WIKI LIVE CHALLENGE: CHALLENGING DEEP RESEARCH AGENTS WITH EXPERT-LEVEL WIKIPEDIA ARTICLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Research Agents (DRAs) show great potential in assisting humans with complex research tasks. Yet existing DRA evaluations are limited by scalability, as models iterate rapidly, benchmarks require continuous updates, while high-quality references and criteria depend on substantial human effort that cannot scale at low cost. Existing benchmarks have yet to achieve both reference quality and scalability simultaneously. To bridge this gap, we introduce **Wiki Live Challenge (WLC)**, a scalable, live benchmark that leverages the newest Wikipedia Good Articles (GAs) as expert-level references. The Wikipedia community continuously produces expert-reviewed GAs, enabling WLC to acquire high-quality new tasks at minimal cost. Grounded in Wikipedia's GA criteria, we collect 100 recent GAs and propose **Wiki Eval**, a comprehensive evaluation framework comprising Wiki Writing, a fine-grained writing assessment with 39 criteria, and Wiki Fact, rigorous metrics for factual verifiability. Extensive experiments on various DRA systems demonstrate a significant gap between current DRAs and human expert-level Wikipedia articles, especially in factual information gathering and coverage.

1 INTRODUCTION

With the explosive growth of Large Language Model (LLM) capabilities, LLM-driven agents have demonstrated remarkable potential in handling expert-level tasks (Achiam et al., 2023; Yang et al., 2025; Team et al., 2025a; Mialon et al., 2024). Among these, the Deep Research Agent (DRA) represents one of the most advanced agent systems (Team et al., 2025b; Qiao et al., 2025; Li et al., 2025a; Zheng et al., 2025). Through multi-step web information retrieval, integration, and reasoning, DRAs can complete research tasks that would typically require significant time and effort from human experts.

Existing DRAs still suffer from hallucinations and biases in research and writing (Zhan et al., 2026). To comprehensively evaluate these systems, high-quality evaluation tasks and assessment criteria are essential. Existing works (Du et al., 2025; Li et al., 2025b) have attempted to use reports generated by strong DRAs as references to evaluate others, scoring them based on manually designed dimensions such as comprehensiveness and writing quality. Other approaches (Sharma et al., 2025; Xu et al., 2025a) design rubrics for reports collected from the web or annotated by humans.

As illustrated in the left panel of Figure 1, these methods are limited by scalability. As models advance rapidly, existing benchmarks can quickly become saturated, creating a growing need for fresh evaluation tasks. Yet constructing high-quality reference articles and evaluation criteria requires substantial human effort and time, making it difficult for benchmark updates to keep pace with model progress. LLM-generated references are easy to scale but lack quality assurance, and their evaluation criteria may deviate from human expert expectations (Fan et al., 2025; Li et al., 2024). Human-annotated rubrics offer higher quality, but are often insufficiently fine-grained or require laborious expert verification, making it difficult to cost-effectively reconstruct evaluation sets after model iterations.

To bridge this gap, we introduce the **Wiki Live Challenge (WLC)**, a scalable, live benchmark that challenges DRAs with the latest Wikipedia Good Articles (GAs). At the task level, the Wikipedia community continuously produces new GAs through expert review, enabling WLC to acquire new

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

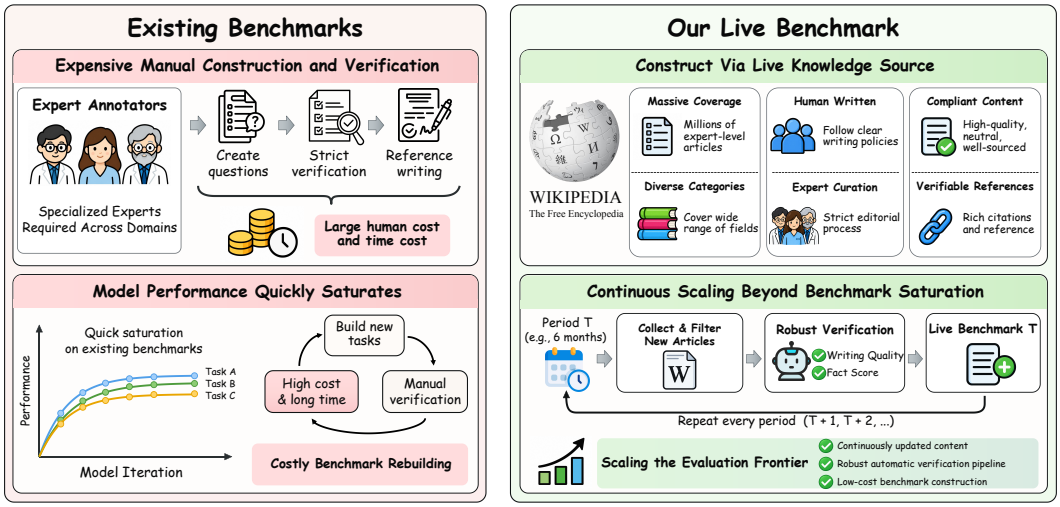


Figure 1: **The Need for Live Benchmarks.** Left: Existing benchmarks require substantial expert effort for construction and verification, while rapid model progress leads to saturation and costly rebuilding. Right: Wikipedia provides diverse, expert-reviewed articles with established editorial criteria and verifiable references. WLC continuously collects new articles each period and evaluates writing quality and factual verifiability through Wiki Eval, enabling scalable live benchmarking.

tasks at near-zero content creation cost, as illustrated in the right panel of Figure 1. All GAs in WLC were newly created after the knowledge cutoff of current mainstream models[1]. At the evaluation level, GA criteria serve as community-refined expert consensus standards that generalize across domains without requiring task-specific rubric design. Wikipedia articles represent comprehensive research on a subject, involving extensive information gathering and organization, written from an objective and neutral perspective with strict verifiability. We posit that producing such content can rigorously test core DRA capabilities, yet current models remain especially limited in factual information gathering.

Specifically, for the main evaluation set, we collect 100 high-quality Wikipedia Good Articles that strictly follow Wikipedia's editorial guidelines and have been reviewed and revised by human experts, providing strong quality assurance. Using these articles as human-expert references, we construct **Wiki Eval**, an evaluation framework with two complementary dimensions. Based on the Wikipedia Good Article criteria, we develop Wiki Writing, a fine-grained writing evaluation protocol with 39 criteria covering writing quality, neutrality, and coverage. We also introduce Wiki Fact, a factual evaluation method with two sub-metrics that measure (i) information richness of generated articles relative to Wikipedia and (ii) whether generated statements are strictly traceable to supporting sources. Our contributions are summarized as follows:

- We introduce **Wiki Live Challenge (WLC)**, a live benchmark for evaluating DRAs. By leveraging the Wikipedia GA review process as a continuously running, low-annotation-cost source of expert-level content, WLC achieves high scalability and timeliness.
- We propose **Wiki Eval**, an evaluation framework covering both writing quality and factuality, with all criteria strictly grounded in Wikipedia Good Article standards.
- We conduct extensive experiments across a diverse set of DRA systems and perform comprehensive analyses and human studies to validate the reliability of our evaluation framework, revealing a significant gap between current DRAs and human expert-level Wikipedia articles, especially in factual information gathering.

[1]This timeframe ensures the content postdates the knowledge cutoff of current mainstream models, mitigating potential data contamination.

2 Related Work

2.1 Deep Research Agent

Deep Research Agents (DRAs) are designed to autonomously explore the web, retrieve information, and synthesize findings into comprehensive reports. Recent progress focuses on improving their reasoning and planning for long-term tasks. Notably, DeepResearcher (Zheng et al., 2025) is the first work to train LLMs via end-to-end reinforcement learning in a real, dynamic web environment for deep information retrieval and integration. Similarly, Tongyi DeepResearch (Team et al., 2025b) employs an end-to-end training framework to enable scalable reasoning. WebResearcher (Qiao et al., 2025) treats research as a decision process, using iterative refinement to manage noise. WebSailor (Li et al., 2025a) tackles uncertainty in web navigation through structured sampling. Furthermore, systems like WebDancer (Wu et al., 2025) have pushed the boundaries of autonomous information seeking. These developments highlight a shift towards agents that can independently verify information and synthesize knowledge.

2.2 Deep Research Benchmarks

Evaluating the capabilities of DRAs requires benchmarks that go beyond simple question answering. Early general agent benchmarks such as GAIA (Mialon et al., 2024) and AgentBench (Liu et al., 2025) assess fundamental abilities like tool use and reasoning but often lack the depth required for evaluating long-form research reports. FreshWiki (Shao et al., 2024) is an early dataset utilizing Wikipedia articles to evaluate generated text, but it lacks fine-grained rubrics tailored for modern DRAs. DeepResearch Bench (Du et al., 2025) provides 100 PhD-level tasks across 22 domains, employing reference-based adaptive criteria. ReportBench (Li et al., 2025b) focuses on report generation quality using survey papers as references, while ResearchRubrics (Sharma et al., 2025) offers expert-written criteria for evaluating open-ended queries. More recent benchmarks further emphasize live tasks and expert-grounded evaluation, including LiveResearchBench (Wang et al., 2025), DeepScholar-Bench (Patel et al., 2025), and DEER (Han et al., 2025). Additionally, BrowseComp (Wei et al., 2025) and CiteEval (Xu et al., 2025b) focus on citation accuracy and source attribution. Despite these efforts, existing benchmarks either rely on model-generated references that lack rigorous human verification, or require costly expert annotation that cannot scale with rapid model iterations.

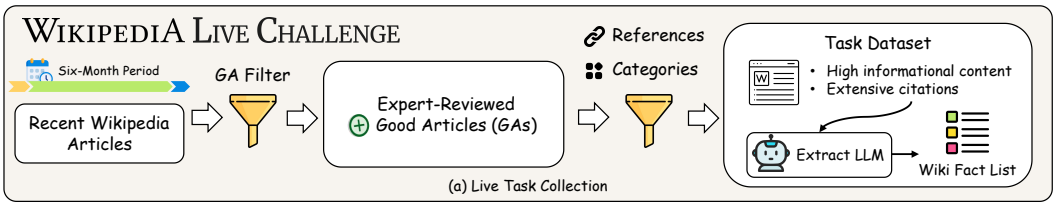
3 Wiki Live Challenge

This section presents WLC in detail, covering our data construction methodology and the Wiki Eval evaluation framework grounded in Wikipedia Good Article criteria. The overall framework is illustrated in Figure 2.

3.1 Data Construction

Existing DRA benchmarks attempt to use reports generated by strong models or collected from the open web as references (Du et al., 2025; Wang et al., 2025; Patel et al., 2025; Han et al., 2025), but struggle to achieve both quality and scalability. These sources often lack expert evaluation and may contain biases or errors from the generative models or human authors. In contrast, Wikipedia is rigorously reviewed and revised by human editors, adhering to strict core policies on neutrality, comprehensiveness, and verifiability. Good Articles (GAs) represent the highest editorial standards through a meticulous expert review process, and the community continuously promotes new GAs across diverse domains, providing inherent scalability. Authoring such content requires extensive research, the synthesis of diverse viewpoints, and the maintenance of objectivity, serving as a rigorous test of DRA capabilities. We therefore adopt GAs as reference standards and ground our evaluation in the official Good Article criteria.

Specifically, in each collection period, we collect and organize newly created Wikipedia articles from a time window chosen to fall after the evaluated models' knowledge cutoffs. From this corpus, we select Good Articles (GAs) that have passed Wikipedia's review process and strictly adhere to the Good Article criteria, thus ensuring high quality. To ensure the complexity of the task, we rank these articles based on the number of reference URLs and structural depth, excluding simple list-based



Good Article Criteria

Eval	Dimension	Criteria
Wiki Writing	Well-written (21)	Articles should be written in encyclopedic style, which differs from the technically dense style...
		Articles should address the topic without twisting the truth or telling "lies-to-children"...
		.....
	Neutral (10)	It addresses the main aspects of the topic.
		The lead, ideally the introductory sentence or at least introductory paragraph, of an article...
		.....
	Broad in its coverage (8)	It represents viewpoints fairly and without editorial bias, giving due weight to each.
		Avoid stating opinions as facts...
		.....
Wiki Fact	Verifiable	It contains no original research...
		All content that could reasonably be challenged...
		.....

(b) GA Criteria List

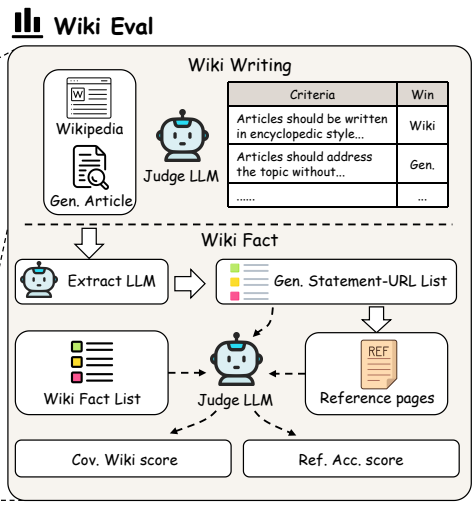


Figure 2: **The overview of our Wiki Live Challenge (WLC) benchmark.** (a) We continuously collect recent Wikipedia articles (e.g., from Mar. 1 to Dec. 1 in first iteration), filter the latest expert-reviewed Good Articles, and build the live task dataset. (b) Our evaluation is grounded in GA criteria: well-written, neutral, broad coverage and verifiable. (c) Our evaluation framework, Wiki Eval, incorporates two key dimensions: Wiki Writing and Wiki Fact.

articles. Ultimately, we curate 100 Wikipedia Good Articles for each collection period. Figure 3 presents the article distribution, dataset statistics, and a representative task case of the main evaluation set, covering 15 major domains. WLC is highly scalable, as updating to new collection periods incurs minimal human effort.

3.2 Evaluation Methodology

Our evaluation framework is grounded in the Wikipedia Good Article criteria, a widely recognized standard established by human experts[2]. We selected four dimensions most relevant to DRA capabilities, focusing on writing style, neutral point of view, broad in coverage, and verifiability. Based on these, we constructed **Wiki Eval**, an evaluation methodology comprising two primary components: Wiki Writing and Wiki Fact.

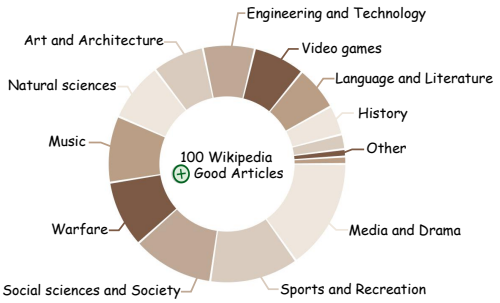
3.2.1 Wiki Writing

To assess writing quality, we construct Wiki Writing, a fine-grained evaluation framework based on the three core dimensions of the Wikipedia Good Article criteria: *Well-written*, *Neutral*, and *Broad in its coverage*. This framework comprises 39 distinct criteria derived directly from official Wikipedia writing guidelines, as shown in Figure 2-*b*. We employ an LLM-as-a-Judge approach: for each criterion, the Judge LLM determines a winner between the Wikipedia article w_i and the generated article g_i:

$$\text{Judge}(w_i, g_i) = \text{Judge-LLM}(w_i, g_i). \tag{1}$$

We aggregate the win rates across all criteria to compute the overall writing score. The complete list of criteria is provided in Appendix C.

[2] https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria



Metric	Min	Max	Average
Words per Article	1012	12047	3052
References per Article	14	281	48.3
Facts per Article	54	459	149

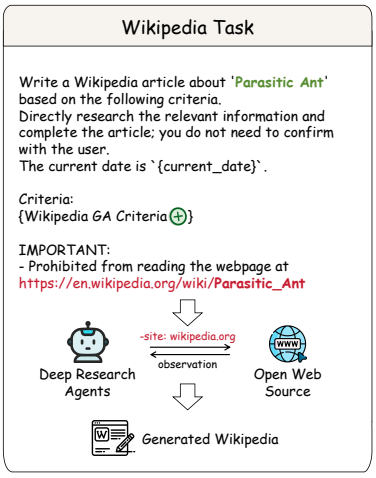


Figure 3: **Overview of the WLC benchmark dataset.** The left panel displays the distribution of collected Wikipedia Good Articles across 15 major categories and the key statistics of the WLC Benchmark Dataset. The right panel illustrates a representative task case.

3.2.2 WIKI FACT

To evaluate the article's factual accuracy (1) relative to Wikipedia and (2) relative to cited references, we introduce two evaluation metrics, as illustrated in Figure 2-*c*. We first preprocess both the Wikipedia article and the generated article by employing an extraction LLM to extract facts from each, yielding a fact list for the Wikipedia article and a Statement-URL pair list for the generated article.

For factual accuracy **relative to Wikipedia**, we measure the coverage of generated statements against Wikipedia facts. For each fact f_i in the Wikipedia fact list, we retrieve the top-10 most relevant statements from the generated article.[3] These statements and the target fact are then input into a fact-checking model to determine a consistency score:

$$\text{Fact}(f_i, G) = \begin{cases} 1, & \text{if consistent} \\ 0, & \text{if inconsistent} \\ 0, & \text{if not_support} \end{cases} \tag{2}$$

We then calculate the coverage score for each article by averaging the scores of all its target facts:

$$\text{Cov. Wiki.} = \frac{1}{|F|} \sum_{f_i \in F} \text{Fact}(f_i, G) \tag{3}$$

where F represents the set of facts extracted from the Wikipedia article, and G represents the set of facts from the corresponding generated article.

For factual accuracy **relative to references**, we assess whether the generated statements are supported by their citations. For each extracted Statement-URL pair, we utilize Jina Reader[4] to retrieve the content of the cited webpage. We then employ the fact-checking model to verify if the statement is supported by source content. Statement–URL pairs whose cited pages cannot be retrieved are excluded from the Ref. Acc. denominator.

The final Reference Accuracy score is calculated as the proportion of statements that are fully supported by their references set R:

$$\text{Ref. Acc.} = \frac{1}{|S|} \sum_{s_i \in S} \text{Fact}(s_i, R) \tag{4}$$

[3] A sensitivity analysis over $k \in \{5, 10, 20, \text{full-article}\}$ is reported in Appendix J.

[4] `https://jina.ai`

where S denotes the list of Statement-URL pairs in the generated article for which the cited page is successfully retrieved.

4 Experimental Settings

4.1 Implementation Details

For Cov. Wiki, candidate statements are retrieved by cosine similarity using OpenAI `text-embedding-3-small`, retaining the top-10 candidates for each Wikipedia fact before verification. For the evaluation of Wiki Writing, we use Gemini-2.5-Pro (Comanici et al., 2025) as the judge LLM. For Wiki Fact, we employ Gemini-2.5-Flash for both fact extraction and verification, balancing performance with the cost of processing large token volumes. Generated reports are evaluated against the corresponding Wikipedia reference snapshots. Table 1 presents the main evaluation results, with results from a new collection period provided in Appendix D. Further implementation details are provided in Appendix A.1.

4.2 Evaluated Models

In this work, we extensively evaluated advanced Deep Research Agent systems, including proprietary systems such as OpenAI o3 Deep Research (OpenAI, 2025), Gemini-2.5-pro Deep Research, Gemini-3-pro Deep Research (Google, 2025), and Qwen-3-max Deep Research (Yang et al., 2025). Regarding open-source frameworks, we selected Tongyi DeepResearch (Team et al., 2025b) and Deep Researcher (Zheng et al., 2025), two state-of-the-art open-source models trained via reinforcement learning. Additionally, to disentangle the underlying LLM's capability from the surrounding agent scaffold, we evaluate an **open-source harness paired with different LLM backbones**: we uniformly adopt LangChain Open Deep Research[5] as the shared deep-research scaffold, holding all tools, prompts, and hyperparameters identical, and swap in different LLM backbones such as GPT-5, Gemini-3-pro, and Kimi-K2.5.

5 Results and Discussions

5.1 Main Results

5.1.1 Evaluation on Wiki Writing

As shown in Table 1, Gemini-3-pro Deep Research and LangChain (GPT-5) demonstrate the strongest Wiki Writing performance, with LangChain (Kimi-K2.5) also competitive within the shared harness. We observe substantial performance disparities across different DRAs. Notably, fully open-source DRA models lag significantly behind: Deep Researcher achieves only 2.28, with manual review revealing that its reports are often incomplete after minimal information-gathering. Tongyi DeepResearch performs relatively better and approaches some proprietary systems like Doubao, yet a gap remains, indicating that matching proprietary long-form report generation is still challenging for smaller end-to-end RL-trained models. Within the controlled LangChain harness, the Wiki Writing ranking still tracks backbone capability (e.g., GPT-5 and Kimi-K2.5 lead; GPT-4.1 and DeepSeek-V3.2 trail), suggesting that the cross-system gaps are not merely a harness artifact.

5.1.2 Evaluation on Wiki Fact

Considering Wiki Fact in Table 1, all DRA systems are far from expert-level coverage: the best-performing agent, LangChain (Kimi-K2.5), reaches only 31.71% on Cov. Wiki, indicating that current models are still far from reaching the expert-level information gathering capabilities of Wikipedia. Under the controlled LangChain harness, scores scale smoothly with backbone capability, and LangChain (Kimi-K2.5) even surpasses every proprietary system on Cov. Wiki, indicating that part of the cross-system variation reflects underlying LLM differences rather than product-specific pipelines. A symmetric precision-side counterpart, Prec. Wiki, is reported in Appendix K as an auxiliary observation.

[5] `https://github.com/langchain-ai/open_deep_research`

Table 1: **Main results of WLC across Wiki Writing and Wiki Fact.** Wiki Writing scores are win rates aggregated over 39 GA-based criteria. Cov. Wiki measures factual coverage against Wikipedia, and Ref. Acc. measures citation supportability. Ref. Acc. is not reported for Deep Researcher and Tongyi DeepResearch because these systems do not produce extractable citation markers.

Models	**Wiki Writing**				**Wiki Fact**	
	Overall	Well-written	Neutral	Broad	Cov. Wiki	Ref. Acc.
Open-Source Agent Framework						
Deep Researcher	2.28	1.90	1.90	3.75	5.62	–
Tongyi Deep Research	15.05	10.90	11.60	30.25	22.73	–
Open-Source Harness with LLMs						
GPT-4.1	20.67	18.76	19.40	27.25	7.08	7.34
GPT-5	<u>53.62</u>	<u>50.95</u>	**54.20**	<u>59.88</u>	20.96	**67.60**
Gemini-2.5-pro	40.92	42.10	30.60	50.75	21.20	59.62
Gemini-3-pro	40.77	40.14	36.20	48.13	16.26	60.74
Qwen-3-max	48.36	47.81	43.20	56.25	25.56	39.96
DeepSeek-V3.2	26.24	24.05	25.05	33.46	14.13	43.63
Kimi-K2.5	52.49	50.78	<u>53.47</u>	55.74	**31.71**	42.58
Proprietary Agent Framework						
Doubao Deep Research	19.13	16.00	16.10	31.13	22.97	37.05
Qwen-3-max Deep Research	25.15	18.29	27.70	40.00	22.22	61.44
Perplexity Deep Research	27.38	26.79	20.40	37.63	29.21	28.27
Grok Deep Search	28.38	27.52	24.30	35.75	20.73	60.63
OpenAI o3 Deep Research	31.08	28.43	24.90	45.75	25.12	57.44
Gemini-2.5-pro Deep Research	35.18	30.10	26.10	<u>59.88</u>	<u>30.76</u>	41.68
Gemini-3-pro Deep Research	**58.33**	**60.81**	46.10	**67.12**	28.83	<u>66.98</u>

To further investigate this gap, we present a case study on the *Parasitic Ant* article in Figure 4. The heatmap reveals a difficulty gradient aligned with article structure: DRAs perform well on procedural sections like *Methods* but struggle with specialized sections such as *Phylogeny* and *Defense*. Analysis shows that while general definitions are universally covered, all systems fail to retrieve precise quantitative data (e.g., gene counts) and domain-specific terminology. This suggests that agents effectively capture broad concepts but lack the precision for granular, high-difficulty details.

Regarding Reference Accuracy, scores for Deep Researcher and Tongyi DeepResearch are omitted due to unreliable citation formatting. Among the remaining systems, LangChain (GPT-5) achieves the highest accuracy, while Perplexity scores lower as it only cites a subset of retrieved content. Due to opacity in internal search mechanisms and issues such as paywalls, this metric serves as a supplementary measure of verifiability rather than definitive grounding. Beyond aggregate Cov. Wiki and Ref. Acc., we further dissect failure modes through Wiki Conf. and Ref. Conf., which respectively measure direct contradictions with Wikipedia facts and with cited references; full results and discussion are deferred to Appendix L.

5.2 Analysis and Discussion

5.2.1 Analysis Across Categories

Difficulty Our benchmark comprises Wikipedia articles from 15 distinct categories, revealing significant performance disparities among models across these domains. Notably, in *History* and *Mathematics*, the average win rate across all systems remained below 20%, whereas in *Natural Sciences* and *Philosophy and Religion*, average scores exceeded 40%. This highlights substantial variation in the difficulty of information retrieval and summarization across fields. We further analyzed the correlation between Wikipedia statistical features and category win rate, as detailed in Table 2. Here, lower category win rates indicate higher difficulty. Results indicate that category win rate has negligible correlation with article length or citation count but is moderately correlated with total page views. This suggests that higher view counts typically correspond to more popular

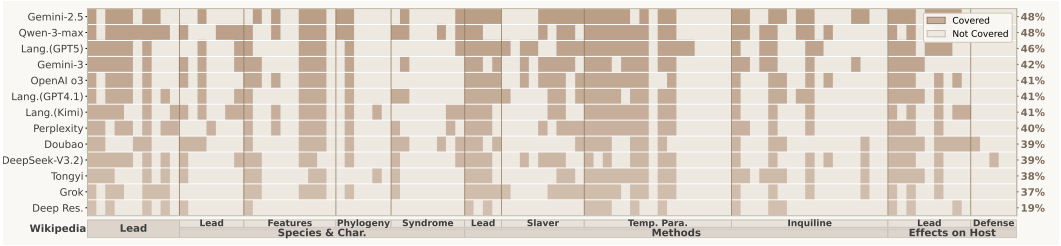


Figure 4: **Fact Coverage Heatmap on Wikipedia Good Article "Parasitic Ant".** The x-axis represents individual facts ordered by their appearance in the article sections, and the y-axis represents different DRAs.

categories, where information is more accessible for DRAs to mine. More information is provided in Appendix M.1.

Table 2: **Correlation between Wikipedia article features and category win rate.** Category win rate shows moderate correlation with popularity (page views) but is independent of article length, statement or link count.

Feature	Coeff.	Strength
Page Views	$r = +0.478$	Moderate
External Links	$r = +0.067$	Negligible
Length (Bytes)	$r = -0.025$	Negligible
Statement Count	$r = +0.058$	Negligible

Robustness To verify the robustness of evaluation across different Wikipedia categories, we further analyzed the performance variation of DRAs across categories. To control for differences caused by inherent category difficulty, we used the deviation of each DRA system's score from the category mean as its relative score. We hypothesized that there is no significant difference in the relative performance of DRAs evaluated across different categories and conducted an ANOVA test. The results indicate that, with the exception of Deep Researcher, there are no significant differences in the relative performance of the systems evaluated across categories, with $p > 0.05$, demonstrating that our evaluation possesses cross-category robustness. Deep Researcher consistently underperformed across varying task difficulties; thus, we attribute the variance in its evaluation across categories to intrinsic performance limitations. Further details are provided in Appendix M.2.

5.2.2 Different Models for Judgement

Table 3: **Pairwise Agreement Rate (PAR) of different Judge LLMs with human annotations.** PAR is computed against a stratified human-reference set of 30 articles, with 2 per category across all 15 domains, corresponding to $\sim$1,170 criterion-level judgments. Cost represents the average cost per article evaluation. Qwen3-80B-A3B is deployed locally, incurring no API costs.

Judge Models	PAR (%)	Cost ($)
Gemini-2.5-pro	**82.13**	0.132
GPT-5	80.67	0.128
Qwen3-235B-A22B	78.72	0.071
DeepSeek-V3.1	76.41	**0.024**
GPT-5-mini	73.59	0.026
Qwen3-80B-A3B	72.31	–

To verify the performance differences among various models serving as Judge-LLMs, we sampled 30 Wikipedia-DRA article pairs using stratified sampling with 2 articles per category across all

15 domains and manually annotated the win rate for each criterion, yielding approximately 1,170 criterion-level annotations. This allowed us to observe the consistency between different judge models and human judgments. We report the Pairwise Agreement Rate (PAR) for criterion-level evaluation across different models, with results shown in Table 3. Among the proprietary models, Gemini-2.5-pro demonstrated superior consistency. In open-source solutions, Qwen3-235B-A22B also shows strong consistency, maintaining competitive performance even compared to proprietary models.

To rule out self-preference bias from using Gemini-2.5-pro as the judge, we conducted a full cross-evaluation with GPT-5. The resulting system rankings achieve a Spearman ρ of 0.876, a Kendall τ of 0.750, and a PAR of 87.5%, confirming that our evaluation reflects consistent system rankings rather than model-family preference. Full results are provided in Appendix G.

5.2.3 Wikipedia Leakage

Because Wikipedia and its mirrors are publicly accessible, agents may access target pages despite the task instruction. Although our prompt prohibits access to Wikipedia, some systems do not fully follow the instruction. For Cov. Wiki, we therefore exclude statements citing the target page or its mirrors before retrieval. We report three leakage signals: **Citation Leakage**, **8-gram Overlap**, and **Near-copy rate** among fact-matched sentence pairs. The representative systems in Table 4 include the low-leakage LangChain (GPT-5) baseline, systems with high citation leakage, and systems whose near-copy rate is high despite limited citation leakage. Perplexity and Gemini-3-pro Deep Research have high citation and near-copy rates, while Deep Researcher, Doubao, and LangChain (Kimi-K2.5) illustrate the difference between the two signals.

Table 4: **Representative Wikipedia leakage results.** Cite Leak, 8-gram, and Near-copy are percentages. Complete system results and measurement details are provided in Appendix N.

System	Cite Leak	8-gram	Near-copy
LangChain (GPT-5)	0.09	0.21	0.18
Gemini-2.5-pro Deep Research	17.32	0.23	0.39
Gemini-3-pro Deep Research	28.85	1.09	2.83
Perplexity Deep Research	40.45	2.47	7.12
Deep Researcher	–	5.25	7.87
Doubao Deep Research	1.31	2.20	5.58
LangChain (Kimi-K2.5)	0.39	2.04	4.31

LangChain (GPT-5) provides a low-leakage convergence baseline, with 0.09% Citation Leakage and 0.18% Near-copy. Perplexity and Gemini-3-pro Deep Research show high values on both signals, whereas Deep Researcher, Doubao, and LangChain (Kimi-K2.5) show elevated Near-copy despite limited Citation Leakage, indicating that the two measurements capture complementary leakage patterns.

6 Conclusion

In this study, we introduce **Wiki Live Challenge (WLC)**, a live benchmark that challenges the ability of Deep Research Agents to write Wikipedia-style articles, using Wikipedia Good Articles as human-expert references. WLC comprises 100 recent Good Articles spanning 15 categories, accompanied by a comprehensive evaluation framework, Wiki Eval. Wiki Eval combines a fine-grained writing assessment, constructing 39 criteria grounded in Wikipedia Good Article criteria, with a factual evaluation that measures both the coverage of Wikipedia facts and verifiability via cited references. Extensive experiments on a diverse set of deep research agents reveal a substantial gap between current agents and human-authored Wikipedia articles, particularly in factual information gathering and coverage. WLC thus provides a highly scalable benchmark for deep research agents from the perspective of authoring Wikipedia-style articles, covering writing quality, factual coverage, and citation grounding, enabling continuous and rigorous evaluation.

AI USE STATEMENT

In this work, we used generative AI tools primarily for translation and manuscript editing, with limited assistance in adapting existing code, processing data, and checking results within a human-designed research workflow. We take responsibility for the final content of this work, including AI-assisted text and artifacts.

REFERENCES

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

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, 2025. URL https://arxiv.org/abs/2506.11763.

Tianyu Fan, Xinyao Niu, Yuxiang Zheng, Fengji Zhang, Chengen Huang, Bei Chen, Junyang Lin, and Chao Huang. Understanding deepresearch via reports, 2025. URL https://arxiv.org/abs/2510.07861.

Google. Gemini deep research. https://gemini.google/overview/deep-research/, 2025.

Janghoon Han, Heegyu Kim, Changho Lee, Dahm Lee, Min Hyung Park, Hosung Song, Stanley Jungkyu Choi, Moontae Lee, and Honglak Lee. Deer: A comprehensive and reliable benchmark for deep-research expert reports, 2025. URL https://arxiv.org/abs/2512.17776.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating super-human reasoning for web agent, 2025a. URL https://arxiv.org/abs/2507.02592.

Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. Reportbench: Evaluating deep research agents via academic survey tasks, 2025b. URL https://arxiv.org/abs/2508.15804.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2025. URL https://arxiv.org/abs/2308.03688.

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*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

OpenAI. Deep research system card. https://openai.com/index/deep-research-system-card/, 2025.

Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. Deepscholar-bench: A live benchmark and automated evaluation for generative research synthesis, 2025. URL https://arxiv.org/abs/2508.20033.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, Rui Min, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents, 2025. URL https://arxiv.org/abs/2509.13309.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.347. URL https://aclanthology.org/2024.naacl-long.347/.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents, 2025. URL https://arxiv.org/abs/2511.07685.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025a.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025b.

Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouthi, Caiming Xiong, and Shafiq Joty. Liveresearchbench: A live benchmark for user-centric deep research in the wild, 2025. URL https://arxiv.org/abs/2510.14240.

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, 2025. URL https://arxiv.org/abs/2504.12516.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webdancer: Towards autonomous information seeking agency, 2025. URL https://arxiv.org/abs/2505.22648.

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry, 2025a. URL https://arxiv.org/abs/2507.16280.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. Citeeval: Principle-driven citation evaluation for source attribution, 2025b. URL https://arxiv.org/abs/2506.01829.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yuhao Zhan, Tianyu Fan, Linxuan Huang, Zirui Guo, and Chao Huang. Why your deep research agent fails? on hallucination evaluation in full research trajectory, 2026. URL https://arxiv.org/abs/2601.22984.

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. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

A Detailed Evaluation Settings

A.1 Implementation Details

Judge LLM for Wiki Writing Based on the agreement with human evaluation results, we selected Gemini-2.5-pro, which demonstrated the best performance, as the Judge-LLM for assessing the win rate. During evaluation, we input the Wikipedia original article, the generated article, and the criteria belonging to the same category as a single prompt into the model for batch evaluation. The evaluation prompt is shown in Figure 5.

Judge LLM for Wiki Fact We employ Gemini-2.5-flash as both the fact extraction model and the fact-checking model. The prompts for fact extraction and fact checking are presented in Figure 6 and Figure 7, respectively.

For Cov. Wiki, we use embedding-based candidate retrieval: cosine similarity ranks generated statements using OpenAI `text-embedding-3-small` against each Wikipedia fact, and the top-10 candidates are passed to the Gemini-2.5-Flash verifier.

A.2 Details of Evaluated DRAs

A.2.1 Data Collection Process

DRAs with Web Services Data for Gemini-2.5-pro Deep Research, OpenAI o3 Deep Research, Perplexity Deep Research, Grok Deep Search, Qwen-3-max Deep Research, and Doubao Deep Research were collected via their respective web services. During collection, some DRA systems required secondary user interaction to confirm the research direction; in these cases, all human annotators are instructed to use a unified prompt instruction: *All content requires you to research and follow the criteria I provide. Always exclude the search or reading of Wikipedia pages.*

Locally Deployed DRAs Data for LangChain Open Deep Research, Deep Researcher, and Tongyi Deep Research were collected in a local environment. For LangChain Open Deep Research, we followed its open-source framework settings, using Tavily[6] as the search engine, and ran it as a shared scaffold with all tools, prompts, and hyperparameters held identical across all backbones (GPT-4.1, GPT-5, Gemini-2.5-pro, Gemini-3-pro, Qwen-3-max, DeepSeek-V3.2, and Kimi-K2.5) for article generation. For Deep Researcher, we deployed its model on a single H20 GPU and used the crawling tools provided in its open-source repository for article generation. For Tongyi Deep Research, we deployed its model on a single H20 GPU, using Serper API[7] as the search engine and Jina[8] as the web crawling tool.

DRAs via API Services The recently released Gemini Deep Research Agent by Google [9], powered by Gemini-3-Pro, is capable of navigating complex information environments using web search to generate detailed reports with citations. We generated articles by invoking the Gemini Deep Research Agent API service.

B Wiki Fact Validation

B.1 Fact Extraction Validation

Wikipedia Good Articles follow a citation-backed structure in which the text between adjacent citation markers provides a model-free reference span for the cited content. We use this structure to validate the fact-extraction stage of Wiki Fact and evaluate coverage over 3,333 citation-backed factual spans.

The extracted facts therefore cover nearly all citation-backed reference content, with the remaining partial cases concentrated in long spans that combine multiple facts.

[6] `https://www.tavily.com/`
[7] `https://serper.dev/`
[8] `https://jina.ai/`
[9] `https://ai.google.dev/gemini-api/docs/deep-research`

Table 5: **Validation of Wikipedia fact extraction.** Coverage is judged at the citation-backed span level.

Measure	Value
Citation-backed factual spans	3,333
Fully covered by extracted facts	91.72%
At least partially covered	99.0%

B.2 Fact Verifier Cross-validation

We further validate the fact-verification stage using Gemini-2.5-Flash as the verifier and two independent judge models, Claude Opus 4.6 and GPT-5. We sample 600 stratified verifier verdicts and present the same candidate facts, context, and decision policy to all judges. The two Wiki Fact checks use the same verification prompt; at evaluation time, the corresponding verification target and evidence are inserted into the target and evidence fields. We use the 2-of-3 majority as a cross-judge reference and separately ask three human annotators to verify 100 sampled verdicts by majority vote. The first two rows of Table 6 report the majority-vote consistency ratios for negative and positive verifier verdicts, respectively; the last row reports the proportion of human-majority decisions that agree with Gemini.

Table 6: **Cross-validation of the Wiki Fact verifier.** The panel contains the verifier and two independent judges.

Measure	Value
Verifier negative verdict	96.9%
Verifier positive verdict	82.8%
Human agreement	92.0%

C Details of Wiki Eval

Our criteria selection is grounded in the Wikipedia Good Article criteria, as shown in Figure 8. We selected the first four dimensions (excluding stability and illustrations), focusing on the textual content. For the dimensions of *Well-written*, *Broad in its coverage*, and *Neutral*, we delved into the specific guidelines for each, strictly adhering to Wikipedia's guidelines on *lead sections*[10], *words to watch*[11], *verifiability*[12], *summary style*[13], and *neutral point of view*[14], collecting a total of 39 fine-grained criteria to construct the **Wiki Writing** evaluation, which are listed in Table 20, Table 21, and Table 22. For the *Verifiable* dimension, we constructed the **Wiki Fact**, an evaluation metric that encompasses both the factual accuracy of the generated article relative to Wikipedia and its consistency with cited references.

D New period Results

D.1 Implementation Details

Following the collection procedure in Section 3.1, we construct a new evaluation set of 100 Good Articles across 14 domains from Wikipedia articles created between December 2025 and August 2026.

We use LangChain Open Deep Research with GPT-4.1, GPT-5, Gemini-2.5-Pro, Gemini-3-Pro, Qwen-3-Max, DeepSeek-V3.2, DeepSeek-V4-Pro, and DeepSeek-V4-Flash. The backbones share the same tools, prompts, and research configuration. Wiki Writing uses Gemini-2.5-Pro, while

[10] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section
[11] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch
[12] https://en.wikipedia.org/wiki/Wikipedia:Verifiability
[13] https://en.wikipedia.org/wiki/Wikipedia:Summary_style
[14] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

fact extraction and verification use Gemini-2.5-Flash. Generated reports are evaluated against the Wikipedia reference snapshots collected for this period, following the evaluation procedure in Appendix A.1.

D.2 Evaluation Results

Table 7 presents the results for the new collection period. DeepSeek-V4 achieves leading performance, with V4-Pro obtaining the highest Wiki Writing score and V4-Flash the highest Wiki Fact coverage. Despite improved writing quality, the models still show a substantial gap in factual coverage relative to Wikipedia.

Table 7: **WLC results in the new collection period (December 2025–August 2026).** All scores are percentages. Wiki Writing reports win rates over 39 GA-derived criteria. Cov. Wiki measures reference-fact coverage, and Ref. Acc. measures citation support. Bold and underlined values indicate the best and second-best scores.

Backbone	Wiki Writing				Wiki Fact	
	Overall	Well-written	Neutral	Broad	Cov. Wiki	Ref. Acc.
GPT-4.1	23.36	20.05	20.80	35.25	16.89	55.74
GPT-5	<u>57.77</u>	<u>53.43</u>	<u>64.90</u>	<u>60.25</u>	18.05	54.25
Gemini-2.5-pro	50.31	50.00	47.10	55.13	22.22	**70.00**
Gemini-3-pro	47.62	49.71	42.00	49.13	23.51	66.64
Qwen-3-max	45.64	44.43	43.80	51.13	21.43	54.88
DeepSeek-V3.2	45.72	41.14	46.80	56.38	24.37	57.35
DeepSeek-V4-Pro	**64.36**	**62.14**	**72.10**	**60.50**	<u>31.12</u>	<u>67.95</u>
DeepSeek-V4-Flash	49.85	44.38	55.90	56.63	**32.04**	62.16

E Graded Scoring

We evaluate a sampled set of criteria on 20 stratified articles for 16 systems using Gemini-2.5-Pro. For each criterion, the judge independently assigns a 1–5 score to the generated article and the Wikipedia reference. The system-level graded score is the mean difference between these two scores. Binary win rates are computed on the same article–criterion subset. Table 8 reports the results.

Table 8: **Binary and graded scoring on the sampled articles.** Binary win rates and graded score differences use the same 20-article sample and criterion set.

System	Binary win (%)	Graded mean (gen − gt)
Gemini-3-pro Deep Research	57.51	$+0.413$
LangChain (GPT-5)	44.33	-0.261
Gemini-2.5-pro Deep Research	35.48	-0.562
LangChain (Gemini-3-pro)	34.87	-0.049
LangChain (Qwen-3-max)	31.90	-0.328
LangChain (Kimi-K2.5)	30.26	-0.326
Perplexity Deep Research	27.11	-0.688
OpenAI o3 Deep Research	26.95	-0.633
LangChain (Gemini-2.5-pro)	25.30	-0.674
Grok Deep Search	21.56	-0.742
LangChain (DeepSeek-V3.2)	15.95	-1.109
Qwen-3-max Deep Research	15.79	-1.777
Doubao Deep Research	15.46	-0.870
Tongyi Deep Research	13.79	-1.223
LangChain (GPT-4.1)	9.40	-1.722
Deep Researcher	2.56	-2.544

System rankings from graded score differences are highly correlated with binary win rates, with Spearman $\rho = 0.941$, $p < 10^{-7}$. Equal scores account for 14.6% of graded judgments.

F FACTS ACROSS RESEARCH STAGES

We analyze retrieved evidence, compressed notes, and final reports from six Open Deep Research backbones on a stratified sample. Table 9 reports the fraction of extracted reference facts covered at each stage, aggregated across the sampled articles. GPT-5 and Kimi-K2.5 retrieve a substantial fraction of reference facts, while their notes and final reports cover fewer of them. Other backbones already have lower coverage at the retrieved-evidence stage.

Table 9: **Reference-fact coverage across research stages.** Evidence, Notes, and Report scores are percentages over the listed reference facts.

Backbone	Evidence	Notes	Report
GPT-5	72.34	36.79	16.49
Kimi-K2.5	79.89	53.87	34.77
DeepSeek-V3.2	32.04	25.46	21.64
Qwen-3-max	28.71	22.54	20.67
Gemini-2.5-pro	19.64	16.03	12.81
Gemini-3-pro	30.09	20.75	16.79

G CROSS-EVALUATION

To rule out self-preference bias from using Gemini-2.5-pro as the primary judge, we conducted a full re-evaluation with GPT-5 as the judge model, as shown in Table 10. We report three consistency metrics between the two sets of rankings: Spearman ρ measures monotonic rank correlation, Kendall τ measures pairwise concordance, and Pairwise Agreement Rate (PAR) measures the fraction of system pairs preserving their relative order between the two judges. The Overall system rankings achieve a Spearman ρ of 0.876, a Kendall τ of 0.750, and a PAR of 87.5%, with 105 of 120 system pairs preserving their relative order, confirming high cross-judge consistency.

Table 10: **Wiki Writing results with GPT-5 as the Judge LLM.** System rankings remain highly consistent with those obtained using Gemini-2.5-pro, with Spearman $\rho = 0.876$ and Kendall $\tau = 0.750$ on Overall scores.

System	Overall (%)	Well-written (%)	Neutral (%)	Broad (%)
Open-Source Agent Framework				
Deep Researcher	7.74	2.48	9.50	19.38
Tongyi Deep Research	12.15	5.43	7.60	35.50
Open-Source Harness with LLMs				
GPT-4.1	16.59	12.05	13.10	32.88
GPT-5	53.15	51.57	47.10	**64.88**
Gemini-2.5-pro	28.74	26.38	27.30	36.75
Gemini-3-pro	34.95	32.67	35.80	39.88
Qwen-3-max	26.59	24.33	25.60	33.75
DeepSeek-V3.2	16.88	13.65	18.27	23.60
Kimi-K2.5	30.01	29.98	29.51	30.71
Proprietary Agent Framework				
Doubao Deep Research	15.97	12.24	13.20	29.25
Qwen-3-max Deep Research	13.18	11.67	9.00	22.38
Perplexity Deep Research	23.62	21.33	26.00	26.62
Grok Deep Search	30.23	31.07	23.54	36.36
OpenAI o3 Deep Research	37.18	34.71	30.70	51.75
Gemini-2.5-pro Deep Research	32.15	30.57	30.00	39.00
Gemini-3-pro Deep Research	**61.10**	**57.86**	**69.50**	59.13

H Position-Swap Robustness

To verify that our binary evaluation design extracts stable comparative signals rather than random noise, we conducted a position-swap experiment across all systems. For each criterion-level judgment, we swapped the presentation order of the Wikipedia article and the generated article, then re-evaluated with the same Judge LLM.

Criteria-level consistency reaches 82.3%, indicating that the judge produces the same winner regardless of input order for the vast majority of comparisons. At the system level, overall win rates before and after the swap are highly linearly correlated with a Pearson r of 0.9611, and the relative rankings remain nearly identical with a Spearman ρ of 0.9455. Most systems maintain their exact original rank. These results confirm that our forced-choice design captures genuine quality differences.

I Length–Score Correlation Analysis

To rule out a potential "longer-is-better" artifact in our evaluation, we analyse the relationship between generated-article length measured in words and each Wiki Eval score across all evaluated systems and 100 articles. We report both the Pearson r and Spearman ρ between length and each score, as shown in Table 11.

As shown in Table 11, the four Wiki Writing dimensions all exhibit negligible-to-weak correlation with length, with $|r| \leq 0.151$, so writing-quality judgments are not driven by output length. Cov. Wiki shows a moderate positive correlation, with $r = 0.310$ and $\rho = 0.409$, which is the expected and benign behaviour for a recall-oriented metric: a more thorough research artifact that covers more expert-verified facts is naturally longer, and that is precisely what the metric is intended to reflect. Ref. Acc. shows only a weak correlation, indicating that citation supportability is largely driven by citation quality rather than length.

Table 11: **Correlation between generated-article length and Wiki Eval scores across all evaluated systems $\times$ 100 articles.** All four Wiki Writing dimensions show negligible-to-weak correlation with length, indicating that the 39 criterion-based judgments are not driven by output length. Cov. Wiki shows a moderate positive correlation, which is consistent with its recall-oriented design.

Score	Pearson r	Spearman ρ	Strength
Overall	$+0.077$	$+0.198$	Negligible
Well-written	$+0.023$	$+0.119$	Negligible
Neutral	$+0.095$	$+0.212$	Negligible
Broad	$+0.151$	$+0.243$	Weak
Cov. Wiki	$+0.310$	$+0.409$	Moderate
Ref. Acc.	$+0.137$	$+0.161$	Weak

J Retrieval Depth Sensitivity for Cov. Wiki

To verify that the top-10 retrieval pipeline used in Cov. Wiki is not a meaningful bottleneck, we conduct a full sensitivity sweep over $k \in \{5, 10, 20, \text{full-article}\}$ on a 20-article random sample averaged across systems. In the *full-article* setting the entire generated article is used as the candidate pool, providing an oracle upper bound on what the metric can achieve. Each Wikipedia fact still triggers a single fact-checking call (one fact against k candidate statements), so the per-call token cost grows linearly in k.

Table 12: **Top-k retrieval sensitivity for Cov. Wiki on a 20-article sample, averaged across systems.** "Top-10" is the default used throughout the paper; "Full" uses the entire generated article as the candidate pool, serving as an oracle upper bound. The default Top-10 setting is within 1.34 points of this upper bound, and system rankings are preserved across all four settings.

Retrieval Setting	Top-5	Top-10	Top-20	Full
Avg. Cov. Wiki (%)	17.91	18.84	20.01	20.18

As shown in Table 12, average Cov. Wiki rises only mildly across the four settings, with the default Top-10 within 1.34 points of the full-article upper bound, and the relative ordering of systems is preserved across all settings. This indicates that the retrieval step is not a meaningful bottleneck for Cov. Wiki: when the retriever misses a fact, it is overwhelmingly because the fact is genuinely absent from the generated article rather than because the retriever fails. Top-10 is therefore a principled cost–quality default for a benchmark that is intended to be re-runnable as the GA pool grows over time.

K Analysis of Wiki Precision

As a complementary signal to the recall-oriented Cov. Wiki, we additionally compute a precision-side quantity, denoted **Prec. Wiki**, by swapping the roles of facts and statements in Eq. 2: for each statement s_j in the generated article we retrieve the top-10 most relevant Wikipedia facts and check whether s_j is consistent with any of them, and average the resulting binary scores over all generated statements:

$$\text{Prec. Wiki} = \frac{1}{|S|} \sum_{s_j \in S} \text{Fact}(s_j, F). \tag{5}$$

Here, S denotes all statements extracted from the generated article, including statements without citation markers. Intuitively, Prec. Wiki reflects what fraction of an agent's generated statements are echoed by the expert-verified Wikipedia fact list, and can therefore serve as a partial supervision signal for filtering or up-weighting Wikipedia-supported statements during training.

Cov. Wiki is recall-oriented and does not penalize correct information that is absent from the Wikipedia fact list. To quantify this case, we identify generated statements outside Wikipedia coverage and verify a stratified sample against the sources cited by the corresponding articles.

Table 13: **Wiki-side precision signal across all systems**, alongside Cov. Wiki for context. Prec. Wiki denotes the fraction of generated statements that are echoed by the expert-verified Wikipedia fact list.

Models	Cov. Wiki (%)	Prec. Wiki (%)
Open-Source Agent Framework		
Deep Researcher	5.62	58.98
Tongyi Deep Research	22.73	54.19
Open-Source Harness with LLMs		
GPT-4.1	7.08	34.15
GPT-5	20.96	43.22
Gemini-2.5-pro	21.20	43.80
Gemini-3-pro	16.26	36.22
Qwen-3-max	25.56	38.10
DeepSeek-V3.2	14.13	38.35
Kimi-K2.5	31.71	37.97
Proprietary Agent Framework		
Doubao Deep Research	22.97	44.29
Qwen-3-max Deep Research	22.22	38.78
Perplexity Deep Research	29.21	55.49
Grok Deep Search	20.73	35.12
OpenAI o3 Deep Research	25.12	50.96
Gemini-2.5-pro Deep Research	30.76	37.58
Gemini-3-pro Deep Research	28.83	55.76

Table 14: **Verification of statements outside Wikipedia coverage.** The final three categories are percentages among statements with a resolved verification outcome.

Measure	Value
Generated statements outside Wikipedia coverage	48.3%
Supported by cited sources	53.5%
Conflicting with cited sources	13.8%
No evidence in cited sources	32.7%

More than half of the resolved statements outside Wikipedia coverage are supported by their cited sources. This complements Cov. Wiki with a source-grounded signal for information that is not present in the reference fact list.

We do not include Prec. Wiki in the main result table because, on its own, a high Prec. Wiki value is hard to interpret as a quality signal: an agent that produces very short outputs covering only a few well-known facts can score high on Prec. Wiki while leaving most of the topic uncovered. A representative example is Deep Researcher, which attains the highest Prec. Wiki at 58.98% yet the lowest Cov. Wiki at 5.62%; manual inspection confirms that its reports are typically very short and contain only a handful of high-frequency facts. We therefore view Prec. Wiki as a useful auxiliary observation when read together with Cov. Wiki and article length, particularly as a partial supervision signal for downstream training, rather than as a stand-alone evaluation metric for end-to-end DRA performance.

L Analysis of Fact Conflicts

We further analyse *conflicts*, where generated statements directly contradict either the established facts in Wikipedia Good Articles (**Wiki Conf.**) or the content of their own cited references (**Ref.**

Conf.). Such conflicts are particularly detrimental as they introduce explicit falsehoods rather than mere omissions.

Table 15: **Conflict rates across different systems.** Wiki Conf. is the fraction of statements conflicting with Wikipedia facts. Ref. Conf. is the fraction of statements conflicting with their cited references.

Models	Wiki Conf. (%)	Ref. Conf. (%)
Open-Source Agent Framework		
Deep Researcher	9.98	–
Tongyi Deep Research	9.89	–
Open-Source Harness with LLMs		
GPT-4.1	24.69	2.94
GPT-5	10.45	3.14
Gemini-2.5-pro	9.68	4.42
Gemini-3-pro	12.86	3.68
Qwen-3-max	10.01	3.81
DeepSeek-V3.2	13.58	3.58
Kimi-K2.5	11.40	3.36
Proprietary Agent Framework		
Doubao Deep Research	9.78	4.65
Qwen-3-max Deep Research	10.41	6.87
Perplexity Deep Research	7.74	3.04
Grok Deep Search	11.40	4.96
OpenAI o3 Deep Research	9.12	3.95
Gemini-2.5-pro Deep Research	7.79	4.58
Gemini-3-pro Deep Research	7.62	4.23

Table 15 illustrates distinct error patterns across systems, decoupling the analysis of verification against ground truth versus cited sources. A high citation conflict rate points to severe hallucination, where the model fabricates information not supported by its claimed references; notably, Qwen-3-max Deep Research exhibits a high citation conflict rate of 6.87%. Conversely, a high conflict rate with Wikipedia implies the inclusion of hallucinations or information from unreliable sources that contradicts established facts. For example, while LangChain (GPT-4.1) achieves the lowest citation conflict at 2.94%, it records the highest Wiki conflict at 24.69%, suggesting it may be incorporating incorrect information despite adhering to its own retrieval context. LangChain (GPT-5) demonstrates superior performance by maintaining low conflict rates across both dimensions.

M Detailed Analysis Across Categories

M.1 Difficulty Analysis

Table 16 presents the category difficulty ranking. We rank categories by their average Wiki Writing win rate; lower win rates indicate higher difficulty. We also list the average article length, statement count, external link count, and total page views for each category.

M.2 Robustness Analysis

To verify the robustness of our evaluation across different Wikipedia categories, we formulate the null hypothesis H_0 that there is no significant difference in the relative performance of a DRA system when evaluated across different categories. The relative performance is calculated as the deviation of the system's score from the category mean to control for inherent category difficulty.

We conducted an ANOVA test for each system to test this hypothesis. The detailed results, including p-values and effect sizes η^2, are presented in Table 17.

Table 16: **Difficulty ranking and statistical features of different Wikipedia categories.** Categories are ranked by average Wiki Writing win rate, with lower win rates indicating higher difficulty.

Rank	Category	Avg. Win Rate (%)	Avg. Bytes	Avg. Stmts	Avg. Links	Avg. Views
1	History	16.6	42,882	175.8	45.5	16,927
2	Mathematics	18.9	29,445	85.0	47.0	5,648
3	Language and literature	19.6	55,728	177.8	94.8	4,786
4	Art and architecture	22.0	49,681	145.7	102.9	10,051
5	Engineering and technology	22.2	57,869	169.1	94.0	6,721
6	Agriculture, food and drink	25.2	37,317	70.0	111.0	5,756
7	Media and drama	27.0	39,891	126.1	70.1	179,130
8	Music	28.4	36,845	111.9	84.9	62,728
9	Video games	29.0	32,491	89.7	56.3	7,907
10	Sports and recreation	30.2	41,862	151.3	73.1	18,342
11	Warfare	32.0	48,352	223.7	37.8	8,757
12	Social sciences and society	32.6	51,052	132.5	85.8	12,414
13	Geography and places	38.9	46,078	162.0	94.0	1,024
14	Natural sciences	41.8	50,660	186.6	90.1	47,984
15	Philosophy and religion	56.2	39,287	142.0	69.0	154,968

Table 17: **ANOVA test results for cross-category robustness.** A p-value above 0.05 indicates that a system's relative performance does not significantly depend on category, supporting cross-category robustness of the evaluation. A low effect size η^2 further confirms that category contributes little to the variance.

System	p-value	η^2
Open-Source Agent Framework		
Deep Researcher	1.11e-11	0.571
Tongyi Deep Research	0.6097	0.097
Open-Source Harness with LLMs		
GPT-4.1	0.5702	0.101
GPT-5	0.5345	0.105
Gemini-2.5-pro	0.8903	0.062
Gemini-3-pro	0.9580	0.048
Qwen-3-max	0.4347	0.117
DeepSeek-V3.2	0.6891	0.090
Kimi-K2.5	0.8827	0.079
Proprietary Agent Framework		
Doubao Deep Research	0.9781	0.040
Qwen-3-max Deep Research	0.5770	0.101
Perplexity Deep Research	0.1615	0.161
Grok Deep Search	0.9065	0.059
OpenAI o3 Deep Research	0.8669	0.066
Gemini-2.5-pro Deep Research	0.8831	0.063
Gemini-3-pro Deep Research	0.3200	0.132

Criterion-level pattern. Across the 39 GA-derived criteria, current DRAs perform worst on neutral wording and editorial restraint, including puffery, euphemisms, and the distinction between facts and opinions. Structural organization is comparatively easier, suggesting that systems more often learn the form of an encyclopedic article than its editorial discipline.

N WIKIPEDIA LEAKAGE ANALYSIS

To measure direct and surface-form leakage from the target Wikipedia articles, we report citation leakage, 8-gram overlap, and the near-copy rate among fact-matched sentence pairs. Citation Leakage is the proportion of statements citing the target Wikipedia page or its mirrors. 8-gram Overlap is the fraction of generated-text 8-grams that appear in the Wikipedia source. Near-copy is the fraction of fact-matched sentence pairs with normalized sequence similarity above 0.8. The complete results are shown in Table 18.

Table 18: **Wikipedia Leakage Rates across different systems.** Cite Leak measures the proportion of statements citing the target Wikipedia page or its mirrors. 8-gram measures the fraction of 8-grams in the generated text that appear in the Wikipedia source. Near-copy measures the fraction of fact-matched sentence pairs with normalized sequence similarity above 0.8.

Models	**Cite Leak (%)**	**8-gram (%)**	**Near-copy (%)**
Open-Source Agent Framework			
Deep Researcher	–	5.25	7.87
Tongyi Deep Research	–	0.74	1.28
Open-Source Harness with LLMs			
GPT-4.1	1.13	0.08	0.17
GPT-5	0.09	0.21	0.18
Gemini-2.5-pro	1.70	0.83	1.20
Gemini-3-pro	0.64	0.56	1.25
Qwen-3-max	1.53	1.07	0.80
DeepSeek-V3.2	4.86	0.73	1.56
Kimi-K2.5	0.39	2.04	4.31
Proprietary Agent Framework			
Doubao Deep Research	1.31	2.20	5.58
Qwen-3-max Deep Research	6.88	0.13	0.14
Perplexity Deep Research	40.45	2.47	7.12
Grok Deep Search	0.82	0.41	0.53
OpenAI o3 Deep Research	1.85	0.49	1.07
Gemini-2.5-pro Deep Research	17.32	0.23	0.39
Gemini-3-pro Deep Research	28.85	1.09	2.83

O DATA COLLECTION COSTS

We report the costs incurred in collecting 100 articles generated by each DRA system. This encompasses both the compensation for human annotators and the operational costs of the DRA systems. Human annotators were paid $1 per article collection for their time. The specific generation costs for each DRA system are detailed in Table 19.

P HUMAN JUDGE-AGREEMENT VALIDATION

To validate agreement between the Wiki Writing judge and human assessments, we recruited five PhD-level annotators to evaluate a stratified sample of 30 Wikipedia–DRA article pairs based on each criterion. The sample includes 2 articles per category across all 15 domains, yielding $\sim$1,170 criterion-level annotations in total. All article data underwent preprocessing to rigorously remove references and inline citation tags, and was presented in Markdown format. The order of articles was randomized to ensure annotators evaluated solely based on writing quality. We strictly instructed annotators to thoroughly read the articles under comparison. For each criterion, in addition to

Table 19: **Cost of collecting 100 articles for each DRA system.** Account/API Cost refers to fees for model access or subscription. Human Cost refers to compensation for annotators collecting data via web services ($1/article). "–" indicates negligible costs (e.g., free access or automated collection).

Models	Account/API Cost ($)	Human Cost ($)	Total ($)
Open-Source Agent Framework			
Deep Researcher	–	–	–
Tongyi Deep Research	–	–	–
Open-Source Harness with LLMs			
GPT-4.1	10.7	–	10.7
GPT-5	460.9	–	460.9
Gemini-2.5-pro	75.4	–	75.4
Gemini-3-pro	45.6	–	45.6
Qwen-3-max	13.5	–	13.5
DeepSeek-V3.2	11.2	–	11.2
Kimi-K2.5	52.4	–	52.4
Proprietary Agent Framework			
Doubao Deep Research	–	100	100
Qwen-3-max Deep Research	–	100	100
Perplexity Deep Research	40	100	140
Grok Deep Search	–	100	100
OpenAI o3 Deep Research	200	100	300
Gemini-2.5-pro Deep Research	100	100	200
Gemini-3-pro Deep Research	157.3	–	157.3

determining the winner, they were required to provide a rationale for their decision. Annotators were compensated at a rate of $10 per hour.

Q Prompt Details

This appendix presents the detailed prompts used in Wiki Eval, including the Wiki Writing comparison prompt, the Wiki Fact extraction and verification prompts, the Good Article criteria prompt, and the fine-grained writing criteria used for evaluation.

Table 20: Detailed fine-grained criteria for **Well-written** (Part 1).

Category 1: Well-written (Part 1)
Clear & Concise:
1. Articles should be written in encyclopedic style, which differs from the technically dense style found in scholarly writing aimed at specialists.
2. Articles should address the topic without twisting the truth or telling "lies-to-children", but should also minimize (unexplained) jargon and not take prior knowledge for granted.
3. Articles should be self-contained as much as possible, rather than relying on excessive links to explain unfamiliar concepts.
Lead section:
1. An article's content should begin with an introductory lead section – a concise summary of the article – which is never divided into sections. The remainder of the article is typically divided into sections.
2. The lead section gives the basics in a nutshell, introduces the article, and cultivates interest in reading on—though not by teasing the reader or hinting at what follows.
3. The lead section should be written in a clear, accessible style with a neutral point of view.
4. The lead should stand on its own as a concise overview of the article's topic. It should identify the topic, establish context, explain why the topic is notable, and summarize the most important points, including any prominent controversies. The notability of the article's subject is usually established in the first few sentences.
5. As in the body of the article itself, the emphasis given to material in the lead should roughly reflect its importance to the topic, according to reliable, published sources. Apart from basic facts, significant information should not appear in the lead if it is not covered in the remainder of the article.
6. A lead section should be carefully sourced as appropriate, although it is common for citations to appear in the body and not the lead.
7. Significant information should not appear in the lead, apart from basic facts, if it is not covered in the remainder of the article, although not everything in the lead must be repeated in the body of the text. Exceptions include specific facts such as quotations, examples, birth dates, taxonomic names and typological classifications, case numbers, and titles.

Table 21: Detailed fine-grained criteria for **Well-written** (Part 2).

Category 1: Well-written (Part 2)
Words to Watch:
1. Puffery: Words such as these are often used without attribution to promote the subject of an article, while neither imparting nor plainly summarizing verifiable information. They are known as "peacock terms" by Wikipedia contributors. Instead of making subjective proclamations about a subject's importance, use facts and attribution to demonstrate it. Words to Watch: legendary, best, great, greatest, acclaimed, iconic, visionary, outstanding, leading, celebrated, popular, award-winning, landmark, cutting-edge, innovative, revolutionary, extraordinary, brilliant, hit, famous, renowned, remarkable, prestigious, world-class, respected, notable, virtuoso, honorable, awesome, unique, pioneering, phenomenal, prominent.
2. Contentious labels: Value-laden labels – such as calling an organization a cult, an individual a racist, sexist, terrorist, or freedom fighter, or a sexual practice a perversion – may express contentious opinion and are best avoided unless widely used by reliable sources to describe the subject, in which case use in-text attribution. Avoid myth in its informal sense, and establish the scholarly context for any formal use of the term. Words to Watch: cult, racist, perverted, sexist, homophobic, transphobic, misogynistic, sect, fundamentalist, heretic, extremist, denialist, terrorist, freedom fighter, bigot, myth, neo-Nazi, -gate, pseudo-, controversial.
3. Unsupported attributions: These words may disguise a biased view. Claims about what people say, think, feel, or believe, and what has been shown, demonstrated, or proved should be clearly attributed. Words to Watch: some people say, many people remember, many scholars state, it is believed/regarded/considered, many are of the opinion, most feel, experts declare, it is often reported, it is widely thought, research has shown, science says, scientists claim, it is often said, officially, is widely regarded as, X has been described as Y.
4. Expressions of doubt: When these are used, ensure that the source of the accusation is clear. Words to Watch: supposed, apparent, purported, alleged, accused, so-called ... Also, scare-quoting: a Yale "report"; undue emphasis: "... a Baptist church".
5. Editorializing: These words should usually be avoided so as to maintain an impartial tone. Words to Watch: notably, it should be noted, arguably, interestingly, essentially, utterly, actually, only, clearly, absolutely, of course, without a doubt, indeed, happily, sadly, tragically, aptly, fortunately, unfortunately, untimely.
6. Synonyms for said: In some types of writing, repeated use of said is considered tedious, and writers are encouraged to employ synonyms. On Wikipedia, it is more important to avoid language that makes undue implications. Words to Watch: reveal, point out, clarify, expose, explain, find, note, observe, insist, speculate, surmise, claim, assert, admit, confess, deny, confirm ...
7. Euphemisms: Euphemisms should generally be avoided in favor of more neutral and precise terms. Words to Watch: passed away, gave her life, eternal rest, make love, an issue with, collateral damage, differently abled.
8. Clichés and idioms: Clichés and idioms should generally be avoided in favor of direct, literal expressions. Words to watch: lion's share, tip of the iceberg, white elephant, gild the lily, take the plunge, ace up the sleeve, bird in the hand, twist of fate, at the end of the day.
9. Relative time references: Prefer specific statements to general ones. Words to watch: recently, lately, currently, today, presently, to date, X years ago, formerly, in the past, traditionally, this/last/next (year/month/winter/spring/summer/fall/autumn), yesterday, tomorrow, in the future, now, to this day, soon, since.
10. Unspecified places or events: Prefer specific statements to general ones. Words to watch: this country, here, there, somewhere, sometimes, often, occasionally, somehow.
11. Survived by: Phrasing such as "Smith died in 1982, survived by her husband Jack and two sons" should be avoided; this information can be made more complete and spread out through the article. Words to watch: is/was survived by, [Name]'s survivors include.

Table 22: Detailed fine-grained criteria for **Broad in its coverage** and **Neutral**.

Category 2: Broad in its coverage
1. It addresses the main aspects of the topic.
2. The lead, ideally the introductory sentence or at least introductory paragraph, of an article, should make clear what the scope of the article is.
3. All material that is notable, referenced, and that a reader would be likely to agree matches the specified scope must be covered (at least in a summarized fashion).
4. What reliable sources say about material that is out of scope for the decided-upon subject is largely irrelevant to that article and can be removed.
5. It stays focused on the topic without going into unnecessary detail.
6. The lead contains a quick summary of the topic's most important points.
7. Each major subtopic is detailed in its own section of the article.
8. It contains no irrelevant (nor only loosely relevant) information.
Category 3: Neutral
1. It represents viewpoints fairly and without editorial bias, giving due weight to each.
2. Avoid stating opinions as facts. Usually, articles will contain information about the significant opinions that have been expressed about their subjects. However, these opinions should not be stated in Wikipedia's voice. Rather, they should be attributed in the text to particular sources, or where justified, described as widespread views, etc. For example, an article should not state that genocide is an evil action but may state that genocide has been described by John So-and-so as the epitome of human evil.
3. Avoid stating seriously contested assertions as facts. If different reliable sources make conflicting assertions about a matter, treat these assertions as opinions rather than facts, and do not present them as direct statements.
4. Avoid stating facts as opinions. Uncontested and uncontroversial factual assertions made by reliable sources should normally be directly stated in Wikipedia's voice, for example the sky is blue not [name of source] believes the sky is blue. Unless a topic specifically deals with a disagreement over otherwise uncontested information, there is no need for specific attribution for the assertion, although it is helpful to add a reference link to the source in support of verifiability. Further, the passage should not be worded in any way that makes it appear to be contested.
5. Prefer nonjudgmental language. A neutral point of view neither sympathizes with nor disparages its subject (or what reliable sources say about the subject), although this must sometimes be balanced against clarity. Present opinions and conflicting findings in a disinterested tone. Do not editorialize. When editorial bias towards one particular point of view can be detected the article needs to be fixed. The only bias that should be evident is the bias attributed to the source.
6. Indicate the relative prominence of opposing views. Ensure that the reporting of different views on a subject adequately reflects the relative levels of support for those views and that it does not give a false impression of parity, or give undue weight to a particular view. For example, to state that According to Simon Wiesenthal, the Holocaust was a program of extermination of the Jewish people in Germany, but David Irving disputes this analysis would be to give apparent parity between the supermajority view and a tiny minority view by assigning each to a single activist in the field.
7. An article should not give undue weight to minor aspects of its subject but should strive to treat each aspect with a weight proportional to its treatment in the body of reliable, published material on the subject.
8. Any inclusion of fringe or pseudoscientific views should not give them undue weight. The fringe or pseudoscientific view should be clearly described as such. An explanation of how experts in the relevant field have reacted to such views should be prominently included.
9. In the case of beliefs and practices, Wikipedia content should not only encompass what motivates individuals who hold these beliefs and practices but also account for how such beliefs and practices developed. Wikipedia articles on history and religion draw from religion's sacred texts as primary sources and modern archaeological, historical, and scientific works as secondary and tertiary sources.
10. Article titles that combine alternative names are discouraged. For example, names such as "Derry/Londonderry", "Aluminium/Aluminum", and "Flat Earth (Round Earth)" should not be used. Instead, alternative names should be given their due prominence within the article itself, and redirects created as appropriate.

System Prompt

```
You are a strict, meticulous, and objective Wikipedia article evaluation expert. You excel at using specific criteria to compare two Wikipedia-style articles on the same topic, and for each criterion you MUST decide a clear winner - either Article 1 wins or Article 2 wins.
```

User Prompt

```
Articles to Evaluate
<article_1>
{article1_text}
</article_1>
<article_2>
{article2_text}
</article_2>
Criterion Category and Criteria
We focus on: {category.name}
{category.description}
Under this category there are {k} criteria:
{criteria_list}
Compare which article better satisfies each criterion.
Output your response in strict JSON format with exactly {k} results (one per criterion):
{
"results": [
{ "criteria_index": 1, "reason": "Brief explanation for criteria 1", "winner": 1 },
{ "criteria_index": 2, "reason": "Brief explanation for criteria 2", "winner": 2 }
]
}
Note: winner value should be:
- 1 = Article 1 wins
- 2 = Article 2 wins
```

Figure 5: **Prompt used for Wiki Writing Evaluation.** The model is tasked with comparing two articles based on specific criteria and determining a winner.

Fact Extraction Prompt

You will be provided with an article. The body of the article contains text with citations to references.
Citations in the main text may appear in the following forms: ...[1], ...1, ...[(1)](https://...)
Your Task:
Please extract **all** informative statements (facts) from the main text. Each statement **must** be associated with its corresponding citation number(s). Return the result as a list of (fact, ref_idx) pairs.
Citation Assignment Rule (Critical):
For each statement, find the **nearest citation(s) that immediately follow it** in the text. This is the citation that covers this statement.
- If a statement is followed directly by one or more citations (e.g., [1] or [1][2][3]), those are the citations for that statement.
- If a statement has no citation immediately after it, look forward in the text to find the **next citation(s)** that appear. All statements between the previous citation and this next citation share this next citation as their source.
- In other words: citations at the end of a sentence or paragraph typically cover all the preceding uncited statements up to the last citation marker.
Extraction Rules:
1. **Completeness & Context (Crucial):** Every extracted fact must be a complete, self-contained statement. Do not extract fragmented phrases.
2. **Handling Citations:** Each statement must have a citation. Use the nearest following citation(s) as described above.
3. **Formatting:** Return a JSON list. Ensure quotation marks in each fact are preserved.
Output Format Example:
Given text: "Li Qiang constructed a socioeconomic status index...[15] It has been validated...[3][7][12]"
Output: [{ "fact": "...", "ref_idx": "15" }, { "fact": "...", "ref_idx": ["3", "7", "12"] }]
Here is the main text of the article:
{article_text}
Please begin the extraction now. Output only the JSON list directly, without any chitchat or explanations.

Figure 6: **Prompt used for Fact Extraction.** The model extracts atomic facts and their associated citation indices from the text.

System Prompt

```
You are a professional data annotator. Given a sentence and a paragraph composed of several gold factual support sentences, your task is to verify whether the sentence is consistent with the facts. Matching any single support sentence is enough for consistency.
Decision policy:
consistent: At least one support sentence directly entails or clearly confirms the statement.
inconsistent: At least one support sentence clearly contradicts the statement.
not_support: No support sentence clearly confirms or contradicts the statement (insufficient evidence).
```

User Prompt

```
Below are the sentence to judge and the factual paragraph:
<judge_sentence>
{gen_sentence}
</judge_sentence>
<factual_paragraph>
{gt_sentence}
</factual_paragraph>
Output Format (strict JSON only):
{
"reason": "...",
"verdict": "consistent | inconsistent | not_support"
}
```

Figure 7: **Prompt used for Fact Checking.** The model verifies the consistency of a generated statement against a gold factual paragraph.

The six Good Article criteria

1. Well-written:
(a) the prose is clear, concise, and understandable to an appropriately broad audience; spelling and grammar are correct;
(b) it complies with the Manual of Style guidelines for lead sections, layout, words to watch, fiction, and list incorporation.
2. Verifiable with no original research:
(a) it contains a list of all references (sources of information), presented in accordance with the layout style guideline;
(b) reliable sources are cited inline. All content that could reasonably be challenged, except for plot summaries and that which summarizes cited content elsewhere in the article, must be cited no later than the end of the paragraph (or line if the content is not in prose);
(c) it contains no original research; and
(d) it contains no copyright violations or plagiarism.
3. Broad in its coverage:
(a) it addresses the main aspects of the topic; and
(b) it stays focused on the topic without going into unnecessary detail (see summary style).
4. Neutral: it represents viewpoints fairly and without editorial bias, giving due weight to each.
5. Stable: it does not change significantly from day to day because of an ongoing edit war or content dispute. (Not applicable for offline evaluation)
6. Illustrated, if possible, by media such as images, video, or audio: (Not applicable for text-only evaluation)
(a) media are tagged with their copyright statuses, and valid non-free use rationales are provided for non-free content; and
(b) media are relevant to the topic, and have suitable captions.

Figure 8: **The six Wikipedia Good Article criteria.** These criteria serve as the foundation for our Wiki Eval framework.