**Dyno-GenQA: Dynamic Summarization for Question-Answering using LLMs**

Jennifer Lo Foh Wei1, a), Hui-Ngo Goh1, 2, b), Choo-Yee Ting1, 3, c), Albert Quek1, 4, d) and Amy Hui-Lan Lim1, 5, e)

1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

2Centre for Natural Language Processing, CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

3Centre for Big Data and Blockchain Technologies, CoE for Advanced Cloud, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

4Centre for Image and Vision Computing, CoE for Artificial Intelligence, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

5Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

*b) Corresponding author: hngoh@mmu.edu.my  
a) jenniferlofw@gmail.com*

*c) cyting@mmu.edu.my*

*d) quek.albert@mmu.edu.my*

*e) amy.lim@mmu.edu.my*

**Abstract.** The development of artificial intelligence (AI) and natural language processing (NLP) has greatly improved how textual information is processed, particularly in simplifying large amounts of unstructured text. However, most document summarization tools and question answering (QA) tools still face important challenges. They often struggle to work well with distinct types of documents, various layouts, and content in multi-languages. Moreover, most tools do not support interactive features that allow users to explore the content more naturally. To address these challenges, this work introduces Dyno-GenQA, an approach that blend generative question answering with dynamic summarisation. The system is designed to produce context-aware summaries and support conversational interactions tailored to diverse document types. Through targeted evaluation on tenancy agreements and news articles, Dyno-GenQA evidence improved adaptability, usability, and relevance, offering a more flexible and user-centric solution for document summarization and question answering processing.

# INTRODUCTION

The evolution of AI and NLP has tremendously changed how human interact with textual data. Despite that, existing document summarizers and question answering tools still suffer limitations. They often struggle in dealing with diverse document types, structures, and languages, and lack the capability in engaging with users in a conversational manner. This shortcoming is particularly obvious in their ineffectiveness in generating summaries dynamically across domain and document types.

Dyno-GenQA, an approach that combine generative QA with dynamic summarisation. It is architecture to provide more flexible document processing solutions by tapping into the advantage of large language models (LLMs). The system can generate summary in the form of templates according to specific content and structure of each document.

The contributions of this work are as follows:

* A dynamic document summarization method integrating with QA capabilities using LLMs.
* A dynamic template generation mechanism that adapts to the content and structure of different document types, ensuring more accurate and tailored summarization
* Overcome the limitations of existing document summarizers by offering a comprehensive, flexible, and adaptive approach to document processing, centred on user needs.

The rest of the section is structured as follows. In section II, we briefly describe the related works in document summarization, question-answering and Retrieval Augmented Generation (RAG). Section III discusses the methodology of Dyno-GenQA approach whereas Section IV points out the results obtained upon experimenting the approach. Finally, Section V concludes the whole work.

# RELATED WORKS

The field of document processing has seen significant advancements with the advent of AI and NLP technologies. Traditional document summarizers and question-answering systems have been widely studied and implemented across various domains. However, these systems often exhibit limitations in handling various document types, structures, and languages, and lack the capability for conversational interaction. In fact, most of the document summarizer either using extractive or abstractive approach still return summary in textual format.

The introduction of LLMs, for instance, GPT-3 [1], [2] and BERT [3], [4] has improved the accuracy and relevancy of document summarization and question answering. Despite the foundational models, fine-tuned LLMs and Retrieval-Augmented Generation (RAG) techniques has also enhance the performance of question answering tasks [5], [6]. There was a study that improved intent recognition and response generation in the financial domain by introducing the LB-KBQA system that integrates LLMs and BERT [7].

Over the years, QA systems have evolved from rule-based approaches [8], [9] to extractive, abstractive till generative models approaches. Extractive QA systems focused on retrieving direct answers from text passage, often lack of context incoherence. As the field of evolved, LLMs brought out new capabilities in understanding and generating human-like responses.

The presence of RAG has enhanced QA systems by integrating external knowledge retrieval from databases with generative models enables more accurate and contextually correct answers [10], [11]. Recent research have also shown the effectiveness of combining LLMs with RAG for domain-specific question-answering. The studies shows that it will significantly improve the quality and relevance of generated answers[12], [13]. This highlights the evolution of generative models will shift towards dynamic, adaptive QA systems, such as the proposed Dyno-GenQA, which leverages LLMs and dynamic summarization to overcome the limitations of pure extractive and abstractive approaches, which in turn provides tailored and context-aware responses.

In the field of document summarization, several innovative ideas have been introduced to improve the handling of complex and lengthy text. This include event-keyed summarization [14], [15] extractive summarization techniques [16], [17], [18]. However, these methods still face challenges in adapting the diversity of content and formatting of different document types.

Prompt engineering is an art of asking questions to LLMs with the intention to obtain desired output. Among all techniques, chain-of-thought and in-context prompting produce more accurate and concise summaries [19], [20]. In such, the proposed Dyno-GenQA aims to fill this gap by leveraging LLMs to provide a tailored document summaries and conversational interactions.

# METHODOLOGY

This section describes the overall system architecture, explains the processes of dynamic template generation, and outlines the implementation of generative question answering component of Dyno-GenQA as can be seen in Figure 1, consisting of the following key components:

1. **Input**: User may upload text document in TXT or DOCX format. The uploaded document temporarily stores as a string variable in the memory for further processing.
2. **Document Type Classification**: The classification process starts with extracting the document content by using the *Text Document Reader* module. The Document Classifier then applies prompt engineering techniques to guide the LLMs in analysing the content and determining the document type into the predefined categories as shown in Table 1.

A diagram of a document

AI-generated content may be incorrect.**FIGURE 1**. Architecture of Dyno-GenQA

|  |  |
| --- | --- |
| TABLE 1. Predefined categories | |
| Document Type | Predefined Categories |
| Tenancy Agreement | Tenancy Agreement |
| News Article | World |
| National |
| Business |
| Technology |
| Sports |
| Politics |
| Culture |

1. **Summarization Template Generation**: This step generates a customized summary template tailored to the specific content and structure of the document once the document type is identified. A template is dynamically generated for one document type using prompt that instruct the LLMs to focus on key aspects relevant to the document type.

For example, a zero-shot prompt for document classification is phrased as: “*Which document type best describes this document? Choose one from the following: Tenancy Agreement, News Article*”.

In the case of news articles from the national category, the LLMs is prompted as follows: *"Summarize the following news into a table with five columns: 'Main Idea,' 'Date and Location,' 'Government Response,' 'Public Reaction,' and 'National Impact.*

1. For more specialised case, few-shot prompting is implemented. For instance, “*Which category best describes this news article? Choose one from the following: World: International events, global affairs, foreign policy. National: News and events within a specific country. Business: Financial markets, companies, economic trends.*”.

It uses a pre-trained LLM, the Gemini 1.5 Flash, to generate answers that match the user’s question and the document context. Google Gemini 1.5 Flash is a state-of-the-art LLMs developed by Google. It is trained on large datasets using deep learning methods. As a result, it performs well in tasks like text generation, summarization, and question answering. Unlike extractive summarization systems that copy the exact text from documents. The Dyno-GenQA system generates responses by understanding the context of the content, which allows it to provide more natural and informative answers.

1. **Summarize the Document Based on the Generated Template**: This step produces a summary of a document using the dynamic template created in step 3. It ensures that the summary captures the important points of the document, providing users with an extensive overview.

The Dyno-GenQA approach is flexible, and its modular architecture allows it to adapt to various document types and user queries, providing an inclusive and adaptive document processing solution.

# EVALUATION

## Exploratory Data Analysis

The proposed Dyno-GenQA was evaluated on two different types of documents, (a) tenancy agreements and (b) news articles. An exploratory data analysis was carried out to examine the structure and content of these documents. This analysis informed the design of summarization and QA templates, allowing them to be tailored to the specific features of each document type.

### Tenancy Agreement

The tenancy agreement dataset consists of 27 documents collected from Kaggle. It served as the foundation for developing both the dynamic template summariser and the document-based generative QA system. These documents vary in length, ranging from 239 to 3,311 words, and collected from three countries of Bangladesh, India, and Malaysia. Although most of the tenancy agreements are in plain text, a few contain images such as signatures and stamp duty marks, which are commonly found in legal documents. Each agreement is structured in paragraphs, which supports consistent processing. Nevertheless, not all documents are complete. Some are missing important details like either the lease term or full address, reflecting the typical inconsistencies found in real-world legal content. Not to compromise, for tenancy agreements, certain elements were found to be important for summarisation, such as *names of the parties involved, the lease terms,* and *payment details*.

### News Articles

The dataset includes 30 English news articles collected from Kaggle, all originally published by The Star, a well-known English-language newspaper in Malaysia. These articles cover a wide range of topics and vary in structure, length, and writing style, offering a realistic view of everyday news reporting. Each article also comes with useful metadata, such as the publication date, location, and source, which supports deeper analysis.

Although all articles were initially placed under the ‘Nation’ section, the Dyno-GenQA reclassified them into more specific categories—National, Culture, and Politics—by applying prompt engineering techniques. The length of the articles ranges from 68 to 350 words, reflecting the different levels of detail typically found in news writing.

One common issue found in the dataset was the inconsistency in reporting format of dates and location. The Dyno-GenQA addressed this by identifying the actual event location and interpreting time references like “yesterday” or “a month ago” to extract the correct incident date.

For summarisation, several elements were found to be especially important. These include the *date, location, main idea, government response, public reaction*, and *national impact*. By focusing on these points, Dyno-GenQA can generate summaries that are clear, relevant, and easy to understand—helping users quickly grasp the key information and structure of each article.

## Evaluation Process

The evaluation focused on how useful and clear the generated summaries were to human readers. It did not measure the technical accuracy of the LLMs. Three human evaluators took part in the assessment. The evaluators reviewed each summary and rated it based on how clear, relevant, and useful it was. The summaries were not judged against a fixed correct version. Instead, the evaluators checked whether the main ideas from the original documents were included. Readability and content organization were also considered in the evaluation as a good summary was expected to be easy to read, well-structured and concise.

Each field in the generated summaries was reviewed on its effectiveness in conveying the essential information, and a satisfaction rating is calculated using the Equation (1). Suggestions for improvement were noted when the summaries did not meet the evaluators' satisfaction criteria. This approach focused on ensuring that the generated summaries were practical and meaningful from the user's perspective.

(1)

In the evaluation form, each generated summary template comprised five distinct fields, resulting in a total of 150 fields to be evaluated. As outlined in Equation (1), the accuracy of the generated summaries was calculated by dividing the number of ticks received by the total number of fields available.

# Results AND DISCUSSION

The Dyno-GenQA demonstrates varying levels of accuracy across different document types, indicating its ability to adapt and perform well in diverse contexts. The accuracy of Dyno-GenQA generated summary template are shown in Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TABLE 2 Evaluation result | | | | |
|  | Dataset | | | |
| Tenancy Agreement | National News | Politics News | Culture News |
| Accuracy | 93% | 93.3% | 88% | 95.3% |

The developed Dyno-GenQA demonstrated an overall satisfaction rate of 92.2% in dynamic template generation and summarization for news articles across various categories, while achieving a 93% human satisfaction rate with the tenancy agreement dataset. These results indicate that the system is highly proficient in extracting and representing the essential information from the provided documents within the generated summary templates. Figures 2 and 3 show the prototypes of Dyno-GenQA on template-based document summary and QA respectively.

## Discussion

Based on the user study, it shows that the Dyno-GenQA system has several strengths:

* **Adaptability**: It can create templates that match the structure and content of different types of documents, without needing human modification.
* **Conversational Engagement**: It allows users to ask questions and get response in a natural chat-like way to helps users understand the content better.
* **Enhanced Summarization**: The system uses template to create short by complete summaries. These summarise highlight the key points without giving too much information.

# Conclusion

Dyno-GenQA serves as a considerable improvement in automated document processing by integrating generative QA with dynamic summarisation using LLms. It tackles the limitations observed in traditional QA and summarization frameworks that rigid and constraint to document and structure specific. The modular architecture of Dyno-GenQA enables the adaptive template formation suited to variety of document types and structures.

The working prototype of the Dyno-GenQA is tested on (a) tenancy agreements and (b) news articles. The evaluation results showed improvements and satisfaction in the quality of the summaries and the user’s engagement with the prototype. This is due to its capability in producing responses that reflect the context of each document. In addition, the produced summaries are also according to the structure and content of the input document. These features make Dyno-GenQA a practical tool for real-world document processing tasks.

Nonetheless, the system does occasionally produce inconsistent answers which tentatively lacks the flexibility needed to handle a wider variety of document types. It highlights the need for further refinement to enhance its reliability and scalability.

In conclusion, Dyno-GenQA lays a foundation for future research in document understanding systems by integrating LLMs for document processing solutions that are more adaptive, accurate, and user centric.

A screenshot of a computer

AI-generated content may be incorrect.

**FIGURE 2**. Document summary with Dyno-GenQA

A screenshot of a computer

AI-generated content may be incorrect.

**FIGURE 3**. Generative QA with Dyno-GenQA

# REFERENCES

1. T. Goyal, J. J. Li, and G. Durrett, “News summarization and evaluation in the era of GPT-3,” (2022).

2. J. Wu, S. Wang, Z. Yu, R. Zhang, and Y. Tian, “Recursively summarizing books with human feedback,” (2021).

3. S. Abdel-Salam and A. Rafea, “Performance study on extractive text summarization using BERT models,” Information **13**, 67 (2022).

4. X. Zhang, F. Wei, and M. Zhou, “HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, Florence, Italy, 2019), pp. 5059–5069.

5. L. Zhang, H. Chen, Y. Wang, and P. Liu, “Enhancing large language model performance to answer questions and extract information more accurately,” (2023).

6. S. Liu and J. Wu, “Towards a robust retrieval-based summarization system,” (2022).

7. Y. Zhao, Z. Li, Y. Pan, J. Wang, and Y. Wang, “LB-KBQA: Large-language-model and BERT based knowledge-based question and answering system,” (2023).

8. P. Katyayan and N. Joshi, “Design and development of rule-based open-domain question-answering system on SQuAD v2.0 dataset,” in Proceedings of the 3rd International Conference on ICT for Digital, Smart, and Sustainable Development (EAI, New Delhi, India, 2022).

9. E. Riloff and M. Thelen, “A rule-based question answering system for reading comprehension tests,” in ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems (Association for Computational Linguistics, Seattle, WA, 2000), pp. 13–19.

10. A. Y. Alan, E. Karaarslan, and O. Aydin, “A retrieval-augmented generation-based question answering system proposal for understanding Islam: MufassirQAS LLM,” (2024).

11. F. Pan, M. Canim, M. Glass, A. Gliozzo, and J. Hendler, “End-to-end table question answering via retrieval-augmented generation,” arXiv:2203.16714 (2022)

12. C. Wang, Y. Liu, M. Li, H. Zhang, and X. Sun, “BioRAG: A RAG-LLM framework for biological question reasoning,” arXiv:2408.01107 (2024)

13. Z. Hei, J. Sun, Y. Zhao, L. Xu, and H. Chen, “DR-RAG: Applying dynamic document relevance to retrieval-augmented generation for question-answering,” (2024).

14. W. Gantt, A. Martin, P. Kuchmiichuk, and A. S. White, “Event-keyed summarization,” arXiv:2402.06973 (2024)

15. J. Zhang, K. Li, C. Yao, and Y. Sun, “Event-based summarization method for scientific literature,” Personal and Ubiquitous Computing **25**, 959–968 (2021).

16. N. Gu, Y. Gao, and R. H. R. Hahnloser, “MemSum-DQA: Adapting an efficient long document extractive summarizer for document question answering,” arXiv:2310.06436 (2023)

17. S. Ghodratnama, A. Beheshti, M. Zakershahrak, and F. Sobhanmanesh, “Extractive document summarization based on dynamic feature space mapping,” IEEE Access **8**, 139084–139095 (2020).

18. W. Xiao and G. Carenini, “Extractive summarization of long documents by combining global and local context,” arXiv:1909.08089 (2019)

19. J. Wang, Y. Lin, M. Zhou, F. Zhang, and T. Li, “Prompt engineering for healthcare: Methodologies and applications,” arXiv:2304.14670 (2024)

20. J. Wei, X. Wang, D. Schuurmans, Q. Le, and T. Chen, “Chain-of-thought prompting elicits reasoning in large language models,” arXiv:2201.11903 (2023)