Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02485-4

RESEARCH

Enhancing large language models: alleviating knowledge deficiency with external knowledge and semantically aware reasoning (SAR)

Virach Sornlertlamvanich[1,2]

Received: 19 August 2024 / Revised: 15 March 2025 / Accepted: 13 May 2025

Abstract

Inspired by the study of the human thought process, which is categorized into two systems reflecting the brain's balancing act between speed and cognition, we propose a dual-process architecture that augments System 1 with additional knowledge akin to System 2 in human cognition. The methodology is demonstrated through FAQ retrieval task, showcasing the potential for human-like cognitive processing. Challenged by current data-driven large language models (LLMs) in reasoning and knowledge depth, this work presents a novel approach to improving conversational understanding. We leverage advanced text analysis to strategically extract key information from FAQs and utilize LLM-generated questions combined with robust semantic similarity metrics to significantly improve the precision of user query matching. The results indicate better semantic understanding and reasoning, offering a promising pathway to advancing LLM capabilities in conversational contexts. The base LLM (SBERT) enhanced with semantic textual similarity using Sentence-BERT (STS-SBERT) achieves a mean Average Precision (mAP) of 0.6165, compared to 0.3600 for SBERT alone. By strategically integrating key sentence extraction during knowledge preparation, generating questions, and applying semantic textual similarity measures, our model achieves a substantial improvement in user query matching precision. However, the activation of semantically aware reasoning (SAR) remains an issue for future research.

Keywords Large language model (LLM) · Dual-process architecture · Important sentence extraction · Semantically aware reasoning (SAR) · Semantic textual similarity sentence-BERT (STS-SBERT)

1 Introduction

"Does an LLM really possess any kind of knowledge?" The question of whether large language models (LLMs) possess any form of knowledge has become a central topic in the

✉ Virach Sornlertlamvanich
virach@musashino-u.ac.jp

1 AAII, Faculty of Data Science, Musashino University, 3-3-3 Ariake, Koto-ku 135-8181, Tokyo, Japan

2 Faculty of Informatics, Burapha University, 169 Long Had Bangsaen Road, Saensuk, Muang 20131, Chonburi, Thailand

Published online: 11 June 2025

```
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
)
```

```
print(completion.choices[0].message.content)
```

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

Fig. 1 ChatGPT's answer to a fact-finding question without context provided

field of artificial intelligence. Although LLMs are highly proficient in tasks such as text generation, translation, summarization, and question answering, the nature of their knowledge representation and reasoning capabilities remains a subject of ongoing debate.

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

On the other hand, some perspectives argue that LLMs possess a distinct form of knowledge. Proponents emphasize the model's ability to learn and adapt to new information, demonstrating a level of understanding beyond simple pattern matching [4]. Moreover, LLMs can occasionally exhibit surprising reasoning abilities, such as inferring implicit relationships or synthesizing conclusions from multiple sources [5]. This suggests that the models may develop internal representations that capture certain aspects of real-world knowledge, albeit in a form that differs from human knowledge in structure and accessibility.

The ongoing debate on 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. LLMs do not reason about occurrences by connecting to any knowledge representation of either language or the real world. It may generate sentences through few-shot, one-shot, or even zero-shot learning from the provided information [6]. It has the capability to compute quickly enough to provide a response within an acceptable time frame with a proper language sequence, but there is no guarantee of the accuracy of the answers. The responses from FAQ retrieval applications are impressive in terms of language capability. The reliability of both the correctness and consistency of the answers is highly questionable.

Consequently, Figure 1 shows the 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".

When searching for uncertain information, verifying its accuracy becomes a challenge. It is undesirable to discover later that the acquired information is incorrect. Hallucinations in

```
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
)
```

```
print(completion.choices[0].message.content)
```

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

Fig. 2 ChatGPT's answer to a fact-finding question with context provided

```
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
)
```

```
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.
```

Fig. 3 ChatGPT's answer to a simple logical test

LLM can result from factors such as incomplete or noisy training data. However, once the factual information is supplied, the LLM can provide accurate responses.

Next, perform the same inquiry as above, but include related factual information. Figure 2 illustrates the result from ChatGPT API call, when accurate factual information is supplied in the "content" parameter for the "assistant," the model generates accurate and appropriate responses. Providing accurate facts effectively enhances the model's response quality.

In addition, a straightforward logical test is now performed on ChatGPT 3.5 to evaluate its reasoning capabilities.

It is not only a matter of language usage or factual information retrieval, but we can also observe the weird response in a symbolic substitution (reasoning) problem as shown in Figure 3. Although this error may occur by chance, the accountability of such mistakes becomes questionable, particularly in tasks that require 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].

Several approaches aim to mitigate LLM hallucinations originating from factual knowledge insufficiency. Retrieval-augmented generation (RAG) methods [8] enhance LLMs by retrieving relevant information from external knowledge bases during inference. This allows the model to access and incorporate up-to-date and specific facts, reducing reliance on its internal knowledge, which may be outdated or incomplete. However, RAG can be computationally expensive and may not always retrieve the most relevant information. Knowledge distillation techniques aim to transfer knowledge from larger, more complex models (teachers) to smaller, more efficient models (students) [9]. This can improve the student model's factual accuracy and reasoning abilities. However, effective knowledge distillation requires careful selection of teachers and optimization of the distillation process. While RAG and knowledge distillation offer valuable approaches to improving LLM factual accuracy, they may have limitations in capturing the nuances of human reasoning and understanding.

Daniel Kahneman [10] introduces a dual-process framework for understanding human cognition, distinguishing between System 1, which operates intuitively and effortlessly, and System 2, which is responsible for deliberate, analytical reasoning. This model underscores the limitations of intuitive judgments, highlighting how biases and heuristics can result in systematic errors. Similarly, current LLMs predominantly function as System 1, excelling in generating rapid, intuitive responses derived from extensive training datasets but lacking the deliberate reasoning and critical evaluation associated with System 2.

To bridge this gap, our approach aligns with Daniel Kahneman's dual-process framework, distinguishing between intuitive, fast-thinking System 1 and deliberate, analytical System 2. The proposed "System 2 mechanism" with semantically aware reasoning (SAR) explicitly embodies System 2 by incorporating semantic awareness into the reasoning process. This structured, context-sensitive evaluation enables LLMs to move beyond the instinctive, pattern-based responses characteristic of System 1, ensuring that answers in FAQ tasks are more factually accurate and resistant to hallucinations.

In the FAQ implementation, the system leverages the LLM, treated as the language engine of fast-thinking System 1, to facilitate natural language communication with users, ensuring fluent and intuitive interactions. Meanwhile, SAR operates as the analytical System 2, conducting semantic-based recognition to evaluate user queries against predefined FAQ knowledge. As a result, the FAQ system not only delivers responses that are naturally articulated but also ensures higher accuracy in answering user inquiries.

The main contributions of the paper are as follows:

- A dual-process architecture is proposed to alleviate the deficiency of LLMs by incorporating external knowledge about truths and an efficient algorithm for semantic localization capabilities.
- Leveraging knowledge distillation from LLMs, based on the essentialization of a knowledge source (e.g., an FAQ dataset), is proposed as a method to create core knowledge grounded in truth.
- SAR is realized by the semantic textual similarity sentence-BERT (STS-SBERT) model.
- A comprehensive evaluation is conducted through the actual FAQ dataset.

In fact, beyond language fluency, remedying faults often requires the incorporation of missing parts of knowledge. The structure of the paper is as follows: Section 2 reviews related works focusing on the limitations of LLMs and the distilled knowledge they provide. Section 3 presents the proposed dual-process architecture, which integrates knowledge extraction and preparation processes into LLMs to enable semantically aware FAQ knowledge retrieval based on user queries. Our proposed dual-process architecture is thoroughly evaluated using an FAQ dataset, as described in Section 4. 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 5. Lastly, Section 6 concludes with the expression of the total integration of LLMs with knowledge engineering. This approach represents a comprehensive solution to alleviate the limitations of LLMs, resulting in an augmented reasoning model that is superior in semantic understanding precision for user queries, as discussed under the topic of augmented language model (ALM) [11].

2 Related works

Although 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 the complexities of distilling LLM knowledge, highlighting the need for careful validation before spreading distilled information [12].

The notion of LLM knowledge itself is debated. Although LLMs are very good at statistical pattern recognition and text generation, their grasp of factual accuracy and real-world understanding remains debatable [1, 13]. 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, 13].
- **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 [11].
- **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 [14].

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 growing interest in leveraging LLMs for widely applicable use cases, such as ChatBots designed to deliver accurate information in response to user queries, highlights the importance of exercising caution. In these contexts, inaccuracies or misleading responses can lead to significant consequences, including public misinformation, erosion of trust, and broader social impacts. As a result, deploying LLMs in such scenarios necessitates careful consideration of the following factors:

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

RAG is a hybrid approach that combines the strengths of retrieval-based models and generative models [8]. This approach addresses some of the limitations inherent in purely generative models, particularly in tasks that require up-to-date, accurate, and contextually relevant information. This approach has been shown to improve the factual accuracy and relevance of generated text, making it particularly effective for tasks such as open-domain question answering and knowledge-intensive tasks. The effectiveness of RAG greatly depends on the quality of the retrieval step. If the retrieved documents are not highly relevant or accurate, the generated output may be misleading or incorrect.

Knowledge distillation approaches, which involve retraining the model by integrating specific knowledge for a particular task, offer a promising strategy for improving reasoning-based reading comprehension through multi-teacher distillation [9]. However, these approaches do not clearly delineate or leverage the cognitive processes associated with System 1 and System 2 in human cognition. As a result, the student model may become more attuned to intuitive responses (System 1) without adequately incorporating the deliberate, reflective reasoning characteristic of System 2. This lack of a structured pathway to engage both systems could limit the model's ability to reason critically and solve complex problems.

3 Dual-process architecture

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. However, while demonstrating high proficiency in languages, its responses may be 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 thought processing [10], System 1 is an automatic system shaped by past experiences. Although it responds quickly, 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 the necessary and sufficient data one day? Regardless of the answer (whether "yes" or "no"), the system will encounter efficiency challenges in knowledge finding process.

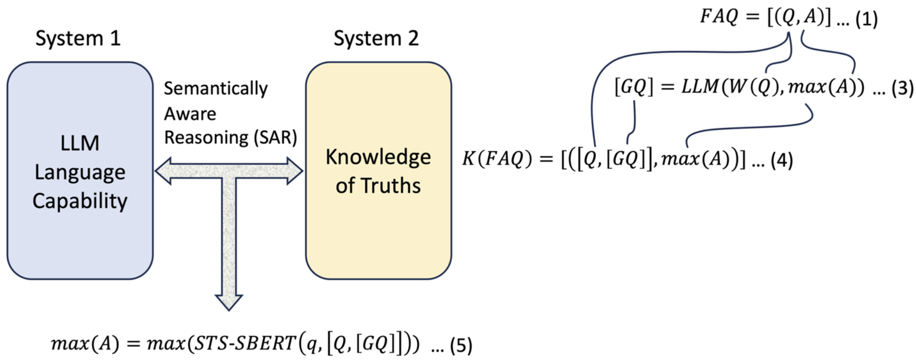

Fig. 4 Dual-process architecture describing semantically aware reasoning (SAR) between LLM as a System 1 and encapsulated knowledge in System 2

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 thought processing, equipped with knowledge about truths and SAR for the results obtained. SAR refers to the cognitive process of interpreting and understanding information using semantic context while being influenced by conscious awareness. It involves a thoughtful consideration of meaning and context in decision-making and problem solving. Figure 4 illustrates the components and the interrelation between System 1 and 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 us define it in the form of the following notation in Equation (1),

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

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. Unexpectedly, even when the FAQ is manually prepared, there is often a tendency to include an overwhelming amount of information in an attempt to be as comprehensive as possible. However, this can have a negative impact, introducing information redundancy and noise that hinder user experience. For example, in response to a simple query about directions to an office, the answer might include all possible routes, parking details, and even the office's postal address with telephone number. Without concise rewriting, this excessive information can overwhelm users and hinder efficient text processing during the knowledge preparation process. Consequently, during the knowledge preparation process, we use text summarization techniques ($f_{summarization}$) applying to the answers in the FAQ to extract the sentence containing the most essential information ($max(A)$) as formalized in Equation (2) [15].

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

As a result, a set of proper knowledge or a 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. However, this method may make users uncomfortable. Therefore, it can be expected that the language capability of LLMs can 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 as defined in Equation (3). T5 is a transformer-based neural network architecture developed by Google Research [16, 17].

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

In Equation (4), the knowledge extracted from the FAQ ($K(FAQ)$) is represented as a list of pairs, consisting of the original FAQ question (Q), the list of generated questions (GQ), and the important sentences in the original FAQ answer or $max(A)$. The generated questions (GQ) play an important role in enriching the information derived from the key sentences extracted from the original FAQ answer. As a result, a more comprehensive and accurate knowledge base of the FAQ can be established.

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

To obtain the correct answer ($max(A)$) from the FAQ, the user query (q) is verified against the list of original questions (Q) and the generated questions (GQ) in the SAR process. The $STS - SBERT$ model, fine-tuned from the BERT base model, is implemented with sentence embeddings using Siamese BERT networks, as formalized in Equation (5) [18]. In the SAR process, semantic textual similarity is effectively utilized to ensure accurate semantic alignment between the user query (q) and the generated questions (GQ), enabling the retrieval of the most appropriate answer from the FAQ.

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

4 Evaluation of the dual-process architecture on an FAQ dataset

Building on the dual-process architecture proposed in Section 3, we implemented the knowledge extraction and SAR processes using the Amagasaki FAQ dataset. The effectiveness of SAR was validated by demonstrating improved accuracy in textual semantic similarity measures compared to traditional text similarity measures. Additionally, during the knowledge preparation phase, the refinement of FAQ answers was evaluated to assess its impact on enhancing the accuracy of retrieving FAQ knowledge based on user queries (q).

4.1 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 a FAQ database containing a set of 1,786 questions and the corresponding answers on the FAQ page of Amagasaki city. The FAQ dataset is quite large and is prepared manually 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). $\langle Revised \rangle$ [Related FAQ] I want to know about the Regional General Center. $\langle Revised \rangle$ [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 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 the similarity measure against the test set of 784 user-generated queries, as shown in Table 2. The test set is prepared by Kyoto University using crowd-sourcing according to the explanatory answers in the FAQ [19].

The expression of a query is different from the question in the 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 candidates of answers in three groups of relation, that is, relevance (correct information), relate (relevant information) and the same group (same group of query but answer contains irrelevant information). For simplicity, we group all related responses into a list of relevant responses to measure similarity in the evaluation process.

4.2 Knowledge extraction from Amagasaki FAQ

In FAQ retrieval 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 [20] 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

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?	82	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]. $\langle Revised \rangle$ [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? $\langle Revised \rangle$ [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)

Table 3 Accuracy in similarity measure between query (q) and question (Q)

sim(q,Q)	mAP	Top5
SBERT	0.3600	0.6543
Fine-tuned SBERT	**0.4757**	0.7577

of words from the synsets. This method is integrated into the system architecture of FAQ database retrieval [19], employing query-question similarity measures in TSUBAKI [21], 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) [17], 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 found that the simple cosine similarity measure between sentences is ineffective in identifying appropriate questions. To evaluate this, we implemented cosine similarity matching between the query and the question using both the original SBERT model and the fine-tuned SBERT model. The results of the similarity measure show comparatively low performance for both models, as presented in Table 3.

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

Table 4 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,

present in the sentences, without consideration of the 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 STS-SBERT model [18] to measure the semantic similarity between the user query and question. In our experiment, we fine-tune the Japanese Sentence-BERT model which is generated from the base model of the Tohoku University NLP Lab. Knowledge of FAQ ($K(FAG)$) can be expressed as a list of pairs of original FAQ question (Q), the list of generated questions (GQ) and the important sentences in the original FAQ answer ($max(A)$).

5 Results and discussion

FAQ database normally contains a large number of pairs of a question and an answer. We search the FAQ database by finding the best match between 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 explain how to extreme the finding of the matched question which is only one representative question of the common questions for an answer. In fact, matching according to 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.

5.1 LLM knowledge distillation activated by question–answer pairs from FAQ

The generative model in LLMs has the potential to generate relevant sentences based on given keywords and context. In the question generation process, content words such as nouns, verbs, adverbs, and adjectives are extracted from the FAQ question sentence to create a list of keywords, and the corresponding FAQ answer sentences are used as context for T5 to generate a new corresponding question. The default hyperparameter is used to generate only one output to reduce evaluation complexity. In practical ChatBot use cases, multiple questions may be needed to increase the possibility of matching with other information in the answer. However, a list of relevant keywords must be prepared to correspond to the 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 a 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 4.

Table 4 shows the generated questions based on the questions and answers in the FAQ. The generated questions request the same information as the original question but express it differently. This is because the keywords from the question are provided during the generation

Table 5 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 6 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 7 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

process. The results of question generation can be used to offer variations of the question in the intent of a ChatBot.

The relationship between the question (Q), the generated question (GQ), and the query (q) is investigated by their similarity measures on the user-generated query test set. The experiments were conducted using both SBERT and fine-tuned SBERT models applied to Semantic Textual Similarity (STS). Table 5 shows how closely the meaning of GQ aligns with the original Q. Tables 6 and 7 show how closely the meanings of Q and GQ align with q, indicating how well the correct answer can be retrieved from the FAQ.

The value of mAP (mean Average Precision) represents the mean of the average precision scores for each query. Top1 refers to the accuracy measured by the correct answer found in the top position, while Top5 measures the accuracy of the correct answer found within the top five answers.

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 7) is quite low comparing to the original Q (Table 6). However, the GQ can somehow play an important role in covering the unseen query expressions of other users that cannot be matched well with the Q as reported in [22].

5.2 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. In addition, the most impactful consequences are the unrelated information texts that may come from the original source of the city guidebook and the service manual.

To refine the FAQ responses, we use a text summarization technique that identifies key sentences through feature-based important sentence extraction [15] and named entity recognition [23]. From a total of 1,786 answers in the Amagasaki FAQ database, 100 answers are randomly selected for annotating significant sentences. Each important sentence is rep-

Table 8 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)

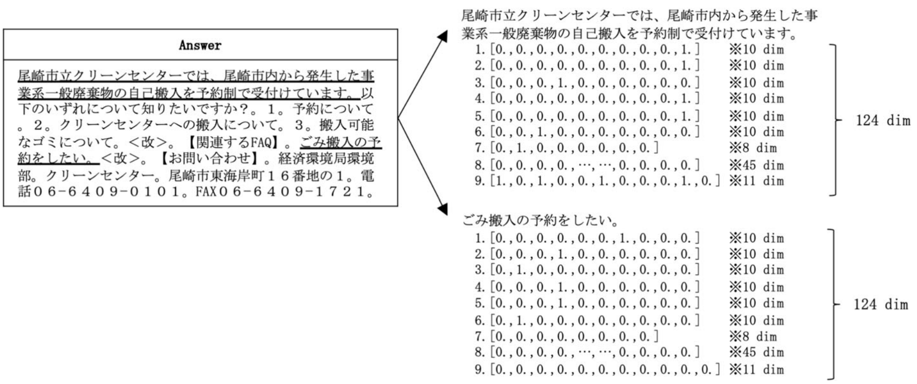

Fig. 5 124-dimensional sentence embedding vector of important sentences, an example from actual implementation on Japanese text

resented by nine features, as specified in Table 8, which form a 124-dimensional sentence embedding vector, as shown in Figure 5.

As a result, the important sentences were successfully extracted with an accuracy of 90.28% using the Light Gradient Boosting Machine (LightGBM), a decision tree-based classifier. The experiment was conducted on a sample of 100 responses, comprising 1,009 sentences for training and 432 sentences for testing.

By applying important sentence extraction to the original FAQ answers, the impact of unrelated sentences can be mitigated. This also improves the quality of the generated questions (GQ). Table 9 presents the new answer with the selected important sentences ($max(A)$) and the improved GQ, which, for instance, now includes a question about ‘the types of waste’ rather than one about ‘language use.’ This demonstrates a better alignment with the answer ($max(A)$).

In the original answer (A), bold text highlights the important sentences, while the underscored text in the last two columns indicates the differences between the two types of GQ.

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 knowledge distilled from LLM, derived from both erroneous and biased data, requires careful evaluation to ensure its functionality. The results presented in Table 9 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 10 demonstrates the improved quality of the generated questions (GQ) in representing Q, achieving a mAP of 0.5635 compared to 0.4852. Table 11 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.

As noted in Section 5.1, GQ plays an important role in increasing the possibility of matching with 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 12 shows how well the GQ can complement Q in finding the best match between Q and the GQ as formalized in Equation (5).

The mAP rate for the matching between q and the best matching of Q and GQ improved to 0.6165 (as shown in Table 12), compared to 0.6144 (in Table 6) and 0.5038 (in Table 11) in cases of individual matching. While the enhancement introduced by GQ is modest, its inclusion helps maintain the quality of query matching. The proposed approach, leveraging LLM for query expansion, demonstrates promising results in improving matching accuracy for previously unseen queries.

Compared to previous state-of-the-art (SOTA) approaches implemented for the same FAQ task, the best results of our proposed dual-process architecture, which implements SAR reasoning between the LLM and domain-encapsulated knowledge, surpass those reported in previous studies on FAQ retrieval, where the LLM is modified through fine-tuning or retraining methods. Specifically, our best result, obtained using Fine-tuned STS-SBERT, achieves a mAP of 0.6165, outperforms the TSUBAKI high-performance computing environment, which employs the WordNet expansion for query and question similarity measurement (0.558 mAP) and the BERT-based query and answer similarity measurement model (0.576 mAP) [19].

6 Conclusion and future works

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 alleviate the data insuf-

Table 9 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. $\langle Change \rangle$ [Related FAQ] **I want to make a reservation for waste delivery.** $\langle Change \rangle$ [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?**

Table 10 Accuracy in similarity measure between question (Q) and generated question (GQ)

sim(Q,GQ)	Answer with only Important Sentences ($max(A)$)			Original Answer (A)		
	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 11 Accuracy in similarity measure between query (q) and generated question (GQ)

sim(q,GQ)	Answer with only Important Sentences ($max(A)$)			Original Answer (A)		
	mAP	Top1	Top5	mAP	Top1	Top5
STS-SBERT	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

Table 12 Accuracy in similarity measure between query (q) and the best match of question (Q) and generated question (GQ)

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

ficiency 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 LLMs with a more holistic and reasoning-based understanding. This paper explores the integration of additional 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 knowledge preparation to improve LLM knowledge distillation, this study has shown the effectiveness of using important sentences, semantic textual similarity measures, and the generation of questions to improve 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. In addition, the results showcase improvements in the semantic representation of questions, leading to a more accurate matching with user queries. The use of SAR reasoning, particularly through the incorporation of important sentences, has proven to be a valuable strategy. The increased accuracy in matching user queries with relevant answers signifies the potential of this approach in refining information query systems.

As we progress, the findings of this research highlight several promising areas for further exploration. The proposed dual-process architecture, in particular, demonstrates the potential to reduce complexity in model preparation and fine-tuning, compared to traditional knowledge distillation approaches that require the creation of additional models for specific tasks. By separating the knowledge of truths from the LLM, which primarily handles language capabilities, this approach has proven effective in managing domain-specific tasks without requiring extensive modifications.

Future work could explore various methods of knowledge extraction and representation. While more complex representations, such as knowledge graphs [24], offer potential benefits, they also introduce additional challenges related to reasoning and practical utilization. Efficiently leveraging the language capabilities of LLMs is strongly recommended, following the principles of the dual-process model inspired by human cognition.

However, the use of semantic textual similarity in our approach has limitations in reasoning, particularly when knowledge is incomplete or absent. Future research could focus on refining semantic reasoning models, incorporating additional features, and addressing challenges across diverse datasets. Ultimately, this study provides valuable insights into optimizing LLM interactions with external knowledge through the integration of semantically aware reasoning mechanisms.

Acknowledgements The contributions of Ryusei Doi in sharing the data analysis results, which provided valuable insights for the study, are sincerely appreciated.

Author contributions V.S. designed the concept and create the architecture to prove in further steps. V.S. conducted the experiments and analyzed the results. V.S. reviewed the related works and prepared the concept for writing. V.S. composed the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Brundage M, Avin S, Clark J, Toner H, Eckersley yP, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen G (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. University of Oxford, Future of Humanity Institute
2. Yuan X, Xu C, Tan Y (2020) Deep Learning from a Statistical Perspective. Statistical Analysis and Data Mining: The ASA Data Science Journal 13(4):349–365
3. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 610–623. ACM, New York
4. Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level Concept Learning through Probabilistic Program Induction. Science 350(6266):1332–1338
5. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, vol 33. Curran Associates Inc, Red Hook, NY, USA, pp 1877–1901

6. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV (2022) Finetuned Language Models Are Zero-Shot Learners. In: Proceedings of the International Conference on Learning Representations (ICLR), pp. 1–20
7. Press O, Zhang M, Min S, Schmidt L, Smith NA, Lewis M (2023) Measuring and Narrowing the Compositionality Gap in Language Models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2023), pp. 5687–5711
8. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA
9. Zhao Z, Xie Z, Zhou G, Huang JX (2024) MTMS: Multi-teacher Multi-stage Knowledge Distillation for Reasoning-Based Machine Reading Comprehension. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1995–2005. Association for Computing Machinery, New York, NY, USA
10. Kahneman D (2011) Thinking, Fast and Slow. Farrar, Straus, and Giroux, New York
11. Mialon G, Dessí R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, Roziére B, Schick T, Dwivedi-Yu J, Celikyilmaz A, Grave E, LeCun Y, Scialom T (2023) Augmented Language Models: a Survey. Trans. Mach. Learn. Res
12. Tan S, Tam WL, Wang Y, Gong W, Zhao S, Zhang P, Tang J (2023) GKD: A General Knowledge Distillation Framework for Large-scale Pre-trained Language Model. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), pp. 134–148. Association for Computational Linguistics, Toronto, Canada
13. Welleck S, Kulikov I, Roller S, Dinan E, Cho K, Weston J (2020) Neural Text Generation with Unlikelihood Training. In: Proceedings of the International Conference on Learning Representations (ICLR)
14. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D, Bosma M, Zhou D, Metzler D, Chi EH, Hashimoto T, Vinyals O, Liang P, Dean J, Fedus W (2022) Emergent Abilities of Large Language Models. Preprint at Transactions on Machine Learning Research (TMLR) https://api.semanticscholar.org/CorpusID:249674500
15. Hirao T, Isozaki H, Maeda E, Matsumoto Y (2003) Key Sentence Extraction Method using Support Vector Machine. Journal of Information Processing Society of Japan 44(8):2230–2243
16. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C (2021) mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Proceedings of the the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 483–498
17. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Preprint at arXiv:1910.10683 (2020)
18. Reimers N, Gurevych I (2019) 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), pp. 3980–3990
19. Sakata W, Shibata T, Tanaka R, Kurohashi S (2019) 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), pp. 1113–1116
20. Fellbaum C (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge
21. Shinzato K, Shibata T, Kawahara D, Hashimoto C, Kurohashi S (2008) TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), pp. 189–196
22. Doi R, Charoenporn T, Sornlertlamvanich V (2022) Automatic Question Generation for Chatbot Development. In: Proceedings of the 7th International Conference on Business and Industrial Research (ICBIR2022), Bangkok, Thailand, pp. 301–305
23. Sornlertlamvanich V, Yuenyong S (2022) Thai Named Entity Recognition using BiLSTM-CNN-CRF enhanced by TCC. Journal of IEEE Access 10:53043–53052. https://doi.org/10.1109/ACCESS.2022.3175201
24. Xie Z, Zhang Y, Zhou G, Liu J, Tu X, Huang JX (2024) One Subgraph for All: Efficient Reasoning on Opening Subgraphs for Inductive Knowledge Graph Completion. IEEE Transactions on Knowledge & Data Engineering 36:8914–8927. https://doi.org/10.1109/TKDE.2024.3432767

Prof. Virach Sornlertlamvanich received his B.Eng. and M.Eng. degrees in Mechanical Engineering from Kyoto University, Japan, in 1984 and 1986, respectively, and later earned his Ph.D. in Computer Engineering from the Tokyo Institute of Technology in 1998. His academic and professional engagement in the fields of Knowledge Engineering and Artificial Intelligence began during his time at Kyoto University in the early 1980s. From 1988 to 1995, he was a key contributor to the Multilingual Machine Translation project, contributing significantly to advancements in language technologies. His long-standing contributions to natural language processing include the creation of Thailand's first part-of-speech tagged Thai corpus (ORCHID) and the design of the first corpus-based Thai-English dictionary (LEXiTRON), both released in 1997. He also led the implementation of the country's first interlingual-based English-Thai online machine translation service, setting a benchmark in applied computational linguistics. In 2009, Prof. Virach designed the "Digitized Thailand" project to establish an intelligent service platform that could serve as a foundational framework for digital content sharing and application integration. The initiative led to the digitization of cultural heritage and local knowledge, with practical applications in tourism, product design, and education. In 2020, he launched the "AI Ready City Networking in RUN" initiative, aimed at energizing cities through an extreme AI platform focused on opportunity and sustainability. Over the course of his career, Prof. Virach has received numerous prestigious honors. He was named the "Researcher of the Year 2001" by The Nation newspaper, received the "National Distinguished Researcher Award" in Information Technology and Communication from the National Research Council of Thailand in 2003, and was awarded the "ASEAN Outstanding Engineering Achievement Award" by the ASEAN Federation of Engineering Organizations in 2011. In recognition of his longstanding contributions, he was also honored with the "Outstanding Alumni Award" by the Tokyo Tech Alumni Association (Thailand Chapter) in 2021.