*Digital Health and Informatics Innovations for Sustainable Health Care Systems*
*J. Mantas et al. (Eds.)*


*doi:10.3233/SHTI240482*

# Well-Tempered Medical Prompt Engineering for Explainable Extubation

Sujung LEE[a,b], Won Ik CHO[c], Chansung PARK[d] and Taehoon KO[a,1]
[a] *Department of medical informatics, College of Medicine, The Catholic University of Korea, Republic of Korea*
[b] *Medical AI LAB, MODULABS, Republic of Korea*
[c] *Seoul National University, Republic of Korea*
[d] *Electronics and Telecommunications Research Institute, Republic of Korea*
ORCiD ID: Sujung Lee https://orcid.org/0000-0001-6151-2497, Won Ik Cho https://orcid.org/0000-0002-8882-9125, Chansung Park https://orcid.org/0009-0003-3310-1617, Taehoon Ko https://orcid.org/0000-0002-4045-0036

**Abstract.** This study investigated whether the large language model (LLM) utilizes sufficient domain knowledge to reason about critical medical events such as extubation. In detail, we tested whether the LLM accurately comprehends given tabular data and variable importance and whether it can be used in complement to existing ML models such as XGBoost.

**Keywords.** Medical Prompt Engineering, Explainability, Large Language Model

## 1. Introduction

Explainable AI (XAI) has become increasingly important in the medical domain, as it helps to build trust and transparency in the decision-making process of AI models. However, many current medical AI models lack sufficient explainability, which hinders their adoption in clinical practice. Large language models (LLMs), trained with massive diverse texts, have the potential to enhance the explainability of medical AI models due to their remarkable reasoning and comprehension abilities [1].

We aim to investigate whether LLMs possess domain-specific knowledge in critical care medicine and if they can offer context-aware reasoning about medical events such as extubation failure. However, LLMs exhibit several limitations when asked to explain model decisions without appropriate prompt engineering (only the clinician's persona and input value). LLMs may stick to specific words (e.g., extubation failure), explain general medical knowledge, repeat some variables, yield columns with missing values, or provide false information. In this study, we aimed to overcome these limitations by thorough prompt engineering and leverage LLM's inherent knowledge and reasoning capability in the medical domain, focusing on prediction of extubation failure.

[1] Corresponding Author: Taehoon Ko, PhD; E-mail: thko@catholic.ac.kr.

## 2. Method

We built an XGBoost model (accuracy 0.900, sensitivity 0.837, specificity 0.923, AUROC 0.966) using MIMIC-III dataset [2] for binary classification of reintubation within 48 hours after extubation. We quantified each variable's contribution using SHapley Additive exPlanations (SHAP) [3]. Prompt engineering was tested by GPT-4. Initially, providing model results, SHAP values, and clinician's persona as LLM input displayed some limitations mentioned above (repetition, hallucination, etc.) To address these issues, we employed tempered prompt engineering techniques: a) Providing variables terminology/descriptions, b) Instructing to choose top 3 variables based on SHAP, c) Avoiding generating random values for missing value variables, d) Focusing on specific topics instead of general medical knowledge.

## 3. Results, Discussion and Conclusions

The prompt was designed with the following criteria: to consistently understand the given data and task, to interpret SHAP values and medically relate the importance of variables, and to provide an explanation if there is a difference between the model and real clinical practice (details to be shared online). Through a qualitative analysis, we found that providing explanations of the model's decisions and processes via LLMs can offer the following benefits. First, it provides insights into the model's behavior, which is helpful when the model's decisions differ from those of clinicians. Second, even in similar environments, model decisions can vary, and additional explanations can help identify the reasoning behind such decisions, allowing experts to improve the AI model.

We investigated LLMs to enhance the explainability of medical AI model for predicting extubation failure. Leveraging LLM's knowledge and reasoning capabilities through prompt engineering, we developed a reliable approach to improve medical AI models. This approach utilizes natural language to enhance human understanding of the model's decision-making process, making the models more trustworthy and acceptable to healthcare professionals and patients. Future research should focus on evaluating the explanations by medical experts and validating the generalizability of this approach.

**Acknowledgement:** This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C1074). Additionally, it received backing from the Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

## References

[1] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023 Aug 03;620:172–180. doi:10.1038/s41586-023-06291-2.

[2] Johnson A, Pollard T, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. doi:10.1038/sdata.2016.35.C

[3] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advanced in Neural Information Processing Systems 30 (NIPS2017).