Embedding-Based Response Blocking Algorithm for Enhancing the Reliability of Domain-Specific Language Models in the Atomic Energy Industry

Seungdon Yeom[a,b] , Soyeon Kim [a,b], Yonggyun Yu[a,b*]
[a]*Korea Atomic Energy Research Institute, 111, Daedeok-daero 989 beon-gil , Yuseong-gu , Daejeon, 34057, Korea*
[b]*Korea National University of Science & Technology, 217, Gajeong-ro, Yuseong-gu , Daejeon, 34113,Korea*
[*]*Corresponding author: ygyu@kaeri.re.kr*

Keywords : atomic energy, domain-specific language model, large language model reliability

1. Introduction

With the continuous advancement of language models, their use in various industrial fields is increasing. Consequently, the development of language models specialized for each industry domain is also being actively pursued. As the application scope of language models expands across different fields, issues of reliability and security in the responses generated by generative language models are also being raised. Due to these emerging reliability issues, research on methodologies to improve the reliability of large-scale language models is continuously being conducted. [1] Among the research aimed at enhancing reliability, studies on building red teaming is the most representative.

Furthermore, just as models are being developed for specific domains, such as healthcare and finance, specialized models are being developed for the atomic energy industry. However, to be used in specialized fields such as the atomic energy industry, where security and confidentiality are important, there is a need for methods to control responses generated based on data. This paper proposes a process for developing atomic energy domain-specific language models and propose an embedding-based response blocking algorithm to enhance the reliability of response generation in atomic energy industry-specific language models.

For this purpose, to train the language model with atomic energy domain knowledge, 18000 datasets were collected from KHNP's “Nuclear Glossary,” “Glossary of Nuclear Laws and Regulations,” Nuclear Safety and Security Commission's “Nuclear Safety Regulation Glossary,” Korea Atomic Energy Research Institute's “Nuclear-related Academic Papers,” and Seoul National University Nuclear Policy Center's “Atomic Wiki,” [2-6] and performed fine-tuning by applying the LoRA (Low-Rank Adaptation) [7] technique. Also, 784 data sets from the Nuclear Safety and Security Commission's "Accident and Failure Investigation Reports” [8] were collected, and these data were assumed to be security data that should not be included in responses. Based on this, the proposed embedding-based response control algorithm using cosine similarity is used to demonstrate the response control process of generative language models. This study aims to explore the potential for improving reliability in large language models through response output control.

2. Proposed Method

In this chapter the process of developing a atomic energy domain-specific model and the design and experimentation of an embedding-based algorithm are described.

2.1 Development of a Atomic Energy Domain-Specific Language Model

In this section, the process of data collection, model training, and performance analysis undertaken to develop a atomic energy domain-specific language model is presented. Initially, atomic energy documents were collected and preprocessed to build a training dataset. This was followed by further pre-training and instructional tuning using a Korean pre-trained language model. Finally, a QA (question-answering) evaluation dataset was developed, and the model's performance was analyzed using key metrics.

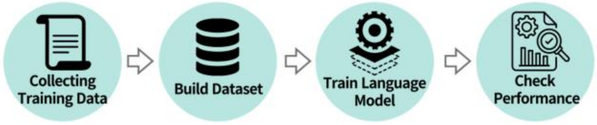


Fig. 1. Training Process of a Domain-Specific Language Model for Atomic Energy

2.1.1 Data acquisition

To develop a language model specific to the atomic energy field, atomic documents were first collected from various sources. A total of 18,000 datasets were gathered from diverse data sources, including KHNP's “Nuclear Glossary” and “Glossary of Nuclear Laws and Regulations,” the Nuclear Safety Commission's “Nuclear Safety Regulation Glossary,” the Korea Atomic Energy Research Institute's “Academic Papers on Nuclear Power,” and the Seoul National University Nuclear Policy Center's “Atomic Wiki.” The collected data underwent preprocessing steps such as deduplication, document cleanup, and format conversion to transform it into a dataset format suitable for language model training. This process ensured data consistency and facilitated efficient learning by the model.

2.1.2 Training Language Model

To develop a language model specialized for the atomic energy domain, a large-scale language model pre-trained in Korean was used as the base model. The model selected for this purpose was “MLP-KTLim/llama-3-Korean-Bllossom-8B,” a Korean language model based on the Llama architecture, known for its excellent performance in Korean natural language processing. The learning process involved two main phases: additional pre-training and instruction-tuning, aimed at enhancing the model's performance in specific tasks. During the additional pre-training phase, the existing Korean language model was further trained to acquire a deeper understanding of the atomic energy domain using a text dataset focused on atomic energy, building on the knowledge gained from general texts. The LoRA technique was applied in this phase to effectively internalize domain-specific information while efficiently utilizing the existing model parameters. This technique maximized the efficiency of further pre-training by reducing the number of parameters and conserving memory and computational resources, all while maintaining the model's performance. In the instruction-tuning phase, the pre-trained model was fine-tuned to enhance its ability to perform specific tasks within the atomic energy domain. Instruction-tuning involved providing the model with targeted instructions or examples to handle tasks such as atomic energy safety assessment and technical report writing, enabling optimal performance for each task. The LoRA technique was again employed to expedite and optimize the tuning process, allowing the model to be trained with clear instructions and appropriate datasets for each task. This approach improved the model's practicality and accuracy in the atomic energy domain.

2.1.3 Analysis Performance

To evaluate the performance of the trained model, a atomic energy related QA evaluation dataset was constructed, and the model's performance was analyzed by comparing the answers generated by the model with the correct answers in the evaluation dataset. In this process, two main metrics were employed to assess the appropriateness of the use of important terms in the atomic energy domain and the similarity of the answers: ROUGE-L and Cosine Similarity.

- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence)** is a metric that measures the textual similarity between the generated answers and the correct answers in the evaluation dataset, with a particular focus on evaluating textual matches based on close sequences. This metric is useful for evaluating whether the model accurately generated atomic energy terms and sentences.
- Cosine Similarity evaluates the similarity between two texts by measuring the cosine angle between the generated answer and the vectorized representation of the correct answer. This metric is useful for assessing the semantic similarity of texts and is used to determine how semantically consistent the model generated answers.

These metrics were converted into percentages to serve as quantitative evaluation metrics.

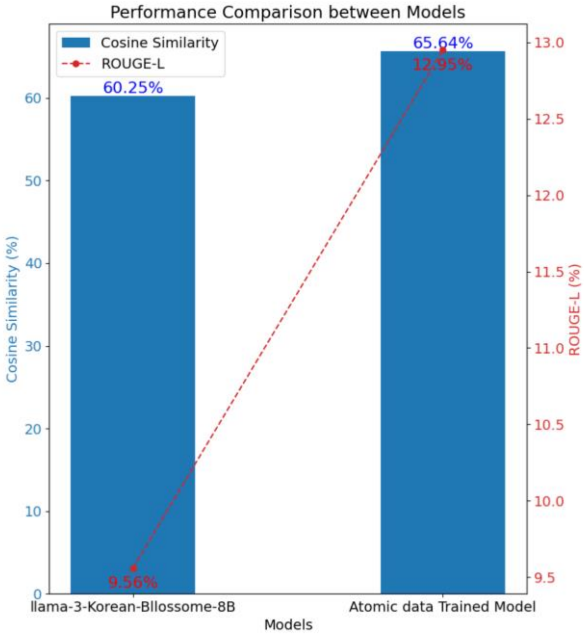


Fig. 2. Performance comparison of the model before and after training on the Atomic energy domain

The evaluation results showed that the performance of the trained model improved by 3.4% in ROUGE-L score and 5.4% in Cosine Similarity compared to the base model. These results suggest that the trained model has improved its ability to use the terminology and understand the context required in the atomic energy domain.

2.2 Embedding-Based Response Blocking Algorithm

In this section, the process for designing and experimenting with the embedding-based response blocking algorithm is detailed. The design of the embedding-based response blocking algorithm comprises three main steps: data acquisition, generation of data embedding, and response blocking through similarity comparison.

2.2.1 Data acquisition

The process of data acquisition for creating embeddings used in the embedding-based response control algorithm is detailed. To generate the "Nuclear Incident Data Set," texts were extracted by crawling the Nuclear Safety and Security Commission's "Accident and Failure Investigation Reports" page. The extracted texts were converted into JSON files to facilitate the embedding generation. The format of the resulting JSON dataset, used for the final embedding creation, is shown in Table I. After preprocessing, a total of 783 entries for the "Nuclear Incident Data Set" were generated.

Table I: Nuclear Incident Data Set

### Nuclear Incident Data Set Example

```
{
        "번호": 3,
        "시설": "고리 1 호기",
        "발생일자": "1979-02-06 12:17",
        "사건제목": "주급수펌프 B 의
Breaker 소손으로 4.16KV bus(XSW-1A)가
저전압되면서 원자로냉각재펌프 A 가 저전압으로
trip 되어 원자로 및 터빈발전기 정지",
        "원자로출력": "100%",
        "발전기출력": "580MWe",
        "계통": "2 차",
        "원인": "전기",
        "정지유형": "없음",
        "등급": "없음",
        "내용": "주급수펌프 B breaker 의
steel strip 과 insulated copper Arc
chute 사이에서 flush over 에 의해 소손이
발생하여 4.16KV Bus 1A(XSW-1A)가 저전압이
되고 원자로냉각재펌프 A 의 저전압에 의한
trip 으로 2.6 12:17 원자로 및 터빈발전기가
정지됨. 480V Bus 3B(XSW-3B) station
transformer 용 breaker 를 주급수펌프 B
breaker 로 대체하고 2.6 14:28 원자로임계 도달
및 19:34 계통병입을 실시함"
    }
```

2.2.2 Embedding-Based Response Blocking Algorithm

The algorithm consists of four main steps : 1) Question embedding, 2) Loading security dataset embeddings, 3) Similarity calculation, and 4) Determining whether to block a response. These steps proceed as follows

- Question embedding:
The question entered by the user is vectorized using a pre-trained embedding model. This vector contains the semantic representation of the question and is used to calculate the similarity to the security dataset in a later step. The embedding model used here is "sentence-transformers/all-MiniLM-L6-v2."

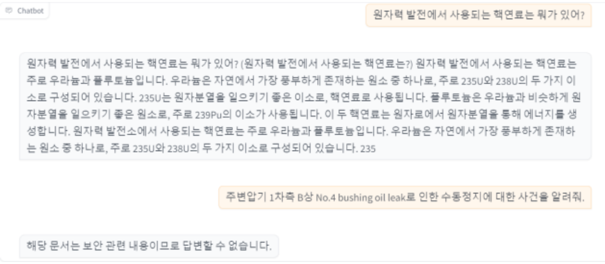
- Loading security dataset embedding:
In this study, we utilize a pre-built security dataset. This dataset contains sensitive data such as atomic energy related incident information, each of which is converted into a vector using an embedding model and stored. When the algorithm is run, these embedding vectors are loaded into memory and used.

- Similarity calculation:
The similarity between the loaded security dataset embedding and the question embedding is calculated. The similarity calculation is done by measuring the angle between the vectors, using the Cosine Similarity method. This similarity value indicates the semantic similarity between the question and the security data.

- Determining whether to block a response:
If the calculated similarity value exceeds a predefined threshold(e.g., 0.8), the question is considered to have a high similarity to the security data. In this case, the algorithm blocks the generation of a response and returns the user with the message "The document is security-related and cannot be answered." On the other hand, if the similarity does not exceed the threshold, the normal response generation process proceeds with the trained language model.

2.2.3 Experimental Analysis

To evaluate the effectiveness of the embedding-based response blocking algorithm, the algorithm was validated to ensure that security-related questions are properly blocked. The goal was to ensure that it appropriately blocks responses to questions containing sensitive information, and to visually verify this.



원자력 발전에서 사용되는 핵연료는 뭐가 있어?

원자력 발전에서 사용되는 핵연료는 뭐가 있어? (원자력 발전에서 사용되는 핵연료는?) 원자력 발전에서 사용되는 핵연료는 주로 우라늄과 플루토늄입니다. 우라늄은 자연에서 가장 풍부하게 존재하는 원소 중 하나로, 주로 235U와 238U의 두 가지 이소로 구성되어 있습니다. 235U는 원자분열을 일으키기 좋은 이소로, 핵연료로 사용됩니다. 플루토늄은 우라늄과 비슷하게 원자분열을 일으키기 좋은 원소로, 주로 239Pu의 이소가 사용됩니다. 이 두 핵연료는 원자로에서 원자분열을 통해 에너지를 생성합니다. 원자력 발전소에서 사용되는 핵연료는 주로 우라늄과 플루토늄입니다. 우라늄은 자연에서 가장 풍부하게 존재하는 원소 중 하나로, 주로 235U와 238U의 두 가지 이소로 구성되어 있습니다. 235

주변압기 1차측 B상 No.4 bushing oil leak로 인한 수동정지에 대한 사건을 알려줘.

해당 문서는 보안 관련 내용이므로 답변할 수 없습니다.

Fig. 3. Results of blocking questions about security data

The Fig. 3 above shows the results of the algorithm successfully blocking responses to security-related questions. As you can see, if the question has a high similarity to security data, the message "This document is security-related and cannot be answered" is displayed.

3. Conclusions

In this study, a specialized language model for atomic energy was developed using atomic energy domain data.

The specialized model was designed to deeply understand the specialized terminology and context used in the atomic energy field, and showed higher accuracy and reliability in processing atomic energy-related information compared to general language models.

Subsequently, a research was conducted to leverage this specialized language model to enable response control based on sentence embedding and cosine similarity. The algorithm was designed to protect sensitive data in the atomic energy domain, and experiments showed that it was successful in blocking answers to questions that exceeded a set threshold. This confirmed that the embedding-based approach is effective in protecting sensitive information.

However, limitations that cosine similarity alone cannot completely block sensitive data was found, especially when sentence structures are complex, contain metaphorical expressions, or when unrelated sentences containing similar terms exist.

Nevertheless, the response control system based on the atomic energy-specific language model demonstrated the feasibility of improving reliability through response output control in large-scale language models, which is an important foundation for security filtering techniques to contribute to improving the response quality and reliability of language models.

ACKNOWLEDGMENTS

This work was supported in part by Korea Atomic Energy Research Institute R&D Program under Grant KAERI-524540-24.

REFERENCES

[1] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, Universal Adversarial Triggers for Attacking and Analyzing NLP, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Nov. 3-7, 2019, Hong Kong, China.

[2] 한국수력원자력(주)_원자력용어집(2019). https://www.data.go.kr/data/15038485/fileData.do?recommendDataYn=Y (accessed Aug, 16, 2024).

[3] 한국수력원자력(주)_원자력관련법령 용어집(2014). https://www.data.go.kr/data/15002295/fileData.do (accessed Aug, 16, 2024).

[4] 원자력안전위원회 원자력안전규제용어사전. https://www.nssc.go.kr/ko/cms/FR_CON/index.do?MENU_ID=2460 (accessed Aug, 14, 2024).

[5] 한국원자력연구원_국내 원자력 관련 최신 동향 발표자료 목록(2020). https://www.data.go.kr/data/3077573/fileData.do (accessed Aug, 16, 2024).

[6] Atomic Wiki (2023) https://atomic.snu.ac.kr/index.php/%EB%8C%80%EB%AC%B8 (accessed Aug, 16, 2024).

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Proceedings of the International Conference on Learning Representations (ICLR 2022), Jan. 29, 2022, Virtual.

[8] 원자력안전정보 모음(KINS) 분야별안전정보. https://nsic.nssc.go.kr/information/reguDataActive.do?nsicDtaTyCode=nppAccient (accessed Aug, 14, 2024).