FORAGE: RELATED WORKS PREDICTION AS A BENCHMARK FOR AGENTIC RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic Retrieval Augmented Generation, where language models autonomously plan queries against a retriever, has several important applications, but lacks rigorous evaluation: single-turn benchmarks ignore iteration, multi-turn benchmarks reward fact lookup over document understanding, and web-based benchmarks are prone to contamination. We introduce FORAGE (**F**inding **O**ut **R**elated **A**rticles via **G**enerative **E**xploration), a live agentic retrieval benchmark built on a related works prediction (RWP) task: given a research paper with its related works section, citations, and bibliography redacted, the system must query a scientific corpus to rank candidate papers by likelihood of being cited as related work. We instantiate FORAGE separately on NeurIPS and ICLR 2025 papers and evaluate three open-source LLMs along with a leading closed-source model across sparse, dense, and late-interaction retrievers. The best configuration recovers nearly a third of each paper's cited related works, and an oracle probe that removes the retriever from the loop recovers substantially more, placing the bottleneck in query generation and retrieval. FORAGE provides substantial headroom for future work on agentic retrieval, and is designed to be refreshed each conference cycle to remain contamination-resistant as models evolve. We release the dataset here.

1 INTRODUCTION

Large language models (LLMs), despite strong reasoning capabilities, are limited by the knowledge encoded in their parameters and struggle on out-of-distribution information (Petroni et al., 2019; Kandpal et al., 2023; Mallen et al., 2023). Retrieval Augmented Generation (RAG) (Izacard et al., 2023; Lewis et al., 2020; Jiang et al., 2023) addresses this limitation by retrieving relevant evidence from an external corpus and conditioning generation on it. A standard RAG pipeline consists of a document retriever that returns chunks of text relevant to the query, optionally followed by post-processing such as re-ranking. *Agentic retrieval* (Asai et al., 2024; Yao et al., 2023) extends this paradigm by allowing the LLM to autonomously decide *when*, *what*, and *how often* to retrieve. This is especially valuable for tasks where a single retrieval is insufficient, or where the user query does not directly correspond to the evidence required.

Despite the rapid adoption of agentic retrieval, the benchmarks used to evaluate retrieval-augmented LLMs were largely designed for a single-shot setting and translate poorly to it. Single-turn benchmarks measure how well a fixed query is matched to a relevant document, but cannot evaluate how a model *formulates* that query or adapts when an initial query returns weak results. Multi-hop benchmarks do support iteration but reward aggregating short factoids from short passages, exercising neither holistic document understanding nor strategic query planning over long-form sources. Benchmarks built on web data, meanwhile, are increasingly susceptible to contamination as models train on ever-larger web crawls. The result is that no current benchmark jointly stresses agentic, multi-step retrieval, holistic understanding of long documents, and resistance to contamination over time, qualities that are stressed by the majority of applications of agentic retrieval, such as prior-art search, deep research tasks, due diligence and so on.

We introduce FORAGE, a live agentic retrieval benchmark that jointly stresses retrieval, query generation, and deep document understanding. FORAGE is built around a novel *related works prediction* (RWP) task: given an appropriately redacted research paper and a retrieval system over a large scientific corpus, the model must rank candidate papers by their likelihood of being cited as

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.

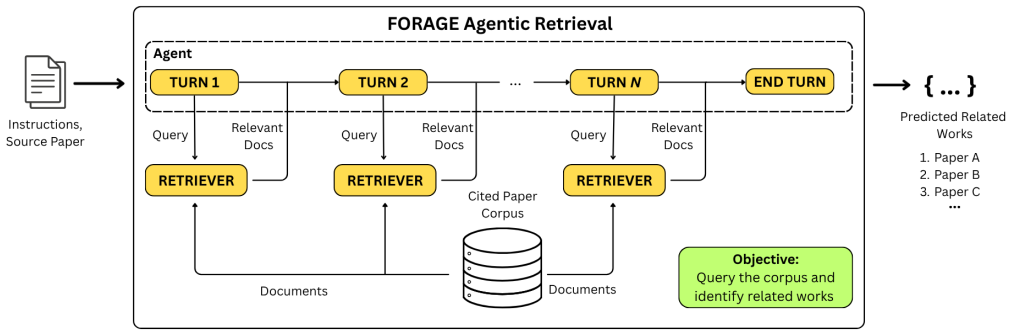


Figure 1: Agentic retrieval setup for FORAGE evaluation. Given a redacted source paper, the agent must query a corpus to ultimately identify the paper's related works. The agent generates queries for a retriever, which returns relevant documents from the corpus. The agent may query the retriever multiple times before finalizing its prediction.

related work. The task naturally rewards an agentic setting: a strong system must read the paper, formulate a retrieval strategy, and execute multiple queries to converge on a high-quality ranking. By sourcing source papers from the most recent edition of major venues, we obtain high-quality, out-of-distribution data that can be refreshed with every conference cycle. While tasks like prior-art search, deep research and due diligence fail as evaluation tasks due to the lack of live, openly available ground-truth citations at scale, RWP is the only task where expert (author) citations are routinely produced (and reviewed) at scale, are available openly, and refreshed every cycle.

This paper makes three main contributions: We propose RWP and instantiate it as the FORAGE benchmark separately on NeurIPS and ICLR 2025 papers with 1,024 and 772 papers respectively. We evaluate three open-source LLMs and a frontier LLM in combined with sparse, dense, and late-interaction retrievers, and report performance against both a no-retrieval baseline and an oracle linear-scan upper bound [1]. We analyze the results and find that different configurations show performance well below the upper-bound, and find potential performance hurdles, establishing substantial headroom and a concrete target for future work.

2 Related Works

Retrieval Augmented Generation. The factual accuracy of LLMs is fundamentally constrained by the knowledge captured during pretraining (Kandpal et al., 2023; Mallen et al., 2023), being especially restricted for specialized domain knowledge, rapidly evolving information, and long-tail entities (Mallen et al., 2023). RAG addresses this by augmenting the model's context with documents retrieved at inference time (Lewis et al., 2020), improving accuracy, reducing hallucination, and enabling generalization to out-of-distribution domains (Izacard et al., 2023; Shuster et al., 2021).

Agentic Retrieval. Agentic retrieval extends RAG by allowing the model to decide when and what to retrieve rather than relying on a single pre-generation retrieval step. Frameworks such as ReAct (Yao et al., 2023) and IRCoT (Trivedi et al., 2023) interleave chain-of-thought reasoning with tool use, enabling the model to formulate retrieval queries as part of an explicit reasoning trace. This iterative formulation is particularly suited to multi-hop reasoning tasks, where a single retrieval is insufficient (Trivedi et al., 2023; Jiang et al., 2023), and to tasks requiring strategic information gathering of which RWP is a natural example.

Retrieval Methods. Retrieval techniques have evolved from term-based methods such as TF-IDF (Sparck Jones, 1972) and BM25 (Robertson & Zaragoza, 2009) to dense neural retrievers (Karpukhin et al., 2020; Sachan et al., 2023). Learned sparse methods such as SPLADE (Formal et al., 2021) combine the efficiency of sparse representations with the generalization of learned models, while late-interaction models such as ColBERT (Khattab & Zaharia, 2020) retain token-level

[1] We release the evaluation code here.

Table 1: Comparison of FORAGE with representative retrieval and RAG benchmarks. "Deep Doc." indicates whether the task fundamentally requires understanding a long document end-to-end, as opposed to localized passage matching. "Refreshable" indicates whether the benchmark is designed to be regenerated to mitigate training-data contamination over time.

Benchmark	Agentic	Deep Doc.	Domain	Refreshable
Retrieval and RAG benchmarks				
BEIR (Thakur et al., 2021)	✗	✗	Heterogeneous	✗
MS MARCO (Bajaj et al., 2016)	✗	✗	Web	✗
NQ (Kwiatkowski et al., 2019)	✗	✗	Wikipedia	✗
BRIGHT (SU et al., 2025)	✗	✗	Heterogeneous	✗
MultiHop-RAG (Tang & Yang, 2024)	✓	✗	News	✗
MuSiQue (Trivedi et al., 2022)	Partial	✗	Wikipedia	✗
FRAMES (Krishna et al., 2025)	✓	✗	Web	✗
SciDocs (Cohan et al., 2020)	✗	Partial	Scientific	✗
RGB (Chen et al., 2024)	✗	✗	News/Wiki	✗
CRAG (Yang et al., 2024)	Partial	✗	Heterogeneous	✗
InfoDeepSeek (Xi et al., 2025)	✓	✗	Web	Partial
LitSearch (Ajith et al., 2024)	✗	✗	Scientific	✗
Deep research benchmarks				
BrowseComp (Wei et al., 2025)	✓	✗	Web	✗
DeepResearch Bench (Du et al., 2025)	✓	Partial	Web	✗
ReportBench (Li et al., 2025)	✓	✗	Scientific	✓
FORAGE (ours)	✓	✓	Scientific	✓

representations and compute relevance via fine-grained interactions at retrieval time. Graph-based retrievers (Edge et al., 2024) organize corpora into structured representations that capture relations beyond what flat embeddings express.

RAG and Retrieval Benchmarks. Several benchmarks evaluate retrieval and RAG systems, but none jointly capture the properties needed to evaluate agentic retrieval over long-form documents. Single-turn benchmarks such as BEIR (Thakur et al., 2021), MS MARCO (Bajaj et al., 2016), and Natural Questions (Kwiatkowski et al., 2019) measure how well a fixed query is matched to a relevant document but do not evaluate iterative query formulation. BRIGHT (SU et al., 2025) raises the difficulty with queries that require reasoning to identify relevant documents, but retains the single-query setting. Multi-hop benchmarks like MultiHop-RAG (Tang & Yang, 2024), MuSiQue (Trivedi et al., 2022), and FRAMES (Krishna et al., 2025) do support iteration, but are framed around aggregating short factoids from short passages rather than holistic comprehension of long documents. More recent agentic benchmarks such as InfoDeepSeek (Xi et al., 2025) and CRAG (Yang et al., 2024) extend evaluation to multi-step retrieval over the web, but draw on heavily-indexed sources that are vulnerable to contamination as models scale. RGB (Chen et al., 2024) similarly focuses on news and Wikipedia content.

Within the scientific domain, LitSearch (Ajith et al., 2024) evaluates retrieval with 597 literature-search questions about recent ML and NLP papers, but each query is a short, self-contained question. The task closest to RWP is citation recommendation, which recommends references given a draft paper (Bhagavatula et al., 2018) or a local citation context (Gu et al., 2022); these methods are non-agentic, and local variants rely on the surrounding text of each citation, which RWP deliberately redacts. Multi-XScience (Lu et al., 2020) instead generates a paper's related-work section given its abstract and the articles it references, i.e., assuming the retrieval step that FORAGE evaluates has already been solved.

Deep research benchmarks. A recent line of work evaluates agents that autonomously browse, retrieve, and synthesise long-form reports. BrowseComp (Wei et al., 2025) poses short questions with hard-to-find, but easy verifiable answers. DeepResearch Bench (Du et al., 2025) evaluates 100 expert-authored PhD-level tasks. ReportBench (Li et al., 2025) also uses a citation-prediction task, with a reverse-engineered prompt in place of the original paper.

3 FORAGE DESIGN

3.1 TASK DEFINITION

We define Related Works Prediction (RWP) as follows: Let P be a research paper from which the related works section, all in-text citations, and the bibliography have been redacted. Let C be a corpus of candidate papers and $\mathcal{R}$ a retrieval tool that, given a query q, returns a ranked list of passages from C. A system is given P and tool access to $\mathcal{R}$, and must output a ranked list of papers from C in decreasing order of likelihood of being cited as related work in P. The system may invoke $\mathcal{R}$ at most n times.

RWP is modeled as a ranking task, where the model must rank papers by their likelihood of being related. The set of related works for a paper is open-ended: many works may be topically related but could not be cited due to other restrictions. Treating the originally cited set as a ground-truth binary label would falsely penalize the system for surfacing genuinely related papers that the authors did not cite. Formulating it as ranking mitigates this to some extent, as the model can surface a larger number of works without the same penalization. In an ideal prediction the originally cited papers occupy the top of the ranking (in any order), with other plausibly related works following.

We deliberately target the related works section rather than performing local citation prediction across the paper. Citations in introductions, methods, and experiments are anchored to surrounding text and are largely predictable from local context, which means they can be solved without holistic comprehension. Related works, by contrast, depend on the contributions and framing of the paper as a whole, which lets us redact the entire section while preserving the signal needed to recover it. RWP therefore stresses document-level understanding and strategic query formulation in a way local citation prediction does not.

3.2 DATASET CONSTRUCTION

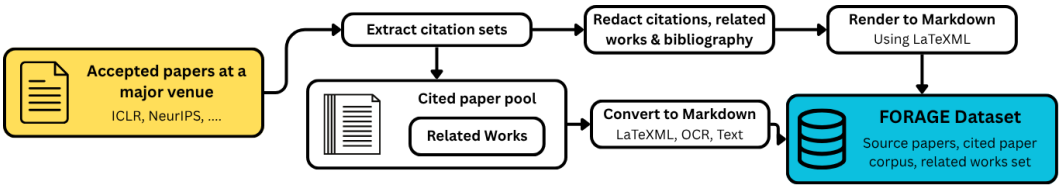


Figure 2: FORAGE dataset construction process. LaTeXsources for the source papers are obtained from arXiv and are used for all processing tasks. Citation sets are extracted from source papers before redacting them and rendering them to markdown using LaTeXML. The cited papers are converted to markdown using a LaTeXML (via the ar5iv dataset) when available, and falling back to OCR and text extraction.

FORAGE's source papers are drawn from the latest edition of major research conferences, ensuring high-quality and out-of-distribution data even for recent models. We instantiate FORAGE on NeurIPS and ICLR 2025 for this paper.

Source paper processing. We obtain LaTeX sources of accepted papers from arXiv. For each paper, we parse the sources to extract (i) A_i, the set of all papers cited anywhere in P, and (ii) $R_i \subseteq A_i$, the subset cited within the related works section. We mask in-text citations with `[citation]` placeholders and remove the bibliography and the related works section in their entirety. Papers without a clearly delineated related works section are excluded. Non-content material (e.g., conference checklists) is also removed. Source papers are then rendered to Markdown using LaTeXML.

Cited paper processing. All cited papers are converted to Markdown. We restrict to papers available on arXiv. Where LaTeX sources are available we use LaTeXML conversions from the ar5iv dataset (Ginev, 2024); otherwise we fall back to text extraction or OCR. The union $\bigcup_i A_i$ across source papers forms the retrieval corpus.

Table 2: Summary statistics for both FORAGE-NeurIPS-25 and FORAGE-ICLR-25 datasets.

	NeurIPS'25 / ICLR'25
Source papers	1,024 / 772
Corpus papers	25,050 / 18,650
Avg. cited / paper	45.5 / 45.5
Avg. rel. works / paper	24.2 / 22.1
Avg. source len. (tok)	19,611 / 20,034

Table 3: Distribution of related works by category (NeurIPS'25 / ICLR'25).

Category	%
Motivation	37.8 / 41.6
Competing	40.5 / 37.4
Backdrop	8.7 / 10.2
Component	4.8 / 4.6
Predecessor	3.6 / 2.6
Benchmark/Dataset	3.3 / 2.1
Survey	1.2 / 1.4

3.3 Dataset Statistics

Table 2 summarizes the resulting FORAGE datasets. NeurIPS and ICLR versions comprise 1,024 and 772 source papers respectively.

Coverage across fields and contribution types. Source papers span a broad cross-section of machine learning research (Table 4). The largest areas are NLP & LLMs ($\sim$20%), Multimodal ($\sim$13%), and Computer Vision ($\sim$12%), with meaningful representation across reinforcement learning, generative models, fairness and safety, and applied science. By contribution type, methodological papers dominate ($\sim$80%), followed by empirical studies ($\sim$10%), theory papers ($\sim$6%), and analysis papers ($\sim$2%). This distribution reflects the underlying composition of the venues rather than a deliberate sampling choice: source papers are drawn uniformly from accepted submissions at the venue, so FORAGE inherits the subfield skew of the venue.

Per-paper length and citation profile. Source papers for both datasets are long, averaging approximately 20,000 tokens, with most of the mass between roughly 12k and 28k tokens. This length is consequential for the task, as the agent must reason over the full document to construct effective retrieval queries, and short documents would not exercise the holistic-comprehension property that motivates RWP . Each source paper cites 45.5 papers on average, of which 24.2 (53%) appear in the related works section. The related-works subset is therefore non-trivially smaller than the full bibliography, which means the task is not simply citation prediction, and the system must distinguish related-work citations from citations elsewhere in the paper.

What kinds of related work do papers cite? Not all related-work citations serve the same rhetorical function, and this affects what a strong agentic retriever must surface. We classify each related-work citation into one of seven functional categories. This categorization was produced by presenting GPT-OSS-20B with the title, abstract and verbatim text of the related works section. A subset was manually reviewed to ensure the model identified the correct signals.

Table 3 reports the distribution. Two categories dominate: Motivation (works cited to establish the importance of the problem) and Competing (works the source paper directly contrasts itself against), which together account for nearly three-quarters of all citations. The remaining citations form a long tail across Backdrop, Component, Predecessor, Benchmark/Dataset, and Survey.

3.4 Evaluation Metrics

We use two metrics to measure performance:

Recall@k. Recall@10 measures the coverage of the originally cited set within the predicted ranking. It is the fraction of correctly predicted papers in the top-k ranking. Let $\hat{R}_i^{(h)}$ denote the top-h predictions for paper i and let $h = \min(k, |R_i|)$. Then

$$\text{Recall@}k = \frac{1}{|D|} \sum_{i \in D} \frac{|\hat{R}_i^{(h)} \cap R_i|}{h}.$$

Table 4: Distribution of papers by area and contribution type in FORAGE. Each cell reports NeurIPS-25 / ICLR-25 counts.

Area	Method	Theory	Empirical	Analysis	Total
Core	64 / 36	37 / 18	– / 2	3 / 2	104 / 58
Generative Models	73 / 65	5 / 7	2 / 1	2 / 1	82 / 74
NLP & LLMs	152 / 141	6 / 3	25 / 31	10 / 8	193 / 183
Reinforcement Learning	78 / 55	8 / 5	– / 2	1 / 2	87 / 64
Computer Vision	118 / 89	– / 2	10 / 3	3 / –	131 / 94
Graphs & Geometry	26 / 18	2 / 3	2 / –	– / –	30 / 21
Fairness & Safety	37 / 34	7 / 3	4 / 2	3 / 1	51 / 40
Multimodal	123 / 81	1 / 1	32 / 13	3 / 2	159 / 97
Systems & Efficiency	15 / 14	1 / –	– / –	– / –	16 / 14
Science Applications	82 / 57	1 / –	20 / 6	1 / 2	104 / 65
Other	52 / 48	3 / 4	12 / 10	– / –	67 / 62
Total	**820 / 638**	**71 / 46**	**107 / 70**	**26 / 18**	**1024 / 772**

Using a fixed top-k cut gives a more interpretable upper bound of 1 across papers with varying $|R_i|$.

Mean Reciprocal Rank (MRR). MRR is the reciprocal of the rank of the first correctly-predicted paper in the system's ranking. For a source paper i with related works R_i, let r_i be the rank of the highest ranked paper from R_i in the predicted ranking.

$$\text{MRR} = \frac{1}{|D|} \sum_{i \in D} \frac{1}{r_i}.$$

These metrics measure the two aspects of a good prediction: large number of related works are surfaced, and the related works are ranked highly. We report $k = 10$ throughout.

4 Experiments

Models. We evaluate three open-source language models: GPT-OSS-120B (OpenAI et al., 2025), Nemotron-3-Super-120B (referred to as Nemotron hereafter) (NVIDIA, 2025), and Gemma 4-31B (Gemma Team et al., 2026). We also evaluate GPT-5.6-Luna (OpenAI, 2026) to gauge task difficulty for frontier models.

Retrievers. We pair each LLM with three retrieval families: a sparse retriever (BM25 (Robertson & Zaragoza, 2009)); two dense retrievers (Qwen3-Embedding-4B with the Qwen3-Reranker-4B (Zhang et al., 2025), and EmbeddingGemma-300M (Schechter Vera et al., 2025)); and a late-interaction retriever (ColBERT (Khattab & Zaharia, 2020)). GPT-5.6-Luna is evaluated with EmbeddingGemma.

No-retrieval baseline. Because the corpus consists of real published papers, some cited works likely appear in each model's pretraining distribution. We therefore include a *no-retrieval* baseline in which the model produces a ranked list directly from parametric knowledge given only the redacted source paper. This baseline isolates the contribution of retrieval and provides a contamination floor.

Oracle linear scan. To estimate an upper bound on performance, we conduct the oracle linear scan experiment which simulates a perfect retriever. The model is presented with every paper in the corpus and then tasked with first classifying the paper as related or not, followed by a ranking of the predicted works. We select a subset of source papers, with the corpus being $\bigcup A_i$, chosen such that the candidate pool exceeds the maximum number of candidate papers the model would encounter after retrieval.

With the maximum 8 tool calls and 8 documents per call, the model sees at most 64 candidate passages drawn from the full corpus. Given an average of 24.2 cited related works per paper, this means the model must, in practice, distinguish a few true positives from up to roughly 40 close negatives. The oracle probe runs on a 16-paper subset with a candidate pool of $\sim$600 documents, aiming to replicate the setup.

Truncated source ablations. To assess whether the task stresses holistic document understanding, we perform experiments where only a truncated portion of the source paper is presented to the model. We evaluate two configurations, one with only the abstract and another with the abstract and introduction.

To showcase the live nature, we instantiate FORAGE on papers from ICLR-25 and evaluate the Gemma family on FORAGE-ICLR-25. The evaluation runs for Gemma on both ICLR and NeurIPS are run thrice to study variation across runs.

Implementation All experiments use LlamaIndex (Liu, 2022) for orchestration. Corpus documents are chunked prior to indexing for retrieval. Chunking is structure-aware. We begin by splitting the document into headings using the markdown heading tokens, followed by paragraph splitting. We use a chunk size of 1024 with chunk overlap 128. Each chunk is stored with metadata including the candidate paper's arXiv ID (for ranking), title, and header path (e.g., `# FORAGE > ## Abstract`), which is presented to the model with the content for structural context. Please refer to Appendix B for detailed parameters and prompts for the experiments.

4.1 Results & Analyses

Table 5 reports the results for the oracle scan, truncated ablations and the GPT-5.6-Luna evaluation. Table 6 lists the results across all (model, retrieval) configurations, with no-retrieval baselines.

Table 5: Results of the oracle probe, frontier model and truncated ablation runs on FORAGE-NeurIPS-25. Input is the portion of the redacted source paper shown to the model. The oracle scan replaces the retriever with the full candidate pool and is run on a 16-paper subset; the no-retrieval relies on the LLM knowledge alone.

Model	Input	Retrieval	Recall@10	MRR
Gemma-4-31B	Full paper	Oracle linear scan	**53.40**	**82.35**
GPT-5.6-Luna	Full paper	EmbGemma	40.60	79.60
Gemma-4-31B	Abstract only	EmbGemma	24.39	75.52
Gemma-4-31B	Abstract + Intro	EmbGemma	27.62	78.05
Gemma-4-31B	Full paper	None	7.00	29.44

Memorization is not driving performance. The no-retrieval baseline achieves Recall@10 $\sim 7-10$ and MRR $\sim$30 across all three models, which is far below any retrieval configuration. Retrieval improves Recall@10 by 3–5$\times$ over this baseline (e.g., Gemma-4-31B: 7.00 $\rightarrow$ 34.75). As source papers may optionally be present as a pre-print before acceptance at the venue, we also calculate the

Table 6: Evaluation results for language model and retrieval system combinations on FORAGE-NeurIPS-25, formatted as Recall@10 / MRR. All retrieval methods under-perform the oracle linear scan baseline.

LLM	No Retrieval	BM25	EmbGemma	ColBERT	Qwen3
GPT-OSS-120B	10.22/33.54	34.51/64.75	**35.53**/68.11	32.49/66.68	35.30/68.33
Nemotron	8.92/32.02	27.80/68.22	30.05/72.43	30.70/71.19	30.46/71.35
Gemma-4-31B	7.00/29.44	32.20/75.84	34.75/**78.15**	32.24/73.21	35.11/77.87

Table 7: Percentage of source papers hypothetically present in LLM training distribution, along with evaluation results with EmbGemma.

LLM	Overlap %age	Recall@10
Nemotron	19.43%	30.05
Gemma	6.05%	34.75
GPT	0.49%	35.53

Table 8: Evaluation on the two datasets, across 3 runs each on Gemma-4 + EmbGemma. The FORAGE benchmark is refreshable.

Dataset	Recall@10	MRR
NeurIPS-25	34.75 ± 1.0	78.15 ± 0.9
ICLR-25	31.5 ± 0.4	72.9 ± 0.4

fraction of papers that could hypothetically be present in the models' training distribution by comparing the model's cutoff date with the paper's date of first publication of arXiv (Table 7). Overlap does not track performance: Nemotron has the highest overlap (19.43%) yet the lowest R@10 with EmbGemma (30.05), while GPT-OSS, with only 0.49% overlap, reaches 35.53. This indicates that parametric knowledge alone is insufficient, and contamination is not driving performance.

Retrieval is the major bottleneck, not ranking. The oracle linear-scan probe is significantly higher than all retriever configurations, including the frontier model which falls behind by $\sim 13\%$ in R@10. With the same LLM, the oracle outperforms Gemma-4-31B + EmbGemma by 18.65 R@10 points (53.40 vs. 34.75): models rank relevant papers well once they are among the candidates, but fail to surface most of them. A stronger LLM closes only part of this gap, as GPT-5.6-Luna (40.60) improves on the best open-source configuration by 5.07 points, less than a third of the oracle gap. As the oracle scan replicates the ideal retrieval scenario, the performance delta with the oracle experiment shows that the major bottleneck does not lie in the final ranking step; instead, the query generation and retrieval performance forms the major hurdle to performance.

Retrieval helps substantially, but with diminishing returns across retrievers. All four retriever families hit around 30 R@10 and 65 MRR. Differences *across* retriever families are relatively modest, with dense retrieval generally edging out sparse and late-interaction methods. Gemma and GPT-OSS-120B paired with EmbeddingGemma achieve the strongest results.

The benchmark is updatable. We apply the FORAGE pipeline unchanged to papers sourced from ICLR'25 (see Table 8 for results). Gemma-4-31B + EmbGemma achieves 31.5 ± 0.4 R@10 and 72.9 ± 0.4 MRR on ICLR-25, comparable to 34.75 ± 1.0 and 78.15 ± 0.9 on NeurIPS-25. The low standard deviations (≤ 1.0) across three runs further show that results are stable across evaluations. This showcases the live nature of FORAGE, and that essential benchmark characters are preserved across instantiations.

RWP stresses holistic document understanding. The truncated ablations under-perform the retrieval setting by $\sim 10\%$, indicating the entire paper provides signals for accurate prediction; holistic understanding of the document is required for performance. With the abstract alone, R@10 drops from 34.75 to 24.39 (-10.36 points), and adding the introduction recovers only 3.23 points (27.62); both drops far exceed the ± 1.0 run-to-run deviation.

Blind human review. There is no exhaustive set of related works for any paper. To study the impacts of this limitation on model performance, we conduct a small focused blind review of the best configuration's top predictions, annotating 120 pairs across 20 different source papers. The review annotated 71.1% of false positives as "not citable as related work" (versus 100% of negatives and 68.3% agreement on true positives). This suggests that the ground-truth is largely complete, and false penalization is low.

Analysis of retrieval. We analyze the full set of documents retrieved by the retriever along with the distribution of retriever calls in one session (Table 9). We measure retrieval recall, the fraction of related works surfaced anywhere across a session's retrieved candidates. Low retrieval recall (22-24%) suggests that the retriever frequently fails to surface relevant documents in its top 8, which places an upper bound on performance, regardless of ranking. With 24.2 related works per paper on

Table 9: Retrieval failure analysis. "Recall" is the percentage of R_i surfaced anywhere in the union of retrieved candidates across a session; "Distinct retrieved / session" and "/ query" are averages. Top half varies the retriever (averaged over LLMs); bottom half varies the LLM.

		Distinct retrieved	
	Recall	/ session	/ query
Retrieval method			
BM25	22.8	26.5	4.6
EmbeddingGemma	22.8	23.4	3.9
ColBERT	23.2	35.5	6.3
Qwen3	24.5	25.4	4.4
Language model			
Gemma-4-31B	23.9	25.4	4.6
GPT-OSS-120B	23.4	27.0	4.5
Nemotron	21.0	29.1	4.1

average, this corresponds to only about 5–6 related works surfaced in an entire session. Retrieved results are also redundant: only 3.9–6.3 of the 8 chunks returned per query come from distinct papers, and ColBERT surfaces 52% more distinct papers per session than EmbeddingGemma (35.5 vs. 23.4) with nearly identical retrieval recall (23.2 vs. 22.8%). Additionally, we observe that Gemma-4-31B issues most of its queries in the first step, while both GPT-OSS and Nemotron query the retriever at roughly one request per step. This reflects a difference in strategy between the models, as Gemma therefore does not exploit the iterative nature of agentic retrieval. These diagnostics indicate query generation and retrieval performance as being the major bottlenecks.

5 Limitations and Future Work

FORAGE has several limitations. First, as discussed in Section 3.1, the originally cited related-works set cannot be treated as exhaustive: many works are plausibly related but were not cited. We mitigate this by framing the task as ranking, and the blind review quantifies the extent of the residual mismatch. Second, our pipeline restricts source and cited papers to those available on arXiv, which under-represents fields and communities that do not routinely post preprints (notably parts of theoretical statistics and some applied disciplines). We aim to address this in the future with the help of more accessible and accurate OCR models and agent-assisted literature searches.

FORAGE is designed as a live, refreshable benchmark. We plan to release FORAGE datasets for the next editions of NeurIPS and ICLR, and, with community support, to expand the benchmark to additional domains beyond machine learning.

6 Conclusion

We introduced FORAGE, a live agentic retrieval benchmark based on the related works prediction task. By sourcing evaluation data from the most recent edition of major conferences, FORAGE provides a continuously refreshable, out-of-distribution challenge that resists contamination, and stresses query generation and holistic document understanding. FORAGE evaluation results show that, across a variety of modern and frontier LLMs and retrieval systems, performance converges to a ceiling well below the oracle. Our analysis identifies query generation and retrieval performance as the major bottlenecks in current agentic retrieval systems that contribute to the performance gap, helping guide future progress in the field. We release FORAGE as a rigorous, refreshable testbed to track that progress.

AI USE STATEMENT

In this work, we used generative AI tools for (1) implementing methods and (2) supporting qualitative and thematic data analysis.

We have **not** used generative AI tools for (1) generating synthetic datasets, (2) helping develop theoretical models or conceptual frameworks, (3) designing or providing feedback on research methodology or experiments, (4) assisting with translation, (5) cleaning and reformatting the dataset or (6) interpreting results. Formulating mathematical claims, providing critical ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work.

Additionally, we used generative AI tools for (1) creating or editing software code, (2) draft parts of the research paper, (3) summarizing or analyzing existing literature, (4) editing the research paper to improve readability and (5) identifying relevant literature.

We have reviewed all AI-assisted work. All LLM-generated code for the dataset creation pipeline and experiment harness was verified. A subset of the LLM-assisted data annotations from Section 3.3 were manually reviewed for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code for the evaluation harness along with detailed steps to reproduce the experiments discussed in the paper are available here. The dataset is released here.

REFERENCES

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15068–15083, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.840. URL https://aclanthology.org/2024.emnlp-main.840/.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 9112–9141, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/25f7be9694d7b32d5cc670927b8091e1-Paper-Conference.pdf.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016. URL https://arxiv.org/abs/1611.09268.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 238–251, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1022. URL https://aclanthology.org/N18-1022/.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/29728/31250.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In Dan Jurafsky,

Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.207. URL https://aclanthology.org/2020.acl-main.207/.

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.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130?

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2288–2292, 2021. URL https://dl.acm.org/doi/abs/10.1145/3404835.3463098.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce

Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Deyan Ginev. ar5iv:04.2024 dataset, an html5 conversion of arxiv.org, 2024. URL https://sigmathling.kwarc.info/resources/ar5iv-dataset-2024/. SIGMathLing – Special Interest Group on Math Linguistics.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking, 2022. URL https://arxiv.org/abs/2112.01206.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435. URL https://www.jmlr.org/papers/volume24/23-0037/23-0037.pdf.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020. URL https://dl.acm.org/doi/abs/10.1145/3397271.3401075.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.243. URL https://aclanthology.org/2025.naacl-long.243/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational*

Linguistics, 7:453–466, 2019. URL https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00276/1923288/tacl_a_00276.pdf.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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

Jerry Liu. LlamaIndex, 11 2022. URL https://github.com/jerryjliu/llama_index.

Yao Lu, Yue Dong, and Laurent Charlin. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8068–8074, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.648. URL https://aclanthology.org/2020.emnlp-main.648/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

NVIDIA. Nvidia nemotron 3: Efficient and open intelligence, 2025. URL https://arxiv.org/abs/2512.20856. White Paper.

OpenAI. Gpt-5.6 luna (gpt-5.6-luna) [large language model], 2026. URL https://openai.com/index/gpt-5-6/.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases?, 2019. URL https://arxiv.org/abs/1909.01066.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. Questions are all you need to train a dense passage retriever. *Transactions of the Association for Computational Linguistics*, 11:600–616, 2023. doi: 10.1162/tacl_a_00564. URL https://aclanthology.org/2023.tacl-1.35/.

Henrique* Schechter Vera, Sahil* Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divya Sreepat, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025. URL https://arxiv.org/abs/2509.20354.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, 2021. URL https://aclanthology.org/2021.findings-emnlp.320.pdf.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972. URL https://www.emerald.com/jd/article-pdf/28/1/11/1336479/eb026526.

Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Arik, Danqi Chen, and Tao Yu. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 48941–48991, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7a0f8055c838df8e62329a76c7c6403d-Paper-Conference.pdf.

Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024. URL https://arxiv.org/abs/2401.15391/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/65b9eea6e1cc6bb9f0cd2a47751a186f-Paper-round2.pdf.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. URL https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00475/2020694/tacl_a_00475.pdf.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023. URL https://aclanthology.org/2023.acl-long.557.

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.

Yunjia Xi, Jianghao Lin, Menghui Zhu, Yongzhao Xiao, Zhuoying Ou, Jiaqi Liu, Tong Wan, Bo Chen, Weiwen Liu, Yasheng Wang, Ruiming Tang, Weinan Zhang, and Yong Yu. Infodeepseek: Benchmarking agentic information seeking for retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2505.15872.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. Crag - comprehensive rag benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10470–10490. Curran Associates, Inc., 2024. doi: 10.52202/079017-0335. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper-Datasets_and_Benchmarks_Track.pdf.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

Table 10: Generation parameters for all models.

Parameter	Value
Temperature	1.0
Top P	1
Context Window	131072
Max Retries	1
Reasoning Effort	High
Max tool calls	8
Chunks per retrieval	8

A Data Verification

There are many points of failure when converting the cited papers to markdown with LaTeXML which can lead to malformed output, e.g. parsing errors, outdated sources, top-level file selection etc. To verify high-quality conversion for such a large corpus, we employ a mix of Word Error Rate (WER) and Word Coverage metrics with respect to the text extracted from the PDF. Word coverage is defined as the fraction of unique words present in the PDF which are also present in the converted markdown. We disregard all non-ASCII symbols and math tokens when calculating these metrics. After calibrating these metrics on a set of known oracle markdown-PDF pairs, they allow us to reliably filter erroneous conversions.

For the source papers, all the markdown was validated by manual review.

B Experimental Details

B.1 Inference Details

Since the models in the study do not specify recommended generation parameters, we use the same parameters for all evaluations, specified in Table 10.

Max tool calls is set to 8 as it was found to be the highest the models could accommodate before reaching approaching context limits.

B.2 Prompts

We use four distinct prompts across our experimental conditions. All prompts share a common output format specification: a JSON object with a `cited_ids` field containing a ranked list of arXiv IDs in descending order of citation likelihood, and a one-sentence `reasoning` field.

The linear scan prompts follow a two-stage design: the first stage elicits a binary yes/no prediction per candidate, and the second stage presents all positively selected candidates for final ranking.

All prompts begin with the same preamble:

```
You are predicting which papers are cited in the 'related works'
section of a research paper.
## Source Paper Context
{src_paper}
```

No Retrieval

```
## Task
```

Table 11: Summary of prompt components across experimental conditions. All prompts include the source paper in context and the shared output format specification described above.

Condition	Additional Context	Instruction	Output
No Retrieval	None	Predict and rank related works from parametric knowledge	Ranked `cited_ids`
Agentic Retrieval	Retrieved chunks (upon tool call)	Query the tool, rank retrieved papers by citation likelihood	Ranked `cited_ids`
Linear Scan Stage 1	Candidate title and abstract	For each candidate, predict if it is a likely related work	Binary prediction (yes/no)
Linear Scan Stage 2	All Stage 1 positives	Rank the selected candidates by citation likelihood	Ranked `cited_ids`

```
Based on your understanding of the paper, list the research papers
that are likely cited as related works.  Reference the papers using
their arXiv ID. Rank the papers on the basis of their likelihood of
being cited.  You may rank as many papers as you wish.
{format_spec}
```

Agentic Retrieval

```
## Task
Based on your understanding of the paper, list the research papers
that are likely cited as related works.  You are provided with a query
tool that you must call with the relevant query to retrieve candidate
papers.  The query tool will retrieve excerpts based on similarity.
Design your queries accordingly.  You may call this tool a maximum of
{max_calls} times.
{format_spec}
```

Linear Scan Stage 1 Prompt

```
## Candidate Paper
Title:  {candidate_title}
Abstract:  {candidate_abstract}
## Task
Based on your understanding of the paper, predict whether the
candidate paper is likely to be cited as a related work.
{format_spec}
```

Linear Scan Stage 2 Prompt

```
## Selected Candidates
You previously selected a collection of papers that you deemed are
likely to be cited as related works in the source paper.  These are
given below:  {selected_candidates}
## Task
```

```
Rank the selected candidates by their likelihood of being cited as
related works.
{format_spec}
```

B.3 HUMAN BLIND REVIEW

The blind review was performed by the co-authors. A screenshot of the interface used for the same is shown in Fig. 3.



Figure 3: Interface used for blind review

The reviewers were instructed to read and analyze a source paper, followed by a classification of a set of candidate papers as being related or not.