Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

Annual Report Summarizer: RAG based Summarizer

1Dr. Jayasudha Subburaj, 2Abinanthan K, 3Dhanush S, 4Gopalakrishnan M, 5Ramana Sri S

1Professor, Sri Krishna College of Engineering and Technology

2,3,4,5Student, Sri Krishna College of Engineering and Technology

Article History:

Received:04/09/2025

Revised:03/10/2025

Accepted:27/11/2025

Abstract:

Introduction: --- Annual reports are complicated, voluminous documents that summarize the financial performance, governance structure and strategic outlook of a company. Manual analysis is time-consuming and overwhelming due to their size and technical density. Advances in Large Language Models (LLMs) and Generative AI enable automated interpretation of such documents. Retrieval-Augmented Generation (RAG) strengthens contextual accuracy by grounding generated text in retrieved evidence. The Annual Report Summarizer (ARS) addresses these challenges using semantic retrieval, section-level summarization and accessibility features such as multilingual translation and text-to-speech.

Objectives: Demonstrate the ARS architecture, including PDF preprocessing, vector storage and RAG summarization. Evaluate contextual accuracy and comprehensibility of summaries using LLM-based tools such as G-EVAL and ROUGE. Generate section-wise human-like summaries for real annual reports. Highlight the impact of ARS on financial literacy, research and analytical efficiency.

Methods: The ARS system integrates PDF text extraction, cleaning, segmentation, embedding generation, semantic retrieval and LLM-based summarization. Text is converted into vector embeddings for retrieval using cosine similarity. The RAG module retrieves relevant chunks and generates summaries through context-aware prompting. The system includes modules for user input, preprocessing, embeddings, summarization and output delivery with multilingual translation and text-to-speech.

Results: ARS accelerates interpretation of lengthy financial documents by combining semantic retrieval and LLM-based summarization. It improves clarity, factual consistency and coherence of generated summaries. Evaluation using G-EVAL and ROUGE supports the system's contextual accuracy. ARS reduces time and cognitive effort while enhancing accessibility for analysts, students and researchers.

Conclusions: ARS automates interpretation of complex annual reports using Retrieval-Augmented Generation. It integrates semantic retrieval, contextual summarization and accessibility features. While limitations exist in scalability, numerical accuracy and domain-specific adaptation, ARS demonstrates strong potential as a reliable framework for financial analysis. Future improvements may include domain-specific fine-tuning and multimodal data processing.

Keywords: Annual Report Summarization, Retrieval-Augmented Generation, Generative AI, LLMs, Financial Document Analysis.

1. Introduction

The rapid rate of the evolution of Artificial Intelligence (AI) and Natural Language Processing (NLP) has transformed the process of interaction of human beings with large amounts of textual data. Publication of annual reports on the electronic platform in the corporate and financial world has led to

the influx of information in the public, yet it is a very complicated, voluminous, and cumbersome document that is not easy to read properly. Annual reports which are summaries of a financial performance, governance structure and strategic outlook of a company are extremely important to their investors, analysts and researches. However, manual analysis is a time-consuming and mind-consuming process due to its size and technical density and therefore it is easy to become information overloaded and the accuracy of the analysis will be compromised.

The recent advances in the Large Language Models (LLM) and Generative AI have opened up new opportunities to use such documents as input toward automating its interpretation. It is possible to feed any form of text into Generative Summarization systems, learn financial language, and summarize it in a sensible and section-level fashion. The reason behind this is that such systems would allow the stakeholders to derive valuable information within a minimum of time by focusing on the most significant sections of the document, such as the Letter of the Chairman, a report of the Board, and Financial Statements. The Retrieval-Augmented Generation (RAG) systems have potential as a contextual summary of corporate texts using a NLP-based approach, as illustrated in Financial NLP works, e.g. Prototype-as-Query RAG as a Framework to Summarise Financial Reports [1] and Auto-Generating Earnings Report Analysis with an Augmented LLM [4].

Other previous summarization methods, such as extractive models, n-gram co-occurrence score (e.g. ROUGE [5]) were more focused on a surface similarity and had little insight into the underlying semantic information. With the introduction of the RAG architectures, one can now combine both the information retrieval with the context-based text generation, and this makes text generation much more factual and contextual. These hybrid systems can identify the right text snippets of a vector database, preserving the summaries generated that are based on the source text, which is also used in recent financial LLM research [2][3]. Nevertheless, there are challenges in areas such as layout-conscious parsing, reduction of hallucinations and multi-section coherence particularly when applied to structured financial disclosures.

The article reveals a smart system, Annual Report Summarizer (ARS), which addresses these concerns by using contextual retrieval and generation of summaries according to RAG. The PDF annual report process is also done in steps of text extraction, cleaning, semantic chunking and embedding generation process followed by contextual query retrieval and summary generation process. Besides, the aspect of such accessibility as multilingual translation and text-to-speech support give the users the opportunity to absorb the insights with ease due to the presence of language and visual barriers.

2. Objectives

The primary research questions of the study are:

- To demonstrate the architecture and the procedure of the ARS system with PDF pre-processing, the storage of vectors and the summarization of the RAG.
- On the contextual accurateness and comprehensibility of generated summaries by comparing the generated summaries with the evaluator using the assessment system using the LLM, such as G-EVAL [3] and quantitative evaluation tools, such as ROUGE [5].

- To demonstrate the applicability of ARS to generate section-wise human-like summaries using the real life annual reports.
- To comment on the implication of introducing ARS into the sphere of financial analysis, research, and education, one needs to pay attention to the fact that it can allow the levels of financial literacy to increase and analytical work to be less complex.

3. Methods

The smart system is Annual Report Summarizer (ARS), which will enable the automatization of the interpretation of expansive corporate annual reports through the presentation of the summary of each section of information as well as the background of the information in a concise form. It is a technique that integrates Retrieval-Augmented Generation (RAG) with Generative AI that enables the system that can access the corresponding pieces of the text and generate the correct and human-like summaries.

The architecture of ARS framework is developed based on scalable and modular pipeline that encourages effective document management, contextual look-up and summary generation. The standard process begins with the upload of an annual report of a company in the PDF format and then continues through a series of central processing layers that collaborate with each other to give the final output summary.

Modules include the User Input Layer, Preprocessing and Text Segmentation Module, Embedding and Semantic Retrieval Module, Summarization and Generation Module, and Output and Accessibility Module. The ARS system is based on an idea of layered integration pipeline where the user uploads an annual report, preprocessing separates text into meaningful paragraphs, chunks are embedded and stored as vectors, the RAG engine retrieves relevant portions, and the LLM generates summaries, followed by translation or audio conversion.

4. Results

Key Implementation Features

- Mandatory Terms & Conditions acceptance.
- PyMuPDF-based text extraction and cleaning.
- RAG using Sentence Transformers + Chroma DB.
- Gemini-based section-wise summarization.
- Streamlit UI for summaries.
- Environment variables for secure API handling.

Performance Metrics

1. Summarization Metrics

- ROUGE-1, ROUGE-2, ROUGE-L
- BERTScore (Precision, Recall, F1)
- LLM Score for coherence and accuracy

2. System Performance

- Processing time: <3 minutes for 250 pages
- Retrieval accuracy: 90–94%

5. Discussion

- Scalability of ARS is a critical problem, which occurs in the annual report processing, which is long and graphically intensive. Size of Document and Token Limits affect processing, and embedding storage overhead increases memory usage. Computation costs limit scalability to real time or enterprise use.
- RAG frameworks assist in minimizing hallucinations and there is always a struggle in making sure that the created summaries are factual. Numerical misinterpretation, context drift, and factual hallucination remain challenges.
- Financial domain is a particular linguistic and contextual issue that is difficult to make the generic LLMs perceive. Terminology ambiguity, table and chart interpretation, and legal sensitivity complicate summarization.
- Evaluation limitations include metric weaknesses, subjective LLM evaluators, and absence of standardized datasets. Translation accuracy, text-to-speech issues, and cross-platform deployment also pose constraints.
- Cloud-based processing risks and compliance concerns affect data privacy and confidentiality.

Refrences

[1] Anonymous, “Prototype-as-Query RAG for Financial Report Summarisation,” ACL Conference Submission, 2024.
[2] X. Yang, S. Zang, Y. Ren, D. Peng, and Z. Wen, “Evaluating Large Language Models on Financial Report Summarization: An Empirical Study,” arXiv preprint arXiv:2411.06852, 2024.
[3] Y. Liu, A. Feng, T. He, W. Liu, and P. Zhou, “G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment,” arXiv preprint arXiv:2303.16634, 2023.
[4] V.-D. Le, “Auto-Generating Earnings Report Analysis via an Augmented LLM,” arXiv preprint arXiv:2412.08179, 2024.
[5] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” Proceedings of the Workshop on Text Summarization Branches Out, 2004.
[6] J. White et al., “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,” arXiv preprint, 2023.
[7] R. Angles, “A Comparison of Current Graph Database Models,” ICDEW, 2012.