DELL: Dual-Knowledge Enhanced LLMs for Precise Decision Making in Healthcare

Danying Mo
Sun Yat-sen University
Guangzhou, China
mody3@mail2.sysu.edu.cn

Chao Yu*
Sun Yat-sen University
Guangzhou, China
yuchao3@mail.sysu.edu.cn

Xuan Lin
Sun Yat-sen University
Guangzhou, China
linx227@mail2.sysu.edu.cn

Zhongqi Wu
South China Normal University
Guangzhou, China
wuzhongqi@m.scnu.edu.cn

Yuheng Luo
Sun Yat-sen University
Guangzhou, China
luoyh226@mail.sysu.edu.cn

Chen Bai
Hong Kong Metropolitan University
Hong Kong, China
baichenjack@aliyun.com

Chaojin Chen*
The Third Affiliated Hospital of Sun Yat-sen University
Guangzhou, China
chenchj28@mail.sysu.edu.cn

ABSTRACT

Despite the significant success in complex natural language understanding and reasoning, existing large language models (LLMs) often struggle to generate accurate outputs in tasks that require precise numerical predictions. This issue becomes even more challenging in healthcare applications, where critical dosage decisions usually demand reliable quantitative reasoning, given no explicit quantitative evidences. To address this challenge, we propose a framework termed **Dual-knowledge Enhanced LLMs (DELL)**, which enhances the precise clinical decision-making capability of LLMs by fusing their internally encoded commonsense knowledge with domain-specific quantitative knowledge automatically mined from medical data using external models. Specifically, DELL employs multiple explainable models to distill multi-source, quantitative knowledge representations from electronic health records, which are then incorporated into the reasoning process of LLMs to resolve conflicts, achieve a consensus, and make final precise decisions. Extensive experiments on intravenous fluid and vasopressor dosing tasks for sepsis in intensive care units are conducted to evaluate the effectiveness of DELL and demonstrate its consistent superiority over widely used reasoning strategies, including chain-of-thought prompting, few-shot learning, and retrieval-augmented generation, in reducing prediction error and improving accuracy.[1]

KEYWORDS

Precise Decision Making; Quantitative Reasoning; Large Language Model

[1]Supplementary materials can be found at https://github.com/mo1d2y3/DELL_for_precise_LLMs_in_Health

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LVFL1061

ACM Reference Format:
Danying Mo, Chao Yu*, Xuan Lin, Zhongqi Wu, Yuheng Luo, Chen Bai, and Chaojin Chen*. 2026. DELL: Dual-Knowledge Enhanced LLMs for Precise Decision Making in Healthcare. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/LVFL1061

1 INTRODUCTION

Large language models (LLMs) have exhibited remarkable capabilities across a wide range of tasks, including translation, summarization, text generation, and complex reasoning involving commonsense and logic[1, 6, 8, 21, 25, 29, 30]. Beyond these language-oriented applications, recent works have explored using LLMs to complete tasks requiring fine-grained numeric outputs [3, 36, 45]. For example, in robotic control, LLMs have been guided to generate continuous control signals or low-level actions through carefully designed prompts [36], task-specific instruction tuning [3, 36], or integration with closed-loop feedback mechanisms [45].

Although prior advances have demonstrated the potential of LLMs in fine-grained numeric outputs generation when supported by human-designed knowledge, their capability to complete precise clinical decision tasks, such as generating numerically accurate predictions to determine appropriate drug dosages or therapy adjustments, remains largely unexplored [23, 32]. Such tasks impose higher precision requirements than other clinical applications, such as clinical report generation [37], medical question answering [31, 35], and evidence summarization [5, 22], which have been extensively explored in prior studies on LLMs in healthcare [13, 18, 20, 23, 42]. Equipping LLMs with the capability to make precise clinical decisions would substantially broaden their applicability in healthcare, enabling them to provide more comprehensive and fine-grained decision support, thereby enhancing their potential value in medical practice. However, existing LLMs generally struggle to be competent in some critical clinical tasks that require

*Corresponding authors.

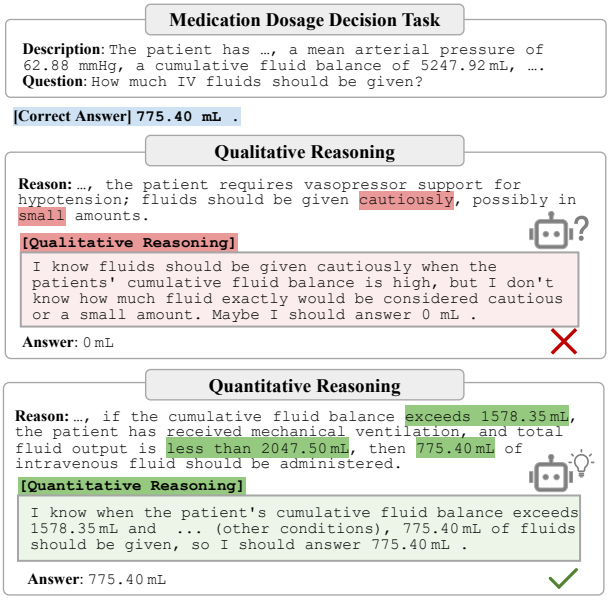

Figure 1: Illustration of a medication dosage problem.

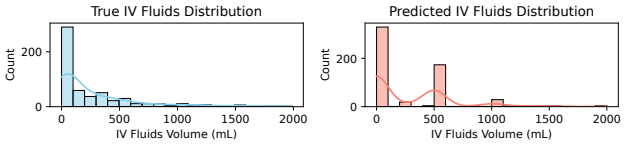

Figure 2: Comparison between distributions of actual administered dosages and predictions by LLMs. LLMs tend to produce a few fixed values (e.g., 0 or 500 mL).

precise numerical predictions. As an illustration, Figure 1 considers a clinical scenario when a precise decision of medication dosage should be made for a sepsis patient. LLMs generally fail to predict the accurate answer of 775.40 *mL* based only on their inherent reasoning capabilities.

The failure of existing LLMs in precise clinical tasks like dosage decision for sepsis can be largely attributed to their lack of quantitative domain knowledge (i.e., the association between the final decisions and their corresponding reasoning evidence) [17, 43]. In particular, normal LLMs, relying on their internal knowledge acquired through large-scale pretraining, can only conduct qualitative reasoning such as "when patients' cumulative fluid is high, fluids administered should be cautious", but without the capabilities of reasoning over quantitative predictions like "how much fluids exactly would be considered cautious" or explicit reasoning evidences like "when the patient's cumulative fluid balance exceeds 1578 *mL* and meets other conditions, 775.40 *mL* of fluids should be administered". The lack of reasoning over quantitative domain knowledge leads LLMs to produce only a few fixed values that frequently appear in their pretraining data, rather than adjusting their output based on a specific inquiry case, which can be supported by the distinct dosage distribution in Figure 2.

In this paper, a novel framework called Dual-knowledge Enhanced LLMs (DELL) is proposed, which enhances the precise decision-making capability of LLMs in healthcare by fusing their internally encoded commonsense knowledge with domain-specific quantitative knowledge automatically mined by external models. DELL comprises three main modules: External Knowledge Distillation, Internal Knowledge Utilization, and Dual-knowledge enhanced LLM Generation. The External Knowledge Distillation module employs multiple explainable methods to automatically distill multi-source quantitative knowledge from clinical data and express it in natural language as external knowledge, while the Internal Knowledge Utilization module uses task prompting to activate domain-related commonsense as internal knowledge. These two knowledge are then jointly processed by the Dual-knowledge enhanced LLM Generation module, where a reasoning paradigm is designed to resolve potential conflicts and synthesize all available evidences into a consensus to support precise decision making.

In summary, our contributions are threefold: (i) We identify and formalize the limitations of LLMs in precise clinical decision-making tasks when quantitative knowledge is unavailable; (ii) We propose DELL, a novel framework that automatically distills quantitative knowledge using explainable methods and leverages a designed reasoning paradigm to make final precise decisions; and (iii) We demonstrate the effectiveness of DELL on a clinical task about sepsis treatment and showcase its consistent superiority over other commonly used reasoning strategies like chain-of-thought, few-shot learning and retrieval-augmented generation.

2 RELATED WORK

2.1 LLMs with Precise Decision Making

LLMs for precise decision-making aim to produce accurate outputs in decision tasks that require precise numeric predictions. Such research has been primarily explored in the field of robotic control. Prompt2walk [36] utilizes LLMs as low-level feedback controllers, generating continuous control signals through few-shot prompts to output precise actions in high-dimensional robotic systems, thus verifying the feasibility of LLMs in outputting precise control actions in robotic control systems. InCoRo [45] integrates LLMs controllers with closed-loop feedback mechanisms and proposes a robotic control system incorporating contextual learning with classical feedback control loops, achieving precise output of low-level control signals in complex tasks. GenChip [3] proposes a framework based on LLM-generated robotic strategy code, via designing a compliant action space, achieving precise control output in high-precision contact operation tasks, thus providing a technical pathway for LLM-driven precise operations. However, these approaches depend heavily on explicitly crafted quantitative priors, such as task-specific prompts, calibrated physical parameters, or control spaces. Similar limitations are also observed in mathematical reasoning tasks that require precise numerical computation. Although LLMs can perform step-by-step symbolic reasoning in arithmetic operations, their performance heavily depends on internal numerical patterns or prompt designs [24, 40]. Consequently, the precision largely stems from these human-specified quantitative rules, limiting generalizability to domains where such well-defined quantitative knowledge is unavailable or costly to obtain.

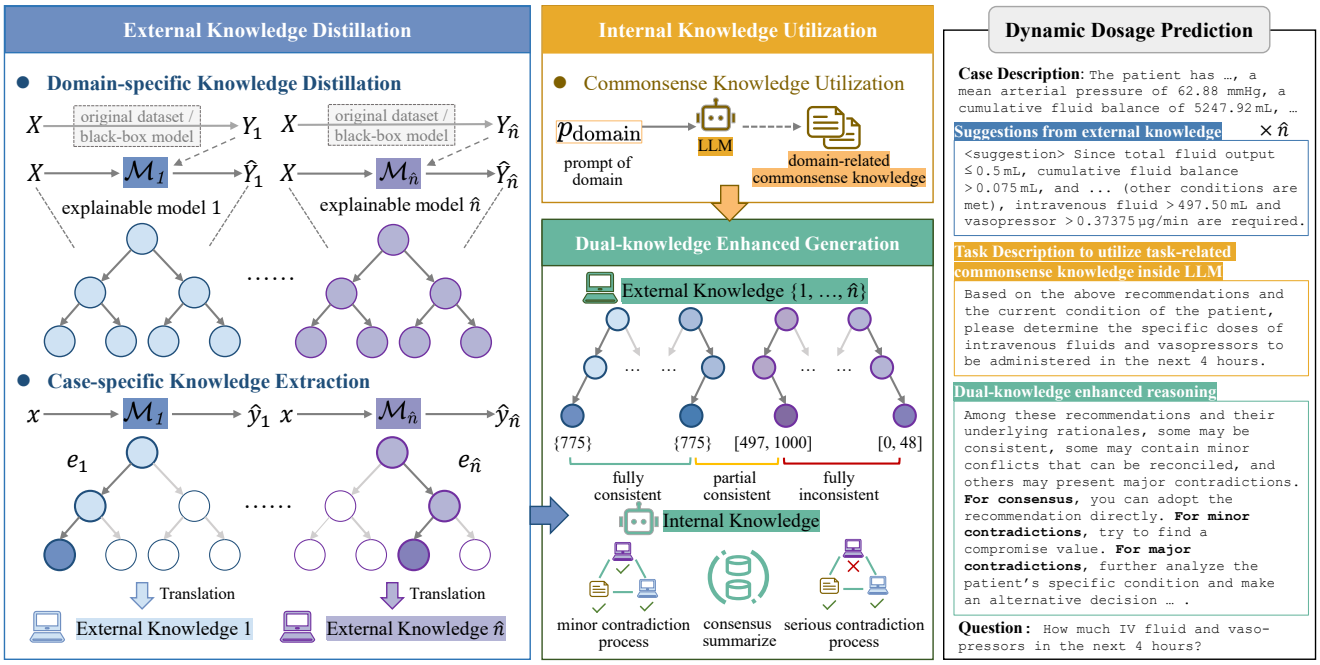

Figure 3: Illustration of the proposed DELL framework. The External Knowledge Distillation module (Left) deploys multiple explainable methods to distill domain-specific quantitative knowledge, extracts case-specific knowledge, and translates them into natural language as external knowledge. The Internal Knowledge Utilization module (Top-Center) activates task-relevant commonsense knowledge as internal knowledge. The Dual-knowledge enhanced LLM Generation module (Bottom-Center) leverages both types of knowledge, resolves contradictions, and makes precise numerical decisions. The right figure illustrates the application of DELL components within the prompt, with each color indicating a corresponding module.

2.2 LLMs with Rule-Based Reasoning

Rule-based methods aim to guide the decision-making process of models via a series of predefined rules, thereby generating more accurate and reliable outputs. Recently, a series of works have integrated rule-based methods to enhance LLMs. A rule-following fine-tuning method (RFFT) [10] is proposed to enhance the performance of LLMs by providing rules and instructing the model to recite and follow, demonstrating that explicit rule teaching facilitates models in learning rule-based reasoning and improving length generalization capability in mathematical tasks. RNR [33] utilizes rule-included datasets for LLM fine-tuning, achieving strong performance in both instruction following and standard natural language processing tasks. This indicates that rule knowledge can complement the fundamental capabilities of LLMs, validating the practical value of rule-based methods in deployment across diverse real-world scenarios. Moreover, other efforts [9, 12, 39, 41, 44] have further validated that rule-based methods can effectively enhance the capability of LLMs to make reliable decisions. Rules adopted in these works are typically qualitative, describing procedural operations, role requirements, or task-related behavioural constraints that guide model behaviour at a programmatic or linguistic level. Some tasks like the airline baggage fee calculation in [44] have human-defined quantitative rules, but domains such as clinical dosage decision (e.g., intravenous fluids or vasopressor adjustment) lack readily accessible, human-defined quantitative rules. This gap highlights the challenge of extending rule-based reasoning from tasks that rely on rules expressed in programs or natural language descriptions to domains where quantitative knowledge and reasoning are required but explicit numerical rules are unavailable.

3 METHODOLOGY

The goal of DELL is to produce precise numeric outputs $\overline{y} \in \mathbb{R}^d$ from input features $x \in \mathbb{R}^m$, where d is the number of target values to be predicted and m is the dimension of the input. As depicted in Figure 3, the framework of DELL consists of three main modules: *External Knowledge Distillation*, *Internal Knowledge Utilization* and *Dual-knowledge enhanced LLM Generation*. The *External Knowledge Distillation* module, which comprises a domain-specific knowledge distillation and a case-specific knowledge extraction component, is responsible for deriving quantitative domain knowledge from original task datasets. In particular, domain-specific knowledge distillation deploys explainable models to distill general domain knowledge from task datasets, including both the predicted outputs and the corresponding reasoning evidence. This general domain knowledge is then utilized by the case-specific knowledge extraction component to derive the reasoning evidence for a given inquiry case, which serves as external domain knowledge for downstream decisions. The *Internal Knowledge Utilization* module then uses task prompting to conduct task-relevant commonsense reasoning

based on the internal knowledge inherently encoded in the LLM. Finally, the *Dual-knowledge enhanced LLM Generation* module guides the reasoning process of LLMs to resolve potential conflicts and synthesize both internal and external knowledge into a coherent consensus, resulting in a precise decision. A concrete example in the right panel of Figure 3 further exemplifies how DELL synergizes the above three modules to enable precise decision making in a clinical dosage prediction problem.

3.1 External Knowledge Distillation

The goal of this module is to distill quantitative domain knowledge from external task-related dataset that can be synthesized into the inherent reasoning process of LLMs. To this end, explainable models are used to learn the association of the predicted outputs and the corresponding reasoning evidence encoded in the external task-related dataset, so as to produce interpretable predictions that LLMs can understand and utilize through in-context learning. Moreover, considering that only one single explainable model could possibly cause biased perspective, it is more reasonable to apply multiple explainable models to capture diverse complementary knowledge, especially for tasks involving predictions of multiple numeric targets ($d \geq 2$).

3.1.1 Domain-specific Knowledge Distillation. Given a dataset $\mathcal{D} = (X, Y)$, where X denotes the input samples and $Y = \{Y_i\}_{i=1}^{n}$ denotes the target values for n prediction targets, $\mathcal{D}_i = (X, Y_i)$ denotes specific dataset associated with the i-th prediction target. Each Y_i can either be the ground truth labels from the original dataset or the predictions generated by an intermediate black-box model. This step distills a set of explainable models $\{\mathcal{M}_i\}_{i=1}^{\hat{n}}$ that map the input samples X to predictions $\hat{Y}_i$ with the reasoning evidence E_i representing the explanations of quantitative associations between X and $\hat{Y}_i$. These explainable models are optimized as follows:

$$\mathcal{M}_i = \arg\min_{\mathcal{M}} \ \mathbb{E}_{(x,y_i)\sim\mathcal{D}_i} \left[\mathcal{L}(\mathcal{M}(x), y_i)\right], \tag{1}$$

where $\mathcal{L}$ represents the objective function used to fit each explainable model. The number of trained explainable models $\hat{n}$ is a hyperparameter that can be determined by two factors: (i) the number of prediction targets n, and (ii) the number of explainable methods applied on each target. Once this stage has been completed, the distilled general domain knowledge can be utilized by the case-specific knowledge extraction component to derive the reasoning evidence for a given inquiry case.

3.1.2 Case-specific Knowledge Extraction. Given a new case input x, this step extracts the case-specific prediction $\hat{y}_i$ and its associated explanation e_i from $\mathcal{M}_i$, which are translated into a form (e.g., natural language) that LLMs can understand and utilize through in-context learning. The predictions and the corresponding reasoning evidence are then stored in an external knowledge set $\overline{K}_{\text{ex}}$:

$$\overline{K}_{\text{ex}} = \{K_{\text{ex}i}\}_{i=1}^{\hat{n}}, \text{ where } K_{\text{ex}i} = Trans(\hat{y}_i, e_i). \tag{2}$$

An example of the extracted external knowledge for a specific case is shown on the right side of Figure 3.

3.2 Internal Knowledge Utilization

In this module, the prompt p_{domain} that provides domain information about the inquiry case and the specific case information $Con(x)$, where Con means translating input into a natural language format, together activate commonsense knowledge $\widetilde{K}_{\text{in}}$ encoded in LLMs:

$$\widetilde{K}_{\text{in}} = \text{LLM}(Con(x), p_{\text{domain}}), \text{ where } \widetilde{K}_{\text{in}} \in \Sigma^*, \tag{3}$$

where the notation Σ^* denotes the space of characters. An illustrative example of p_{domain} is shown on the right side of Figure 3.

This internal knowledge typically comprises commonsense about the task or domain described in p_{domain}, but is insufficient for quantitativeness if LLMs are not inherently encoded with quantitative knowledge about the task. For example, when presented with a sepsis patient who has a high cumulative fluid balance, the LLM may understand the definition of "sepsis" and the qualitative relationship between high fluid balance and the requirement for reduced intravenous fluids (patients with high cumulative fluid balance should be administered less intravenous fluid because excessive fluid accumulation can lead to tissue edema, impaired organ function, and increased risk of complications like heart failure), yet is unable to perform precise calculations due to lack of explicit quantitative reasoning evidences.

3.3 Dual-knowledge Enhanced LLM Generation

This module proceeds to generate outputs for the precise decision tasks, using commonsense knowledge activated by p_{domain} [11, 27], the input case $Con(x)$ and the external knowledge set $\overline{K}_{\text{ex}}$:

$$\begin{aligned} a &= \text{LLM}(Con(x), p_{\text{domain}}, \overline{K}_{\text{ex}}), \\ \overline{y} &= \text{RE}(a), \text{ where } \overline{y} \in \mathbb{R}^d, \end{aligned} \tag{4}$$

where $a \in \Sigma^*$ represents the generated answer in natural language format, $\overline{y}$ is the final numerical output that can be extracted by regular expressions RE.

Given that the external knowledge set $\overline{K}_{\text{ex}}$ is generated automatically by explainable models, conflicts may arise between different elements $K_{\text{ex}i}$. To address this, a reasoning paradigm is designed and implemented via prompts, enabling LLMs to classify the external knowledge into three categories based on their degree of divergence, apply processing strategies corresponding to each category to mitigate inconsistencies among knowledge statements, and finally derive a consensus output to support precise decisions. Specifically, the descriptions and corresponding processing strategies of the aforementioned categories, namely *Consistency*, *Minor Contradiction*, and *Major Contradiction*, are as follows:

- **Consistency**: Consistent information found across multiple knowledge statements is aggregated in this category, and no contradictions need to be processed.
- **Minor Contradiction**: This category identifies minor discrepancies among knowledge statements that are not identical but are potentially compatible. For example, as illustrated in Figure 3, one knowledge source might provide an exact value, while another provides an interval containing that value. In such cases, the LLM is expected to use its internal knowledge to reconcile the information and produce a compatible decision.
- **Major Contradiction**: This category identifies significant, incompatible conflicts among knowledge statements. As shown in

Table 1: Performance comparison across LLMs of different parameter scales and baseline strategies. For each setting, experiments are repeated three times, and both the average and standard deviation are reported. The RMSE values for IV fluids and VP differ significantly in magnitude, which is expected, as the dosage scale of IV fluids is inherently larger than that of VP. Scientific notation is used for all values exceeding 10,000. For each model, the best-performing reasoning strategy is highlighted in bold, with our framework, DELL, consistently achieving the best results.

LLM	Setting	$\text{RMSE}_{\text{IV}} \downarrow$	$\text{RMSE}_{\text{VP}} \downarrow$	$\text{ACC}_{\text{IV}} \uparrow$	$\text{ACC}_{\text{VP}} \uparrow$	$\text{ACC}_{\text{TOTAL}} \uparrow$
R1-671B	Standard	879.05±48.46	20.12±14.37	0.18±0.01	0.57±0.00	0.09±0.01
	CoT [38]	843.93±10.89	117.42±183.20	0.16±0.01	0.59±0.00	0.10±0.01
	Few-shot [2]	934.61±313.58	1.68±0.48	0.15±0.00	0.46±0.00	0.09±0.01
	RAG-G [16]	764.17±31.45	92.55±110.83	0.16±0.01	0.60±0.00	0.11±0.00
	RAG-Q [16]	732.38±0.00	0.53±0.14	0.16±0.00	0.69±0.00	0.15±0.00
	DELL (ours)	**626.70±7.35**	**0.49±0.10**	**0.35±0.01**	**0.71±0.01**	**0.26±0.01**
R1-32B	Standard	1485.71±318.51	261.63±271.44	0.22±0.01	0.41±0.03	0.09±0.01
	CoT [38]	1188.75±34.39	16.37±5.52	0.23±0.01	0.40±0.02	0.08±0.01
	Few-shot [2]	702.16±15.38	0.92±0.13	0.19±0.01	0.17±0.02	0.03±0.00
	RAG-G [16]	1323.64±175.19	63.94±11.73	0.21±0.01	0.55±0.01	0.09±0.01
	RAG-Q [16]	755.51±36.69	0.28±0.03	0.16±0.00	0.68±0.00	0.15±0.00
	DELL (ours)	**649.62±7.62**	**0.22±0.00**	**0.30±0.01**	**0.73±0.01**	**0.23±0.00**
R1-7B	Standard	6.15e4±4.6e4	1.37e13±2.37e13	0.22±0.01	0.41±0.03	0.10±0.00
	CoT [38]	1.40e5±1.82e5	7.02e7±1.17e8	0.22±0.01	0.43±0.01	0.09±0.01
	Few-shot [2]	2743.91±251.60	19.32±26.47	0.17±0.00	0.24±0.02	0.04±0.01
	RAG-G [16]	8383.48±2803.79	2.32e4±1.98e4	0.20±0.01	0.53±0.00	0.10±0.01
	CoTD [34]	2.69e5±1.78e5	1.77e6±2.86e6	0.22±0.01	0.40±0.00	0.09±0.01
	DELL (ours)	**1538.77±409.36**	**9.17±6.35**	**0.28±0.00**	**0.67±0.01**	**0.19±0.00**
Multi-source Voting		664.95	2.96	0.33	0.60	0.18

Figure 3, one source might suggest a range of high values while another suggests a non-overlapping range of low values. In this situation, the LLM is expected to use its general knowledge to assess which source is more consistent with common sense and typical circumstances, thereby deciding which reasoning path to follow and which should be dropped. When appropriate, the LLM is allowed to propose a new value beyond the given suggestions, grounded in its internal knowledge and the patient context.

By identifying the three types of divergence, the LLM leverages internal knowledge $\tilde{K}_{\text{in}}$ to assess and prioritize external knowledge, resolve potential conflicts and produce a consensus output. This reasoning process enables a more comprehensive integration between internal and external knowledge sources, allowing the LLM to fully utilize the available knowledge for precise decision making.

4 EXPERIMENTS

To evaluate the precise decision making capability of DELL in healthcare, we conduct a series of experiments on a critical dosage decision task, i.e, the individualized IV fluids and vasopressor (VP) dosages in the sepsis, which is the third leading cause of death worldwide and the main cause of mortality in hospitals [28].

4.1 Experimental Settings

4.1.1 Dataset. The dataset used for the sepsis treatment task is derived from MIMIC-III, an open-access database across six ICUs at a Boston teaching hospital [14]. Data extraction and processing follow the protocol of [15, 26], where each patient's health information is described by demographics, vital signs, laboratory values, and administered IV fluids and VP. Patient data are represented as multidimensional discrete time series with a 4-hour time step, where the total volume of IV fluids and maximum dose of VP administered are calculated to represent the treatments given within that period. Subsequently, both IV fluids and VP are discretized into 5 dosage ranges according to their quantiles. Combining each pair of IV fluids and VP dosage ranges further yields a discrete action space of 25 possible treatment decisions. Further details of dataset processing, the dataset split, and test data selection are provided in Appendix A.

4.1.2 Metrics. In this study, the precision of predicted IV fluid and VP dosages is evaluated using Root Mean Squared Error (RMSE), while prediction correctness is assessed using Accuracy (ACC). Specifically, five metrics are reported: RMSE_{IV}, RMSE_{VP}, ACC_{IV}, ACC_{VP} and $\text{ACC}_{\text{TOTAL}}$, which respectively measure the numerical precision of IV and VP dosage predictions, categorical correctness of IV and VP dosage ranges, and the accuracy of joint IV-VP dosage decisions within a 25-class space. The joint use of RMSE and ACC enables a comprehensive evaluation of LLM performance from complementary perspectives, with RMSE measuring the deviation between predicted dosages and the ground truth, and ACC evaluating whether predictions fall within the predefined dosage ranges.

4.1.3 Implementation Details. Based on the original dataset described above, five labeled datasets are first constructed by assigning

Table 2: Performance comparison across LLMs of different parameter scales and external knowledge combinations. For each setting, experiments are repeated three times, and both the average and standard deviation are reported. Scientific notation is used for all values exceeding 10,000. For each model, the best-performing reasoning strategy is highlighted in bold, and the second-best is shown in *italics*. Combination of all available knowledge K_{ALL} consistently achieves the best results.

LLM	Setting	$\text{RMSE}_{\text{IV}} \downarrow$	$\text{RMSE}_{\text{VP}} \downarrow$	$\text{ACC}_{\text{IV}} \uparrow$	$\text{ACC}_{\text{VP}} \uparrow$	$\text{ACC}_{\text{TOTAL}} \uparrow$
R1-671B	$K_{\text{R-IV}}$ + $K_{\text{R-VP}}$	*579.85±9.08*	3.47±2.82	0.32±0.03	0.39±0.21	0.12±0.08
	$K_{\text{C-IV}}$ + $K_{\text{R-IV}}$	**573.40±5.04**	*2.09±1.61*	*0.34±0.01*	*0.61±0.02*	*0.20±0.01*
	$K_{\text{C-VP}}$ + $K_{\text{R-VP}}$	734.06±142.37	3.68±3.05	0.20±0.13	0.55±0.07	0.11±0.09
	$\boldsymbol{K}_{\text{ALL}}$	626.70±7.35	**0.49±0.10**	**0.35±0.01**	**0.71±0.01**	**0.26±0.01**
R1-32B	$K_{\text{R-IV}}$ + $K_{\text{R-VP}}$	**613.34±12.29**	*0.98±0.18*	*0.27±0.01*	0.13±0.00	0.02±0.00
	$K_{\text{C-IV}}$ + $K_{\text{R-IV}}$	*616.13±10.96*	44.67±8.24	0.27±0.02	*0.55±0.02*	*0.16±0.01*
	$K_{\text{C-VP}}$ + $K_{\text{R-VP}}$	1297.68±182.46	1.81±0.48	0.22±0.01	0.50±0.01	0.10±0.00
	$\boldsymbol{K}_{\text{ALL}}$	649.62±7.62	**0.22±0.00**	**0.30±0.01**	**0.73±0.01**	**0.23±0.00**
R1-7B	$K_{\text{R-IV}}$ + $K_{\text{R-VP}}$	*2521.38±2742.47*	*31.55±30.88*	**0.29±0.01**	0.21±0.02	0.05±0.00
	$K_{\text{C-IV}}$ + $K_{\text{R-IV}}$	1.41e8±2.44e8	4.84e5±6.56e5	0.28±0.01	*0.47±0.01*	*0.12±0.00*
	$K_{\text{C-VP}}$ + $K_{\text{R-VP}}$	1.22e5±1.81e5	434.13±678.73	0.20±0.01	0.36±0.01	0.08±0.00
	$\boldsymbol{K}_{\text{ALL}}$	**1538.77±409.36**	**9.17±6.35**	*0.28±0.00*	**0.67±0.01**	**0.19±0.00**

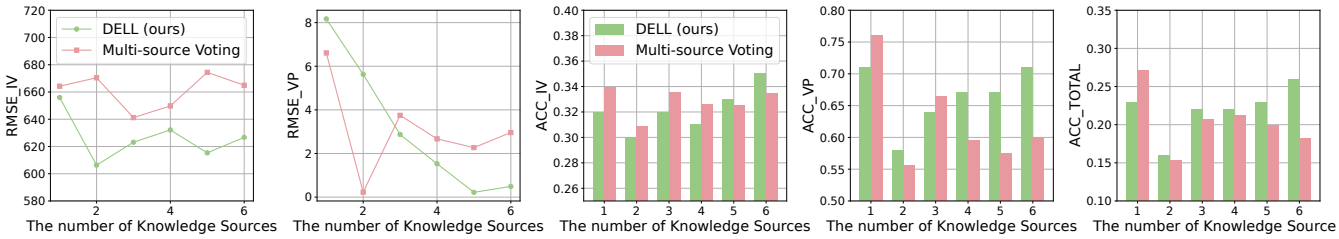

Figure 4: Performance comparison between DELL (perform on R1-671B) and Multi-source Voting across different numbers of quantitative knowledge sources. DELL shows a consistent advantage when at least four distinct quantitative knowledge sources are available.

different prediction targets to the same set of feature inputs. The targets include: (1) 25 discrete dosage classes formed by combinations of IV and VP ranges, (2) 5 discrete IV dosage classes, (3) 5 discrete VP dosage classes, (4) continuous IV values, and (5) continuous VP values. To distill external domain knowledge, explainable models including classification-type decision trees, regression-type decision trees, and a local rule-based explanation method (LORE) [7] are employed. These models are chosen due to their ability to provide clear decision paths that explain how predictions are derived from input features, and their ease of translation into human-readable forms. Among them, LORE is used to explain the 25 discrete dosage classes predicted by a black-box model, showing the usage of quantitative knowledge distilled from black-box models. Details of the black-box model used are shown in Appendix B. Based on the defined targets, 6 explainable models are trained: 4 for classification and 2 for regression, and are denoted as $\mathcal{M}_{\text{C-joint-D}}$ (decision tree for joint dosage classification), $\mathcal{M}_{\text{C-joint-L}}$ (LORE for joint dosage classification), $\mathcal{M}_{\text{C-IV}}$ (decision tree for IV dosage classification), $\mathcal{M}_{\text{C-VP}}$ (decision tree for VP dosage classification), $\mathcal{M}_{\text{R-IV}}$ (decision tree for IV dosage regression) and $\mathcal{M}_{\text{R-VP}}$ (decision tree for VP dosage regression), respectively. Detailed description and performance of these models can be seen in Appendix B.

4.1.4 Baselines. The proposed DELL framework is compared with several widely adopted strategies commonly used to enhance LLM performance in reasoning and decision-making tasks: Standard Prompting (**Standard**), Chain-of-Thought Prompting (**CoT**) [38], Few-shot Prompting (**Few-shot**) [2], Retrieval-Augmented Generation (**RAG**) [16] and Chain-of-Thought Distillation (**CoTD**) [4, 19, 34]. For RAG-based methods, they can be further distinguished according to the type of external knowledge used for retrieval: **RAG-G** retrieves relevant ICU guidelines[2], while **RAG-Q** retrieves qualitative knowledge distilled from explanation models and translated into natural language. Besides, to assess whether the performance gains of DELL arise simply from the incorporation of external knowledge, Multi-source External Knowledge Voting **(Multi-source Voting)**, which predicts results directly by aggregating the outputs of multiple explainable models through a voting

[2]The ICU guidelines used in RAG-G are obtained from a public repository (https://www.icuguideline.com) compiling evidence-based clinical guidelines and expert consensuses published by recognized critical care societies and medical associations.

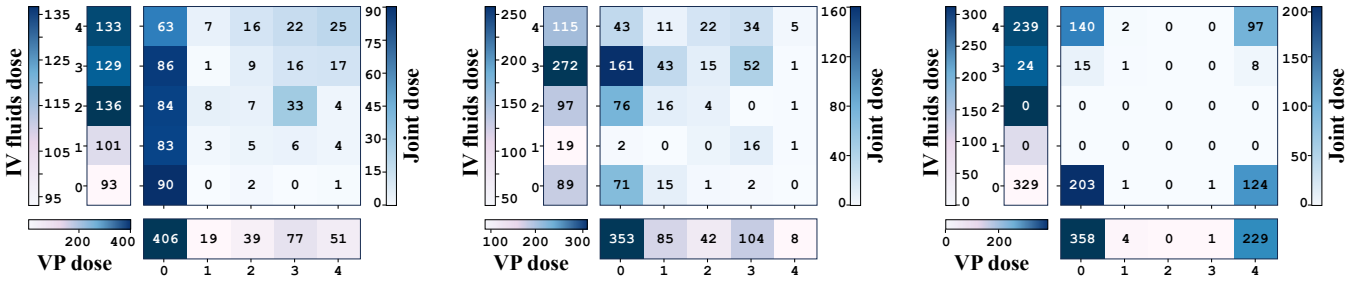

Figure 5: Alignment between the distributions of predicted and actual dosages in a discretized space (R1-671B). The distributions of actual dosages (Left), DELL predictions (Middle), and Standard predictions (Right) are plotted. DELL produces a distribution that more closely resembles the ground truth.

mechanism, without involving LLMs in the decision-making process, is also included. Details about these baselines are provided in Appendix C. For the LLMs used, models from the DeepSeek-R1 family (671B, 32B and 7B) are evaluated across multiple parameter scales, and more experimental results can be seen in Appendix D.

4.2 Results

4.2.1 Effectiveness Evaluation of DELL. Results in Table 1 show the performance of DELL on precise decision-making for sepsis, compared to other reasoning strategies. It is clear that DELL consistently achieves the best performance across all model sizes and evaluation metrics. On R1-671B, DELL achieves an RMSE of 626.70 for IV fluids and 0.49 for VP, corresponding to reductions of 17.99% and 70.83%, respectively, compared to the best-performing baselines that do not incorporate quantitative knowledge (RAG-G for IV fluids and Few-shot for VP). In terms of total accuracy, DELL reaches 26%, which is more than twice the RAG-G baseline performance of 11%. Such improvements are also achieved by R1-32B and R1-7B. Besides DELL, standard prompting consistently performs the worst across all metrics, suggesting internal knowledge alone is insufficient for accurate value prediction.

Among the methods that incorporate external knowledge, RAG-G, RAG-Q, and DELL exhibit distinct ways of leveraging such information for decision making. RAG-G, which retrieves ICU guidelines as relevant external texts, improves upon Few-shot prompting (e.g., in ACC_{VP} under R1-671B) but still lacks sufficient precision compared with DELL, owing to the absence of explicit quantitative reasoning knowledge related to the task. RAG-Q, equipped with retrieval access to quantitative knowledge distilled from explainable models, further enhances performance beyond RAG-G, yet still falls short of DELL due to the inherent limitations of retrieval-based systems in selecting appropriate quantitative knowledge for specific cases. These results confirm the superiority of DELL, which effectively coordinates the internally encoded commonsense knowledge of LLMs with domain-specific quantitative knowledge automatically mined by external models. More results on DELL's performance across demographic subpopulations can be found in Appendix D.

To assess whether the performance gains of DELL arise simply from the incorporation of external domain-specific quantitative knowledge, we compare it with the multi-source external knowledge voting baseline, where predictions are made purely based on the outputs of multiple explainable models, without the inherent reasoning of LLMs. As shown in Table 1, the voting baseline achieves only sub-optimal performance in terms of RMSE_{IV} and ACC_{IV} under R1-671B. In contrast, DELL not only achieves substantially lower RMSEs ($p < 0.05$), but also maintains comparable overall accuracy. The comparison highlights that the LLMs do not merely follow external suggestions but instead perform deeper integration and reasoning over both external and internal knowledge, leading to more accurate and reliable decision making.

4.2.2 Analysis of Contributions of Multiple Knowledge. To explore how different external knowledge contributes to the precise decision making of DELL, an analysis of multiple external knowledge sources is performed and reported in Table 2. Specifically, $K_{\text{R-IV}}$, $K_{\text{R-VP}}$, $K_{\text{C-IV}}$, and $K_{\text{C-VP}}$ denote the quantitative knowledge extracted from $\mathcal{M}_{\text{R-IV}}$, $\mathcal{M}_{\text{R-VP}}$, $\mathcal{M}_{\text{C-IV}}$, and $\mathcal{M}_{\text{C-VP}}$, respectively. K_{ALL} represents the combination of all available quantitative knowledge. Regardless of model sizes, K_{ALL} consistently yields the best overall performance across nearly all metrics. These findings indicate that diverse and complementary knowledge helps LLMs make more accurate and stable decisions, especially in smaller models.

Beyond the advantage of leveraging all available knowledge, we further investigate which external quantitative knowledge is most critical for LLMs. To quantify their effectiveness, we count how many times each knowledge combination setting ranks as the best (bold) or second-best (italic) across the five evaluation metrics. Results across all model scales in Table 2 show that in addition to K_{ALL}, the combination $K_{\text{C-IV}}$ + $K_{\text{R-IV}}$ ranks second overall, followed by $K_{\text{R-IV}}$ + $K_{\text{R-VP}}$. Notably, knowledge combinations involving IV fluids, on which LLMs perform relatively poorly compared to VP, as shown in Table 1 (ACC_{VP} is always higher than ACC_{IV}), tend to yield better performance. This suggests that providing knowledge in areas where LLMs are weak in making a proper prediction is more efficient than reinforcing knowledge they already possess.

4.2.3 Sensitivity Analysis on the Number of Quantitative Knowledge Sources. To examine how the number of provided quantitative knowledge sources affects the LLMs' capability for precise decision making within the DELL framework, we conduct a corresponding sensitivity analysis. The number of provided knowledge sources

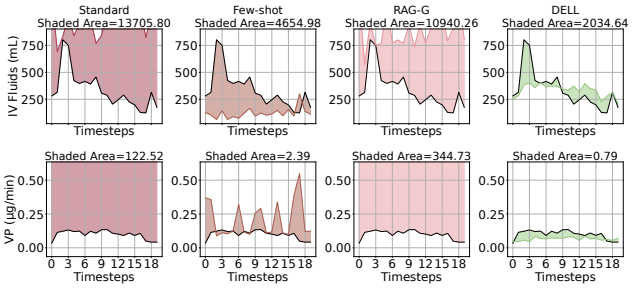

Figure 6: Alignment between predicted and actual dosage trends over time (R1-32B). Predictions at the same time step are aggregated to compute mean values. In each subfigure, black and colored lines represent the actual and predicted dosages, respectively, and highlight the gap between predicted and actual curves. The shaded area represents the total absolute difference between the predicted and true curves, computed as the sum of pointwise deviations over all time steps. Ranges of some curves exceed the figure bounds and are partially truncated. DELL's predictions most closely follow the ground-truth curves.

$\hat{n}$ is varied from 1 to 6 across experimental settings. For each experimental setting, we randomly select the required number of knowledge sources, repeat this random selection three times, and report the averaged results in Figure 4. When the number of quantitative knowledge sources is relatively small (e.g., 1–3), DELL does not consistently outperform the voting mechanism. Specifically, while DELL achieves lower RMSE_{IV} values, it performs worse than voting on the remaining metrics. This can be attributed to the limited external information available to guide the reasoning process. With insufficient quantitative knowledge inputs, DELL relies more heavily on the LLMs' internal knowledge, leading to unstable predictions across different targets. In contrast, the voting mechanism benefits from its variance-reduction property when aggregating knowledge sources, yielding more stable yet less adaptive results. As the number of quantitative knowledge sources increases, the diversity and potential inconsistency among the sources also grow. The voting mechanism, which assumes equal reliability of all sources, becomes less effective as it cannot distinguish between informative and noisy knowledge. Conversely, DELL leverages the LLMs' intrinsic reasoning ability to evaluate the semantic reliability and consistency of each knowledge source, and integrates complementary information while suppressing misleading knowledge. These results indicate that DELL benefits from a sufficiently diverse and informative set of quantitative knowledge sources to fully leverage its reasoning-based integration capability. When provided diverse knowledge for the LLMs to discern reliability and complementarity among them, DELL can consistently outperform the combinations themselves. More results on stress tests, such as varying training sample sizes of external models, can be found in Appendix D.

4.2.4 Visualization of DELL Prediction. To further examine the effectiveness of DELL from fine-grained perspectives, we visualize whether DELL generates predictions that are more consistent with real-world clinical decisions from two perspectives: (1) The alignment between the distributions of predicted and actual dosages in a discretized space, and (2) the alignment between predicted and actual dosage trends over time. For perspective (1), we discretize dosages into 25 bins and observe that DELL produces distributions that more resemble the ground truth, as shown in Figure 5. For perspective (2), Figure 6 compares the temporal dosage curves of Standard, Few-shot, RAG, and DELL with ground truth IV fluids and VP dosages, where the predicted dosage curves produced by DELL most closely follow the ground truth curve across time, as indicated by the smallest shaded area between two curves. These visualization results demonstrate that DELL can not only make precise decisions but also achieve finer-grained alignment with clinical practice.

4.2.5 A Case Study of the Reasoning Process. To better illustrate how DELL integrates internal and external quantitative knowledge within its reasoning process, a representative case is presented and can be seen in Appendix E. DELL first distills external knowledge statements, then identifies consensus, minor contradictions, and major contradictions among them. With the incorporation of internal knowledge, it evaluates these suggestions and decides whether to reject, retain, or synthesize them into a final treatment decision.

5 CONCLUSIONS

In this paper, we propose a novel framework DELL to enhance the precise decision-making capability of LLMs in healthcare. DELL distills multi-source, quantitative knowledge from explainable models and coordinates them with LLMs' internal knowledge under the guidance of a designed reasoning paradigm. We evaluate the effectiveness of DELL in the dosage decision task for sepsis and verify its consistent superiority over other commonly used reasoning strategies. By leveraging these two kinds of knowledge, DELL can reduce the prediction errors and improve overall accuracy. With precise decision-making and quantitative reasoning as fundamental abilities for the future development of intelligent agents, our work equips LLM-based agents with these capabilities by integrating external quantitative knowledge with internal commonsense knowledge. Besides, DELL is designed as a transparent decision-support tool that provides clinicians with quantitative evidence distilled from interpretable models and explanations for reconciling conflicting suggestions, thereby enabling human oversight and override of any recommendation. In the future, we plan to evaluate DELL in other real-life domains (e.g., finance) where quantitative reasoning is critical for precise decision making.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from the Distinguished Young Scholars Project funded by the Natural Science Foundation of Guangdong Province (No. 2025B1515020060), the Special Support Project of Guangdong Province (Grant No.0720240209), the Basic and Applied Basic Research Program of the Guangzhou Science and Technology Plan (No. 2025A04J7141), the Science and Technology Program of Guangzhou City (No.2024A04J4246), and the Xiaomi Young Talents Program.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Kaylee Burns, Ajinkya Jain, Keegan Go, Fei Xia, Michael Stark, Stefan Schaal, and Karol Hausman. 2024. GenCHiP: Generating Robot Policy Code for High-Precision and Contact-Rich Manipulation Tasks. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*. 9596–9603. https://doi.org/10.1109/IROS58592.2024.10801525

[4] Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. 2024. Learning to maximize mutual information for chain-of-thought distillation. *arXiv preprint arXiv:2403.03348* (2024).

[5] Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273* (2023).

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[7] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[9] Yi Hu, Shijia Kang, Haotong Yang, Haotian Xu, and Muhan Zhang. 2025. Training Large Language Models to be Better Rule Followers. *arXiv preprint* (2025), arXiv–2502.

[10] Yi Hu, Xiaojuan Tang, Haotong Yang, and Muhan Zhang. 2024. Case-based or rule-based: How do transformers do the math? *arXiv preprint* (2024).

[11] Shijue Huang, Wanjun Zhong, Deng Cai, Fanqi Wan, Chengyi Wang, Mingxuan Wang, Mu Qiao, and Ruifeng Xu. 2025. Empowering Self-Learning of LLMs: Inner Knowledge Explicitation as a Catalyst. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 24150–24158.

[12] Adam Ishay and Joohyung Lee. 2025. Llm+ al: Bridging large language models and action languages for complex reasoning about actions, Vol. 39. 24212–24220.

[13] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[14] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[15] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. 2018. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine* 24, 11 (2018), 1716–1720.

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[17] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems* 35 (2022), 3843–3857.

[18] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2023), 28541–28564.

[19] Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050* (2023).

[20] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* 15, 6 (2023).

[21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[22] June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461* (2023).

[23] Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. 2025. A generalist medical language model for disease diagnosis assistance. *Nature medicine* 31, 3 (2025), 932–942.

[24] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of transformers with simple arithmetic tasks. *arXiv preprint arXiv:2102.13019* (2021).

[25] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2025-08-02.

[26] Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. 2018. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602* (2018).

[27] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019* (2023).

[28] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. 2016. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama* 315, 8 (2016), 801–810.

[29] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261* (2022).

[30] Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).

[31] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical AI. *Nejm Ai* 1, 3 (2024), AIoa2300138.

[32] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. 2025. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671* (2025).

[33] Kuan Wang, Alexander Bukharin, Haoming Jiang, Qingyu Yin, Zhengyang Wang, Tuo Zhao, Jingbo Shang, Chao Zhang, Bing Yin, Xian Li, et al. 2024. RNR: Teaching large language models to follow roles and rules. *arXiv preprint* (2024).

[34] Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879* (2023).

[35] Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2024. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering* 3, 1 (2024), 133.

[36] Yen-Jen Wang, Bike Zhang, Jianyu Chen, and Koushil Sreenath. 2024. Prompt a Robot to Walk with Large Language Models. In *Proceedings of the Conference on Decision and Control (CDC)*. 1531–1538. https://doi.org/10.1109/CDC56724.2024.10885862

[37] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology* 1, 3 (2023), 100033.

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[39] Wenkai Yang, Yankai Lin, Jie Zhou, and Ji-Rong Wen. 2025. Distilling rule-based knowledge into large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 913–932.

[40] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015* (2023).

[41] Brian Zhang and Zhuo Zhang. 2024. Detecting bugs with substantial monetary consequences by LLM and rule-based reasoning. *Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), 133999–134023.

[42] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* (2023).

[43] Jin Peng Zhou, Charles Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. 2024. Don't Trust: Verify–Grounding LLM Quantitative Reasoning with Autoformalization. *arXiv preprint arXiv:2403.18120* (2024).

[44] Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. 2024. RuleArena: A Benchmark for LLM Rule-Guided Reasoning in Real-World Scenarios. In *Workshop on Reasoning and Planning for Large Language Models*.

[45] Jiaqiang Ye Zhu, Carla Gomez Cano, David Vazquez Bermudez, and Michal Drozdzal. 2024. Incoro: In-context learning for robotics control with feedback loops. *arXiv preprint* (2024).