Information Modelling and Knowledge Bases XXXVI
Y. Kiyoki et al. (Eds.)

doi:10.3233/FAIA241572

AI Model Deficiency in Knowledge Insufficiency

Virach SORNLERTLAMVANICH[1],
Ryusei Doi[2] and Thatsanee CHAROENPORN[3]
Asia AI Institute (AAII), Faculty of Data Science, Musashino University, Japan.
ORCID ID: Virach SORNLERTLAMVANICH https://orcid.org/0000-0002-6918-8713
ORCID ID: Thatsanee CHAROENPORN https://orcid.org/0000-0002-9577-9082

Abstract. Challenged by data-driven AI limitations in reasoning and knowledge depth, this work presents a novel approach for enhanced conversational understanding. We leverage advanced text analysis to strategically extract key information from FAQs, then utilize AI-generated questions and robust semantic similarity metrics to significantly improve user query matching precision. Through the strategic integration of important sentence extraction in knowledge preparation, coupled with question generation and the application of semantic textual similarity measures, our model achieves a substantial improvement in user query matching precision. We propose a dual-system architecture—augmenting System 1 with additional knowledge akin to System 2 in human cognition. The methodology is exemplified through chatbot correction using FAQs, demonstrating the potential for human-like mind processing. Results showcase improved semantic understanding and reasoning, offering a promising path for advancing AI capabilities in conversational contexts.

Keywords. Large Language Model (LLM), Text summarization, Semantically Conscious Reasoning (SCR), Semantic Textual Similarity (STS), Sentence-BERT (SBERT)

1. Introduction

"Does a LLM really occupy any kinds of knowledge?" The question of whether large language models (LLMs) possess any forms of knowledge has become a central topic in the field of artificial intelligence. While LLMs are very fluent in tasks like text generation, translation, summarization, and question answering, the nature of their knowledge representation and reasoning capabilities remains under debate.

Several arguments suggest that LLMs do not hold true knowledge in the traditional sense. First, their training data can be vast and unstructured, encompassing factual information, fictional narratives, and even contradictory viewpoints [1]. This makes it difficult to distinguish reliable knowledge from mere statistical associations within the

---

[1] Corresponding Author: Virach Sornlertlamvanich, Asia AI Institute (AAII), Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan; E-mail: virach@musashino-u.ac.jp

[2] Ryusei Doi, Asia AI Institute (AAII), Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan; E-mail: s2022048@stu.musashino-u.ac.jp

[3] Thatsanee Charoenporn, Asia AI Institute (AAII), Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan; E-mail: thatsane@musashino-u.ac.jp

model's internal representations [2]. Additionally, LLMs often struggle with tasks requiring factual consistency or logical reasoning, indicating a reliance on statistical patterns rather than genuine understanding [3].

However, other perspectives argue that LLMs do possess a distinct form of knowledge. Proponents highlight the model's ability to learn and adapt to new information, demonstrating a level of understanding beyond simple pattern matching [4]. Additionally, LLMs can sometimes exhibit surprising reasoning abilities, such as inferring implicit relationships or drawing conclusions from multiple sources [5]. This suggests that the model may be developing internal representations that capture some aspects of real-world knowledge, even if it differs from human knowledge in its structure and accessibility.

The ongoing debate about LLM knowledge reflects the complexity of understanding intelligence in artificial systems. It is likely that LLMs possess a unique type of knowledge that is neither identical to nor entirely separate from human knowledge. Further research is needed to elucidate the nature of this knowledge and its implications for the field of artificial intelligence.

LLMs are foundational machine learning models that employ deep learning algorithms to process and comprehend natural language. These models undergo training with extensive amounts of text data to acquire knowledge of patterns and entity relationships within the language. The underpinnings of LLM behavior can be summarized as follows:

- LLMs are not developed based on the process of language understanding.
- LLMs do not reason about occurrences by connecting to any knowledge representation of either language or the real world.
- LLMs are developed through training on massive amounts of text data using an artificial neural network.
- LLMs have the capability to compute quickly enough to provide a response within an acceptable timeframe with a proper language sequence, but there is no guarantee of the accuracy of the answers.

Therefore,

- LLMs generate sentences in a sequence based on the weight-evaluated order.
- LLMs may generate sentences through few-shot, one-shot, or even zero-shot learning from the provided information [6].
- LLMs are observed to possess the intelligence to converse fluently with humans.

Consequently, the following result obtained from a straightforward fact-finding question on the ChatGPT 3.5 API was the inquiry, "Who is the founder of Musashino University?" defined as the "content" of "user".

```
[60]: completion = openai.ChatCompletion.create(
          model="gpt-3.5-turbo",
          messages=[
              {"role": "system", "content": "Answer the question as truthfully as possible, \
              and if you're unsure of the answer, say 'Sorry, I don't know."},
              {"role": "user", "content": "Who is the founder of Musashino University?"}
          ],
          temperature = 0
      )

[61]: print(completion.choices[0].message.content)

      The founder of Musashino University is Yasujiro Shimazu.
```

When seeking information of which you are unsure, verifying its accuracy becomes challenging. It is undesirable to discover later that the acquired information is incorrect.

Hallucinations in LLM can result from factors like incomplete or noisy training data. However, once factual information is supplied, the LLM can furnish accurate responses.

In the following ChatGPT API call, when factual information is supplied in the "content" parameter for the "assistant," the model provides accurate and appropriate responses. The model's response can be remedied by providing it with the accurate factual information.

```
[64]: completion = openai.ChatCompletion.create(
          model="gpt-3.5-turbo",
          messages=[
              {"role": "system", "content": "Answer the question as truthfully as possible, \
              and if you're unsure of the answer, say 'Sorry, I don't know."},
              {"role": "assistant", "content": "In 1924, Dr. Junjiro Takakusu, a Buddhist scholar, \
              founded Musashino Women's Academy within Tsukiji Honganji Temple. After that, \
              he took over Chiyoda Women's College, which was run by Chiyoda Women's College, and \
              opened Musashino Women's Junior College in 1950. In 1965, the predecessor Musashino \
              Women's University was founded. In 2003, the school name was changed to Musashino University."},
              {"role": "user", "content": "Who is the founder of Musashino University?"}
          ],
          temperature = 0
      )

[67]: print(completion.choices[0].message.content)

      The founder of Musashino University is Dr. Junjiro Takakusu.
```

Now, let's conduct a simple logical test on the same ChatGPT 3.5.

```
[69]: completion = openai.ChatCompletion.create(
          model="gpt-3.5-turbo",
          messages=[
              {"role": "system", "content": "The following is a quiz testing reasoning skills, please answer."},
              {"role": "user", "content": "Let's represent a digit of 1 with 'banana' and \
              a digit of 0 with 'orange.' If the sequence 'banana, orange, orange' is equivalent \
              to 100, then what is the number represented by 'orange, banana, banana, orange, banana'?"}
          ],
          temperature = 0
      )

[70]: print(completion.choices[0].message.content)

      The number represented by 'orange, banana, banana, orange, banana' can be determined by substituting the dig
      its with their corresponding values.

      Given that 'banana' represents 1 and 'orange' represents 0, we can rewrite the sequence as follows: 01011.

      Converting this binary number to decimal, we get 11.

      Therefore, the number represented by 'orange, banana, banana, orange, banana' is 11.
```

It is not a matter of language usage or factual information retrieval, but it is a symbolic substitution (reasoning) problem. While this error may occur by chance, the accountability of such mistakes becomes questionable, particularly in tasks that demand high reliability. Similarly, LLM may correctly predict the dates of birth and death of a celebrity but may not correctly predict the age. This discrepancy is called the compositionality gap for language models [7].

Indeed, beyond language fluency, the remedy for faults often requires the incorporation of the missing parts of knowledge. Section 2 discusses the limitation of LLM and the distilled knowledge it offers. By strategically incorporating important sentence extraction into knowledge preparation, along with the implementation of question generation and semantic textual similarity measures, our model not only achieves a significant enhancement in precision for user query matching but also provides a robust foundation for improved semantic understanding, as elaborated in Section 3. Lastly, Section 4 concludes with the expression of the total integration of LLM with knowledge engineering. This approach represents a comprehensive solution to address the limitations of LLM, resulting in a reasoning augmented model that is superior in precision of semantic understanding for user queries as discussed under the topic of augmented language model (ALM) [8].

2. Knowledge Distillation from LLMs

While research on utilizing LLMs for knowledge distillation has gained significant traction, several critical questions remain regarding the nature and reliability of the extracted knowledge. This section investigates into the complexities of LLM knowledge distillation, highlighting the need for careful validation before propagating distilled information [9].

The very notion of LLM knowledge is itself debated. While LLMs are very good at statistical pattern recognition and text generation, their grasp of factual accuracy and real-world understanding remains debatable [1,10]. This raises the question of what actual knowledge we can hope to distill from such models.

Current approaches to LLM knowledge distillation focus on extracting factual information embedded within the model's internal representations. Techniques like attention visualization and explainable AI methods offer glimpses into these representations, potentially revealing semantic relationships and factual nuggets. However, the extracted knowledge often suffers from limitations inherent to the LLM itself:

- **Data biases**: LLMs trained on vast and potentially biased internet data may contain inconsistencies, factual errors, discrimination, toxic content and misleading information. Distilling such knowledge can perpetuate these biases, leading to unreliable and potentially harmful results [1,10].
- **Statistical associations**: LLMs often rely on statistical associations identified within their training data, not necessarily representing true understanding. They are generally trained to perform statistical language modeling given a single parametric model, and a limited context, typically the n previous or surrounding tokens. Distilling these associations as factual knowledge can lead to spurious correlations and unreliable inferences [8].
- **Limited reasoning**: LLMs often struggle with tasks requiring logical reasoning or commonsense knowledge [4]. Distilling knowledge from such models might lack the necessary depth and context to be truly informative or reliable. Strategically prompting in LLM is used to enhance its reasoning ability. It typically takes one of the two forms: zero-shot, where the model is directly prompted with a test example's input; and few-shot, where few examples of a task are prepended along with a test example's input. This few-shot prompting is also known as in-context learning or few-shot learning [11].

Given these limitations, validating the reliability of distilled knowledge becomes a big concern before its propagation or application. This validation process should encompass several key aspects:

- **Fact-checking**: Extracted factual claims should be rigorously cross-referenced with trusted sources and expert knowledge to ensure accuracy and prevent the spread of misinformation.
- **Bias detection and mitigation**: Techniques to identify and mitigate data biases within the LLM and the distilled knowledge must be employed to avoid perpetuating harmful stereotypes or discriminatory tendencies.

- **Logical consistency and plausibility**: Distilled knowledge should be evaluated for logical consistency and real-world plausibility. Techniques like commonsense reasoning evaluation and domain-specific knowledge verification can help identify inconsistencies and potential errors.

The desire to leverage LLMs for business applications, such as chatbots capable of conveying facts relevant to user queries, further underscores the need for caution. In such scenarios, factual inaccuracy or misleading information can have significant consequences, impacting brand reputation, customer trust, and even financial losses. Therefore, deploying LLMs in business settings necessitates:

- **Domain-specific training**: LLMs should be fine-tuned on datasets specific to the business domain, ensuring the distilled knowledge aligns with industry standards and best practices.
- **Human oversight and control**: Ultimately, human oversight and control mechanisms are crucial to ensure responsible use of LLMs in business applications. This includes establishing clear guidelines for content generation, implementing robust fact-checking procedures, and providing avenues for user feedback and error correction.

3. Fulfillment of AI model efficiency by incorporating knowledge about truths and efficient algorithm for semantically locating capability

The LLM possesses the significant potential to replace human call centers due to its conversational fluency in multiple languages, stemming from extensive training on a vast array of texts across various domains of knowledge. Nonetheless, while it demonstrates high proficiency in languages, its responses may questionable, especially when addressing factual queries. The details of its responses are generated from the trained dataset using statistical associations, casting uncertainty on the reliability of the information provided. Consequently, employing it for fact-finding tasks is not deemed plausible.

In Daniel Kahneman's framework of mental processing [12], System 1 is an automatic system shaped by past experiences. While it responds swiftly, it can introduce errors due to its less conscious operation. The LLM demonstrates its System 1-like capability in human mind processing. The question arises: Can the model be trained effectively with necessary and sufficient data one day? Regardless of the answer (whether "yes" or "no"), the system will encounter efficiency challenges in knowledge finding process.

In this research, we propose a model that harnesses LLM's language capabilities as the equivalent of System 1. This model is then enhanced with factual knowledge and an efficient algorithm for semantically locating capability. The final component serves as a System 2 counterpart in human mental processing, equipped with knowledge about the truths and semantically conscious reasoning for the obtained results. Semantically Conscious Reasoning (SCR) refers to the cognitive process of interpreting and understanding information using semantic context while being influenced by conscious awareness. It involves the thoughtful consideration of meaning and context in decision-making and problem-solving. Figure 1 illustrates the components and the interrelation between System 1 and System 2.

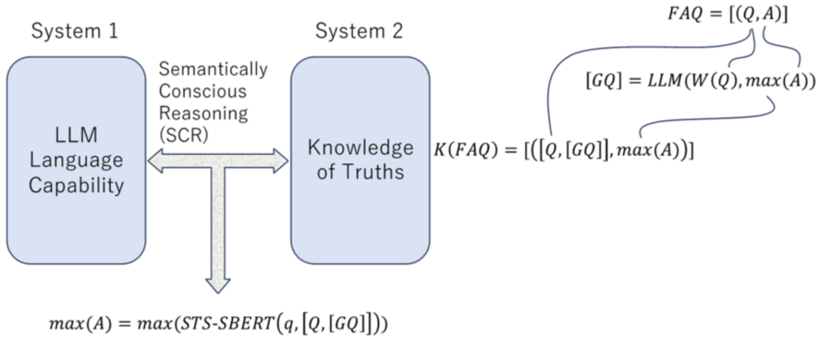

Figure 1. Semantically Conscious Reasoning (SCR) between LLM as a System 1 and knowledge encapsulated System 2

In this study, we work on Amagasaki FAQ (FAQ) which is a collection of pairs of a frequently asked question (Q) and its corresponding answer (A). Let's define it in a form of following notation,

$$FAQ = [(Q, A)] \tag{1}$$

It is the knowledge of the truths that Amagasaki city provides in response to queries about the city and its administrative services. The FAQ is manually generated, and most answers are derived from the city guidebook and service manual. During the knowledge preparation process, we refine the answers using the text summarization technique ($f_{summarization}$) to extract the sentence (max(A)) containing the essential information [13].

$$max(A) = f_{summarization}(A) \tag{2}$$

As a result, a set of proper knowledge or FAQ with max(A) is well-prepared. However, since the question (Q) in the FAQ is limited to a representation of the frequently asked question, it is not plausible to restrict users to asking only from a prepared list of questions. This approach is commonly implemented in some systems, allowing users to select questions from a list to receive answers. Nevertheless, this method may make users feel uncomfortable. Therefore, the language capability of LLMs can be expected to expand the original question by incorporating information from max(A).

In the experiment, we utilize the Text-to-Text Transfer Transformer (T5) to generate a question (GQ) by providing the list of keywords (W(Q)) extracted from the question and max(A) as the context. T5 is a transformer-based neural network architecture developed by Google Research [14, 15].

$$[GQ] = LLM(W(Q), \max(A)) \tag{3}$$

The knowledge from the FAQ (K(FAQ)) can be represented as a list of pairs, including the original FAQ question (Q), the list of generated questions (GQ), and the important sentences in the original FAQ answer (max(A)).

$$K(FAQ) = [([Q, [GQ]], \max(A))] \tag{4}$$

To obtain the correct answer (max(A)) from the FAQ, the user query (q) will be verified against the list of questions from the original question (Q) and the list of generated questions (GQ). The Textual Similarity (STS) Sentence-BERT (SBERT) model (STS-SBERT) is generated by fine-tuning the BERT base model and implemented with sentence embeddings using Siamese BERT-networks [16].

$$\max(A) = \max\left(STS - SBERT\big(q, [Q, [GQ]]\big)\right) \tag{5}$$

3.1. LLM Knowledge Distillation for Domain-Specific Chatbot Generation

Preparing a question list for the intents in creating a chatbot is a labor-intensive task. It involves the challenge of looking up sentences in the FAQ database, and it is not trivial to assume a set of variations of questions that can be properly matched with user queries.

Word expansion is a common approach used to broaden the matching coverage between user queries and questions in the FAQ database. This method aims to address the problem of mismatching due to word variation in expressions or synonyms. For instance, a query like "What is the price of ...?" might be expressed as "How much is ...?" or "What does it cost ...?". In our preliminary experiment, we employed the synset of WordNet [17] to expand word forms, disregarding the multiple word sense problem by including all possible words found in the synsets. However, the results did not show a significant improvement in question matching rate, and it consumed considerable time and memory to include all combinations of words from the synsets. This method is integrated into the system architecture of FAQ database retrieval [18], employing query-question similarity measures in TSUBAKI [19], where synonyms and sentence dependency structures are considered.

Rather than expanding the word by its synonyms, we generate other related questions from the question and answer in the FAQ database. Text-To-Text Transfer Transformer (T5), as demonstrated by Raffel et al. (2020) [15], facilitates the generation of queries from extracted important sentences. In a scenario involving legal FAQs, T5 converted key legal principles within answers into informative queries, improving the model's ability to provide legally sound responses. It is expected that based on the large scale pre-trained model, the questions in other variation of expressions can be generated.

Moreover, we discovered that the simple cosine similarity measure between sentences is ineffective in identifying appropriate questions. This is attributed to differences in expression and word form used in the sentences being compared. The cosine similarity method computes the similarity of the sum of word vectors present in the sentences, lacking consideration for word context, which is crucial for word sense disambiguation. This limitation is particularly evident in the case of user free input queries, where sentences can vary significantly in expressing a specific question.

To improve the matching rate between the user query and questions in FAQ, we utilize Semantic Textual Similarity (STS) Sentence-BERT (SBERT) model [16] to measure the semantic similarity between the user query and question. In our experiment, we fine-tune the Japanese Sentence-BERT model[4] which is generated from the base model by Tohoku University NLP Lab[5]. The knowledge of FAQ can be expressed as a list of pairs of original FAQ question (Q), list of generated question (GQ), and the important sentences in original FAQ answer (max(A)).

---

[4] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

[5] https://github.com/cl-tohoku/bert-japanese

3.2. FAQ database (Amagasaki FAQ) and User Generated Query Test Set

Amagasaki FAQ is the Japanese administrative municipality domain FAQ database which is prepared by the Amagasaki city local government. It is an FAQ database containing a set of 1,786 questions and the corresponding answers in FAQ page of Amagasaki city. The FAQ dataset is quite large and manually prepared to give the responsive answer about the city.

Table 1. An example of a pair of question (Q) and answer (A) in Amagasaki FAQ.

No.	Question (Q)	Answer (A)
1	How do I get to the Imakita Regional General Center?	Imakita Regional General Center does not have enough parking lots, so please use the city bus. Please come to “Tachibana Station” by the JR line, “Tsukaguchi Station” and “Mukonoso Station” by the Hankyu Line, and “Amasaki Station”, “Mukogawa Station” and “Deyashiki Station” by the Hanshin Line, and then use the city bus. Which station are you from? 1. From JR Tachibana Station (location is about a 10-minute walk to the southwest). 2. From Hankyu Tsukaguchi Station (south). 3. From Hankyu Mukonoso Station (south). 4. From Hanshin Amagasaki Station (north). 5. From Hanshin Mukogawa Station. 6. From Hanshin Deyashiki Station (north). <Revised> [Related FAQ] I want to know about the Regional General Center. <Revised> [Inquiry] Imakita Regional General Center 3-14-1 Nishitachibanacho, Amagasaki City. Phone 06-6416-5729.

Table 1 shows an example of a pair of question and answer. Though there is no detail of how the FAQ is prepared, it can be observed that the questions are manually prepared based on the given answers of the city related information. Almost all the questions are to ask about a part of the information in the given answers.

To test our proposed method in preparing questions for intent development for a chatbot, we apply our approach to evaluate the accuracy of similarity measure against the test set of 784 user generated queries as shown in Table 2. The test set is prepared by Kyoto University from crowdsourcing according to the FAQ explanatory answers [18].

Table 2. An example of a pair of user query (q), the matched question (Q) and answer (A) in Amagasaki FAQ.

No.	Query (q)	No.	Question (Q)	Answer (A)
86	Can you mail me a copy of my resident card?	62	Can I have a copy of my resident card mailed to me?	A copy of the residence certificate can be requested by mail from the person or a person in the same household. In the case of a request from a third party (other than the person or a person in the same household as the person), a power of attorney from the person is required. If you have not been delegated by the person, or if you are requesting mail from a corporation, public service, lawyer, etc., please contact the Citizens Division. However, the resident’s card with my number can only be obtained by the person or a member of the same household. Please see the following link for details. [URL]. <Revised> [Related FAQ] What kind of content is included in the copy of the resident’s card, and how much is the fee? Can an agent obtain a resident card with my number? <Revised> [inquiry] Citizen Service Department,

				Citizen Collaboration Bureau. Citizens Division. Phone 06-6489-6408. Inquiry time. From 8:45 am to 5:30 pm. However, the counter handling hours are from 9:00 am to 5:30 pm. holiday. Saturdays, Sundays, national holidays, year-end and New Year holidays (December 29-January 3).

The expression of query is different from the question in FAQ but they convey exactly the same meaning. However, the answer shows much more information about the detail condition in mailing the resident card.

The test set gives more candidate of answers in three groups of relation, that is relevance (correct information), relate (relevant information), and same group (same group of query but answer contains irrelevant information). For simplicity, we group all the related answers into a list of relevant answers to measure the similarity in the evaluation process.

3.3. LLM Knowledge Distillation Activated by Question-Answer Pairs from FAQ

FAQ database normally contains a large number of pairs of a question and an answer. We search the FAQ database by finding the best matching of the user query and the question in the database. Then the answer of the matched question is returned to the user. The problem is to how to extreme the finding of the matched question which is only one representative question of the common questions for an answer. Actually, the matching by their meaning is preferred. To do so, we have to prepare a set of sentence variation or an algorithm that can cover the intentional meaning.

Generative model in LLMs has potential of generating relevance sentences of given keywords and context. In the question generation process, we extract content words such as noun, verb, adverb, and adjective from the FAQ question sentence to create a list of keywords, and use the corresponding FAQ answer sentences as the context for T5 to generate a new corresponding question. We use the default hyperparameter to generate only one output to reduce the complexity in evaluation. In the practical use case in chatbot, some numbers of questions are needed to extend the possibility to match with other information in the answer. However, list of keywords in concern is needed to prepare corresponding the part of information provided in the answers.

From Table 1, the list of content words, ("How", "get to", "Imakita" "Regional", "General", "Center"), is extracted from the question (Q) to use as the keyword list, and the answer (A) is used as the context for T5 to generate a question which is shown in the generated question (GQ) in Table 3.

Table 3. An example of generated question (GQ) according to the question (Q) and answer (A) in Amagasaki FAQ.

No.	Generated Question (GQ)	Question (Q)	Answer (A)
1	What bus stops are there for the Imakita Regional General Center?	How do I get to the Imakita Regional General Center?	Imakita Regional General Center does not have enough parking lots, so please use the city bus. Please come to "Tachibana Station" by the JR line, ….

Table 3 shows the generated question according to the question and answer in the FAQ. The generated question still requests for the same information as in the question but has the different expression. This is because the keywords from the question are provided in the generation process. The result of question generation can be used to serve the variants of question in the intent of the chatbot.

The relationship between the question (Q), generated question (GQ), and query (q) are investigated by their similarity measure on the user generated query test set. The experiments have been conducted on both SBERT and fine-tuned SBERT models applied on Semantic Textual Similarity (STS). Table 4 shows how close the meaning of GQ is to the original Q. Tables 5 and 6 show how close the meanings of Q and GQ are to the q, which means how good the proper answer from the FAQ can be retrieved.

The value of *mAP* (mean Average Precision) is the mean of the average precision scores for each query, Top1 is the accuracy measured on the correct answer found in the top position and Top5 is the one found within the top five answers.

Table 4. Accuracy in similarity measure between question (Q) and generated question (GQ).

sim(Q, GQ)	*mAP*	Top1	Top5
STS-SBERT	0.4594	0.3628	0.5671
Fine-tuned STS-SBERT	**0.4852**	0.4072	0.5963

Table 5. Accuracy in similarity measure between query (q) and question (Q).

sim(q, Q)	*mAP*	Top1	Top5
STS-SBERT	0.5202	0.4018	0.6505
Fine-tuned STS-SBERT	**0.6144**	0.5064	0.7398

Table 6. Accuracy in similarity measure between query (q) and generated question (GQ).

sim(q, GQ)	*mAP*	Top1	Top5
STS-SBERT	0.3922	0.2832	0.5217
Fine-tuned STS-SBERT	**0.4842**	0.3661	0.6250

Certainly, the improvement of STS-SBERT model after fine-tuning can be confirmed in all cases. Furthermore, we found that based on the similarity measure with q, the contribution of GQ (Table 6) is quite low comparing to the original Q (Table 5). However, the GQ can somehow play an important role to cover the unseen query expressions from other users which cannot be matched well to the Q as reported in [20].

3.4. Information Purification for Better Knowledge Distillation

After conducting an in-depth analysis of the errors, we found that the FAQ answer we use as a context for T5 to generate the new corresponding question are not consistently assigned. The answer contains multiple sentences together with some remarks. Moreover, the most impactful consequences are the unrelated information texts which may come from original source of city guidebook and service manual.

To clean up the FAQ answer, we apply a text summarization technique to create a model for identifying the important sentences based on the feature-based for important sentence extraction in text summarization [13]. One hundred out of 1,786 answers are randomly selected from the Amagasaki FAQ database for annotating important sentences. Nine features of each important sentence, as defined in Table 7, are generated to form a 124-dimensional sentence embedding vector as illustrated in Figure 2.

Table 7. Nine features of the important sentence.

No.	Feature	Dimension	Description
1	Sentence relation position	10	MinMax normalized value in the range of [0,1]
2	Sentence length	10	MinMax normalized value in the range of [0,1]
3	TF-IDF	10	A measure of importance of a word to a document in a collection
4	Dependency-structure based TF-IDF 1	10	Dependency structure based TF-IDF, taking the longest dependency path
5	Dependency-structure based TF-IDF 2	10	Dependency structure based TF-IDF, taking the predicate path
6	Okapi-BM25	10	A type of TF-IDF, taking document length into account, shorter document gets higher value
7	Named Entity	8	Named entity type i.e. person, location, organization, artifact, date, money, percent, time
8	Conjunction word	45	A set of 45 conjunction words (Japanese)
9	Auxiliary word	11	A set of 11 auxiliary words (Japanese)

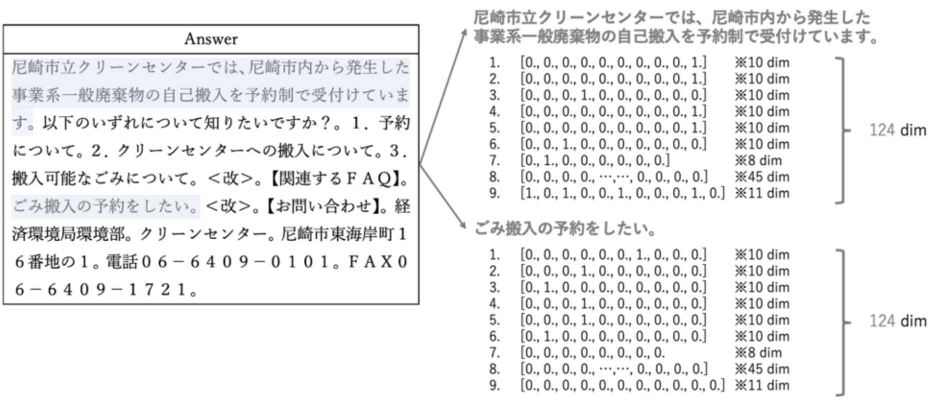

Figure 2. 124-dimensional sentence embedding vector of important sentences, an example from actual implementation on Japanese text.

As a result, the important sentences are successfully extracted with an accuracy of 90.28 per cents in case of using Light Gradient Boosting Machine (lightgbm), a decision tree based classifier. The experiment is conducted using a sample of 100 answers, comprising 1,009 sentences for training and 432 sentences for testing.

After applying important sentence extraction to the original FAQ answer, the effect of the unrelated sentences can be mitigated. The result of generated question (GQ) can also be improved. Table 8 displays the new answer with the selected important sentences (max(A)) and the improved GQ, specifically a question about "the types of waste" rather than the question about "the language use." This aligns well with the answer (max(A)).

The bold text in the Original Answer (A) column represents the important sentences, and the underscored text in the last two columns indicates the difference between the two types of GQ.

Table 8. Result after applying important sentence extraction.

No.	Question (Q)	Original Answer (A)	Answer with only Important Sentences (max(A))	Generated Question (GQ)	Generated Question (GQ) with max(A)
3	I want to know about the direct delivery of business waste within Amagasaki City.	**At the Amagasaki Municipal Clean Center, we accept reservations for the self-delivery of business general waste generated within Amagasaki City**. What would you like to know about? 1. Regarding reservations. 2. About the delivery to the Clean Center. 3. Information on waste that can be delivered. <Change> [Related FAQ] **I want to make a reservation for waste delivery.** <Change> [Contact] Economic Environment Bureau, Environmental Department, Clean Center, 16-1 Higashi Kaigancho, Amagasaki City. Phone: 06-6409-0101. FAX: 06-6409-1721."	At the Amagasaki Municipal Clean Center, we accept reservations for the self-delivery of business general waste generated within Amagasaki City. I want to make a reservation for waste delivery.	At the Amagasaki Municipal Clean Center, what language should be used to make a reservation for self-delivery of general waste?	At the Amagasaki Municipal Clean Center, what types of waste can be self-delivered?

With the generative model of LLM, it becomes possible to generate relevant questions. However, these generated questions are shaped by the model's statistical associations, potentially introducing biases derived from the provided information. The distilled knowledge from LLM, stemming from both erroneous and biased data, requires careful evaluation to ensure its functionality. The results presented in Table 8 demonstrate that appropriate questions can be generated when provided with suitable information.

As anticipated, the results of user query matching have been improved in all cases when applied to the answers with the selected important sentences (max(A)). Table 9 demonstrates the improved quality of the generated questions (GQ) in representing Q, achieving a *mAP* of 0.5635 compared to 0.4852. Table 10 illustrates how the answers with only important sentences (max(A)) enhance the quality of GQ in terms of matching with user queries, yielding a *mAP* of 0.5038 compared to 0.4842.

Table 9. Accuracy in similarity measure between question (Q) and generated question (GQ).

	Answer with only Important Sentences (max(A))			Original Answer (A)		
sim(Q, GQ)	*mAP*	Top1	Top5	*mAP*	Top1	Top5
STS-SBERT	0.5368	0.4423	0.6484	0.4594	0.3628	0.5671
Fine-tuned STS-SBERT	**0.5635**	0.4642	0.6792	**0.4852**	0.4072	0.5963

Table 10. Accuracy in similarity measure between query (q) and generated question (GQ).

	Answer with only Important Sentences (max(A))			Original Answer (A)		
sim(q, GQ)	*mAP*	Top1	Top5	*mAP*	Top1	Top5
STS-BERT	0.4220	0.3087	0.5485	0.3922	0.2832	0.5217
Fine-tuned STS-SBERT	**0.5038**	0.3801	0.6594	**0.4842**	0.3661	0.6250

As noted in Subsection 3.3, GQ plays an important role in increasing possibility to match with the unseen queries. With the higher quality of GQ, the accuracy of matching with q is improved.

In the actual chatbot implementation, the user query q will be matched over the list of questions in a collection of intents. GQ which is the result of question expansion is also included in the FAQ search space. Table 11 shows how well the GQ can complement Q in finding the best match among Q and GQ as formalized in Equation (5).

Table 11. Accuracy in similarity measure between query (q) and the best match of question (Q) and generated question (GQ).

	Answer with only Important Sentences (max(A))		
max(sim(q, Q), sim(q, GQ))	*mAP*	Top1	Top5
STS-SBERT	0.5275	0.4171	0.6722
Fine-tuned STS-SBERT	**0.6165**	0.5113	0.7423

The *mAP* for the best match of q and GQ has shown improvement to 0.6165, compared to 0.6144 or 0.5038 in the case of individual matching. Although the enhancement introduced by GQ is not substantial, the addition of GQ helps maintain the quality of query matching. The proposed approach, utilizing LLM for query expansion, demonstrates a promising outcome in improving matching for unseen queries.

4. Conclusion

As data-driven AI models, exemplified by the proficient LLM in human conversation, have become increasingly successful, their limitations in reasoning function and comprehensive knowledge have become apparent. Despite their expertise in language generation, these models often struggling with deficiencies in rational decision-making and accessing nuanced knowledge. In response to this gap, our designed concept aims to address the data insufficiency in training models and solving problems. By introducing an additional layer of knowledge akin to System 2 in human cognition, our approach seeks to fulfill AI models with a more holistic and reasoning-based understanding. This paper explores the integration of supplementary knowledge to bridge the gap between the data-driven capabilities of System 1 and the reasoning functions of System 2. To demonstrate the practical feasibility of our approach, we employ a chatbot correction methodology using FAQs, showcasing the potential to emulate human-like mind processing in both System 1 and System 2.

In the knowledge preparation to improve LLM's knowledge distillation, this study has shown the effectiveness of leveraging important sentences, semantic textual similarity measures, and the generation of questions to enhance the quality of user query matching. The application of important sentence extraction to FAQ answers has demonstrated a notable reduction in the impact of irrelevant information. Moreover, the results showcase improvements in the semantic representation of questions, leading to more accurate matching with user queries.

The utilization of semantic consciousness in reasoning, particularly through the incorporation of important sentences, has proven to be a valuable strategy. The enhanced accuracy in matching user queries with relevant answers signifies the potential of this approach in refining information retrieval systems.

As we move forward, the findings of this research suggest promising issues for further exploration. Future work could work deeper into refining semantic reasoning models, exploring additional features, and addressing potential challenges in diverse datasets. Ultimately, this study contributes valuable insights into optimizing user interactions with information retrieval systems through the integration of semantic consciousness in reasoning.

References

[1] Brundage M, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228 (2018).

[2] Marcus G. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00639 (2018).

[3] Mitchell M, Sandra W. Machine learning and human explanation. In: Proceedings of the conference on fairness, accountability, and transparency; 2019; pp. 387-396. ACM.

[4] Lake BM, et al. Human-inspired machine learning. arXiv preprint arXiv:1909.09670 (2019).

[5] Tenney I, et al. Joint understanding of multiple documents by machines and humans. arXiv preprint arXiv:1909.07705 (2019).

[6] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned Language Models Are Zero-Shot Learners. In: Proceedings of the International Conference on Learning Representations (ICLR); 2022.

[7] Press O, Zhang M, Min S, Schmidt L, Smith NA, Lewis M. Measuring and Narrowing the Compositionality Gap in Language Models. 2022.

[8] Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, Rozière B, Schick T, Dwivedi-Yu J, Celikyilmaz A, Grave E, LeCun Y, Scialom T. Augmented Language Models: A Survey. arXiv preprint. 2023. ArXiv:2302.07842.

[9] Hsu KJ, et al. Large language model knowledge distillation: promises and perils. arXiv preprint arXiv:2306.07053 (2023).

[10] Welleck S, Kulikov I, Roller S, Dinan E, Cho K, Weston J. Neural Text Generation with Unlikelihood Training. In: Proceedings of the International Conference on Learning Representations (ICLR); 2020.

[11] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, et al. Emergent Abilities of Large Language Models. In: Transactions on Machine Learning Research (TMLR); 2022.

[12] Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus, and Giroux; 2011.

[13] 平尾 努, 磯崎 秀樹, 前田 英作, 松本 祐治. Support Vector Machine を用いた重要文 抽出法. 情報処理学会論文誌. 2003;44(8):2230-2243. Japanese.

[14] Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL) Conference; 2021; pp. 483-498.

[15] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683 (2020). pp. 1-67.

[16] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP/IJCNLP); 2019; pp. 3980-3990.

[17] Fellbaum C, editor. WordNet: An Electronic Lexical Database. MIT Press; 1998.

[18] Sakata W, Shibata T, Tanaka R, Kurohashi S. FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance. In: Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR); July 2019; pp. 1113-1116.
[19] Shinzato K, Shibata T, Kawahara D, Hashimoto C, Kurohashi S. TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP); 2008; pp. 189-196.
[20] Doi R, Charoenporn T, Sornlertlamvanich V. Automatic Question Generation for Chatbot Development. In: Proceedings of the 7th International Conference on Business and Industrial Research (ICBIR2022); 2022 May 19-20; Bangkok, Thailand. pp. 301-305.