferred method among the three for generating high-quality academic surveys, particularly in terms of structure, due to the lack of outline.

Meta Evaluation To verify the consistency between our evaluation method and human evaluation, we conduct a meta-evaluation involving human experts and our automated evaluation system. Human experts judge pairs of generated surveys to determine which one is superior. This process, referred to as a "which one is better" game, serves as the golden standard for evaluation. We compare the judgments made by our evaluation method against those made by human experts. Specifically, we provide the experts with the same scoring criteria used in our evaluation for reference. The experts rank the generated 20 surveys, and we compare these rankings with those generated by LLM using Spearman's rank correlation coefficient to measure consistency between human and LLM evaluations.

The results of this meta-evaluation are presented in Figure 3. The table shows the Spearman's rho values, indicating the degree of correlation between the rankings given by each LLM and the human experts. The Spearman's rho values indicate a moderate positive correlation between the rankings provided by the LLMs and those given by the human experts, with the mixture of models achieving the highest correlation at 0.5429. These results suggest that our evaluation method aligns well with human preferences, providing a reliable proxy for human judgment.

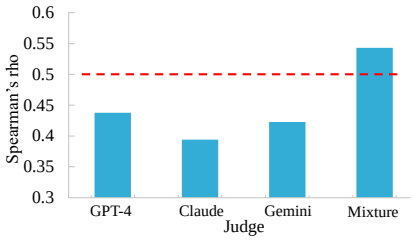


Figure 3: Spearman's rho values indicating the degree of correlation between rankings given by LLMs and human experts. *Note that A value over 0.3 indicates a positive correlation and over 0.5 indicates a strong positive correlation.*

Ablation study To assess the impact of various components on the performance of AutoSurvey, we conduct an ablation study by systematically removing key components: the retrieval mechanism, the reflection phase, and iterations. Additionally, we evaluate the influence of using different base LLMs to demonstrate that even with a less optimal LLM like Claude-haiku, AutoSurvey's performance remains comparable to human-generated surveys.

Table 3 demonstrates that removing the retrieval mechanism significantly degrades citation quality, highlighting its critical role in ensuring accurate and relevant references. The absence of the reflection phase slightly impacts the overall content quality, particularly in structure and coherence.

https://doi.org/10.52202/079017-3655

Table 3: Ablation study results for AutoSurvey with different components removed.

Methods	Citation Quality		Content Quality			
	Recall	Precision	Coverage	Structure	Relevance	Avg.
AutoSurvey	$83.48_{\pm 5.05}$	$77.15_{\pm 6.05}$	$4.7_{\pm 0.45}$	$4.16_{\pm 0.73}$	$4.93_{\pm 0.30}$	4.57
AutoSurvey w/o retrieve	$60.11_{\pm 6.42}$	$51.65_{\pm 6.33}$	$4.51_{\pm 0.49}$	$4.01_{\pm 0.74}$	$4.88_{\pm 0.32}$	4.44
AutoSurvey w/o reflection	$83.23_{\pm 3.82}$	$76.36_{\pm 4.08}$	$4.76_{\pm 0.42}$	$4.13_{\pm 0.76}$	$4.88_{\pm 0.32}$	4.56

Table 4: Performance of AutoSurvey with different base LLM writers.

Base LLM writer	Citation Quality		Content Quality			
	Recall	Precision	Coverage	Structure	Relevance	Avg.
GPT-4	$80.25_{\pm 4.19}$	$78.83_{\pm 7.00}$	$4.8_{\pm 0.54}$	$4.46_{\pm 0.49}$	$4.86_{\pm 0.33}$	4.70
Claude-haiku	$82.45_{\pm 2.77}$	$76.31_{\pm 2.18}$	$4.66_{\pm 0.47}$	$4.26_{\pm 0.67}$	$4.86_{\pm 0.33}$	4.58
Gemini-1.5-pro	$78.13_{\pm 2.39}$	$71.24_{\pm 3.28}$	$4.86_{\pm 0.33}$	$4.33_{\pm 0.78}$	$4.93_{\pm 0.25}$	4.69
Human	85.86	80.51	4.71	4.43	5	4.70

Table 6 shows the performance of AutoSurvey when using different LLMs as the base writer. The results indicate that all three LLMs (GPT-4, Claude-haiku, and Gemini-1.5-Pro) perform well, with GPT-4 slightly outperforming the others in terms of overall content quality. Importantly, even with the less optimal Claude-haiku, AutoSurvey's performance remains competitive with human standards.

Figure 4 presents the effect of different iteration counts on the performance of AutoSurvey. The results show that increasing the number of iterations from 1 to 5 leads to a slight improvement in overall content quality, with diminishing returns after the second iteration.

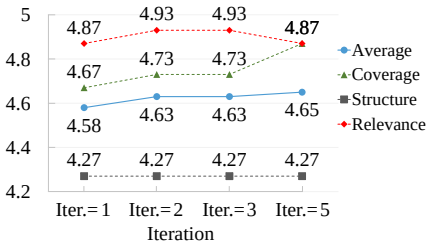


Figure 4: Impact of Iteration on AutoSurvey Performance.

To assess whether the generated survey can provide useful information to enrich the knowledge, we created 50 multiple-choice questions about 5 topics. These questions primarily involve knowledge related to literature, such as identifying which paper proposed a particular method. We compared the accuracy of the Claude model under the following conditions: (1) directly chooses the answer without providing any reference materials, (2) has access to a 32k survey generated by naive RAG-based LLMs, (3) has access to a 32k survey generated by AutoSurvey, and (4) can refer to 20 papers (30k tokens in total) retrieved using the options provided (Upper-bound, directly retrieving the answers).

The results are shown in Table 5 and we find providing topic-related materials can effectively improve the accuracy of answers. Providing option-related papers can be considered an upper bound for performance (73.60%). AutoSurvey improves accuracy by 9.2% compared to directly answering and is 2.4% higher than using naive RAG-based LLM-generated surveys. This demonstrates that our method can effectively provide topic knowledge.

Table 5: Performances given different references.

Methods	Accuracy
Direct	$58.40_{\pm 4.96}$
Naive RAG-based LLMs	$65.20_{\pm 8.06}$
Upper-bound	$73.60_{\pm 3.44}$
AutoSurvey	$67.60_{\pm 4.96}$

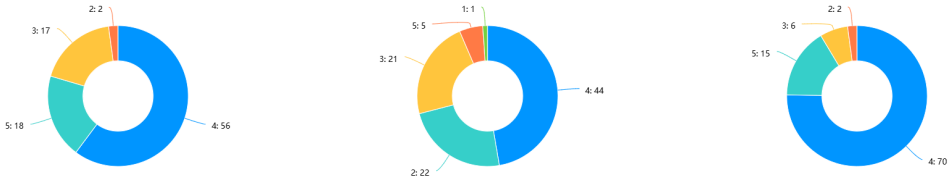
As mentioned in the main results, the naive RAG-based generation method demonstrates a notable decline in citation quality as the survey length increases. In contrast, AutoSurvey maintains stable citation quality across varying survey lengths. Such phenomena may be attributed to the streaming generation process, where each step must reference previous content, leading to the accumulation of errors [28]. To validate this, we segment the extracted claims into 20% intervals and calculate the citation recall for each segment. The results in Table 6 indicate that the recall of Naive RAG gradually decreases as the generated text length increases, while AutoSurvey maintains stable performance.

Table 6: Performance of AutoSurvey with different base LLM writers.

Method	20%	40%	60%	80%	100%
Naive RAG	76.79	73.17	71.52	64.08	49.85
AutoSurvey	82.86	84.89	79.04	82.27	82.29

In summary, the ablation study underscores the critical role of the retrieval mechanism and reflection phase in AutoSurvey. Furthermore, the performance is influenced by using different LLMs as the base writer and varying the iteration count. Nevertheless, AutoSurvey consistently performs well across various configurations, showcasing its robustness and efficiency.

User study To evaluate the real-world performance of AutoSurvey, we launch a free trial and distribute an anonymous survey to 141 users. The survey primarily assesses user perceptions of generation quality and whether the generated surveys contribute to their practical work. Users are asked to rate the relevance, structure, and usefulness to their actual work of the generated survey on a scale from 1 (poor) to 5 (excellent). A total of 93 valid responses are collected. Results reveal that users generally perceive AutoSurvey as relevant to topic, well structured, and useful, with most ratings skewed towards 4 and 5. Compared to other aspects, the proportion of scores below 5 in the "Structure" is higher, indicating that there remains room for improvement in further refining the structure. We are pleased to observe that the majority of users think the generated surveys beneficial to their practical work, indicating the utility and practical value of AutoSurvey.



(a) Relevance (b) Structure (c) Usefulness

Figure 5: Distribution of user ratings in terms of relevance, structure and usefulness.

4 Related Work

Long-form Text Generation The ability to effectively process and generate long-form text is a critical challenge for large language models (LLMs) due to the need to maintain coherence and logical flow over extended passages of text [29, 30, 31, 32]. Several works try to address the challenge by directly extending the context window with different Positional Encoding Techniques[33, 34]. However, modifying position encoding strategies requires retraining the model, which is costly. Another solution is using memory-augmented techniques. RecurrentGPT [35] enables the generation of arbitrarily long texts by simulating the recurrence mechanism of RNNs using natural language prompts to store previous contextual information. Temp-Lora [36] enables long text generation by embedding context information into a temporary Lora module updated progressively during generation rather than relying on an extensive context window. These methods effectively establish relationships among tokens and maintain contextual understanding, but still face the issue of long generation times. To further accelerate the generation process, Hierarchical Modeling Techniques have been explored extensively to capture the inherent hierarchical nature of long-form text [37, 38]. Despite such efficiency, it ignores the long dependency of text and may degrade the content quality [39]. To tackle the drawbacks, AutoSurvey, similarly using a Hierarchical generation paradigm, creates a well-organized outline for guidance and refines the generated content to improve the quality.

https://doi.org/10.52202/079017-3655

Automatic Writing Due to the high costs associated with manual writing, automated writing has attracted substantial research interest in recent years. Compared to traditional methods, which primarily focus on training models to generate linguistically coherent text [40, 41], the emergency of large language models (LLMs) has opened up new possibilities for automated writing, drawing more attention to broader aspects like faithfulness, logical structure, style, and ethics [42, 43, 44, 45]. For example, Retrieval-Augmented Generation techniques are useful for generating claims with citations [27, 46]. IRP framework [47] generates expository text by iteratively performing content planning, fact retrieval, and paraphrasing to ensure factuality and stylistic consistency. Several works focus on the outline creation to improve the structure of generated content. PaperRobot [48] incrementally writes key elements to generate a paper abstract. STORM [20] designs a refined outline based on multiple rounds of wiki-page-related Q&A to facilitate wiki-like article generations. These methods have only been explored in shorter texts (<4k). In contrast, Autosurvey shows its effectiveness in generating long content (64k), with a focus on academic reviews.

5 Limitation

In addition to directly using recall and precision to evaluate citations, we also perform a manual analysis, providing a more comprehensive view of the citation quality. We examine 100 unsupported claims and their corresponding references and find that the errors mainly fall into three categories: (1) Misalignment, (2) Misinterpretation, and (3) Overgeneralization. Misalignment occurs when the connection between them is incorrectly made, such as an irrelevant citation. Misinterpretation happens when the claim and source are related, but the claim incorrectly represents the information from the source. Overgeneralization occurs when a claim extends the conclusions of the source material to a broader context than is supported. Among the three types of errors, overgeneralization accounts for the largest proportion (51%), indicating that LLMs still rely heavily on their parametric knowledge for writing. Misinterpretation has a small proportion (10%), suggesting that LLMs are capable of understanding the content of the references in most cases, avoiding the creation of claims that significantly deviate from the references.

Misalignment (39%): An example is citing the "General Data Protection Regulation (GDPR)" in a context where the referenced paper does not propose GDPR but merely mentions it in the content.

Misinterpretation (10%): An example is claiming that "In-context learning allows LLMs to adapt to new tasks by simply conditioning on a few demonstration examples, without the need for any parameter updates or fine-tuning," based on a paper that focuses on meta-out-of-context learning and mentions the limitations of in-context learning.

Overgeneralization (51%): An example is that "in-context learning can also benefit from advancements in other learning paradigms, such as multi-task learning," based on a paper that discusses multi-task few-shot learning but does not explicitly address its influence on in-context learning.

Among the three types of errors, overgeneralization accounts for the largest proportion (51%), indicating that LLMs still rely heavily on their parametric knowledge for writing. Misinterpretation has a small proportion (10%), suggesting that LLMs are capable of understanding the content of the references in most cases, avoiding the creation of claims that significantly deviate from the references. Additional potential societal impact and ethical considerations are discussed in Appendix E.

6 Conclusion

In this paper, we introduce AutoSurvey, a novel methodology leveraging large language models to automate the creation of comprehensive literature surveys. AutoSurvey addresses key challenges such as context window limitations and parametric knowledge constraints through a systematic approach involving initial retrieval, outline generation, parallel subsection drafting, integration, and rigorous evaluation. Our experiments show that AutoSurvey significantly outperforms Naive RAG and matches human performance in content and citation quality, while also being highly efficient. This advancement offers a scalable and effective solution for synthesizing research literature, providing a valuable tool for researchers in rapidly evolving fields like artificial intelligence.

References

[1] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5):1–36, 2018.

[2] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 2023.

[3] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. arXiv preprint arXiv:2303.18223, 2023.

[4] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. ACM computing surveys (CSUR), 54(10s):1–41, 2022.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[7] 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.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023.

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[10] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595, 2023.

[11] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. In The Twelfth International Conference on Learning Representations, 2023.

[12] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. Advances in Neural Information Processing Systems, 36, 2024.

[13] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024.

[14] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. arXiv preprint arXiv:2307.10169, 2023.

[15] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In International Conference on Machine Learning, pages 31210–31227. PMLR, 2023.

[16] Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How long can context length of open-source llms truly promise? In NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following, 2023.

 https://doi.org/10.52202/079017-3655

[17] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. arXiv preprint arXiv:2404.02060, 2024.

[18] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv preprint arXiv:2310.07521, 2023.

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

[20] 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. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2024.

[21] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. 2024.

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

[23] Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. Kieval: A knowledge-grounded interactive evaluation framework for large language models. 2024.

[24] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2023.

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

[26] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7969–7992, 2023.

[27] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[28] Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. Long text generation by modeling sentence-level and discourse-level coherence. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), August 2021.

[29] Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric P. Xing, and Zhiting Hu. Progressive generation of long text with pretrained language models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 4313–4324. Association for Computational Linguistics, 2021.

[30] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. CoRR, abs/2308.14508, 2023.

[31] Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. CoRR, abs/2309.13345, 2023.

[32] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? CoRR, abs/2311.04939, 2023.

[33] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pages 464–468. Association for Computational Linguistics, 2018.

[34] Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. Self-attention with structural position representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 1403–1409. Association for Computational Linguistics, 2019.

[35] Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. Recurrentgpt: Interactive generation of (arbitrarily) long text. arXiv preprint arXiv:2305.13304, 2023.

[36] Y Wang, D Ma, and D Cai. With greater text comes greater necessity: Inference-time training helps long text generation. arXiv preprint arXiv:2401.11504, 2024.

[37] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. arXiv preprint arXiv:1805.04833, 2018.

[38] Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. arXiv preprint arXiv:2109.10862, 2021.

[39] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Booookscore: A systematic exploration of book-length summarization in the era of llms. arXiv preprint arXiv:2310.00785, 2023.

[40] Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. Towards coherent and cohesive long-form text generation. arXiv preprint arXiv:1811.00511, 2018.

[41] Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. Discourse-aware neural rewards for coherent text generation. arXiv preprint arXiv:1805.03766, 2018.

[42] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. arXiv preprint arXiv:2303.11315, 2023.

[43] Jinxin Liu, Shulin Cao, Jiaxin Shi, Tingjian Zhang, Lei Hou, and Juanzi Li. Probing structured semantics understanding and generation of language models via question answering. arXiv preprint arXiv:2401.05777, 2024.

[44] Chiyu Zhang, Honglong Cai, Yuexin Wu, Le Hou, Muhammad Abdul-Mageed, et al. Distilling text style transfer with self-explanation from llms. arXiv preprint arXiv:2403.01106, 2024.

[45] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence, 4(3):258–268, 2022.

[46] Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. arXiv preprint arXiv:2203.11147, 2022.

 https://doi.org/10.52202/079017-3655

[47] Nishant Balepur, Jie Huang, and Kevin Chen-Chuan Chang. Expository text generation: Imitate, retrieve, paraphrase. arXiv preprint arXiv:2305.03276, 2023.

[48] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. Paperrobot: Incremental draft generation of scientific ideas. arXiv preprint arXiv:1905.07870, 2019.

[49] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2024.

A Detail of Topics and Human-writing Surveys

We select 20 surveys from different topics within the LLM field. During the selection process, we prioritize both the breadth of the topics and the citation count (from google scholar) of the surveys. The basic information of surveys are listed in Table 7.

Table 7: Survey Table

Topic	Survey Title	Citations
In-context Learning	A survey for in-context learning	323
LLMs for Recommendation	A Survey on Large Language Models for Recommendation	55
LLM-Generated Texts Detection	A Survey of Detecting LLM-Generated Texts	42
Explainability for LLMs	Explainability for Large Language Models	25
Evaluation of LLMs	A Survey on Evaluation of Large Language Models	183
LLMs-based Agents	A Survey on Large Language Model based Autonomous Agents	101
LLMs in Medicine	A Survey of Large Language Models in Medicine	234
Domain Specialization of LLMs	Domain Specialization as the Key to Make Large Language Models Disruptive	14
Challenges of LLMs in Education	Practical and Ethical Challenges of Large Language Models in Education	53
Alignment of LLMs	Aligning Large Language Models with Human	53
ChatGPT	A Survey on ChatGPT and Beyond	144
Instruction Tuning for LLMs	Instruction Tuning for Large Language Models	45
LLMs for Information Retrieval	Large Language Models for Information Retrieval	22
Safety in LLMs	Towards Safer Generative Language Models: Safety Risks, Evaluations, and Improvements	17
Chain of Thought	A Survey of Chain of Thought Reasoning	13
Hallucination in LLMs	A Survey on Hallucination in Large Language Models	116
Bias and Fairness in LLMs	Bias and Fairness in Large Language Models	12
Large Multi-Modal Language Models	Large-scale Multi-Modal Pre-trained Models	61
Acceleration for LLMs	A Survey on Model Compression and Acceleration for Pretrained Language Models	22
LLMs for Software Engineering	Large Language Models for Software Engineering	49

B Details of Implementations

We adopt nomic-embed-text-v1.5 [49], a widely used embedding model in RAG applications. To build our database, we store the embeddings of the title and abstract for each paper. Since the context window length is 8k, which is longer than any individual abstract, we embed the raw text directly without chunkings. During generation, related papers are retrieved by the abstract and ranked by their similarity to the query. When generating subsection content, the model needs to write the corresponding paper titles where citations are required. After generation, each title will be embedded as a query and be mapped to the closest paper title in our database. This approach allows the LLMs to use their own parameter knowledge to generate citations without references while ensuring the existence of the generated citations. When calling API, we set temperature = 1 and other parameters as default. Even with the same parameters, the final length of the generated surveys can vary. Therefore, papers with lengths from 8k to 16k are classified into the 8k category, those from 16k to 32k into the 16k category, and so on.

C Details of Evaluation

For citation quality, we define a sentence with at least one citation as a claim and extract all the claims from the generated survey. For human evaluations, we invite three PhD students and all of them have experience in writing LLMs-related surveys. We provide them with the same scoring criteria, along with explanations of the specific metrics. They are asked to score based on these criteria, and the final rankings of the generated surveys are determined by the total scores.

D Cost Analysis

We present the average number of tokens to generate a 32k-tokens survey, along with the cost of using different LLMs in Table 8.

Table 8: Cost of AutoSurvey

Input tokens	Output tokens	Claude-haiku	Gemini-1.5-pro	GPT-4
3009.7K	112.9K	0.89$	11.72$	33.48$

E Societal Impact and Ethical Considerations

By integrating various specialized databases, our approach can generate academic surveys across different fields, potentially filling the gaps in existing reviews. However, as our method relies on the performance of large models, it inevitably contains citation errors. Therefore, the generated survey content is intended for reference only. All personnel involved in the evaluation process participated voluntarily and received ample compensation. All data used in our experiment is sourced from arXiv and is allowed for non-commercial use.

F Prompt used in AutoSurvey

```
ROUGH_OUTLINE_PROMPT =
    '''
    You want to write a overall and comprehensive academic survey
    about [TOPIC].
    You are provided with a list of papers related to the topic below:
    ---
    [PAPER LIST]
    ---
    You need to draft a outline based on the given papers.
    The outline should contains a title and several sections.
    Each section follows with a brief sentence to describe what to
    write in this section.
    The outline is supposed to be comprehensive and contains [SECTION
    NUM] sections.

    Return in the format:
    <format>
    Title: [TITLE OF THE SURVEY]
    Section 1: [NAME OF SECTION 1]
    Description 1: [DESCRIPTION OF SENTCTION 1]

    ...

    Section K: [NAME OF SECTION K]
    Description K: [DESCRIPTION OF SENTCTION K]
    </format>
    The outline:
    '''

SUBSECTION_OUTLINE_PROMPT =
    '''
    You want to write a overall survey about [TOPIC].
    You have created a overall outline below:
    ---
    [OVERALL OUTLINE]
    ---
    The outline contains a title and several sections.
    Each section follows with a brief sentence to describe what to
    write in this section.
```


https://doi.org/10.52202/079017-3655

```
    You need to enrich the section [SECTION NAME].
    The description of [SECTION NAME]: [SECTION DESCRIPTION]
    You need to generate the framwork containing several subsections
    based on the overall outlines.
    Each subsection follows with a brief sentence to describe what to
    write in this subsection.
    These papers provided for references:
    ---
    [PAPER LIST]
    ---
    Return the outline in the format:
    <format>
    Subsection 1: [NAME OF SUBSECTION 1]
    Description 1: [DESCRIPTION OF SUBSENTCTION 1]

    ...

    Subsection K: [NAME OF SUBSECTION K]
    Description K: [DESCRIPTION OF SUBSENTCTION K]
    </format>
    Only return the outline without any other informations:
    '''

MERGING_OUTLINE_PROMPT =
    '''
    You want to write a overall survey about [TOPIC].
    You are provided with a list of outlines as candidates below:
    ---
    [OUTLINE LIST]
    ---
    Each outline contains a title and several sections.
    Each section follows with a brief sentence to describe what to
    write in this section.
    You need to generate a final outline based on these provided
    outlines to make the final outline show comprehensive insights of
    the topic and more logical.
    Return the in the format:
    <format>
    Title: [TITLE OF THE SURVEY]
    Section 1: [NAME OF SECTION 1]
    Description 1: [DESCRIPTION OF SENTCTION 1]

    ...

    Section K: [NAME OF SECTION K]
    Description K: [DESCRIPTION OF SENTCTION K]
    </format>
    Only return the final outline without any other informations:
    '''

SUBSECTION_WRITING_PROMPT =
    '''
    You  wants to write a overall and comprehensive survey about [
    TOPIC].
    You have created a overall outline below:
    ---
    [OVERALL OUTLINE]
    ---
    Below are a list of papers for reference:
    ---
    [PAPER LIST]
    ---

    Now you need to write the content for the subsection:
    "[SUBSECTION NAME]".
```

```
    The details of what to write in this subsection called [SUBSECTION
     NAME] is in this descripition:
    ---
    [DESCRIPTION]
    ---
    Here is the requirement you must follow:
    1. The subsection is recommended to contain more than [WORD NUM]
    words.
    2. When writing sentences that are based on specific papers above,
     you cite the "paper_title" in a '[]' format to support your
    content.

    Here's a concise guideline for when to cite papers in a survey:
    ---
    1. Summarizing Research: Cite sources when summarizing the
    existing literature.
    2. Using Specific Concepts or Data: Provide citations when
    discussing specific theories, models, or data.
    3. Using Established Methods: Cite the creators of methodologies
    you employ in your survey.
    4. Supporting Arguments: Cite sources that back up your
    conclusions and arguments.
    ---
    Only return the content more than [WORD NUM] words you write for
    the subsection [SUBSECTION NAME] without any other information:
    '''

CITATION_REFLECTION_PROMPT =
    '''
    You want to write a overall and comprehensive survey about [TOPIC
    ].
    Below are a list of papers for references:
    ---
    [PAPER LIST]
    ---
    You have written a subsection below:
    ---
    [SUBSECTION]
    ---
    The sentences that are based on specific papers above are followed
     with the citation of "paper_title" in "[]".
    For example 'the emergence of large language models (LLMs) [PaLM:
    Scaling language modeling with pathways]'

    Here's a concise guideline for when to cite papers in a survey:
    ---
    1. Summarizing Research: Cite sources when summarizing the
    existing literature.
    2. Using Specific Concepts or Data: Provide citations when
    discussing specific theories, models, or data.
    3. Using Established Methods: Cite the creators of methodologies
    you employ in your survey.
    4. Supporting Arguments: Cite sources that back up your
    conclusions and arguments.
    ---

    Now you need to check whether the citations of "paper_title" in
    this subsection is correct.
    Once the citation can not support the sentence you write, correct
    the paper_title in '[]' or just remove it.

    Do not change any other things except the citations.
    Only return the subsection with correct citations:
    '''
```


https://doi.org/10.52202/079017-3655

```
COHERENCY_REFINEMENT_PROMPT =
    '''
    You want to write a overall and comprehensive survey about [TOPIC
    ].

    Now you need to help to refine one of the subsection to improve th
     ecoherence of your survey.

    You are provied with the content of the subsection along with the
    previous subsections and following subsections.

    Previous Subsection:
    ---
    [PREVIOUS]
    ---

    Following Subsection:
    ---
    [FOLLOWING]
    ---

    Subsection to Refine:
    ---
    [SUBSECTION]
    ---

    Now refine the subsection to enhance coherence, and ensure that it
     connects more fluidly with the previous and following subsections
    .
    Remember that keep all the essence and core information of the
    subsection intact. Do not modify any citations in [] following the
     sentences!!!!
    Only return the whole refined content of the subsection without
    any other informations:
    '''

NLI_PROMPT =
    '''
    ---
    Claim:
    [CLAIM]
    ---
    Source:
    [SOURCE]
    ---
    Claim:
    [CLAIM]
    ---
    Is the Claim faithful to the Source?
    A Claim is faithful to the Source if the core part in the Claim
    can be supported by the Source.\n
    Only reply with 'Yes' or 'No':
    '''


NAIVE_RAG_PROMPT =
    '''
    You are an expert in artificial intelligence who wants to write a
    overall and comprehensive survey about [TOPIC].
    You are provided with a list of papers related to [TOPIC] below:
    ---
    [PAPER LIST]
    ---

    Here is the survey content you have written:
```

```
---
[SURVEY CONTENT]
---

Hers is the requirement of the survey:
1. The survey must be more than [SURVEY LEN] tokens!
2. containing serval sections. Each section contains several
subsections.
5. Cite several paper provided above to support the content you
write.

Here is the format of your writing:
1. '##' indicates the section title
2. '###' indicates the subsection title
3. Only cite the "paper_title" in '[]'. An example of citation: '
the emergence of large language models (LLMs) [Language models are
 few-shot learners; Language models are unsupervised multitask
learners; PaLM: Scaling language modeling with pathways]'

You need to continue writing the survey by adding a new section or
 subsection.

Do not stop until the length of survey is more than [SURVEY LEN]
tokens!!!

Return the content you write:
'''
```


https://doi.org/10.52202/079017-3655

```
# Comprehensive Survey on Emotion Recognition using Large Language
    Models

## 1. Introduction to Emotion Recognition and Large Language Models
     Emotion recognition has been a crucial and active research area in
     the field of affective computing, which aims to enable machines
    to understand, interpret, and respond to human emotions [1].
    Emotions play a fundamental role in human cognition, decision-
    making, and social interaction [2], and the ability to
    automatically recognize and interpret emotions has a wide range of
     applications, including healthcare, education, entertainment, and
     human-computer interaction [3]. The importance of emotion
    recognition is evident in various real-world applications. In
    healthcare, emotion recognition can be used to monitor patient
    mental health, provide personalized therapy, and improve doctor-
    patient communication [4]. In education, emotion recognition can
    help identify student engagement and frustration levels, enabling
    adaptive learning environments that cater to individual needs [5].
     In the entertainment industry, emotion recognition can be used to
     analyze viewer responses and tailor content to evoke desired
    emotional responses [6]. Despite the significant benefits of
    emotion recognition, the field faces several challenges that have
    hindered its widespread adoption and implementation [7]. One of
    the primary challenges is the inherent complexity and subjectivity
     of emotions, which can vary across individuals, cultures, and
    contexts [8]. Emotions are often expressed through multiple
    modalities, including facial expressions, vocal cues, body
    language, and physiological signals, and integrating these diverse
     sources of information is a significant challenge [9].
    Additionally, the availability of high-quality, diverse, and
    annotated emotion datasets is a persistent challenge in the field
    [10]. Many existing datasets are limited in size, lack diversity,
    or have inconsistent or subjective emotion labeling, which can
    lead to biases and poor generalization of emotion recognition
    models [11].
     ...
### 1.1 Background on Emotion Recognition

### 1.2 Large Language Models and their Capabilities

### 1.3 Emotion Representation in LLMs

### 1.4 Multimodal Emotion Recognition using LLMs

## 2. Techniques and Approaches for Emotion Recognition using LLMs

### 2.1 Fine-tuning LLMs on Emotion Datasets

### 2.2 LLM-based Prompting Methods for Emotion Recognition

### 2.3 Integrating LLMs with Other Modalities for Multimodal Emotion
    Recognition

## 3. Enhancing LLM-based Emotion Recognition

### 3.1 Data Augmentation for Improving Emotion Recognition

### 3.2 Prompt Engineering for Emotion Recognition

### 3.3 Integrating External Knowledge for Emotion Recognition

## 4. Challenges, Limitations, and Ethical Considerations

### 4.1 Model Biases and Hallucinations
```

```
### 4.2 Interpretability and Explainability

### 4.3 Ethical Considerations

## 5. Applications and Future Directions

### 5.1 Assistive Robotics

### 5.2 Mental Health Assessment

### 5.3 Customer Service and User Experience

### 5.4 Symbolic Reasoning and Long-tailed Emotions

### 5.5 Robust Evaluation Frameworks

## References

[1] Affective Computing for Large-Scale Heterogeneous Multimedia Data
    A  Survey
[2] Emotion Recognition in Conversation  Research Challenges, Datasets
   , and  Recent Advances
[3] A Comprehensive Survey on Affective Computing; Challenges, Trends,
     Applications, and Future Directions
[4] Affective Computing for Healthcare  Recent Trends, Applications,
   Challenges, and Beyond
[5] Automatic Sensor-free Affect Detection  A Systematic Literature
   Review
[6] Affective Video Content Analysis  Decade Review and New
   Perspectives
[7] Emotion Recognition from Multiple Modalities  Fundamentals and
   Methodologies
[8] The Ambiguous World of Emotion Representation
[9] Multimodal Affective Analysis Using Hierarchical Attention
   Strategy with  Word-Level Alignment
[10] Expression, Affect, Action Unit Recognition  Aff-Wild2, Multi-
   Task  Learning and ArcFace
[11] Feature Dimensionality Reduction for Video Affect Classification
    A  Comparative Study
```

 https://doi.org/10.52202/079017-3655

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] See limitation section

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] No theory in this paper

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] All details provided

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-3655

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] See experiment section

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-3655

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We just use apis of LLMs, and their owners have safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our research are not research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3655