

REPORTBENCH: EVALUATING DEEP RESEARCH AGENTS VIA ACADEMIC SURVEY TASKS

000
001
002
003
004
005 **Anonymous authors**
006 Paper under double-blind review
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030

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 publicly.

1 INTRODUCTION

031
032
033
034
035
036 The rapid development of LLM-powered Deep Research agents has revolutionized the process of
037 knowledge synthesis by enabling autonomous execution of extensive research tasks, including aca-
038 demic literature surveys, industry analyses, and market assessments (Chen et al., 2025; Gottweis
039 et al., 2025; Lu et al., 2024; Tang et al., 2025; Yamada et al., 2025; Zheng et al., 2025; Li et al.,
040 2025). Tasks that traditionally required days or weeks of manual effort can now be completed
041 within minutes. Notable examples include advanced systems such as OpenAI (OpenAI, 2025) and
042 Google’s Gemini Deep Research (Google, 2025), which effectively integrate various external tools
043 and perform multiple rounds of deep reasoning. Despite their promising capabilities, widespread
044 practical adoption critically depends on their ability to consistently deliver research reports with
045 high factual accuracy and comprehensive content quality. Therefore, it is essential to monitor and
046 ensure the quality of generated reports through evaluation. However, defining what constitutes a
047 good report is challenging and lacks broad consensus, resulting in the current absence of mature
048 evaluation methodologies for research report generation.

049 In addressing this challenge, we decompose the evaluation of research reports generated by LLMs
050 into two core dimensions: writing quality and report content. Due to the subjectivity of writing-
051 style evaluation, while the criteria for assessing content quality can be more clearly defined, this
052 work focuses primarily on the evaluation of report content, leaving the assessment of writing quality
053 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.

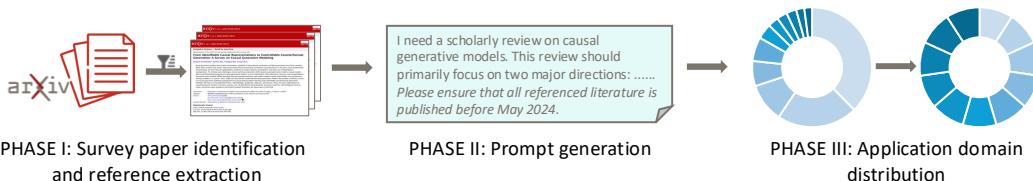
054 To establish a high-quality benchmark capable of rigorously assessing research reports, we propose
 055 ReportBench, a novel evaluation framework leveraging expert-generated literature reviews. Given
 056 the constraints of relying on human annotators, who typically vary in expertise, we propose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

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

071 In the second phase of our validation pipeline, we design two different verification procedures based
 072 on whether a statement includes an explicit citation to external literature. Specifically, for cited
 073 statements, the system identifies all in-text citations within the report, maps each citation to its
 074 corresponding source document, and employs semantic matching to ensure factual support from
 075 the cited literature. For non-cited statements, the framework employs a voting mechanism across
 076 multiple web-connected models to verify the factuality of these statements. By combining these
 077 complementary validation procedures, ReportBench delivers a systematic and detailed assessment
 078 of AI-generated research reports, ensuring the relevance and quality of cited literature and the factual
 079 accuracy of all claims through citation-based and web-based validation. [To validate these automatic](#)
 080 [metrics and better understand errors, we further conduct a human evaluation and qualitative error](#)
 081 [analysis on a subset of reports. The human evaluation results exhibit high agreement with our](#)
 082 [automatic evaluation pipeline, further confirming the quality of ReportBench.](#)

083 Our contributions can be summarized as follows:

- 084 • We present **ReportBench**, a systematic benchmark designed to evaluate the quality of research
 085 reports generated by Deep Research agents, with a focus on the quality of references and the
 086 factual accuracy of all statements presented in the report.
- 087 • We propose an automated and scalable data synthesis method for constructing academic sur-
 088vey tasks, including prompts and ground truth, from expert-authored survey papers on arXiv.
 089 Additionally, we introduce an automatic agentic evaluation framework that evaluates the pre-
 090 cision and recall of the generated report with respect to the ground-truth references and performs
 091 factual verification of individual claims made within the report.
- 092 • We release a comprehensive benchmark suite—datasets, prompts, and evaluation scripts—to
 093 support reproducible research and community-driven progress in evaluating LLM-based
 094 knowledge synthesis.



102 Figure 1: Overall benchmark data construction workflow.
 103
 104

2 METHODOLOGY

105 We introduce **ReportBench**, a comprehensive evaluation framework designed to rigorously assess
 106 Deep Research agents through two interconnected components: (i) the automated construction of

108 high-quality benchmark datasets derived from expert-authored survey papers, and (ii) a systematic
 109 validation pipeline that evaluates the quality and factual consistency of AI-generated research
 110 reports. In the following sections, we detail the processes that underlie the synthesis of the dataset and
 111 the design of our evaluation workflow.
 112

113 2.1 DATASET CONSTRUCTION 114

115 In this section, we detail the end-to-end pipeline to construct high-quality deep research questions
 116 along with ground-truth answers based on published survey papers. This workflow comprises three
 117 consecutive phases: (i) survey paper identification and reference extraction, (ii) prompt generation,
 118 and (iii) application domain distribution. A diagram illustrating the data construction process is
 119 presented in Figure 1.
 120

121 2.1.1 PHASE I: SURVEY PAPER IDENTIFICATION AND REFERENCE EXTRACTION 122

123 The first step is to identify high-quality survey papers to create evaluation tasks. We start from the
 124 complete arXiv metadata snapshot (arXiv.org submitters, 2024) and retain papers submitted on or
 125 after 2020-01-01. To ensure the quality of papers, we only select those that have undergone peer
 126 review and have been formally published. We achieve this by using regular expressions, *i.e.*, querying
 127 over titles to match “*survey*” or “*review*” to filter survey papers and searching “*published*” or
 128 “*accepted*” in the comments field of a submission. To reduce systematic false positives in domains
 129 such as astronomy, we prompted GPT-4○ (Hurst et al., 2024) with each paper’s title and abstract to
 130 produce a binary classification of whether the paper is a literature survey.
 131

132 For each survey paper, we analyze its LaTeX source file to extract cited references. Specifically, we
 133 parse LaTeX citation commands, identify and retrieve relevant bibliographic entries from associated
 134 bibliography databases, and filter these to retain only references explicitly cited in the main text.
 135 Hence, the extracted bibliography mirrors the true citation pattern of the paper. The resulting dataset
 136 constitutes a gold-standard benchmark for evaluating retrieval precision. Finally, we retained 678
 137 papers.
 138

139 2.1.2 PHASE II: PROMPT GENERATION 140

141 Survey papers can be regarded as a great depth of research work focused on a specific topic at a
 142 specific time, making it possible to create deep research questions in a *reverse prompt engineering*
 143 manner. In other words, given the publication date and the full text of a survey paper obtained
 144 through a PDF parsing tool, we prompt an LLM to generate a query whose ideal answer is precisely
 145 that paper. Hence, we obtain a query and its ground truth (the survey paper itself). To increase the
 146 diversity of prompts, we design three types of prompt templates:
 147

148 **Sentence-level prompt**

149 A single sentence that succinctly defines the overarching
 150 academic field covered by the survey.
 151

152 **Paragraph-level prompt**

153 A short paragraph elaborating the research area, its main
 154 subtopics, and the methodological perspectives covered in the
 155 survey.
 156

157 **Detail-rich prompt**

158 A detailed question that comprehensively describes the
 159 specific research domain, key research directions, and
 160 the methodological approaches of interest. Additional
 161 constraints may be included, such as preferred conferences
 162 or journals, language of the cited literature (*e.g.*, English,
 163 Chinese), participating institutions or laboratories.
 164

165 In addition, to ensure that LLMs’ retrieval window matches the survey’s citation horizon and pre-
 166 vents leakage of post-publication knowledge, we require each generated prompt to include a cut-off
 167

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.”

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. As some tested systems integrate search tools internally, tool-side restrictions cannot be applied for fair comparison. 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

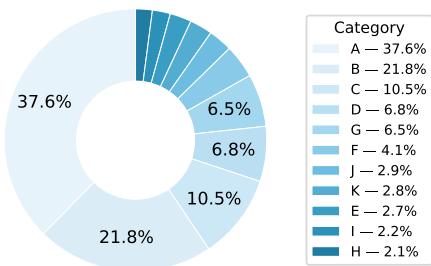
2.1.3 PHASE III: APPLICATION DOMAIN DISTRIBUTION

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		

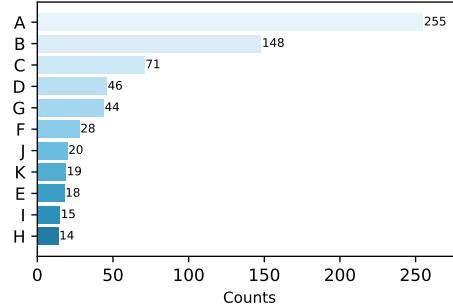
The distribution of prompts across these domains is inherently biased due to the specific disciplinary focus of the arXiv corpus, as shown in Figure 2. [From the pool of 678 candidate surveys, we assign each survey to an application-domain bucket and perform stratified random sampling so that domains contribute roughly equally.](#) To create a balanced and general test set, we sample a fixed number of surveys per domain (10 in our implementation) using uniform random sampling with a fixed seed (random seed = 42), yielding 100 tasks in total without any additional hand-picking [beyond the earlier filters](#). 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 **ReportBench**. The quality of the classification of this subset was then reviewed and validated by four research experts.

Category Distribution (%)



(a) Category distribution (pie).

Category Counts



(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.

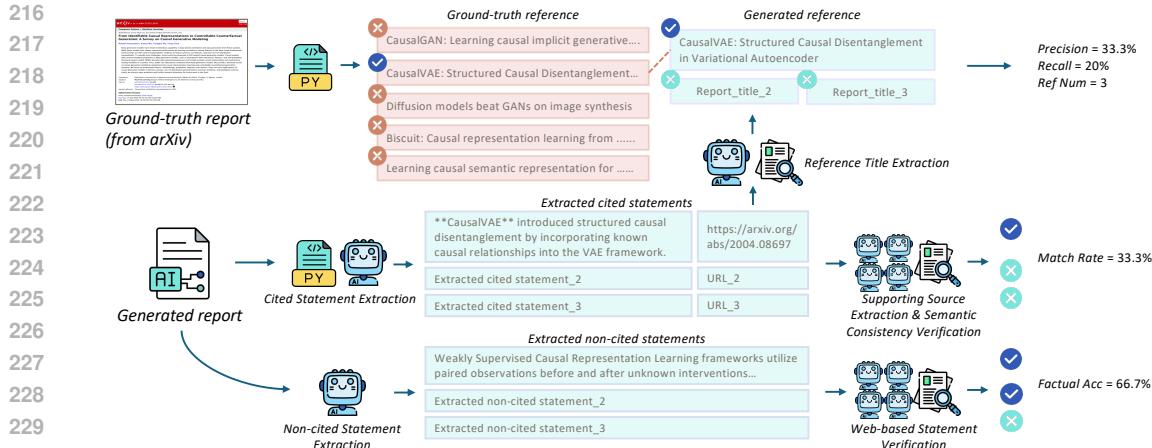


Figure 3: Evaluation Process.

2.2 EVALUATION PROCESS

Our evaluation process, as shown in Figure 3, 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. **Content quality** is first evaluated by assessing the cited references: we compare the reference list 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. **Note that the time limit is enforced via the ground-truth bibliography: for each task we only include references published on or before the source survey date and exclude the survey itself, so any citation to post-cutoff papers or to the source survey lies outside the ground-truth set is counted as an error, automatically lowering precision/recall even if the model ignores the instruction.** **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. Prompts for evaluation are presented in Appendix A.4.

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 models. 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.

In our current implementation, we normalize and match URL-style citations, as models are instructed to cite using URLs and the released evaluator assumes this format. This is a design choice rather than a fundamental limitation: the same extraction–normalization pipeline can be extended to BibTeX entries, arXiv IDs, and DOIs by adding format-specific extractors and mapping them to canonical URLs (e.g., [https://doi.org/...](https://doi.org/), [https://arxiv.org/abs/...](https://arxiv.org/abs/)) before comparison with the ground-truth bibliography, enabling systems to use their native citation styles without changing the evaluation protocol.

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

270 comparing the statement with the retrieved content, and the results are aggregated to compute an
 271 overall citation consistency score for the report. Unlike traditional “LLM-as-a-judge” approaches,
 272 which often suffer from instruction-following issues or biased scoring, our method decomposes the
 273 evaluation into fine-grained, interpretable, and verifiable steps. All intermediate outputs are retained
 274 for optional human inspection, thereby maximizing the reliability and transparency of the evaluation
 275 process.

276
 277 **Non-cited Statements.** We use a simple two-step validation process. First, we extract all factual
 278 statements in the report that do not have any citations, and remove content that is general common
 279 sense or already supported by references. Then, we ask several web-connected LLMs to check
 280 each statement independently. Each model looks up information online and gives its judgment. We
 281 combine their answers using a voting mechanism to decide whether the statement is likely to be
 282 correct. This approach avoids relying on a single model and makes the validation more reliable.

283 3 EXPERIMENT

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

290 3.1 SETTINGS

291 Our evaluation pipeline uses different LLMs for distinct components. For statement extraction,
 292 supporting source extraction, and semantic consistency verification, we adopt GPT-4o. For the fact-
 293 checking of non-cited statements, we employ two web-connected models: Gemini-2.5-Pro and
 294 Gemini-2.5-Flash. Each model performs three independent judgments per statement, resulting
 295 in a total of six verdicts. The final decision is determined by majority voting, and the proportion of
 296 votes is recorded as a confidence score. In the evaluation of base models, we integrated search and
 297 link-reading tools using each model’s native function call interface. Specifically, we used SerpAPI¹
 298 for Google Search access and Firecrawl² for retrieving web pages in Markdown format. Due to
 299 context length limitations, we capped the maximum number of tool calls at five per instance.

300 To evaluate the performance of both Deep Research agents and base models, we manually collected
 301 responses from the web-based interfaces of OpenAI and Gemini, as well as batch-executed outputs
 302 from the base models, [during the time window from July 14–25, 2025. Hence, the present results](#)
 303 [correspond to the July 2025 snapshot](#). During data collection, we ensured that OpenAI was using the
 304 standard version of Deep Research, powered by the o3 model. For Gemini, we made sure that both
 305 the “Gemini 2.5 Pro” and “Deep Research” toggles were enabled on the web interface to activate its
 306 full research capabilities.

307 3.2 EVALUATION METRICS

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

318 ¹<https://serpapi.com/>

319 ²<https://www.firecrawl.dev/>

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
o3	0.299	0.031	12.26	31.43%	16.16	82.22%	11.51
Claude-4-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 over three times as many references per report (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 in generating reports. OpenAI Deep Research achieves a higher citation match rate than Gemini (78.87% vs 72.94%) while producing 88.2 cited statements on average, 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.

Our evaluation pipeline relies on automatic URL-to-paper mapping and LLM-based judgments of citation correctness and factuality. To assess the reliability, we also conduct a small-scale human expert study to validate our automatic evaluation pipeline and find high agreement (typically 84–96%) between expert judgments and our URL mapping, statement-level factuality checks, and citation-level precision/recall metrics, indicating that the pipeline is well-aligned with domain experts. For more details, please refer to Appendix A.2.3.

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 (OpenAI, 2025).

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 $\circ 3$, 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.

382 **Gemini Deep Research vs. Gemini-2.5-Pro**

383 Similarly, Gemini Deep Research and its base model Gemini-2.5-Pro diverge significantly
 384 across multiple dimensions. Gemini Deep Research trades off some precision (0.145 vs. 0.269)
 385 to achieve much higher recall (0.036 vs. 0.010) and generates far more references per report (32.42
 386 vs. 4.27). In terms of statement volume, it produces many more cited statements (96.2 vs. 6.58)
 387 and non-cited statements (49.6 vs. 9.35). Despite this increase in volume, its citation alignment re-
 388 mains strong (72.94% vs. 59.24%), while its non-cited statement accuracy is slightly lower than the
 389 base model (92.21% vs. 96.08%). These pronounced gaps—in precision/recall trade-off, citation
 390 count, and overall coverage—mirror the contrast observed between OpenAI Deep Research and $\circ 3$,
 391 and suggest that the system has undergone targeted optimization for thorough research and report
 392 generation. Taken together with the visible “plan” and “step-by-step reasoning” phases presented in
 393 the Gemini Deep Research web interface, it seems plausible that the system functions more like a
 394 thoughtfully constructed multi-agent workflow or pipeline.

395 **Base-Model Comparison**

396 Among the four base models, Claude-4-Sonnet demonstrates the most balanced perfor-
 397 mance—achieving a precision of 0.337, a recall of 0.021, an average of 6.74 reference documents
 398 per report, a high citation semantic consistency (73.67%), and a strong non-cited statement factual
 399 accuracy (92.64%). In contrast, Gemini-2.5-Pro attains higher precision (0.269) at the expense
 400 of recall (0.010) and generates fewer reference documents on average (4.27 per report), limiting its
 401 coverage. Gemini-2.5-Flash underperforms on both precision (0.237) and recall (0.012), with
 402 lower citation semantic consistency (44.88%), indicating poorer citation relevance. Meanwhile, $\circ 3$
 403 produces the most references (12.26 per report) and moderate recall (0.031), but its citation semantic
 404 consistency (31.43%) and non-cited statement accuracy (82.22%) lag behind.

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

408 4 ANALYSIS

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

416 **Statement Hallucination.** In our manual audit of arXiv:2407.15186 test cases, we identified rep-
 417 resentative errors in statement generation. For example, OpenAI Deep Research generated the fol-
 418 lowing claim:

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

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

429 **Citation Hallucination.** During our evaluation of arXiv:2009.12619, we observed a clear instance
 430 of link hallucination in the generated report from Gemini-2.5-Pro. The model generated the
 431 claim:

432 In-vehicle Crowd Monitoring: The use of surveillance cameras inside buses and
 433 trains for passenger counting is a well-established practice. Advanced image pro-
 434 cessing and computer vision techniques can automatically analyze video feeds to
 435 estimate the passenger load. For instance, a system was proposed to estimate the
 436 number of passengers in a bus using image processing techniques on the captured
 437 video frames, achieving high accuracy. [Vision-Based In-Vehicle Crowd Moni-
 438 toring](https://www.researchgate.net/publication/224217198_A_vision-based_system_for_in-vehicle_crowd_monitoring).
 439

440 However, the cited URL does not exist and appears to be entirely fabricated by the model. Because
 441 the link cannot be resolved, no supporting text or evidence can be retrieved to validate the state-
 442 ment, resulting in a citation mismatch. This example highlights a common error mode in function-
 443 call-driven retrieval: the model confidently invents plausible-looking reference links that nonethe-
 444 less point to nothing, undermining factual grounding.
 445

446 These examples demonstrate that even advanced Deep Research agents remain susceptible to hallu-
 447 cinating author names, misaligning citations, and fabricating links. Crucially, our evaluation met-
 448 rics—especially citation semantic consistency—are sensitive to such discrepancies, allowing us to
 449 quantitatively capture and penalize these hallucination phenomena across model outputs.
 450

451 5 RELATED WORK

452 Long-standing interest has been in the use of AI to synthesize information, not only in the writing
 453 of scientific articles (Chen et al., 2025; Gottweis et al., 2025; Lu et al., 2024; Tang et al., 2025;
 454 Yamada et al., 2025), but also in the search for information and the generation of reports in the
 455 general domains (Zheng et al., 2025; Li et al., 2025). With the rapid advancement of information
 456 synthesis research, the evaluation of long-form reports has become increasingly important.
 457

458 **Fact Checking Evaluation** Driven by efforts from both academia and industry, automated fact
 459 checking has evolved into a well-established multistage pipeline, which has become the dominant
 460 research paradigm in the field (Eldifrawi et al., 2024). Claim detection aims to identify factual
 461 statements worth verifying from large volumes of text (Guo et al., 2022; Panchendarajan & Zubiaga,
 462 2024), while evidence retrieval focuses on retrieving relevant documents or textual snippets that
 463 support or refute a given claim (Eldifrawi et al., 2024; Nanekhan et al., 2025). Building on this
 464 pipeline, several benchmarks have been proposed to evaluate the performance of fact checking in
 465 both the general domain (Thorne et al., 2018; Ma et al., 2024) and the scientific domain (Wadden
 466 et al., 2020; 2022; Ho et al., 2025). However, these benchmarks focus solely on fact-checking
 467 components, rather than evaluating the synthesized information as a whole.
 468

469 **Citation Evaluation** Research reports often include a substantial amount of citation-related content,
 470 and evaluating the precision and standardization of these citations plays a crucial role in assessing
 471 the overall quality of the report (Sarol et al., 2024). Given a report with citation content, tasks such as
 472 cited context identification, evidence sentence retrieval, and citation accuracy classification are
 473 commonly used to analyze citation quality (Sarol et al., 2024). Widely applied in assisted paper writing
 474 and review systems, citation verification tools are designed from multiple perspectives, including
 475 syntactic verification, existence verification, and semantic verification (Barrot, 2025; Bairagi & Li-
 476 hitkar, 2024). While citation correctness and existence have been well-studied, the completeness of
 477 citations remains underexplored.
 478

479 **Survey Generation** With the advent of LLMs, automated survey generation has seen rapid progress.
 480 Early works leveraged LLMs to improve literature comprehension and survey writing (Wang et al.,
 481 2024; Hu et al., 2025), achieving better coherence compared to sentence extraction methods. Sub-
 482 sequent research explored structured and hierarchical organization with fixed references. Other
 483 approaches focused on modeling paper relationships via citation networks, including AutoSurvey
 484 (Wang et al., 2024) with a two-stage LLM pipeline and HiReview (Hu et al., 2025) with a taxonomy-
 485 driven framework, though both faced limitations in capturing human writing styles or relying on
 486 restricted citation scopes. More recently, SurveyForge (Yan et al., 2025) combines human outline
 487 structure analysis with high-quality literature retrieval, generating and refining full survey content
 488 through a scholar navigation agent. Compared with SurveyBench, ReportBench focuses solely on
 489 well-defined and automatically verifiable dimensions of evaluation. In addition, through an au-
 490

486 tomatized construction pipeline, it ensures data quality while offering clear scalability advantages,
 487 enabling it to serve as a potential source of training data for report optimization in future work.
 488

489 **Deep Research Evaluation** The rise of deep research agents (DRAs), driven by powerful models
 490 such as ChatGPT (OpenAI, 2025) and Gemini (Google, 2025), has underscored the urgent need
 491 for robust and targeted evaluation methodologies. While existing benchmarks evaluate capabilities
 492 such as web retrieval (Wei et al., 2025; Zhou et al., 2025; Wu et al., 2025), multi-hop factual reasoning
 493 (Wei et al., 2024; Mialon et al., 2024; Phan et al., 2025), and end-to-end report generation (Du
 494 et al., 2025; Bosse et al., 2025). These methods often operate at a surface level and fall short of evaluating
 495 the core competencies essential for rigorous and reliable research. **Compared to DeepResearch**
 496 **Bench** (Du et al., 2025), **ReportBench** differs in three key ways: (i) it proposes a largely automatic,
 497 survey-driven data construction pipeline that is easier to scale than the manually authored tasks in
 498 DeepResearch Bench; (ii) it provides a citation-level gold bibliography for each task, enabling precise
 499 measurement of reference precision/recall rather than relying solely on LLM-as-a-judge scores;
 500 and (iii) it performs statement-level factuality checking for both cited and non-cited claims, yielding
 501 more fine-grained diagnostics of hallucination, over-citation, and under-citation.

502 6 CONCLUSION

503 In this paper, we present **ReportBench**, a comprehensive benchmark for evaluating the quality of
 504 references and the factual accuracy of all statements in reports generated by Deep Research agents.
 505 By leveraging expert-authored survey papers as ground truth and reverse prompt engineering, we enable
 506 consistent evaluation of AI-generated research reports across multiple dimensions. Our framework
 507 introduces a fine-grained validation workflow that separately assesses cited and non-cited statements,
 508 combining citation semantic consistency checks and web-based factual verification. Through
 509 large-scale experiments on leading LLM-based research agents and the base models, we demonstrate
 510 that Deep Research products can outperform base models in content coverage and factual grounding,
 511 but still face challenges in hallucination, over-citation, etc. We hope that ReportBench will serve as
 512 a valuable tool for the research community to monitor, compare, and further improve the reliability
 513 of AI systems designed for academic survey tasks.

516 7 ETHICS STATEMENT

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

521 **Data Distribution.** The benchmark is predominantly constructed from peer-reviewed STEM (Science,
 522 Technology, Engineering, and Mathematics) survey papers on arXiv, which induces a STEM-centric
 523 bias and means that our results should not be over-interpreted as measuring general deep-
 524 research ability in social sciences, humanities, or other under-represented domains. At the same
 525 time, the data-construction pipeline itself is domain-agnostic: given a corpus of survey-like papers
 526 and basic metadata, the procedure in Section 2.1 can be applied to other disciplines. We view instantiating
 527 domain-specific variants (e.g., for social sciences and law) as important future work.

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

533 **Efficiency and cost.** ReportBench is intentionally a quality-focused benchmark: we evaluate report
 534 correctness and depth, but do not report comparative latency or cost, as token usage and end-to-end
 535 delays were not systematically logged during the original runs and cannot be reliably reconstructed
 536 ex post. A systematic study of efficiency—including proper instrumentation, request-level logging,
 537 and controlled load conditions—is an important direction for future work and a natural extension of
 538 the current benchmark.

540 **8 REPRODUCIBILITY STATEMENT**
 541

542 We will open-source ReportBench in full, including all constructed prompts, the ground-truth refer-
 543 ence list for each example, and metadata of the source surveys (arXiv ID, title, authors, comments,
 544 etc.). We will also release the complete evaluation code used in this work; users only need to provide
 545 API keys for the external services specified in the repository (e.g., search and web retrieval) to run
 546 end-to-end assessments of their generated reports on ReportBench. The repo will include configura-
 547 tion files and scripts to reproduce our pipelines, along with instructions to re-run the tested baselines
 548 and to evaluate new model outputs. To comply with the terms of service of the evaluated products
 549 and models, we will not publish our generated reports during evaluation; instead, we provide the ex-
 550 act prompts, evaluation scripts, and scoring logic so that others can independently obtain evaluated
 551 model outputs under their own accounts and reproduce the paper’s results.
 552

553 **REFERENCES**
 554

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

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

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

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

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

568 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
 569 Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
 570 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
 571 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

572 Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench:
 573 A comprehensive benchmark for deep research agents. *CoRR*, abs/2506.11763, 2025. doi: 10.
 574 48550/ARXIV.2506.11763. URL <https://doi.org/10.48550/arXiv.2506.11763>.

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

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

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

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

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

600 Yunlong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjani, Boxin Zhao, and Liang
 601 Zhao. Taxonomy tree generation from citation graph, 2025. URL <https://arxiv.org/abs/2410.03761>.

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

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

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

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

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

620 Kevin Nanhekhan, Venkatesh V, Erik Martin, Henrik Vatndal, Vinay Setty, and Avishek Anand.
 621 Flashcheck: Exploration of efficient evidence retrieval for fast fact-checking. In Claudia
 622 Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio
 623 Pinelli, Fabrizio Silvestri, and Nicola Tonello (eds.), *Advances in Information Retrieval -*
 624 *47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10,*
 625 *2025, Proceedings, Part IV*, volume 15575 of *Lecture Notes in Computer Science*, pp. 385–399.
 626 Springer, 2025. doi: 10.1007/978-3-031-88717-8_28. URL https://doi.org/10.1007/978-3-031-88717-8_28.

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

628 Rubaa Panchendarajan and Arkaitz Zubiaga. Claim detection for automated fact-checking: A
 629 survey on monolingual, multilingual and cross-lingual research. *Nat. Lang. Process. J.*, 7:
 630 100066, 2024. doi: 10.1016/J.NLP.2024.100066. URL <https://doi.org/10.1016/j.nlp.2024.100066>.

631 Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael
 632 Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Jason Hausenloy, Oliver Zhang,
 633 Mantas Mazeika, Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, Steven Y.
 634 Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Jessica P. Wang,
 635 Pawan Kumar, Oleksandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Mstyslav
 636 Kazakov, Johannes Schmitt, Geoff Galgon, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott

648 Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Zachary Giboney,
 649 Gashaw M. Goshu, Joan of Arc Xavier, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah
 650 Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, John Wydallis,
 651 Mark Nandor, Ankit Singh, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Darling Duclosel, Jung-
 652 bae Nam, Jennifer Zampese, Ryan G. Hoerr, Aras Bacho, Gautier Abou Loume, Abdallah Galal,
 653 Hangrui Cao, Alexis C. Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Us-
 654 man Qazi, Lianghui Li, Sumeet Motwani, Christian Schröder de Witt, Edwin Taylor, Johannes
 655 Veith, Eric Singer, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G.
 656 Willcocks, Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont,
 657 Justine Leon Uro, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward
 658 Vendrow, Kengo Zenitani, Julien Guillod, Yuqi Li, Joshua Vendrow, Vladyslav Kuchkin, and
 659 Ng Ze-An. Humanity’s last exam. *CoRR*, abs/2501.14249, 2025. doi: 10.48550/ARXIV.2501.
 660 14249. URL <https://doi.org/10.48550/arXiv.2501.14249>.

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

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

666 James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-
 667 scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent
 668 (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association
 669 for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp.
 670 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:
 671 10.18653/v1/N18-1074. URL <https://aclanthology.org/N18-1074/>.

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

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

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

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

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

694 Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang,
 695 Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web
 696 traversal. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
 697 (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

702 (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pp. 10290–
703 10305. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.acl-long.508/>.

704

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

709

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

713

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

717

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

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A APPENDIX**
757758 **A.1 USE OF LLMs**
759760 This work involved both human effort and the assistance of large language models (LLMs) in sev-
761 eral stages, with all outputs subject to human oversight and verification. During experiment im-
762 plementation, humans designed the codebase framework and core logic, while AI-based IDE tools
763 in combination with Claude-4-Sonnet and Claude-4-Opus were used to generate code. To ensure
764 quality, intermediate results were logged and manually inspected at each evaluation stage to prevent
765 the propagation of error. For related work discovery, we used OpenAI and Gemini’s Deep Research
766 to assist in surfacing potentially relevant papers, followed by manual reading, summarization, and
767 selection. In the writing process, OpenAI’s GPT-4o and GPT-4.1 models were employed to polish
768 drafts written by humans, focusing on improving grammar and clarity.
769770 **A.2 ADDITIONAL ANALYSES**
771772 **A.2.1 CITATION RECALL BY REFERENCE IMPORTANCE**
773774 To better understand which papers current agents tend to recover, we complement the global citation
775 recall with a stratified analysis by reference importance. Here we use citation count as a simple
776 proxy for how central or influential a paper is.
777778 For each survey-level task, we take the ground-truth reference list and look up the citation count of
779 every paper in a standard scholarly index³. Within that single survey, we then sort its references by
780 citation count and cut at the 25%, 50%, and 75% quantiles. This yields four groups: Q1 contains
781 the least-cited 25% of references within that survey, Q2 the next 25%, Q3 the 50–75% range, and
782 Q4 the most-cited 25%. We deliberately define Q1–Q4 per survey (rather than globally) because
783 citation distributions can vary greatly across domains and we want “Q4” to always mean “the most
784 central references for this particular survey”, independent of field-specific scale differences.
785786 Given these per-task quartiles, we compute stratified recall for each model by aggregating over all
787 tasks: for Q1, for example, we collect all references that fall into Q1 across all surveys and measure
788 what fraction were correctly cited by the model; we repeat the same calculation for Q2, Q3, and Q4.
789 Table 2 reports the resulting recall values (in percent).
790788 **Table 2: Citation recall by reference-importance quartile on ReportBench. Q1–Q4 partition refer-
789 ences within each survey by citation count (from lowest to highest). Values are recall in %.**
790791

Model	Q1	Q2	Q3	Q4
Gemini-2.5-Flash	0.4	0.7	0.8	1.8
Gemini-2.5-Pro	0.3	0.6	0.9	1.3
o3	1.0	1.8	3.0	3.5
Claude-4-Sonnet	1.0	1.8	2.3	2.3
Gemini Deep Research	1.8	2.7	3.0	3.3
OpenAI Deep Research	1.3	2.3	3.2	3.3

801 Across all systems, recall consistently increases from Q1 to Q4: models recover substantially more
802 of the most highly cited references than of the long tail. For instance, o3 improves from 1.0% recall
803 in Q1 to 3.5% in Q4, and similar trends hold for the other systems. This stratified view shows that
804 current agents are much more likely to retrieve the core, high-impact references than obscure or
805 marginal ones.
806807 We also observe that the two deep-research products (Gemini Deep Research and OpenAI Deep
808 Research) achieve uniformly higher recall across all quartiles than base-models. This indicates
809³<https://www.semanticscholar.org/product/api>

810
 811 Table 3: Subjective content quality scores (1–5) and gaming rates (%) for each model on Report-
 812 Bench. Values are means across tasks; higher is better for D, S, I, F, and SV. Abbreviations: D
 813 = Content Depth, S = Synthesis & Structure, I = Insight & Trends, F = Future Directions, SV =
 Survey-ness, G = Gaming Rate.

Model	D	S	I	F	SV	G (%)
Gemini-2.5-Pro	3.137	2.756	2.397	2.023	2.779	6.1
Claude-4-Sonnet	3.930	3.364	3.039	3.248	3.519	1.6
o3	4.066	3.582	3.270	3.041	3.721	2.3
Gemini Deep Research	4.560	4.160	4.070	4.070	4.480	0.0
OpenAI Deep Research	4.107	3.756	3.550	3.756	3.954	0.8

825 that ReportBench is sensitive enough to capture progress in literature coverage, rather than merely
 826 penalizing systems on the very long tail of rare citations.

828 A.2.2 SUBJECTIVE CONTENT QUALITY EVALUATION

830 Our original design deliberately focused on objectively checkable aspects of quality, emphasizing
 831 the quality of references and the factual accuracy of all statements in the report. To further assess
 832 content quality and understand how it correlates with these initially defined objective dimensions,
 833 we introduce an additional subjective content quality evaluation component to capture survey-style
 834 qualities that are difficult to measure purely through reference- and fact-level checks.

835 Concretely, we define six dimensions intended to characterize the extent to which a report functions
 836 as a genuine survey:

- 838 • **Content Depth (1–5):** How thoroughly the report covers the main subtopics of the field, dis-
 839 cussing key methods, trade-offs, and limitations beyond superficial descriptions.
- 840 • **Synthesis & Structure (1–5):** How well prior work is organized into coherent themes or tax-
 841 onomies and related to each other, rather than being listed in isolation.
- 842 • **Insight & Trend Analysis (1–5):** Whether the report draws non-trivial patterns and trends across
 843 works, explaining underlying design principles and helping readers quickly understand the land-
 844 scape.
- 845 • **Future Directions & Open Problems (1–5):** How concretely and convincingly the report pro-
 846 poses future research directions or open problems grounded in the surveyed literature.
- 848 • **Survey-ness (1–5):** A holistic judgment of the extent to which the report functions as a genuine
 849 survey paper in the ML/CS sense, integrating depth, synthesis, and insight into a useful starting
 850 point for researchers.
- 851 • **Gaming / Degenerate Behavior (binary):** Whether the report resembles a trivial “bag-of-
 852 sentences” output (e.g., near-pure per-paper listing without synthesis) versus a genuine attempt
 853 at survey-style writing.

855 In Table 3, we abbreviate these six dimensions as *D* (Content Depth), *S* (Synthesis & Structure), *I*
 856 (Insight & Trends), *F* (Future Directions), *SV* (Survey-ness), and *G* (Gaming Rate).

857 We then use LLM-as-a-judge to score each report along these six axes. For the first five dimensions
 858 we use a 1–5 scale; for the last dimension we record a binary *Gaming* flag and report its empirical
 859 rate. Table 3 shows the aggregated results over all tasks, and we highlight three key findings:

- 861 1. **Deep-research products score highest on content quality.** Gemini Deep Research and OpenAI
 862 Deep Research achieve the best scores across all content dimensions (with Gemini generally first
 863 and OpenAI second), indicating that the products are also better at producing genuinely survey-
 like, synthesized reports—not just at citing correctly.

864 2. **Gaming behavior is rare and detectable.** The *Gaming* flag is triggered only in a small fraction
 865 of cases (0–6.1% across systems), with the specialized deep-research agents having near-zero
 866 gaming rates. This suggests that, in practice, systems that would try to game the factual met-
 867 rics by stitching together isolated sentences are both uncommon and explicitly penalized by this
 868 additional rubric.

869 3. **Content quality and objective metrics are strongly aligned.** When we rank systems by the
 870 holistic *SV* score and by our original citation/factuality metrics, the induced partial order over
 871 systems is nearly identical. This indicates that our objective, reference- and factuality-based
 872 metrics already capture much of the underlying report quality, and that the new content-quality
 873 scores provide an interpretable confirmation rather than contradicting the original evaluation.

874 **A.2.3 HUMAN EXPERT VALIDATION OF THE EVALUATION PIPELINE**

875 Our evaluation pipeline relies on automatic URL-to-paper mapping and LLM-based judgments of
 876 citation correctness and factuality. To assess the reliability of this pipeline, we conduct a small-scale
 877 human validation study in which domain experts manually re-evaluate a stratified sample of system
 878 outputs.

879 We randomly sample results across systems and tasks (with stratification by model), and for each
 880 sampled report we evaluate the following components:

881 (a) **URL-to-paper mapping accuracy.** Experts manually resolve each system-generated reference
 882 URL to a canonical scholarly record and compare it with our automatic mapping. The effective
 883 agreement is 96.7%, combining (i) 80% strict agreement on verifiable references and (ii) 16.7%
 884 cases where the model produces unverifiable or fabricated URLs, which should not be counted
 885 as disagreement with the mapping pipeline because these hallucinated URLs have no correct
 886 resolution to begin with. Only 3.3% of cases correspond to genuine technical retrieval failures.
 887 Since unverifiable URLs are treated as incorrect references by construction, they do not inflate
 888 citation scores, and these residual failures have negligible impact on system ranking.

889 (b) **Cited-statement consistency.** For statements in the report that explicitly cite a paper, experts
 890 read both the statement and the cited paper and decide whether the statement faithfully repre-
 891 sents the cited work. We then compare these labels with the outputs of our LLM-based citation
 892 judge. The human and LLM labels agree on roughly **90%** of evaluated statements, suggesting
 893 that the automatic judge is reasonably aligned with expert judgments in this setting.

894 (c) **Factuality of non-cited statements.** For a subset of statements without explicit citations,
 895 experts manually fact-check the content (using the web and standard scholarly search engines)
 896 and assign a binary factuality label. We compare these labels against our multi-model, web-
 897 augmented factuality pipeline. Agreement is high, on the order of **96%**, indicating that our
 898 factuality assessment is reliable even when no explicit reference is present.

899 (d) **Citation-level metrics.** We then ask experts to manually compute citation-level precision and
 900 recall on the same sampled outputs and compare these values with the corresponding Report-
 901 Bench scores produced by our pipeline. The two sets of results show an agreement rate of
 902 84%, indicating that the final citation metrics closely track expert judgments despite minor
 903 intermediate errors.

904 (e) **Prompt and gold-label validation.** Finally, experts examine the automatically constructed
 905 prompts and gold bibliographies for a subset of tasks, checking for faithfulness to the underly-
 906 ing source surveys (no leakage, correct temporal cutoff). In roughly **95%** of the sampled tasks,
 907 experts fully agree with the automatically derived prompt and ground-truth bibliography; the
 908 remaining cases are minor edge conditions that do not affect our main conclusions.

909 Taken together, these results suggest that our automatic evaluation pipeline is well-aligned with
 910 expert judgments across all of its major components, and that residual discrepancies are small com-
 911 pared to the performance gaps observed between systems.

912 **A.2.4 JUDGE-FAMILY SENSITIVITY: GEMINI VS. GPT**

913 One concern with using a Gemini-family model as the LLM-as-a-judge is the possibility of a hidden
 914 family-level bias in favor of Gemini-generated reports. To probe this, we conduct a sensitivity study
 915 with an independent judge model from a different provider.

918
 919 Table 4: Comparison of Gemini vs. GPT as LLM-as-a-judge on three representative systems. Left
 920 block: Gemini judge (same as in the main paper). Right block: GPT judge. Prec. = Precision, Rec.
 921 = Recall, MR = Match Rate, Fact. = Factual Accuracy.
 922

923 924 Model	925 Gemini judge				926 GPT judge			
	Prec.	927 Rec.	928 MR (%)	929 Fact. (%)	Prec.	930 Rec.	931 MR (%)	932 Fact. (%)
Gemini-2.5-Pro	0.269	0.010	59.24	96.08	0.290	0.011	52.55	91.44
o3	0.299	0.031	31.43	82.22	0.321	0.031	32.01	72.61
Claude-4-Sonnet	0.337	0.021	73.67	92.64	0.365	0.024	64.21	84.06

933
 934 Specifically, we re-run our citation–factuality evaluation using a GPT-series model as the LLM-
 935 as-a-judge, while keeping everything else fixed: the system reports, the gold references, and the
 936 evaluation pipeline (URL mapping, statement extraction, and scoring logic) are identical to the main
 937 experiments. Table 4 compares the original Gemini-judge setup (as in the paper) with the new GPT-
 938 judge setup on three representative systems. Precision/Recall are reported on the 0–1 scale; Match
 939 Rate and Factual Accuracy are reported in percent.
 940

941 We obtain two key observations. First, the induced model ranking is unchanged. Under both Gemini
 942 and GPT judges, the relative ordering of systems on citation precision and recall is identical, and
 943 consistent with the overall benchmark ordering. In other words, swapping the judge from Gemini to
 944 GPT perturbs the absolute scores slightly but does not change which systems are better or worse on
 945 our main metrics.
 946

947 Second, GPT is somewhat stricter at statement-level evaluation. For all three test models, the GPT-
 948 based judge yields lower match rates and factuality accuracies than the Gemini-based judge, indi-
 949 cating a more conservative standard for accepting statements as supported. This suggests that our
 950 original Gemini-based evaluation is, if anything, slightly optimistic rather than biased in favor of
 951 Gemini-generated reports.
 952

953 Overall, this sensitivity check shows that our main findings are robust to swapping the judge family,
 954 and we do not see evidence that using Gemini-based judges materially advantages Gemini systems
 955 relative to others.
 956

957 A.2.5 TOOL-CALL BUDGET ABLATION

958 In the main experiments, we cap the number of tool calls at 5 per task, primarily for practical reasons
 959 such as context length, latency, and cost. To understand how sensitive our results are to this design
 960 choice, we run an explicit ablation on a representative baseline (Gemini-2.5-Pro), keeping everything
 961 else fixed and varying the maximum number of tool calls.
 962

963 Table 5: Effect of the maximum tool-call budget B on citation and factuality metrics for Gemini-
 964 2.5-Pro on ReportBench. Precision/Recall are on the 0–1 scale; Match Rate and Factual Accuracy
 965 are in %.
 966

967 Max Tool Calls B	Precision	968 Recall	969 Match Rate (%)	970 Factual Acc. (%)
968 3	0.249	0.008	50.32	96.79
969 5 (paper setting)	0.269	0.010	59.24	96.08
970 10	0.275	0.008	57.20	95.71

971 Table 5 reports the resulting citation and factuality metrics when the maximum tool-call budget B
 972 is set to 3, 5, or 10. Precision and recall are reported on the 0–1 scale, while Match Rate and Factual
 973 Accuracy are reported in percent.
 974

975 We observe two main trends. First, increasing the budget from 3 to 5 tool calls yields a clear but
 976 modest improvement across citation precision, recall, and match rate, indicating that allowing a few
 977

972 additional searches helps the baseline discover more relevant literature. Second, while moving from
 973 5 to 10 calls does slightly increase precision and occasionally match rate, the overall gains are unsta-
 974 ble: recall and factual accuracy fluctuate or even decline. In practice, repeatedly fetching full papers
 975 quickly pushes against the model’s context-length limit, and without additional mechanisms for con-
 976 text compression or longer-term memory, simply raising the tool-call budget does not translate into
 977 consistently better use of the retrieved evidence.

978 Collectively, these results suggest that the baselines reach a practical performance plateau at around
 979 5 tool calls. This configuration captures most of the benefit from additional search, while higher bud-
 980 getts offer diminishing and noisy returns under our current architecture. We therefore treat the 5-call
 981 setting as a reasonable, saturated operating point; extending the budget further does not materially
 982 change the comparative conclusions of the study.

983 A.2.6 FAILURE MODES AND ERROR TAXONOMY

984 To complement the quantitative metrics, we add a qualitative error taxonomy based on manual in-
 985 spection of representative failure cases from multiple systems. Starting from our original distinc-
 986 tion between statement- vs. citation-level hallucinations, we further refine errors into more concrete cat-
 987 egories. In a manually annotated sample of problematic reports, we observe three dominant types.

988 **Temporal-cutoff violations (~42%).** The agent cites papers that clearly post-date the survey’s
 989 publication (e.g., referencing 2024–2025 work in a task whose cutoff is 2021). These are often
 990 otherwise reasonable references, but they break the historical constraint and indicate that the agent
 991 is effectively “peeking into the future” instead of reconstructing the literature as of the survey date.
 992 In one representative case, the agent cites the paper “*Federated Learning Security and Privacy-
 993 Preserving Algorithm and Experiments Research Under Internet of Things Critical Infrastructure*”
 994 as part of the core literature. However, the task explicitly enforces a temporal cutoff of July 2022,
 995 while this paper appears to have been published around September 2023. This citation is therefore
 996 not counted as a valid match in ReportBench and is categorized as a temporal-cutoff violation: the
 997 reference is thematically relevant but violates the historical constraint on what was knowable at the
 998 time of the original survey.

999 **Unverifiable references (~21%).** The report contains citations that appear plausible in style (au-
 1000 thors, venue, year) but cannot be resolved to any real paper (no DOI/arXiv/URL match), or whose
 1001 content contradicts the summary in the text. These are classic citation hallucinations and remain a
 1002 major source of error. For example, in one failure case the agent writes a plausible paragraph on
 1003 manifold learning for multimedia and cites a paper titled “*Manifold Learning for Music Information
 1004 Retrieval*” with a link to a ResearchGate page. When we attempt to resolve this citation against
 1005 standard bibliographic sources, we cannot find a corresponding, stable publication record with full
 1006 metadata. In our pipeline, such references are treated as fabricated or unverifiable: the citation looks
 1007 syntactically reasonable and thematically relevant, but does not map to a concrete paper in the gold
 1008 bibliography or in standard indices, and therefore counts as a hallucinated citation rather than valid
 1009 prior work.

1010 **Misaligned research direction (~9%).** The agent drifts to a neighboring but different topic, re-
 1011 sulting in citations and discussion that are coherent in themselves but misaligned with the intended
 1012 survey focus. Typical cases include focusing on generic foundation models when the task is specif-
 1013 ically about long-context retrieval models, or emphasizing broad “AI in healthcare” literature when
 1014 the survey is about a particular subproblem such as continual learning for medical imaging. In such
 1015 cases, many cited papers are real and technically relevant to the broader area, but they do not answer
 1016 the concrete survey question posed in the task.

1017 This expanded taxonomy clarifies not only that systems fail, but also how they fail, and it highlights
 1018 concrete targets for future improvement—for example, stronger temporal control, stricter reference
 1019 verification, and better task grounding at the prompt and planning stages.

1026 A.3 EXAMPLE PROMPTS IN REPORTBENCH
1027

1028

1029

Sentence-level prompt

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

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.

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080
 1081 **Detail-rich prompt**
 1082 I need a detailed academic research report on using Graph
 1083 Neural Networks (GNN) for text classification. The report
 1084 should systematically review advancements in this field, with
 1085 a focus on the following aspects:
 1086 1. **Core Methodology**: Provide a detailed explanation
 1087 and comparison of two main approaches: corpus-level GNNs
 1088 and document-level GNNs. For each method, thoroughly analyze
 1089 graph construction strategies (e.g., defining nodes and edges
 1090 using PMI, TF-IDF, etc.), representation methods for nodes
 1091 and edges, and graph learning algorithms (e.g., GCN, GAT,
 1092 etc.).
 1093 2. **Key Model Analysis**: List and analyze representative
 1094 models, such as TextGCN, SGC, BertGCN (corpus-level), and
 1095 Text-Level-GNN, TextING (document-level).
 1096 3. **Evaluation and Challenges**: Summarize commonly
 1097 used benchmark datasets in this field (e.g., 20NG, R8,
 1098 MR) and evaluation metrics (e.g., Accuracy, F1-score), and
 1099 discuss major challenges faced by current research, such
 1100 as scalability, computational costs, and integration with
 1101 pre-trained language models.
 1102 **Restrictions**:
 1103 - Only refer to and cite papers published **before July 2024**.
 1104 - Focus on English literature published in top
 1105 conferences/journals in natural language processing and
 1106 artificial intelligence (e.g., ACL, EMNLP, NAACL, AAAI, WWW,
 1107 ICLR).
 1108 You also need to follow the following rules:
 1109 - Do not refer to the survey titled 'Graph Neural Networks
 1110 for Text Classification: A Survey'.
 1111 - Responses are given in the form of an English language
 1112 survey with citations where appropriate.

1114 A.4 PROMPTS IN EVALUATION

1116 A.4.1 CITED STATEMENT EXTRACTION

1118 You are given a research report delimited by triple
 1119 backticks.
 1120 Identify every statement that cites an external source (e.g.
 1121 has a URL, DOI, or explicit citation marker) and pair it with
 1122 the corresponding URL.
 1123 Return a JSON list where each item has two keys:
 1124 - "statement": the single-sentence claim, stripped of
 1125 leading/trailing whitespace
 1126 - "url": the canonical URL that supports that claim
 1127 If a citation contains multiple URLs, duplicate the statement
 1128 for each URL.
 1129 ONLY return valid JSON. Report: '```{report}```'

1130
 1131
 1132
 1133

1134 A.4.2 NON-CITED STATEMENT EXTRACTION
 1135
 1136 You are given a research report delimited by triple
 1137 backticks.
 1138 You are also given a list of statements that already have
 1139 citations.
 1140 Your task is to identify factual claims or statements that:
 1141 1. Make specific assertions about facts, data, or events
 1142 2. Are NOT already included in the cited statements list
 1143 3. Could potentially be verified through external sources
 1144 4. Are NOT common knowledge or widely accepted facts
 1145 Exclude:
 1146 - Opinions, analysis, or subjective interpretations
 1147 - Statements that are already cited
 1148 - Common knowledge or universally accepted facts
 1149 - Vague or general statements
 1150 Return a JSON list where each item has one key:
 1151 - "statement": the factual claim that lacks citation support
 1152 ONLY return valid JSON.
 1153 Report:
 1154 ```{report}```
 1155 Already cited statements:
 1156 {cited_statements}
 1157

A.4.3 SUPPORTING SOURCE EXTRACTION

1158
 1159
 1160 You are provided with
 1161 Statement: {statement}
 1162
 1163 Source Document:
 1164 {source_text}
 1165
 1166 Return any relevant content from the source document that
 1167 supports the statement. This can be a sentence, paragraph,
 1168 or even the entire text if necessary.
 1169 If no content supports it, return "NOT_FOUND".
 1170 Return plain text only.
 1171
 1172

A.4.4 SEMANTIC CONSISTENCY VERIFICATION

1173
 1174
 1175 You will decide whether a claim is correctly supported by a
 1176 source sentence.
 1177
 1178 Claim from report:
 1179 {statement}
 1180
 1181 Source Sentence from original source:
 1182 {source_sentence}
 1183
 1184 Respond with JSON containing:
 1185 - "reason": one short sentence explaining your decision
 1186 - "match": true or false // true if the source sentence
 1187 faithfully supports the claim
 1188 Return ONLY the JSON.

1188 A.4.5 WEB-BASED STATEMENT VERIFICATION
 1189
 1190 You are tasked with verifying the accuracy of a factual
 1191 statement using web search capabilities.
 1192
 1193 Statement to verify:
 1194 {statement}
 1195
 1196 Please:
 1197 1. Use web search to find reliable, authoritative sources
 1198 about this statement
 1199 2. Analyze the information you find from multiple sources
 1200 3. Determine if the statement is factually correct or
 1201 incorrect based on your research
 1202
 1203 Respond with JSON containing:
 1204 - "reason": a detailed explanation of your verification
 1205 process and findings (2-3 sentences)
 1206 - "decision": true if the statement is correct, false if it
 1207 is incorrect
 1208
 1209 Only return the JSON response.

1210 A.4.6 REFERENCE TITLE EXTRACTION
 1211

1212 Please analyze the following academic survey and extract all
 1213 cited academic paper titles and author information.
 1214
 1215 Survey content:
 1216 {response}
 1217
 1218 Please reply in JSON format, containing an array named
 1219 'papers', where each paper object includes the following
 1220 fields:
 1221 - title: the title of the paper
 1222 - authors: a list of authors
 1223 - is_academic_paper: true (indicating this is an academic
 1224 paper)
 1225
 1226 Example format:
 1227 {
 1228 "papers": [
 1229 {
 1230 "title": "Deep Learning for Natural Language
 1231 Processing",
 1232 "authors": ["John Smith", "Jane Doe"],
 1233 "is_academic_paper": true
 1234 },
 1235 ...
 1236 }
 1237
 1238 Note: Only extract explicitly mentioned academic papers.
 1239 Do not include books, websites, or other types of references.