ReportBench: Evaluating Deep Research Agents via Academic Survey Tasks

Minghao Li[*†], **Ying Zeng**[*], **Zhihao Cheng**[*], **Cong Ma**, **Kai Jia**
ByteDance BandAI
{liminghao.bd,zengying.ss,zhihao.cheng,macong.13,jiakai}@bytedance.com

Abstract

The advent of Deep Research agents has substantially reduced the time required for conducting extensive research tasks. However, these tasks inherently demand rigorous standards of factual accuracy and comprehensiveness, necessitating thorough evaluation before widespread adoption. In this paper, we propose ReportBench, a systematic benchmark designed to evaluate the content quality of research reports generated by large language models (LLMs). Our evaluation focuses on two critical dimensions: (1) the quality and relevance of cited literature, and (2) the faithfulness and veracity of the statements within the generated reports. ReportBench leverages high-quality published survey papers available on arXiv as gold-standard references, from which we apply reverse prompt engineering to derive domain-specific prompts and establish a comprehensive evaluation corpus. Furthermore, we develop an agent-based automated framework within ReportBench that systematically analyzes generated reports by extracting citations and statements, checking the faithfulness of cited content against original sources, and validating non-cited claims using web-based resources. Empirical evaluations demonstrate that commercial Deep Research agents such as those developed by OpenAI and Google consistently generate more comprehensive and reliable reports than standalone LLMs augmented with search or browsing tools. However, there remains substantial room for improvement in terms of the breadth and depth of research coverage, as well as factual consistency. The complete code and data will be released at the following link: https://github.com/ByteDance-BandAI/ReportBench.

1 Introduction

The rapid development of LLM-powered Deep Research agents has revolutionized the process of knowledge synthesis by enabling autonomous execution of extensive research tasks, including academic literature surveys, industry analyses, and market assessments Chen et al. (2025); Gottweis et al. (2025); Lu et al. (2024); Tang et al. (2025); Yamada et al. (2025); Zheng et al. (2025); Li et al. (2025). Tasks that traditionally required days or weeks of manual effort can now be completed within minutes. Notable examples include advanced systems such as OpenAI OpenAI (2025) and Google's Gemini Deep Research Google (2025), which effectively integrate various external tools and perform multiple rounds of deep reasoning. Despite their promising capabilities, widespread practical adoption critically depends on their ability to consistently deliver research reports with high factual accuracy

[*]Equal contribution
[†]Corresponding author

and comprehensive content quality. Therefore, it is essential to monitor and ensure the quality of generated reports through evaluation. However, defining what constitutes a good report is challenging and lacks broad consensus, resulting in the current absence of mature evaluation methodologies for research report generation.

In addressing this challenge, we decompose the evaluation of research reports generated by LLMs into two core dimensions: writing quality and report content. Due to the subjectivity of writing-style evaluation, while the criteria for assessing content quality can be more clearly defined, this work focuses primarily on the evaluation of report content, leaving the assessment of writing quality to future work. Specifically, we assert that the content quality of research reports hinges on two critical factors: (1) the quality and relevance of cited literature, and (2) the faithfulness and veracity of generated statements, whether derived from cited references or produced by the model.

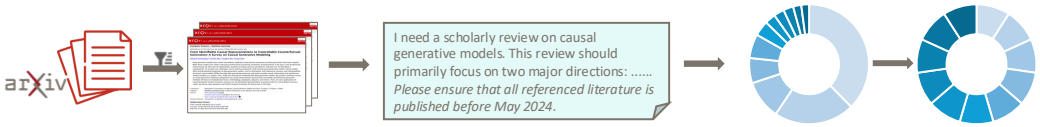
To establish a high-quality benchmark capable of rigorously assessing research reports, we propose ReportBench, a novel evaluation framework leveraging expert-generated literature reviews. Given the constraints of relying on human annotators, who typically vary in expertise, we propose using published survey papers available on arXiv as gold-standard references. Published survey papers are typically written by domain experts and have undergone a peer review process that provides additional expert-level validation, considered among the highest-quality research reports currently available.

In practice, our methodology unfolds in two phases. First, we generate domain-specific retrieval prompts directly from expert-authored survey papers on arXiv: by analyzing each paper's publication date and full text, we generate three granularity levels of prompts (sentence-level, paragraph-level, and richly detailed versions) that precisely capture the scope, methods, and temporal constraints of the original research. These prompts form the backbone of our evaluation corpus, ensuring that downstream agents search and synthesize information within the exact topical and chronological boundaries of each survey. We extract the list of cited references from the arXiv surveys as the ground truth. Given the synthesized prompts as test inputs, Deep Research agents conduct research and generate reports, which are then evaluated based on the reference overlap with the ground truth, serving as a measure of the research skills.

In the second phase of our validation pipeline, we design two different verification procedures based on whether a statement includes an explicit citation to external literature. Specifically, for cited statements, the system identifies all in-text citations within the report, maps each citation to its corresponding source document, and employs semantic matching to ensure factual support from the cited literature. For non-cited statements, the framework employs a voting mechanism across multiple web-connected models to verify the factuality of these statements. By combining these complementary validation procedures, ReportBench delivers a systematic and detailed assessment of AI-generated research reports, ensuring the relevance and quality of cited literature and the factual accuracy of all claims through citation-based and web-based validation.

Our contributions can be summarized as follows:

- We present **ReportBench**, a systematic benchmark designed to evaluate the quality of research reports generated by Deep Research agents, with a focus on the quality of references and the factual accuracy of all statements presented in the report.
- We propose an automated and scalable data synthesis method for constructing academic survey tasks, including prompts and ground truth, from expert-authored survey papers on arXiv. Besides, we introduce an automatic agentic evaluation framework that evaluates the precision and recall of the generated report with respect to a ground-truth reference and performs factual verification of individual claims made within the report.
- We release a comprehensive benchmark suite—datasets, prompts, and evaluation scripts—to support reproducible research and community-driven progress in evaluating LLM-based knowledge synthesis.



PHASE I: Survey paper identification and reference extraction

PHASE II: Prompt generation

PHASE III: Application domain distribution

Figure 1: Overall benchmark data construction workflow.

2 Methodology

We introduce **ReportBench**, a comprehensive evaluation framework designed to rigorously assess Deep Research agents through two interconnected components: (i) the automated construction of high-quality benchmark datasets derived from expert-authored survey papers, and (ii) a systematic validation pipeline that evaluates the quality and factual consistency of AI-generated research reports. In the following sections, we detail the processes that underlie the synthesis of the dataset and the design of our evaluation workflow.

2.1 Dataset Construction

In this section, we detail the end-to-end pipeline to construct high–quality deep research questions along with ground-truth answers based on published survey papers. This workflow comprises three consecutive phases: (i) survey paper identification and reference extraction, (ii) prompt generation, and (iii) application domain distribution. A diagram illustrating the data construction process is presented in Figure 1.

2.1.1 Phase I: Survey Paper Identification and Reference Extraction

The first step is to identify high-quality survey papers to create problems. We start from the complete arXiv metadata snapshot arXiv.org submitters (2024) and only reserve papers with a submission date later than `2020-01-01`. To ensure the quality of papers, we only select those that have undergone peer review and have been formally published. We achieve this by using regular expressions, *i.e.*, querying over titles to match *"survey"* or *"review"* to filter survey papers and searching *"published"* or *"accepted"* in the comments field of a submission. To reduce systematic false positives in domains such as astronomy, we prompted `GPT-4o` Hurst et al. (2024) with each paper's title and abstract to produce a binary classification of whether the paper is a literature survey.

For each survey paper, we analyze its LaTeX source file to extract cited references. Specifically, we parse LaTeX citation commands, identify and retrieve relevant bibliographic entries from associated bibliography databases, and filter these to retain only references explicitly cited in the main text. Hence, the extracted bibliography mirrors the true citation pattern of the paper. The resulting dataset constitutes a gold-standard benchmark for evaluating retrieval precision. Finally, we ultimately retained 678 papers.

2.1.2 Phase II: Prompt Generation

Survey papers can be regarded as a great depth of research work focused on a specific topic at a specific time, making it possible to create deep research questions via a *reverse prompt engineering* manner. In other words, given the publication date and the full text of a survey paper obtained through a PDF parsing tool, we prompt an LLM to generate a query whose ideal answer is precisely that paper. Hence, we obtain a query and its ground truth (the survey paper itself). To increase the diversity of prompts, we design three types of prompt templates:

Sentence-level prompt

```
A single sentence that succinctly defines the overarching academic
field covered by the survey.
```

Paragraph-level prompt

```
A short paragraph elaborating the research area, its main
subtopics, and the methodological perspectives covered in the
survey.
```

Detail-rich prompt

```
A detailed question that comprehensively describes the specific
research domain, key research directions, and the methodological
approaches of interest.  Additional constraints may be included,
such as preferred conferences or journals, language of the cited
literature (e.g., English, Chinese), participating institutions or
laboratories.
```

Besides, to ensure temporal consistency, we require each generated prompt to include a cut-off date corresponding to the most recent update of the paper. For example, an expression like the following is needed.

> "Ensure only papers published before April 2025 are referenced."

This requirement ensures that LLMs' retrieval window matches the survey's citation horizon and prevents leakage of post-publication knowledge and hacking of finding the exact same paper. Nevertheless, we still observe a phenomenon akin to prompt hacking during model evaluation, *i.e.*, the model disregards the imposed temporal constraints and directly retrieves the original source paper. To address this issue, we augment the prompt with an additional explicit instruction, stipulating that the model must refrain from citing the original paper corresponding to the given prompt. We present three prompt examples in Appendix A.3

2.1.3 Phase III: Application Domain Distribution

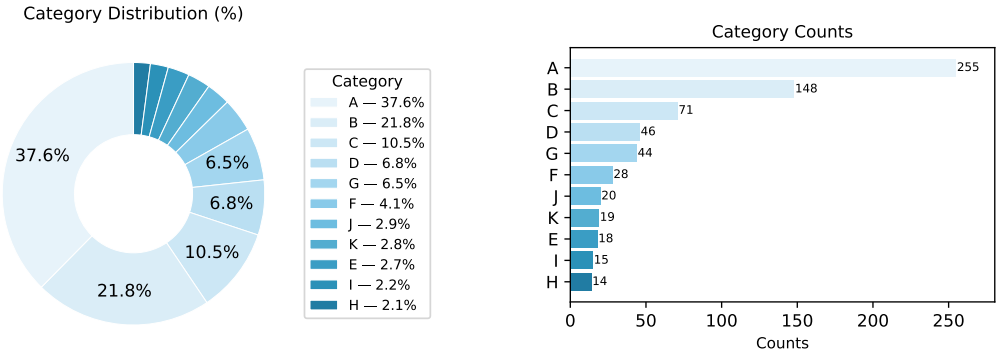
To facilitate a more granular analysis of tested models, we classified the prompts into distinct application domains. Specifically, we utilize Gemini 2.5 Pro Comanici et al. (2025) to classify each paper based on the title and abstract. This process yields ten distinct categories, as shown in the following box. To reduce misclassification, we introduce an unknown category, allowing the model to assign uncertain cases to this class.

A	Basic Research and Scientific Exploration	**F**	Transportation and Smart Mobility
B	Information and Communications Technology	**G**	Public Safety and Social Governance
C	Artificial Intelligence and Data Intelligence	**H**	Finance and Business Services
D	Healthcare and Biomedicine	**I**	Energy and Environmental Sustainability
E	Manufacturing and Smart Manufacturing	**J**	Culture, Media, and Digital Content
K	Unknown Category		

The distribution of prompts across these domains is inherently biased due to the specific disciplinary focus of the arXiv corpus, as shown in Figure 2. To create a balanced and general test set, we down-sample a total of 100 papers. As we have mentioned before, we create three types of prompts for each paper. Thus, we randomly sample from these three types to obtain the final prompt with diversity. In other words, a dataset with 100 prompts is created, which we name as **ReportBench**. The quality of the classification of this subset was then reviewed and validated by four research experts.

2.2 Evaluation Process

Our evaluation process uses test prompts derived from reverse prompt engineering, which require models to generate complete research reports under two constraints: a time limit and a restriction against referencing the original report, which is presented in Figure 3. **Content quality** is first evaluated by assessing the cited references: we compare the reference list



(a) Category distribution (pie). (b) Category counts (bar).

Figure 2: Application domain distribution of the 678 filtered ReportBench prompts: (a) a pie chart showing the proportion of each application domain, (b) a bar chart illustrating the total task counts across all 11 categories.

in the generated report with that of the ground truth, and the overlap ratio between the two lists serves as an indicator of the report's overall quality. **Statement factuality** is further assessed through two complementary validation procedures. For cited statements, we verify alignment with source documents via semantic matching, while for non-cited statements, we adopt a multi-model voting mechanism to assess factual correctness. This dual strategy ensures both the faithfulness of cited content and the veracity of non-cited claims in evaluating Deep Research reports.

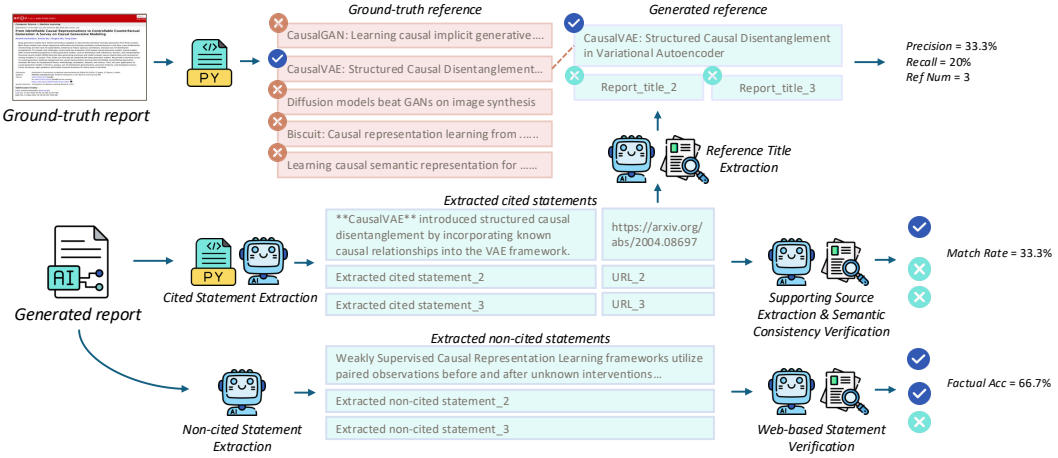


Figure 3: Evaluation Process.

Content quality We first extract all URLs from the report. Since most reports generated by the Deep Research products use URL links to cite web pages, we adopt the same citation format throughout our evaluation, including when assessing the base model. While this approach results in longer text, it offers the advantage of placing the citation immediately adjacent to the corresponding statement, which ensures consistent performance even under chunked evaluation settings. After normalizing and deduplicating them, we retrieve the content of each web page. An LLM is then used to determine whether each page corresponds to a scholarly article and, if so, to extract the article title. Finally, we compute the overlap between the extracted document titles and the ground-truth reference titles to produce a quality score.

Cited statements We design a three-stage structured validation pipeline. First, an LLM automatically identifies all statements in the generated report that contain explicit citation links, establishing a mapping between each statement and its referenced source. Second, we retrieve the full content of each cited webpage via web scraping and prompt the LLM to locate the most semantically relevant passage that supports the original statement. Finally, the LLM performs consistency verification by comparing the statement with the retrieved content, and the results are aggregated to compute an overall citation consistency score for the report. Unlike traditional "LLM-as-a-judge" approaches, which often suffer from instruction-following issues or biased scoring, our method decomposes the evaluation into fine-grained, interpretable, and verifiable steps. All intermediate outputs are retained for optional human inspection, thereby maximizing the reliability and transparency of the evaluation process.

Non-cited statements We use a simple two-step validation process. First, we extract all factual statements in the report that do not have any citations, and remove content that is general common sense or already supported by references. Then, we ask several web-connected LLMs to check each statement independently. Each model looks up information online and gives its judgment. We combine their answers using a voting mechanism to decide whether the statement is likely to be correct. This approach avoids relying on a single model and makes the validation more reliable.

3 Experiment

In this section, we present the performance of a diverse set of models evaluated on Report-Bench. Specifically, we examine specialized Deep Research agents from OpenAI and Google Gemini. Additionally, we assess several state-of-the-art (SOTA) base models, originally lacking native Internet access, by augmenting them with an external search engine and link reader to enable the web-retrieval capabilities essential for completing our evaluation tasks. These enhanced base models are then benchmarked alongside the native Deep Research agents.

3.1 Setttings

In our evaluation workflow, different LLMs are used for different components. For statement extraction, supporting source extraction, and semantic consistency verification, we adopt `gpt-4o`. For the fact-checking of non-cited statements, we employ two web-connected models: `gemini-2.5-pro` and `gemini-2.5-flash`. Each model performs three independent judgments per statement, resulting in a total of six verdicts. The final decision is determined by majority voting, and the proportion of votes is recorded as a confidence score. In the evaluation of base models, we integrated search and link-reading tools using each model's native function call interface. Specifically, we used SerpAPI[1] for Google Search access and Firecrawl[2] for retrieving web pages in Markdown format. Due to context length limitations, we capped the maximum number of tool calls at five per instance.

To evaluate the performance of both Deep Research agents and base models, we manually collected responses from the web-based interfaces of OpenAI and Gemini, as well as batch-executed outputs from the base models, during the period from July 14 to July 25. During data collection, we ensured that OpenAI was using the standard version of Deep Research, powered by the `o3` model. For Gemini, we made sure that both the "Gemini 2.5 Pro" and "Deep Research" toggles were enabled on the web interface to activate its full research capabilities.

[1] `https://serpapi.com/`
[2] `https://www.firecrawl.dev/`

3.2 Evaluation Metrics

As described in our evaluation logic, we define three sets of metrics to assess a model's performance in conducting scientific research tasks. First, we compute the **precision** and **recall** of retrieved references against groundtruth references. Precision reflects the proportion of cited references that are relevant, while recall measures the proportion of ground-truth references successfully retrieved. We also report the average number of references per report to capture the model's reference density. To evaluate statement-level performance, we measure the average number of cited statements and non-cited statements per report. For cited statements, we compute the **match rate**, i.e., the proportion of statements that are semantically consistent with their cited sources. For non-cited statements, we compute the **factual accuracy**, defined as the proportion of statements that are verified to be factually correct via web-connected LLMs.

Test Model	Reference			Cited statements		Non-cited statements	
	Precision	Recall	Ref Num	Match Rate	Count	Factual Acc	Count
OpenAI Deep Research	**0.385**	0.033	9.89	**78.87%**	88.2	95.83%	38.9
Gemini Deep Research	0.145	**0.036**	32.42	72.94%	96.2	92.21%	49.6
gemini-2.5-flash	0.237	0.012	5.47	44.88%	12.1	**98.52%**	11.5
gemini-2.5-pro	0.269	0.010	4.27	59.24%	6.58	96.08%	9.35
openai-o3	0.299	0.031	12.26	31.43%	16.16	82.22%	11.51
claude4-sonnet	0.337	0.021	6.74	73.67%	14.93	92.64%	17.07

Table 1: Performance metrics of OpenAI Deep Research, Gemini Deep Research, and the base models. "Ref Num" denotes the average number of references per report, and "Count" denotes the average number of cited or non-cited statements.

3.3 Product-Level Comparative Analysis

Table 1 presents the performance metrics of OpenAI Deep Research and Gemini Deep Research. In terms of retrieval performance, OpenAI achieves significantly higher precision (0.385) compared to Gemini (0.145), indicating that the references it retrieves are more likely to match the gold-standard set. Gemini shows a slightly higher recall (0.036 vs. 0.033), but this gap is negligible in practical terms. As shown in the table, Gemini generates more than three times as many cited statements on average (32.42 vs. 9.89), yet this increase does not translate into a significant improvement in recall. This suggests that Gemini tends to over-generate citations without proportionally improving the coverage of high-quality references. In some cases, excessive citation may even introduce redundancy or dilute the relevance of retrieved content. Given that the ground truth from ReportBench includes an average of 153 references per paper, with many citations supporting the same or overlapping statements, we believe recall should be considered a secondary signal rather than the primary focus of evaluation.

In terms of statement quality, both products demonstrate strong performance on generating reports, with OpenAI achieving better citation alignment (Match Rate 78.87% vs. 72.94%). Despite citing fewer sources, OpenAI maintains a high average alignment score (88.2), suggesting stronger precision in citation usage. For non-cited statements, Gemini produces more such content (49.6 vs. 38.9), while OpenAI achieves better factual accuracy (95.83% vs. 92.21%), indicating its stronger calibration in generating reliable citation-free content.

3.4 Model-Level Comparative Analysis

We now analyze the results across several foundation models and compare them with the corresponding Deep Research agents.

OpenAI Deep Research vs. `o3`
OpenAI Deep Research and `o3` exhibit similar retrieval performance, with precision (0.385 vs. 0.299) and recall (0.033 vs. 0.031) showing only slight differences. Meanwhile, the average number of references per report is also comparable (9.89 vs. 12.26). This observation aligns well with OpenAI's official disclosure that the retrieval and synthesis backbone of Deep Research is powered by the `o3` model.

However, we observe substantial differences in the number and quality of generated statements. OpenAI Deep Research produces significantly more cited statements on average (88.2 vs. 16.16) and more non-cited statements (38.9 vs. 11.51), while achieving a notably higher citation match rate (78.87% vs. 31.43%) and factual accuracy (95.83% vs. 82.22%). This suggests that Deep Research is not a direct output of `o3`, but rather likely incorporates an additional writing module, possibly optimized via fine-tuning or structured pipelines. Such a pipeline may be responsible for structuring retrieved content into a more coherent, citation-aligned report.

Gemini Deep Research vs. `gemini-2.5-pro`
Similarly, Gemini Deep Research and its base model `gemini-2.5-pro` diverge significantly across multiple dimensions. Gemini Deep Research trades off some precision (0.145 vs. 0.269) to achieve much higher recall (0.036 vs. 0.010) and generates far more references per report (32.42 vs. 4.27). In terms of statement volume, it produces many more cited statements (96.2 vs. 6.58) and non-cited statements (49.6 vs. 9.35). Despite this increase in volume, its citation alignment remains strong (72.94% vs. 59.24%), while its non-cited statement accuracy is slightly lower than the base model (92.21% vs. 96.08%). These pronounced gaps—in precision/recall trade-off, citation count, and overall coverage—mirror the contrast observed between OpenAI Deep Research and `o3`, and suggest that the system has undergone targeted optimization for thorough research and report generation. Taken together with the visible "plan" and "step-by-step reasoning" phases presented in the Gemini Deep Research web interface, it seems plausible that the system functions more like a thoughtfully constructed multi-agent workflow or pipeline.

Base-Model Comparison
Among the four base models, `claude4-sonnet` demonstrates the most balanced performance—achieving a precision of 0.337, a recall of 0.021, an average of 6.74 reference documents per report, a high citation semantic consistency (73.67%), and a strong non-cited statement factual accuracy (92.64%). In contrast, `gemini-2.5-pro` attains higher precision (0.269) at the expense of recall (0.010) and generates fewer reference documents on average (4.27 per report), limiting its coverage. `gemini-2.5-flash` underperforms on both precision (0.237) and recall (0.012), with lower citation semantic consistency (44.88%), indicating poorer citation relevance. Meanwhile, `o3` produces the most references (12.26 per report) and moderate recall (0.031), but its citation semantic consistency (31.43%) and non-cited statement accuracy (82.22%) lag behind.

Overall, Deep Research products significantly outperform their base models in coverage and factual grounding, pointing to the value of task-specific model fine-tuning or pipeline design beyond standalone LLM capabilities.

4 Analysis

It is notable that many models exhibit low citation semantic consistency, particularly when relying on function-call mechanisms to retrieve and cite literature. In our manual inspection of evaluation results, we identified two representative failure types: **statement hallucination**, where the content deviates from the cited source, and **citation hallucination**, where the reference itself is fabricated.

Statement Hallucination. In our manual audit of arXiv:2407.15186 test cases, we identified representative errors in statement generation. For example, OpenAI Deep Research generated the following claim:

> Kulkarni *et al.* (2025) and others introduced RL fine-tuning where the model gets a reward of +1 if its SQL yields the correct answer when run, and 0 otherwise (arXiv:2503.23157v2, §3.2).

Upon inspection, the cited part indeed describes a reasoning-enhanced RL reward scheme for Text-to-SQL; however, the list of authors does not include "Kulkarni.". In fact, Kulkarni did publish a paper on reinforcement learning and Text-to-SQL, but it was not among the references cited in the generated report. We speculate that the model may have encountered similar data during training and mistakenly attributed Kulkarni's contribution to this cited paper.

Citation Hallucination. During our evaluation of arXiv:2009.12619, we observed a clear instance of link hallucination in the generated report from `gemini-2.5-pro`. The model generated the claim:

> In-vehicle Crowd Monitoring: The use of surveillance cameras inside buses and trains for passenger counting is a well-established practice. Advanced image processing and computer vision techniques can automatically analyze video feeds to estimate the passenger load. For instance, a system was proposed to estimate the number of passengers in a bus using image processing techniques on the captured video frames, achieving high accuracy. [Vision-Based In-Vehicle Crowd Monitoring](`https://www.researchgate.net/publication/224217198_A_vision-based_system_for_in-vehicle_crowd_monitoring`).

However, the cited URL does not exist and appears to be entirely fabricated by the model. Because the link cannot be resolved, no supporting text or evidence can be retrieved to validate the statement, resulting in a citation mismatch. This example highlights a common error mode in function-call–driven retrieval: the model confidently invents plausible-looking reference links that nonetheless point to nothing, undermining factual grounding.

These examples demonstrate that even advanced Deep Research agents remain susceptible to hallucinating author names, misaligning citations, and fabricating links. Crucially, our evaluation metrics—especially citation semantic consistency—are sensitive to such discrepancies, allowing us to quantitatively capture and penalize these hallucination phenomena across model outputs.

5 Related Work

Long-standing interest has been in the use of AI to synthesize information, not only in the writing of scientific articles Chen et al. (2025); Gottweis et al. (2025); Lu et al. (2024); Tang et al. (2025); Yamada et al. (2025), but also in the search for information and the generation of reports in the general domains Zheng et al. (2025); Li et al. (2025). With the rapid advancement of information synthesis research, the evaluation of long-form reports has become increasingly important. Existing evaluation benchmarks focus primarily on individual aspects such as fact checking, citation judgment, or overall writing quality. Although these benchmarks can assess the quality of the report to some extent, several limitations remain unaddressed.

Fact Checking Evaluation Driven by efforts from both academia and industry, automated fact checking has evolved into a well-established multistage pipeline, which has become the dominant research paradigm in the field Eldifrawi et al. (2024). Claim detection aims to identify factual statements worth verifying from large volumes of text Guo et al. (2022); Panchendrarajan & Zubiaga (2024), while evidence retrieval focuses on retrieving relevant documents or textual snippets that support or refute a given claim Eldifrawi et al. (2024); Nanhekhan et al. (2025). Building on this pipeline, several benchmarks have been proposed to evaluate the performance of fact checking in both the general domain Thorne et al. (2018); Ma et al. (2024) and the scientific domain Wadden et al. (2020; 2022); Ho et al. (2025). However, these benchmarks focus solely on fact-checking components, rather than

evaluating the synthesized information as a whole, limiting their ability to assess recent long-form output from large language models (LLMs), such as full research reports.

Citation Evaluation Research reports often include a substantial amount of citation-related content, and evaluating the precision and standardization of these citations plays a crucial role in assessing the overall quality of the report Sarol et al. (2024). Given a report with citation content, tasks such as cited context identification, evidence sentence retrieval, and citation accuracy classification are commonly used to analyze citation quality Sarol et al. (2024). Widely applied in assisted paper writing and review systems, citation verification tools are designed from multiple perspectives, including syntactic verification, existence verification, and semantic verification Barrot (2025); Bairagi & Lihitkar (2024). While citation correctness and existence have been well-studied, the aspect of citation completeness—i.e., whether all relevant prior work on a given research topic has been cited—remains under-explored. To address this gap, our ReportBench incorporates large language models and search tools to evaluate and verify the completeness of related work coverage, offering a more comprehensive perspective for research report quality assessment.

Survey Generation With the advent of LLMs, automated survey generation has seen rapid progress. Early works leveraged LLMs to improve literature comprehension and survey writing Wang et al. (2024); Hu et al. (2025), achieving better coherence compared to sentence extraction methods. Subsequent research explored structured and hierarchical organization, such as hierarchical catalogue generation with semantic and structural metrics, though these remained limited to outline generation with fixed references. Other approaches focused on modeling paper relationships via citation networks, including AutoSurvey Wang et al. (2024) with a two-stage LLM pipeline and HiReview Hu et al. (2025) with a taxonomy-driven framework, though both faced limitations in capturing human writing styles or relying on restricted citation scopes. More recently, SurveyForge Yan et al. (2025) combines human outline structure analysis with high-quality literature retrieval, generating and refining full survey content through a scholar navigation agent. The accompanying SurveyBench evaluates generated surveys across reference, outline, and content quality, showing significant improvements over prior methods. Compared with SurveyBench, ReportBench focuses solely on well-defined and automatically verifiable dimensions of evaluation—namely, factual faithfulness and correctness. In addition, through an automated construction pipeline, it ensures data quality while offering clear scalability advantages, enabling it to serve as a potential source of training data for targeted report optimization in future work.

Deep Research Evaluation The rise of deep research agents (DRAs), driven by powerful models such as ChatGPT OpenAI (2025) and Gemini Google (2025), has underscored the urgent need for robust and targeted evaluation methodologies. While existing benchmarks evaluate capabilities such as web retrieval Wei et al. (2025); Zhou et al. (2025); Wu et al. (2025), multi-hop factual reasoning Wei et al. (2024); Mialon et al. (2024); Phan et al. (2025), and end-to-end report generation Du et al. (2025); Bosse et al. (2025), they often operate at a surface level and fall short of evaluating the core competencies essential for rigorous and reliable research. In contrast, our ReportBench is specifically designed to assess two critical pillars of trustworthy and practical DRA outputs: factual accuracy and citation behavior.

6 Conclusion

In this paper, we present **ReportBench**, a comprehensive benchmark for evaluating the quality of references and the factual accuracy of all statements in reports generated by Deep Research agents. By leveraging expert-authored survey papers as ground-truth and reverse prompt engineering, we enable consistent evaluation of AI-generated research reports across multiple dimensions. Our framework introduces a fine-grained validation workflow that separately assesses cited and non-cited statements, combining citation semantic consistency checks and web-based factual verification. Through large-scale experiments on leading LLM-based research agents and the base models, we demonstrate that Deep Research products can outperform base models in content overage and factual grounding, but still face challenges in hallucination, over-citation, etc. We hope that ReportBench will serve as

a valuable tool for the research community to monitor, compare, and further improve the reliability of AI systems designed for academic survey tasks.

References

arXiv.org submitters. arxiv dataset, 2024. URL https://www.kaggle.com/dsv/7548853.

Mandira Bairagi and Shalini R Lihitkar. Revolutionizing research writing and publishing by using ai-powered tools and techniques. In *International Conference on AI-Driven Advancements in Research and Publications: Intellectual Property Rights, Knowledge Management, and Beyond, Ashok Goel Library, Rishihood University, NCR of Delhi, Sonipat, Haryana, India*, 2024.

Jessie S Barrot. Trinka: Facilitating academic writing through an intelligent writing evaluation system. *Assessing Writing*, 65:100953, 2025.

Nikos I. Bosse, Jon Evans, Robert G. Gambee, Daniel Hnyk, Peter Mühlbacher, Lawrence Phillips, Dan Schwarz, and Jack Wildman. Deep research bench: Evaluating AI web research agents. *CoRR*, abs/2506.06287, 2025. doi: 10.48550/ARXIV.2506.06287. URL https://doi.org/10.48550/arXiv.2506.06287.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, et al. Ai4research: A survey of artificial intelligence for scientific research. *arXiv preprint arXiv:2507.01903*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *CoRR*, abs/2506.11763, 2025. doi: 10.48550/ARXIV.2506.11763. URL https://doi.org/10.48550/arXiv.2506.11763.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6679–6692, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.361. URL https://aclanthology.org/2024.acl-long.361/.

Google. Gemini deep research — your personal research assistant. *Google*, 2025. URL https://gemini.google/overview/deep-research/.

Juraj Gottweis, Wei-Hung Weng, Alexander N. Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist. *CoRR*, abs/2502.18864, 2025. doi: 10.48550/ARXIV.2502.18864. URL https://doi.org/10.48550/arXiv.2502.18864.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206, 2022. doi: 10.1162/TACL_A_00454. URL https://doi.org/10.1162/tacl_a_00454.

Xanh Ho, Sunisth Kumar, Yun-Ang Wu, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. Table-text alignment: Explaining claim verification against tables in scientific papers. *CoRR*, abs/2506.10486, 2025. doi: 10.48550/ARXIV.2506.10486. URL https://doi.org/10.48550/arXiv.2506.10486.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. Taxonomy tree generation from citation graph, 2025. URL https://arxiv.org/abs/2410.03761.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *CoRR*, abs/2504.21776, 2025. doi: 10.48550/ARXIV.2504.21776. URL https://doi.org/10.48550/arXiv.2504.21776.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292, 2024. doi: 10.48550/ARXIV.2408.06292. URL https://doi.org/10.48550/arXiv.2408.06292.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. EX-FEVER: A dataset for multi-hop explainable fact verification. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 9340–9353. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.556. URL https://doi.org/10.18653/v1/2024.findings-acl.556.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Kevin Nanhekhan, Venktesh V, Erik Martin, Henrik Vatndal, Vinay Setty, and Avishek Anand. Flashcheck: Exploration of efficient evidence retrieval for fast fact-checking. In Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (eds.), *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part IV*, volume 15575 of *Lecture Notes in Computer Science*, pp. 385–399. Springer, 2025. doi: 10.1007/978-3-031-88717-8_28. URL https://doi.org/10.1007/978-3-031-88717-8_28.

OpenAI. Introducing deep research. *OpenAI*, 2025. URL https://openai.com/index/introducing-deep-research/.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Nat. Lang. Process. J.*, 7:100066, 2024. doi: 10.1016/J.NLP.2024.100066. URL https://doi.org/10.1016/j.nlp.2024.100066.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Jessica P. Wang, Pawan Kumar, Oleksandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Mstyslav Kazakov, Johannes Schmitt, Geoff Galgon, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Zachary Giboney, Gashaw M. Goshu, Joan of Arc Xavier, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, John Wydallis, Mark Nandor, Ankit Singh, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Darling Duclosel, Jungbae Nam, Jennifer Zampese, Ryan G. Hoerr, Aras Bacho, Gautier Abou Loume, Abdallah Galal, Hangrui Cao, Alexis C. Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Lianghui Li, Sumeet Motwani, Christian Schröder de Witt, Edwin Taylor, Johannes Veith, Eric Singer, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks,

Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont, Justine Leon Uro, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Julien Guillod, Yuqi Li, Joshua Vendrow, Vladyslav Kuchkin, and Ng Ze-An. Humanity's last exam. *CoRR*, abs/2501.14249, 2025. doi: 10.48550/ARXIV.2501.14249. URL https://doi.org/10.48550/arXiv.2501.14249.

Maria Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, and Halil Kilicoglu. Assessing citation integrity in biomedical publications: corpus annotation and nlp models. *Bioinformatics*, 40(7):btae420, 2024.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *CoRR*, abs/2505.18705, 2025. doi: 10.48550/ARXIV.2505.18705. URL https://doi.org/10.48550/arXiv.2505.18705.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.609. URL https://doi.org/10.18653/v1/2020.emnlp-main.609.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.347. URL https://aclanthology.org/2022.findings-emnlp.347/.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. Autosurvey: Large language models can automatically write surveys, 2024. URL https://arxiv.org/abs/2406.10252.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *CoRR*, abs/2411.04368, 2024. doi: 10.48550/ARXIV.2411.04368. URL https://doi.org/10.48550/arXiv.2411.04368.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *CoRR*, abs/2504.12516, 2025. doi: 10.48550/ARXIV.2504.12516. URL https://doi.org/10.48550/arXiv.2504.12516.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web traversal. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 10290–10305. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.508/.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob N. Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *CoRR*, abs/2504.08066, 2025. doi: 10.48550/ARXIV.2504.08066. URL https://doi.org/10.48550/arXiv.2504.08066.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing, 2025. URL https://arxiv.org/abs/2503.04629.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *CoRR*, abs/2504.03160, 2025. doi: 10.48550/ARXIV.2504.03160. URL https://doi.org/10.48550/arXiv.2504.03160.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *CoRR*, abs/2504.19314, 2025. doi: 10.48550/ARXIV.2504.19314. URL https://doi.org/10.48550/arXiv.2504.19314.

A Appendix

A.1 Limitations

ReportBench constructs 100 research tasks closely aligned with real-world scientific inquiry by reverse prompt engineering expert-written survey papers. It evaluates generated reports comprehensively along two axes: content quality and statement factuality. Despite its strengths, several limitations remain:

Data Distribution. The benchmark primarily draws from peer-reviewed survey papers on arXiv, most of which are concentrated in STEM fields. This domain skew may limit the applicability of evaluations to other research areas. Future iterations will incorporate a broader set of source domains to improve coverage and generalization.

Copyright Constraints. To mitigate legal risk, we only include papers under permissive licenses (CC BY 4.0, CC BY-SA 4.0, CC0 1.0, and the arXiv.org Non-exclusive license to distribute). The dataset is released under CC0 1.0 and contains only essential metadata (e.g., title, abstract, and references). Further narrowing the license scope would compromise domain balance. Authors who wish to opt out, please contact us for removal.

A.2 Prompts in Evaluation

A.2.1 Cited Statement Extraction

```
You are given a research report delimited by triple backticks.
Identify every statement that cites an external source (e.g.  has
a URL, DOI, or explicit citation marker) and pair it with the
corresponding URL.
Return a JSON list where each item has two keys:
- "statement":  the single-sentence claim, stripped of
leading/trailing whitespace
- "url":  the canonical URL that supports that claim
If a citation contains multiple URLs, duplicate the statement for
each URL.
ONLY return valid JSON. Report:  "‘{report}"‘
```

A.2.2 Non-cited Statement Extraction

```
You are given a research report delimited by triple backticks.
You are also given a list of statements that already have
citations.
Your task is to identify factual claims or statements that:
1.  Make specific assertions about facts, data, or events
2.  Are NOT already included in the cited statements list
3.  Could potentially be verified through external sources
4.  Are NOT common knowledge or widely accepted facts
Exclude:
- Opinions, analysis, or subjective interpretations
- Statements that are already cited
- Common knowledge or universally accepted facts
- Vague or general statements
Return a JSON list where each item has one key:
- "statement":  the factual claim that lacks citation support
ONLY return valid JSON.
Report:
"‘{report}"‘
Already cited statements:
{cited_statements}
```

A.2.3 Supporting Source Extraction

```
You are provided with
Statement: {statement}

Source Document:
{source_text}

Return any relevant content from the source document that supports
the statement. This can be a sentence, paragraph, or even the
entire text if necessary.
If no content supports it, return "NOT_FOUND".
Return plain text only.
```

A.2.4 Semantic Consistency Verification

```
You will decide whether a claim is correctly supported by a source
sentence.

Claim from report:
{statement}

Source Sentence from original source:
{source_sentence}

Respond with JSON containing:
- "reason": one short sentence explaining your decision
- "match": true or false // true if the source sentence faithfully
supports the claim
Return ONLY the JSON.
```

A.2.5 Web-based Statement Verification

```
You are tasked with verifying the accuracy of a factual statement
using web search capabilities.

Statement to verify:
{statement}

Please:
1. Use web search to find reliable, authoritative sources about
this statement
2. Analyze the information you find from multiple sources
3. Determine if the statement is factually correct or incorrect
based on your research

Respond with JSON containing:
- "reason": a detailed explanation of your verification process
and findings (2-3 sentences)
- "decision": true if the statement is correct, false if it is
incorrect

Only return the JSON response.
```

A.2.6 Reference Title Extraction

```
Please analyze the following academic survey and extract all cited
academic paper titles and author information.

Survey content:
{response}

Please reply in JSON format, containing an array named ‘papers‘,
where each paper object includes the following fields:
- title: the title of the paper
- authors: a list of authors
- is_academic_paper: true (indicating this is an academic paper)

Example format:
{
  "papers": [
    {
      "title": "Deep Learning for Natural Language Processing",
      "authors": ["John Smith", "Jane Doe"],
      "is_academic_paper": true
    },
    ...    ]
}

Note: Only extract explicitly mentioned academic papers. Do not
include books, websites, or other types of references.
```

A.3 Prompt in Data Construction

Sentence-level prompt

```
Please help me research the academic advancements in different
radar data representation methods in the field of autonomous
driving, and ensure only papers published before April 2025 are
referenced.

You also need to follow the following rules:
- Do not refer to the survey titled "Exploring Radar Data
Representations in Autonomous Driving: A Comprehensive Review".
- Responses are given in the form of an English language survey
with citations where appropriate.
```

Paragraph-level prompt

```
I am conducting a literature review on 3D LiDAR localization
technology for autonomous vehicles. I hope you can summarize and
analyze the major research directions and methods in this field,
particularly methods based on 3D point cloud registration, methods
based on 3D features, and emerging methods based on deep learning.
Please ensure that all the referenced literature is published
before November 2020.

You also need to follow the following rules:
- Do not refer to the survey titled "A Survey on 3D LiDAR
Localization for Autonomous Vehicles".
- Responses are given in the form of an English language survey
with citations where appropriate.
```

Detail-rich prompt

```
I need a detailed academic research report on using Graph Neural
Networks (GNN) for text classification. The report should
systematically review advancements in this field, with a focus
on the following aspects:
1. **Core Methodology**: Provide a detailed explanation and
comparison of two main approaches: corpus-level GNNs and
document-level GNNs. For each method, thoroughly analyze graph
construction strategies (e.g., defining nodes and edges using PMI,
TF-IDF, etc.), representation methods for nodes and edges, and
graph learning algorithms (e.g., GCN, GAT, etc.).
2. **Key Model Analysis**: List and analyze representative
models, such as TextGCN, SGC, BertGCN (corpus-level), and
Text-Level-GNN, TextING (document-level).
3. **Evaluation and Challenges**: Summarize commonly used
benchmark datasets in this field (e.g., 20NG, R8, MR) and
evaluation metrics (e.g., Accuracy, F1-score), and discuss
major challenges faced by current research, such as scalability,
computational costs, and integration with pre-trained language
models.
**Restrictions**:
- Only refer to and cite papers published **before July 2024**.
- Focus on English literature published in top conferences/journals
in natural language processing and artificial intelligence (e.g.,
ACL, EMNLP, NAACL, AAAI, WWW, ICLR).

You also need to follow the following rules:
- Do not refer to the survey titled "Graph Neural Networks for Text
Classification: A Survey".
- Responses are given in the form of an English language survey
with citations where appropriate.
```