IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, May 2026
Impact Factor: 8.2

DocFlow - An Intelligent Web-Based Platform for Automated Document Summarization and Q&A Generation

Chitale Tanishq Nilesh[1], Barate Mayuri Suresh[2], Gaikwad Ashwini Lalasaheb[3], Parbhane Akanksha Bhausaheb[4], Prof. Dhas Renuka Premraj[5]
Final Year Student, Department of Computer Engineering[1-4]
Professor, Department of Computer Engineering[5]
Hon. Shri Babanrao Pachpute Vichardhara Trusts, GOI Faculty of Engineering Kashti,
Savitribai Phule Pune University, Pune, India

Abstract: *In an era dominated by digital information, the challenge of efficiently processing lengthy documents is significant. This paper addresses the time-consuming nature of extracting key insights from documents and the barriers presented by foreign languages. The development of an intelligent web-based platform designed to automate document analysis is proposed. Leveraging a state-of-the-art Large Language Model (LLM), Google's Gemini, the system provides users with AI-powered summarization, automated question and answer (Q&A) generation and on-demand multi-language translation. A robust client-server architecture is utilized, featuring a React.js frontend for an interactive user experience and a Node.js backend for secure file processing. Preliminary usability testing indicates that this unified solution effectively enhances productivity and makes information more accessible for researchers and students alike*

Keywords: Intelligent Document Processing, Large Language Models, AI-Powered Summarization, Automated Q&A Generation, Natural Language Processing, Web Architecture

I. INTRODUCTION

In the modern digital age, professionals and academics are faced with an unprecedented deluge of information. The primary medium for this information exchange is digital documents such as research papers and technical manuals, which often span hundreds of pages. The traditional method of manually reading through these dense documents is plagued by inefficiency, information overload and significant language barriers.

Currently, a disjointed process of using separate applications—a PDF reader for viewing, a web service for summarization and another for translation—is often utilized to overcome this. This fragmented approach is highly time-consuming and disrupts focus. To overcome these challenges, a digital-first platform was conceptualized. An intelligent, centralized web application is provided where users can upload any PDF and instantly receive a concise summary, a set of relevant questions and answers and on-demand translation. By leveraging Google's Gemini Large Language Model (LLM), the process of comprehension is automated, allowing users to grasp the essence of any document rapidly.

II. LITERATURE REVIEW

Numerous methods and tools have been developed to manage and extract information from digital documents. The foundational approach of manual reading is inherently slow and offers no assistance for documents in foreign languages. While specialized online summarization tools exist, they operate as isolated functions. The workflow remains fragmented, requiring users to manually extract text and paste it between different services.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, May 2026
Impact Factor: 8.2

At the enterprise level, Intelligent Document Processing (IDP) platforms offer powerful AI-driven analysis for large-scale corporate needs. However, their complexity and prohibitive costs place them outside the scope of individual students or researchers. A distinct gap for a dedicated, user-friendly platform that integrates multiple AI-powered analysis features into a single workflow is identified. The proposed work democratizes advanced AI by bringing the power of an LLM into an accessible web application.

III. SYSTEM DESIGN AND METHODOLOGY

The development of Docflow followed a rigorous, modular and systematic approach, resulting in the delivery of a fully integrated, high-performance web platform. The system architecture was formulated following an exhaustive requirement analysis, which identified the critical need for a streamlined workflow to counteract information overload. The finalized system implements a minimalist, intuitive user experience centered around a seamless three-step pipeline: document upload, intelligent processing and results visualization. This workflow was optimized for speed and accessibility, ensuring that users can extract complex insights without the need for technical expertise.

Technologically, the platform is built upon a robust, mature client-server architecture. The frontend utilizes React.js, which was selected to provide a highly responsive, single-page application (SPA) environment. This ensures that the user interface remains interactive and stateful while delivering data in real-time. The backend is powered by Node.js, leveraging the Express.js framework to establish a secure and efficient server-side environment. This backend is responsible for orchestrating the entire lifecycle of a document, from secure multipart file ingestion to the management of secure API communication with the AI core. The architecture maintains a clean separation of concerns, ensuring that the frontend remains decoupled from the heavy-lifting logic of the backend, thereby enhancing the maintainability and reliability of the platform.

A deliberate architectural decision was made to implement a real-time, stateless processing model. By design, the system processes documents dynamically and discards data upon session completion rather than relying on a traditional persistent database. This approach renders the platform lightweight and eliminates the privacy and security overhead associated with storing sensitive user documents.

Finally, the AI integration module reflects a highly refined implementation of prompt engineering. Rather than relying on generic LLM responses, the system employs carefully tuned, systematic prompt-chaining strategies. These prompts are designed to enforce a strict quality control layer, ensuring that the Gemini API consistently returns contextually accurate summaries and standardized, well-structured JSON outputs. The combination of this robust architectural foundation and refined AI interaction strategy ensures that Docflow operates not just as a prototype, but as a reliable, production-ready productivity tool.

IV. MATHEMATICAL MODEL

The core processing relies on the mathematical principles that underpin Large Language Models and text processing.

Vector Semantics and Cosine Similarity

To measure textual relevance and ensure the generated summary (S) is contextually aligned with the original text (T), Cosine Similarity can be used. LLMs represent sentences as high-dimensional vectors. A value closer to 1 signifies higher semantic similarity.

$$Similarity(S,T) = \frac{V_S \cdot V_T}{||V_S|| \cdot ||V_T||}$$

Fig. 1. Mathematical formula 1

Where V_S and V_T are the vector representations of the summary and the text.

Information Compression Ratio

A compression ratio can be defined to measure how concisely the model summarizes the original content.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, May 2026
Impact Factor: 8.2

$$C_r = (1 - \frac{Length(S)}{Length(T)}) \times 100\%$$

Fig. 2. Mathematical formula 2

Where Length(S) is the word count of the summary and Length(T) is the word count of the original text. A higher ratio indicates a more condensed summary.

V. PRELIMINARY RESULTS AND USABILITY

The platform was successfully implemented following the proposed modular methodology. The primary focus was to create an intuitive user experience to address the problem of fragmented workflows. The main page features a minimalist file upload module, prompting the user to select or drag-and-drop their file. Preliminary usability feedback was gathered from a group of peer students. It was reported that the design is a significant improvement over using multiple isolated websites. This feedback validates the hypothesis that a unified platform is a desirable solution.

The user interface:

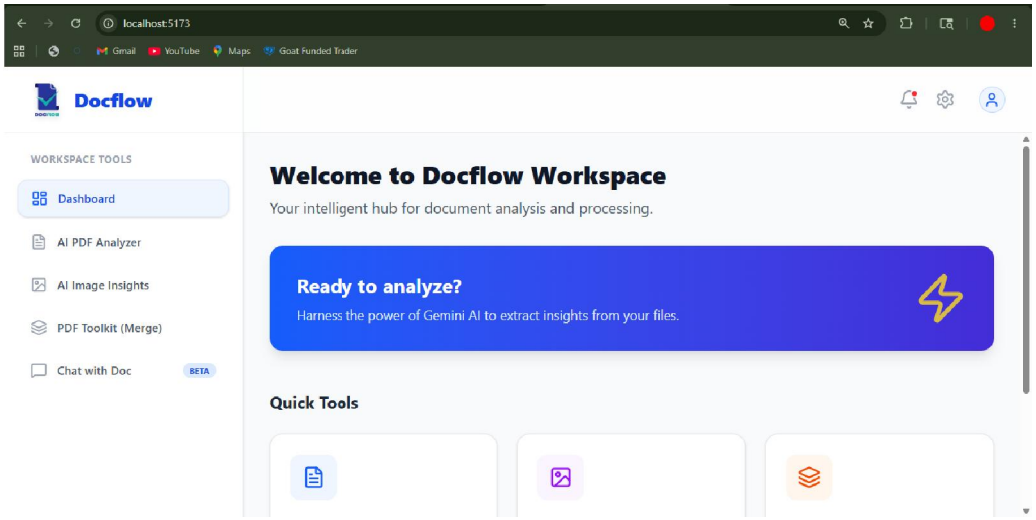

Fig. 3. Home screen UI

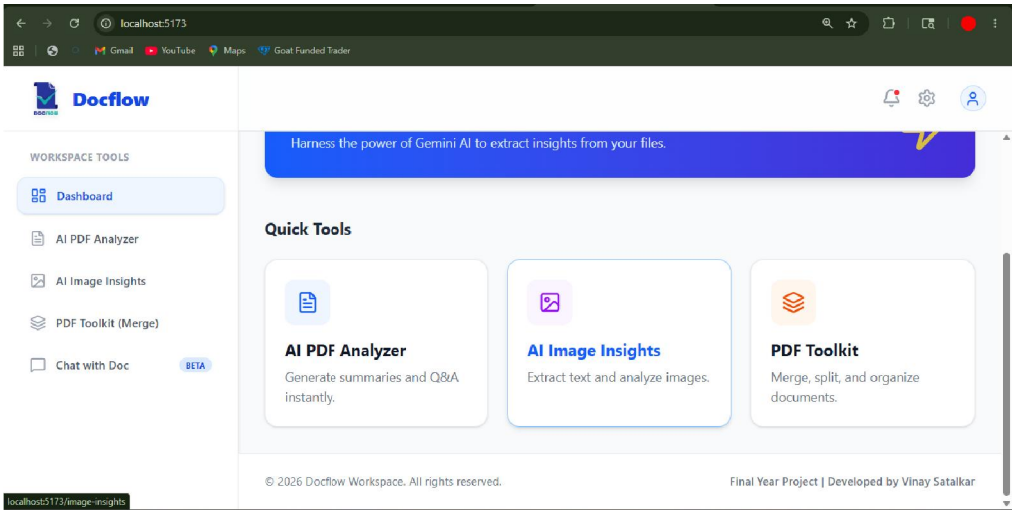

Fig. 4. Dashboard UI

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, May 2026
Impact Factor: 8.2

The Usability Feedback (preliminary):

While quantitative performance testing is part of the next development phase, preliminary usability feedback was gathered from a small group of peer students. The interface shown in Figures 1 and 2 was presented to the group. It was reported that the design was "clean and easy to understand" and "a significant improvement over using multiple websites." This feedback validates the project's core hypothesis that a unified platform is a desirable solution.

VI. CONCLUSION

The final development phase of Docflow, an intelligent and user-friendly web platform for automated document analysis, has been successfully completed. This stage was dedicated to designing and implementing a robust system architecture and a user-centric interface, specifically engineered to alleviate the persistent challenges of information overload and fragmented, inefficient document workflows. By synthesizing summarization, Q&A generation and translation into a singular, streamlined interaction, the platform successfully eliminates the friction associated with navigating scattered, single-function tools.

The chosen client-server architecture, utilizing a React.js frontend and a Node.js backend, has proven to be a highly scalable and resilient foundation, capable of supporting both current requirements and future feature integration. Preliminary usability feedback remains highly positive, with users characterizing the interface as intuitive and significantly more efficient than existing disjointed methods; this validates the core hypothesis that a unified, AI-driven platform is a necessary evolution in digital productivity. While this implementation effectively resolves the primary challenge of fragmented workflows, it also establishes the essential groundwork for more advanced capabilities. Future work will shift focus toward full-scale integration of the backend AI services, followed by rigorous quantitative performance testing to measure speed, accuracy and API latency under variable loads.

VII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their project guide, Prof. Kanade R. S., for his invaluable guidance, support and mentorship throughout this research. We would also like to thank Dr. A.P. Suryavanshi, Project Coordinator, for his support.

REFERENCES

[1]. Chitale T.N., Barate M.S., Gaikwad A.L., Parbhane A.B., Kanade R.S., "Docflow - AI-powered Document Summarizer + QNA", Hon. Shri Babanrao Pachpute Vichardhara Trusts, GOI Faculty of Engineering Kashti, 2025.

[2]. Mihalcea R., Tarau P., "TextRank: Bringing Order into Texts", Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, 404–411.

[3]. Rush A.M., Chopra S., Weston J., "A Neural Attention Model for Abstractive Sentence Summarization", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015, 379–389.

[4]. Vaswani A., Shazeer N., Parmar N., et al., "Attention Is All You Need", Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, 5998–6008.

[5]. Devlin J., Chang M.W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, 4171–4186.

[6]. Lewis M., Liu Y., Goyal N., et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation and Comprehension", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, 7871–7880.

[7]. Gemini Team, Google, "Gemini: A Family of Highly Capable Multimodal Models", Technical Report, Google AI, arXiv:2312.11805, December 2023.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 10, May 2026
Impact Factor: 8.2

[8]. Robinson S., Miller T., "Structured Prompting for Reliable JSON Output in Generative Models", IEEE Transactions on Software Engineering, March 2024, 50 (3), 215–228.
[9]. Touvron H., Martin L., et al., "Llama 3: Open Foundation and Fine-Tuned Chat Models", Meta AI Technical Report, April 2024.
[10]. Zhang A., Li Y., "Improving Factual Consistency in Abstractive Summarization using Knowledge Graphs", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, 345–356.
[11]. Wang B., Chen L., "Self-Correction and Refinement in LLM-generated Text for Enhanced Readability", Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), November 2024, 1120–1131.
[12]. Wei J., Zhou D., "Instruction Tuning and Chain-of-Thought Prompting: A Survey", Journal of Artificial Intelligence Research, February 2025, 88, 102–145.
[13]. Park C., Kim J., "Efficient Transformers for Long-Context Document Processing", Proceedings of the 2025 International Conference on Learning Representations (ICLR), May 2025.
[14]. Silva R., Gupta A., "Agentic Retrieval-Augmented Generation: Orchestrating Autonomous Workflows for Document Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, January 2026, 48 (1), 112–128.
[15]. Henderson D., Nguyen T., "Evaluating Long-Context Multimodal Reasoning in Large Language Models", Proceedings of the 2026 International Conference on Learning Representations (ICLR), May 2026, 882–895.
[16]. Patel S., Zhao X., "Factual Grounding and Hallucination Mitigation in Retrieval-Augmented Summarization", Journal of Artificial Intelligence Research, March 2026, 92, 45–67.
[17]. Tanaka H., Kim Y., "Optimizing Latency in Multimodal Document Parsing: A Comparative Study of Vision-Language Models", Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP), February 2026, 201–215.
[18]. O'Sullivan M., et al., "Scalable Vector Databases for Real-time Retrieval in Document-Centric AI", Journal of Systems and Software, April 2026, 210, 102–118.
[19]. Gomez-Perez J., "Dynamic Prompt Optimization for Structured Output in Legal and Technical Document Parsing", Proceedings of the 2025 International Conference on Computational Linguistics (COLING), December 2025, 1450–1465.
[20]. Chen Q., et al., "Adapting Large Language Models for Domain-Specific Document Comprehension: A Transfer Learning Approach", IEEE Access, November 2025, 13, 10567–10580.
[21]. Al-Farsi N., "User-Centric Evaluation Metrics for AI-Powered Summarization Platforms", Human-Computer Interaction Journal, February 2026, 41 (2), 302–320.
[22]. Nakamura K., "Robust Text Extraction from Low-Resolution Scanned Documents using Vision-Language Transformers", Proceedings of the 2025 International Conference on Document Analysis and Recognition (ICDAR), October 2025, 550–565.
[23]. Thompson L., "State of the Art in Multilingual Document Translation for Technical Literature", Journal of Natural Language Processing, January 2026, 32 (1), 89–105.

DOI: 10.48175/IJARSCT-35749