# Precision Retrieval and Comprehensive Analysis of Corporate Culture Information in Financial Documents Using a RAG Framework

Anonymous CVPR submission

Paper ID 9277

## Abstract

*Corporate culture plays a pivotal role in shaping organizational behavior and decision-making, ultimately influencing financial performance. Traditional methods for analyzing corporate culture are often resource-intensive and lack scalability, particularly for cross-sectional comparisons across multiple companies. Our study addresses these limitations by leveraging advanced text-mining techniques and large language models (LLMs) to analyze corporate culture in Chinese AI-related enterprises. We propose a novel framework based on the LLAMA model, which utilizes specialized agents to evaluate ten cultural dimensions, including adaptability, innovation, and results orientation. To mitigate the hallucination problem inherent in traditional LLMs, we introduce the ReRAG model, a dual-layer Retrieval-Augmented Generation approach built on the Faiss vector library. Our model enhances the accuracy and reliability of corporate culture analysis by grounding responses in relevant knowledge retrieved from financial documents. Using a dataset comprising 255 annual reports and 87 industry reports, we demonstrate that the proposed framework significantly improves the efficiency and precision of corporate culture analysis. Our results indicate @, highlighting the transformative potential of LLMs and RAG techniques in financial document analysis. This research contributes to the field by providing a scalable, data-driven approach to understanding corporate culture and its implications for financial decision-making.*

## 1. Introduction

### 1.1. corporate culture

Corporate culture embodies the core values and behavioral norms that underpin an organization's operations, exerting a subtle yet profound influence on employees' attitudes, behaviors, and decision-making processes. This, in turn, significantly impacts a company's financial performance [4]. Notably, Li et al. demonstrated a significant positive correlation between analysts' optimistic tone in discussing corporate culture and their stock recommendations and target price estimates [6]. For financial analysts, comprehending a company's cultural framework is critical as it facilitates accurate predictions of management's decision-making tendencies, risk appetite, and strategic commitment to long-term objectives.

This study investigates the corporate culture of Chinese enterprises, employing advanced text-mining methodologies to uncover cultural signals embedded in financial documents such as annual reports and press releases. By analyzing these linguistic patterns, this research aims to provide insights into how corporate culture is communicated and its implications for financial analysis and decision-making.

### 1.2. LLMs

Traditional methods for corporate culture analysis often demand extensive time and human resources to review large volumes of material manually. These methods also face limitations in performing cross-sectional comparisons across multiple companies efficiently. In contrast, large language models (LLMs) have emerged as transformative tools within financial services and investment management, enabling the extraction of valuable insights with enhanced efficiency and accuracy from structured and unstructured data sources such as annual reports and press releases.

This study introduces a novel corporate culture analysis framework based on LLAMA. Building upon the ten cultural dimensions proposed by Li et al. [6], including adaptability, innovation, and results orientation, this framework leverages specialized agents for each cultural dimension. These agents evaluate a company's culture from positive and negative perspectives, systematically analyzing textual data. The outputs from these agents are subsequently integrated to produce a holistic assessment of the company's cultural profile, offering a scalable and data-driven approach to corporate culture analysis.

### 1.3. RAG

Traditional large language models (LLMs) are heavily reliant on their training data, making text generators such as GPT and BERT particularly vulnerable to hallucinations—producing seemingly plausible but factually incorrect or fabricated information [3]. To mitigate this issue, Retrieval-Augmented Generation (RAG) has emerged as an innovative hybrid architecture designed to enhance the reliability of LLMs. RAG consists of two primary components: a retrieval module and a generation module. The retrieval module utilizes dense vector representations to identify relevant documents from vast datasets, which are then passed to the generation module. This module uses the retrieved information to produce grounded and factually accurate responses. By incorporating external knowledge retrieval, the RAG framework significantly reduces the hallucination problem inherent in traditional LLMs [5].

In recent years, RAG models have been widely deployed across a variety of domains, including open-domain question answering, conversational agents, and personalized recommendation systems [1]. These applications demonstrate the versatility and scalability of RAG, positioning it as a robust solution for tasks requiring accurate, context-sensitive information generation.

To address the hallucination issue in the precise analysis of financial documents, we propose the ReRAG model, a dual-layer RAG approach built on the Faiss vector library. This model is specifically designed for analyzing financial documents by creating a summarized dataset for each document, which consolidates and indexes the multiple topics discussed within. Using RAG technology, ReRAG retrieves the five most relevant documents related to a query on corporate culture. Subsequently, a second retrieval round is performed on these documents to refine the response further.

Through extensive empirical analysis, we demonstrate that the adoption of multi-layer RAG techniques significantly enhances model accuracy, achieving improvements of compared to , and relative to . In conclusion, this study makes several key contributions:

1.Enhanced Text Mining for Financial Data

By applying advanced text-mining techniques, this research uncovers valuable cultural signals embedded within financial documents, such as annual reports and press releases. These insights contribute to a deeper understanding of corporate communication and its implications for financial analysis.

2.Addressing Limitations of Traditional Methods

The study underscores the inefficiencies of traditional approaches, which are often resource-intensive and unable to perform cross-sectional analyses, while highlighting the transformative potential of large language models in extracting meaningful insights from financial documents.

3.Innovation in Mitigating LLM Hallucination

The proposed ReRAG approach alleviates hallucination by grounding the model's responses in the relevant knowledge retrieved during the search process, thereby improving the reliability and accuracy of LLM-generated outputs.

## 2. Related work

### 2.1. Application of Large Language Models in financial field

Large language models have been widely used in the financial domain, and research [9] has trained language models focused on the financial domain and achieved significant results on financial NLP tasks. In the study, an open-source framework for large language modeling in the financial domain named FinGPT is being constructed. The study [2] develops the FinBERT model, which effectively extracts information from financial texts by pre-training on large-scale financial texts and outperforms traditional lexicon methods and machine learning algorithms in sentiment categorization and ESG topic recognition tasks, proving the potential of the application of LLMs to the financial domain. In the study [10], an open-source framework for large language modeling in the financial domain named FinGPT is constructed and adapted to the financial domain by low-rank adaptation and reinforcement learning techniques, demonstrating the potential of LLMs to be applied in financial domains such as smart investment, quantitative trading, and so on. A survey [7] reviews the state-of-the-art of LLMs applications in finance, demonstrates their performance enhancement on financial natural language processing tasks, and proposes a decision framework to guide the application of large language models. The research [11] showed that through instruction fine-tuning and retrieval enhancement, large language models can be effectively applied to financial sentiment analysis and achieve better performance than traditional models.

### 2.2. Retrieval-Augmented Generation(RAG)

RAG is a technique that combines retrieval and generation to enhance the generation of large language models (LLMs) by retrieving relevant information from external databases. RAG [5] has shown superior performance in generating specific, diverse, and factual language compared to traditional models. Financial documents typically contain domain-specific language, multiple data formats, and unique contextual relationships that general purpose-trained LLMs do not handle well. The specialized terminology and complex data formats in financial documents make it difficult for models to extract meaningful insights, in turn, causing inaccurate predictions, overlooked insights, and unreliable analysis, which ultimately hinder the ability to make wellinformed decisions. Hence, the research [8] have introduced a novel approach that significantly advances the field of information extraction from financial documents through the development of a hybrid RAG system.

## 3. dataset description

In this study, the Sentiment Analysis for Financial News dataset was used as a benchmark to assess the performance of the selected embedding model. To evaluate the effectiveness of the entire financial document analysis system, it was crucial to incorporate a diverse range of financial document data, along with a corresponding financial document QA dataset. However, a number of publicly available financial datasets were found to be inadequate for the specific requirements of this research. The dataset required for this work includes corporate annual reports, financial news, and financial reports.
To address this gap, we collected data from two primary sources: the Cninfo platform and the Djyanbao platform. The Cninfo platform is an official information disclosure platform authorized by the China Securities Regulatory Commission (CSRC) and operated by Shenzhen Securities Information Co., Ltd. It is widely regarded for providing reliable and authoritative information about China's capital markets, including data on listed companies and associated securities. The Djyanbao platform, on the other hand, specializes in providing industry-specific reports. Additionally, we gathered company annual financial reports from the official websites of the respective companies.
During the course of the research, we identified that quarterly reports lacked the necessary level of detail for in-depth analysis, while semi-annual reports showed considerable similarity to annual reports. Consequently, only annual reports were selected for this study to ensure that the data provided comprehensive and distinct insights.
We assembled a dataset comprising 255 annual reports and 87 industry reports, with data primarily drawn from the years 2023 and 2024, to construct the core dataset for our experiments. These reports formed the foundation for validating and refining the proposed methodologies.
Based on the data outlined above, we developed summary datasets for each document, which were used in the first layer of the RAG process in the experimental setup. In parallel, we constructed a related QA dataset, leveraging the same set of documents, to evaluate the feasibility of the proposed financial document analysis system.

## 4. Methodology

### 4.1. ReRAG

The experiment commences by posing a question related to the financial documents, thereby initiating the first round

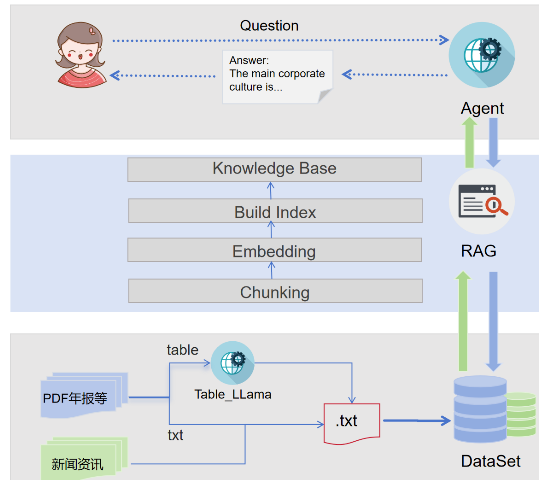


Figure 1. ReRAG Overall Workflow Diagram.

of retrieval. During this phase, a query is made against the established summary dataset of financial documents to identify the most relevant sentences, based on their semantic similarity to the input question. This retrieval process utilizes dense vector representations of the document summaries. Specifically, the input question is encoded into a high-dimensional vector, which is then compared to vectors representing individual document summaries. The sentences exhibiting the highest cosine similarity scores are selected as the most contextually relevant content.
For efficient similarity search, the FAISS library is employed. FAISS, developed by Facebook AI, is a high-performance tool for similarity search and dense vector clustering. It leverages two key principles: inverted file indexing (IVF) and product quantization (PQ). IVF divides the vector space into multiple subspaces, or ”buckets,” which allows the search to be conducted within smaller, more focused regions, thereby expediting the search process. Concurrently, PQ compresses vectors, reducing both storage and computational costs, while maintaining approximate distances between vectors, thus enhancing retrieval efficiency.
In the second round of retrieval, we process the documents retrieved in the first phase. To extract the textual content from the identified PDF documents, the pdfplumber library is utilized. Pdfplumber enables the accurate extraction of data from complex document layouts, including parsing non-linear content such as tables, graphs, and other elements commonly found in financial documents. Upon extraction, the text is subjected to document segmentation, or ”chunking,” which divides the content into smaller, semantically meaningful chunks. These chunks are determined based on natural boundaries, such as paragraphs or

key information units, ensuring that each chunk represents a distinct idea or topic. This segmentation process enhances the model's ability to efficiently process and retrieve specific information in subsequent stages.

To further refine the relevance of the extracted text, the BGE-large-zh model is applied. BGE-large-zh, a pre-trained Chinese language model, is designed for efficient and precise text matching. It encodes the document chunks into dense vector representations, capturing the semantic content of the text. The model then calculates the semantic similarity between the question and each chunk, facilitating the identification of the most relevant sections. These relevant chunks are subsequently forwarded to the generation and inference modules for further processing.

Finally, the generation and inference modules synthesize the information retrieved during both the first and second rounds of retrieval to produce a precise, contextually grounded response. The final answer is derived from the most pertinent knowledge extracted from the document corpus, ensuring that the output is both accurate and firmly grounded in the content of the financial documents.

### 4.2. Generation and Inference

The experiment is grounded in the ten cultural dimensions proposed by Li et al., which include adaptability, innovation, results orientation, among others. These dimensions collectively form a comprehensive framework for analyzing organizational culture. To facilitate the analysis and synthesis of responses from multiple cultural perspectives, a multi-agent model is employed, utilizing Chain-of-Thought (COT) techniques in conjunction with the Retrieval-Augmented Generation (ReRAG) approach.

The multi-agent model comprises ten distinct agents, each dedicated to analyzing one of the ten cultural dimensions identified by Li et al. These dimensions cover various facets of organizational culture, such as adaptability, innovation, and results orientation. Each agent is constructed using a Large Language Model (LLM) trained to focus specifically on its assigned cultural dimension. These agents are tasked with processing the content of financial documents and extracting insights relevant to their designated cultural aspect.

The analysis process commences with the ReRAG module, which retrieves relevant content from the financial documents based on the input question. The retrieved content is then provided to the LLM-based agents. Each agent analyzes both the question and the document content from the perspective of its corresponding cultural dimension.

To enhance the depth of reasoning, the COT methodology is integrated within each agent. This methodology enables each agent to break down the question into smaller sub-tasks, considering how different aspects of the company's culture relate to the inquiry. This structured approach promotes logical organization and articulation of reasoning, thereby improving the accuracy and relevance of the agent's responses.

After each agent has generated a response based on its cultural perspective, the individual answers are integrated into a final synthesized output. This aggregation process involves consolidating insights from all ten agents, carefully considering the nuances of each cultural dimension to ensure a holistic and contextually grounded response. The integration process guarantees that the final answer captures the full spectrum of cultural factors while maintaining coherence and accuracy.

Subsequently, the multi-agent model's output undergoes further refinement through additional rounds of interaction and inference among the agents. This iterative process enhances the overall precision and relevance of the final response, enabling a multidimensional understanding of the input question. The ultimate goal is to generate a response that is not only grounded in the context of the financial documents but also reflective of the ten cultural dimensions, thereby providing a comprehensive and robust answer to the query.

## 5. Result

To evaluate the accuracy of the ReRAG model, we constructed a QA dataset based on financial document data. The dataset contains a series of questions that are designed to assess the model's ability to extract relevant information and generate accurate responses from financial documents. In the experiment, we compare the performance of the ReRAG model against two other models, namely Llama3 and Qwen2.5, by asking each model to answer the questions from the QA dataset.

The evaluation process involves calculating the cosine similarity between the ground truth answers and the answers generated by each model. Specifically, for each question in the dataset, the correct answer is compared with the generated response from each model by encoding both answers into dense vector representations. The cosine similarity score is then computed to quantify the degree of similarity between the generated answer and the correct answer. A higher cosine similarity indicates that the generated answer is more semantically similar to the correct one, which serves as an indicator of the model's accuracy.

This method allows for a direct and quantitative comparison of the models' performance, providing insights into how well each model can handle the specific challenges posed by financial document analysis. By evaluating ReRAG alongside other state-of-the-art models, such as Llama3 and Qwen2.5, we can assess the relative strengths and weaknesses of each approach in terms of answering financial-related questions accurately. The results of this evaluation will contribute to understanding the effectiveness of the ReRAG model in generating grounded and contextually relevant answers within the financial domain.

| ReRAG | Llmam3 | Qwen2.5 |
|---|---|---|
| 0.84228515625 | 0.50732421875 | 0.54647111875 |

Table 1. The cosine similarity between the answers obtained by each model for the QA dataset questions and the standard answers.

Based on the experimental results, we observed that the average semantic similarity between the answers generated by the ReRAG model and the ground truth answers, measured by cosine similarity, was 0.842. In contrast, the average semantic similarity between the answers generated by the Llama3 model and the ground truth answers was 0.507, while for the Qwen2.5 model, the average similarity with the correct answers was 0.546.

To ensure a fair and accurate comparison, all models were evaluated using the same QA dataset, which was derived from financial documents. The cosine similarity scores were computed by encoding both the generated answers and the correct answers into dense vector representations and calculating the cosine similarity between them. Higher cosine similarity scores indicate a greater degree of semantic alignment between the model's generated responses and the ground truth answers.

Upon comparing the results, the ReRAG model exhibited a significantly higher accuracy compared to both Llama3 and Qwen2.5, as evidenced by the consistently higher average semantic similarity scores. Specifically, the ReRAG model outperformed Llama3 and Qwen2.5 by a margin of 0.335 and 0.296, respectively. This demonstrates that the ReRAG model is more effective in generating responses that are contextually and semantically aligned with the correct answers, indicating its superior capability in understanding and generating accurate answers to questions based on financial documents.

## 6. Conclcusion

Our study presents a novel framework for analyzing corporate culture in Chinese AI-related enterprises by integrating advanced text-mining techniques, large language models (LLMs), and Retrieval-Augmented Generation (RAG). The proposed ReRAG model, featuring a dual-layer retrieval process and multi-agent analysis, addresses the limitations of traditional methods, such as inefficiency and resource intensity, while significantly enhancing the accuracy and reliability of corporate culture analysis in financial documents. By leveraging FAISS for efficient similarity search, BGE-large-zh for precise text matching, and a multi-agent framework with Chain-of-Thought (COT) techniques, our study captures the nuances of ten cultural dimensions, providing deeper insights into corporate communication and its financial implications. The ReRAG model also mitigates LLM hallucination by grounding responses in retrieved knowledge, ensuring contextually accurate outputs. @ (Results will be inserted here). Overall, our research establishes a scalable and adaptive solution for corporate culture analysis, setting a new benchmark for future studies in multi-domain retrieval and financial document analysis.

## 7. Division of Labor

Research and Idea Retrieval: Jiayang Yao, Jiajia Ye
Data Collection: Jiayang Yao
Text Reading and Table Extraction from PDFs: Jiayang Yao
Multi-Agent Model Development: Jiajia Ye
ARG Framework Development: Jiajia Ye
Mid-Term Report (Literature Review, Text on Slide 4 of PPT): Jiayang Yao
Mid-Term Report (Other Sections): Jiajia Ye
Report Compilation: Jiajia Ye, Jiayang Yao
PPT Creation: Jiajia Ye

## References

[1] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions, 2024. 2

[2] Allen H Huang, Hui Wang, and Yi Yang. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841, 2023. 2

[3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023. 2

[4] Campbell R. Harvey Shivaram Rajgopal Shivaram John R. Graham, Jillian Grennan. Corporate culture: Evidence from the field. *Journal of financial economics*, 2022. 1

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 2

[6] Kai Li, Feng Mai, Rui Shen, Chelsea Yang, and Tengfei Zhang. Dissecting Corporate Culture Using Generative AI – Insights from Analyst Reports. *SSRN Electronic Journal*, 2023. 1

[7] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382, 2023. 2

[8] Bhaskarjit Sarmah, Dhagash Mehta, Benika Hall, Rohan Rao, Sunil Patel, and Stefano Pasquali. Hybridrag: Integrat-

ing knowledge graphs and vector retrieval augmented generation for efficient information extraction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 608–616, 2024. 3

[9] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023. *ArXiv preprint: https://arxiv.org/pdf/2303.17564. pdf*, 2024. 2

[10] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023. 2

[11] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 349–356, 2023. 2