MedICL: In-Context Learning for Semantically Enhanced AKI Prediction in Cardiac Surgery

Chenyang Su[1,2⋆], Yishun Wang[5], Boqiang Xu[2,3], Rong Feng[1,2], Lei Du[5⋆⋆], Hongbin Liu[2,3], and Gaofeng Meng[2,3,4]

[1] City University of Hong Kong, Hong Kong SAR
[2] Centre for Artificial Intelligence and Robotics, HK Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong SAR
[3] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[5] West China Hospital, Sichuan University, Chengdu, China

Abstract. Cardiac surgery is associated with the risk of acute kidney injury (AKI), which can lead to prolonged hospital stays and increased mortality. Accurate prediction of AKI before its onset could significantly improve patient outcomes. However, existing AKI prediction models primarily focus on numerical features such as laboratory values and vital signs, while overlooking textual features, including preoperative diagnoses and surgical procedures. To address this limitation, we propose MedICL, which applies in-context learning (ICL) to the cardiac surgery domain. By leveraging the powerful comprehension and reasoning capabilities of large language models, MedICL enables the integration of textual and numerical features for AKI prediction. Nevertheless, the performance of ICL is highly sensitive to the quality of the provided examples, potentially limiting its effectiveness. To overcome this challenge, we introduce a Semantic Matching Unit (SMU), which selects semantically relevant examples for each sample, thereby significantly enhancing the model's performance. Furthermore, we observed that ICL-based AKI predictions often suffer from instability and exhibit suboptimal performance on downstream tasks. To address these issues, we developed the Task Adaptability Enhancer (TAE), which calibrates the prediction probabilities generated by ICL on the validation set. This approach not only stabilizes the model's outputs but also enhances its adaptability to specific task scenarios. A series of experiments on the datasets collected from West China Hospital (WCH) demonstrated that MedICL achieved state-of-the-art performance. These results highlight the indispensable role of medical text data in AKI prediction for cardiac surgery scenarios, showcasing its potential to improve clinical practice.

Keywords: In-context Learning · Surgical Data Science · Acute Kidney Injury (AKI) Prediction.

⋆ C. Su and Y. Wang—Contributed equally.
⋆⋆ Corr. authors: dulei@scu.edu.cn, gfmeng@nlpr.ia.ac.cn.

1 Introduction

Acute kidney injury (AKI) is a common and potentially life-threatening complication of cardiac surgery, with an incidence ranging from 5% to 42% [2]. It significantly increases mortality and healthcare costs [23]. Since no effective treatments currently exist for established AKI, early and accurate identification of high-risk patients is critical for prevention through proactive management. Therefore, real-time, personalized predictions of AKI during or after cardiac surgery are essential to mitigating its impact and reducing related complications [1].

In recent years, machine learning (ML) methods have been widely applied to AKI prediction. The extensive adoption of electronic medical record (EMR) systems in hospitals has significantly enhanced the efficiency and accuracy of patient clinical data collection, providing a rich source of data for developing ML-based AKI prediction models [3]. Most of these methods have demonstrated promising predictive performance, with studies covering both critically ill adults [4] and pediatric patients [5]. However, research on AKI prediction in the population undergoing cardiac surgery is limited. Additionally, conventional machine learning methods for AKI prediction primarily rely on numerical data (such as laboratory values and vital sighs) [1] and make limited use of textual data (such as preoperative diagnoses and surgical procedures).

To address the above limitations, We propose **MedICL**, a novel framework built upon the powerful text understanding and few-shot learning capabilities of large language model [6], making it highly suitable for cardiac surgery scenarios with abundant textual data but limited overall data availability [7]. Our contributions are summarized as follows: First, to the best of our knowledge, we are the first to propose applying in-context learning (ICL) to the AKI prediction task, enabling the complete dataset, including textual data, to be input into the model for end-to-end AKI prediction. Second, considering that ICL heavily relies on the quality of examples[8], this framework introduces a Semantic Matching Unit (SMU) a plug-and-play module that selects the most semantically relevant examples for each sample based on embedding similarity, thereby improving ICL prediction performance. Third, we introduce the Task Adaptability Enhancer (TAE), a corrective mechanism that enables ICL to output a probability distribution for classification instead of directly predicting a single label. The probability distribution is further calibrated to ensure more stable outputs and better adaptability to downstream tasks[9]. Finally, extensive experiments on a real-world dataset highlight the promising performance of the proposed method, **MedICL**, while ablation studies further validate the effectiveness of each module.

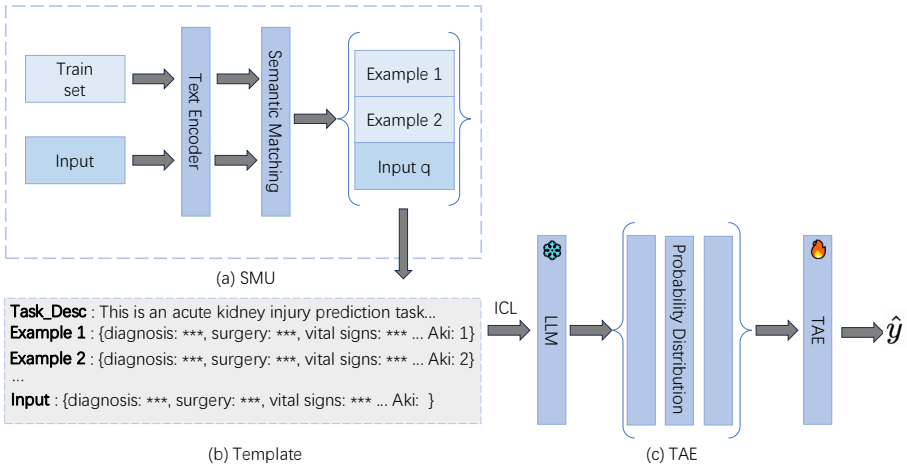


Fig. 1: The overview of our proposed MedICL framework. Panel (a) illustrates the workflow of the Semantic Matching Unit (SMU), where the input and training set are processed by a text encoder to perform semantic matching. Panel (b) shows how the matching results are filled into a prompt template to structure the input. In Panel (c), multiple outputs from the large language model(LLM) are averaged and refined by the Task Adaptability Enhancer (TAE) to produce the final output.

2 Methodology

2.1 Preliminaries

In-Context Learning(ICL) is a framework that enables language models to learn tasks using only a few examples provided as demonstrations [6]. Formally, given an input query text q and a set of candidate answers $A = \{a_1, \ldots, a_m\}$, a pretrained language model $\mathcal{L}$ selects the candidate answer with the highest score as its prediction, conditioned on a demonstration set D. The set D consists of an optional task instruction T and k demonstration examples. Thus, D can be expressed as $\{T, e(q_1, a_1), \ldots, e(q_k, a_k)\}$ or $\{e'(q_1, a_1, T), \ldots, e'(q_k, a_k, T)\}$, where $e'(q_i, a_i, T)$ represents an example formatted in natural language according to the task. Depending on whether the k demonstration examples belong to the same task, the problem can be categorized as task-specific in-context learning (ICL) or cross-task ICL. In the cross-task ICL setting, each example may have its own distinct instruction.

The likelihood of a candidate answer a_j is defined as:

$$P(a_j \mid q, D) = g_{\mathcal{L}}(a_j, D, q),$$

where $g_{\mathcal{L}}$ is a scoring function computed by the pretrained language model $\mathcal{L}$, given the candidate answer a_j, the context D, and the input q. The final predicted

label $\hat{a}$ is the candidate answer with the highest likelihood:

$$\hat{a} = \arg\max_{a_j \in A} P(a_j \mid q, D).$$

From the above, we can summarize that:(1) ICL shares similarities with few-shot learning, as both involve the use of a small number of examples to perform a task. However, the key difference lies in their approach: few-shot learning typically requires updating model parameters to adapt to the task [10], whereas ICL operates directly on pretrained large language models (LLMs) without any parameter updates. (2) Similarly, ICL is closely related to prompt learning, as it incorporates demonstration examples into the prompt to guide the model's predictions. Despite this resemblance, the distinction is that prompt learning may involve either discrete templates or continuous soft prompts, which are designed to elicit desired outputs [11], while ICL specifically relies on natural language examples formatted as part of the input prompt.

2.2 Proposed Methodology

Semantic Matching Unit The choice of examples has a significant impact on the performance of in-context learning (ICL). Randomly selecting examples can lead to substantial fluctuations in prediction results and often produces suboptimal outcomes. Previous studies have demonstrated that selecting examples from the training set that are semantically more similar to the query q improves the model's performance [12]. Therefore, it is crucial to match each q with the top-k semantically most similar examples in a personalized manner.

We propose the **SMU** to address this challenge. Firstly, we preprocess the data to extract only the textual components, denoted as x_i^{text}, which include information such as the medical histories, preoperative diagnoses, and intraoperative procedures. Both the test query q^{text} and the extracted textual components x_i^{text} from the training set are then fed into a sentence encoder $\mu_\theta(\cdot)$ to obtain their vector representations $v_q = \mu_\theta(q^{\text{text}})$ and $v_i = \mu_\theta(x_i^{\text{text}})\,(i = 1, 2, \ldots, N)$.

Secondly, the similarity between v_q and v_i is computed using measures such as cosine similarity, $s_i = \frac{v_q \cdot v_i}{\|v_q\| \|v_i\|}$, or negative Euclidean distance, $s_i = -\|v_q - v_i\|_2$. Then, the k training examples with the highest similarity scores $s_{\sigma(1)} \geq s_{\sigma(2)} \geq \cdots \geq s_{\sigma(k)}$ are selected.

Afterward, these selected examples $\{x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(k)}\}$ and their corresponding outputs $\{y_{\sigma(1)}, y_{\sigma(2)}, \ldots, y_{\sigma(k)}\}$ are concatenated to form the in-context learning demonstration set $D = \{(x_{\sigma(1)}, y_{\sigma(1)}), (x_{\sigma(2)}, y_{\sigma(2)}), \ldots, (x_{\sigma(k)}, y_{\sigma(k)})\}$. Finally, the constructed demonstration set D, along with the test query q, is fed into a pretrained language model $\mathcal{L}$ to generate the prediction $\hat{y}$, where:

$$\hat{y} = \mathcal{L}(q, D).$$

Task Adaptability Enhancer ICL lacks robustness to variations in prompt templates and the order of demonstrations in practical applications, leading to unreliable predictions [27]. Moreover, pretrained large language models may not

be well-adapted to downstream tasks [13], such as predicting AKI in cardiac surgery patients, which we aim to address.

To tackle these issues, we propose a probabilistic calibration approach, **TAE**, which adjusts the output probability distribution rather than directly predicting discrete labels. First, the ICL model generates a raw probability distribution $\mathbf{p} \in \mathbb{R}^m$ over possible classes for a given query q and a demonstration set D. Next, the raw probabilities $\mathbf{p}$ are calibrated using an affine transformation:

$$\tilde{\mathbf{p}} = \text{softmax}(\mathbf{A}\mathbf{p} + \mathbf{b}),$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a weight matrix and $\mathbf{b} \in \mathbb{R}^n$ is a bias vector. To optimize the calibration parameters $\mathbf{A}$ and $\mathbf{b}$, we utilize a small validation set $\{(\mathbf{x}_i^v, y_i^v)\}_{i=1}^{D_v}$, generating corresponding prompts P_i^v for each validation sample. The optimization objective minimizes the loss function, such as cross-entropy, between the calibrated probabilities and the true labels:

$$\min_{\mathbf{A},\mathbf{b}} \sum_{i=1}^{D_v} \text{Loss}(\theta^*, \mathbf{A}, \mathbf{b}; P_i^v).$$

This optimization is performed using gradient-based methods, initialized with zeros or random values, and exhibits robustness to initialization choices. Finally, the calibrated probabilities $\tilde{\mathbf{p}} \in \mathbb{R}^m$ are used to make the final prediction on test set:

$$\hat{y} = \arg\max \tilde{\mathbf{p}},$$

ensuring the output is robust, reliable, and better aligned with downstream tasks.
Overall Structure An overview of the proposed framework is illustrated in Figure 1. Our framework consists of a Semantic Matching Unit (SMU) for selecting semantically relevant examples and a Task Adaptability Enhancer (TAE) for calibrating output probabilities. The SMU ensures contextual alignment of the demonstration set, while the TAE adjusts raw probabilities using an affine transformation. Additionally, we designed a task-specific prompt template that incorporates the task description, background knowledge, and relevant examples, enabling the large language models (LLMs) to better understand the task [24].

3 Experiments

We conduct experiments to evaluate the performance of our proposed framework, MedICL, in predicting AKI in patients who underwent cardiac surgery. Additionally, we conduct ablation studies to verify the contribution of each module in the framework and evaluate its performance under various experimental settings, further enhancing its practicality.
Dataset. We analyzed and systematically extracted medical records of patients who underwent cardiac surgery at WCH. From this process, we derived the Adult

Table 1: Performance comparison between MedICL and conventional methods under the Numerical-Only and Text-Augmented scenarios.

Methods	**Numerical-Only**			**Text-Augmented**		
	Prec ↑	**Rec ↑**	**F1 ↑**	**Prec ↑**	**Rec ↑**	**F1 ↑**
Logistic Regression	0.58	0.62	0.60	0.66	0.63	0.64
Random Forest	**0.68**	0.62	**0.65**	0.75	0.70	0.72
XGBoost	0.60	**0.66**	0.63	0.73	0.68	0.70
MedICL (Ours)	0.70	0.58	0.63	**0.80**	**0.75**	**0.77**

Table 2: Ablation experiments of MedICL on ACSD. To evaluate the impact of different components, we: (i) remove the Semantic Matching Unit (SMU), (ii) remove the Task Adaptability Enhancer (TAE), and (iii) disabled the multi-sampling with probability averaging (PA) strategy.

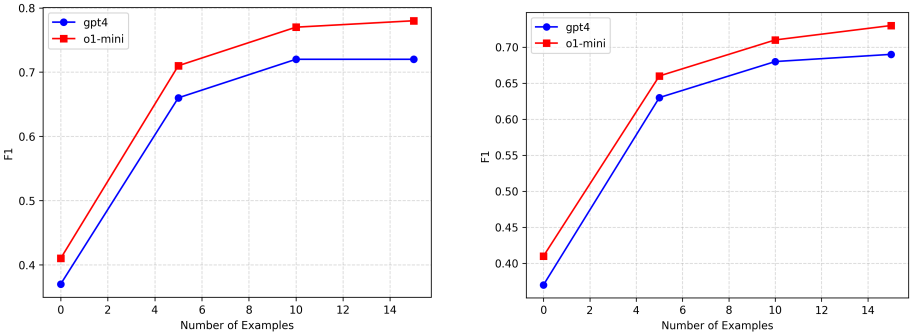
Method	**SMU**	**TAE**	**PA**	**Prec ↑**	**Rec ↑**	**F1 ↑**
Baseline	×	×	×	0.62	0.54	0.58
+ SMU	✓	×	×	0.75	0.70	0.72
+ TAE	×	✓	×	0.68	0.63	0.66
MedICL	✓	✓	×	0.80	0.75	0.77
MedICL+PA	✓	✓	✓	**0.87**	**0.78**	**0.82**

cardiac surgery Dataset (ACSD), which contains detailed information on 5,104 adult patients (aged ≥ 18 years) who received cardiac surgery. Certain patients were excluded from the study, including those with preoperative renal dysfunction (serum creatinine $> 176\,\mu$mol/L or requiring renal replacement therapy), those undergoing emergency surgery, those who died in the operating room, and those requiring intra-aortic balloon pump (IABP) or extracorporeal membrane oxygenation (ECMO) to discontinue cardiopulmonary bypass (CPB) during the procedure.

The data were collected using a standardized form and included patient demographics, NYHA (New York Heart Association) classifications, ASA (American Society of Anesthesiologists) physical status, preoperative laboratory results, medical history, preoperative medications, and intraoperative details. The primary focus of this study was the severity of acute kidney injury (AKI), which was defined and categorized based on the AKIN classification system [14].

Evaluation Metrics Following previous works[16], for AKI prediction, we evaluate our framework using the macro F1 score across all AKI categories (0, 1, 2, 3). Additionally, we report precision and recall to further assess its performance.

Implementation Details The dataset we used includes laboratory values, vital signs, preoperative diagnoses, surgical procedures, and the final AKI severity levels. The dataset is split into training, validation, and test sets in a 6:2:2 ratio. To achieve semantic matching, we generate text embeddings using the text-embedding-ada-002 model [20]. In the prompt template for in-context learning



(a) Cosine Similarity Results (b) Euclidean Distance Results

Fig. 2: Ablation experiments of MedICL on ACSD were conducted to evaluate its performance under different settings: (1) varying models (GPT-4 and O1-mini), (2) the number of demonstration examples (ranging from 0 to 15), and (3) different semantic similarity calculation methods (cosine similarity and euclidean distance).

(ICL), We explicitly define the task, provide the necessary medical background knowledge, emphasize the progressive worsening of conditions as the AKI classification level increases, and require the LLM to output probability distributions. We implement ICL by calling the API and leveraging the o1-mini[15] and gpt-4 models.

4 Results and Discussion

AKI Prediction In previous studies on AKI prediction tasks across various populations, conventional methods such as Random Forest [19], XGBoost[18], and Logistic Regression [17] have been widely used. However, these approaches often focus solely on numerical features while neglecting textual information. Textual information, such as clinical notes, often contains rich contextual and domain-specific knowledge that can significantly improve predictive performance. Therefore, we compared our proposed MedICL method with conventional methods on ACSD under two scenarios: the Numerical-Only scenario, which uses only numerical features, and the Text-Augmented scenario, which incorporates both numerical and textual information.

As shown in Table 1, under the Numerical-Only setting, Random Forest achieved the best performance, with a macro F1 score of 0.65, while MedICL ranked second with a macro F1 score of 0.63, lagging by 0.02. In this setting, textual information was removed from both the demonstration set D and the test query q. The absence of textual information likely hindered the large language model's ability to interpret the numerical features, resulting in suboptimal performance of MedICL.

Under the Text-Augmented setting, textual embeddings generated by text-embedding-ada-002[20] were incorporated into the training process of conventional methods. MedICL achieved the best performance, with a macro F1 score of 0.77, outperforming the best conventional method (Random Forest) by 0.05. Moreover, all methods showed performance improvements compared to the previous setting. These results highlight the significance of medical textual data in AKI prediction for cardiac surgery scenarios. Benefiting from the extensive pre-trained knowledge and reasoning capabilities of large language models [25, 26], ICL excels at understanding and analyzing tasks with only a few examples, leading to the best performance.

Ablation Study The ablation experiments conducted on the ACSD dataset highlight the importance of each component within our proposed MedICL framework, as shown in Table 2. We evaluated all methods using precision, recall, and F1. The baseline refers to randomly matching each query q with a demonstration set D. The Semantic Matching Unit (SMU) significantly improved performance, increasing the F1 score from 0.58 (baseline) to 0.72, demonstrating the critical role of semantically relevant demonstrations in ICL. Meanwhile, the Task Adaptability Enhancer (TAE) achieved a substantial performance boost through domain alignment, raising the F1 from 0.58 (baseline) to 0.66. Lastly, we validated the effectiveness of the multi-sampling with probability averaging (PA) strategy, which further improved the F1 from 0.77 to 0.82. By integrating these three components, our proposed MedICL framework achieved the best performance in predicting AKI among patients undergoing cardiac surgery.

As shown in Figure 2, we evaluated the performance of MedICL under different settings. When using o1-mini for ICL, it outperformed GPT-4, despite GPT-4 achieving excellent performance on the Open Medical-LLM Leaderboard [21, 22]. This is likely because o1-mini demonstrates superior reasoning capabilities, leading to better overall performance. We also analyzed the impact of the number of demonstration examples matched to each query. From the results, performance improves significantly when the number of demonstration examples increases from 0 to 5. However, beyond 10 demonstration examples, the performance gain becomes marginal. From a cost-effectiveness perspective, 10 examples are sufficient to achieve optimal results. Lastly, we evaluated the effect of different semantic similarity calculation methods. Cosine similarity outperformed Euclidean distance, as it is not affected by normalization and accounts for vector direction differences, making it more suitable for semantic similarity calculations.

5 Conclusion

In this study, we propose MedICL, a novel framework based on in-context learning (ICL) that leverages the powerful understanding and reasoning capabilities of large language models (LLMs) for AKI prediction in patients undergoing cardiac surgery. MedICL integrates the Semantic Matching Unit (SMU), which personalizes the demonstration set D by selecting the most relevant examples based

on semantic similarity for each query q in ICL. Additionally, it incorporates the Task Adaptability Enhancer (TAE), which adjusts the probability distribution to ensure that the output is robust, reliable, and better aligned with the context of cardiac surgery. Experimental results demonstrate that MedICL achieves superior performance on the ACSD dataset compared to conventional AKI prediction methods. By introducing medical text data into the AKI prediction task through ICL, we believe that MedICL paves a new pathway for AKI prediction. In the future, we plan to apply our method to larger datasets and more diverse medical scenarios, such as ICU patients and hospitalized patients.

Acknowledgements

This research was supported by the innoHK project and partially by the National Natural Science Foundation of China (Grant No. 62376267). We also thank West China Hospital (WCH) for providing the clinical data used in this study.

Disclosure of Interests

The authors declare no competing interests.

References

1. Zhang, Y., Xu, D., Gao, J. et al. Development and validation of a real-time prediction model for acute kidney injury in hospitalized patients. Nat Commun **16**(68) (2025).
2. Wang, Y., Bellomo, R. cardiac surgery-associated acute kidney injury: risk factors, pathophysiology and treatment. Nat Rev Nephrol **13**, 697–711 (2017).
3. Hu, Junlong et al. Identification and validation of an explainable prediction model of acute kidney injury with prognostic implications in critically ill children: a prospective multicenter cohort study. EClinicalMedicine **68** 102409. 5 Jan. (2024)
4. Shawwa, Khaled, et al. Predicting Acute Kidney Injury in Critically Ill Patients Using Comorbid Conditions Utilizing Machine Learning. Clinical Kidney Journal 1428–35 Apr. (2021)
5. Dong, Junzi, et al. Machine Learning Model for Early Prediction of Acute Kidney Injury (AKI) in Pediatric Critical Care. Critical Care, Dec. (2021)
6. Dong, Qingxiu et al. A Survey for In-context Learning. ArXiv abs/2301.00234 (2023)
7. Wang, Yi Shun et al. Prediction of the severity of acute kidney injury after cardiac surgery. Journal of clinical anesthesia **78** (2022)
8. Min, Sewon, et al. Rethinking the role of demonstrations: What makes in-context learning work?. arXiv preprint arXiv:2202.12837 (2022).
9. Abbas, Momin, et al. Enhancing in-context learning via linear probe calibration. International Conference on Artificial Intelligence and Statistics. PMLR, (2024).
10. Wang, Yaqing, et al. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) 1-34 53.3 (2020)
11. Liu, Pengfei, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys 1-35 55.9 (2023)

12. Liu, Jiachang, et al. What Makes Good In-Context Examples for GPT-3?. arXiv preprint arXiv:2101.06804 (2021)
13. Abbas, Momin, et al. Enhancing in-context learning via linear probe calibration. International Conference on Artificial Intelligence and Statistics. PMLR, (2024)
14. Schneider, Antoine, and Marlies Ostermann.The AKI Glossary. Intensive Care Medicine **43**(6) 893-97 (2017)
15. OpenAI. Openai o1-mini: Advancing cost-efficient reasoning (2024). URL https://openai.com/index/ openai-o1-mini-advancing-cost-efficient-reasoning/
16. Lin, Y., et al.: Acute Kidney Injury Prognosis Prediction Using Machine Learning Methods: A Systematic Review. Kidney Medicine **7**(1), 100936 (2024)
17. Song, X., et al.: Comparison of Machine Learning and Logistic Regression Models in Predicting Acute Kidney Injury: A Systematic Review and Meta-Analysis. International Journal of Medical Informatics **151**, 104484 (2021)
18. Song, X., et al.: Comparison of Machine Learning and Logistic Regression Models in Predicting Acute Kidney Injury: A Systematic Review and Meta-Analysis. International Journal of Medical Informatics **151**, 104484 (2021)
19. Lin, K., et al.: Predicting In-Hospital Mortality of Patients with Acute Kidney Injury in the ICU Using Random Forest Model. International Journal of Medical Informatics **125**, 55–61 (2019)
20. OpenAI. New and improved embedding model (2022). URL https://openai.com/index/new-and-improved-embedding-model/
21. Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, Beatrice Alex: Open Medical-LLM Leaderboard (2024). URL https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard
22. Singhal, K., et al.: Large Language Models Encode Clinical Knowledge arXiv preprint arXiv:2212.13138 (2022)
23. Hobson, C. E., et al. Acute kidney injury is associated with increased long-term mortality after cardiothoracic surgery. Circulation **119**(18), 2444–2453 (2009).
24. Brown, T., Mann, B., Ryder, N., Subbiah, M. et al. Language models are few-shot learners. Adv Neural Inf Process Syst **33**, 1877–1901 (2020).
25. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving language understanding by generative pre-training. (2018).
26. Radford, A., Wu, J., Child, R., Luan, D. et al. Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019).
27. Xie, S.M., Lin, H.W., Savarese, P., Tambe, M. et al. An explanation of in-context learning as implicit Bayesian inference. arXiv preprint arXiv:2111.02080 (2021).