Article

Optimizing QA Systems: Evaluating Row-Based and Traditional Chunking in Structured-Data-Aware Retrieval-Augmented Generation for University Virtual Assistants

Maksat Maratov [1*] and Selchuk Cankurt [2]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan
[2]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/7we5dg32

Abstract

This paper presents the development of a question-answering system that can assist university students with academic and administrative questions. We present a new approach that examines various chunking approaches to the Retrieval-Augmented Generation process. Although RAG is typically used with standard chunking methods, this paper presents row-based chunking, tailored to structured question-answer datasets, in order to enhance context retrieval for large language models. To establish its effectiveness, we conducted a human evaluation to compare the outputs it generated with those generated using standard and row-based chunking. The individuals who tested our system were both students and educators at the university. We concluded that row-based chunking gives more coherent and relevant contextual data than standard ways of chunking when applied to structured data sets. This work highlights the potential of using chunking methods to improve RAG-based systems for domain-specific applications, paving the way towards more accurate and context-sensitive AI-based aid in educational settings.

Keywords: Q&A system, Virtual Assistant, ChatBot, RAG, NLP, Chunking strategy, LLM powered Chatbots

*Corresponding author: maksat.maratov@sdu.edu.kz
Email: maksat.maratov@sdu.edu.kz ORCID: 0009-0004-8511-5014
Email: selcuk.cankurt@sdu.edu.kz ORCID: 0000-0003-0581-1913

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

I. Introduction

In businesses and other sizable organizations, efficient question-answering (QA) systems and support services are essential to manage large amounts of information and user interactions. While most organizations maintain specialized support centers, the quality and usefulness of such services diminish with an increase in the volume of inquiries. This issue is particularly acute at universities, where official support centers do address academic and administrative inquiries but may not be familiar with student-generated topics such as campus events, extracurricular activities, and student life dynamics. Therefore, students rely on peer-shared knowledge, which is decentralized and difficult to formalize within traditional support systems.

In recent years, Large Language Models (LLMs) have become deeply integrated into various aspects of daily life [1], with users increasingly preferring text-generation tools over conventional search engines for information retrieval. However, LLMs struggle with domain-specific, private, or real-time data, leading to hallucinations and misinformation in cases where such knowledge is not explicitly encoded in their training corpus [2]. To address this limitation, the Retrieval-Augmented Generation (RAG) framework was introduced. RAG enhances LLMs by integrating an external knowledge retrieval mechanism, typically consisting of three key components: indexing, retrieval, and generation [3].

RAG has been widely adopted across large enterprises and knowledge-driven organizations to improve factual accuracy and provide real-time, dynamic responses from large text corpora. Various RAG implementations exist, differing primarily in their approaches to indexing, retrieval, and response generation MMed-RAG [4] HiTA [5] FinTMMBench [6] and OmniEval [7]. Some systems employ different indexing techniques, such as keyword-based, dense vector-based, or hybrid search methods, while others vary in their retrieval strategies or choice of language model (LLM) for generation.

However, most existing RAG methods are optimized for unstructured text, relying on general-purpose chunking strategies such as RecursiveTextSplitter, fixed-length chunking, and semantic-based chunking. These chunking approaches divide text into predefined sizes or semantically coherent segments, which may work well for free-form documents but are poorly suited for structured data such as spreadsheets, relational databases, or question-answer (Q&A) tables. In structured datasets, preserving the integrity of data relationships is crucial, as conventional chunking methods risk fragmenting semantically dependent information, leading to retrieval mismatches and inaccurate responses.

In contrast, this work introduces a structured-data-aware RAG approach that optimally handles tabular data by treating each row as a single chunk, rather than using arbitrary chunk size constraints. This approach ensures:

- Preservation of data integrity – Each Q&A pair remains intact, avoiding fragmented information retrieval.
- Efficient retrieval and alignments of the embedding – By using whole rows as a chunk the similarity search operations become more precise.
- Reduction in unnecessary processing overhead – Eliminates the need for reconstructing structured data from fragmented chunks.

Our approach is particularly beneficial for scenarios involving structured Q&A datasets, where maintaining the original structure of the data is crucial for accurate retrieval and answer generation from the LLMs.

II. Review of Related Works

It is essential to learn about the history of question-answering systems prior to addressing chunking and information retrieval strategies. Computer question answering benchmarks were defined at the Text Retrieval Conference (TREC) in 1999, one of the initial benchmarks for QA as a field. No matter which subject is envisioned, the purpose was to return short answers to factoid and enumerative questions [8]. In an effort to make the output more precise, conventional QA systems employ structured information retrieval and categorization [9]. Our study expands on this by including structured data retrieval for university-based Q&A into a RAG framework.

The utilization of structured knowledge sources to increase answer accuracy has been the main focus of recent developments in QA systems. Derici and associates. [10] proposed HazırCevap, a closed-domain QA framework that retrieves answers from reliable educational resources while also utilizing multilingual support through translation. Unlike open-domain QA systems, HazırCevap specifically caters to students by ensuring accuracy through a curated knowledge base. However, it relies on document summarization rather than dynamic retrieval-augmented generation

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

(RAG), which limits its ability to adapt to diverse and evolving queries. Our work extends this by leveraging RAG to retrieve and generate answers in real-time, ensuring both accuracy and contextual relevance.

Retrieval-Augmented Generation (RAG) integrates parametric (pre-trained LLM) and non-parametric (retrieved external data) memory to improve knowledge-intensive tasks [11]. The retrieval module locates relevant information, while the generation module conditions on retrieved context to generate more factual responses. Such patterns of work are observed in most RAG methods. In more detail, there are three main components: indexing, retrieval, and generation. The retrieval module locates relevant information using dense or sparse search, while the generation module integrates this context to produce an accurate response.

Despite the effectiveness of RAG in enhancing factual consistency, not all RAG models are well-suited for structured data retrieval. Many existing implementations are optimized for unstructured text, where chunking strategies such as RecursiveTextSplitter [12] or fixed-length segmentation are commonly employed. While these methods work well for free-form documents, they introduce fragmentation issues when applied to structured datasets like university Q&A tables. For instance, RAG implementations that rely on naive text chunking may separate a question from its corresponding answer, leading to retrieval mismatches and incoherent responses. Furthermore, models such as Hybrid-RAG [13] and ActiveRAG [14] attempt to improve retrieval by incorporating iterative refinement, but they remain inefficient when handling structured data fields due to their reliance on unconstrained semantic search [15].

BM25, a probabilistic information retrieval model, ranks documents based on query terms but may not effectively handle the nuances of structured data [16]. In contrast, our approach preserves the Q&A pair data by treating each row as a single retrieval unit, ensuring accurate and contextually consistent responses.

While various RAG implementations focus on enhancing accuracy, retrieval mechanisms, and source attribution, they do not consider structured Q&A pair data. Traditional RAG frameworks primarily process unstructured documents, making them unsuitable for applications where preserving data relationships—such as university Q&A datasets—is critical. Some studies focus on reducing contradictions in retrieved knowledge and self-reflecting on results, improving reliability like SelfRAG [17], ActiveRAG [14], and InstructRAG [18].

However, these approaches do not address the challenges of structured data retrieval, particularly in handling Q&A pair formats. While some works focus on structured or semi-structured data, they primarily target entity-based retrieval, tabular knowledge representation THoRR [19], or knowledge graphs FastRAG [20], rather than optimizing chunking strategies for structured text. Existing methods fail to consider how row-wise chunking can preserve data integrity in structured datasets, such as university Q&A tables, where each row represents a complete and independent knowledge unit.

III. Methods

In this study, we address the challenge of structured data retrieval in Retrieval-Augmented Generation (RAG) systems by leveraging a university-specific Question-Answer (QA) dataset. Unlike traditional RAG models that process unstructured text, our approach preserves the integrity of structured QA pairs, ensuring accurate and contextually relevant responses.

A. Dataset

The dataset consists of approximately 20,000 QA pairs collected from university students. It covers a wide range of university-related topics, including academic inquiries, student life, event details, club activities, and problem-solving scenarios such as lost ID cards or course registration procedures. Since the dataset is user-generated, it includes variations in phrasing, with some questions appearing in multiple interpretations or with additional details. These variations enhance the model's ability to retrieve contextually appropriate responses.

The dataset includes three languages: Kazakh, Russian, and English. Before training, the data underwent preprocessing, including the removal of stopwords, conversion to lowercase, and other standard NLP cleaning techniques to ensure consistency. Duplicate entries were filtered, while semantically similar but non-identical questions were retained to improve retrieval diversity.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

B. Structured Retrieval and Chunking Strategy

Our RAG implementation deviates from traditional chunking methods, such as character-based, recursive, or semantic splitting. Instead, given the structured nature of our dataset—comprising QA pairs—we treat each row of data as an independent chunk. This ensures that the full context of each question-answer pair remains intact, preventing the fragmentation issues commonly observed in unstructured chunking approaches. By maintaining complete QA pairs as single retrieval units, we preserve the semantic integrity of responses, which positively impacts retrieval accuracy.

After chunking, we proceed with embedding the data for vector-based retrieval. Since our dataset contains content in three languages (Kazakh, Russian, and English), we employ a multilingual embedding model, intfloat/multilingual-e5-large, which is widely adopted for cross-lingual tasks due to its strong performance across a broad range of languages. This model was chosen for its balance between quality and efficiency, and because it has demonstrated robust multilingual retrieval capabilities in both academic benchmarks and practical applications. Although we did not conduct an independent embedding evaluation, we employ it because it is widely adopted for cross-lingual tasks and performs strongly across languages.

For indexing, we utilize VectorStoreIndex, a widely used vector database approach that allows efficient similarity-based retrieval. Each QA pair is stored as an embedding, enabling rapid lookup of semantically similar chunks during the retrieval process.

During retrieval, an input question is first embedded using the same multilingual embedding model. The system then computes the cosine similarity between the query embedding and all indexed QA pair embeddings, selecting the top-K most relevant results. These retrieved QA pairs serve as context for the final answer generation, ensuring that the response is based on the most semantically similar knowledge available.

1) Response Generation:

For the response generation phase, we integrate OpenAI's GPT-4o as the language model. To ensure the model behaves as a university virtual assistant, we apply prompt engineering techniques. The prompt includes:

- Zero-shot learning strategies to help the model generalize across diverse university-related queries.
- Background information about the university to provide institution-specific responses.
- Rules and regulations for handling specific student-related scenarios (e.g., lost ID cards, course registration issues).

This carefully designed prompt ensures consistency in the responses and is used uniformly across all evaluated methods to maintain fairness in comparisons.

The primary motivation behind our chunking strategy is to preserve the full context of each QA pair, avoiding the fragmentation issues introduced by traditional chunking methods. Standard approaches such as RecursiveTextSplitter segment documents based on arbitrary character or semantic boundaries, often leading to incomplete or disjointed retrieval results. In contrast, our row-wise chunking ensures that each QA pair remains intact, providing a more semantically meaningful retrieval unit.

Moreover, while semantic chunking techniques attempt to create contextually coherent splits, they often struggle with multilingual datasets due to limitations in cross-lingual sentence embedding models. This challenge can lead to poor retrieval performance when queries and indexed documents exist in different languages. In our study, we will empirically compare the top-K retrieval performance metrics between our structured chunking method and conventional approaches. Specifically, the top-K value will be set to 20. This will allow us to evaluate and demonstrate the effectiveness of our method in enhancing retrieval accuracy and performance, particularly in the context of university Q&A datasets.

C. Comparison with Traditional Chunking

To establish a meaningful comparison, we evaluate our structured row-based chunking method against the conventional traditional chunking approach, which segments text into fixed-sized chunks or employs semantic splitting strategies. Traditional chunking methods, while widely used, often introduce inconsistencies by fragmenting contextually related information, potentially leading to loss of coherence in retrieval tasks.

Since our dataset is inherently structured in a row-based format, a direct comparison requires adapting the traditional chunking method to a relevant representation. For this, we approximate an unstructured document

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

format through the conversion of the dataset to a continuous text-based QA format, approximating how data would typically be retained in unstructured documents. This step takes care that both chunking methodologies are evaluated on the same premises.

By presenting the dataset in this form, we are able to test how well the traditional methodology recovers useful responses and preserves contextual coherence compared to our row-based method. By doing so, we are highlighting the limitations of applying the traditional chunking method to structured data and comparing them with the advantages of a retrieval-nominated chunking method.

IV. Results and Discussion

To evaluate the performance of proposed chunking method against the traditional approach of RAG, a small-scale study was conducted involving 10 participants, comprising both students and teachers at SDU Univeristy. Participants were asked a series of questions related to university life, operations, and logistics, in three languages: English, Kazakh, and Russian. The primary goal was to test whether our chunking strategy—where each chunk is a full Question-Answer (QA) pair—leads to more relevant and precise information retrieval compared to traditional chunking, which segments the text arbitrarily or by fixed-length windows.

A total of 6 multilingual questions were used as test inputs:

1) What is EPT?
2) Потерял ID карту. Что мне делать?
3) СДУ университетінде көлік тұрағы бар ма? Егер бар болса, студенттер көліктерін қоя алады ма?
4) Өзімнің ағылшын деңгейімді қалай көрсем болады?
5) Что если у меня есть пересечения в расписании файналов?
6) Give me step-by-step instructions of how to get the book from the SDU library.

The results presented below highlight only the cases where there was a noticeable difference between the two retrieval methods. In some cases, such as locating specific teacher offices or retrieving department-specific information, both methods failed to retrieve a relevant answer, which indicates a lack of data coverage rather than method inadequacy.

A. Traditional Methodology

User Request 1: What is EPT? [label=chunk1] Q: What is an EPT in SDU? A: English Proficiency Test: SDU offers an English Proficiency Test (EPT) to assess students' ... [56 words]

Q: SDU-да EPT дегеніміз не? [7 words]

Q: Что такое олимпиада SPT? [7 words]

They also give grades for the task that were loaded. Q: What benefits and opportunities does SPT (Profile Testing System) provide for school and college graduates? A: SPT — is a test that is conducted in the specialized subjects of UNT for high school and college students... [90 words]

nursing homes, orphanages 3)marvel - Organizer 4)handmade - Craft Direction 5)cooking-cooking a variety of meals together Q: What kind of race is SPT? A: SPT is a competition for 11th grade Students... [135 words]

User Request 2: Потерял ID карту. Что мне делать?

Q: Что делать если забыл ID карту? [9 words]

Q: Если я забыл дома свою ID карту, то как могу войти в универ? [16 words]

Q: В каких местах я могу использовать ID-карту? [9 words]

Q: Что будет, если студент потерял свою карту ID? Будет ли допуск на экзамены? [16 words]

Q: Если вы потеряли ID-карту, вы можете пойти в центр обслуживания студентов, чтобы создать ее... [132 words]

User Request 3: Сду университетінде көлік тұрағы барма? Егер бар болса студенттер көліктерін қоя аладыма?

Q: Салынынпты әлі білмесе, СДУ-да кеңселерді қалай табуға болады? A: Егер студент кабинетті таба алмаса, ол [17 words]

Q: Студенттер үшін көлік қандай нұсқалар бар? [9 words]

Q: Ата-анам маған көлік сатып алды. Мен университеттің тұрағына қоя аламын ба? A: SDU аумағындағы автотұрақ Қызметкерлер мен қонақтарға арналған... [31 words]

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

Q: Шегуге бола ма? A: Университет университет аумағында темекі шегуге немесе электронды құрылғыларды [12 words]

Q: Ғимаратында не орналасқан? A: SDU Life ғимаратында студенттерге [6 words]

B. Row-Based Chunking Method

In contrast, the row-based chunking method significantly enhances retrieval by preserving complete question-answer pairs:

User Request 1: What is EPT?

[label=chunk1] SDU offers an English Proficiency Test (EPT) to assess students' English language skills. The test is typically required for admission to English-medium programs or for students seeking exemptions from English language courses. It evaluates ... the test if needed. Additionally, SDU may offer English language clubs or resources to help students improve their language skills and prepare for the EPT. [76 words] '

What is an English proficiency test? It's an exam to take for an exchange program. [9 words]

SPT is a competition for 11th grade students. Through the competition, you can win an internal grant. Even on the day you don't win, you will be given a discount on paid education. At first, the competition will be based on math literacy, and then you will pass it according to your professional subject. [44 words]

When students apply to the university, they need to take an English test to see how well they know the language. There are two parts to the test: 1) Grammar test: consists of 50 questions; 2) A speaking part ... If a student's English level is below that, they'll have to pay for courses to improve it. Currently, each level costs 102,000 KZT. [91 words]

Typically, 1 ECTS is equal to 25-30 hours. This is an indicator of the value of the course. [20 words]

User Request 2: Потерял ID карту. Что мне делать?

Что делать если забыл ID карту? [9 words]

Сначала нужно оповестить эдвайзера и сделать запрос на восстановление ID карты. [16 words]

Студентам необходимо посетить сервисный центр. [7 words]

Вам следует обратиться эдвайзеру, через некоторое время вам выдадут новую. [15 words]

ID-карта, центр обслуживания студентов, QR-код, 2000 тенге. [8 words]

User Request 3: Сду университетінде көлік тұрағы барма? Егер бар болса студенттер көліктерін қоя аладыма?

Университетте көлік тұрағы бар, университеттің ауласында орналасқан. Өкінішке орай студенттер көліктерін қоя алмайды. Только преподавателям разрешается парковать свои автомобили. [33 words]

Иә, SDU университетінің ауласында көлік тұрағы бар. Бірақ та студенттер ол көлік тұрағын қолдана алмайды. Тұрақ мұғалімдерге арналған. [22 words]

SDU аумағындағы автотұрақ қызметкерлер мен қонақтарға арналған. Студенттер оған көліктерін қалдыра алмайды. [18 words]

Өкінішке орай, егер сіз университет қызметкері болмасаңыз, университет аумақтарына көлік қоя алмайсыз. Бірақ сіз көлігіңізді университеттің жанында арнайы тұрақ орындарына қоя аласыз. [27 words]

Иә, биыл тегін. [5 words]

As shown above, the row-based chunking method proves to be more effective in retrieving relevant information aligned with user queries. By maintaining the integrity of each question-answer pair, it allows the retriever to surface more meaningful and contextually complete results.

C. Generation

The next step in the pipeline is generation, where the impact of the input format becomes even more evident. Below, we compare the responses generated by the language model using the two different retrieval inputs: Traditional Chunking [T] and Row-Based Chunking [R].

Question 1

Q: What is EPT? Response [T]: EPT is the English Proficiency Test offered to assess students' English language skills. It is typically required for admission to English-medium programs or for students seeking exemptions from English language courses. The test evaluates proficiency in listening, reading, writing, and speaking. Response [R]: The English Proficiency Test (EPT) at SDU is designed to assess students' English language skills. It is typically

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

required for admission to English-medium programs or for students seeking exemptions from English language courses. The test evaluates proficiency in listening, reading, writing, and speaking. Students can register for the test through the university's website or designated testing centers. The format may include multiple-choice questions, writing essays, and oral interviews. Results are usually available within a specified timeframe, and students may have the opportunity to retake the test if needed. Additionally, SDU may offer English language clubs or resources to help students improve their language skills and prepare for the EPT. The test consists of two parts: a grammar test with 50 questions and a speaking part where an examiner asks questions on random topics to evaluate the student's English level. After the test, students receive a certificate indicating their English level. Most study programs require at least a B1 level. If a student's English level is below that, they will need to pay for courses to improve it, with each level costing 102,000 KZT.
Question 2

Q: Потерял ID карту. Что мне делать? Response [T]: Если вы потеряли ID карту, вам следует обратиться в Центр обслуживания студентов, чтобы запросить изготовление новой карты. Это будет стоить 2000 тенге, и процесс может занять некоторое время, так как карты изготавливаются партиями. Response [R]: Если вы потеряли ID карту в SDU, выполните следующие шаги: 1. Свяжитесь с консультантом (Advising Desk) и сообщите о потере карты. 2. Оплатите штраф за утерю карты, который составляет 2000 тенге. Оплату можно произвести через Kaspi QR или Halyk Bank. 3. Заполните форму через QR-код. 4. Когда карта будет готова, вам сообщат об этом по электронной почте. 5. Также вы можете обратиться в Студенческий сервисный центр (Student Service Center) для получения новой карты. Учтите, что карты изготавливаются партиями, и процесс может занять до двух недель.
Question 3: Сду университетінде көлік тұрағы барма?

Q: Сду университетінде көлік тұрағы барма? Егер бар болса студенттер көліктерін қоя аладыма? Response [T]: СДУ аумағындағы автотұрақ қызметкерлер мен қонақтарға арналған, сондықтан студенттер көліктерін ол жерге қоя алмайды. Алайда, СДУ аумағына жақын жерде көптеген тегін тұрақ орындары бар. Response [R]: Иә, SDU университетінде көлік тұрағы бар, ол университеттің ауласында орналасқан. Алайда, өкінішке орай, студенттерге бұл көлік тұрағын қолдануға рұқсат жоқ. Тұрақ тек мұғалімдер мен қызметкерлерге арналған. Студенттер көліктерін университеттің жанында арнайы тұрақ орындарына қоя алады. Бұрын студенттерге SDU ішінде тұрақ қоюға рұқсат етілген, бірақ орын тапшылығына байланысты уақытша тыйым салынды. Жақында жаңа блок ашылғанда, тұрақ қолжетімді болуы мүмкін.
Question 4: Өзімнің ағылшын деңгейімді қалай көрсем болады?

Q: Өзімнің ағылшын деңгейімді қалай көрсем болады? Response [T]: Өзіңіздің ағылшын деңгейіңізді көру үшін oldmy.sdu.edu.kz парақшасына кіріп, негізгі беттегі негізгі ақпарат бөлімінен ағылшын деңгейі туралы ақпаратты таба аласыз. Сіздің ағылшын деңгейіңіз сол бетте көрсетулі болады. Response [R]: Өз ағылшын деңгейіңізді көру үшін oldmy.sdu.edu.kz парақшасына кіріп, негізгі бетке өту арқылы негізгі ақпараттың ішінен ағылшын деңгейі туралы ақпаратты көре аласыз. Сіздің ағылшын деңгейіңіз сол бетте көрсетулі.
Question 5: Что если у меня есть пересечения в расписании файналов?

Q: Что если у меня есть пересечения в расписании файналов? Response [T]: Вам следует обратиться к вашему эдвайзеру, который может перенести один из экзаменов на другой день, чтобы устранить конфликт в расписании. Response [R]: Если у вас есть пересечения в расписании файналов, вам следует предпринять следующие шаги: 1. Обратитесь к вашему эдвайзеру, чтобы обсудить проблему. 2. Вместе с эдвайзером посетите деканат и напишите заявление о пересечении времени экзаменов. 3. Деканат рассмотрит ваше заявление и обсудит возможность изменения времени одного из экзаменов. 4. Если студентов с такой проблемой немного, для вас могут выделить отдельное время для сдачи экзамена с участием учителя и прокторов. 5. В некоторых случаях может потребоваться заполнение специальной формы и предоставление доказательств пересечения времени экзаменов. 6. Также можно обратиться в MDE центр в университете для назначения другого дня сдачи экзамена или сдачи с другой группой, но это нужно сделать заранее.
Question 6: Give me step by step instructions of how to get the book from the SDU library.

Q: Give me step by step instructions of how to get the book from the SDU library. Response [T]: To get a book from the SDU library, follow these steps: 1. Visit the SDU library website. 2. Use the search function to find the book you want by entering the book's name and author. 3. Check the availability status of the book. If it's available, note the location details such as the floor and key letters. 4. Go to the library with your ID card. 5. Locate the book using

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

the information you noted. 6. Use your ID card to check out the book at the library's checkout desk. Response [R]: To get a book from the SDU library, follow these step-by-step instructions: 1. Visit the SDU Library: Go to the SDU library, which is a 3-story building located on the university campus. 2. Find the Book: You can either search for the book online on the SDU library website or physically browse the library shelves. If searching online, enter the book's name and author in the search bar to check its availability and location within the library. 3. Locate the Book: If the book is available, note the floor and section where it is located. Use the key letters provided to find the exact shelf. 4. Borrow the Book: Once you have the book, go to the librarian or a consultant at the library. Present your student ID card to the librarian. This is necessary to borrow the book. 5. Discuss the Borrowing Period: Talk to the librarian or consultant about how long you wish to borrow the book. Typically, books can be borrowed for a period of 3 weeks. 6. Take the Book Home: After the librarian processes your request, you can take the book home. Remember, you can borrow up to five books at a time. 7. Return the Book: Ensure you return the book by the due date to avoid fines. If you lose the book, you will need to pay a fine or replace it.

D. Evaluation

Apart from retrieval and generation measures, a human test was conducted to study the quality and usability of the assistant's responses. This test focused on the way users felt about the accuracy, relevance, and usefulness of the answers they received.

1) Participants:
The evaluation involved approximately 50 participants, comprising both students and staff members from various departments within the university. It is important to note that this survey is distinct from an earlier preliminary survey, which involved 10 participants. The initial survey was conducted to collect user-asked questions and evaluate the quality of LLM-provided answers based on a chunking strategy approach. In contrast, the current survey focuses on a comparative evaluation of two different approaches—traditional chunking and row-based chunking—using a structured set of evaluation criteria.

2) Manual Evaluation:
A manual evaluation was performed to assess the virtual assistant's capabilities.

TABLE I
Structured Evaluation Approach for AI-Generated Responses

No.	Evaluation Question	Grading Method
1	Did the prediction contain any hallucinations?	Binary response: Yes or No
2	Assess the relevance of the response to the question.	Rating scale: 1 to 5
3	Evaluate the content size and structure of the response.	Rating scale: 1 to 5
4	Did you identify any logical inconsistencies in the response?	Binary response: Yes or No
5	What is your overall evaluation of the responses?	Rating scale: 1 to 5

Table I. presents a structured approach to the quality assessment of AI responses. It contains both binary (Yes/No) tests and scaled ratings (1 to 5) to ensure a complete analysis.

- Binary questions help identify significant issues such as hallucinations (false or misleading information) and fallacies.
- Rating questions allow for a finer-grained assessment of aspects like relevance, content structure, and overall quality.

These evaluation criteria can be used to score a single answer or to compare several answers from an AI system. When comparing, the more accurate, coherent, and complete answer is scored higher. This systematic process makes AI-generated answers factually correct, logically sound, and well-organized, thus making them credible and useful sources of information for users.

Both virtual assistants' (R and T) performance on five criteria in evaluation is plotted above. The results were divided into two parts depending on their evaluation method: Binary and Scaled evaluations. The Row-Based Chunking RAG (R) does better than the Traditional Chunking RAG (T) on the most important ones.

Figure 1 presents the binary evaluation results.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

- Avoidance of hallucination (Q1): R performed better due to improved chunking, which positively impacted the generation phase.
- Logical consistency (Q4): Both approaches maintained strong logical consistency in their responses.

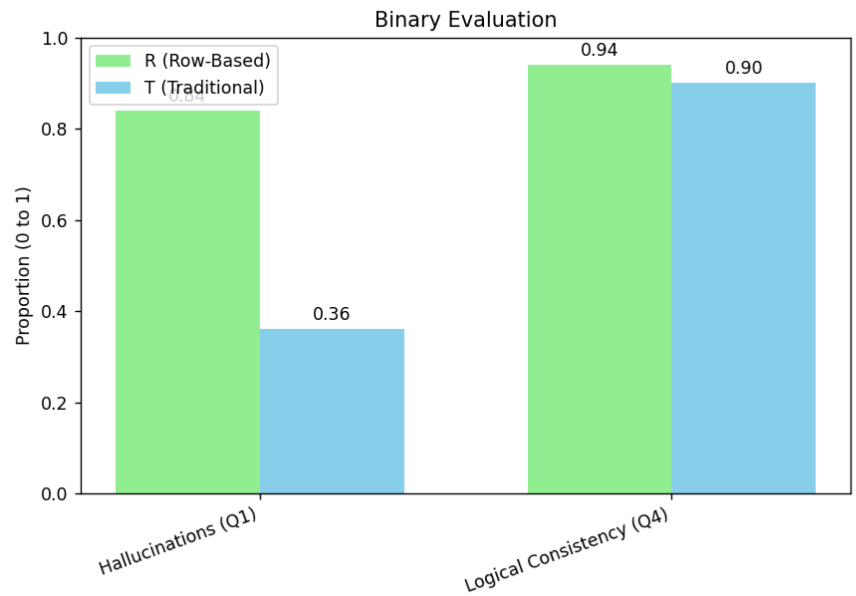

Fig. 1. Binary evaluation results for virtual assistants R and T across two criteria: hallucination avoidance (Q1) and logical consistency (Q4).

Also the Figure 2 shows the scaled evaluation results.

- Response relevance (Q2): R had a higher relevance score.
- Content size and structure (Q3): R was rated more positively than T.
- Overall evaluation (Q5): R scored higher than T, indicating better overall quality of response.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

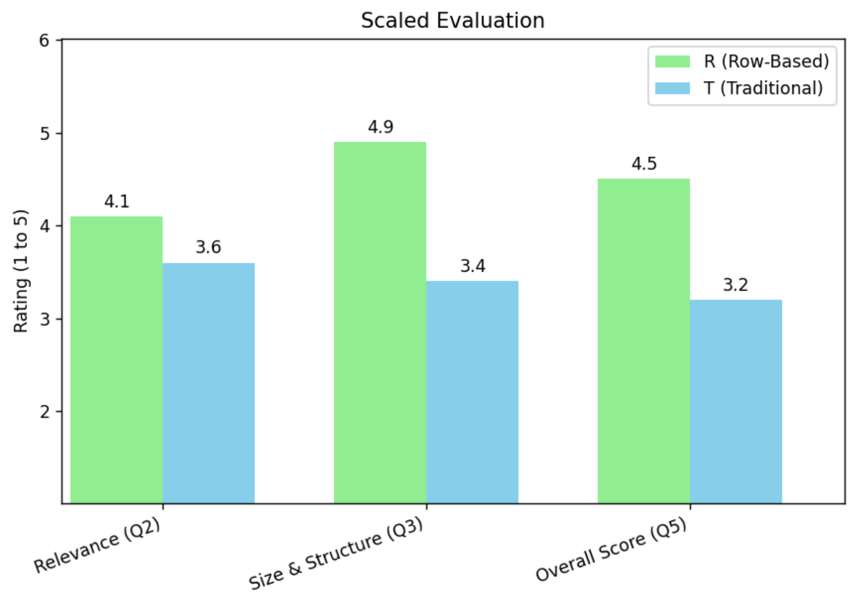

Fig. 2. Scaled evaluation results for virtual assistants R and T across three criteria: response relevance (Q2), content size and structure (Q3), and overall evaluation (Q5).

The no-response rates for each assistant are visualized in Figure 3:

- R had a 9.1% no-response rate.
- T had a 14.3% no-response rate.

This data indicates that R provided responses more frequently than T when evaluated across multiple queries.

E. Evaluation Metrics

We applied our row-based chunking method to a collection of 50 queries collected in a user study. As was mentioned before, every one of the 50 volunteers asked 3 different questions but we took 1 from each and which was tackled by two distinct RAG systems: one with traditional chunking and the other with our row-based chunking method. For each question, we retrieved top-k chunks from both the systems and manually judged their relevance to the question context.

Based on this human-judgment, Precision@k, Recall@k, and F1@k values were computed and compared with returned chunks versus information needed to answer each query. On k=10, our row-based model achieved Precision of 0.58, Recall of 0.67, and F1 score of 0.62, which was considerably higher than the baseline paragraph-based scheme (Precision@10 = 0.41, Recall@10 = 0.44, F1@10 = 0.42). The outcomes indicate that chunk alignment with semantically similar rows within structured data leads to more accurate and comprehensive retrieval, which ultimately improves answer quality in RAG systems.

F. Discussion

These results reinforce the central significance of chunking strategies to retrieval performance, elucidating why distinct methods yield varied outcomes. Standard chunking practices have a propensity to cause incoherences by disunifying contextually coherent information, whereas Row-Based Chunking RAG preserves total context units. This is specifically beneficial for structured information, such as FAQs, where coherence should be preserved so that correct retrieval can be supported.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

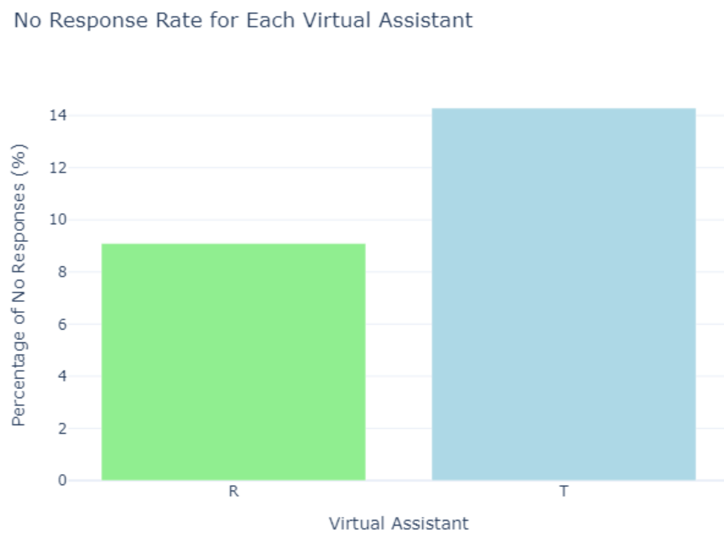

Fig. 3. No-response rates for virtual assistants R and T.

One key thing to note is that if Row-Based Chunking RAG retrieval fails, Traditional Chunking RAG will also fail. But not vice versa—Row-Based Chunking RAG can pass when Traditional Chunking RAG fails. This is due to the fact that traditional chunking techniques sometimes dismember logically related information, and retrieval models struggle more to generate effective responses from them. Such findings suggest that for structured datasets, Row-Based Chunking RAG provides a more solid and context-aware solution.

The evaluation also indicates that Row-Based Chunking RAG always produces more structured and richer answers than Traditional Chunking RAG. This is because of the following reasons:

- Preservation of context: Row-Based Chunking RAG retains full rows as single chunks, ensuring more cohesive retrieval.
- Reduced fragmentation: Traditional Chunking RAG sometimes splits related information into multiple smaller chunks, leading to a loss of coherence in responses.

The no-response rate difference reinforces these findings. While both methods fail in some cases, Row-Based Chunking RAG consistently outperforms Traditional Chunking RAG in retrieving relevant content. This suggests that inefficient chunking in Traditional Chunking RAG contributes to response failures, whereas Row-Based Chunking RAG's structured approach improves retrieval even in challenging cases.

G. Implications

The findings we obtained can be implemented in such structured data QA systems. For RAG systems, choosing an effective chunking strategy is crucial to enhance the response relevance. It can be applied in domains where structured knowledge is key—such as academic assistants, customer support bots, or legal document retrieval—Row-Based Chunking RAG could enhance accuracy and reduce hallucinations.

H. Limitations and Future Considerations

The advantages of our Row-Based Chunking RAG are most noticeable in structured datasets, even if it increases retrieval effectiveness. It's still unclear how well it performs in texts that are more narrative or unstructured.

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

Furthermore, even though Row-Based Chunking RAG performs better than Traditional Chunking RAG in our evaluation, more study is required to determine whether it can scale to bigger and more varied datasets. Hybrid techniques that dynamically modify chunking algorithms according to query context should also be investigated in future work.

In the end, our findings emphasize how crucial careful data architecture is for retrieval-based AI systems, confirming that the quality of generated responses can be greatly impacted by the way information is chunked.

V. Conclusion and Future Work

This paper investigated the impact of chunking methods on retrieval performance on an academic Q&A dataset. Our results show that Row-Based Chunking significantly improves response completeness and coherence over traditional chunking methods. Through retaining the full context units, this approach reduces inconsistency and improves retrieval accuracy, particularly for structured data such as FAQs.

In addition, the study highlights the point that traditional chunking often leads to disconnected responses due to random text splitting. In contrast, Row-Based Chunking is logically consistent, which allows for more effective retrieval of semantic information. The aspect that it possesses a lower no-response rate also bears witness to its application in structured data retrieval.

However, despite these developments, some of the limitations still remain, including the scope of our evaluation and the challenge of handling unstructured or multilingual data. Future research can explore hybrid chunking methods that adapt dynamically to different types of data and retrieval needs. The integration of user feedback and real-world testing will also help in advancing the practical applicability of structured chunking RAG systems.

Our findings enhance knowledge on chunking techniques in information searching, offering a yet more systematic and effective approach to be applied in university virtual assistants and further beyond.

VI. Acknowledgment

The author would like to express sincere gratitude Selcuk Cankurt, Kamila Orynbekova, Ardak Shalkarbay, Yernar Akhmetbek, Aiken Kazin, Inkar Shoganova, Ualikhan Sadyk and Mukhtar Amirkumar for their valuable guidance, unwavering support, and insightful feedback throughout the entire process of completing this thesis.

References

[1] Chen, J., Liu, Z., Huang, X., Wu, C., Liu, Q., Jiang, G., ... & Chen, E. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. World Wide Web, 27(4), 42. https://doi.org/10.1007/s11280-024-01276-1

[2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1–55. https://doi.org/10.1145/3703155

[3] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., ... & Li, Q. (2024, August). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6491–6501). https://doi.org/10.1145/3637528.3671470

[4] Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., ... & Yao, H. (2024). MMED-RAG: Versatile multimodal RAG system for medical vision language models. arXiv preprint arXiv:2410.13085. https://doi.org/10.48550/arXiv.2410.13085

[5] Liu, C., Hoang, L., Stolman, A., & Wu, B. (2024, July). HiTA: A RAG-Based Educational Platform that Centers Educators in the Instructional Loop. In International Conference on Artificial Intelligence in Education (pp. 405–412). Cham: Springer. https://doi.org/10.1007/978-3-031-64299-9_37

[6] Zhu, F., Li, J., Pan, L., Wang, W., Feng, F., Wang, C., ... & Chua, T. S. (2025). FinTMM-Bench: Benchmarking Temporal-Aware Multi-Modal RAG in Finance. arXiv preprint arXiv:2503.05185. https://doi.org/10.48550/arXiv.2503.05185

[7] Wang, S., Tan, J., Dou, Z., & Wen, J. R. (2024). OmniEval: An Omnidirectional and Automatic RAG Evaluation Benchmark in Financial Domain. arXiv preprint arXiv:2412.13018. https://doi.org/10.48550/arXiv.2412.13018

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.

[8] Olvera-Lobo, M. D., & Gutiérrez-Artacho, J. (2015). Question answering track evaluation in TREC, CLEF and NTCIR. In New Contributions in Information Systems and Technologies: Volume 1 (pp. 13–22). Springer. https://doi.org/10.1007/978-3-319-16486-1_2

[9] Allam, A. M. N., & Haggag, M. H. (2012). The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3).

[10] Derici, C., Aydin, Y., Yenialaca, Ç., Aydin, N. Y., Kartal, G., Özgür, A., & Güngör, T. (2018). A closed-domain question answering framework using reliable resources to assist students. Natural Language Engineering, 24(5), 725–762. https://doi.org/10.1017/S1351324918000141

[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 33, 9459–9474.

[12] LangChain. (2023).

[13] Yuan, Y., Liu, C., Yuan, J., Sun, G., Li, S., & Zhang, M. (2024). A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. arXiv preprint arXiv:2408.05141. https://doi.org/10.48550/arXiv.2408.05141

[14] Xu, Z., Liu, Z., Liu, Y., Xiong, C., Yan, Y., Wang, S., ... & Yu, G. (2024). ActiveRAG: Revealing the treasures of knowledge via active learning. arXiv preprint arXiv:2402.13547. https://doi.org/10.48550/arXiv.2402.13547

[15] Fan, Y., Yan, Q., Wang, W., Guo, J., Zhang, R., & Cheng, X. (2025). TrustRAG: An Information Assistant with Retrieval Augmented Generation. arXiv preprint arXiv:2502.13719. https://doi.org/10.48550/arXiv.2502.13719

[16] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333–389. http://dx.doi.org/10.1561/1500000019

[17] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023, October). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations.

[18] Wei, Z., Chen, W. L., & Meng, Y. (2024). InstructRAG: Instructing retrieval-augmented generation with explicit denoising. arXiv preprint arXiv:2406.13629. https://doi.org/10.48550/arXiv.2406.13629

[19] Kim, K., Kim, M., Lee, H., Park, S., Han, Y., & Jeon, B. K. (2024). THoRR: Complex Table Retrieval and Refinement for RAG. In IR-RAG 2024 Workshop Proceedings, 3784, 50–55.

[20] Abane, A., Bekri, A., & Battou, A. (2024). FastRAG: Retrieval Augmented Generation for Semi-structured Data. arXiv preprint arXiv:2411.13773. https://doi.org/10.48550/arXiv.2411.13773

Received: April 22, 2025. Reviewed: May 30, 2025. Accepted: June 3, 2025.