Document Q&A Research Paper

Abstract

This document presents a comprehensive study on document-based question answering systems. We explore novel approaches that leverage large language models to extract precise answers from unstructured text documents. Our method achieves significant improvements over traditional information retrieval techniques by combining semantic understanding with direct evidence extraction.

The key contribution of this work is a framework that not only provides answers but also identifies and cites the exact textual evidence supporting those answers. This approach enhances transparency and allows users to verify the accuracy of generated responses.

1. Introduction

Document question answering has emerged as a critical task in natural language processing. Unlike open-domain QA, document QA requires systems to ground their answers in specific source material.

Traditional approaches rely on keyword matching and term frequency analysis, which often fail to capture the semantic nuances of complex questions. Recent advances in transformer-based models have opened new possibilities for deeper document understanding.

In this paper, we propose a methodology that processes documents page by page, extracting structured text and then reasoning over the full content to produce evidence-backed answers. Our system is designed for single-document analysis with a focus on accuracy and verifiability.

2. Methodology

Our approach consists of three main stages: text extraction, semantic analysis, and evidence mapping.

2.1 Text Extraction
We extract text from PDF documents using optical character recognition and text layer parsing. Each page is processed independently to maintain page-level attribution.

2.2 Semantic Analysis
The extracted text is analyzed using a large language model that understands context across pages. Our method achieves a 15% improvement over the baseline across all evaluated metrics. The model receives the full document context along with the user's question.

2.3 Evidence Mapping
For each answer, the system identifies specific quotes from the document that support the conclusion. These quotes must be exact substrings of the original text, ensuring verifiability.

3. Results

We evaluated our system on a benchmark of 500 document-question pairs across multiple domains.

3.1 Accuracy
Our system achieved 92.3% accuracy on factual questions and 87.1% on inferential questions. This represents a significant advancement over prior methods.

3.2 Evidence Quality
In 94% of cases, the cited evidence directly supported the generated answer. Human evaluators rated the evidence relevance at 4.2 out of 5.

3.3 Performance Comparison
Compared to baseline retrieval methods, our approach showed improvements in both precision and recall. The system processes an average document in under 3 seconds, making it suitable for interactive use.

4. Discussion

The results demonstrate that combining semantic understanding with explicit evidence extraction produces more trustworthy QA systems.

4.1 Strengths
The proposed framework consistently outperforms existing approaches in both accuracy and efficiency. Users particularly valued the ability to see exactly which parts of the document supported each answer.

4.2 Error Analysis
Most errors occurred when questions required synthesizing information from multiple non-adjacent pages. The system occasionally selected evidence that was topically related but did not directly support the answer.

4.3 Conflicting Evidence
In some documents, we found evidence that both supported and contradicted the primary conclusion. However, the limitations discussed in Section 5 suggest that further validation is needed before deploying this system in high-stakes environments.

5. Limitations

While our results are promising, several limitations should be noted.

5.1 Document Length

The current system is optimized for documents of up to 50 pages. Performance may degrade with significantly longer documents due to context window constraints.

5.2 Language Support

Our evaluation was conducted exclusively on English-language documents. Multilingual support remains an area for future work.

5.3 Domain Specificity

The system performed best on academic and technical documents. Performance on highly specialized legal or medical texts was somewhat lower, suggesting that domain adaptation may be necessary for certain use cases.

5.4 Reliability Concerns

Some reviewers noted that the system may produce overconfident answers in cases where the document provides insufficient information.

6. Conclusion

This paper presented a document QA system that provides evidence-backed answers with page-level attribution. Our experiments demonstrate strong performance across multiple metrics and document types.

Key takeaways:
- Evidence-based QA significantly improves user trust and answer verifiability.
- Page-level text extraction enables precise source attribution.
- The system achieves state-of-the-art results on standard benchmarks.

Future work will focus on extending the system to support multi-document queries, improving performance on specialized domains, and adding multilingual capabilities.

References

[1] Smith et al., 'Advances in Document Understanding', NLP Conference 2024.
[2] Johnson & Lee, 'Evidence Extraction for QA Systems', AI Review 2023.
[3] Zhang et al., 'Transformer Models for Document Analysis', ICML 2024.
[4] Brown et al., 'Large Language Models as Knowledge Bases', Nature ML 2024.