

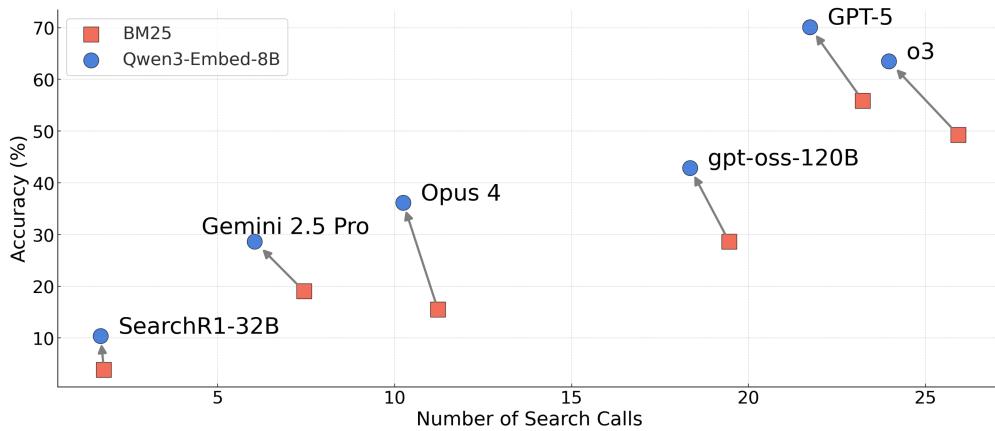
000 BROWSECOMP-PLUS: A MORE FAIR AND TRANS- 001 PARENT EVALUATION BENCHMARK OF DEEP SEARCH 002 AGENTS

003 **Anonymous authors**

004 Paper under double-blind review

010 ABSTRACT

013 Deep search agents, which integrate large language models (LLMs) with search
014 tools, have shown success in improving the effectiveness of handling complex
015 queries that require iterative search planning and reasoning over search results.
016 Evaluations on current benchmarks like BrowseComp relies on black-box live
017 web search APIs, have notable limitations in (1) *fairness*: dynamic and opaque
018 web APIs hinder fair comparisons and reproducibility of deep search agent meth-
019 ods; (2) *transparency*: lack of control over the document corpus makes it diffi-
020 cult to isolate retriever contributions. To address these challenges, we introduce
021 BROWSECOMP-PLUS, a benchmark derived from BrowseComp, employing a
022 fixed, carefully curated corpus. Each query in BROWSECOMP-PLUS includes
023 human-verified supporting documents and mined challenging negatives, enabling
024 controlled experimentation. The benchmark is shown to be effective in distin-
025 guishing the performance of various deep search agents. For instance, the fully
026 open-sourced method Search-R1, when paired with the BM25 retriever, achieves
027 3.86% accuracy, whereas the GPT-5 achieves 55.9%. Integrating the GPT-5 with
028 the Qwen3-Embedding-8B retriever further enhances its accuracy to 70.1% with
029 fewer search calls. This benchmark allows comprehensive evaluation and disentan-
030 gled analysis of deep search agents and retrieval methods, fostering insights into
031 retrieval effectiveness, citation accuracy, and context engineering in deep search
032 agents. Code and data will be released.



049 Figure 1: Accuracy vs. number of search calls for deep search agents with different retrievers. GPT-5,
050 o3, gpt-oss are evaluated with high reasoning effort. The figure shows that **deep search agents mostly**
051 **improve the final accuracy at a cost of more search calls**, whereas **better retrieval systems not**
052 **only improve the overall accuracy but also reduce the number of search calls**. For reference,
053 GPT-5 achieves 59.9% accuracy when evaluated using the Google Search API.

054 **1 INTRODUCTION**

055

056 Recent benchmarks for evaluating deep search agents, such as BrowseComp (Wei et al., 2025), have
 057 showcased the impressive capabilities of combining large language models (LLMs) with web search
 058 tools in solving complex, reasoning-intensive queries. These benchmarks typically provide sets
 059 of queries paired directly with answers, where agents are employed with live web search APIs to
 060 retrieve supporting documents in real time (Zhou et al., 2025; Chen et al., 2025). While this approach
 061 effectively assesses the end-to-end performance of deep search agents, it introduces several critical
 062 limitations that impede systematic analysis and evaluation of individual system components.

063

- 064 • **Fair Comparison of Deep Search Agents.** Current evaluations of deep search agents often
 065 conflate agent system performance with the effectiveness of their retrieval components, making
 066 it difficult to achieve fair and consistent comparisons across systems. This entanglement also
 067 severely undermines the reproducibility of experiments, which is a key requirement for rigorous
 068 evaluation (Voorhees, 2019).
- 069 • **Transparency of Retrieval Process.** The transparency of the retrieval process comes from two
 070 aspects: the retrieval algorithm and the target retrieval corpus. In the current evaluation pipelines,
 071 supporting documents are obtained through black-box web search APIs that operate over the
 072 entire internet, which are highly dynamic in content and consistently evolving over time. The
 073 lack of a controlled retrieval process hinders the evaluation of retrieval models' contribution to
 074 deep-research agents.
- 075 • **Accessibility:** The dependence on commercial web search APIs introduces substantial practical
 076 constraints, including high operational costs and variability in retrieval quality. These issues not
 077 only limit accessibility but also introduce unnecessary complexity and uncertainty.

078 To address these limitations and enable precise, reproducible, transparent, and component-focused
 079 evaluation of deep search agents, we introduce `BROWSECMP-PLUS`, a new benchmark dataset.
 080 `BROWSECMP-PLUS` extends the original BrowseComp dataset (Wei et al., 2025) by providing a
 081 fixed and curated corpus of documents specifically selected and verified by human annotators. Each
 082 query in `BROWSECMP-PLUS` is accompanied by explicitly identified supportive documents and
 083 hard negative documents. This carefully collected document corpus allows researchers to evaluate the
 084 retrieval and LLM agent components independently, facilitating detailed analysis of each component's
 085 impact on the final answer quality. Additionally, by eliminating reliance on dynamic web APIs,
 086 `BROWSECMP-PLUS` significantly reduces costs, enhances reproducibility, and improves the overall
 087 robustness of benchmarking in deep search agents.

088 To demonstrate the utility of `BROWSECMP-PLUS`, we conduct comprehensive evaluations by
 089 pairing various open- and closed-source LLMs with a range of retrieval models on our curated corpus.
 090 This setup allows us to systematically analyze how different combinations affect answer quality
 091 and to identify where performance bottlenecks lie, whether in the retriever or the language model.
 092 We find that even when equipped with state-of-the-art retrievers, Deep-Research agents still face
 093 substantial challenges in consistently surfacing all necessary evidence, for reasoning-intensive queries.
 094 These findings motivate the need for evaluation frameworks that disentangle retrieval from reasoning,
 095 support fine-grained component analysis, and remain fully reproducible.

096 Furthermore, we extend our evaluation to test retrieval models directly on the original BrowseComp
 097 queries, an analysis that was previously infeasible due to the absence of a fixed corpus and grounded
 098 relevant document judgments. Our findings reveal that even state-of-the-art retrieval models struggle
 099 to retrieve relevant documents for these complex, reasoning-intensive queries.

100 In summary, our contributions are threefold:

101

- 102 • We present `BROWSECMP-PLUS`, a fair and transparent benchmark for deep search agents,
 103 featuring a fixed, human-verified corpus with supporting and challenging negative documents.
- 104 • We provide the first systematic analysis of retrieval–agent interactions under controlled conditions,
 105 evaluating a broad range of retrievers and LLM-based agents.
- 106 • We release all benchmark data, evaluation scripts, and baselines to facilitate reproducible research
 107 and foster future advances in various dimensions to improve the deep-research system.

108 **2 RELATED WORKS**109 **2.1 DEEP SEARCH AGENT**

110 Deep search agents conduct tasks through iterative query reasoning, search planning, and reflection
 111 on retrieved results (Asai et al., 2024), outperforming the traditional single-round retrieval-augmented
 112 generation paradigm (Lewis et al., 2020). Commercial closed-source models such as Gemini (Gemini
 113 2.5 Team, 2025), Opus (Anthropic Team, 2024b), and o3 (OpenAI Team, 2025a), as well as open-
 114 source models like GPT-OSS (OpenAI Team, 2025b), allow access to external retrievers via tool-use
 115 APIs or MCP (Anthropic Team, 2024a). Recent research works such as Search R1 (Jin et al.,
 116 2025b) and WebSailor (Li et al., 2025), both based on the Qwen (Yang et al., 2025) model, leverage
 117 reinforcement learning to further enhance search tool capabilities. Fair evaluation of such agents,
 118 however, requires a fixed retriever system to make comparisons meaningful.
 119

120 **2.2 NEURAL RETRIEVAL**

121 Neural retrieval methods, such as Dense Passage Retrieval (Karpukhin et al., 2020), encode queries
 122 and documents into dense vectors using transformer models, and perform retrieval through nearest-
 123 neighbor search (Douze et al., 2024). These methods have significantly improved retrieval ef-
 124 fectiveness compared to traditional lexical-based methods like BM25 (Robertson, 1994). Recent
 125 improvements in neural retrievers include advanced training strategies such as continuous pretrain-
 126 ing (Chen et al., 2024; Gao & Callan, 2022), data augmentation (Li et al., 2023; Ma et al., 2025b;
 127 Shao et al., 2025), integration of large language models as backbones (Ma et al., 2024; Wang et al.,
 128 2023), and LLM distillation techniques (Lee et al., 2024; Zhang et al., 2025). While retrievers are
 129 a critical component of deep search agents, the contribution of different retrievers to the overall
 130 performance of these agents remains underexplored.
 131

132 **2.3 DEEP SEARCH BENCHMARKS**

133 Traditional benchmarks such as NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi
 134 et al., 2017) have significantly contributed to evaluating retrieval and retrieval-augmented generation
 135 systems (Lewis et al., 2020; Karpukhin et al., 2020; Lin et al., 2024). However, these benchmarks
 136 primarily feature single-hop questions, which typically do not require multi-step reasoning or iterative
 137 retrieval. Although datasets like HotpotQA (Yang et al., 2018) offer multi-hop questions, they are
 138 shallow in depth (2 hops), and the corpus is limited to Wikipedia. To robustly evaluate deep search
 139 agents capable of complex reasoning and iteratively retrieve many turns, benchmarks requiring
 140 deep multi-turn query interactions are essential. BrowseComp (Wei et al., 2025) stands out as a
 141 benchmark explicitly designed for this purpose, offering complex queries paired with verifiable
 142 answers. Recent extensions of BrowseComp concepts, such as ZH-BrowseComp (Zhou et al., 2025)
 143 and MedBrowseComp (Chen et al., 2025), further expand to multilingual queries and domain-specific
 144 challenges. Mind2Web2 (Gou et al., 2025) on the other hand proposed to evaluating time-varied
 145 questions with agent-as-judge.
 146

147 Existing benchmarks primarily focus on question-answer evaluations of integrated systems with-
 148 out standardized corpora, complicating comparative assessments of retrieval methodologies.
 149 BROWSECMP-PLUS facilitates fair and comprehensive evaluations by providing human-verified
 150 corpus, expanding the classic Cranfield paradigm (Voorhees, 2002) to modern deep search agent
 151 evaluation.
 152

153 **3 BROWSECMP-PLUS**154 **3.1 PRELIMINARY: BROWSECMP**

155 The BrowseComp benchmark contains 1,266 challenging fact-seeking questions specifically designed
 156 to assess the capability of deep search agents to interactively and creatively navigate the web for
 157 complex, hard-to-find information (Wei et al., 2025). The questions are deliberately constructed
 158 to be difficult for both humans and LLMs, yet they feature verifiable, concise answers, enabling
 159 straightforward evaluation through simple answer matching. While effective and widely employed
 160

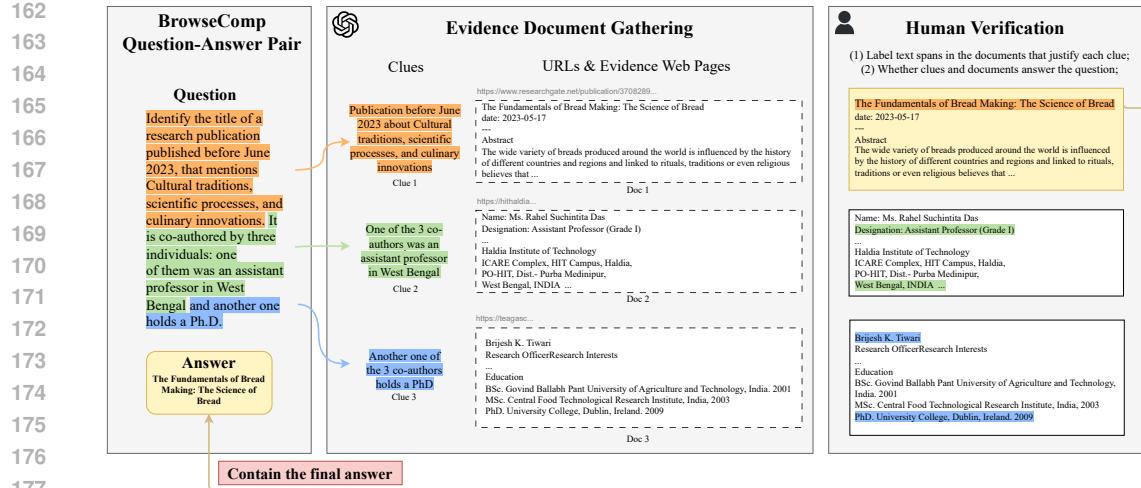


Figure 2: The two-stage pipeline of collecting evidence documents in the corpus (Section 3.2).

for end-to-end evaluation of deep search agents with web search access, this approach complicates the isolated measurement of retrieval effectiveness within these frameworks.

3.2 BUILDING THE DOCUMENT CORPUS

Constructing a corpus for BrowseComp questions is non-trivial. Three key challenges need to be addressed:

1. **Comprehensive coverage:** The corpus must provide complete evidence to support the entire reasoning chain required to answer each question.
2. **Retrieval difficulty:** The corpus should contain enough distracting negative documents so that search agents and retrievers are challenged in locating the correct evidence.
3. **Practical size:** The corpus should be large enough to yield reliable research insights, while avoiding overly-large computation costs for research purposes.

To meet these criteria, we curate evidence documents through a two-stage pipeline involving automated evidence mining followed by human verification, and perform hard-negative mining via web search to attach challenging, distracting documents to each query. The sections below describe this process in detail and present a 100k-document corpus that effectively supports the study of deep search agents.

3.2.1 EVIDENCE DOCUMENT GATHERING

The original BrowseComp dataset contains only question-answer pairs, without the URLs of the web pages that support these answers. To build a document collection with supporting evidence, the first step involves identifying relevant web pages for each question.

To achieve this, we leverage the OpenAI o3 model with web search enabled. Since the datasets intentionally make direct retrieval of relevant documents difficult, we adopt a *reverse-engineering* strategy: We provide the answer together with the question and instruct the model to search the web for pages that have evidence supporting the answers. We also ask the model to structure the output in a table format with three columns: (1) Clue: the part of the question to address; (2) URL: the web page link containing evidence supporting the clue; and (3) Evidence: the content from the web page that supports the clue. The purpose of this table format is to facilitate human annotators in verifying each clue and its corresponding web page in the next step. An example prompt for this step is provided in Appendix A.

216 Of the 1,266 original question-answer pairs in BrowseComp, the OpenAI o3 model fails to provide
 217 supporting evidence for 124 pairs, either due to output formatting errors or because the model abstains
 218 from answering due to low confidence. For the remaining 1,142 pairs, we scrape the URLs cited as
 219 evidence using Selenium,¹ and parse them with Trafilatura (Barbaresi, 2021). However, a combination
 220 of hallucinated URLs and scraping challenges prevents us from successfully scraping all of them. As
 221 a result, we exclude 137 question-answer pairs that contain at least one URL where we are unable to
 222 scrape, as missing a URL for a clue will make the question incomplete to answer.

223 This leaves us with 1,005 queries for the next stage: human verification.
 224

225 3.2.2 EVIDENCE DOCUMENT VERIFICATION 226

227 In this stage, we aim to verify that the documents contain sufficient evidence for each clue in the
 228 questions. For each question-answer pair, we present human annotators with the output table from
 229 OpenAI o3 in the previous stage, with URLs replaced by the corresponding processed documents.

230 Annotators are asked to:
 231

- 232 1. Confirm that each clue is sufficiently justified by the supporting documents. Instead of simply
 233 confirming the match, annotators must label the text spans in the documents that justify each clue,
 234 as this explicit step encourages high-quality verification.
- 235 2. Determine whether the combination of clues and supporting evidence enables a human to answer
 236 the *entirety* of the question correctly. For instance, if a query asks for an individual matching five
 237 characteristics, all five must be verifiable from the documents.

238 If the original output from OpenAI o3 fails to meet both criteria, annotators are instructed to revise
 239 the clues and search the web for additional supporting documents for at least 20 minutes, before
 240 concluding that the desired evidence documents cannot be collected.
 241

242 In addition to constructing the evidence document set, annotators also label which documents directly
 243 contain the final answer; these are designated as *gold documents*. Note that a gold document is not
 244 defined merely by containing the ground-truth answer as an exact substring; in some cases, the answer
 245 is included in the document in an implicit way. For example, a question might ask for the number
 246 of publications by a particular author, with the ground-truth answer being “7”. A gold document in
 247 this case could be the author’s personal webpage listing their publications; while it may not contain
 248 the string “7” explicitly, it logically contains the answer. Similarly, there are many cases where the
 249 answer appears in the document in a variant form, such as a different date format or a paraphrased
 250 phrase, rather than an exact string match. Our goal in constructing the gold document set is to provide
 251 a more robust and semantically meaningful alternative to the simple substring-based approach in
 252 identifying documents that contain the final answer.

253 Figure 2 illustrates the complete evidence document collection process. A detailed example, including
 254 a screenshot of the labeling interface shown to human annotators, is provided in Appendix B.

255 For quality control, we sample each annotator’s labeled data and cross-validate them among annota-
 256 tors, showing over 80% of agreement on average. Overall, of the 1,005 question-answer pairs from
 257 the previous stage, 830 passed human verification. The most common failure mode occurs when the
 258 documents provided by OpenAI o3 do not satisfy the two verification criteria, and human annotators
 259 are unable to gather sufficient additional evidence within a reasonable effort. In addition to these, we
 260 identify and exclude several other categories of problematic cases as detailed in Appendix C.

261 The entire labeling process involved 14 university student annotators and required over 400 hours of
 262 manual effort.

263 3.3 HARD NEGATIVE MINING 264

265 To ensure the collected corpus remains a reasonable size while still being challenging enough for
 266 search systems to identify correct answers among distracting documents, we mine hard negative
 267 documents via web search to form the corpus. This has proven to be effective in evaluating information
 268 retrieval systems using a sub-sampled corpus (Fröbe et al., 2025; Zhuang & Zuccon, 2022).
 269

¹<https://www.selenium.dev/documentation>

270 Specifically, we take each question from BrowseComp and prompt GPT-4o to break it down into
 271 simpler, self-contained sub-queries. On average, this results in about seven sub-queries per original
 272 query. Each sub-query is then sent to a Google Search API provider (SerpAPI), which returns up to
 273 100 search results. We scrape these results using the same process used for collecting documents
 274 during positive example construction. We illustrate this hard negative document collecting process in
 275 Figure 4. The prompt used to create these sub-queries is provided in Appendix D.

277 3.4 FINAL CORPUS STATISTICS

279 After deduplicating the positive and negative documents collected as above, we arrive at a corpus
 280 of 100,195 documents, along with 830 queries. On average, each query contains 6.1 evidence
 281 documents, 76.28 negatives, and 2.9 gold documents. Each document averages 5179.2 words and
 282 32296.2 characters.

283 4 EXPERIMENTS

285 4.1 EXPERIMENT SETUP

287 **Search Agents** We list the agent baseline models in Appendix H.1. To perform agentic search with
 288 the LLMs, we provide the LLM with a retriever tool as tool use. We follow the original prompt from
 289 BrowseComp (Wei et al., 2025), which instructs the model to answer a given question along with
 290 a confidence estimate (expressed as a percentage). There are two revisions of the original prompts:
 291 (1) We explicitly prompt the LLM to use the provided tools to adapt to our custom search tool; (2)
 292 We instruct the model to cite the sources when generating the final answer, enabling the evaluation
 293 of citation quality. The complete prompt is included in Appendix E. We use this prompt across all
 294 models except Search-R1, which uses the prompt aligned with its original fine-tuning.

295 **Retriever** We list the retriever baseline models in Appendix H.2. The retriever tool is set to retrieve
 296 the top $k = 5$ search results, where each result is truncated to the first 512 token of the corresponding
 297 document. This truncation is due to budget constraints, which prevent us from providing full document
 298 content. To assess the impact of this design choice, we analyze the distribution of the number of
 299 tokens required to include the ground-truth answer for each query. As illustrated in Figure 5 (b), when
 300 documents are truncated to the first 512 tokens, 86.5% of queries still contain the ground-truth answer
 301 in at least one of their gold documents. Further ablations exploring alternative tool configurations are
 302 discussed in Section 4.7.

304 4.2 EVALUATION METRICS

306 **Deep Search Agent Effectiveness** We report end-to-end effectiveness of the deep search agents
 307 with three metrics: Accuracy, Recall, and Search Calls. Accuracy follows BrowseComp: an LLM-
 308 as-judge (GPT-4.1) compares the model’s final answer against the ground truth using the evaluation
 309 prompt listed in Appendix F. Recall measures how many human-verified evidence documents the
 310 agent retrieved during its entire interaction. Search Calls is the average number of search API
 311 invocations per query. In addition, following BrowseComp, we compute calibration error using the
 312 confidence estimates produced by the search agents, in the same way as Humanity’s Last Exam (Phan
 313 et al., 2025), measuring how closely a model’s predicted confidence matches the actual accuracy of its
 314 predictions. For Search-R1, we do not report calibration error because the input and output format of
 315 this model are fixed without a confidence source output. Lastly, to understand whether the accuracy
 316 obtained by each agent stems from its agentic ability or merely its parametric knowledge, we also
 317 evaluate each LLM’s accuracy when directly prompted with the question, without any retriever or
 318 external knowledge.

319 **Retrieval Effectiveness** For evaluating retriever effectiveness, our `BROWSECOMP-PLUS` bench-
 320 mark provides human-verified evidence documents and gold documents, along with a fixed test
 321 document collection, enabling evaluation under the Cranfield paradigm (Voorhees, 2019). Specifi-
 322 cally, we follow standard TREC practice to create a query-document relevance label file² for both

323 ²Known as a qrel file.

Table 1: End-to-end agent accuracy on `BROWSECOMP-PLUS` across LLMs and retrievers. All agents are prompted with the same tool-use prompt, except for Search-R1, which uses the prompt identical to its training.

LLM	Retriever	Accuracy	Recall	Search Calls	Calibration Error
GPT-4.1	None	3.86%	N/A	N/A	73.83%
	BM25	14.58%	16.42%	10.35	68.96%
	Qwen3-Embed-8B	35.42%	36.89%	8.67	54.67%
o3	None	19.52%	N/A	N/A	14.07%
	BM25	49.28%	56.64%	25.93	12.58%
	Qwen3-Embed-8B	63.49%	73.24%	23.97	16.77%
GPT-5	None	26.18%	N/A	N/A	24.57%
	BM25	55.90%	61.70%	23.23	13.50%
	Qwen3-Embed-8B	70.12%	78.98%	21.74	9.11%
Sonnet 4	None	1.69%	N/A	N/A	40.92%
	BM25	14.34%	21.31%	9.95	29.79%
	Qwen3-Embed-8B	36.75%	47.33%	9.03	24.51%
Opus 4	None	2.42%	N/A	N/A	11.95%
	BM25	15.54%	22.96%	11.22	22.00%
	Qwen3-Embed-8B	36.14%	50.84%	10.24	12.79%
Gemini 2.5 Flash	None	3.13%	N/A	N/A	79.01%
	BM25	15.54%	21.45%	10.56	29.28%
	Qwen3-Embed-8B	33.01%	40.19%	9.77	21.63%
Gemini 2.5 Pro	None	7.47%	N/A	N/A	76.72%
	BM25	19.04%	22.81%	7.44	51.58%
	Qwen3-Embed-8B	28.67%	35.31%	6.04	44.08%
gpt-oss-120B-high	None	3.13%	N/A	N/A	48.89%
	BM25	28.67%	35.50%	19.45	46.48%
	Qwen3-Embed-8B	42.89%	52.63%	18.35	40.34%
Qwen3-32B	None	0.96%	N/A	N/A	67.98%
	BM25	3.49%	3.12%	0.92	57.41%
	Qwen3-Embed-0.6B	4.10%	3.45%	0.91	60.71%
	Qwen3-Embed-4B	7.83%	6.20%	0.89	61.06%
	Qwen3-Embed-8B	10.36%	7.80%	0.94	59.84%
	ReasonIR	9.16%	7.59%	0.91	55.15%
SearchR1-32B	None	0.48%	N/A	N/A	N/A
	BM25	3.86%	2.61%	1.78	N/A
	Qwen3-Embed-0.6B	5.66%	5.30%	1.73	N/A
	Qwen3-Embed-4B	9.40%	7.90%	1.68	N/A
	Qwen3-Embed-8B	10.36%	10.17%	1.69	N/A
	ReasonIR	9.43%	8.37%	1.74	N/A

evidence documents and gold documents separately, and then compute Recall@k and nDCG@k to assess the effectiveness of retrievers.

4.3 END-TO-END DEEP SEARCH AGENTS PERFORMANCE

Table 1 summarizes the overall deep search performance across different LLMs and retrievers. Proprietary models (GPT-4.1, o3, GPT-5, Sonnet-4, Opus-4, Gemini) demonstrate high answer accuracy, with OpenAI’s GPT-5 achieving the highest accuracy (70.12%) when paired with the Qwen3-Embedding-8B retriever. Open-source models such as Qwen3-32B and SearchR1-32B lag behind proprietary models. With Qwen3-Embedding-8B as the retriever, Qwen3-32B achieves only 10.36% accuracy, compared to 35.42% for GPT-4.1 and 63.49% for o3. Notably, the only high-performing open-source model we studied is gpt-oss-120B in its high reasoning mode, which achieves 42.89% accuracy, surpassing Opus 4 when both are paired with Qwen3-Embedding-8B.

In general, closed-source agents call the search tool more frequently than open-source models. For instance, OpenAI’s GPT-5 and o3 issue an average of more than 20 search calls per query, while Qwen3-32B and SearchR1-32B make fewer than 2, despite being explicitly prompted to use the tool. This reflects a test-time scaling effect: more exhaustive search correlates with better outcomes and aligns with prior findings that reasoning-intensive queries benefit from exploratory retrieval.

378 Table 2: Effectiveness of retrievers. The complete question is used as the query for all retrieval
 379 methods for fair comparison.

381 Retriever	382 Recall@5	382 Recall@100	382 Recall@1000	382 nDCG@10
Evidence Document Retrieval				
383 BM25	384 1.2	384 4.7	384 13.7	384 1.6
385 jina-colbert-v2	386 5.7	386 18.1	386 35.7	386 7.9
387 Qwen3-Embed-0.6B	388 6.2	388 26.5	388 59.7	388 8.0
389 Qwen3-Embed-4B	390 9.8	390 40.2	390 71.8	390 14.0
391 Qwen3-Embed-8B	392 14.5	392 47.7	392 76.7	392 20.3
393 ReasonIR-8B	394 12.2	394 43.6	394 73.9	394 16.8
Gold Document Retrieval				
395 BM25	396 1.4	396 6.1	396 17.3	396 1.7
397 jina-colbert-v2	398 6.6	398 20.4	398 39.7	398 6.8
399 Qwen3-Embed-0.6B	400 8.5	400 30.5	400 66.2	400 7.4
401 Qwen3-Embed-4B	402 13.0	402 47.3	402 77.0	402 13.6
403 Qwen3-Embed-8B	404 18.5	404 55.8	404 83.5	404 19.5
405 ReasonIR-8B	406 15.3	406 49.7	406 78.9	406 15.5

395 In the parametric-only setting where no retrieval of external knowledge is used, most LLMs show
 396 very limited accuracy. Only o3 and GPT-5 perform notably better, correctly answering about 20% of
 397 the questions; this may suggest that these models were trained on BrowseComp. When comparing
 398 across different LLM agents, this potential contamination is another important factor to keep in mind.
 399

400 4.4 EFFECT OF RETRIEVAL QUALITY

402 A consistent trend observed across all models is that stronger retrieval leads to higher final accuracy.
 403 First, consider the retriever’s effectiveness on our dataset. We evaluate retrieval performance using
 404 the original full queries, with results shown in Table 2. Compared to BM25, Qwen3-Embedding-8B
 405 and ReasonIR-8B achieve substantially higher recall and nDCG for both evidence document retrieval
 406 and gold document retrieval. Notably, we observe a model size scaling law within the Qwen3
 407 embedding family; larger models consistently perform better, with Qwen3-Embedding-8B surpassing
 408 ReasonIR-8B at the 8B scale.

409 Now, as indicated in Table 1, replacing the BM25 retriever with a stronger retriever leads to significant
 410 accuracy gains across all LLM agents. For instance, OpenAI’s GPT-5’s accuracy improves from
 411 55.9% to 70.12%, while Sonnet 4 and Opus 4 both achieve more than double their BM25 accuracy.
 412 This suggests a strong positive correlation between retrieval effectiveness and research agent accuracy.

413 Moreover, stronger retrievers potentially reduce the number of search calls. For most proprietary
 414 models, Qwen3-Embedding-8B reduces search calls by approximately 1–3 compared to BM25. This
 415 shows that better retrieval not only improves effectiveness (accuracy) but also efficiency (fewer tool
 416 calls). In Appendix K, we also report differences in proprietary agent API costs when using different
 417 retrievers. Agents using Qwen3-Embedding-8B incur lower costs due to fewer input and output
 418 tokens, further supporting the efficiency gains enabled by stronger retrieval.

419 In addition, Appendix I reports the coverage, average number, precision, and recall of the document
 420 citations attributed by the agent during answer generation, highlighting opportunities for future
 421 improvements in long-answer generation with proper evidence attribution via citations. We also
 422 assess the role of LLM rerankers as part of the retrieval module in Appendix J, showing the potential
 423 of further improving the effectiveness of deep search agents through reranking.

425 4.5 ORACLE RETRIEVAL

427 We evaluate effectiveness in an extreme oracle setting, where search agents are prompted with all
 428 labeled positive documents to answer the questions. In this setup, GPT-4.1 achieves an accuracy
 429 of 93.49%. This highlights two key points. First, it showcases the importance of the retriever: if
 430 the retriever is of perfect quality, search agents can attain substantially high accuracy on complex
 431 reasoning tasks in BROWSECOMP-PLUS, in contrast to the 14.58% baseline accuracy of GPT-4.1
 when using BM25 as the retriever. Second, it validates the quality of the BROWSECOMP-PLUS

432
433
434 Table 3: OpenAI gpt-oss models in different reasoning effort settings
435
436
437
438
439
440
441
442
443
444
445
446
447
448

LLM	Retriever	Accuracy	Recall	Search Calls	Calibration Error
gpt-oss-20B-low	BM25	4.11%	5.36%	1.89	40.89%
	Qwen3-Embed-8B	13.37%	17.37%	1.87	36.34%
gpt-oss-20B-medium	BM25	16.39%	21.96%	13.72	41.78%
	Qwen3-Embed-8B	29.88%	41.31%	13.64	35.99%
gpt-oss-20B-high	BM25	21.08%	31.98%	26.87	33.42%
	Qwen3-Embed-8B	34.58%	49.29%	23.87	27.81%
gpt-oss-120B-low	BM25	9.52%	8.54%	2.06	43.59%
	Qwen3-Embed-8B	24.94%	22.50%	2.21	40.96%
gpt-oss-120B-medium	BM25	23.73%	27.02%	9.73	45.78%
	Qwen3-Embed-8B	37.59%	43.45%	9.64	41.77%
gpt-oss-120B-high	BM25	28.67%	35.50%	19.45	46.48%
	Qwen3-Embed-8B	42.89%	52.63%	18.35	40.34%

449
450
451
452
453
454
455
456 Table 4: Comparison of Qwen3-32B and GPT-4.1 with and without get-document tool, using Qwen3-
457 Embedding-8B as retriever.
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

Model	Accuracy	Search Calls	Get Document Calls	Calibration Error
GPT-4.1	35.42%	8.67	N/A	54.67%
GPT-4.1 + get-doc	43.61%	10.03	1.85	54.28%
Qwen3-32B	10.36%	0.94	N/A	59.84%
Qwen3-32B + get-doc	11.69%	1.01	0.27	56.47%

corpus itself: GPT-4.1, a non-reasoning model, is able to correctly answer 93.49% of questions using only the evidence documents in the corpus. For the remaining 6.51% of cases, human annotators reviewed each instance and confirmed that the answers are indeed answerable from the positive documents; the errors stem solely from GPT-4.1’s failure to reason correctly.

4.6 IMPACT OF REASONING EFFORT

We evaluate how the reasoning effort of LLMs influences answer quality and retrieval behavior. To isolate this effect, we focus on the gpt-oss family, which offers three reasoning modes: *low*, *medium*, and *high*. As shown in Table 3, increasing the reasoning effort leads to substantial improvements in accuracy and recall across all model sizes and retrievers. For example, gpt-oss-20B with Qwen3-Embed-8B improves from 13.37% accuracy in *low* mode to 34.58% in *high* mode, along with a recall jump from 17.37% to 49.29%. Similarly, gpt-oss-120B with Qwen3-Embed-8B’s accuracy rises from 24.94% to 42.89%. These gains, however, come with a trade-off: higher reasoning modes dramatically increase the average number of search calls (e.g., from ≈ 2 to ≈ 24 for gpt-oss-20B with Qwen3-Embed-8B), implying higher computational and latency costs. Interestingly, calibration error tends to decrease with higher reasoning effort, suggesting that the models become more aligned between confidence and correctness as they reason more extensively.

4.7 EFFECT OF DOCUMENT READING STRATEGY

In previous experiments, we always presented only the first 512 tokens of each retrieved document as a preview to the LLM during each round of search and reasoning, due to token budget constraints. However, in realistic deep search scenarios, agents often have access to a document reader tool that enables reading the full content of a document. To evaluate the potential benefit of such a tool, we conduct experiments with GPT-4.1 and Qwen3-32B, both with and without access to a whole-document reader (referred to as the get-document tool). Appendix G contains the revised prompt used when the get-document tool is added.

Results are shown in Table 4. For GPT-4.1, enabling the get-document tool improves accuracy from 35.42% to 43.61%, with a modest increase in search calls (from 8.67 to 10.03) and an average of 1.85 full-document reads per query, confirming that full-document access provides additional useful context that enhances decision-making. For Qwen3-32B, which performs worse overall, the benefit is

486 more modest. Accuracy improves slightly from 10.36% to 11.69%, and the number of get-document
 487 calls remains low (0.27 per query on average). This suggests that while the tool can help, the model’s
 488 limited reasoning and tool-use ability constrain its ability to exploit the additional information.
 489

490 5 CONCLUSION

491
 492 We introduced `BROWSECOMP-PLUS`, a new benchmark designed to address the reproducibility,
 493 fairness, and transparency challenges in evaluating deep search agents. By grounding each query in a
 494 fixed, human-verified corpus containing both positive and hard-negative documents, our framework
 495 enables the independent and controlled assessment of retrieval and agent components.
 496

497 Through extensive experiments, we demonstrate that retrieval quality substantially impacts both the
 498 effectiveness and efficiency of deep search agents. `BROWSECOMP-PLUS` provides a robust platform
 499 for probing these dynamics and paves the way for future research on co-optimizing retrievers and
 500 agents, improving out-of-distribution tool-use generalization, and advancing context engineering
 501 frameworks. By making our benchmark and baselines publicly available, we aim to catalyze the next
 502 generation of deep search agents.
 503

504 REPRODUCIBILITY STATEMENT

505
 506 The primary motivation for constructing `BROWSECOMP-PLUS` is to enable fair and reproducible
 507 comparisons of deep search agents. To this end, we will release the constructed corpus, pre-built
 508 retrieval indexes, and one-click reproducible code for evaluating all combinations of deep search
 509 agents and retrievers presented in this work. In addition, we plan to open source the full execution
 510 traces of our experiments, since some baselines are expensive to reproduce (e.g., running Opus 4 on
 511 the 830 `BROWSECOMP-PLUS` queries can incur approximately USD \$2,000). By releasing these
 512 traces, we hope to help lower barriers for future researchers and support more efficient development
 513 of deep search agents.
 514

515 ETHICS STATEMENT

516
 517 The `BROWSECOMP-PLUS` dataset extends OpenAI’s `BrowseComp`, which is released under the MIT
 518 license. The augmented corpus was obtained by scraping documents from publicly accessible web
 519 sources searched via a Google API provider. As the data is drawn solely from open web content, we
 520 assess the ethical and legal risks to be minimal.
 521

522 LIMITATION

523
 524 `BROWSECOMP-PLUS` has several limitations that we acknowledge and hope future work can address.
 525 First, although the corpus is constructed through careful human verification, it is difficult to guarantee
 526 the absence of false negatives, relevant documents that are missing from the corpus. This limitation is
 527 inherent to all large-scale retrieval benchmarks, but it may still create a gap between our benchmark
 528 and ideal evaluation. Second, the initial evidence-gathering step uses an OpenAI model (o3) to
 529 propose candidate URLs, which may introduce bias toward distributions that are more easily surfaced
 530 by that model; although humans subsequently edited or replaced many documents, this potential
 531 bias should be noted. Third, `BROWSECOMP-PLUS` primarily evaluates textual evidence and does
 532 not fully capture the diversity of real-world web content, such as interactive pages, dynamic layouts,
 533 multimedia, or unparsed PDFs. Finally, in this work we focus on evaluation based on short, concluded
 534 answers and cited documents within long-form responses. Comprehensive evaluation of generated
 535 reports for complex, deep-search agent tasks remains an open direction for future work.
 536

537 REFERENCES

538
 539 Anthropic Team. Introducing the model context protocol. November 2024a. URL <https://www.anthropic.com/news/model-context-protocol>.

540 Anthropic Team. The claude 3 model family: Opus, sonnet, haiku. 2024b. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

541

542

543

544 Akari Asai, Ziqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning
545 to retrieve, generate, and critique through self-reflection. In *The Twelfth International Confer-
546 ence on Learning Representations*, 2024. URL <https://openreview.net/forum?id=hSyW5go0v8>.

547

548 Adrien Barbaresi. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery
549 and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the
550 Association for Computational Linguistics and the 11th International Joint Conference on Natural
551 Language Processing: System Demonstrations*, pp. 122–131. Association for Computational
552 Linguistics, 2021. URL <https://aclanthology.org/2021.acl-demo.15>.

553

554 Antoine Chaffin and Raphaël Sourty. Pylate: Flexible training and retrieval for late interaction models,
555 2024. URL <https://github.com/lightonai/pylate>.

556

557 Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding:
558 Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge
559 distillation, 2024.

560

561 Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas
562 Hartvigsen, Jack Gallifant, and Danielle S. Bitterman. Medbrowsecomp: Benchmarking medical
563 deep research and computer use. *arXiv:2505.14963*, 2025. URL <https://arxiv.org/abs/2505.14963>.

564

565 Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel
566 Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv:2401.08281*,
567 2024.

568

569 Maik Fröbe, Andrew Parry, Harrisen Scells, Shuai Wang, Shengyao Zhuang, Guido Zuccon,
570 Martin Potthast, and Matthias Hagen. Corpus subsampling: Estimating the effectiveness of
571 neural retrieval models on large corpora. In *Advances in Information Retrieval: 47th Eu-
572 ropean Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025,
573 Proceedings, Part I*, pp. 453–471, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-
031-88707-9. doi: 10.1007/978-3-031-88708-6_29. URL https://doi.org/10.1007/978-3-031-88708-6_29.

574

575 Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense pas-
576 sage retrieval. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings
577 of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Pa-
578 pers)*, pp. 2843–2853, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:
579 10.18653/v1/2022.acl-long.203. URL <https://aclanthology.org/2022.acl-long.203>.

580

581 Gemini 2.5 Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long
582 context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.

583

584 Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Botao Yu, Andrei Kopanev, Weijian
585 Qi, Yiheng Shu, Jiaman Wu, Chan Hee Song, Bernal Jimenez Gutierrez, Yifei Li, Zeyi Liao,
586 Hanane Nour Moussa, TIANSHU ZHANG, Jian Xie, Tianci Xue, Shijie Chen, Boyuan Zheng,
587 Kai Zhang, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2web
588 2: Evaluating agentic search with agent-as-a-judge. In *The Thirty-ninth Annual Conference on
589 Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL <https://openreview.net/forum?id=AUaW6DS9si>.

590

591 Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas
592 Koukounas, Mohammad Kalim Wang, Nan Wang, and Han Xiao. Jina-ColBERT-v2: A general-
593 purpose multilingual late interaction retriever. In Jonne Sälevä and Abraham Owodunni (eds.),
Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024), pp.

594 159–166, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi:
 595 10.18653/v1/2024.mrl-1.11. URL <https://aclanthology.org/2024.mrl-1.11/>.
 596

597 Bowen Jin, Jinsung Yoon, Priyanka Kargupta, Sercan O. Arik, and Jiawei Han. An empirical study
 598 on reinforcement learning for reasoning-search interleaved llm agents, 2025a. URL <https://arxiv.org/abs/2505.15117>.
 599

600 Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and
 601 Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement
 602 learning. *arXiv:2503.09516*, 2025b. URL <https://arxiv.org/abs/2503.09516>.
 603

604 Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale
 605 distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and
 606 Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Com-
 607 putational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July
 608 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147/>.
 609

610 Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi
 611 Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie
 612 Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on
 613 Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November
 614 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL
 615 [https://aclanthology.org/2020.emnlp-main.550/](https://aclanthology.org/2020.emnlp-main.550).
 616

617 Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris
 618 Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion
 619 Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav
 620 Petrov. Natural questions: A benchmark for question answering research. *Transactions of the
 621 Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL
 622 <https://aclanthology.org/Q19-1026/>.
 623

624 Jinyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael
 625 Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez
 626 Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda,
 627 Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large
 628 language models. *arXiv:2403.20327*, 2024. URL <https://arxiv.org/abs/2403.20327>.
 629

629 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
 630 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela.
 631 Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th
 632 International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY,
 633 USA, 2020. Curran Associates Inc. ISBN 9781713829546.
 634

634 Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan
 635 Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong
 636 Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating super-human
 637 reasoning for web agent. *arXiv:2507.02592*, 2025. URL <https://arxiv.org/abs/2507.02592>.
 638

639 Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards
 640 general text embeddings with multi-stage contrastive learning. *arXiv:2308.03281*, 2023. URL
 641 <https://arxiv.org/abs/2308.03281>.
 642

643 Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo
 644 Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse
 645 and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on
 646 Research and Development in Information Retrieval*, SIGIR '21, pp. 2356–2362, New York, NY,
 647 USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.
 3463238. URL <https://doi.org/10.1145/3404835.3463238>.
 648

648 Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Ro-
 649 driguez, Jacob Kahn, Gergely Szilvassy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. RA-DIT:
 650 Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning*
 651 *Representations*, 2024. URL <https://openreview.net/forum?id=220Tbutug9>.

652 Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou.
 653 ReasonRank: Empowering passage ranking with strong reasoning ability. *arXiv preprint*
 654 *arXiv:2508.07050*, 2025.

655 Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-shot listwise document reranking
 656 with a large language model. *arXiv:2305.02156*, 2023. URL <https://arxiv.org/abs/2305.02156>.

657 Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage
 658 text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and*
 659 *Development in Information Retrieval, SIGIR '24*, pp. 2421–2425, New York, NY, USA, 2024.
 660 Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657951.
 661 URL <https://doi.org/10.1145/3626772.3657951>.

662 Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy
 663 Lin. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality.
 664 SIGIR '25, pp. 4061–4065, New York, NY, USA, 2025a. Association for Computing Machinery.
 665 ISBN 9798400715921. doi: 10.1145/3726302.3730135. URL <https://doi.org/10.1145/3726302.3730135>.

666 Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, and Xilun Chen. DRAMA:
 667 Diverse augmentation from large language models to smaller dense retrievers. In Wanxiang Che,
 668 Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd*
 669 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.
 670 30170–30186, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN
 671 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1457. URL <https://aclanthology.org/2025.acl-long.1457/>.

672 Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embed-
 673 ding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th*
 674 *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–
 675 2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/
 676 v1/2023.eacl-main.148. URL <https://aclanthology.org/2023.eacl-main.148/>.

677 OpenAI Team. OpenAI o3 and o4-mini system card. 2025a. URL <https://cdn.openai.com/o3-mini-system-card-feb10.pdf>.

678 OpenAI Team. GPT-OSS-120B & 20B model card. 2025b. URL https://cdn.openai.com/pdf/419b6906-9da6-406c-a19d-1bb078ac7637/oai_gpt-oss_model_card.pdf.

679 Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin
 680 Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for
 681 the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing*
 682 *Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=n6SCkn2QaG>.

683 Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin
 684 Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra,
 685 Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry
 686 Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes,
 687 Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav
 688 Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou,
 689 Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon,
 690 Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala,
 691 Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney,
 692 Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng,

702 Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben
 703 McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui
 704 Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G.
 705 Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward
 706 Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow,
 707 Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis
 708 Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod,
 709 Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang,
 710 Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin
 711 Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel
 712 Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad
 713 Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov,
 714 Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den
 715 Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust,
 716 Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong
 717 Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu,
 718 Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner,
 719 James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene,
 720 Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan,
 721 Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova,
 722 Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster,
 723 Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes,
 724 Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider,
 725 Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser,
 726 Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam
 727 Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim
 728 Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef
 729 Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Mont-
 730 gomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan
 731 Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz,
 732 Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K
 733 Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng,
 734 Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger,
 735 Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev,
 736 Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis,
 737 Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu,
 738 Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hen-
 739 drik Kirchner, Maksym Ovchinnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou
 740 Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani,
 741 Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémie Andréeletti, Raghav Singhal, Jacob
 742 Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson,
 743 Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu,
 744 Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William
 745 Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob
 746 Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio,
 747 Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah
 748 Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam
 749 Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad,
 750 Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang
 751 Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang,
 752 Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan,
 753 Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep,
 754 Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching,
 755 Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank
 Agnihotri, Pavel Zhelnov, Mohammadreza Mofayzezi, Alexander Piperski, David K. Zhang, Kos-
 756 tantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho,
 757 Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah
 758 Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta,
 759 Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidamaram, Timothy Manik, Hector

756 Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor
 757 Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez,
 758 Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik
 759 Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh,
 760 Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda,
 761 Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike
 762 Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon
 763 Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld,
 764 Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mar-
 765 iana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah
 766 Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie
 767 Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin
 768 Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo
 769 Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir
 770 Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat,
 771 Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan
 772 Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau,
 773 Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna
 774 Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg
 775 Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie,
 776 Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu,
 777 Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor
 778 Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding
 779 Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámarra,
 780 Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben
 781 Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner,
 782 Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica
 783 Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk
 784 Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc,
 785 Tania C. B. Santos, Samir Shamseldeen, Loukmene Karim, Anna Liakhovitskaia, Nate Resman,
 786 Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira
 787 Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail
 788 Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan
 789 Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel
 790 Ramirez-Trinidad, Yana Malyshева, Daphny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel
 791 Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy
 792 Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric
 793 Vergo, Juehang Qin, Benjamín Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis,
 794 Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Duc Huy,
 795 Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber,
 796 Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus
 797 Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin
 798 Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong,
 799 Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh,
 800 Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang
 801 Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim
 802 Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will
 803 Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya
 804 Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gun-
 805 jan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky,
 806 Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright,
 807 Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Re-
 808 han Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze,
 809 Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin,
 Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahalooohoreh,
 Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro,
 Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini
 Jain, Ashley Aaron, Murat Tirayioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin,
 Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph

810 McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert,
 811 Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved,
 812 Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar,
 813 Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu,
 814 Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok,
 815 Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis,
 816 Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata,
 817 Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang,
 818 Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel
 819 Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser,
 820 Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng
 821 Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina
 822 Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt,
 823 Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong,
 824 Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang,
 825 Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang,
 826 Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao
 827 Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo,
 828 Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Sha-
 829 bani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven,
 830 Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao
 831 Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos,
 832 Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen
 833 Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee,
 834 Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamoun, Shuyu Wu,
 835 Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi,
 836 Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera
 837 Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E.
 838 Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee,
 839 Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park,
 840 Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhannany, Han Lin, Philipp D. Siedler, Jichao
 841 Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar
 842 Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin,
 843 Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster,
 844 Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz,
 845 Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej
 846 Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal,
 847 Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song,
 848 Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak,
 849 Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha,
 850 Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung,
 851 Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis
 852 Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin
 853 Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan
 854 Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain,
 855 Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan,
 856 Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo
 857 Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang
 858 Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash
 859 Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason
 860 Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith
 861 Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert
 862 Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin
 863 Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy
 864 Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi
 865 Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti
 866 Sundarapandiyar, Ashley Zhang, Andrew Le, Zafir Nasim, Srikanth Yalam, Ritesh Kasamsetty,
 867 Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen,
 868 Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer

864 Yue, Alexandr Wang, and Dan Hendrycks. Humanity’s last exam. *arXiv:2501.14249*, 2025. URL
 865 <https://arxiv.org/abs/2501.14249>.

866

867 Stephen E. Robertson. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference*
 868 (*TREC-3*). NIST, 1994.

869 Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford.
 870 Okapi at TREC-2. In Donna K. Harman (ed.), *Proceedings of The Second Text REtrieval Confer-*
 871 *ence, TREC 1993, Gaithersburg, Maryland, USA, August 31 - September 2, 1993*, volume 500-215
 872 of *NIST Special Publication*, pp. 21–34. National Institute of Standards and Technology (NIST),
 873 1993. URL <http://trec.nist.gov/pubs/trec2/papers/ps/city.ps>.

874

875 Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Col-
 876 BERTv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat,
 877 Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022*
 878 *Conference of the North American Chapter of the Association for Computational Linguistics: Human*
 879 *Language Technologies*, pp. 3715–3734, Seattle, United States, July 2022. Association for Compu-
 880 tational Linguistics. doi: 10.18653/v1/2022.naacl-main.272. URL [https://aclanthology.org/2022.naacl-main.272/](https://aclanthology.org/2022.naacl-main.272).

881

882 Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan
 883 Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. Reasonir:
 884 Training retrievers for reasoning tasks. *arXiv:2504.20595*, 2025. URL <https://arxiv.org/abs/2504.20595>.

885

886 Sahel Sharifmoghaddam, Ronak Pradeep, Andre Slavescu, Ryan Nguyen, Andrew Xu, Zijian Chen,
 887 Yilin Zhang, Yidi Chen, Jasper Xian, and Jimmy Lin. RankLLM: A python package for reranking
 888 with llms. In *Proceedings of the 48th International ACM SIGIR Conference on Research and*
 889 *Development in Information Retrieval, SIGIR ’25*, pp. 3681–3690, New York, NY, USA, 2025.
 890 Association for Computing Machinery. ISBN 9798400715921.

891

892 Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu
 893 Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik,
 894 Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive
 895 retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025. URL
 896 <https://openreview.net/forum?id=ykuc5q381b>.

897

898 Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin,
 899 and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking
 900 agents. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference*
 901 *on Empirical Methods in Natural Language Processing*, pp. 14918–14937, Singapore, December
 902 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923. URL
 903 <https://aclanthology.org/2023.emnlp-main.923/>.

904

905 Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin
 906 Braschler, Julio Gonzalo, and Michael Kluck (eds.), *Evaluation of Cross-Language Information*
 907 *Retrieval Systems*, pp. 355–370, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN
 978-3-540-45691-9.

908

909 Ellen M. Voorhees. *The Evolution of Cranfield*, pp. 45–69. Springer International Publishing,
 910 Cham, 2019. ISBN 978-3-030-22948-1. doi: 10.1007/978-3-030-22948-1_2. URL https://doi.org/10.1007/978-3-030-22948-1_2.

911

912 Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving
 913 text embeddings with large language models. *arXiv:2401.00368*, 2023.

914

915 Shuai Wang, Ekaterina Khramtsova, Shengyao Zhuang, and Guido Zuccon. Feb4rag: Evaluating
 916 federated search in the context of retrieval augmented generation. In *Proceedings of the 47th*
 917 *International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, pp. 763–773, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657853. URL <https://doi.org/10.1145/3626772.3657853>.

918 Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won
 919 Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsec comp: A simple yet
 920 challenging benchmark for browsing agents. *arXiv:2504.12516*, 2025. URL <https://arxiv.org/abs/2504.12516>.

921
 922 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,
 923 Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge,
 924 Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi
 925 Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao
 926 Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize
 927 Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu
 928 Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan,
 929 Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3
 930 technical report. *arXiv:2505.09388*, 2025. URL <https://arxiv.org/abs/2505.09388>.

931
 932 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and
 933 Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answer-
 934 ing. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceedings
 935 of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380,
 936 Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi:
 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259/>.

937
 938 Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie,
 939 An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advanc-
 940 ing text embedding and reranking through foundation models. *arXiv:2506.05176*, 2025. URL
 941 <https://arxiv.org/abs/2506.05176>.

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

946 Shengyao Zhuang and Guido Zuccon. Asyncval: A toolkit for asynchronously validating dense
 947 retriever checkpoints during training. In *Proceedings of the 45th International ACM SIGIR
 948 Conference on Research and Development in Information Retrieval*, SIGIR ’22, pp. 3235–3239,
 949 New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi:
 950 10.1145/3477495.3531658. URL <https://doi.org/10.1145/3477495.3531658>.

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972 A OPENAI o3 EVIDENCE DOCUMENT GATHERING PROMPT

975 I will give you a question and a correct answer, and you are to search online for evidence
 976 that supports the answer. List the evidence you've used to justify this answer step-by-step,
 977 including their urls in your output. Your final list of urls should be in the order such that a
 978 human can visit them in order to justify the answer.

979 Question: {question}

980 Answer: {answer}

983 This is all the information you have to work with to produce the final list of urls. Format your
 984 answer in a table with 3 columns:

985 - clue: the clue mentioned in the question
 986 - url: the http web url of the evidence you've found
 987 - evidence: the content in the url page that supports the clue

990 B LABELLING UI EXAMPLE

994 Question:
 995 Please identify the fictional character who occasionally breaks the fourth wall with the audience, has a backstory involving help from selfless ascetics, is known for his humor, and had a TV show that aired between the 1960s and 1980s with fewer than 50 episodes.

996 Answer:
 997 Plastic Man

998 Evidence/Clues: [Add Clue](#)

Clue 1	(Matched ✓)	Edit Delete Reset Decline No Match
Breaks the fourth wall		
Likely from doc 1		
Plastic Man's "Powers and Abilities" list explicitly includes "Breaking the Fourth Wall" among his skills, confirming he sometimes addresses the audience directly. character-level.fandom.com		
Linked to: Doc 1: "Breaking the Fourth Wall."		
Clue 2	(Matched ✓)	Edit Delete Reset Decline No Match
Nursed by selfless ascetics (monks) in his origin		
Likely from doc 2		

999 Documents: [Add Document](#)

Document 1	Edit Delete Highlight Mark Err
https://character-level.fandom.com/wiki/Plastic_Man_%28Post-Crisis%29	
Gender: Male	
Age: Unknown, At least 90+ years	
Classification: Human, Mutate, Former Criminal, Superhero	
Powers and Abilities: Superhuman Physical Characteristics, Elasticity, Toon Force, Shapeshifting, Camouflage, Stealth Mastery, Voice Mimicry, Size Manipulation, Body Control, Breaking the Fourth Wall , Immortality (Types 1, 2 and 3), Regeneration (High, regenerated from mere molecules, although it required someone to collect at least 80% of his body mass) and Ultrasonic Detection. Immune to Mind Manipulation, Transmutation and Telepathy. Resistance to Acid, Blunt Attacks, Piercing Attacks, Energy Projection, and Magic.	
Attack Potency: Solar System level (Could trade blows with a bloodied Fernus)	

1000 Document 2

1001 <https://www.britannica.com/topic/Plastic-Man>

1002 madcap genius of his creator, Jack Cole. Cole had led a colourful life, including cycling across America at the age of 18, before moving to New York in 1935 and dedicating himself to his true passion of cartooning. After a fitful start as a gag cartoonist, he found himself in at the beginning of the nascent comics explosion, working for Centaur Publishing and Lev Gleason Publications before joining Quality Comics. In mid-1941, owner Everett "Bury" Arnold asked Cole to create a new hero for Quality's upcoming new Police Comics title—something in the tradition of Will Eisner's Spirit. Cole responded with his own sort of super-detective, a hero who always got his man in his own way: Plastic Man.

1003 In August 1941, the first issue of Police Comics introduced a hoodlum called Eli O'Brian, hard at work cracking a safe at the Crawford Chemical Works. Disturbed by a guard, O'Brian and his gang flee the building, but a stray bullet hits a large chemical vat, showering the thief with acid. Injured and desperate, O'Brian runs for miles before reaching a mountain retreat called Rest-Haven, where [he is tended to by kind monks who shield him from the police](#). Inspired by their trust in him, he decides to turn over a new leaf and vows to change his ways. Only then does he discover that the acid has affected his body in such a way that he can now stretch it into any shape he can think of. Thrilled by that discovery ("Great guns! I'm stretchin' like a rubber-band!"), he dons a red bodysuit, trimmed with a yellow belt and topped off with wraparound sunglasses, and begins his new life's work as a crime fighter.

1004 The evidence above suffices to fully derive the answer from scratch?

1005 True False

1006 Which documents contain the final answer "Plastic Man"? (Select all that apply)

1007 Document 1

1008 Document 2

1009 Document 3

1010 Document 4

1011 Document 5

1012 Please verify docs 1, 2, 3, 4, 5 contain the final answer.

1017 Figure 3: A screenshot of the annotation interface.

1026 **C PROBLEMATIC CASES**
1027

1028

- 1029 • **BrowseComp Errors:** During the verification process, we discover that some question-answer
1030 pairs in BrowseComp are inherently flawed. For example, one question asks for the name of
1031 a book whose author later returned to acting. Using the ground-truth answer, we can identify
1032 the intended book and its listed author. However, upon further investigation, we find that the
1033 individual who wrote the book and the one who returned to acting are two different people who
1034 happen to share the same name.
- 1035 • **Extensive Use of Google Maps:** 42 queries in BrowseComp require distance-related information
1036 that explicitly prompt multiple calls to Google Maps. These are removed because high-quality
1037 documents discussing specific Google Maps distances between arbitrary locations are difficult to
1038 obtain. Moreover, scraping static snapshots of Google Maps pages to include in the corpus is not
1039 a valid substitute; answering such questions as intended should require agents to be augmented
1040 with access to the Google Maps API, rather retrieving from a corpus. However, this capability
1041 lies outside the scope of our objective to build a static, document-based dataset.
- 1042 • **Ambiguous or Non-Unique Answers:** Some question-answer pairs are well-supported by doc-
1043 uments, but suffer from ambiguity in the expected answer format or the existence of multiple
1044 valid answers. For instance, one question asks for the username of an individual who authored a
1045 specific story on an internet forum. While the ground-truth answer is correct, it is only one of
1046 three usernames credited as authors. We remove 13 such queries due to this kind of ambiguity.

1047

1048 **D NEGATIVE MINING QUERY DECOMPOSITION PROMPT**
1049

1050

1051 You are an expert at breaking down complex, multi-part questions into simpler, self-contained
1052 subqueries.
1053

1054 Your task is to analyze the given question and decompose it into a series of smaller, more
1055 manageable subqueries that, when answered together, would provide all the information
1056 needed to answer the original question.
1057

1058 Guidelines:
1059

- 1060 1. Each subquery should focus on a single piece of information or concept
- 1061 2. Subqueries **MUST** be completely self-contained and answerable independently - do not
1062 use pronouns or references like "this person", "the author", "these conditions", "they", "the
1063 movie", etc.
- 1064 3. Each subquery should include all necessary context and constraints from the original query
- 1065 4. Preserve all important details and constraints from the original query
- 1066 5. Return only the subqueries as a JSON array of strings

1067 Example:
1068

1069 Original: "Please identify the fictional character who occasionally breaks the fourth wall
1070 with the audience, has a backstory involving help from selfless ascetics, is known for his hu-
1071 mor, and had a TV show that aired between the 1960s and 1980s with fewer than 50 episodes."
1072

1073 Subqueries: ["Which fictional characters occasionally break the fourth wall with the
1074 audience?", "Which fictional characters have a backstory involving help from selfless
1075 ascetics?", "Which fictional characters are known for their humor?", "Which TV shows aired
1076 between the 1960s and 1980s?", "Which TV shows had fewer than 50 episodes?]

1077 Please decompose this query into subqueries:
1078 {query}
1079

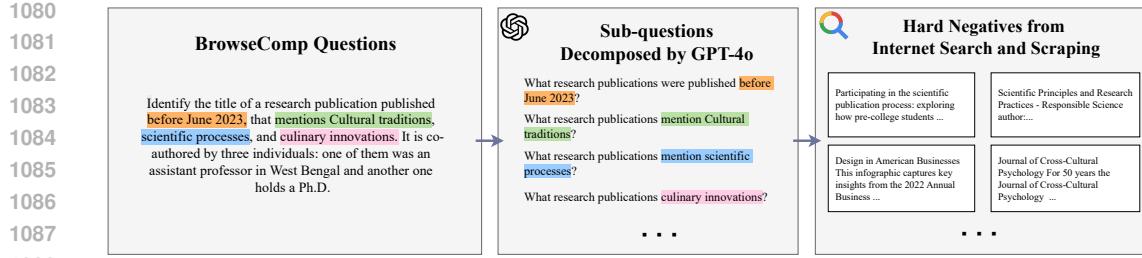


Figure 4: The pipeline of collecting hard negative documents in Section 3.3.

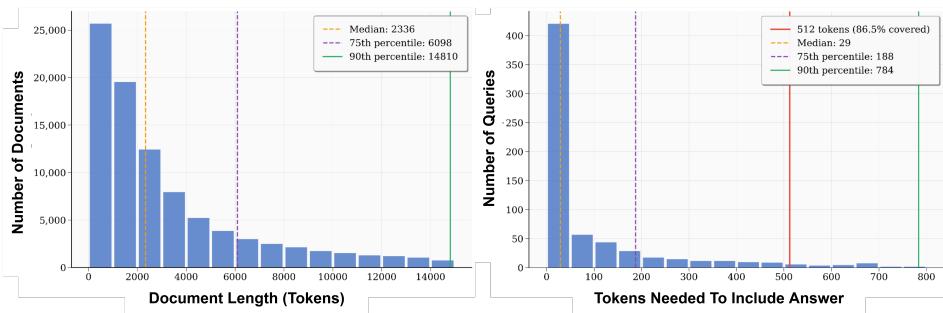


Figure 5: (a) Token distribution of corpus length, showing up to 90th percentile for display; (b) Distribution of tokens needed to include answer in gold documents per query, showing up to 90th percentile for display

E MAIN SEARCH PROMPT

You are a deep research agent. You need to answer the given question by interacting with a search engine, using the search tool provided. Please perform reasoning and use the tool step by step, in an interleaved manner. You may use the search tool multiple times.

Question: {Question}

Your response should be in the following format:

Explanation: {{your explanation for your final answer. For this explanation section only, you should cite your evidence documents inline by enclosing their docids in square brackets [] at the end of sentences. For example, [20].}}

Exact Answer: {{your succinct, final answer}}

Confidence: {{your confidence score between 0% and 100% for your answer}}

1134 F EVALUATION PROMPT
1135

1136

1137 Judge whether the following [response] to [question] is correct or not based on the precise
1138 and unambiguous [correct_answer] below.

1139

1140 [question]: {question}

1141

1142 [response]: {response}

1143

1144 Your judgement must be in the format and criteria specified below:

1145

1146 extracted_final_answer: The final exact answer extracted from the [response]. Put the
1147 extracted answer as ‘None’ if there is no exact, final answer to extract from the response.

1148

1149 [correct_answer]: {correct_answer}

1150

1151 reasoning: Explain why the extracted_final_answer is correct or incorrect based on [cor-
1152 rect_answer], focusing only on if there are meaningful differences between [correct_answer]
1153 and the extracted_final_answer. Do not comment on any background to the problem, do not
1154 attempt to solve the problem, do not argue for any answer different than [correct_answer],
1155 focus only on whether the answers match.

1156

1157 correct: Answer ‘yes’ if extracted_final_answer matches the [correct_answer] given above,
1158 or is within a small margin of error for numerical problems. Answer ‘no’ otherwise, i.e. if
1159 there if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is
1160 incorrect.

1161

1162 confidence: The extracted confidence score between 0% and 100% from [response]. Put
1163 100 if there is no confidence score available.

1164

1165 G SEARCH PROMPT WITH GET-DOC
1166

1167

1168 You are a deep research agent. You need to answer the given question by interacting with a
1169 search engine, using the search and get_document tools provided. Please perform reasoning
1170 and use the tools step by step, in an interleaved manner. You may use the search and
1171 get_document tools multiple times.

1172

1173 Question: {Question}

1174

1175 Your response should be in the following format:

1176

1177 Explanation: {{your explanation for your final answer. For this explanation section only, you
1178 should cite your evidence documents inline by enclosing their docids in square brackets [] at
1179 the end of sentences. For example, [20].}}
1180 Exact Answer: {{your succinct, final answer}}
1181 Confidence: {{your confidence score between 0% and 100% for your answer}}

1182

1183 H BASELINES
11841185 H.1 LLM SEARCH AGENTS
11861187 We evaluate several representative commercial models with strong agentic search capabilities, ranging
1188 from the most advanced reasoning models to cost-effective ones: GPT-5, o3, GPT-4.1 (OpenAI
1189 Team, 2025a), claude-opus-4, claude-sonnet-4 (Anthropic Team, 2024b), gemini-2.5-pro, gemini-2.5-
1190 flash (Gemini 2.5 Team, 2025).

1188 We also assess leading open-source efforts. This includes Qwen3-32B (Yang et al., 2025), a popular
 1189 open-source reasoning LLM, and Search-R1 (Jin et al., 2025b;a), a model fine-tuned for agentic
 1190 search based on the Qwen backbone. Specifically, we use the 32B checkpoint released in (Jin et al.,
 1191 2025a). Finally, we evaluate the recent advanced gpt-oss-120B (OpenAI Team, 2025b), a reasoning
 1192 LLM optimized for search tool usage that offers multiple reasoning effort settings, ranging from low
 1193 to high.

1194 H.2 RETRIEVERS

1195 In our study, we compared a range of retrieval methods from a traditional lexical baseline to modern
 1196 state-of-the-art dense embedding retrievers:

- 1200 • BM25 (Robertson et al., 1993): The classic sparse lexical retriever, which matches queries to
 1201 documents based on term statistics.
- 1202 • Qwen3-Embedding (Zhang et al., 2025): A dense embedding retriever, available in sizes 0.6B,
 1203 4B, and 8B, built on the Qwen3 foundation model family (Yang et al., 2025). It achieves
 1204 state-of-the-art performance on retrieval benchmarks such as MTEB (Muennighoff et al., 2023).
- 1205 • ReasonIR (Shao et al., 2025): A dense embedding specifically trained for reasoning-intensive
 1206 retrieval via synthetic data generation, setting a new state-of-the-art on reasoning-intensive
 1207 information retrieval benchmark BRIGHT (SU et al., 2025).
- 1208 • Jina-ColBERT-v2 (Jha et al., 2024): A late-interaction retriever that trains ColBERTv2 (San-
 1209 thanam et al., 2022) from a newer BERT backbone to support much longer contexts.

1211 We use the Pyserini IR toolkit (Lin et al., 2021) to serve the BM25 retriever, the Tevatron dense re-
 1212 trieval toolkit (Ma et al., 2025a) to serve Qwen3-Embedding and ReasonIR, along with PyLate (Chaf-
 1213 fin & Sourty, 2024) to serve Jina-ColBERT-v2.

1214 I CITATION QUALITY

1217 Table 5 reports the coverage, average number, precision, and recall of the document citations attributed
 1218 by the agent during answer generation. As the results show, although agents using BM25 issue
 1219 more search calls, nearly all metrics are lower than those achieved with Qwen3-Embedding-8B. This
 1220 indicates that documents returned by BM25 are less useful in the iterative deep research process,
 1221 whereas Qwen3-Embedding-8B provides more relevant and informative documents.

1224 J EFFECT OF RERANKING

1226 To evaluate the impact of reranking, we apply listwise reranking (Sun et al., 2023; Ma et al., 2023)
 1227 over the top-20 and top-100 retrieved candidates using RankLLM (Sharifmoghaddam et al., 2025)
 1228 with Qwen3-8B/32B and ReasonRank-7B/32B (Liu et al., 2025) models. The reranker operates with
 1229 a sliding window of 20 candidates and a stride of 10, using a 16k-token context and a 16k-token
 1230 thinking budget (output token count) to balance coverage and compute. Longer candidates are
 1231 truncated to fit within the context window as needed.

1232 Table 6 reports the effect of reranking after first-stage retrieval with Qwen3-Embed-8B, in the retrieval-
 1233 only setting. For like-sized models, Qwen3 and ReasonRank perform similarly, with differences
 1234 typically within 1 point. Overall, reranking yields sizable gains, improving Recall@5 by 8.4–24.0
 1235 points. With top-20 reranking, model size matters little (only ~2–3 points difference). Expanding the
 1236 reranking candidate set to 100 improves all models, with larger gains for the 32B models, thereby
 1237 widening the effectiveness gap between 8B and 32B models at higher rerank depths.

1238 Table 7 reports the effect of integrating reranking into end-to-end performance of two search agents,
 1239 GPT-4.1 and gpt-oss-20B (high reasoning effort), using Qwen3-Embed-8B as the first-stage retriever
 1240 and Qwen3-8B to rerank the top 20 candidates. For both models, Accuracy (judged by GPT-4.1) and
 1241 Recall improve substantially. This further indicates that reranking improves the precision and recall
 of retrieved evidence at higher ranks, helping the agent surface more relevant information.

1242 Table 5: Per-query averages of citation coverage, citation count, precision, and recall for labeled
 1243 evidence documents. Search-R1 is excluded because its fine-tuned outputs do not contain citations.
 1244

1245 LLM	1246 Retriever	1247 Coverage	1248 Avg # Citations	1249 Precision	1250 Recall
1247 GPT-4.1	BM25	57.0%	1.92	37.0%	16.1%
	Qwen3-Embedding-8B	79.2%	2.54	58.5%	28.2%
1249 o3	BM25	63.5%	3.27	86.7%	51.0%
	Qwen3-Embedding-8B	78.0%	3.51	91.8%	56.2%
1251 GPT-5	BM25	94.9%	3.89	71.8%	51.3%
	Qwen3-Embedding-8B	98.0%	4.28	83.4%	62.3%
1253 Sonnet 4	BM25	76.1%	3.19	31.9%	21.3%
	Qwen3-Embedding-8B	90.7%	4.19	52.4%	39.9%
1255 Opus 4	BM25	74.9%	3.03	35.1%	22.3%
	Qwen3-Embedding-8B	86.1%	3.82	58.9%	42.6%
1257 Gemini 2.5 Flash	BM25	74.2%	4.89	34.2%	21.7%
	Qwen3-Embedding-8B	89.2%	4.75	51.5%	35.1%
1259 Gemini 2.5 Pro	BM25	53.9%	3.03	52.1%	31.4%
	Qwen3-Embedding-8B	59.4%	3.49	64.9%	41.5%
1261 gpt-oss-120B-high	BM25	62.5%	3.55	50.8%	31.5%
	Qwen3-Embedding-8B	76.9%	3.88	60.8%	38.2%
1264 Qwen3-32B	BM25	87.0%	1.85	8.9%	2.6%
	Qwen3-Embedding-0.6B	90.1%	1.79	8.7%	2.5%
	Qwen3-Embedding-4B	91.7%	1.84	16.1%	4.9%
	Qwen3-Embedding-8B	90.2%	1.78	20.0%	6.6%
	ReasonIR	95.8%	1.74	18.0%	5.7%

1268 Table 6: Effectiveness of rerankers with Qwen3-Embed-8B in retriever-only evaluation. The full
 1269 question is used as the query in both stages. Reranking is applied to the top-20 and top-100 candidates.
 1270 Scores in parentheses denote improvements over the base retriever (Δ vs. first stage).
 1271

1272 Reranker	1273 Top-20		1274 Top-100	
	1275 Recall@5 (Δ)	1276 nDCG@10 (Δ)	1277 Recall@5 (Δ)	1278 nDCG@10 (Δ)
1275 Qwen3-Embed-8B	14.5 (-)	20.3 (-)	14.5 (-)	20.3 (-)
Evidence Document Retrieval				
1277 ReasonRank-7B	22.9 (+8.4)	29.5 (+9.2)	29.5 (+15.0)	38.0 (+17.7)
1278 Qwen3-8B	23.3 (+8.8)	30.0 (+9.7)	29.6 (+15.1)	37.7 (+17.4)
1279 ReasonRank-32B	24.9 (+10.4)	32.1 (+11.8)	34.4 (+19.9)	43.8 (+23.5)
1280 Qwen3-32B	24.7 (+10.2)	31.8 (+11.5)	35.0 (+20.5)	44.3 (+24.0)
Gold Document Retrieval				
1282 ReasonRank-7B	28.7 (+10.2)	28.9 (+9.4)	36.8 (+18.3)	37.1 (+17.6)
1283 Qwen3-8B	29.2 (+10.7)	29.6 (+10.1)	36.7 (+18.2)	36.6 (+17.1)
1284 ReasonRank-32B	30.7 (+12.2)	31.5 (+12.0)	42.5 (+24.0)	43.5 (+24.0)
1285 Qwen3-32B	30.5 (+12.0)	31.3 (+11.8)	42.2 (+23.7)	43.0 (+23.5)

1287 K API COST

1289 Table 8 Shows the API costs of the experiments in Table 1.
 1290

1292 L FUTURE WORK AND DISCUSSION

1294 We believe that our BROWSECOMP-PLUS opens new avenues for advancing research in the Deep-
 1295 Research area. BROWSECOMP-PLUS retains the challenging nature of the original BrowseComp
 while providing a more controlled and transparent experimental setup similar to early pivotal evalua-

1296 Table 7: Effect of reranking on end-to-end agent performance. Qwen3-Embed-8B is used as the
 1297 first-stage retriever and Qwen3-8B is used for reranking top 20 retrieved candidates.
 1298

1299 LLM	1300 Retriever/Reranker	1301 Accuracy	1302 Recall	1303 Search Calls	1304 Calibration Error
1300 GPT-4.1	Qwen3-Embed-8B	35.42%	36.89%	8.67	54.67%
	+Qwen3-8B	47.11%	51.46%	8.77	49.86%
1302 gpt-oss-20B-high	Qwen3-Embed-8B	34.58%	49.29%	23.87	27.81%
	+Qwen3-8B	40.24%	57.98%	21.98	21.47%

1305 Table 8: Overall API costs of proprietary agents for the experiments in Table 1.
 1306

1307 LLM	1308 Retriever	1309 Accuracy	1310 Price (USD)
1309 GPT-4.1	BM25	14.58%	\$106.96
	Qwen3-Embed-8B	35.42%	\$89.81
1311 o3	BM25	49.28%	\$836.35
	Qwen3-Embed-8B	63.49%	\$740.79
1313 GPT-5	BM25	55.90%	\$400.36
	Qwen3-Embed-8B	70.12%	\$360.71
1315 Sonnet 4	BM25	14.34%	\$352.04
	Qwen3-Embed-8B	36.75%	\$325.75
1317 Opus 4	BM25	15.54%	\$2,043.95
	Qwen3-Embed-8B	36.14%	\$1,842.48
1319 Gemini 2.5 Flash	BM25	15.54%	\$47.32
	Qwen3-Embed-8B	33.01%	\$41.29
1321 Gemini 2.5 Pro	BM25	19.04%	\$138.64
	Qwen3-Embed-8B	28.67%	\$99.92

1325 tion benchmarks like Natural Question (NQ) (Kwiatkowski et al., 2019) and HotpotQA (Yang et al.,
 1326 2018). Like how NQ and HotpotQA have facilitated the design, comparison, and of modern neural
 1327 QA systems, we hope that `BROWSECOMP-PLUS` will serve similar roles for Deep-Research agent
 1328 studies. Here, we list some immediate research directions.

1329 While our current work focuses on how different retrievers influence inference performance, a
 1330 promising future direction is to examine the role of the retriever during agent optimization. For
 1331 example, optimizing a search agent may be more challenging when paired with BM25 than with
 1332 a modern embedding-based retriever, simply because BM25 surfaces fewer relevant documents.
 1333 Understanding how retriever quality affects the learning dynamics of an agent remains an open
 1334 question.

1335 Another important extension is to study the agent’s “out-of-distribution” tool-use capabilities. For
 1336 instance, if an agent is optimized using a BM25 search tool, how well does its performance generalize
 1337 when switched to an embedding-based search tool?

1339 A more creative research could be an attempt on a breakdown of the commercial search engine. As
 1340 much as a folktale, a commercial search solution employs tiered, composed, and multi-facet search
 1341 solution. Is the LLM able to orchestrate a set of search tools to perform federated search (Wang et al.,
 1342 2024), or even a sub-agent, to get quality results similar to those from Google?

1343 A further direction is to design retrieval models that are tolerant of, or even adaptive to, a specific
 1344 agent. In the Deep Research setting, the primary consumer of retrieved documents is no longer a
 1345 human, but a tool-augmented LLM agent. This raises the possibility that retrieval models could
 1346 be co-optimized with the agent for achieving overall answer accuracy, rather than developed and
 1347 evaluated in isolation.

1348 Finally, as shown in this work, an oracle retriever capable of surfacing gold or highly relevant
 1349 documents can greatly improve accuracy. Such retrievers may also reduce the number of search
 iterations required, improving the overall efficiency of the research process. Developing high-

1350 Table 9: Evidence document retrieval effectiveness on the Fineweb 10BT corpus.
1351

1352 Retriever	1353 Corpus	1354 Recall@5	1355 Recall@100	1356 Recall@1000	1357 nDCG@10
1352 BM25	1353 Original	1354 1.2	1355 4.7	1356 13.6	1357 1.6
1352 BM25	1353 Original + Fineweb	1354 2.2	1355 8.0	1356 19.4	1357 3.1
1352 Qwen3-Embed-8B	1353 Original	1354 14.5	1355 47.7	1356 76.7	1357 20.3
1352 Qwen3-Embed-8B	1353 Original + Fineweb	1354 11.6	1355 37.6	1356 64.2	1357 16.4
1352 ReasonIR-8B	1353 Original	1354 12.2	1355 43.6	1356 73.9	1357 16.8
1352 ReasonIR-8B	1353 Original + Fineweb	1354 8.6	1355 30.7	1356 56.3	1357 11.8

1360 Table 10: Accuracy of end-to-end search agents on our BROWSECOMP-PLUS original 100k corpus
1361 vs. FineWeb 10BT corpus.
1362

1363 LLM	1364 Retriever	1365 Corpus	1366 Accuracy
1365 SearchR1-32B	1366 BM25	1367 Original	1368 3.86%
	1366 BM25	1367 Original + Fineweb	1368 4.72%
	1367 Qwen3-Embed-8B	1368 Original	1369 10.36%
	1367 Qwen3-Embed-8B	1368 Original + Fineweb	1369 8.33%
1369 Qwen3-32B	1370 BM25	1371 Original	1372 3.49%
	1370 BM25	1371 Original + Fineweb	1372 5.42%
	1371 Qwen3-Embed-8B	1372 Original	1373 10.36%
	1371 Qwen3-Embed-8B	1372 Original + Fineweb	1373 7.11%

1374 precision retrieval systems for reasoning-intensive, complex queries could yield substantial benefits
1375 for real-world applications.

1376 Overall, BROWSECOMP-PLUS serves as an ideal testbed for pursuing these directions, enabling sys-
1377 tematic and fine-grained analyses of agent–retriever interactions within the Deep-Research paradigm.
1378

1379 M EFFECT OF CORPUS SIZE

1380 The corpus in BROWSECOMP-PLUS contains approximately 100K documents. While real-world
1381 agents often operate over much larger, web-scale corpora, we aim to assess whether our designed
1382 corpus size is sufficient to support valid experimental observations. To this end, we augment our
1383 benchmark corpus with the Fineweb-edu (Penedo et al., 2024) document collection (10 billion
1384 tokens),³ deduplicated by URL. This expansion results in a significantly larger corpus of 9,771,311
1385 documents—roughly 10 times larger than the original.

1386 Table 9 shows retrieval performance before and after adding Fineweb documents. For BM25, retrieval
1387 effectiveness improves across all metrics, likely due to better inverse document frequency (IDF)
1388 estimation in the larger corpus, which strengthens BM25’s lexical scoring.

1389 In contrast, neural retrievers (Qwen3-Embedding-8B and ReasonIR-8B) show degraded performance
1390 on the Fineweb-augmented corpus. This drop is theoretically expected: the relative ranking of
1391 documents from the original small corpus remains unchanged, but the newly added Fineweb doc-
1392 uments can now appear in the top ranks. Since these additional documents are unjudged, they are
1393 treated as non-relevant under standard TREC-style evaluation, inevitably lowering measured retrieval
1394 effectiveness.

1395 It is important to note that lower retrieval scores for embedding models on Fineweb do not necessarily
1396 indicate worse final answers, some unjudged, top-ranked Fineweb documents may be “false negatives”
1397 that still provide useful evidence. However, as shown in Table 10, adding Fineweb does not improve
1398 answer accuracy for embedding-based retrievers. For example, Qwen3-32B with Qwen3-Embedding-
1399 8B drops from 10.36% to 7.11% accuracy.

1400
1401
1402
1403 ³[https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu/viewer/
sample-10BT](https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu/viewer/sample-10BT)

Overall, expanding the corpus size by a factor of 10 does not lead to different conclusions about the ranking or effectiveness level among the retrievers and LLM search agents, supporting our claim that the original 100K corpus offers both strong positive coverage and sufficient challenge for robust evaluation.

N USAGE OF LLM

ChatGPT is used during the writing to polish text (e.g., correct grammar) and format tables.

O SIGNIFICANT TEST OF MAIN RESULTS

In Table 6, we present the visualization of the significance test on the answer accuracy of each search agent integrated with different retrievers. The methods are ordered by their accuracy scores. The upper-right triangle indicates which pairwise comparisons reach the significance level of $p \leq 0.05$.

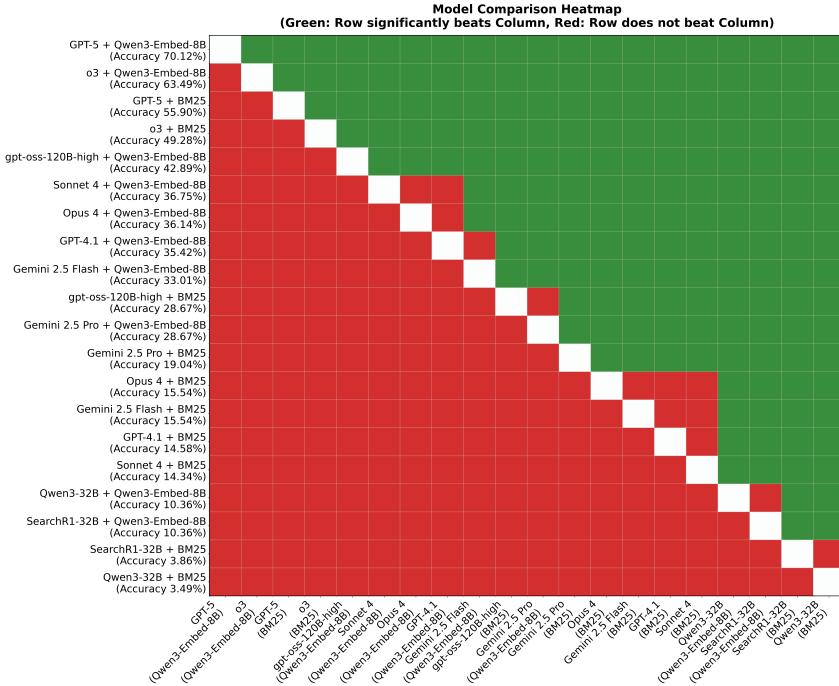


Figure 6: Pairwise McNemar’s tests with Bonferroni correction were conducted at a significance level of ≤ 0.05 . A green cell at Row (i), Column (j) indicates that the method in Row (i) performs significantly better than the method in Column (j).

P ANSWER ACCURACY WITH DIFFERENT JUDGMENT METHODS

In Table 11, we report answer-accuracy measurements using LLM-as-judge with GPT-4.1, Qwen3-32B, and substring matching. Across these evaluation methods, we observe consistent trends. Notably, the LLM-as-judge approach is more robust in handling cases where the predicted answers differ in format from the ground-truth labels.

1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470

1471	LLM	Retriever	Substring Match	GPT-4.1 Judge	Qwen3-32B Judge
1472	GPT-4.1	bm25	14.58	14.58	15.30
1473	GPT-4.1	Qwen3-Embedding-8B	34.46	35.42	36.39
1474	o3	bm25	45.78	49.28	50.48
1475	o3	Qwen3-Embedding-8B	60.48	63.49	65.90
1476	Sonnet 4	bm25	13.37	14.34	14.70
1477	Sonnet 4	Qwen3-Embedding-8B	33.73	36.75	37.35
1478	Opus 4	bm25	15.18	15.54	15.54
1479	Opus 4	Qwen3-Embedding-8B	33.13	36.14	36.75
1480	Gemini 2.5 Flash	bm25	15.54	15.54	16.27
1481	Gemini 2.5 Flash	Qwen3-Embedding-8B	31.45	33.01	34.58
1482	Gemini 2.5 Pro	bm25	17.71	19.04	19.88
1483	Gemini 2.5 Pro	Qwen3-Embedding-8B	27.83	28.67	29.52
1484	Qwen3-32B	bm25	3.25	3.49	3.61
1485	Qwen3-32B	Qwen3-Embedding-0.6B	4.22	4.10	4.22
1486	Qwen3-32B	Qwen3-Embedding-4B	8.43	7.83	8.07
1487	Qwen3-32B	Qwen3-Embedding-8B	9.76	10.36	10.72
1488	Qwen3-32B	ReasonIR	8.67	9.16	9.28
1489	SearchR1-32B	bm25	3.86	3.86	4.11
1490	SearchR1-32B	Qwen3-Embedding-0.6B	6.27	5.66	6.02
1491	SearchR1-32B	Qwen3-Embedding-4B	10.60	9.40	9.28
1492	SearchR1-32B	Qwen3-Embedding-8B	11.81	10.36	11.08
1493	SearchR1-32B	ReasonIR	10.64	9.43	9.31
1494	oss-20b-low	bm25	3.51	4.11	3.99
1495	oss-20b-low	Qwen3-Embedding-8B	11.93	13.37	14.10
1496	oss-20b-medium	bm25	15.54	16.39	16.87
1497	oss-20b-medium	Qwen3-Embedding-8B	26.87	29.88	30.48
1498	oss-20b-high	bm25	19.76	21.08	21.45
1499	oss-20b-high	Qwen3-Embedding-8B	31.93	34.58	35.06
1500	oss-120b-low	bm25	8.80	9.52	9.76
1501	oss-120b-low	Qwen3-Embedding-8B	22.41	24.94	25.54
1502	oss-120b-medium	bm25	21.33	23.73	24.58
1503	oss-120b-medium	Qwen3-Embedding-8B	33.49	37.59	38.55
1504	oss-120b-high	bm25	26.99	28.67	29.16
1505	oss-120b-high	Qwen3-Embedding-8B	40.24	42.89	44.10
1506	GPT-5	bm25	51.69	55.90	57.59
1507	GPT-5	Qwen3-Embedding-8B	65.18	70.12	71.69

Table 11: Comparison of accuracy measurement based on LLM-as-judge with GPT4.1, LLM-as-judge with Qwen3-32B, and sub-string Matching.

1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511