

CORPUSQA: A 10 MILLION TOKEN BENCHMARK FOR CORPUS-LEVEL ANALYSIS AND REASONING

Zhiyuan Lu¹, Chenliang Li¹, Yingcheng Shi¹, Weizhou Shen¹, Ming Yan¹, Fei Huang¹

¹Tongyi Lab, Alibaba Group

Correspondence: zylu738@gmail.com, ym119608@alibaba-inc.com

ABSTRACT

While large language models now handle million-token contexts, their capacity for reasoning across entire document repositories remains largely untested. Existing benchmarks are inadequate, as they are mostly limited to single long texts or rely on a "sparse retrieval" assumption—that answers can be derived from a few relevant chunks. This assumption fails for true corpus-level analysis, where evidence is highly dispersed across hundreds of documents and answers require global integration, comparison, and statistical aggregation. To address this critical gap, we introduce CorpusQA, a new benchmark scaling up to 10 million tokens, generated via a novel data synthesis framework. By decoupling reasoning from textual representation, this framework creates complex, computation-intensive queries with programmatically guaranteed ground-truth answers, challenging systems to perform holistic reasoning over vast, unstructured text without relying on fallible human annotation. We further demonstrate the utility of our framework beyond evaluation, showing that fine-tuning on our synthesized data effectively enhances an LLM’s general long-context reasoning capabilities. Extensive experiments reveal that even state-of-the-art long-context LLMs struggle as input length increases, and standard retrieval-augmented generation systems collapse entirely. Our findings indicate that memory-augmented agentic architectures offer a more robust alternative, suggesting a critical shift is needed from simply extending context windows to developing advanced architectures for global information synthesis.

1 INTRODUCTION

Recent advancements in large language models have pushed the boundaries of context processing, with models now capable of ingesting millions of tokens in a single prompt (Comanici et al. (2025); Yang et al. (2025b); Chen et al. (2025)). More recently, efforts have been made to extend the complex reasoning capabilities of reinforcement learning models to long-context scenarios (Wan et al. (2025)). This has unlocked remarkable capabilities in long-document understanding and summarization. However, a critical gap persists between processing a single long document and performing analysis across an entire document repository. In high-stakes professional domains such as finance, law, and scientific research, real-world tasks often demand reasoning not over one text, but over a collection of hundreds or thousands of documents to uncover insights that are only visible at a corpus level. For instance, a financial analyst might need to determine how many companies generated more net cash from financing activities than from operating activities after reviewing dozens of financial reports (Figure 1), or a legal researcher might need to synthesize precedents from an entire case law database. We term this challenging paradigm Corpus-Level Analysis.

A defining characteristic of this paradigm is the high dispersion of evidence, where critical information is not concentrated in a few documents but is finely scattered across the entire repository. Consequently, deriving a correct answer demands a holistic processing of the complete, ultra-long context, coupled with complex reasoning, statistical aggregation, and precise computation. This requirement directly invalidates the core assumption of Retrieval-Augmented Generation (RAG) (Lewis et al. (2020); Zhao et al. (2024); Yu et al. (2024)), which presupposes that answers can be synthesized from a small number of relevant retrieved chunks, rendering it inherently unsuitable for such globally-aware tasks.

Benchmark	Up to 10M Length	Reasoning Intensive	Controllable Context	High Evidence Dispersion	Multilingual
InfiniteBench Zhang et al. (2024)	✗	✗	✗	✗	✓
LongBench v2 Bai et al. (2024)	✗	✓	✗	✗	✓
Loong Wang et al. (2024)	✗	✗	✓	✓	✓
FanOutQA Zhu et al. (2024)	✗	✗	✓	✓	✗
FRAMES Krishna et al. (2024)	✗	✓	✗	✗	✗
CRAG Yang et al. (2024)	✗	✗	✗	✗	✗
CorpusQA (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison of CorpusQA with existing long-context and RAG benchmarks. CorpusQA is the first benchmark designed to simultaneously evaluate reasoning over massive-scale corpora (up to 10M tokens) with highly dispersed evidence, addressing a key gap in prior work.

Existing evaluation benchmarks Krishna et al. (2024); Yang et al. (2024) are inadequate to measure this capability. They are largely confined to single long texts or rely on a "sparse retrieval" assumption that is invalidated by the high evidence dispersion in corpus-level analysis. Even recent multi-document benchmarks Wang et al. (2024); Zhu et al. (2024) that require integrating information across all provided texts operate at a scale too small to pose a true corpus-level challenge, as their entire context often fits within a single prompt for modern LLMs. This leaves a critical gap in assessing holistic reasoning over vast repositories, a gap CorpusQA is designed to fill. Table 1 summarizes these limitations and contrasts them with our approach.

To bridge this gap and steer research towards this critical challenge, we introduce CorpusQA, the first benchmark designed for large-scale corpus-level analysis. Our benchmark makes several key departures from prior work by combining three core design principles. It features **unprecedented scale and complexity**, scaling to 10M tokens with reasoning-intensive queries that necessitate architectures for global synthesis over simple context expansion. This scale enables our second principle: **high information dispersion**, where deliberately scattered evidence invalidates RAG’s sparse retrieval assumption and necessitates a holistic understanding. Finally, to ensure reliable evaluation against this challenge, we introduce **guaranteed factual grounding**. We developed a novel schema-driven pipeline that programmatically computes ground-truth answers from an underlying structured representation of the documents, thus providing verifiable correctness without relying on a fallible LLM for annotation.

We conduct extensive experiments on CorpusQA, evaluating a wide array of state-of-the-art long-context LLMs, RAG systems, and more advanced agentic architectures. Our findings reveal a clear picture of the current landscape: while leading LLMs perform reasonably well on smaller (128K) corpora, their performance degrades significantly as the scale approaches 1M tokens. More strikingly, standard RAG systems fail catastrophically on our benchmark, as their retrieval mechanisms are unable to gather the globally distributed evidence required. In contrast, memory-augmented agentic systems demonstrate substantially greater resilience, highlighting a promising direction for future research.

Our main contributions are:

1. We introduce **CorpusQA**, the first large-scale benchmark for corpus-level analysis, featuring diverse domains, complex reasoning tasks, and a scale up to 10M tokens.

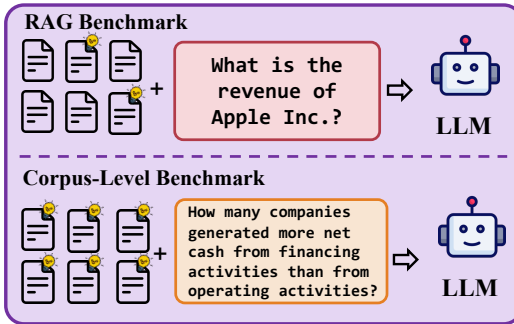


Figure 1: RAG-style benchmarks assume sparse retrieval, where the answer can be derived from a few retrieved documents. In contrast, corpus-level benchmarks require holistic reasoning over an entire document repository with highly dispersed evidence.

2. We present a novel long-context data synthesis framework that programmatically generates complex questions with verifiable ground-truth answers. We further demonstrate that data synthesized by this framework can effectively enhance an LLM’s long-context reasoning capabilities when used for training.
3. Through comprehensive experiments, we demonstrate the severe limitations of current LLMs and RAG systems on this new task paradigm and provide strong evidence that memory-based architectures are a more promising path forward for true corpus-level reasoning.

2 RELATED WORK

2.1 RAG EVALUATION

Retrieval-Augmented Generation (RAG) has become a standard paradigm for grounding LLMs in external knowledge. Consequently, a variety of benchmarks have been developed to evaluate RAG systems. For instance, **FRAMES** Krishna et al. (2024) assesses multi-hop reasoning over structured and unstructured knowledge bases, while **CRAG** Yang et al. (2024) introduces a more dynamic evaluation of RAG pipelines, including self-correction and web search capabilities. However, these benchmarks are premised on a "sparse retrieval" assumption—that answers can be synthesized from a few retrieved document chunks. This paradigm fails for corpus-level analysis, where critical information is highly dispersed across the entire document set. Such tasks demand a global understanding, rendering sparse retrieval inherently insufficient.

2.2 LONG CONTEXT EVALUATION

Recent advancements have spurred the creation of benchmarks targeting long-context understanding. One category, including **LongBenchV2** Bai et al. (2024) and **InfiniteBench** Zhang et al. (2024), primarily evaluates a model’s ability to recall facts or reason within a single, concatenated text. While they push the boundaries of context length, they do not address the challenge of integrating information scattered across a multitude of discrete documents.

Another category, multi-document QA benchmarks, takes a step closer to our paradigm. For example, **Loong** Wang et al. (2024) and **FanOutQA** Zhu et al. (2024) require reasoning across several documents. However, these benchmarks typically involve fewer than a dozen documents, and the total context size often remains within the capacity of modern long-context LLMs. In contrast, **CorpusQA** scales to hundreds of documents and millions of tokens, where evidence is so dispersed that simple retrieval or direct ingestion fails, necessitating more advanced, system-level architectures for global information synthesis.

3 THE CORPUS-LEVEL QA BENCHMARK

3.1 OVERVIEW

Creating a benchmark for corpus-level analysis presents a significant challenge: how to generate complex, reasoning-intensive questions with verifiably correct answers when the task itself is designed to be difficult for even the most advanced LLMs. To overcome this, we designed a novel data generation pipeline that decouples the complex reasoning from the unstructured text. By first structuring the information, programmatically deriving ground-truth answers, and then re-presenting the original unstructured documents to the models being evaluated, we ensure 100% factual accuracy and logical consistency. As illustrated in Figure 2, our pipeline is composed of six main stages, which can be grouped into four key phases.

3.2 DOCUMENT CURATION AND STRUCTURING

The foundation of **CorpusQA** lies in high-quality, data-rich source documents. This phase involves carefully selecting appropriate documents and transforming their key information into a machine-readable format.

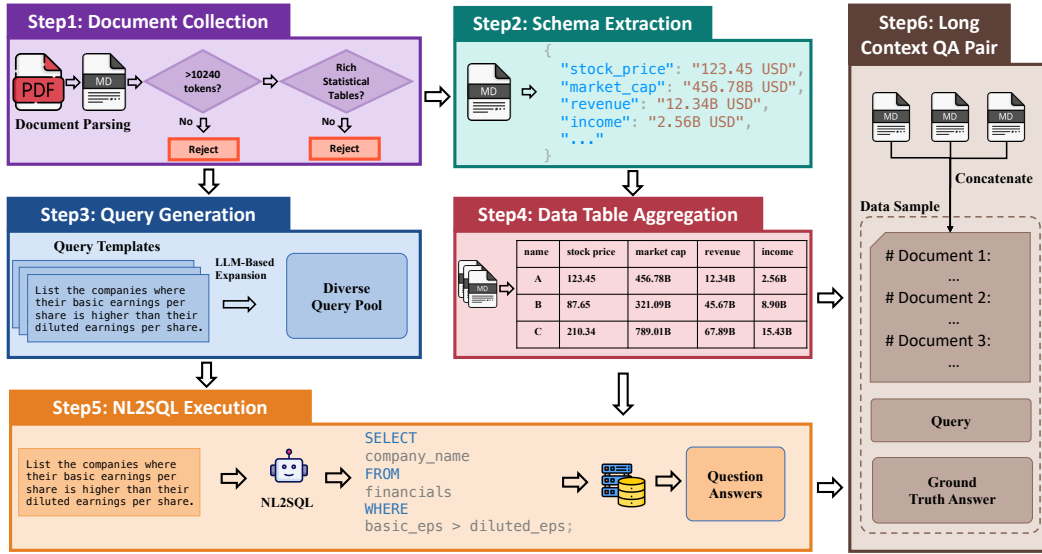


Figure 2: Overall data generation pipeline of CorpusQA. The process consists of six steps: (1) document collection and filtering to ensure long and data-rich inputs, (2) schema extraction for structured representation, (3) query generation via templates and LLM-based expansion, (4) data table aggregation across documents, (5) programmatic ground-truth generation using NL2SQL execution, and (6) final QA pair assembly by concatenating original documents with generated queries and answers. This design guarantees both realism and verifiable correctness for corpus-level reasoning tasks.

Step 1: Document Collection: We first collect a large set of real-world PDF documents from diverse domains such as finance and academia. Each PDF is then parsed into Markdown to obtain a uniform, text-centric representation for downstream processing. To ensure the benchmark’s focus on long-context and quantitative reasoning, we apply a two-stage filtering process. First, documents with fewer than 10,240 tokens are discarded. Second, we retain only those documents containing rich statistical tables and numerical data, as these are essential for constructing the complex analytical queries that define our benchmark.

Step 2: Schema Extraction We perform schema extraction using a robust, multi-model voting approach to guarantee data quality. For each document, we query an ensemble of three state-of-the-art language models (Gemini 2.5 Pro Comanici et al. (2025), GPT-4.1 OpenAI (2025a), and Qwen-Max Team) (2025)) multiple times for each target field. A value is only considered valid if it represents a clear majority across all extraction results. This consensus-driven method filters out hallucinations and inconsistencies; documents that fail to produce a consensus for any field are discarded entirely. The resulting validated key-value pairs (e.g., `stock_price: "123.45 USD"`, `revenue: "12.34B USD"`) are then structured into a consistent JSON-based schema for programmatic use in subsequent steps.

3.3 LLM-AUGMENTED QUERY GENERATION

To ensure a diverse and challenging set of questions, we generate queries through a semi-automated process that combines human expertise with LLM capabilities.

Step 3: Query Generation We first manually author a set of high-quality query templates that encapsulate various reasoning patterns, including comparison, ranking, conditional filtering, and multi-step calculations. An example template is: "List the companies where their basic earnings per share is higher than their diluted earnings per share." To increase linguistic diversity and prevent models from overfitting to specific phrasings, we leverage an LLM to paraphrase and expand these templates. This process results in a large and varied pool of natural language queries that are semantically grounded in the initial templates and categorized by difficulty (Easy, Medium, Hard)

based on their logical and computational complexity. For a detailed breakdown of these difficulty levels with examples, please see Appendix B.

3.4 PROGRAMMATIC GROUND TRUTH GENERATION

This phase is the cornerstone of our methodology, guaranteeing the correctness of our benchmark’s answers without relying on a fallible LLM’s reasoning over long text.

Step 4: Data Table Aggregation The structured JSON objects extracted from every document in a given corpus are aggregated to construct a single, global data table. In this table, each row typically represents an entity (e.g., a company, a university), and each column corresponds to a specific attribute from our schema. This table serves as the canonical, structured knowledge base for the entire document set.

Step 5: NL2SQL Execution To generate a verifiable answer for each query, we adopt a robust programmatic approach. A natural language query from our pool is first translated into an executable SQL statement using a state-of-the-art LLM. For instance, the example query from Step 3 would be converted to `SELECT company_name FROM financials WHERE basic_eps > diluted_eps;`. This SQL statement is then executed directly against the aggregated data table from Step 4. This method ensures that every ground truth answer is 100% accurate and logically derived from the complete source data, establishing an unimpeachable standard for evaluation.

Category	Avg T.	Lang.	#Inst.
<i>Domain</i>			
Financial	2.71M	ZH, EN	664
Education	3.12M	EN	328
Real Est.	3.59M	EN	324
<i>Length Set</i>			
Set1 (128K)	92.5K	EN, ZH	329
Set2 (1M)	905K	EN, ZH	329
Set3 (4M)	3.71M	EN, ZH	329
Set4 (10M)	7.84M	EN, ZH	329
<i>Difficulty</i>			
Easy	3.11M	EN, ZH	296
Medium	3.11M	EN, ZH	648
Hard	3.20M	EN, ZH	372

Table 2: Statistics of the dataset.

3.5 FINAL QA PAIR ASSEMBLY

In the final stage, we assemble the components into the complete benchmark instance that will be presented to the models.

Step 6: Long Context QA Pair: The original, full-text Markdown documents selected in Step 1 are concatenated in their entirety to form the final long-context input. Each instance in CorpusQA is thus a triplet comprising: (1) the concatenated document corpus as the context, (2) a natural language query from our generated pool, and (3) the programmatically derived ground truth answer. This final structure challenges a model to perform the sophisticated, corpus-level reasoning over raw, unstructured text that we have systematically executed on a structured backend.

3.6 DATA QUALITY VERIFICATION

To validate the integrity of our data generation pipeline, we conducted a rigorous quality check on the schema extraction stage, which forms the basis for our programmatic ground-truth generation. For the `education_en` and `financial_zh` domains, which rely on model-based extraction from unstructured text, we randomly sampled 1,000 extracted schemas per domain for manual verification. The `financial_en` and `real_estate_en` domains were exempt, as their data was derived directly from structured sources, ensuring its fidelity. The initial consistency between the model’s extractions and human verification was very high: 94.13% for `education_en` and 98.97% for `financial_zh`. Critically, a secondary review of all inconsistent cases by different human annotators revealed that the LLM’s original extraction was correct in every instance of disagreement. This finding underscores the robustness of our multi-model voting approach and confirms the near-perfect accuracy of the structured data used to generate the benchmark’s ground-truth answers.

Model	Financial-zh	Financial-en	Education-en	Real Estate-en	Overall
Set1 (128K)					
Gemini-2.5-Pro	89.16	80.72	76.83	74.07	80.19
Gemini-2.5-Flash	83.13	74.70	74.39	79.01	77.81
GPT-5	93.98	78.31	81.71	74.07	82.02
GPT-5-Mini	93.98	78.31	74.39	72.84	79.88
Qwen3-235B-A22B-Thinking	90.36	74.70	70.73	76.54	78.08
Qwen3-30B-A3B-Thinking	83.13	61.45	64.63	69.14	69.59
DeepSeek-R1-0528	83.13	73.49	78.05	64.20	74.72
MiniMax-M1-80k	73.49	67.07	76.83	66.67	71.02
Set2 (1M)					
Gemini-2.5-Pro	50.60	48.19	65.85	38.27	50.73
Gemini-2.5-Flash	26.25	43.37	48.78	30.86	37.31
MiniMax-M1-80k	12.05	9.64	7.32	17.50	11.63

Table 3: Accuracy (%) of base LLMs across four domains under two context-length settings. Set1 (128K) corresponds to the maximum context length supported by most frontier models, while Set2 (1M) includes the few base LLMs that claim million-token support. Results show strong performance at 128K but substantial degradation at 1M.

Model	Financial-zh	Financial-en	Education-en	Real Estate-en	Overall
Set1 (128K)					
Memory Agent - Gemini-2.5-Pro	79.01	57.58	65.82	67.90	67.58
RAG - Gemini-2.5-Pro	71.08	51.81	58.54	41.98	55.85
Set2 (1M)					
Memory Agent - Gemini-2.5-Pro	56.79	45.78	41.46	35.80	44.95
RAG - Gemini-2.5-Pro	10.39	18.07	25.00	16.05	17.38
Set3 (4M)					
Memory Agent - Gemini-2.5-Pro	19.28	12.05	32.93	28.12	22.14
RAG - Gemini-2.5-Pro	1.20 [†]	3.61	3.70 [†]	1.22 [†]	2.43
Set4 (10M)					
Memory Agent - Gemini-2.5-Pro	7.89	9.64	14.63	12.35	11.13
RAG - Gemini-2.5-Pro	3.61 [†]	1.20 [†]	0.00 [†]	0.00 [†]	1.20

Table 4: Accuracy (%) of Memory Agent and RAG, both built on Gemini-2.5-Pro, across four domains under four different corpus scales (128K–10M tokens). Higher values indicate better performance. [†] denotes cases where the number of retrieved chunks by RAG was smaller than the number of documents, making accurate answers impossible.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation Targets Our benchmark is divided into four context lengths: 128k, 1M, 4M, and 10M. Among them, the 128k and 1M settings are mainly used for directly testing the long-context capabilities of LLMs, while the 4M and 10M settings are primarily used to evaluate the effectiveness of advanced strategies such as retrieval-augmented generation and memory agents. For long-context evaluation, we consider both closed-source and open-source models. Closed-source models include Gemini-2.5-Pro Comanici et al. (2025), Gemini-2.5-Flash Comanici et al. (2025), GPT-5-2025-08-07 OpenAI (2025b), GPT-5-Mini-2025-08-07 OpenAI (2025b). Open-source models include the Qwen3 family Yang et al. (2025a) (Qwen3-235B-Thinking-2507 and Qwen3-30B-Thinking-2507, DeepSeek-R1-0528 Guo et al. (2025), and Minimax-M1-80k Chen

et al. (2025). All of the above models are evaluated on the 128k setting, while the 1M setting is tested with Gemini-2.5-Pro, Gemini-2.5-Flash, and Minimax-M1 due to their extended context window support.

For the 4M and 10M settings, where the total corpus length significantly exceeds any available model’s context window, we focus on system-level solutions:

Retrieval-Augmented Generation. RAG systems operate by first segmenting the corpus into retrievable chunks, then retrieving a small subset relevant to the query, and finally generating answers conditioned on both the retrieved evidence and the query. This approach reduces the effective input size but relies heavily on retrieval recall and chunk-level information density. Moreover, because corpus-level tasks often involve evidence scattered across the entire context, a large number of chunks must be retrieved to ensure sufficient coverage and accuracy of the final answer.

Memory Agents. The Memory Agent Yu et al. (2025) framework is designed for ultra-long text processing. It operates by processing a document in chunks, iteratively updating a fixed-size memory buffer. The final output is generated solely from the information consolidated in this memory. We evaluate two implementations of this approach: (1) MemAgent-14B, a specialist model fine-tuned for this task by Yu et al. (2025), and (2) general-purpose LLM (Gemini-2.5-Pro) instructed to follow the same workflow without task-specific tuning.

Evaluation Metrics Many answers in our benchmark involve numbers, percentages, or named entities that can appear in different but equivalent forms. This makes exact string matching unreliable for correctness evaluation.

To address this, we use an LLM-as-a-Judge Zheng et al. (2023); Gu et al. (2024) approach: a strong LLM is prompted to assess whether a model’s prediction is semantically equivalent to the reference answer. We then compute accuracy as the percentage of predictions judged correct. This method provides a more flexible and robust evaluation than strict matching, especially for reasoning-heavy and numerically precise tasks. We use DeepSeek-V3Liu et al. (2024) as the judge model, and the prompt template is provided in the Appendix.

Implement Details We set temperature = 0 to eliminate randomness and keep other hyperparameters default. For API-Based LLMs, we directly utilize the official API for testing. As for open-source models, we conduct experiments on a server with 8xA100 80GB.

4.2 MAIN RESULTS

Most base LLMs currently support context lengths of up to 128k tokens, and we evaluate ten recent models under this setting. In addition, a few base LLMs are able to process 1M-token contexts, for which we evaluate three models. The results across both the 128k and 1M settings are reported in Table 3. For larger-scale settings (4M and 10M tokens), which exceed the context limits of base LLMs, we rely on system-level methods—retrieval-augmented generation and Memory Agents to handle ultra-long contexts. Their performance is evaluated across the 128k to 10M settings, with results summarized in Table 4.

Table 3 shows that closed-source models achieve the best results at 128K, with GPT-5 (82.02) and Gemini-2.5-Pro (80.19) leading the benchmark. When extended to 1M tokens, performance drops substantially for all models, yet Gemini-2.5-Pro remains comparatively strong at 50.73, outperforming other baselines by a large margin. Among open-source models, Qwen3-235B-Thinking (78.08) and DeepSeek-R1 (74.72) perform competitively at 128K, but instruction-tuned variants and smaller models degrade severely, with MiniMax-M1-80k collapsing to 11.63 at 1M. Overall, while both closed- and open-source LLMs handle 128K contexts reasonably well, only top-tier models such as Gemini-2.5-Pro show partial resilience at the million-token scale.

4.3 THE COLLAPSE OF RAG AND THE PROMISE OF MEMORY AGENTS

Table 4 compares Memory Agent and RAG, both built on Gemini-2.5-Pro, from 128K to 10M tokens. At 128K, Memory Agent already outperforms RAG (67.58 vs. 55.85). As scale increases, the gap widens: at 1M tokens, Memory Agent retains moderate performance (44.95), while RAG drops sharply (17.38); at 4M and 10M tokens, RAG nearly fails, whereas Memory Agent remains functional.

These results reveal a core challenge of corpus-level analysis: **high information dispersion**. Retrieval-only methods cannot capture evidence scattered across thousands of documents, leading to collapse under ultra-long contexts. In contrast, memory mechanisms provide more robust aggregation, though still far from solving the task. By scaling to 10M tokens, our benchmark not only reflects the growing context windows of base LLMs but also offers a platform to evaluate advanced system-level methods such as memory, compression, and agentic approaches.

4.4 IMPACT OF CORPUS-LEVEL TRAINING

A key advantage of our data synthesis framework is its ability to generate not only evaluation benchmarks but also high-quality training data. To demonstrate this, we conducted a pilot experiment to assess the impact of training on data synthesized by our framework. We synthesized a new training set of 480 samples and used it to fine-tune the open-source Qwen3-4B-Thinking model for 25 steps with Group Relative Policy Optimization (GRPO) Shao et al. (2024). We then evaluated the model’s performance on our CorpusQA test set (in-domain) and on two external long-context benchmarks, LongBenchV2 and FRAMES, to measure out-of-domain generalization.

The results, presented in Figure 3, are promising. The fine-tuned model achieved a significant gain on CorpusQA, demonstrating its ability to learn the corpus-level analysis task. More interestingly, the model also showed notable improvements on both LongBenchV2 and FRAMES. This suggests that training on the complex, globally-aware reasoning tasks generated by our framework can enhance a model’s broader long-context capabilities.

4.5 HUMAN PERFORMANCE

To contextualize the benchmark’s difficulty, we conducted a comparative study between a human expert and two AI systems: a base LLM (Qwen3-30B-A3B-Thinking) and a Memory Agent built upon the same model. We sampled 12 question-answer pairs at each scale (128K, 1M, 4M, and 10M tokens). To prevent humans from spending too much time on a single problem, we set a time limit of 90 minutes, after which they give up. The results are summarized in Table 5.

The findings reveal a clear performance hierarchy that shifts dramatically with scale. At 128K tokens, the task is challenging but manageable: the human expert is most accurate (75.0%) but slow, while the base LLM offers a strong balance of speed and accuracy (66.7%). However, the 1M token mark represents a critical tipping point where human and base LLM performance plummets, and beyond this, the task becomes demonstrably superhuman. At 4M and 10M tokens, the human expert failed every task and the base LLM is architecturally incapable of processing the input, leaving the Memory

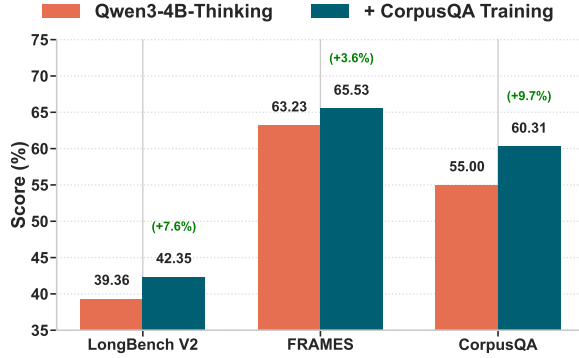


Figure 3: Impact of fine-tuning with synthesized data. Training a model on data generated by our synthesis framework not only improves in-domain performance (CorpusQA) but also generalizes to enhance performance on external long-context benchmarks (LongBenchV2 and FRAMES).

Context	Method	Acc.(%)	Time (min)	Timeout (%)
128K	Human	75.0	17.8	6.6
	Base LLM	66.7	2.1	0
	Memory Agent	50.0	1.4	0
1M	Human	50.0	31.9	33.3
	Base LLM	20.0	120.7	0
	Memory Agent	33.3	69.7	0
4M	Human	0.0	–	100.0
	Base LLM	Context Limit Exceeded		
	Memory Agent	25.0	323.5	0
10M	Human	0.0	–	100.0
	Base LLM	Context Limit Exceeded		
	Memory Agent	16.7	789.3	0

Table 5: **Human on CorpusQA**. Comparison of a human expert, a base LLM, and our agent. **Acc.**: Accuracy. **Time**: Average time in minutes. **Timeout**: Percentage of tasks where the subject timed out (>90min) or gave up. Base LLMs cannot process 4M/10M contexts due to architectural limits.

This suggests that training on the complex, globally-aware reasoning tasks generated by our framework can enhance a model’s broader long-context capabilities.

Agent as the only viable approach, which remained operational and achieved non-trivial accuracy (25.0% at 4M and 16.7% at 10M). This experiment powerfully underscores two conclusions: first, CorpusQA presents a challenge that rapidly exceeds human cognitive limits, validating its difficulty; second, it provides compelling evidence that for truly large-scale corpus analysis, simply expanding an LLM’s context window is an inefficient strategy. Scalable, resilient architectures like the Memory Agent are not just an alternative but a necessity to tackle reasoning at this scale.

5 CONCLUSION

In this work, we introduced CorpusQA, a new benchmark scaling to 10 million tokens designed to evaluate corpus-level reasoning, a task defined by high evidence dispersion where traditional benchmarks fall short. Our experiments reveal that while even top-tier LLMs struggle at scale, standard RAG systems collapse entirely, proving inadequate for holistic analysis. In contrast, memory-augmented agentic architectures demonstrate far greater resilience, suggesting a promising direction for future research. Our findings underscore a critical need to shift focus from merely extending context windows to developing novel architectures capable of global information synthesis, a challenge for which CorpusQA now provides a crucial yardstick.

LIMITATIONS

This work has two primary limitations. First, our evaluation relies on an LLM-as-a-Judge, which is less deterministic than a fully rule-based system. Second, our experiments did not extend to more complex agentic architectures, such as those performing deep research Li et al. (2025b;a); Team (2025), leaving their performance on CorpusQA as an important area for future investigation.

REFERENCES

- Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.
- Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation, 2024. URL <https://arxiv.org/abs/2409.12941>.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.
- Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, et al. Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning. *arXiv preprint arXiv:2509.13305*, 2025a.

- Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025b.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- OpenAI. Introducing GPT-4.1 in the api. <https://openai.com/index/gpt-4-1/>, 2025a.
- OpenAI. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/>, 2025b.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Alibaba (Qwen Team). Qwen-max (snapshot version 2025-01-25). <https://www.alibabacloud.com/help/en/model-studio/models>, 2025.
- Tongyi DeepResearch Team. Tongyi deepresearch: A new era of open-source ai researchers. <https://github.com/Alibaba-NLP/DeepResearch>, 2025.
- Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. Qwenlong-1l: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*, 2025.
- Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *arXiv preprint arXiv:2406.17419*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, et al. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025b.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490, 2024.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pp. 102–120. Springer, 2024.
- Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiyang Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, et al. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *arXiv preprint arXiv:2507.02259*, 2025.
- Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. Infinitebench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024.
- Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*, 2024.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zou, Zhuohan Liu, Ion Stoica, Eric P Xing, Joseph E Gonzalez, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.
- Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. *arXiv preprint arXiv:2402.14116*, 2024.

A REASONING MODELS VS. INSTRUCTION MODELS

Our benchmark specifically targets long-context reasoning, and the results in Figure 4 clearly show that reasoning-oriented models substantially outperform their instruction-tuned counterparts (e.g., Qwen3-235B-Thinking: 78.08 vs. Qwen3-235B-Instruct: 61.41). This consistent gap highlights a unique feature of our benchmark: it is sensitive to reasoning ability rather than alignment alone, making it particularly effective at revealing the trade-off introduced by instruction tuning in long-text scenarios.

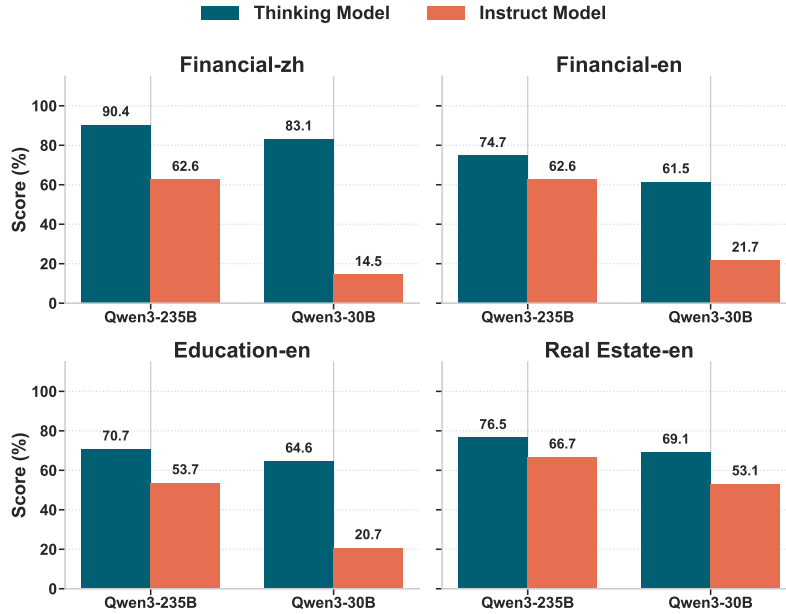


Figure 4: Comparison of Thinking vs. Instruct models on the 128K setting. Reasoning-oriented models (Thinking) consistently outperform their instruction-tuned counterparts, highlighting the benchmark’s sensitivity to complex reasoning over simple instruction following.

B QUERY DIFFICULTY LEVELS

The queries in CorpusQA are classified into three difficulty levels to systematically evaluate different aspects of corpus-level reasoning. Below are the definitions and representative examples for each category.

Easy: Basic Aggregation and Single-Dimension Analysis This level tests for simple data retrieval, filtering, and sorting.

- **Simple Aggregation:** e.g., "List the top 5 universities with the highest total enrollment."
- **Single-Condition Filtering:** e.g., "Find universities with more than 1,000 part-time faculty members."
- **Basic Ranking:** e.g., "Rank the top 10 universities by the number of transfer students."

Medium: Numerical Calculation and Multi-Condition Analysis This level requires combining multiple criteria and performing basic arithmetic operations.

- **Ratio Calculation:** e.g., "Calculate the admission rate for each university."
- **Multi-Condition Combination:** e.g., "Which universities have a graduation rate above 90% and fewer than 5,000 undergraduate students?"

- **Simple Custom Metrics:** e.g., "What is the student-to-faculty ratio for each institution?"
- **Grouped Comparison:** e.g., "Compare the average graduation rates of large versus small universities."

Hard: Custom Metrics and Multi-Step Reasoning This level involves complex, multi-step calculations, custom-defined business logic, and advanced analytical reasoning.

- **Multi-Step Calculation:** e.g., "For all universities with a first-year admission rate below 50%, calculate the average six-year graduation rate. Then, find the universities in this group whose individual graduation rate is more than 15 percentage points above this calculated average."
- **Complex Business Logic:** e.g., "Which companies in Q1 2025 had a positive 'operating safety margin' (defined as Net cash from operations - Cash paid to employees - Financial expenses), where this margin also exceeded their net profit attributable to shareholders?"
- **Advanced Metric Construction:** e.g., "Identify companies with high financial risk, defined by the criterion: $(X_1 \times 1.2 + X_2 \times 1.4 + X_3 \times 3.3) < 1.8$, where $X_1 = (\text{Current Assets} - \text{Accounts Payable} - \text{Contract Liabilities}) / \text{Total Assets}$, $X_2 = \text{Undistributed Profit} / \text{Total Assets}$, and $X_3 = \text{Operating Revenue} / \text{Total Assets}$."

C DATA TEMPLATE

To ensure consistent and fair evaluation, all models were tested using a standardized prompt structure. The template below details the format, which includes the concatenated corpus of documents, the specific question, and a strict set of instructions on output formatting.

```
# Document 1:
<document1>

# Document 2:
<document2>

# Document 3:
<document3>

...

# Question:
<question>

# Output requirements:

1. Your answer can only be one of the following types: a list of strings (including empty list "[]"), a single short string, a number (floating point numbers should retain 2 decimal places, without units), or a percentage (retain 2 decimal places).

2. To output the date, use the format "YYYY-MM".

3. If multiple files contain a conflicting value for the same item, please use the value from the latest file.

4. Please place the final answer at the end, and strictly use the format: The answer is: xxx.

# Answer format examples:

• For a list of strings: The answer is: ["NETFLIX INC", "H&R BLOCK INC"]
• For a single string: The answer is: "South"
• For a number: The answer is: 1234567.89
• For a percentage: The answer is: 12.34%
```

D LLM-AS-A-JUDGE PROMPT

Given that answers can be numerically complex or expressed in varied but equivalent phrasings, simple string matching is insufficient for accurate evaluation. We therefore employ an LLM-as-a-Judge approach. The following prompt instructs the judge model to assess the semantic equivalence between a model’s generated answer and the ground truth, ensuring a robust and fair scoring process.

You are an expert in verifying if two answers are the same. Your input is a problem and two answers, Answer 1 and Answer 2. Your task is to determine if they are equivalent, without attempting to solve the original problem.

Compare the answers to verify they represent identical values or meaning, even when written in different forms or notations.

Your output **must** follow the following format:

1. Provide an explanation for why the answers are equivalent or not.
2. Then provide your final answer in the form of: [[YES]] or [[NO]]

Problem: <question>

Answer 1: <model_answer>

Answer 2: <ground_truth>