CFQA: A Chinese Financial Question Answering Benchmark From Corporate Annual Reports

Tianning Zhu[1], **Mo Liu**[2], **Murathan Kurfalı**[3]
[1] Department of Linguistics and Philology, Uppsala University, Sweden
[2] Business Department, CITIC Securities, Beijing, China
[3] RISE Research Institutes of Sweden, Stockholm, Sweden
zackzhu00@foxmail.com, liumo@citics.com, murathan.kurfali@ri.se

Abstract

We present CFQA, a Chinese financial question answering benchmark constructed from 50 publicly listed companies' annual reports spanning 2023–2025. The benchmark comprises 500 questions, derived by applying 10 question templates to each source document, and covers five categories: fact extraction, enumeration, comparative calculation, judgment verification, and reasoning analysis. All gold-standard answers are manually annotated and grounded in the source reports. To illustrate benchmark utility, we evaluate a retrieval-augmented generation (RAG) system against a no-retrieval baseline, and introduce a rule-based consistency detector that distinguishes fabricated content from other error types. RAG improves average answer accuracy from 7.53% to 8.07%, with the most consistent gains observed in fact extraction and judgment verification tasks for domain-adapted models. Crucially, by decoupling exact-match accuracy from evidence-support judgments, our detector reveals that despite low absolute scores, RAG architectures successfully constrain model confabulation, exhibiting remarkably low true fabrication rates. However, performance gains in higher-order cognitive tasks, such as comparative calculation and reasoning analysis, remain non-significant across evaluated models, highlighting the boundaries of current retrieval-augmented systems in complex financial reasoning. The dataset, annotation guidelines, and evaluation code are publicly released.

Keywords: Chinese Financial QA, Benchmark, Financial Reports, Retrieval-Augmented Generation, Error Analysis, faithfulness evaluation

1. Introduction

Financial document understanding is a long-standing challenge in natural language processing. Corporate annual reports are a central form of financial narrative but they are particularly demanding because of unique and challenging characteristics: multimodal content mixing text with charts and tables, numerical values (amounts, ratios, percentages, dates), high timeliness and dynamic interrelations, and extensive use of specialized accounting terminology. Moreover, despite growing interest in financial NLP, the majority of existing benchmarks focus on English documents, whereas high-quality benchmarks for Chinese annual-report question answering remain limited.

To address this shortcoming, we present *CFQA*[1], a Chinese financial question answering benchmark constructed from 50 publicly available annual-report PDFs (2023–2025). CFQA contains 500 question–answer instances, created by instantiating 10 question templates for each report (10 × 50). The benchmark covers five question types: fact extraction, list enumeration, comparative calculation, judgment verification, and reasoning analysis. All gold answers are manually annotated. CFQA is designed to support research on retrieval-augmented generation (RAG) and systematic error analysis for Chinese financial documents.

The main contributions of this paper are: (i) a 500-question Chinese financial QA benchmark with five question types and manually annotated, evidence-grounded answers; (ii) a template-driven construction methodology that ensures type diversity and reproducibility; (iii) a rule-based consistency detector that categorises answer errors as fabrication, retrieval gap, or calculation error; and (iv) a comparative evaluation of six open-source language models under baseline (no-retrieval) and RAG conditions, demonstrating the utility of the benchmark for measuring retrieval benefit across question types.

2. Related Work

Research on Chinese financial NLP has expanded in recent years, supported by the development of both general Chinese reading-comprehension resources and finance-specific corpora. CMRC 2018 established a widely used Chinese machine reading comprehension benchmark, but it is built from Wikipedia paragraphs and therefore is not a financial-domain dataset (Cui et al.,

[1] *Our dataset is avaiable at: https://github.com/zhutianning/Hallucination-detection-for-RAG*

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 28–38

2019). In the financial domain, Zheng et al. introduced Doc2EDAG, which includes a large-scale dataset of Chinese financial announcements for document-level event extraction (Zheng et al., 2019). For sentiment-focused research, Yuan et al. presented a target-based sentiment annotation corpus for Chinese financial news (Yuan et al., 2020). More recently, OmniEval proposed a financial-domain RAG benchmark spanning multiple task classes, but it is oriented toward evaluating RAG systems rather than document-grounded question answering over complete annual reports (Wang et al., 2025). Compared with these resources, CFQA focuses specifically on full regulatory filings, covers five template-driven question types, and requires answers to be explicitly grounded in the source document or marked as absent.

3. Benchmark Construction

The system is divided into five major modules: Data Extraction (MinerU) → Indexing/Retrieval (text/tables/images) → RAG Generation → Hallucination Detection (rule-based). We build a knowledge extraction pipeline, using MinerU to extract structured content from financial report PDFs (tables converted to Markdown, images, and hierarchical text—headings/paragraphs); implement retrieval (text/tables/images) and the RAG workflow (Lewis et al., 2020), ensuring that generation can cite specific evidence locations (page numbers/table cells/images) (Suri et al., 2025). The core objective is to automatically determine whether each key assertion in the generated answer is supported by the retrieved evidence.

3.1. Source Document Collection

We collect 50 annual-report PDFs from publicly listed Chinese companies covering fiscal years 2023–2025. To ensure a representative and diverse evaluation of financial natural language processing, these companies were purposefully selected from major economic sectors, such as internet and financial industry. We release the question–answer data and document metadata, while the original PDFs remain available from their public disclosure channels.

3.2. Annotation

To evaluate the hallucination detection capabilities of RAG systems in financial scenarios, we constructed a specialized, high-complexity benchmark. All gold-standard answers and question formulations were manually annotated by a professional investment advisor from a Chinese securities company. The expert annotation process was strictly governed by our comprehensive project guidelines (which we released on our GitHub repository), which were specifically designed to make the benchmark hallucination-detection friendly.

3.3. Document Processing Pipeline

We employ a dual-pipeline data processing framework for extracting and merging content from PDF annual reports. The process begins with the parallel execution of two complementary parsing strategies on source documents: PyMuPDF ensures the integrity and character-level accuracy of the raw text stream, while MinerU performs layout analysis to precisely extract structured elements such as tables, charts, and hierarchical headings. A subsequent alignment and merging module uses a dynamic programming algorithm to map MinerU's structured blocks with PyMuPDF's page-level text, resolving parser-induced page offsets. During merging, redundant headers and footers are removed, and long texts are re-chunked based on semantic completeness, resulting in a cleaned, deduplicated JSON corpus from 50 reports that serves as the external knowledge source for retrieval-augmented generation.

To enable the retrieval of non-textual content, a standardized text-centric processing pipeline is implemented. Tables and images extracted by MinerU are converted into HTML and image snapshots. For elements lacking captions, a multimodal vision-language model (Qwen2.5-VL-32B-Instruct) generates descriptive textual captions. All content is then consolidated into page-based Markdown strings, where visual elements are embedded via their textual descriptions and file paths. These unified chunks are vectorized using the BGE-M3 text embedding model, which encodes the semantic information of visuals through their captions. During retrieval, the system performs a cosine similarity search over this text-based index. Retrieved chunks provide the LLM with descriptive captions and asset paths, thereby enabling a "text-as-proxy" paradigm where all reasoning about visual content is mediated through pre-generated text, ensuring efficiency and compatibility with standard text-RAG architectures.

3.4. Data Generation

To construct the dataset, we utilized a structured pool of 50 templates, allocating 10 templates to each of the five question categories. For each annual report, candidate questions were programmatically generated by instantiating these templates with extracted metadata, such as the company name and fiscal year. Finally, the generation script sampled from this candidate pool to con-

Type	Design Purpose	Example Question Form
Fact Extraction	Retrieve a specific value, name, or date from a financial statement or disclosure.	What was [Company]'s total operating revenue in FY[year]?
Enumeration	List all members of a specified set disclosed in the report.	List the names and shareholding ratios of the top-ten shareholders of [Company] as of [year-end].
Comparative Calculation	Compute a year-on-year change or ratio using values from the report.	Calculate the year-on-year growth rate of [Company]'s net profit attributable to shareholders in FY[year].
Judgment Verification	Determine whether a stated event or condition is disclosed; extract details if so.	Did [Company] implement an equity incentive plan in FY[year]? If so, what were the number of grant recipients and the number of shares granted?
Reasoning Analysis	Attribute a trend or evaluate a causal explanation using evidence from the report.	What are the primary factors cited by management to explain the change in [Company]'s operating margin in FY[year]?

Table 1: Question types used in our benchmark.

struct the final benchmark, ensuring a strictly balanced distribution across all question types.

This study employed the open-source Qwen3-8B-Instruct model (Qwen Team, 2024) to facilitate question generation, resulting in a curated set of 500 non-repetitive questions derived from 50 annual reports (2023–2025). The dataset is characterized by a strictly balanced distribution across five critical question types—*fact extraction*, *enumeration*, *comparative calculation*, *judgment verification*, and *reasoning analysis*—each constituting 20% of the total (see Table 1). This design ensures a multidimensional assessment of system performance on retrieval precision, structured extraction, temporal comparison, conditional logic, and causal reasoning.

The raw set contained placeholder metadata and invalid answers, rendering it unsuitable for rigorous evaluation. In contrast, our refined set incorporates authentic filenames and page references, features precise, verifiable answer formulations, and is explicitly structured to support hallucination detection. Questions demand multi-dimensional analysis, often requiring the integration of data across income statements, balance sheets, and cash flow statements. The use of specific numerical queries and conditional qualifiers (e.g., "if any") enhances clarity and testability.

Our design is intrinsically "hallucination-detection friendly." All answers are grounded in reported evidence, ensuring verifiability. Explicit numerical requests simplify the detection of fabrication, while judgment verification questions test logical integrity. Enumeration tasks target completeness of answers, and calculation questions permit direct mathematical verification. This dataset provides a robust framework for quantifying and analyzing hallucinations in later work.

4. Evaluation Metrics

One of the contributions of CFQA is a rule-based consistency detector that supports systematic error analysis on model outputs. The detector does not label answers simply as correct or incorrect; instead, it categorises each answer along several dimensions to distinguish fabrication-type errors from other failure modes. This distinction strengthens our benchmark: a system that retrieves no evidence and fabricates an answer fails differently from a system that retrieves correct evidence but makes an arithmetic error. Distinguishing these error types helps us reveal the apparent hallucination rates and provides diagnostic information. We regard a generated answer as a hallucination if it (a) directly contradicts the retrieved evidence, or (b) makes a specific factual claim about an entity or value that does not appear anywhere in the evidence. Errors that arise from incomplete retrieval, incorrect calculation, or paraphrase misalignment are categorised as non-hallucination errors.

1. **Numerical Verification**: Given the high sensitivity to numerical data in financial QA, we extract numerical expressions (including percentages, decimals, etc.) from both the answer and the evidence. A relative error threshold ($\tau = 0.01$) is used for tolerance matching. If a key numerical value in the answer cannot be matched to a corresponding value within the error range in the evidence, a "numerical unsupported" signal is triggered.

2. **Text Coverage**: To measure the faithfulness of non-numerical facts, we normalize the answer and evidence (remove punctuation, unify character forms) and calculate the coverage ratio of the answer's token set within the evidence's token set. A coverage ratio below 0.4

is flagged as a low-coverage signal, suggesting the answer may contain statements not present in the evidence. Similarly, low coverage can also stem from evidence truncation, misaligned retrieval, or paraphrasing in the answer. Thus, this dimension alone is not equivalent to "hallucination" but indicates a risk of "lack of evidential support."

3. **Reference Consistency**: We verify whether the source (filename and page number) cited in the answer matches the metadata of the evidence used for verification. Filename matching allows for a degree of fuzziness (e.g., removing date prefixes), and a deviation of ± 2 pages is permitted to tolerate pagination errors from PDF parsing. This module identifies "citation errors/evidence misalignment," which are primarily retrieval or citation errors and should not be classified as hallucinations.

4. **Calculation Verification**: For comparative calculation questions (e.g., growth rate, change magnitude), the detector uses regular expression templates to extract computational expressions from the answer and attempts to verify them against relevant numerical values in the evidence (implemented as heuristic rules, returning medium-confidence outcomes to avoid over-assertion). By definition, if the original data exists in the evidence but the calculation result is wrong, it should be classified as a calculation error (non-hallucination). Only when the answer uses original numerical values not present in the evidence or asserts a conclusion directly conflicting with the evidence should it be considered a "hallucination/contradiction."

5. **List Completeness vs. Fabricated Items**: For list/enumeration tasks, mere "incompleteness" is not equivalent to hallucination. We categorize list-related errors into two types:

   - Fabricated Item (Hallucination): The answer contains list items that do not exist in the evidence (e.g., fictitious customer or product names). This is a typical hallucination.
   - Omitted Item (Incompleteness): All items listed in the answer can be found in the evidence, but the answer fails to cover all items that should be listed according to the evidence. This is an "incomplete retrieval/answer" and is a non-hallucination error.

   In practice, the detector checks both whether answer items match those in evidence (for fabrication) and whether evidence items are covered by the answer (for omission) to avoid misclassifying "under-listing" as "fabrication."

Scores from the five dimensions are combined via fixed weights (numerical: 0.30; text coverage: 0.25; reference: 0.20; calculation: 0.125; list: 0.125) to produce a composite confidence score, from which a three-way classification is derived: Evidenced, Partially Evidenced, or No Evidence. This final label is intended as a diagnostic guide rather than a ground-truth correctness judgement.

5. Experimental Setup

To demonstrate the utility of CFQA for benchmarking retrieval-based systems, we compare a RAG pipeline against a no-retrieval baseline across six open-source language models. All experiments use the 500-question CFQA test set.

Models. We evaluate the following models: Qwen3-8B, Qwen3-14B, Mixtral-8×7B, Mixtral-8×22B, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. These models were selected to cover a range of parameter scales and represent both Chinese-pretrained (Qwen) and multilingual (Mixtral, Llama) architectures, providing a diverse picture of how retrieval benefit varies across model families.

Baseline. The vanilla (no-retrieval) baseline receives no document context. It uses a maximum output length of 512 tokens. Its prompt instructs the model to answer based on its pretraining knowledge of Chinese financial reporting conventions and to explicitly acknowledge when specific values are not known.

RAG Pipeline. The retrieval component uses the BAAI/bge-m3 embedding model to encode document chunks as 1024-dimensional vectors. The knowledge corpus is indexed with page-level chunking: each page forms a distinct chunk paired with its source filename and page number. At inference time, the top-k = 5 chunks most similar to the query (cosine similarity) are retrieved from an in-memory vector store. The RAG system uses a temperature of 0.2 and a maximum output length of 1024 tokens; its prompt instructs the model to ground answers in the retrieved passages and to return a structured JSON object containing the answer text and a provenance record (filename and page number).

Evaluation metrics Our evaluation framework assesses system performance through two complementary layers: Answer Correctness and Hallucination Detection. First, Answer Correctness,

Model	Baseline Acc.	RAG Acc.	Δ
Qwen3-8B	0.096	0.120	+0.024
Qwen3-14B	0.086	0.134	+0.048
Mixtral-8×7B	0.098	0.066	-0.032
Mixtral-8×22B	0.086	0.102	+0.016
Llama-3.1-8B-Instruct	0.066	0.040	-0.026
Llama-3.1-70B-Instruct	0.020	0.022	+0.002

Table 2: Overall accuracy on CFQA (500 questions) for baseline vs RAG settings.

defined as Accuracy, measures the alignment between the generated output and the manually annotated Gold Standard. To mitigate the risk of underestimating semantically correct answers expressed in varied surface forms, this correctness score eschews strict exact-matching in favor of flexible heuristics, incorporating numeric matching with tolerance and text coverage thresholds. Second, independently of the gold reference, we evaluate generation faithfulness via a deterministic rule-based Hallucination Detector that quantifies answer–evidence consistency. This module applies multi-dimensional checks to classify outputs into three evidence-support risk categories: Evidenced, Partially Evidenced, or No Evidence. By explicitly decoupling the gold-standard accuracy from the evidence-support judgment, our methodology strictly isolates fabrication-type hallucinations from non-hallucinatory errors, such as incomplete retrieval or calculation mistakes, thereby providing a highly rigorous and nuanced assessment of RAG performance.

6. Results

6.1. Overall Performance

Table 2 reports the baseline and RAG accuracy for each model. RAG improves accuracy for four of the six models. The mean accuracy across models increases from 7.53% (baseline) to 8.07% (RAG), confirming that retrieved evidence from source reports provides a modest overall benefit on this benchmark. The two models for which RAG does not improve—Mixtral-8×7B and Llama-3.1-8B-Instruct—show a slight accuracy drop, suggesting that these smaller multilingual models may struggle to faithfully incorporate long retrieved Chinese passages into their outputs. The Qwen-series models benefit the most from retrieval, consistent with their stronger Chinese-language pretraining.

6.2. Experiments on Category-specific Subsets

Table 3 shows per-category accuracy for each model under both conditions. For most models, RAG provides noticeable gains on fact extraction and judgment verification, moderate gains on comparative calculation, and minimal to no gains on reasoning analysis and enumeration. For fact extraction, accuracy improves from near 0% in the baseline to up to 7% with RAG for most models, with Qwen3-8B reaching 7% and Qwen3-14B reaching 5%. Judgment verification shows the strongest absolute accuracy overall (up to 33% RAG accuracy for Qwen3-8B and 32% for Qwen3-14B), as these questions require a binary determination that is often explicitly stated in the source document. Conversely, enumeration performance was surprisingly higher in the baseline setting for some models, reaching 23.0% for Llama-3.1-8B-Instruct, 17.0% for Mixtral-8x7B, and 14.0% for Qwen3-8B, but this performance generally decreased or remained stagnant under RAG conditions. Retrieval helps more on tasks with clearly stated answers, such as fact extraction and judgment verification. By contrast, enumeration and reasoning are harder because they require more complete coverage and better integration of evidence. Reasoning analysis shows minimal to no improvement, suggesting that retrieving relevant evidence alone is not enough for these tasks.

6.3. Error Analysis

Our error analysis using the multi-dimensional detector reveals that the majority of RAG failures stem not from outright fabrications, but from incomplete extraction or semantic misalignment, categorized predominantly as partially evidenced. For example, Llama-3.1-8B and Mixtral-8x7B generated partially evidenced answers for 75.4% and 65.2% of queries, respectively. Conversely, severe Unsupported hallucinations—indicating direct contradictions or fabricated claims—are concentrated in structurally and computationally demanding tasks such as Enumeration and Comparative Calculation. In Qwen3-14B, 51.0% of enumeration answers and 52.0% of calculation answers were strictly unsupported, frequently triggered by the model hallucinating non-existent list items or inventing arithmetic results. Meanwhile, explicit knowledge tasks like Judgment Verification and Fact Extraction yielded the highest proportions of fully Supported outputs.

Model	Fact Extraction	Comp. Calculation	Jud. Verification	Reasoning	Enumeration
Qwen3-8B	0.020 / **0.070**	0.000 / **0.030**	0.260 / **0.330**	**0.060** / 0.040	**0.140** / 0.130
Qwen3-14B	0.010 / **0.050**	0.000 / **0.030**	0.240 / **0.320**	0.070 / **0.080**	0.110 / **0.190**
Mixtral-8×7B	0.000 / **0.010**	**0.000** / **0.000**	**0.270** / 0.250	**0.050** / 0.020	**0.170** / 0.050
Mixtral-8×22B	0.000 / **0.020**	0.000 / **0.020**	0.260 / **0.310**	**0.080** / 0.050	0.090 / **0.110**
Llama-3.1-8B-Instruct	**0.000** / **0.000**	**0.000** / **0.000**	0.090 / **0.150**	**0.010** / **0.010**	**0.230** / 0.040
Llama-3.1-70B-Instruct	**0.000** / **0.000**	**0.000** / **0.000**	0.030 / **0.090**	**0.020** / 0.000	**0.050** / 0.020

Table 3: Performance comparison between Baseline and RAG systems across question types for different models. Each cell shows *Baseline / RAG*. The better score is boldfaced and underlined.

7. Discussion

The Asymmetric Impact of Retrieval Augmentation. A key finding of our evaluation is the asymmetric impact of retrieval augmentation across different model architectures. While RAG significantly improved the overall accuracy of domain-adapted models like Qwen3-14B ($p=0.000126$), it actively degraded the performance of specific multilingual architectures. Most notably, Mixtral-8x7B experienced a statistically significant decrease in accuracy from 9.8% to 6.6% ($p=0.012$), and Llama-3.1-8B exhibited a sharp, significant drop in enumeration accuracy ($p<0.001$). Notably, the largest model evaluated, Llama-3.1-70B, exhibited anomalously low accuracy (2%). This does not necessarily indicate severe fabrication. Instead, many of its responses were partially grounded in evidence but were written as open-ended analyses rather than extractive answers, often with vague qualifiers such as "approximately" and "possibly," which reduced agreement with the gold answers. This performance degradation suggests that feeding long, dense Chinese financial passages into the context windows of these smaller multilingual models may overwhelm their instruction-following capabilities, leading to distracted generations where relying on their internal parametric knowledge occasionally yielded better heuristic guesses.

Cognitive Boundaries and Faithfulness Evaluation. Furthermore, the varied improvements across cognitive categories highlight the boundaries of standard RAG pipelines. While retrieval augmentation successfully surfaces explicit facts to drive statistically significant accuracy gains in Judgment Verification for both Qwen3-8B ($p=0.023$) and Qwen3-14B ($p=0.013$), improvements in tasks requiring multi-step reasoning or comparative calculation were universally non-significant across all six evaluated models. This demonstrates that merely supplying correct financial data does not endow the LLM with the necessary symbolic logic or arithmetic capabilities. Crucially, by decoupling gold-standard accuracy from our multi-dimensional evidence-support judgments, we prove that a large portion of "incorrect" answers are non-hallucinatory processing failures. Despite achieving low absolute accuracy scores, RAG models exhibited remarkably low true hallucination (fabrication) rates—such as 7.4% for Qwen3-14B and 8.2% for Qwen3-8B, confirming that multimodal RAG remains highly effective at constraining model confabulation in the financial domain.

8. Conclusion

Our study introduces CFQA, a benchmark for Chinese financial annual-report question answering, and presents baseline and retrieval-augmented results with detailed error analysis. Our evaluation demonstrates that RAG effectively improves answer accuracy for well-adapted models, most notably increasing Qwen3-14B's accuracy from 8.6% to 13.4%, and Qwen3-8B's accuracy from 9.6% to 12.0%. McNemar's tests reveal that the statistical significance of these improvements varies by model: Qwen3-14B achieved a highly significant overall gain ($p = 0.000126$), with notable task-specific significance in judgment verification ($p=0.013$) and enumeration ($p=0.043$). For Qwen3-8B, while the overall gain was not statistically significant ($p=0.082$), it still achieved statistically significant improvements specifically in the judgment verification task ($p=0.023$). Conversely, models like Mixtral-8x7B showed a statistically significant decrease in accuracy under RAG conditions ($p=0.012$). Furthermore, the system successfully controls hallucination rates, with RAG models exhibiting low fabrication rates ranging from 7.4% (Qwen3-14B) to 16.0% (Mixtral-8x7B). The proposed rule-based hallucination detector enables automated fact-checking and identifies evidence-contradictory claims with high precision, consistent with recent benchmarks (Bang et al., 2025; Sok et al., 2025). By explicitly addressing table and chart data alignment (Yang et al., 2025; Suri et al., 2025), this work helps address a gap in existing evaluations and offers a practical framework for developing more reliable domain-specific intelligent assistants.

Limitations

Despite the effectiveness of our RAG framework, several limitations warrant consideration. First, the evaluation is confined to financial annual reports and has a modest sample size ($N = 500$), which may limit generalization to open-domain contexts. Second, statistical analysis reveals that across all evaluated models, improvements in tasks requiring higher-order logic, such as "Comparative Calculation" and "Reasoning Analysis" consistently failed to reach statistical significance. This suggests that retrieval augmentation alone is insufficient to address LLMs' inherent deficits in multi-step logical reasoning and arithmetic operations. Furthermore, the system's approach to fusion remains preliminary, lacking deep semantic alignment between charts and text. The trade-off between retrieval precision and recall also requires optimization, as overly aggressive filtering may discard critical context.

Finally, our evaluation primarily relies on deterministic, rule-based metrics. While effective for verifiable factual consistency, this approach may not fully capture semantically correct answers with varied surface forms, potentially underestimating semantic faithfulness. Incorporating LLM-as-a-Judge frameworks could complement our current evaluation.

Ethics Statement

This research complies with ethical guidelines. All experimental data are from publicly available annual reports of listed companies and contain no Personally Identifiable Information (PII) or unauthorized trade secrets. Although our system improves financial fact-checking accuracy, given the probabilistic nature of Large Language Models, its outputs are for auxiliary reference only. They are not legally binding investment advice or audit conclusions.

9. Bibliographical References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. Benchmarking retrieval-augmented multimodal generation for document question answering. *arXiv preprint arXiv:2505.16470*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jeanie Genesis. 2025. Retrieval-augmented text generation: Methods, challenges, and applications. *Preprints*.

Ziyu Gong, Yihua Huang, and Chengcheng Mai. 2025. Mmrag-docqa: A multi-modal retrieval-augmented generation method for document question-answering. *arXiv preprint arXiv:2508.00579*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. 2025. Unidoc-bench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2407.10671*.

Channdeth Sok, David Luz, and Yacine Haddam. 2025. Metarag: Metamorphic testing for hallucination detection in rag systems. *arXiv preprint arXiv:2509.09360*.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of NAACL*, pages 6088–6109.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pages 5737–5762.

Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. 2025. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*.

Chaofa Yuan, Yuhan Liu, Rongdi Yin, Jun Zhang, Qinling Zhu, Ruibin Mao, and Ruifeng Xu. 2020. Target-based sentiment annotation in chinese financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5040–5045, Marseille, France. European Language Resources Association.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. In *Proceedings of EMNLP-IJCNLP*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

A. Design Characteristics of Each Question Type

A.1. Fact Extraction

- **Original (Chinese):** 2022 年中国人保的营业总收入是多少亿元?
- **Transliteration (Pinyin):** 2022 nián Zhōngguó Rénbǎo de yíngyè zǒng shōurù shì duōshǎo yì yuán?
- **Translation (English):** What was the total operating revenue of PICC in 2022, in hundreds of millions of yuan?

A.2. List Enumeration

- **Original (Chinese):** 列举 2022 年中国人保前五大客户的名称及其销售收入占比。
- **Transliteration (Pinyin):** Lièjǔ 2022 nián Zhōngguó Rénbǎo qián wǔ dà kèhù de míngchēng jí qí xiāoshòu shōurù zhànbǐ.
- **Translation (English):** List the names and sales revenue percentages of PICC's top five customers in 2022.

A.3. Comparison & Calculation

- **Original (Chinese):** 计算 2023 年中国神华营业收入的同比增长率和增长金额。
- **Transliteration (Pinyin):** Jìsuàn 2023 nián Zhōngguó Shénhuá yíngyè shōurù de tóngbǐ zēngzhǎng lǜ hé zēngzhǎng jīn'é.
- **Translation (English):** Calculate the year-on-year growth rate and growth amount of China Shenhua's operating revenue in 2023.

A.4. Judgment & Verification

- **Original (Chinese):** 2024 年建设银行是否存在商誉减值？如果是，减值金额是多少?
- **Transliteration (Pinyin):** 2024 nián Jiànshè Yínháng shìfǒu cúnzài shāngyù jiǎnzhí? Rúguǒ shì, jiǎnzhí jīn'é shì duōshǎo?
- **Translation (English):** Did CCB have goodwill impairment in 2024? If so, what is the impairment amount?

A.5. Reasoning & Analysis

- **Original (Chinese):** ROE 分解：分析 2024 年杜邦三因素对工商银行 ROE 变动的贡献。
- **Transliteration (Pinyin):** ROE fēnjiě: fēnxī 2024 nián Dùbāng sān yīnsù duì Gōngshāng Yínháng ROE biàndòng de gòngxiàn.
- **Translation (English):** ROE decomposition: Analyze the DuPont three-factor contribution to ICBC's ROE change in 2024.

B. Examples of the Hallucination Detector

Tables 4–6 provide illustrative examples of the hallucination detector's judgments, covering a correct answer, a potential hallucination, and a severe hallucination.

Component	Content
Question	**Original (Chinese):** 2024 年中国工商银行的营业收入是多少? **Transliteration (Pinyin):** 2024 nián Zhōngguó Gōngshāng Yínháng de yíngyè shōurù shì duōshǎo? **Translation (English):** What was the operating revenue of Industrial and Commercial Bank of China (ICBC) in 2024?
Question Type	Fact Extraction
Retrieved Evidence	**[Filename]** ICBC_2024_Annual_Report.pdf **[Page]** 15 **Original (Chinese):** 根据年报，2024 年工商银行实现营业收入 8234.56 亿元，同比增长 3.2%。其中，利息净收入为 6789.12 亿元，手续费及佣金净收入为 1445.44 亿元。 **Transliteration (Pinyin):** Gēnjù niánbào, 2024 nián Gōngshāng Yínháng shíxiàn yíngyè shōurù 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. Qízhōng, lìxī jìng shōurù wéi 6789.12 yì yuán, shǒuxùfèi jí yōngjīn jìng shōurù wéi 1445.44 yì yuán. **Translation (English):** According to the annual report, ICBC achieved an operating revenue of 8,234.56 billion yuan in 2024, a year-on-year increase of 3.2%. Among this, net interest income was 6,789.12 billion yuan, and net fee and commission income was 1,445.44 billion yuan.
RAG Model Answer	**Original (Chinese):** 2024 年工商银行的营业收入为 8234.56 亿元，同比增长 3.2%。 **Transliteration (Pinyin):** 2024 nián Gōngshāng Yínháng de yíngyè shōurù wéi 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. **Translation (English):** The operating revenue of ICBC in 2024 was 8,234.56 billion yuan, representing a year-on-year growth of 3.2%.
Rule-Based Judgment	**Numerical Verification:** PASS (8,234.56 billion yuan found in evidence). **Text Coverage:** 0.85 (High coverage). **Reference Consistency:** PASS (Filename and page match). **Composite Score:** 0.92.
Final Verdict	**Verdict:** Evidenced. **Confidence:** 0.92. **Decision:** The answer is well-supported by evidence, no hallucination.

Table 4: Example 1: Correct Answer

Component	Content
Question	**Original (Chinese):** 2024 年中国工商银行的净利润增长率是多少? **Transliteration (Pinyin):** 2024 nián Zhōngguó Gōngshāng Yínháng de jìnglìrùn zēngzhǎng lǜ shì duōshǎo? **Translation (English):** What was the net profit growth rate of ICBC in 2024?
Question Type	Fact Extraction
Retrieved Evidence	**[Filename]** ICBC_2024_Annual_Report.pdf **[Page]** 15 **Original (Chinese):** 根据年报，2024 年工商银行实现营业收入 8234.56 亿元，同比增长 3.2%。净利润为 3456.78 亿元。 **Transliteration (Pinyin):** Gēnjù niánbào, 2024 nián Gōngshāng Yínháng shíxiàn yíngyè shōurù 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. Jìnglìrùn wéi 3456.78 yì yuán. **Translation (English):** According to the annual report, ICBC achieved an operating revenue of 8,234.56 billion yuan in 2024, a year-on-year increase of 3.2%. The net profit was 3,456.78 billion yuan.
RAG Model Answer	**Original (Chinese):** 2024 年工商银行的净利润增长率为 5.8%，与上年相比有所改善。 **Transliteration (Pinyin):** 2024 nián Gōngshāng Yínháng de jìnglìrùn zēngzhǎng lǜ wéi 5.8%, yǔ shàngnián xiāngbǐ yǒu suǒ gǎishàn. **Translation (English):** The net profit growth rate of ICBC in 2024 was 5.8%, showing an improvement compared to the previous year.
Rule-Based Judgment	**Numerical Verification:** FAIL (5.8% not found in evidence). **Text Coverage:** 0.35 (Low coverage). **Reference Consistency:** PASS (Filename and page match). **Composite Score:** 0.48.
Final Verdict	**Verdict:** No Evidence. **Confidence:** 0.48. **Decision:** The answer contains information not found in the evidence, indicating hallucination risk.

Table 5: Example 2: Potential Hallucination

Component	Content
Question	**Original (Chinese):** 2024 年中国工商银行的营业收入是同比增长还是下降? **Transliteration (Pinyin):** 2024 nián Zhōngguó Gōngshāng Yínháng de yíngyè shōurù shì tóngbǐ zēngzhǎng háishì xiàjiàng? **Translation (English):** Did ICBC's operating revenue increase or decrease year-over-year in 2024?
Question Type	Judgment Verification
Retrieved Evidence	**[Filename]** ICBC_2024_Annual_Report.pdf **[Page]** 15 **Original (Chinese):** 根据年报，2024 年工商银行实现营业收入 8234.56 亿元，同比增长 3.2%。 **Transliteration (Pinyin):** Gēnjù niánbào, 2024 nián Gōngshāng Yínháng shíxiàn yíngyè shōurù 8234.56 yì yuán, tóngbǐ zēngzhǎng 3.2%. **Translation (English):** According to the annual report, ICBC achieved an operating revenue of 8,234.56 billion yuan in 2024, a year-on-year increase of 3.2%.
RAG Model Answer	**Original (Chinese):** 2024 年工商银行的营业收入同比下降了 2.5%，这主要是受市场环境的影响。 **Transliteration (Pinyin):** 2024 nián Gōngshāng Yínháng de yíngyè shōurù tóngbǐ xiàjiàng le 2.5%, zhè zhǔyào shì shòu shìchǎng huánjìng de yǐngxiǎng. **Translation (English):** ICBC's operating revenue decreased by 2.5% year-over-year in 2024, primarily due to market conditions.
Rule-Based Judgment	**Numerical Verification:** CONTRADICTION (Evidence indicates +3.2% growth, answer states -2.5% decline). **Text Coverage:** 0.25 (Very low coverage). **Reference Consistency:** PASS. **Composite Score:** 0.15.
Final Verdict	**Verdict:** Contradiction. **Confidence:** 0.15. **Decision:** The answer directly contradicts the evidence, indicating severe hallucination.

Table 6: Example 3: Severe Hallucination