QuantMind: A Context-Engineering Based Knowledge Framework for Quantitative Finance

Haoxue Wang[1, 2*], **Keli Wen**[1*], **Yuante Li**[1, 3*], **Qu Qiancheng** [1, 4*], **Xiangxu Mu** [1, 5*]
Xinjie Shen [1], **Jiaqi Gao** [1], **Chenyang Chang** [1], **Chuhan Xie** [1, 6], **San Yu Cheung** [1, 7]
Zhuoyuan Hu [1, 8], **Xinyu Wang**[1, 9], **Sirui Bi** [1, 2], **Bi'an Du** [1†]
[1] LLMQuant Research, [2] University of Cambridge, [3] Carnegie Mellon University
[4] National University of Singapore, [5] IEIT SYSTEMS CO., LTD., [6] Peking University
[7] The Chinese University of Hong Kong, [8] Shanghai Jiao Tong University, [9] Beihang University
{haoxue, keli, bian, sirui}@llmquant.com, yuantel@cs.cmu.edu
e1486355@u.nus.edu, muxiangxu@ieisystem.com

Abstract

Quantitative research increasingly relies on unstructured financial content such as filings, earnings calls, and research notes, yet existing LLM and RAG pipelines struggle with point-in-time correctness, evidence attribution, and integration into research workflows. To tackle this, We present `QuantMind`, an intelligent knowledge extraction and retrieval framework tailored to quantitative finance. `QuantMind` adopts a two-stage architecture: (i) a **knowledge extraction** stage that transforms heterogeneous documents into structured knowledge through multi-modal parsing of text, tables, and formulas, adaptive summarization for scalability, and domain-specific tagging for fine-grained indexing; and (ii) an **intelligent retrieval** stage that integrates semantic search with flexible strategies, multi-hop reasoning across sources, and knowledge-aware generation for auditable outputs. A controlled user study demonstrates that `QuantMind` improves both factual accuracy and user experience compared to unaided reading and generic AI assistance, underscoring the value of structured, domain-specific context engineering for finance.

1 Introduction

The growing reliance on unstructured financial data, including SEC filings, earnings call transcripts, and broker research notes, has fundamentally influenced quantitative finance. Domain-specific language models like FinBERT [1], BloombergGPT [2], and FinGPT [3], together with retrieval-augmented generation (RAG) pipelines [4, 5, 6], have advanced financial NLP and retrieval. However, they continue to face persistent challenges: ❶ lack of point-in-time correctness, ❷ insufficient evidence traceability, and ❸ limited integration with quantitative research workflows. These limitations restrict their reliability in practical financial applications.

The rise of long-context and agent-oriented large language models (LLMs), such as GPT-4 [7] and Claude [8], highlights the need for structured, agent-readable knowledge frameworks in finance. While general frameworks such as `agents.md` [9] and `claude.md` [10] provide broad design principles, they do not address the domain-specific requirements of financial research. Meeting these demands requires transforming heterogeneous and dynamic financial artifacts into structured, auditable knowledge that supports reproducible reasoning and seamlessly integrates with customizable research workflows.

[*] Equal Contribution.
[†] Corresponding Author.

To address these gaps, we propose `QuantMind`, an intelligent knowledge extraction and retrieval framework tailored to quantitative finance. Our contributions are threefold:

- We design a two-stage decoupled architecture comprising **knowledge extraction** and **intelligent retrieval**, enabling point-in-time correctness, provenance preservation, and reproducibility.
- In the extraction stage, we introduce AI-driven multi-modal parsing, adaptive summarization with cost-optimized chunking, and domain-specialized tagging for fine-grained indexing.
- In the retrieval stage, we develop a flexible retrieval pattern architecture supporting both DeepResearch and RAG, unified knowledge representation with vectorization readiness, and a flow-based orchestration layer inspired by multi-agent systems.

2 Related Work

Retrieval-Augmented Generation and Context Engineering. RAG combines a retriever and a generator, typically built on dense encoders (e.g., SBERT [11]) and ANN indexes (e.g., FAISS [12]). Standard pipelines operate in a straightforward manner: documents are chunked, embedded, and the top-k results are concatenated [4, 11, 13]. This design often produces weak provenance. To address these limitations, context engineering introduces reranking [14, 15], multi-hop retrieval [16, 17], and structure-aware methods [18, 19]. Nevertheless, existing systems remain fragment-centric and version-insensitive. In financial applications, where critical evidence is distributed across tables, footnotes, and figure captions, such fragmentation undermines reference integrity and disrupts.

Scientific Knowledge Management Systems. Beyond finance, infrastructures such as the Semantic Scholar Literature Graph, S2ORC [20], and ORKG [21] demonstrate how structured corpora and knowledge graphs can facilitate search and enable limited forms of reasoning through explicit structure and versioning [22]. However, these systems are domain-general and do not provide an agentic, finance-aware layer capable of multistep, auditable, retrieval-augmented reasoning over continuously evolving documents.

3 QuantMind

QuantMind adopts a two-stage decoupled architecture that separates knowledge extraction from intelligent retrieval, as illustrated in Figure 1. The framework consists of two main stages: 1) Knowledge Extraction Stage, which transforms unstructured financial content into structured knowledge units, and 2) Intelligent Retrieval Stage, which enables dynamic retrieval patterns for knowledge access and analysis.

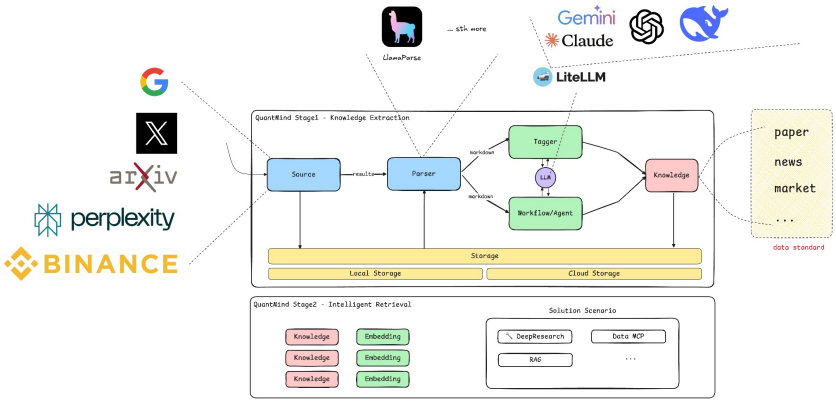

Figure 1: Architecture of `QuantMind`. The framework follows a two-stage design: (i) **knowledge extraction**, which structures heterogeneous financial documents via multi-modal parsing, summarization, and tagging; and (ii) **intelligent retrieval**, which supports semantic search, multi-hop reasoning, and knowledge-aware generation for quantitative research.

Stage1: Knowledge Extraction. Reliable LLM analysis, reasoning, and QA depend on the quality of source data [23, 24]. We therefore build a knowledge extraction pipeline that starts with large scale crawling and continues with multimodal parsing and enrichment:

➟ *Multi-Modal Parsing.* In quantitative finance, key insights are often encoded not only in text but also in tables of results, figures of market dynamics, formulas defining models, and even podcast and video modalities, ignoring these modalities risks of incomplete or biased knowledge extraction. To address this, we adopt a multi-modal parsing operator $\mathcal{P}$ (implemented with LlamaParse that decomposes a financial document D into $\mathcal{P}(D) = \{T, F, M, S\}$, where T denotes textual passages, F visual and tabular elements, M symbolic mathematical content, and S the semantic organization that integrates them. By preserving cross-modal dependencies, the extracted knowledge serves as an abstraction that encodes both the content and the underlying logical–empirical relations of the paper.

➟ *Adaptive Summarization.* Long documents are costly to be processed directly. To improve scalability, we split the content $C = \{c_1, \ldots, c_n\}$ into semantically coherent segments and summarize each segment with a lightweight model $s_i = M_{\text{cheap}}(c_i)$. These local summaries are then aggregated by a more expressive model $S_{\text{final}} = M_{\text{powerful}}(s_1 \oplus \cdots \oplus s_n)$, where $\oplus$ denotes ordered concatenation. This two-stage design reduces cost since $\text{Cost}_{\text{total}} = n \cdot \text{Cost}(M_{\text{cheap}}) + \text{Cost}(M_{\text{powerful}}) \ll n \cdot \text{Cost}(M_{\text{powerful}})$, while retaining fidelity at the global level.

➟ *Domain-Specialized Tagging.* To better capture relationships across papers, we introduce a high-dimensional tagging scheme that provides fine-grained indexing. Each paper P is enriched with structured labels covering its primary research area, secondary topics, methodological orientation, and application domain. For robustness, each tag t_i is associated with a confidence score $c_i = f(C, t_i, M_{\text{tag}}) \in [0, 1]$, computed by a discriminative tagging model M_{tag}, which supports domain-aware retrieval and systematic cross-paper comparison.

Stage2: Intelligent Retrieval. Once high-quality knowledge units are extracted, the next challenge lies in retrieving and integrating them for downstream reasoning and question answering. This stage equips the framework with flexible retrieval patterns, multi-hop reasoning, and knowledge-aware generation.

➟ *Adaptive Retrieval Strategies.* Given a query q and a knowledge base K, the framework dynamically selects retrieval patterns. For simple factual queries, a lightweight RAG-style approach retrieves once, $R = \text{Retrieve}(q, K)$, and augments generation with this context to produce an answer. For more complex reasoning, a DeepResearch process iteratively expands context, $R_t = \text{Retrieve}(q \oplus R_{t-1}, K), \quad R_{\text{final}} = M_{\text{reason}}(q, R_1, \ldots, R_n)$, balancing efficiency with the need for depth.

➟ *Multi-Hop Reasoning.* Certain queries require connections that span multiple documents or methodologies. To support this, the system performs iterative retrieval and synthesis across hops, progressively enriching the query until it captures the necessary context. This mechanism enables conceptual linking (e.g., relating market regimes), methodological comparison (contrasting algorithms), and empirical validation across independent studies.

➟ *Knowledge-Aware Generation.* The retrieved context R_{context} is integrated with the query to enhance both the generation of questions and answers. Enriched questions $Q_{\text{enhanced}} = M_{\text{gen}}(q \oplus R_{\text{context}})$ incorporate broader connections, while answers $\text{Ans} = M_{\text{ans}}(Q_{\text{enhanced}}, q \oplus R_{\text{context}})$ combine local content with the retrieved knowledge. This allows the system to move beyond factual responses, supporting comparative reasoning, cross-domain insights, and temporally informed analysis.

4 User Study

Goal and Design. The aim of this study is to assess whether `QuantMind` improves quantitative finance research performance relative to unaided reading and a generic AI assistant. We define research performance as the ability to: (i) extract factual information from academic papers (e.g., factor definitions, alpha sources, return characteristics), and (ii) perform higher-level reasoning (e.g., evaluating factor generalization across markets, horizons, or methodological extensions).

We employed a within-subjects repeated measures design with counterbalancing to control for potential learning and fatigue effects. Each participant completed six tasks, each derived from a distinct finance paper, under one of the following three conditions: ❶ **Without AI**: Participants relied exclusively on their own reading and research skills; ❷ **With AI Assistant**: Participants were provided with assistance generated by a generic LLM (`GPT-4o`), using a fixed prompt created prior to the study to supply supplementary information without offering direct answers; ❸ **With QuantMind**: Participants received assistance generated by the proposed `QuantMind` framework, which employed

domain-specific RAG from a structured knowledge base. The prompting strategy was optimized to extract and synthesize relevant information from both the focal paper and related research.

To balance exposure across conditions, we applied a Latin Square design [25] with participants and papers as blocking factors, ensuring equal representation of each condition and controlling for order effects. The detailed assignment is provided in Appendix A.2.

Corpus. We curated a set of six seminal papers in quantitative finance (see Appendix A.3), selected to capture both historical breadth and methodological diversity. The corpus spans foundational contributions in financial economics, canonical studies in factor modeling, and recent works situated at the intersection of machine learning and quantitative finance. For each paper, we designed two distinct categories of evaluation tasks: ❶ **Information-extraction questions** (3–4 per paper): objective queries with verifiable ground-truth answers, such as precise definitions of factors, data sources, or performance statistics. Responses were scored against a pre-defined answer key. ❷ **Analytical questions** (1–2 per paper): open-ended tasks requiring higher-order reasoning, including the evaluation of factor generalizability across markets, time horizons, or methodological extensions. Responses were assessed using an LLM-as-a-judge framework [26], which enabled systematic comparison of logical coherence, depth, and interpretive accuracy.

Metrics. We evaluated research performance along two principal dimensions: `Quality`: quality of answers judged upon subjects' comprehension, logic, as well as the breadth and depth of their insights. `UX Rating`: subjective evaluation of AI assistance on a 5-point Likert scale covering 4 dimensions: relevance, accuracy, helpfulness, and clarity.

Statistical Analysis. We analyzed both performance metrics (`Quality` and `UX Rating`) using linear mixed-effects models to account for the within-subjects repeated measures design:

$$Y_{ijk} = \mu + \tau_i + s_j + p_k + \epsilon_{ijk} \,, \tag{1}$$

where Y_{ijk} is the observed value for treatment i, subject j, and paper k; μ is the grand mean; τ_i denotes the fixed effect of treatment condition; $s_j \sim N(0, \sigma_s^2)$ and $p_k \sim N(0, \sigma_p^2)$ are random intercepts for subject and paper, respectively; and $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$ is the residual error.

Such model specification allows us to isolate the treatment effect while accounting for subject- and paper-level variability. To further assess differences between assistance conditions, we conducted post-hoc pairwise comparisons using Tukey's HSD ($\alpha = .05$), and quantified effect sizes with Cohen's d.

Results. Table 4 reports descriptive statistics and pairwise differences. Each treatment condition includes approximately equal numbers of observations (66–68), and the variance structure indicates heterogeneity across both participants and papers. For `Quality`, `QuantMind` significantly outperforms both baselines: a 1.14-point gain over the no-AI condition ($p < 0.001$) and a 0.43-point gain over the generic AI assistant ($p = 0.001$). For `UX Rating`, while the overall treatment effect was not significant at the conventional level ($p = 0.108$), `QuantMind` improved average ratings by 0.38 points and yielded a statistically significant enhancement in perceived helpfulness ($p = 0.003$). These findings suggest that `QuantMind` not only enhances the accuracy and depth of research outputs but also provides a more supportive user experience compared to both unaided reading and a generic AI assistant. Complete statistical outputs, including confidence intervals and additional pairwise contrasts, are provided in Appendix A.4.

Metric	Treatment	n	Mean	SD	Median	Min–Max
Accuracy (0–5)	Without AI	66	3.11	0.68	3.0	1.0–4.5
	AI Assistant	68	3.82	0.64	4.0	2.0–5.0
	QuantMind	66	4.25	0.55	4.5	3.0–5.0
UX Rating (1–5)	Without AI	66	—	—	—	—
	AI Assistant	68	3.78	0.62	4.0	2.0–5.0
	QuantMind	68	4.21	0.55	4.0	3.0–5.0

Table 1: Descriptive statistics of accuracy and UX ratings across treatment conditions, showing n, mean, SD, median, and range.

5 Conclusion

We introduced QuantMind, a framework that converts heterogeneous financial artifacts into a structured, agent-readable knowledge base with domain-aware retrieval. Its decoupled extraction and retrieval stages ensure point-in-time correctness, provenance, and reproducibility, while supporting DeepResearch, RAG, and structured natural language access. A controlled study confirms that QuantMind improves both accuracy and efficiency in quantitative research.

References

[1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019. URL https://arxiv.org/abs/1908.10063.

[2] Shijie Wu, Joshua Bolton, Yousef Rizk, et al. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023. URL https://arxiv.org/abs/2303.17564.

[3] Xiao-Yang Yang et al. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023. URL https://arxiv.org/abs/2306.06031.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. 2021. URL https://arxiv.org/abs/2005.11401.

[5] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://arxiv.org/abs/2004.04906.

[6] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference*, 2020. URL https://arxiv.org/abs/2004.12832.

[7] OpenAI. Gpt-4 technical report. 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

[8] Anthropic. Introducing the next generation of claude, 2024. URL https://www.anthropic.com/news/claude-3-family.

[9] OpenAI. Agents.md, 2025. URL https://agents.md/.

[10] 2025. URL https://www.anthropic.com/engineering/claude-code-best-practices.

[11] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982, 2019.

[12] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2025. URL https://arxiv.org/abs/2401.08281.

[13] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, July 2017.

[14] Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, Chenlin Zhou, Jiayi Mao, Tianze Xia, Jiafeng Guo, and Shenghua Liu. A survey of context engineering for large language models, 2025. URL https://arxiv.org/abs/2507.13334.

[15] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert, 2020. URL https://arxiv.org/abs/1901.04085.

[16] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, 2022.

[17] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, 2018.

[18] Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2590–2602, 2019.

[19] Wanjun Zhong, Junjie Huang, Qian Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. Reasoning over hybrid chain for table-and-text open domain question answering. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4531–4537, 2022.

[20] Kyle Lo, Lucy Lu Wang, Mark Neumann, et al. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4969–4983, 2020. URL https://arxiv.org/abs/1911.02782.

[21] Mehdi Ali Afzal Jaradeh and Sören Auer. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture (K-CAP)*, pages 243–246, 2019. URL https://doi.org/10.1145/3360901.3364435.

[22] Jason Priem and Heather Piwowar. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022. URL https://arxiv.org/abs/2205.01833.

[23] Xuanhe Zhou, Junxuan He, Wei Zhou, Haodong Chen, Zirui Tang, Haoyu Zhao, Xin Tong, Guoliang Li, Youmin Chen, Jun Zhou, Zhaojun Sun, Binyuan Hui, Shuo Wang, Conghui He, Zhiyuan Liu, Jingren Zhou, and Fan Wu. A survey of llm $\times$ data, 2025. URL https://arxiv.org/abs/2505.18458.

[24] Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *CoRR*, 2024.

[25] Ronald A Fisher. The design of experiments. 1949.

[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

[27] Maarten Jansen, Laurens Swinkels, and Weili Zhou. Anomalies in the china a-share market. *Pacific-Basin Finance Journal*, page 101607, 2021.

[28] Tim Bollerslev, Sophia Zhengzi Li, and Bingzhi Zhao. Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis*, 55:751–781, 2020.

[29] Ferhat Akbas, Ekkehart Boehmer, Chao Jiang, and Paul D. Koch. Overnight returns, daytime reversals, and future stock returns. *Journal of Financial Economics*, pages 850–875, 2022.

[30] Andrii Babii, Ryan T. Ball, Eric Ghysels, and Jonas Striaukas. Panel data nowcasting: The case of price-earnings ratios, 2023. URL https://arxiv.org/abs/2307.02673.

[31] Penggan Xu. Replication of reference-dependent preferences and the risk-return trade-off in the chinese market, 2025. URL https://arxiv.org/abs/2505.20608.

[32] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, pages 637–654, 1973.

A Detailed Experimental Results

A.1 Prompt Design

We detail the prompts used in the experiment to ensure consistency across conditions. The *AI Assistant Prompt* provides generic background information, the *QuantMind Structured Prompt* adds domain-specific context from the knowledge base, and the *LLM Judge Prompt* specifies the evaluation criteria for answer quality. Full prompt texts are shown below.

AI Assistant Prompt

Based on the provided research paper, generate helpful background information and relevant concepts that could assist in answering the following question, but do not provide direct answers.

QuantMind Structured Prompt

Retrieve and synthesize relevant concepts, methodologies, and contextual information from the quantitative finance literature to provide comprehensive background support for addressing the following question, ensuring no direct answers are provided.

LLM Judge Prompt for Answer Quality Evaluation

Please evaluate the quality of answers from the following 6 participants for the same academic question.

Paper Context: {paper_context}
Question: {question_text}
Answers: {formatted_answers}

Please rate each participant's answer on a scale of 1-5 (with one decimal point allowed) based on the following criteria:
Scoring Criteria:
1 point: Answer is completely irrelevant, incorrect, or incomprehensible
2 points: Answer is mostly irrelevant, showing little understanding of the question
3 points: Answer is partially relevant with some understanding but lacks depth
4 points: Answer is relevant with reasonable depth and clear logic
5 points: Answer is highly relevant, demonstrates deep analysis, rigorous logic, and unique insights
Note:
- For participants marked as "No response," please return "No response"
- Only rate participants who provided substantive answers

Please return the scoring results in the following format:
Participant 1: X.X points
Participant 2: X.X points
Participant 3: X.X points
Participant 4: X.X points
Participant 5: X.X points
Participant 6: X.X points

Please ensure that your evaluations are objective and impartial, based on the content quality, logical coherence, and relevance of each answer.

A.2 Counterbalanced Assignment in Within-subjects User Study

To mitigate order and sequence effects, we adopted a counterbalanced assignment scheme based on an incomplete Latin square design. This approach ensured that each assistance condition (Without AI

[W], Baseline AI Assistant [B], and QuantMind [C]) appeared with approximately equal frequency across participants and papers, while preventing systematic biases due to task ordering. Table 2 illustrates the assignment matrix, where rows correspond to participants and columns to the six curated papers.

Participant	Paper 1	Paper 2	Paper 3	Paper 4	Paper 5	Paper 6
Subject 1	W	W	B	B	C	C
Subject 2	W	B	B	C	C	W
Subject 3	B	B	C	C	W	W
Subject 4	B	C	C	W	W	B
Subject 5	C	C	W	W	B	B
Subject 6	C	W	W	B	B	C

Table 2: Counterbalanced assignment under an incomplete Latin square. Rows denote participants, columns denote papers, and entries indicate assistance conditions: Without AI (W), Baseline AI Assistant (B), or `QuantMind` (C).

A.3 Selected Research Papers and Question Design

Table 3 presents the mapping between paper IDs and the corresponding evaluation questions used in the user study. For clarity, the six paper IDs correspond to the following works. **P1**: Anomalies in the China A-share market [27]; **P2**: Good volatility, bad volatility, and the cross section of stock returns [28]; **P3**: Overnight returns, daytime reversals, and future stock returns [29]; **P4**: Panel data nowcasting: The case of price–earnings ratios [30]; **P5**: Replication of Reference-Dependent Preferences and the Risk-Return Trade-Off in the Chinese Market [31]; and **P6**: The pricing of options and corporate liabilities [32].

Paper ID	**Question ID**	**Question**
1	1	Based on the paper, which factors do you think perform strongest in the Chinese A-share market? How would you explain the source of these factor returns?
1	2	Which factors underperform in the paper? In your research or practice, how would you improve these factors or explore their potential value?
1	3	Apart from short-selling constraints, state-owned enterprises, and market reforms, what other China-specific factors might affect factor performance?
1	4	How do you assess the robustness and practicality of a factor in quantitative investment practice?
1	5	Does this paper provide clues about future research directions for factors? What new ideas do you have for future research?
1	6	If you were to construct a new factor from scratch, how would you use the methodology in this paper to validate its effectiveness?
2	7	What is the core idea of the paper?
2	8	What data is needed to construct trading signals? What are the time frequencies of these data?
2	9	Which stock market data was used for backtesting in the paper? What metrics were selected for the backtest?
2	10	Pseudocode for calculating the simplest version of the factor?

Table Continued

Paper ID	Question ID	Question
2	11	What is the source of returns for this factor? What type of factor is it?
2	12	Can other factors be derived from this factor?
3	13	What is the core idea of the paper?
3	14	What data is needed to construct trading signals? What are the time frequencies of these data?
3	15	Which stock market data was used for backtesting in the paper? What metrics were selected for the backtest?
3	16	Pseudocode for calculating the simplest version of the factor?
3	17	What is the source of returns for this factor? What type of factor is it?
3	18	Can other factors be derived from this factor?
4	19	What is the approach of the study for constructing predictive factors?
4	20	How is the sparse group LASSO (sg-LASSO) regularization method defined in the paper? Provide the formula and explain the parameters.
4	21	What are the inputs of the machine learning models used in the paper?
4	22	What are the limitations of the models in the paper?
5	23	What is the core idea of the paper?
5	24	What data is needed to construct trading signals? What are the time frequencies of these data?
5	25	Which stock market data was used for backtesting in the paper? What metrics were selected for the backtest?
5	26	Write pseudocode for calculating the simplest version of the factor.
5	27	What is the source of returns for this factor? What type of factor is it?
5	28	Can other factors be derived from this factor?
6	29	Describe the basic assumptions and derivation idea of the Black-Scholes option pricing model.
6	30	What parameters are set in the model? Please explain their economic meaning.
6	31	When the market exhibits a "volatility smile," how does the Black-Scholes model explain this deviation?
6	32	Provide the formula used for pricing with the Black-Scholes model.
6	33	What are the limitations of the Black-Scholes model in actual markets?
6	34	Assuming you are an options trader, how would you use the Black-Scholes model to identify overvalued or undervalued options?

Table 3: Mapping of Questions to Paper IDs.

A.4 Raw Experimental Results

A.4.1 Per-Participant Results

Table 4 presents the UX rating result. The subject and paper ID columns in this table correspond to the respective columns in Table 2.

subject	paper ID	assistance lvl	relevance	accuracy	helpfulness	clarity	average
1	3	2	3	4	4	3	3.5
1	4	2	4	4	4	3	3.75
1	5	3	5	4	5	2	4
1	6	3	5	4	5	3	4.25
2	2	2	4	5	4	3	4
2	3	2	5	4	4	4	4.25
2	4	3	4	4	4	3	3.75
2	5	3	4	4	5	4	4.25
3	2	2	2	2	1	1	1.5
3	1	2	4	4	2	2	3
3	3	3	2	3	2	2	2.25
3	4	3	4	3	4	3	3.5
4	1	2	4	4	3	4	3.75
4	6	2	5	4	4	5	4.5
4	2	3	5	5	5	5	5
4	3	3	5	5	5	5	5
5	5	2	4	5	5	5	4.75
5	6	2	4	3	4	3	3.5
5	1	3	5	4	5	4	4.5
5	2	3	5	5	5	5	5
6	4	2	4	5	4	5	4.5
6	5	2	4	4	4	4	4
6	1	3	3	4	4	3	3.5
6	6	3	5	4	4	5	4.5

Table 4: Raw results about UX rating.

Table A.4.1 presents the accuracy scores for user responses, categorized by question. The first two columns, "Question ID" and "Paper ID," identify each question and its corresponding source paper. The subsequent columns, labeled "Score 1" through "Score 6," report the accuracy scores, providing a quantitative measure of the correctness of the answers.

question ID	paper ID	participant 1	participant 2	participant 3	participant 4	participant 5	participant 6
1	1	3.5	4.0	4.5	4.0	3.5	4.0
2	1	4.0	3.5	5.0	3.5	3.0	4.0
3	1	3.0	3.5	4.5	4.0	4.0	3.5
4	1	3.5	3.0	4.5	4.0	3.5	4.5
5	1	3.0	2.5	4.5	4.0	3.5	3.5
6	1	3.5	3.0	5.0	4.5	3.0	4.5
7	2	1.0	3.0	4.0	3.5	4.5	4.0
8	2	3.0	3.5	4.5	4.0	2.0	3.0
9	2	4.0	4.5	4.5	5.0	4.0	2.0
10	2	3.0	4.0	4.5	4.0	4.5	1.0
11	2	4.0	3.5	4.0	4.5	4.0	2.0
12	2	2.0	3.5	4.0	5.0	4.5	1.0
13	3	3.5	4.0	4.5	4.0	4.5	3.5
14	3	3.0	4.0	4.5	5.0	2.0	3.0
15	3	4.0	4.5	3.5	5.0	2.0	3.0
16	3	4.0	3.5	4.5	5.0		2.0
17	3	4.0	3.5	4.5	5.0	2.0	3.0
18	3	4.0	4.5	3.5	5.0	1.0	2.0
19	4	4.0	4.5	4.5	2.5	3.0	2.0
20	4	4.5	4.0	4.5	4.0	2.0	3.5
21	4	4.0	4.5	3.5	2.0	1.0	4.0
22	4	4.0	3.5	4.0	3.0	2.0	3.5
23	5	4.0	4.5	4.0	4.5	3.0	4.0
24	5	4.0	4.5	3.0	4.0	2.0	3.5
25	5	4.0	4.5	3.5	5.0	3.5	4.5
26	5	5.0	4.0	3.0	3.0		2.0
27	5	4.5	4.0	4.5	3.5	4.0	4.0
28	5	4.0	3.5	3.0	3.5	4.5	4.0
29	6	5.0	4.0	4.5	3.5	4.5	4.0
30	6	4.5	4.0	4.5	4.0	4.0	5.0
31	6	5.0	3.0	4.5	3.5	3.0	4.5
32	6	5.0	4.5	4.5	3.5	2.0	5.0
33	6	5.0	3.5	4.0	4.5	3.0	4.5
34	6	5.0			3.5	4.0	4.5

Table 5: Raw result of Answers Quality.

A.4.2 Aggregated Summary Statistics

	Answer Quality			User Experience		
Metric	Estimate	t-value	p-value	Estimate	t-value	p-value
Fixed Effects						
Intercept (Lv1/Lv2)	3.106	15.60	5.25×10^{-8}	3.750	11.96	1.19×10^{-5}
Lv2 vs Reference	0.711	6.07	6.97×10^{-9}	–	–	–
Lv3 vs Reference	1.139	9.76	$< 2 \times 10^{-16}$	0.375	1.70	.108
ANOVA						
F-value	48.44		$< 2.2 \times 10^{-16}$	2.89		.1075
DenDF	187.0			17.0		
Random Effects (Variance)						
Subject ID	0.1485			0.4441		
Paper ID	0.0475			0.0000		
Residual	0.4509			0.2923		

Table 6: Summary of mixed-effects model results.

Model Specifications:
Answer Quality: REML criterion at convergence = 434.7, N = 200
User Experience: REML criterion at convergence = 50.1, N = 24
Note:
Reference levels: Lv1 (Without AI) for Answer Quality; Lv2 (Generic AI) for User Experience.
Significance codes: $^{***}p < .001$; User experience model showed singular fit (Paper ID variance = 0).

	Answer Quality		User Experience	
Comparison	Estimate	p-value	Dimension	p-value
Pairwise Comparisons			**Dimension Analysis**	
Lv1 - Lv2	-0.711	$< .0001$	Relevance	.2053
Lv1 - Lv3	-1.139	$< .0001$	Accuracy	.7577
Lv2 - Lv3	-0.428	.0009	Helpfulness	.0028
			Clarity	.5984
Residuals Distribution			**Descriptive Stats**	
Minimum	-3.094		Lv2 Mean	3.75
1Q	-0.615		Lv3 Mean	4.12
Median	0.139		Improvement	+0.37
3Q	0.623			
Maximum	3.040			

Table 7: Post-hoc comparisons and descriptive statistics.

B Discussion

Limitations. `QuantMind` faces several constraints: (i) potential retrieval drift under temporal or domain shifts; (ii) a limited sample size in the user study; and (iii) incomplete coverage beyond equities, with credit, rates, and derivatives requiring broader taxonomies and schemas.

Ethics. All sources are restricted to publicly available documents with licensing/ToS respected. User-study participants provided informed consent. Only minimal telemetry (timing and UX ratings) was logged, and all reported results are aggregated.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: Yes, the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Yes, we discuss the limitations in Appendix B.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [NA]
   Justification: No new theorems are proposed. Because this paper is a LLM-based framework paper with a comprehensive evaluation.
   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Yes. Study protocol, metrics, and analysis plan are detailed in Section 4.
   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: Yes. In the final camera-ready version we will release all data and code.
   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: Yes. The paper specifies all materials, tasks, and evaluation metrics in Section 4.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: Statistical significance was assessed with linear mixed-effects models, post-hoc tests, and effect sizes with confidence intervals. Variability and small-N limitations are explicitly noted.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: Details on hardware specifications, indexing, and query execution time are provided in the Appendix A to ensure reproducibility.
   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: Yes, we have reviewed the NeurIPS Code of Ethics and ensured full compliance throughout our research process.
   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: Positive impacts (faster, auditable research) and risks (misinterpretation, licensing) discussed in Ethics in Appendix B.
    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
    Answer: [NA]
    Justification: No risky pretrained models or scraped image datasets are released; source ToS respected.
    Guidelines:
    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [Yes]
    Justification: All papers used in the user study are properly cited and listed in Appendix A.3.
    Guidelines:
    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [Yes]
    Justification: Documentation for indices, prompts, and taxonomies will accompany anonymized release.
    Guidelines:
    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [Yes]
    Justification: Not crowdsourcing; small-sample user study with instructions, compensation (if any), and consent in supplement.
    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [NA]
    Justification: Institutional requirements vary; if applicable, we will disclose IRB/ethics review in camera-ready while keeping anonymity.
    Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
    Answer: [Yes]
    Justification: LLM components (prompting, tool-augmented retrieval, and evaluation scaffolds) are part of the core system and are described with pointers to Sections 3-4 and Appendix A.
    Guidelines:
    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.