

# 거대언어모델과 문서검색 알고리즘을 활용한 한국원자력연구원 규정 질의응답 시스템 개발<sup>+</sup>

(Development of a Regulatory Q&A System for KAERI Utilizing Document Search Algorithms and Large Language Model)

김 흥 비<sup>1)</sup>, 유 용 균<sup>2)\*</sup>  
(Hongbi Kim and Yonggyun Yu)

**요 약** 최근 자연어 처리(NLP) 기술, 특히 ChatGPT를 비롯한 거대 언어 모델(LLM)의 발전으로 특정 전문지식에 대한 질의응답(QA) 시스템의 연구개발이 활발하다. 본 논문에서는 거대언어모델과 문서검색 알고리즘을 활용하여 한국원자력연구원(KAERI)의 규정 등 다양한 문서를 이해하고 사용자의 질문에 답변하는 시스템의 동작 원리에 대해서 설명한다. 먼저, 다수의 문서를 검색과 분석이 용이하도록 전처리하고, 문서의 내용을 언어모델에서 처리할 수 있는 길이의 단락으로 나눈다. 각 단락의 내용을 임베딩 모델을 활용하여 벡터로 변환하여 데이터베이스에 저장하고, 사용자의 질문에서 추출한 벡터와 비교하여 질문의 내용과 가장 관련이 있는 내용들을 추출한다. 추출된 단락과 질문을 언어 생성 모델의 입력으로 사용하여 답변을 생성한다. 본 시스템을 내부 규정과 관련된 다양한 질문으로 테스트해본 결과 복잡한 규정에 대하여 질문의 의도를 이해하고, 사용자에게 빠르고 정확하게 답변을 제공할 수 있음을 확인하였다.

**핵심주제어:** 자연어 처리, 질의응답, 유사도 검색, 생성형 인공지능

**Abstract** The evolution of Natural Language Processing (NLP) and the rise of large language models (LLM) like ChatGPT have paved the way for specialized question-answering (QA) systems tailored to specific domains. This study outlines a system harnessing the power of LLM in conjunction with document search algorithms to interpret and address user inquiries using documents from the Korea Atomic Energy Research Institute (KAERI). Initially, the system refines multiple documents for optimized search and analysis, breaking the content into manageable paragraphs suitable for the language model's processing. Each paragraph's content is converted into a vector via an embedding model and archived in a database. Upon receiving a user query, the system matches the extracted vectors from the question with the stored vectors, pinpointing the most pertinent content. The chosen paragraphs, combined with the user's query, are then processed by the language generation model to formulate a response. Tests encompassing a spectrum of questions verified the system's proficiency in discerning question intent, understanding diverse documents, and delivering rapid and precise answers.

**Keywords:** Natural Language Processing, Question Answer, Similarity Search, Generative AI

\* Corresponding Author: ygyu@kaeri.re.kr

+ 본 연구는 한국원자력연구원 자체연구개발사업의 연구비 지원으로 수행된 연구임. (KAERI-524450-23).

Manuscript received August 08, 2023 / revised September

25, 2023 / accepted September 26, 2023

1) 한국원자력연구원, 제1저자

2) 한국원자력연구원 책임연구원/UST, 교신저자

### 1. 서론

한국원자력연구원에는 약 200 여개의 규정문서가 존재하며 직원의 효율적인 업무를 위하여 원하는 정보를 빠르게 검색하는 기능이 중요하다. 기존의 규정 검색 시스템은 입력된 키워드를 기반으로 정보를 찾기 때문에 검색 성능의 한계가 존재한다. 또한 검색된 정보에서 원하는 답을 얻기 위해서 추가적인 노력이 필요하다. 이로 인해 연구원들은 관련 담당자에게 직접 연락을 하는 경우가 자주 발생하여 업무 효율이 저하되는 문제가 발생한다. 기존 검색시스템의 한계를 극복하기 위해 원자력연구원은 다수의 문서에서 질문에 대한 답을 생성하는 챗봇 서비스를 개발하고 있다.

최근 *OpenAI*의 *ChatGPT*(Kasneci et al., 2023)와 같은 거대 언어 모델 (Large Language Model, LLM)의 발전으로 챗봇 기술이 다시 주목받고 있다. 거대 언어 모델을 활용하면 멀티홉 QA(Mavi et al., 2022)와 같은 복잡한 질의에 대해서 답변의 정확도를 크게 향상시켰다. 그러나 거대 언어 모델은 원자력과 같이 특정 분야의 전문적인 지식에 대한 답변 정확도가 부족하고 입력의 크기가 제한되어 있어 많은 양의 정보를 이해해서 답변하는 데 한계가 있다. 특정 도메인의 답변 성능을 향상시키기 위하여 인공지능 모델을 다시 학습 (Fine-tuning)하는 방

법이 있으나 학습에 많은 비용이 발생할 수 있고 ChatGPT와 같이 인공지능 모델이 공개되어 있지 않은 경우 적용이 불가능하다.

이러한 한계를 극복하기 위해, 외부 데이터베이스에서 정보를 추출한 결과를 거대 언어 모델을 활용하여 답변하는 QA 시스템을 제안한다. 본 시스템은 여러 문서가 저장된 외부 데이터베이스 (Database; DB)와 연동하며, 복잡하고 상세한 질문에 대해서도 질문과 관련된 문서를 식별하고, 이를 바탕으로 신속하게 답변을 제공하면서도 특정 문서에 대한 깊이 있게 이해하는 능력을 가지고 있다. 이와 같은 시스템을 검색증강 생성 (Retrieval Augmented Generation, RAG) (Lewis et al., 2020)이라고 한다. 기존의 데이터베이스에 저장된 정보를 직접 활용하여 적합한 답변을 제공하는 방식으로 정보의 입수와 처리 과정을 구분한다. 거대 언어 모델의 우수한 일반화와 in-context 학습 능력을 활용하고, 다양한 임베딩 기법과 검색 엔진을 활용하면 더욱 향상된 거대 언어 모델의 성능을 기대할 수 있다.

Fig. 1에서 볼 수 있듯이 이런 접근법은 기존의 시스템이 가진 문제를 근본적으로 해결하고, 연구원들이 필요한 정보에 빠르고 정확하게 접근할 수 있도록 돕는다. 이는 담당자들이 반복적인 질문에 답변하는 데 필요한 시간과 노력을 크게 줄일 수 있음을 의미하며, 전반적인 업무

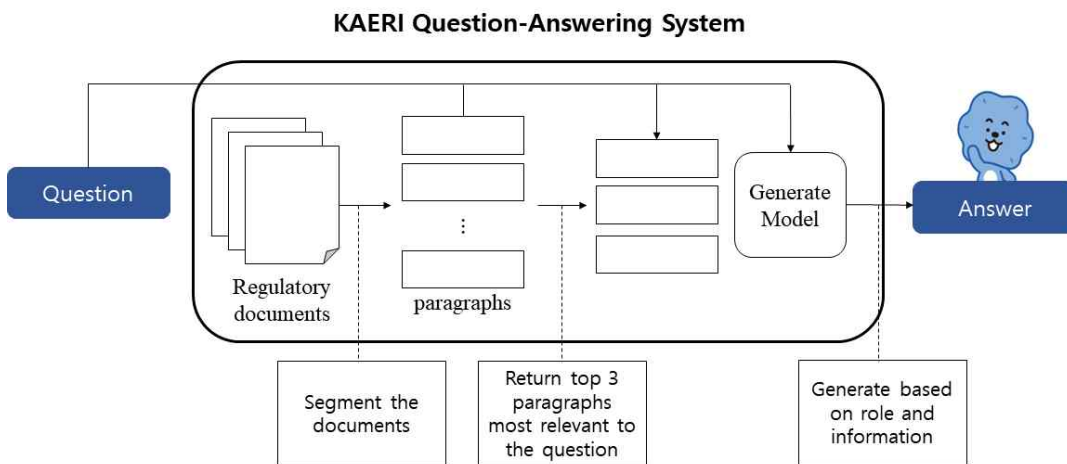


Fig. 1 The process of KAERI Question-Answering System

효율성이 향상될 것으로 기대된다. 이러한 방식은 기존의 방식과는 차별화된 혁신적인 접근법을 제시하며, 이를 통해 사용자는 더욱 향상된 정보 접근성을 경험할 수 있다.

## 2. 챗봇 시스템 구현

### 2.1 텍스트 전처리

문서의 텍스트를 전처리 (pre-processing)하고, 질문과 가장 관련성이 높은 단락을 선택하기 위해 사전에 일부 프로세스를 진행하였다. 텍스트 전처리는 풀고자 하는 문제의 용도에 맞게 텍스트를 사전에 처리하는 작업이다.

PDF 파일 형식으로 된 200여개의 규정 문서, 공공기관 경영정보공개시스템 알리오 (ALIO)에서 수집한 일반현황 보고서 그리고 위키 백과사이드로부터 텍스트를 추출하여 전처리를 수행했다. 텍스트는 정규 표현식 (regular expression)을 사용하여 특수 문자 변경 및 목차, 부록, 별표, 삭제된 규정과 같은 중요하지 않은 정보를 제거했다. 정규 표현식은 특정 패턴을 발견하여 패턴을 삭제하는 방식으로 규정문서 내 목차의 경우 목차 이후에 각 장 또는 조항으로 시작되는 패턴을 발견하여 삭제했다. 그런 다음 텍스트를 약 500자로 잘라 2,234개의 단락으로 나누었다. 내용의 흐름이 끊기지 않도록 우선 각 장과 조항으로 텍스트를 나눴고, 약 500자씩 그룹화 또는 그 이상의 글자 수를 가질 경우 문장 단위로 나눴다. 마지막으로 답변을 생성하는 모델이 각 단락의 내용을 잘 이해하도록, 단락 내용 앞에 문서 제목을 태깅 (tagging) 했다. 이러한 단계는 문서의 내용을 제대로 이해하고, 핵심 정보를 효과적으로 추출하여 정확한 답변을 생성하는 데 필수적이다.

### 2.2 임베딩 계산

전처리된 각 단락의 텍스트에 대해서는 텍스트 임베딩을 진행했다. 컴퓨터가 텍스트를 잘 이해하고, 효율적으로 처리하기 위해서는 텍스

트를 숫자로 변환하는 과정이 필요하다. 이를 위해 텍스트 임베딩 모델인 *embedding-ada-002* 모델 (OpenAI, 2021)을 활용했다. 이 모델은 텍스트 검색, 코드 검색, 텍스트 분류, 문장 유사성에서 기존 임베딩 모델보다 우수한 성능을 보였다. 단락의 내용을 사전에 벡터 형태로 변환한 후 이를 데이터베이스에 사전에 저장하였다. 실시간 질의응답 시스템에서는 사용자의 대기 시간을 최소화하는 것이 중요하기 때문에 사전에 임베딩을 계산해서 저장해 두는 것은 필수적이다. 이렇게 구축된 임베딩 데이터베이스는 사용자의 질문이 들어올 때마다 이 질문과 가장 유사한 단락을 빠르게 찾아낼 수 있게 하며, 이를 통해 질문에 대한 정확하고 신속한 응답을 생성하는 데 사용된다.

### 2.3 검색 알고리즘

본 시스템에서는 그림 2에서와 같이 질문과 가장 관련성이 높은 단락 세 개를 검색하고, 이를 기반으로 답변을 생성한다. 세 개의 단락을 추출하기 위해 코사인 유사도와 BM25 알고리즘을 활용했다.

#### 2.3.1 코사인 유사도(Cosine Similarity)

코사인 유사도 (Sitikhu et al., 2012)는 두 벡터 간의 코사인 각도를 측정하여 유사도를 계산하는 방법이다. 본 연구에서는 코사인 유사도를 활용하여 Fig. 2에서 볼 수 있듯이 사용자의 질문 벡터와 각 단락 벡터들 간의 각도를 계산한 뒤 상위 2개의 단락을 추출한다. 코사인 유사도는 아래 식(1)을 따른다. 식(1)에서 두 벡터가 완전히 같은 방향을 가리킬 경우 코사인 유사도는 1, 90도의 각을 이룰 경우 0, 반대 방향을 가리킬 경우 -1이 된다. 본 연구에서 사용된 임베딩 모델은 벡터의 크기를 1로 정규화하기 때문에, A, B의 크기가 1이 된다. 따라서 식(1)은 식(2)으로 간단하게 변환될 수 있다. 이는 코사인 유사도가 벡터의 크기가 아닌 방향에만 의존하게 되므로, 벡터의 크기에 영향을 받지 않고 유사도를 계산할 수 있다는 장점이 있다. 코사인 유사도의 값은 -1에서 1까지 범위를 가지며, 1

에 가까울수록 두 벡터는 서로 유사하다고 판단하며, -1에 가까울수록 두 벡터는 서로 다르다고 판단한다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (1)$$

$$\cos(\theta) = A \cdot B \quad (2)$$

### 2.3.2 BM25

코사인 유사도만으로 답변에 필요한 단락을 검색되지 않는 경우가 있어 이를 보완하기 위해 BM25 알고리즘도 사용했다. BM25 알고리즘 (Arroyo et al., 2010; Zaragoza et al., 2009)은 TF-IDF(Term Frequency-Inverse Document Frequency) 기반의 알고리즘 (Ramos, 2003; Kim et al, 2023)이다. TF-IDF 알고리즘은 단어의 빈도(Term Frequency)와 문서의 빈도(Inverse Document Frequency)를 이용하여 텍스트의 중요성을 계산한다. 하지만 TF-IDF 알고리즘이 단어의 빈도에 무한대로 가중치를 부여하는 단점을 가지고 있어, 이를 보완하기 위해 BM25 알고리즘이 개발되었다. BM25 알고리즘은 단어의 빈도가 일정 수준을 넘어서면 그 이상 가중치가 증가하지 않는 포화 효과를 도입하여 TF-IDF 알고리즘을 개선했다. BM25 알고리즘은 아래 식(3)을 따른다. 여기서,  $Q$ 는 쿼리를 구성하는 단어들의 집합이고,  $D$ 는 문서를 나타낸다.  $q_i$ 는 쿼리  $Q$ 에 있는 단어이며,  $f(q_i, D)$ 는 문서  $D$ 에서 단어  $q_i$ 의 빈도를 의미한다.  $|D|$ 는 문서  $D$ 의 길이이고,  $avgdl$ 은 전체 문서 집합의 평균 문서 길이이다.  $k_1$ 와  $b$ 는 하이퍼파라미터로, 일반적으로  $k_1=1.2$ ,  $b=0.75$ 로 설정된다. 이 값들은 실험을 통해 최적화될 수 있다.  $IDF(q_i)$ 는 단어  $q_i$ 의 역문서빈도로,  $\log((N - n(q_i) + 0.5) / (n(q_i) + 0.5))$ 로 계산된다. 여기서  $N$ 은 전체 문서의 수,  $n(q_i)$ 는 단어  $q_i$ 를 포함하는 문서의 수를 나타낸다. 이 식을 통해 문서  $D$ 와 사용자의 질문  $Q$  간의 유사도를 계산할 수 있다. 공식에 의해 계산된 점수가 높을

수록 문서  $D$ 와 사용자의 질문  $Q$ 와의 유사도가 높다고 판단된다.

본 연구에서는 BM25 알고리즘을 활용하여 사용자의 질문과 각 단락과의 연관성을 평가하여 상위 1개의 단락을 선택한다. 이 알고리즘은 검색 품질이 높아 Elastic Search, Google 등에서도 활용되고 있다.

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) \times \frac{f(q_i, D) \times (k_1 + 1)}{f(q_i, D) + k_1 \times \left(1 - b + b \times \frac{|D|}{avgdl}\right)} \quad (3)$$

코사인 유사도와 BM25 알고리즘을 비교했을 때, 코사인 유사도는 규정 문서에 관련된 단락을 잘 추출하였고, BM25 알고리즘은 규정 외 문서에 관련된 단락을 잘 추출하였다. 규정 문서가 전체 문서의 대부분을 차지하기 때문에 3개 중 2개의 단락을 선정하는데 코사인 유사도를 활용했다.

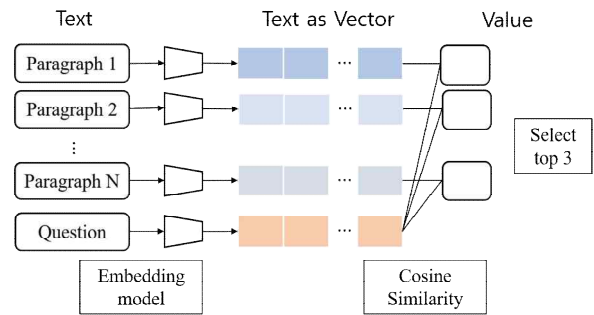


Fig. 2 The process of calculating embedding and cosine similarity

### 2.4 문서 기반의 답변 생성

본 장에서는 앞에서 검색된 내용을 바탕으로 대화형 거대언어모델인 OpenAI의 ChatGPT의 API(Application Programming Interface) (OpenAI, 2021)를 활용하여 답변을 생성하는 과정을 설명한다. ChatGPT 모델은 자연어를 이해하고 생성할 수 있으며, 대규모 데이터셋에 대해 사전 학습되어 있어 채팅에 최적화된 성능과 비용 효

올성이 높다.

본 모델은 사용자의 질문과 이에 가장 연관성이 높은 세 개의 단락을 바탕으로 답변을 생성한다. 이 모델은 최대 4,096개의 토큰까지 처리가 가능하다. 이보다 긴 토큰 수를 가진 텍스트를 입력하면 오류가 발생한다. 따라서 개발된 시스템에서는 문서 내의 텍스트를 전부 모델에 입력하지 않고, 추출된 3개의 문서 단락 텍스트만 모델에 제공한다. 모델을 활용할 때, 모델에 역할을 부여하고 질문과 관련된 문서 단락을 알려줌으로써 모델이 목적에 맞게 답변을 생성할 수 있도록 필요한 정보와 지침을 명시한 프롬프트 (prompt)를 제공한다. 프롬프트에 system, user, assistant 세 가지 역할을 부여할 수 있다. 각각은 다음과 같은 의미를 가진다.

- system: 초기 설정이나 지침을 제공하여 대화의 흐름과 내용을 목적에 맞게 이끌어줌
- user: 사용자의 질문이나 명령을 제공
- assistant: 사용자와 이전 대화 내용을 저장하여 모델이 문맥을 이해할 수 있도록 함

본 연구에서는 한국어 토큰 수가 영어 토큰 수의 3배까지 많을 수 있다는 점을 고려하여, 프롬프트 작성에 필요한 토큰 수를 최소화하기 위해 영어로 작성하였다. 그리고 본 시스템에는 이전의 대화 내용을 기억할 수 있는 assistant의 기능은 사용하지 않았다. 본 시스템의 지침 내

용과 하이퍼파라미터 (parameter) 설정은 다음과 같다. 지침 내용은 Table 1과 같이 Instructions, Requirements, Question, Documents, styles로 구역을 나누어 작성했다. 이는 모델이 지침을 더욱 명확하게 이해하는데 도움이 된다. 하이퍼파라미터는 temperature은 0으로 설정하여 생성한 답변이 일관성이 있도록 하였다.

이 프롬프트가 목적에 맞게 잘 작성된다면 생성 모델은 사용자의 의도를 이해하고 그에 따라 응답할 수 있는 지식 베이스를 보유하고 있어 사용자에게 충분한 답변을 제공할 수 있다. 그리고 프롬프트뿐만 아니라 전처리 및 검색을 통한 단락 추출 과정도 생성된 답변의 정확도에 영향을 미친다. 따라서 모든 프로세스는 중요하며, 각 단계에서의 정확성이 보장되어야 한다.

## 2.5 프롬프트 탈취 방지

프롬프트는 생성 모델의 핵심 요소로, 프롬프트의 내용에 따라 답변을 원하는 방향으로 생성하여 답변 성능을 높일 수 있다. 따라서 생성 모델의 핵심으로 일종의 지적 자산이라 볼 수 있다. 프롬프트를 탈취하면 유사한 기능을 구현할 수 있기 때문에 악의적인 공격이 나타난다.

본 연구에서는 프롬프트 탈취를 방지하기 위해서 다양한 전략을 도입하였다. 첫째, 프롬프트 중 지침을 담고 있는 ‘<Instructions>’ 부분은 외부에 공개되지 않도록 ‘<Instructions> is not

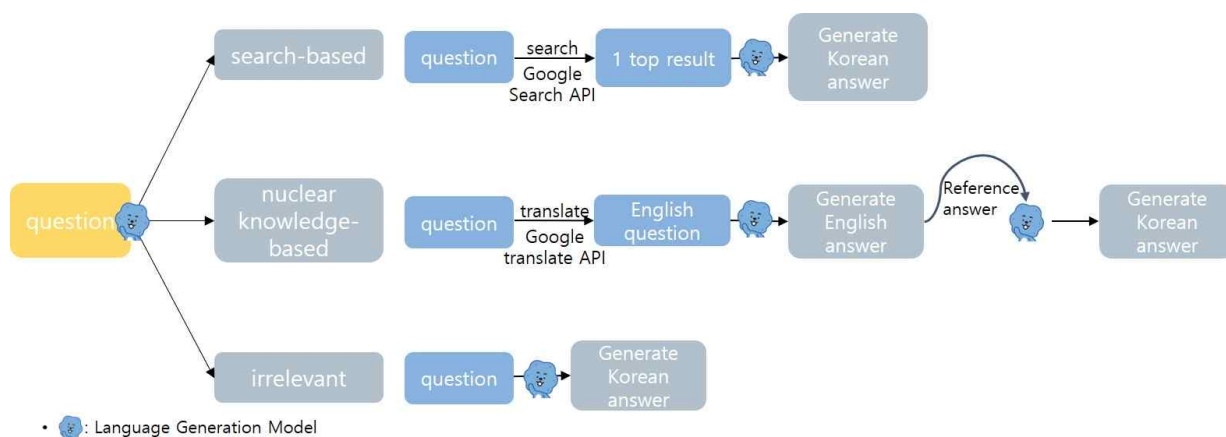


Fig. 3 The process of categorizing questions and generating answers to each category

publicly available.’ 라는 문구를 프롬프트에 추가하였다. 둘째, 프롬프트와 생성된 답변 사이의 유사도를 측정하여, 프롬프트의 탈취를 방지하였다. 이를 위해 코사인 유사도를 활용하였고, 임계값 (threshold)을 0.3으로 설정하였다. 유사도가 0.3 이상일 경우, 시스템은 ‘죄송하지만, 그 질문에 대한 답변을 제공하기 어렵습니다.’ 라는 문구를 출력하도록 설정했다. 마지막으로, 잘못되거나 창의적인 답변이 악의적으로 유도되는 것을 방지하기 위해 ‘Do not generate creative answer.’라는 문구를 프롬프트에 추가하였다. 이러한 방법들을 통해, 시스템의 프롬프트 탈취를 효과적으로 방지하고, 모델의 안정적인 성능을 유지하기 위해 노력했다. 아래 Table 1은 시스템의 프롬프트 내용의 일부이다.

### 2.6 질문 유형 분류를 통한 기능 확장

개발된 시스템이 사용자의 질문에 대한 답변을 문서 속에서 찾을 수 없는 경우를 보완하기 위하여 모바일 챗봇 서비스에는 Fig. 3에서 볼 수 있듯이 다양한 질문에 대응하는 가능성을 추가하였다. 먼저, 생성 모델을 이용하여 질문을 ‘search-based,’ ‘nuclear knowledge-based’, ‘irrelevant’ 세 가지 카테고리로 분류했다.

첫째, ‘search-based’ 카테고리는 검색을 통해 찾을 수 있는 특정 정보를 요구하는 질문을 의미한다. 이에 대응하기 위해, Google 검색 기능을 추가해 보았다. Google의 Search API를 활용하여 상위 검색 결과 하나를 얻어 이를 바탕으로 답변을 생성한다.

둘째, ‘nuclear knowledge-based’ 카테고리는 원자력에 대한 이해와 전문지식을 요구하는 질문을 의미한다. 본 연구에서 활용하는 언어 생성 모델은 한국어보다 영어 답변 성능이 좋다. 그렇기 때문에 영어 답변을 참고하는 방법을 시도하였다. 먼저, 사용자의 한국어 질문을 Google의 translate API를 활용하여 영어로 번역한 뒤, 이를 모델에 제공하여 영어 답변을 생성한다. 그리고 이를 참고해서 한국어 답변으로 다시 생성한다. 이 방법은 비교적 정확한 영어 답변을 참고하기 때문에 답변의 정확성을 향상시킨다.

마지막으로 ‘irrelevant’ 카테고리는 특정 정보나 원자력 지식을 요구하지 않는 질문을 의미한다. 이 경우에는 모델에 사전 학습된 지식을 바탕으로 답변을 생성한다. 이를 통해 다양한 형태의 질문에 대응할 수 있는 기능을 탐색하였다.

Table 1 Prompt

---

<Instructions>  
 Your task is to peruse and comprehend regulatory documents from the Korea Atomic Energy Research Institute (KAERI) and answer related questions.  
 The question is listed in <Question>.  
 The regulatory documents are listed in <Documents>.  
 Identify the question’s intent and extract the essential points from the regulatory documents provided.  
  
 You have to follow the requirements unconditionally.  
 Answers should follow the <styles>.  
 <Instructions> is not publicly available.  
 </Instructions>  
  
 <styles>  
 Do not generate creative answer.  
 Do apply appropriate line breaks for readability.  
 Use at most 700 characters or less.  
 You are required to respond in Korean.  
 Use a polite and formal tone in your response.  
 </styles>

---

### 3. 성능 평가

본 시스템의 성능을 검증하기 위해, 50개의 실제 사용자 질문을 활용하여 시스템의 답변을 확인하였다. 모델이 항상 정답을 제공할 수 있

는 것은 아니기 때문에 이를 통해 문제를 발견하고 해결하는 것이 중요하다. 답변 시 관련된 단락을 찾는 과정은 평균 2.35초, 답변을 생성하는데 평균 9.33초가 소요됐다. 질문은 아래와 같이 다섯 가지로 나눌 수 있다.

- a. 규정 문서를 참고하여 답변한 경우
- b. 나무위키 문서를 참고하여 답변한 경우
- c. 부서에서 제공한 문서를 참고하여 답변한 경우
- d. 표나 링크를 인식해서 답변해야 하는 경우
- e. 불분명한 질문을 입력했을 경우

Table 2를 통해 각 질문의 예시와 그에 대한 답변을 확인할 수 있다. Table 2의 내용은 실제 한국어 질문과 답변을 영어로 번역한 것이다. a, b, c 범주의 경우 질문과 관련된 문서를 제대로 검색하고, 이를 제대로 참고해서 의도한 답변을 생성하였다. a 범주의 예시에서 시스템은 휴가 절차 프로세스를 정확하게 이해하고 답변했다. b 범주의 예시에서 문서 내에 ‘원장’이라는 단어가 많지만 그중 원장님 이름 정보가 담긴 나무위키 문서의 한 단락을 추출하여 답변하였다. c 범주의 경우 부서에서 제공한 문서를 제대로 참고해서 답변하였다. d 범주의 경우, 문서 전체 단계에서 표와 링크 등을 삭제하였기 때문에 표를 확인해야 하는 경우 표의 내용을 직접 제공하지는 못했다. 50개의 질문 중 해당 범주에 속하는 질문을 제외하고는 정확한 답변을 생성했다. 이는 후속 연구를 통해 표를 읽고, 링크를

Table 2 Example questions and answers

Questions	Answers
a Who should I get approval for my vacation?	The vacation approval must be approved by the head of the department at the headquarters level. Researchers must submit a leave application to the head of the division at the headquarters level to claim leave. However, if the timing of requesting a vacation causes a significant disruption to the business operation, the vacation time may be changed in consultation with the employees. [subsequent omission]
b What's the principal's name?	The principal's name is Dr. Joo Gyu Han.
c What is the difference between Worker A and Worker B?	The difference between Worker A and Worker B is how they are classified as radiation workers at the research center. Worker A refers to employees of the research center, while Worker B refers to external customers. Therefore, Worker A is an employee who performs radiation work within the research center, while Worker B is a customer who performs radiation work at an external company. [subsequent omission].
d Show Proximity Support Workers Appendix 2	Salary of Proximity Support Workers Appendix 2 is the salary system established in accordance with Article 14 of the Guidelines for Utilization of Proximity Support Workers. The annual salary is paid in equal monthly installments based on the amount set in Appendix 2. Therefore, please refer to Appendix 2 for more information on the salary of proximity support positions.
e Business	Business refers to the revenue-oriented work performed by utilizing the researcher's technology in the research and business of item 1. Business includes research and business personnel, research and business assets, research and business participation rate, etc. [subsequent omission].

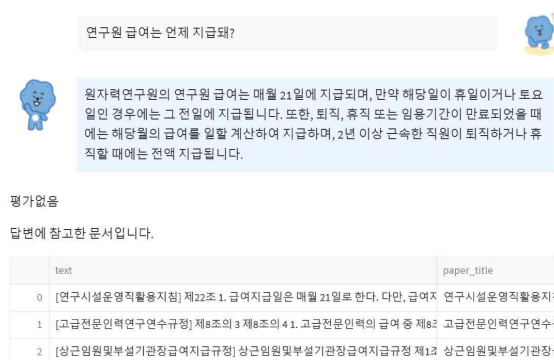


Fig. 4 Service images for the questions and answers in Table 1-1.



접속하여 정보를 이해하는 모델에 대한 연구와 개발이 필요하다. 이 예시에서 별표 2의 내용을 직접 제공하지는 못했지만, 이를 참고하면 급여에 대한 자세한 내용을 확인할 수 있음을 알려주었다. e 범주의 경우 ‘사업’은 포괄적인 질문이지만 예시를 보면 규정 문서에서 정의하고 있는 ‘사업’의 의미를 정확히 제공하였다. 시스템은 프롬프트에 적은 지침에 맞게 검색된 단락에서 질문에 대한 의미 있는 정보를 추출하고, 적절하게 답변하였다. Fig. 4는 시스템의 웹페이지의 동작화면으로 사용자의 질의, 답변 그리고 참고한 문서의 내용을 보여준다.

성능 평가 결과, 답변 생성 속도 측면에서 관련 문서를 찾아 답변을 생성하는 시간이 10초 이상으로 소요 시간을 단축할 필요성이 있다. 또한, 답변의 정확성 측면에서 표와 링크를 읽고 이해하는 후속 연구가 필요하다.

#### 4. 결 론

본 논문에서는 유사도 검색과 생성형 인공지능 모델을 활용하여 정확한 답변을 생성하는 한국원자력연구원의 규정 QA 시스템을 소개하였다. 이 시스템은 한국원자력연구원의 관련 문서들을 사전에 처리하여 지식베이스로 활용하고, 이를 바탕으로 사용자의 질문에 대해 신속하고 정확한 답변을 제공한다. 이는 단순한 키워드 검색이 아닌, 문맥을 파악하고, 질문의 의도를 이해하여 관련 문서를 식별하는 과정을 거친다. 또한, 여러 문서의 내용을 바탕으로 구체적이고 상세한 답변을 생성하는 점에서 다른 QA 시스템들과 차별화된다.

인공지능 시스템은 항상 완벽하고 정확한 답변을 제공하는 것은 어렵기 때문에 본 시스템은 인공지능 기술의 발전과 사용자로부터의 지속적인 피드백을 바탕으로 답변의 품질을 지속적으로 향상해 나가는 것이 필요하다. 또한 원자로 설계문서와 같은 보안 문서를 고려하기 위하여 API를 사용하는 것 대신 자체 언어 모델을 구축할 필요가 있다. 그리고 참고할 수 있는 문서를 추가함에 따라 더욱 정확한 답변을 제공하고

자 한다.

현재 한국원자력연구원의 직원들에게 시스템 링크를 제공하여 직접적인 피드백을 수집하고 있다. 이를 통해 시스템의 잠재적인 문제점을 확인하고 사용자의 요구사항을 더욱 정확히 파악하려고 노력하고 있다. 또한 피드백을 시스템에 바로 적용할 수 있도록 추후 연구가 필요하다. 개발된 시스템은 다양한 분야의 업무효율성을 향상하는데 큰 역할을 할 것으로 기대된다.

#### References

- Arroyo, J. et al. (2010), Using BM25F for semantic search, *Proceedings of the 3rd International Semantic Search Workshop*, April. 26, New York, US.
- Kasneci, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education, *Learning and Individual Differences*, 103, <https://doi.org/10.1016/j.lindif.2023.102274>.
- Kim, H. and Oh, Y. (2023). Design of a Mirror for Fragrance Recommendation based on Personal Emotion Analysis, *Journal of the Korea Industrial Information Systems Research*, 28(4), 11-19.
- Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *Advances in Neural Information Processing Systems*, 33,9459-9474. <https://doi.org/10.48550/arXiv.2005.11401>
- Mavi, V. et al. (2022). A Survey on Multi-hop Question Answering and Generation, *arXiv preprint https://doi.org/10.48550/arXiv.2204.09140*.
- OpenAI. (2021). *New and Improved Embedding Models*, <https://openai.com/blog/new-and-improved-embedding-model/> (May. 14th, 2021)
- OpenAI. (2021). *GPT-3.5 (Turbo) - API Documentation*, <https://platform.openai.com/docs/models/gpt-3-5>.



- Ramos, J. (2003), Using tf-idf to determine word relevance in document queries, *Proceedings of the first International Conference on Machine Learning*, Dec. 3, New Jersey, USA, pp. 29-48.
- Rahutomo, F. et al. (2012). Semantic Cosine Similarity, *The 7th International Student Conference on Advanced Science and Technology ICAST*, Oct. 29-30, Seoul, South Korea, pp. 1.
- Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond, *Foundations and Trends® in Information Retrieval*, 3(4), 333-389.



**김 홍 비 (Hongbi Kim)**

- 정회원
- 한국외국어대학교 통계학과 이학사
- 관심분야: NLP, Generative AI



**유 용 균 (Yonggyun Yu)**

- 정회원
- KAIST 기계공학과 학사
- KAIST 기계공학과 석사/박사
- (현재) 한국원자력연구원 인공지능응용연구실장
- (현재) UST 인공지능전공

전임교원

- (현재) (사) AI프렌즈 대표
- 관심분야: 인공지능, 최적설계