PretextTrans: Investigating Medical Factual Knowledge Mastery of LLMs with Predicate-text Dual Transformation

Anonymous ACL submission

Abstract

In the study, we aim to investigate current LLMs' mastery of medical factual knowledge with a dynamic evaluation schema, which can automatically generate multiple test samples for each medical factual knowledge point. Test samples produced directly by LLMs always introduce factual errors and lack diversity in the manner of knowledge expression. To overcome the drawbacks, here we propose a novel evaluation method, Predicate-text Dual Transformation (PretextTrans), by introducing predicate transformations into the dynamic evaluation schema. Specifically, each medical knowledge point is firstly transformed into a predicate expression; then, the predicate expression derives a series of variants through predicate transformations; lastly, the produced predicate variants are transformed back into textual expressions, resulting in a series of test samples with both factual reliability and expression diversity. Using the proposed PretextTrans method, we systematically investigate 12 well-known LLMs' mastery of medical factual knowledge based on two medical datasets. The comparison results show that current LLMs still have significant deficiencies in fully mastering medical knowledge, which may illustrate why current LLMs still perform unsatisfactorily in real-world medical scenarios despite having achieved considerable performance on public benchmarks. Our proposed method serves as an effective solution for evaluation of LLMs in medical domain and offers valuable insights for developing medical-specific LLMs.

1 Introduction

Recent years have witnessed the rapid advancement of large language models (LLMs), which have exhibited potential across various domains (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; OpenAI, 2023; Madani et al., 2023; Boiko et al., 2023), including medicine. Solving medical problems requires LLMs to master medical factual knowledge comprehensively and in-depth. Recent studies (Singhal et al., 2023; Nori et al., 2023; Pal and Sankarasubbu, 2024) showed that some LLMs (e.g., GPT-4) encode substantial medical factual knowledge, significantly outperforming previous SOTAs across multiple medical benchmarks (e.g., MedQA (Jin et al., 2021)). However, these LLMs are found to perform unsatisfactorily on real-world medical tasks (Thirunavukarasu et al., 2023; Clusmann et al., 2023; Wornow et al., 2023), falling far short of their benchmark performance. This indicates that current benchmarks do not accurately and comprehensively reflect LLMs' mastery of medical factual knowledge. Therefore, we aim to develop a new evaluation method that more precisely and comprehensively investigates LLMs' mastery of medical factual knowledge.

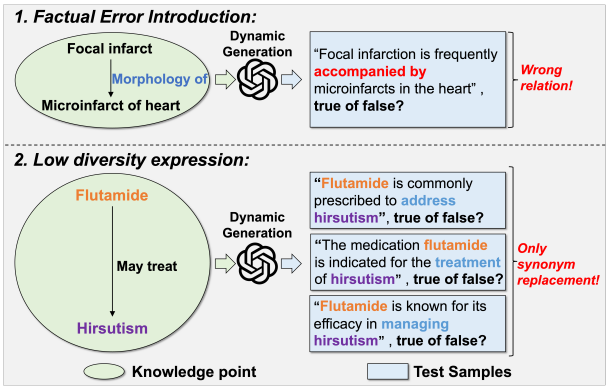

Figure 1: Drawbacks of test samples produced directly by LLMs.

Current evaluations of LLMs' medical knowledge mastery primarily rely on medical benchmarks (Jin et al., 2019, 2021; Pal et al., 2022; Singhal et al., 2023; Sung et al., 2021; Meng et al., 2022), which are reliable but not comprehensive enough for LLM evaluation. Although some newer benchmarks (He et al., 2023; Cai et al., 2024) address this issue by collecting the latest data from diverse sources, constructing these benchmarks can

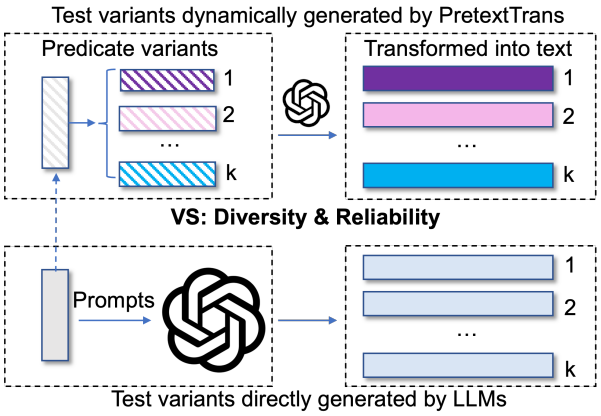

Figure 2: Schema of the proposed Predicate-text Dual Transformation (**PretextTrans**) method (Top) compared with directly generating test variants by LLMs (Bottom).

be costly, and they will face problems such as becoming outdated or leaked to LLMs over time. Recently, several researchers have developed a series of methods (Zhu et al., 2023; Li et al., 2024; Zhu et al., 2024b) to dynamically generate test samples for LLM evaluation, effectively avoiding issues of outdated data and leakage. Therefore, dynamically generating multiple test samples based on each knowledge point in medical knowledge resources is a promising way to comprehensively evaluate LLMs' medical knowledge mastery. A straightforward method is to directly generate test samples using LLMs based on knowledge points. However, this method has two drawbacks as illustrated in Figure 1: (1) **factual error introduction**: factual errors (e.g., incorrect relations) may be introduced during sample generation, affecting the reliability of evaluation; and (2) **low diverse expression**: samples generated from the same knowledge point primarily differ in wording (e.g., synonym replacement) rather than in knowledge expression structure, compromising the diversity of evaluation.

The purpose of this study is to comprehensively investigate LLMs' mastery of medical factual knowledge using a dynamical evaluation method. Because medical factual knowledge primarily involves relationships between medical entities, it can be effectively expressed through predicates. Inspired by this, we propose a **Pre**dicate-**text** Dual **Trans**formation method (**PretextTrans**) that dynamically generates multiple test samples based on the medical knowledge points being evaluated. Figure 2 presents the schema of our method. Specifically, we first express each knowledge point using a predicate expression. Then, we derive a series of structurally diverse variants from this predicate expression through logical implication. Finally, an LLM is employed to transform these variants back to the textual space for generating test samples. The logical implication process ensures the structural diversity of generated test samples and also effectively prevents the introduction of factual errors. Additionally, the LLM-based predicate-to-text transformation ensures that the generated samples are fluent and natural, while also enhancing their syntactic and lexical diversity.

Using the proposed method, we conduct a systematic medical knowledge evaluation of current LLMs based on two medical datasets. Experimental results show that the performance of current LLMs on the multi-sample datasets generated by our method, where each knowledge point is evaluated by multiple samples, is much lower than those on the original single-sample datasets. Furthermore, these LLMs exhibit inconsistency in handling test samples derived from the same knowledge point, failing to achieve the expected performance. These findings indicate that current LLMs have not comprehensively mastered medical factual knowledge, failing to perform satisfactorily in real-world medical scenarios. Our contributions are summarized as follows:

- We introduce a dynamic evaluation method (**PretextTrans**) for comprehensively evaluating LLM medical factual knowledge mastery. Our method generates a series of diverse and reliable test samples for each knowledge point using predicate-text dual transformation.
- Employing the proposed method, we systematically investigate the medical factual knowledge mastery of 12 well-known LLMs.
- Furthermore, we compare LLMs' performance on samples derived from different types of logical implications, shedding light on developing medical foundation models.

2 Related Work

LLM Medical Evaluation Current medical evaluation benchmarks for LLMs can be divided into two categories: (1) QA datasets that evaluate LLMs' comprehensive medical capabilities with questions collected from medical literature (Jin et al., 2019), exams (Jin et al., 2021; Pal et al., 2022), or online websites (Singhal et al., 2023); (2)

datasets for probing LLM medical knowledge mastery (Sung et al., 2021; Meng et al., 2022). These static benchmarks are meticulously created by medical experts and possess high reliability. However, they may face problems such as becoming outdated or leaked to LLMs, affecting the comprehensiveness of evaluation. While constructing new benchmarks can alleviate these problems, they will also become obsolete over time.

Dynamic Evaluation Schema Several studies have proposed dynamic evaluation methods that automatically generate new test samples, effectively avoiding data obsolescence and leakage issues. Some works leverage algorithms to dynamically generate test samples for specific tasks, such as mathematics (Zhu et al., 2024a) and SQL execution (Lei et al., 2023). Others (Zhu et al., 2023, 2024b) generate test samples by paraphrasing existing benchmarks. However, there is currently no related work utilizing dynamic evaluation methods to evaluate LLMs' factual knowledge mastery. To our knowledge, our proposed method is the first to apply the dynamic evaluation schema for evaluating LLMs' mastery of medical factual knowledge.

3 Method

3.1 Evaluation Schema

In this section, we introduce the schema of our PretextTrans method, which generates more diverse and reliable test samples for LLM factual knowledge evaluation. Given a knowledge point P, a straightforward idea is to directly generate a test sample using an LLM:

$$\mathrm{S} = \mathrm{G}_{LLM}(\mathrm{P}) \tag{1}$$

Here, G_{LLM} denotes the LLM generation process, and S refers to the generated test sample. As introduced above, G_{LLM} may create samples that lack diversity and reliability. In contrast, our method first expresses the knowledge point using a predicate expression and then derives a series of variants via logical implication:

$$\mathrm{p} = \mathrm{T}_{text2pre}(\mathrm{P}) \tag{2}$$

$$[\mathrm{q}_1, \mathrm{q}_2, \cdots, \mathrm{q}_K] = \mathrm{T}_{Imp}(\mathrm{p}) \tag{3}$$

Here, $\mathrm{T}_{text2pre}$ denotes a mapping that projects the original knowledge point P into the predicate expression p. T_{Imp} refers to the logical implication, and $\{\mathrm{q}_i\}_{i=1}^{K}$ are the variants derived from the original expression p. The property of logical implication ensures the reliability of these variants, provided that the original expression p is true:

$$(\mathrm{p} = \mathbf{T}) \Rightarrow (\mathrm{q}_i = \mathbf{T}), \quad 1 \leq i \leq K \tag{4}$$

Finally, we convert each predicate variant back to a textual test sample for evaluation:

$$\mathrm{S}_i = \mathrm{T}_{pre2text}(\mathrm{q}_i), \quad 1 \leq i \leq K \tag{5}$$

Here, $\mathrm{T}_{pre2text}$ maps each predicate variant q_i into a corresponding test sample (textual variant). Since these samples are derived from predicate variants with diverse structures, the predicate-text duality ensures they exhibit substantial diversity while maintaining reliability.

Types	Form
Origin	$\mathcal{R}(\mathrm{A}, \mathrm{B})$
Inversion	$\mathcal{R}^{-1}(\mathrm{B}, \mathrm{A})$
Instantiation	$\mathcal{P}(\mathrm{A}, x) \Rightarrow \mathcal{Q}(x, \mathrm{B})$
Double Negation	$\neg(\neg\mathcal{R}(\mathrm{A}, \mathrm{B}))$

Table 1: Three types of logical implication employed in PretextTrans. Here, x is a specific entity (e.g., a patient), and $\mathcal{P}, \mathcal{Q}$ describe the relations between x and A, B (e.g., has a disease, may be treated by a drug).

3.2 Evaluation Framework

Building on the proposed evaluation schema, we develop a novel evaluation framework to evaluate LLMs' mastery of medical factual knowledge comprehensively. Figure 3 presents an overview of this framework.

3.2.1 Predicate Variant Generation

A single knowledge point can be denoted as $\mathrm{P} = (\mathrm{A}, \mathrm{R}, \mathrm{B})$, where A, R, and B refer to the head entity, the relation, and the tail entity, respectively. In predicate logic, such a relation can be effectively presented by:

$$\mathrm{p} = \mathcal{R}(\mathrm{A}, \mathrm{B}) \tag{6}$$

Here, $\mathcal{R}(x, y)$ is a predicate derived from the relation R, representing the statement "x has the relation R with y". p represents its value at the point (A, B). Next, the framework employs three types of logical implications that are widely employed in practical medical applications. Table 1 lists the forms of these implications, including:

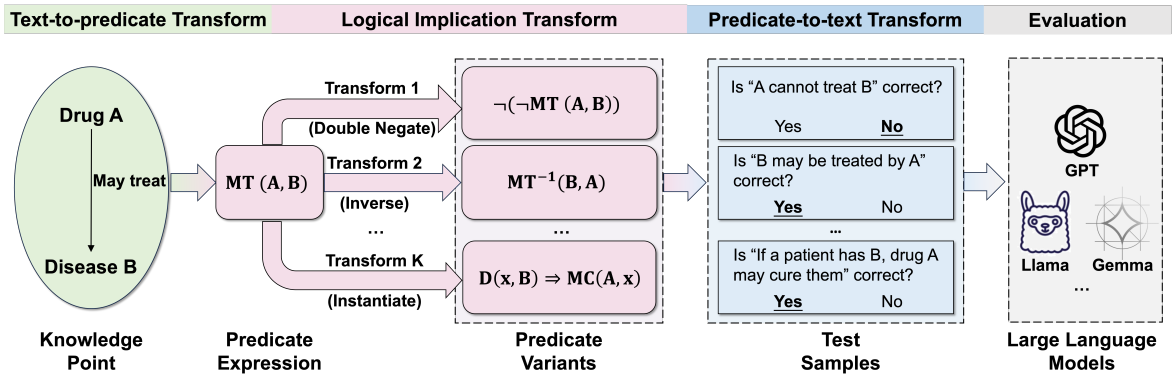

Figure 3: An overview of the proposed framework using PretextTrans for LLM medical factual knowledge evaluation.

- **Inversion**: The inverse expression presents the original expression from another direction. For example, if the statement "Drug A may treat disease B" holds, then "Disease B's prescribed drug includes drug A" also holds.
- **Instantiation**: This type of logical implication applies a general knowledge point to a specific case. For example, the statement "Drug A may treat disease B" can be instantiated as "If a patient has disease B, drug A may cure them." Such transformation is commonly used in disease diagnosis and treatment.
- **Double Negation**: The double negation rule is widely utilized to obtain logically equivalent expressions. In our framework, this rule is applied to construct **negative** expressions. For example, if "Drug A may treat disease B" holds, then "Drug A cannot treat disease B" must be false.

It is noteworthy that these three types of logical implication can be further combined to produce additional expressions based on the transitive property of logical implication. As a result, a total of K variants are generated in this process:

$$\mathrm{q}_i = \mathrm{T}^i_{Imp}(\mathcal{R}(\mathrm{A}, \mathrm{B})), \quad 1 \leq i \leq K \qquad (7)$$

where T^i_{Imp} denotes the i^{th} logical implication.

3.2.2 Textual Sample Generation

A straightforward method to generate test samples from predicate variants is by directly prompting LLMs. However, this method may also introduce factual errors, affecting the reliability of the generated samples. To address this issue, we designed a prototype-based sample generation strategy, as depicted in Figure 4. Specifically, for each predicate variant $\mathrm{T}^i_{Imp}(\mathcal{R}(\mathrm{A}, \mathrm{B}))$, we initially retrieve the

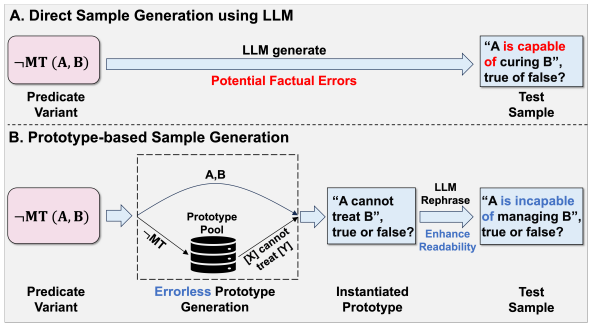

Figure 4: Test sample construction process in the proposed framework. Up: directly generating samples by LLMs may affect the reliability; Down: the proposed prototype-based sample construction strategy.

corresponding prototype from a pre-constructed prototype pool based on the predicate $\mathrm{T}^i_{Imp} \cdot \mathcal{R}$. For predicate variants obtained through double negation, we retrieve prototypes based on their negated form (i.e., single negation form) to generate **negated samples** for LLM evaluation. Subsequently, the prototype is instantiated by the arguments (A, B). The instantiated prototype precisely conveys the predicate variant in the textual space. Finally, the prototype is further rephrased by an LLM to obtain the final test sample S_i. Since current LLMs possess strong language capabilities and rarely make mistakes in sentence rephrasing, the proposed sample generation strategy can ensure the reliability and diversity of the generated samples.

3.2.3 Evaluation Metrics

In our framework, we evaluate LLMs using statement verification tasks, asking them to determine whether a given statement is true or false:

$$\mathrm{Score}(\mathrm{M}, \mathrm{S}_i) = \mathbb{1}(\mathrm{M}(\mathrm{S}_i) = l_i), 1 \leq i \leq K \quad (8)$$

Here, M is the evaluated LLM, S_i is the textual variant (statement) generated by our framework, and $\mathrm{M}(\mathrm{S}_i) \in \{\mathbf{T}, \mathbf{F}\}$ denotes LLM's prediction for S_i. $l_i \in \{\mathbf{T}, \mathbf{F}\}$ is the label of S_i, and the function $\mathbb{1}(\cdot)$ is a characteristic function that equals 1 when the enclosed expression is true, and 0 otherwise.

For a dataset with N knowledge points $\{\mathrm{P}_j\}_{j=1}^{N}$, we initially use the metric *average accuracy* to compute the accuracy across all test samples:

$$a_{\text{avg}} = \frac{1}{N}\frac{1}{K}\sum_{j=1}^{N}\sum_{i=1}^{K}\text{Score}(\mathrm{M}, \mathrm{S}_i^j) \qquad (9)$$

Here, S_i^j denotes the i^{th} test sample derived from the j^{th} knowledge point P_j. While this metric is widely applied in various benchmarks, it cannot evaluate the **consistency** of LLMs in predicting all test samples derived from the same knowledge point, which is crucial for high-risk applications in the medical domain. Therefore, we also utilize another metric, *joint accuracy*, which considers a knowledge point as mastered if **all the related samples** are predicted correctly:

$$a_{\text{joint}} = \frac{1}{N}\sum_{j=1}^{N}\prod_{i=1}^{K}\text{Score}(\mathrm{M}, \mathrm{S}_i^j) \qquad (10)$$

By applying these metrics, we can achieve a comprehensive evaluation of LLMs' mastery of medical factual knowledge.

4 Experiments

4.1 Experiment Setup

Datasets Introduction To investigate the mastery of medical factual knowledge in current LLMs, we applied the proposed framework to two datasets: a biomedical evaluation benchmark MedLAMA (Meng et al., 2022) and a clinical knowledge base DiseK (Zhou et al., 2024). MedLAMA is a large-scale biomedical evaluation benchmark containing 39,053 knowledge triplets across 19 relations, all manually selected from the UMLS Metathesaurus (Bodenreider, 2004) to ensure high quality. DiseK is a clinical knowledge base containing 24,413 triplets, covering 1,000 high-frequency diseases across four crucial relations related to disease diagnosis and treatment. Mastering this disease-related knowledge is essential for LLMs to be applicable in real medical scenarios.

Considering computational costs and dataset size, we select a subset from each dataset for evaluation. Specifically, we randomly select a single entity from the corresponding tail entities for each pair of a head entity and a relation. This approach aims to reduce the evaluation scale while maximizing the diversity of the evaluated knowledge. We also excluded two relations in MedLAMA, which are the inversion of the other two relations in MedLAMA. Furthermore, for each head-relation pair (A, R), we randomly sample a negative entity C that satisfies $\neg\mathcal{R}(\mathrm{A}, \mathrm{C})$ to create a negative triplet $(\mathrm{A}, \mathrm{R}, \mathrm{C})$. Test samples generated from this negative triplet possess a similar structure to those generated from the positive triplet but with opposite labels. By introducing negative triplets, we can further evaluate the ability of LLMs to discern non-knowledge, which is also essential for practical application. Table 2 presents the basic statistics of the sampled datasets. More detailed statistics about these datasets and the relation types involved are provided in Appendix A.

Dataset	MedLAMA	DiseK
Type	Biomedical	Clinical
# Rel Types	17	4
# Triplets	34,000	6,348

Table 2: Statistics of the sampled datasets.

Method Setting To ensure the diversity of evaluation, we combined the three types of logical implication and generated $K = 8$ expressions (variants) for each knowledge point, including the original expression. We crafted a prototype for each combination of relation and logical implication type to generate test samples. Moreover, we utilize Llama3-70B-Instruct (AI@Meta, 2024) to rephrase the instantiated prototypes since it exhibits strong performance on LLM leaderboards. More details of the logical implication process, prototypes, and the prompt format are provided in Appendix B.

For LLM evaluation, we employ the popular 5-shot in-context learning setting (Brown et al., 2020), where five examples are presented before the test sample, guiding LLMs to produce answers in consistent format with the provided examples. We calculate the average and joint accuracies (introduced in Sec 3.2.3) for each LLM. Appendix C provides more details, including the prompt format.

Baselines We initially compare our method with the original datasets. For original datasets, we leverage the templates provided in the benchmarks to generate statements for evaluation. We also im-

Model	MedLAMA			DiseK		
	Origin	LLMEval	PretextTrans	Origin	LLMEval	PretextTrans
Random	50.0	50.0	50.0	50.0	50.0	50.0
ChatGLM3-6B	72.4	$64.1_{\downarrow 8.2}$	$\underline{55.0}_{\downarrow 17.4}$	76.1	$68.5_{\downarrow 7.6}$	$\underline{56.1}_{\downarrow 20.0}$
Llama2-7B	56.4	$58.3_{\uparrow 1.9}$	$\underline{53.1}_{\downarrow 3.4}$	61.7	$\underline{52.7}_{\downarrow 9.0}$	$52.8_{\downarrow 8.9}$
Vicuna-7B	76.4	$68.0_{\downarrow 8.4}$	$\underline{57.5}_{\downarrow 18.9}$	59.9	$60.9_{\uparrow 1.0}$	$\underline{53.9}_{\downarrow 6.0}$
Vicuna-13B	77.0	$69.3_{\downarrow 7.7}$	$\underline{60.7}_{\downarrow 16.3}$	62.5	$57.4_{\downarrow 5.0}$	$\underline{55.7}_{\downarrow 6.7}$
Gemma-7B	73.3	$61.1_{\downarrow 12.2}$	$\underline{59.4}_{\downarrow 13.9}$	59.0	$\underline{54.8}_{\downarrow 4.2}$	$55.0_{\downarrow 4.1}$
Llama3-8B	78.5	$69.1_{\downarrow 9.4}$	$\underline{66.6}_{\downarrow 11.9}$	67.9	$65.3_{\downarrow 2.6}$	$\underline{59.3}_{\downarrow 8.6}$
Llama2-70B	82.0	$69.2_{\downarrow 12.8}$	$\underline{63.8}_{\downarrow 18.2}$	70.5	$67.3_{\downarrow 3.2}$	$\underline{59.0}_{\downarrow 11.5}$
ClinicalCamel-70B	84.8	$73.7_{\downarrow 11.1}$	$\underline{71.9}_{\downarrow 13.0}$	74.5	$70.6_{\downarrow 3.8}$	$\underline{66.1}_{\downarrow 8.4}$
Meditron-70B	79.4	$70.0_{\downarrow 9.4}$	$\underline{64.7}_{\downarrow 14.6}$	71.1	$62.8_{\downarrow 8.3}$	$\underline{60.2}_{\downarrow 10.9}$
Med42-70B	81.8	$\underline{69.3}_{\downarrow 12.5}$	$70.0_{\downarrow 11.8}$	73.3	$69.1_{\downarrow 4.2}$	$\underline{64.8}_{\downarrow 8.5}$
GPT-3.5-turbo	82.1	$76.7_{\downarrow 5.4}$	$\underline{66.2}_{\downarrow 16.0}$	73.5	$67.6_{\downarrow 6.0}$	$\underline{60.3}_{\downarrow 13.3}$
Llama3-70B	**86.6**	$\mathbf{76.9}_{\downarrow 9.7}$	$\underline{\mathbf{76.9}}_{\downarrow 9.7}$	**79.7**	$\mathbf{78.2}_{\downarrow 1.5}$	$\underline{\mathbf{70.9}}_{\downarrow 8.8}$

Table 3: Performance (**average accuracy**) of LLMs on the original datasets (Origin), datasets directly generated by LLM (LLMEval), and datasets generated by our framework (PretextTrans). Bold: Best performance under the same evaluation method; Underline: LLM achieved the lowest performance in this evaluation method.

plemented a dynamic evaluation baseline (named as **LLMEval**) that directly generates test samples from triplets using an LLM. Specifically, we prompt Llama3-70B-Instruct[1] to generate $K = 8$ statements, presenting the given triplet in different ways. We carefully crafted the prompt to ensure maximum diversity in generated samples. Appendix D details the prompt and other settings.

Evaluated LLMs In our study, we evaluate 12 well-known general and medical-specific LLMs: (1) general LLMs: ChatGLM3-6B (Du et al., 2022), Gemma-7B (Team et al., 2024), Llama2 (7B,70B) (Touvron et al., 2023), Llama3 (8B,70B) (AI@Meta, 2024), Vicuna (7B,13B) (Zheng et al., 2023), and GPT-3.5-turbo (Ouyang et al., 2022); (2) medical-specific LLMs: ClinicalCamel-70B (Toma et al., 2023), Meditron-70B (Chen et al., 2023) and Med42-70B (Christophe et al., 2023). We have not evaluate LLMs that are either too expensive (e.g., GPT-4 (OpenAI, 2023)) or not publicly available (e.g., MedPaLM (Singhal et al., 2023)).

4.2 Results

4.2.1 Comparison Study

We first conduct a comparison study across different evaluation methods and LLMs. Table 3 lists LLMs' performance (average accuracy) on the MedLAMA and DiseK datasets evaluated by different methods. The experimental results demonstrate that all evaluated LLMs achieve much lower performance on datasets generated by PretextTrans compared to the original datasets. This suggests that **dynamically generating multiple samples for each knowledge point can significantly enhance the comprehensiveness of evaluation**. Moreover, compared to datasets directly generated by an LLM (LLMEval), almost all LLMs achieve lower performance on datasets created by PretextTrans, with some models (e.g., ChatGLM3-6B and GPT-3.5-turbo) experiencing over 10% degradation. These findings indicate that **PretextTrans is capable of generating test samples that are more comprehensive than those directly generated by LLMs.**

Among all the evaluated LLMs, Llama3-70B outperforms the others across all datasets and evaluation methods, achieving accuracies of 76.9 and 70.9 evaluated by PretextTrans. Llama3-8B also performs best on PretextTrans-generated datasets among LLMs with around 10B parameters, even slightly surpassing the 10x larger Llama2-70B. These results indicate that **Llama3 model series encodes significantly more medical knowledge than other evaluated LLMs**. Additionally, while some medical-specific LLMs (ClinicalCamel, Med42) perform similarly to their backbone model (Llama2-70B) on original datasets, they notably outperform it by around 7% on PretextTrans-generated datasets. This suggests that **training on medical corpora can notably improve the depth of medical knowledge mastery**.

We also study the joint accuracies of LLMs

[1]We choose the same LLM utilized in our framework to make a fair comparison.

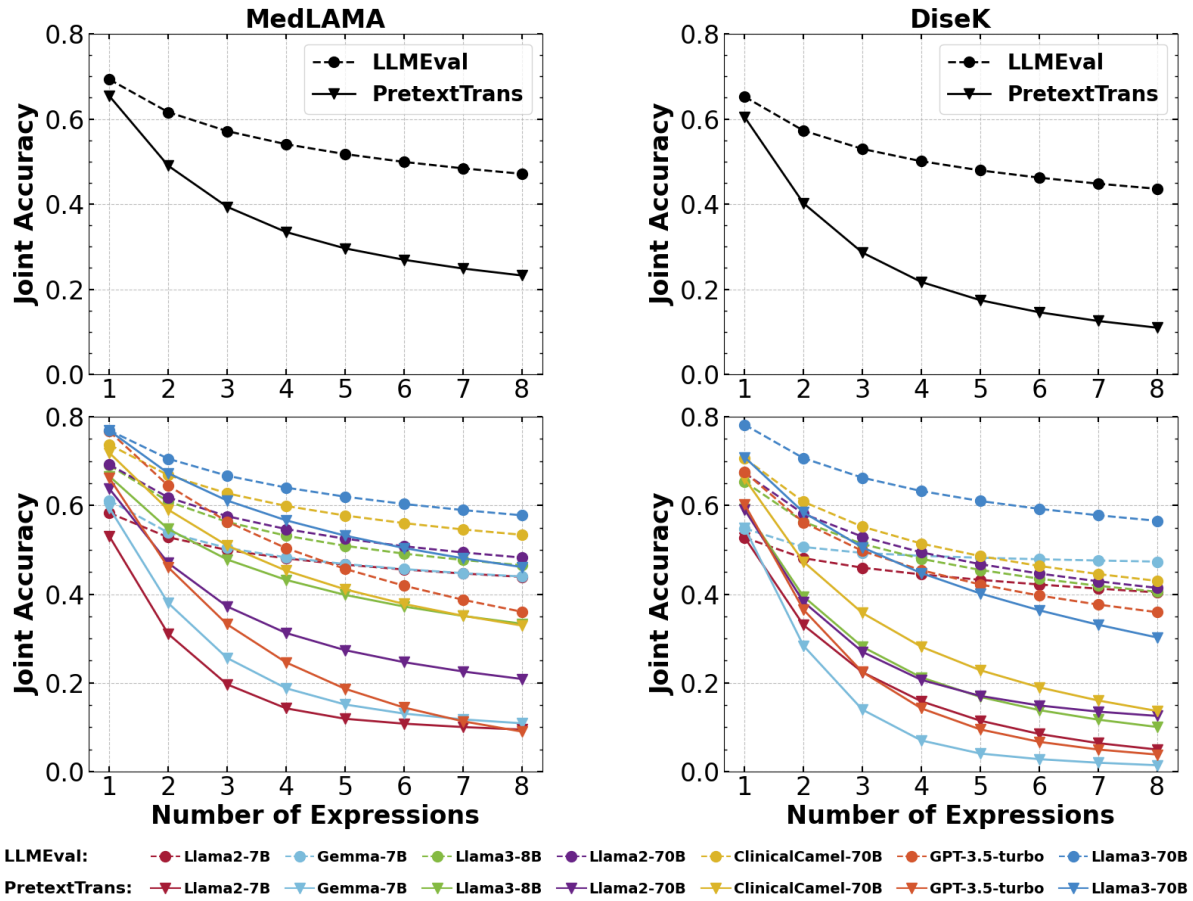

Figure 5: Performance (**joint accuracy**) of 7 LLMs evaluated by increasing the number of expressions per knowledge point. Top: overall performance trend averaged across LLMs; bottom: detailed performance for each LLM.

evaluated by increasing numbers of expressions per knowledge point. The results of seven typical LLMs are illustrated in Figure 5, with the full results provided in Appendix E. **To eliminate the impact of sample addition orders, we enumerate all possible orders and averaged the results.** Therefore, the value at $x = i$ corresponds to the expected joint accuracy evaluated with any i samples. We observe that the results from LLMEval and PretextTrans are quite close when using a single sample for evaluation. However, as the number of test samples increases, the difference between the results from the two methods grows notably larger. This phenomenon indicates that **current LLMs generally exhibit significant lower consistency when confronted with structurally diverse test samples generated by our method** compared to samples directly generated by LLMs. Moreover, as the number of expressions increases, Llama3-70B exhibits a slower decline in performance compared to other LLMs, indicating a more consistent understanding of diverse expression structures from the same knowledge points. Nevertheless, there is still room for improvement in current LLMs' mastery of medical knowledge.

4.2.2 Effectiveness Analysis

Effect of framework components First, we conduct an ablation study to analyze the contribution of each component to our proposed framework. Table 4 presents the ablation results of two typical LLMs, and the full results are listed in Appendix E. Here, we focus on the logical implication (LogImp) and the LLM rephrasing (LMReph) modules that are designed to increase the diversity of test samples. We observe that removing these two modules results in higher evaluation performance, especially when the logical implication module was removed (around 7%). These results indicate that the **logical implication module contributes most to the evaluation diversity in the proposed framework**.

Effect of Implication Types We further conduct a fine-grained analysis of the logical implication types applied in our framework, with results pre-

Datasets	Method	Model	
		ClinicalCamel	Llama3-70B
MedLAMA	PretextTrans	71.9	76.9
	-LogImp	$80.6_{\uparrow 8.8}$	$83.0_{\uparrow 6.1}$
	-LMReph	$72.8_{\uparrow 1.0}$	$80.4_{\uparrow 3.6}$
DiseK	PretextTrans	66.1	70.9
	-LogImp	$73.1_{\uparrow 7.1}$	$77.8_{\uparrow 7.0}$
	-LMReph	$68.0_{\uparrow 1.9}$	$74.0_{\uparrow 3.1}$

Table 4: Ablation results of two typical LLMs for key components of the proposed PretextTrans framework. LogImp: the logical implication module; LMReph: the LLM rephrasing module for generating test samples.

Datasets	ImpType	Model	
		ClinicalCamel	Llama3-70B
MedLAMA	None	80.6	83.0
	+DN	$73.8_{\downarrow 6.9}$	$80.6_{\downarrow 2.4}$
	+DN+Inv	$73.2_{\downarrow 7.4}$	$78.6_{\downarrow 4.3}$
	+All	$71.9_{\downarrow 8.8}$	$76.9_{\downarrow 6.1}$
DiseK	None	73.1	77.8
	+DN	$68.9_{\downarrow 4.2}$	$72.3_{\downarrow 5.6}$
	+DN+Inv	$67.9_{\downarrow 5.2}$	$72.3_{\downarrow 5.5}$
	+All	$66.1_{\downarrow 7.1}$	$70.9_{\downarrow 7.0}$

Table 5: Ablation results of two typical LLMs for different types of logical implication applied in PretextTrans. DN: double negation; Inv: inversion; All: the combination of instantiation, inversion, and double negation.

sented in Table 5. Experimental results show that LLM performance continually declines as more logical implication types are added, indicating their effectiveness. Furthermore, the inclusion of double negation (+DN) leads to a more significant performance degradation (around 5%) than other implication types. This suggests that current LLMs exhibit relatively **less proficiency in understanding negated expressions** compared to instantiated and inverted statements of medical knowledge.

4.2.3 Case Study

We also conduct a case study to examine the effectiveness of the proposed PretextTrans framework. Figure 6 illustrates an example of LLM evaluation by PretextTrans compared with the LLMEval method. The case shows that Llama3-70B correctly answers LLMEval-generated samples that have the same knowledge expression structure. In contrast, the PretextTrans-generated samples possess distinct expression structures, and some of them cannot be correctly answered by Llama3-70B. These findings indicate that the proposed PretextTrans framework **effectively increases the diversity of knowledge expression structures in generated samples**, enabling a more comprehensive evaluation of LLMs' true mastery of medical knowledge.

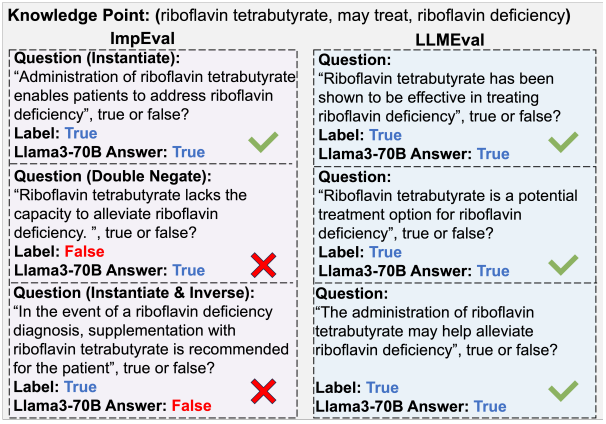

Figure 6: A case of evaluating LLMs using the proposed PretextTrans framework (left) compared with the LLMEval method (right).

5 Conclusion and Discussion

In this paper, we comprehensively investigate LLMs' mastery of medical factual knowledge by designing a dynamical evaluation method named PretextTrans. The proposed method leverages predicate-text dual transformation to dynamically generate multiple test samples for each knowledge point in medical knowledge resources, ensuring their reliability and structural diversity. The experimental results indicate that current LLMs lack comprehensive mastery of medical factual knowledge; thus, they are not yet competent for real-world medical tasks. Furthermore, these LLMs exhibit inconsistency in understanding diverse expressions derived from the same medical knowledge point, thus limiting their practical application in the medical domain. These findings demonstrate that our method can serve as an effective solution to comprehensively evaluate LLMs' medical knowledge mastery. Our study may also shed light on developing medical foundation models. For example, incorporating content that presents the same medical knowledge in diverse ways into the training data may improve LLMs' consistency and comprehensiveness in understanding medical concepts. In the future, we aim to integrate this method with other evaluation forms (e.g., question answering) and medical datasets to conduct a more comprehensive evaluation of LLM medical knowledge mastery.

Limitations

One limitation of our study is that, despite evaluating several well-known general and medical-domain-specific LLMs, we excluded some notable models like GPT-4 and MedPaLM. This was due to either their high costs (it would require $1200 to evaluate GPT-4 on MedLAMA) or their unavailability for public access (e.g., MedPaLM). We plan to evaluate other LLMs in the future if feasible. Additionally, although our evaluation method has the potential to be applied in other domains, it was initially devised and validated for the medical domain. Applying it to other domains may require further validation.

References

AI@Meta. 2024. Llama 3 model card.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

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*, 33:1877–1901.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Avani Gupta, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Marco Pimentel, Tathagata Raha, Ronnie Rajan, and Shadab Khan. 2023. Med42 - a clinical large language model.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. MedEval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8725–8744, Singapore. Association for Computational Linguistics.

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):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. 2023. S3eval: A synthetic, scalable, systematic evaluation suite for large language models. *arXiv preprint arXiv:2310.15147*.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18600–18607.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos Jr, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pages 1–8.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, Dublin, Ireland. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Yuxuan Zhou, Xien Liu, Chen Ning, and Ji Wu. 2024. Multifaceteval: Multifaceted evaluation to probe llms in mastering medical knowledge.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2024a. Dyval: Dynamic evaluation of large language models for reasoning tasks. In *The Twelfth International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024b. Dyval 2: Dynamic evaluation of large language models by meta probing agents. *arXiv preprint arXiv:2402.14865*.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yunze Song, Yumeng Zhang, Hanxu Hu, Yiran Wei, Rui Wang, and Hongyuan Lu. 2023. Clean-eval: Clean evaluation on contaminated large language models. *arXiv preprint arXiv:2311.09154*.

A Details of Datasets

We validate the proposed framework on two datasets: a biomedical evaluation benchmark, MedLAMA, and a disease-centric clinical knowledge base, DiseK. Given the large scale of these datasets, we sample a subset of knowledge points from each by selecting a single tail entity for each 1-to-N relation. Additionally, we sample negative triplets to increase the evaluation difficulty. Table 8 and 9 list the relation types involved in the sampled datasets. The sampled MedLAMA dataset includes 1,000 positive triplets and 1,000 negative triplets for each relation, while the detailed statistics for DiseK are presented in Table 6.

Relation Type	# Positive	# Negative
#Symptoms	987	987
#Affected Sites	745	745
#Therapeutic Drugs	836	836
#Surgical Procedures	599	599

Table 6: Statistics of the sampled DiseK dataset. # Positive: the number of positive triplets extracted from DiseK. # Negative: the number of negative triplets sampled from DiseK.

B Details of Method Setting

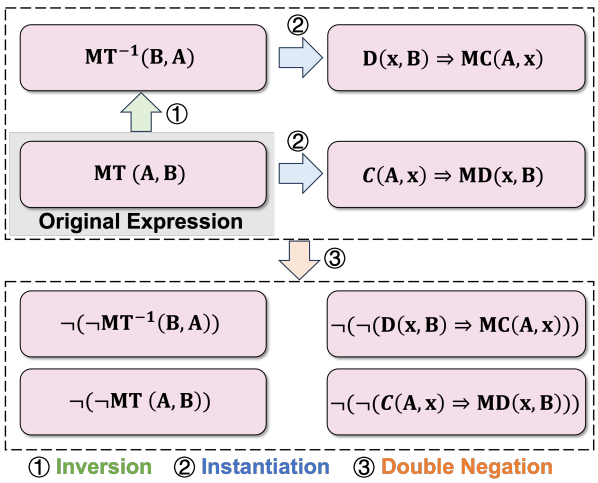

Figure 7: An example of the logical implication procedure implemented in this study.

Details of Logical Implication An example of the logical implication procedure applied in this study is illustrated in Figure 7. First, the inversion operation is applied to the original expression to create a new expression. Subsequently, these two expressions are instantiated into two additional expressions. Finally, double negation is used to generate four more expressions.

Categories	Keywords
True	True, Entailed, Correct, Yes
False	False, Contradicted, Wrong, No

Table 7: The keywords we utilize to extract answers from LLMs' responses.

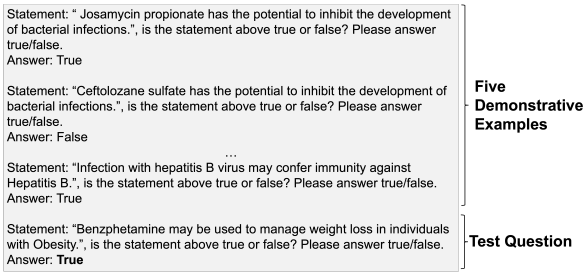

Figure 8: An example of the five-shot in-context learning process applied in our evaluation.

Details of Prototypes-based Generation As introduced before, we designed a prototype-based sample generation strategy to ensure the reliability of the generated samples and crafted a prototype for each combination of relation type and logical implication type by discussing with clinicians. We list all the crafted prototypes in Table 10, 11, and 12 for reproducing our experiments.

For LLM rephrasing, we prompt the Llama3-70B-Instruct model with the following instruction: "*Please paraphrase the following statement to present the same concept in a different way. DO NOT change the basic sentence structure. Directly output the paraphrased statement without other text. Statement: [prototype]*". In our experiments, we found that statements rephrased using this method effectively preserve the original meaning of the prototypes.

C Details of Evaluation Setting

In our implementation, we form test samples based on the following format: "*[Statement], is the statement above true or false? Please answer True or False.*" For the five-shot setting, we randomly select five question-answer pairs for each combination of relation and logical implication type to create demonstrative examples, as depicted in 8. Complex prompting strategies such as chain-of-thought are not applied in our study, as the evaluation statements are crafted to be straightforward and easily understandable, allowing for verification without

Relation Type	Description
associated morphology of	A particular morphology (structural feature or form) is associated with another concept, often a disease.
disease has abnormal cell	A disease is characterized by the presence of abnormal cells.
disease has associated anatomic site	A disease occurs or has an impact at an anatomic site.
disease has normal cell origin	A disease originates from a type of normal cell.
disease has normal tissue origin	A disease originates from a type of normal tissue.
disease mapped to gene	A gene is associated with a specific disease.
disease may have associated disease	A disease may be associated with another disease.
disease may have finding	A possible clinical finding or symptom is observed in a disease.
disease may have molecular abnormality	A potential molecular abnormalities may be present in a disease.
gene encodes gene product	A particular gene encodes a specific gene product, such as protein.
gene product has associated anatomy	A gene product is associated to an anatomical structure.
gene product has biochemical function	A gene product is associated to a biochemical function.
gene product plays role in biological process	A gene product plays a role in a biological process.
has physiologic effect	A substance or process has a physiological effect on the body.
may prevent	A substance may prevent a disease.
may treat	A substance may treat a disease.
occurs after	A event or condition occurs after another.

Table 8: Relation types in the MedLAMA dataset that involve in our study.

the need for complex logical reasoning. In the inference process, we use greedy search for most of LLMs. However, commercial LLMs like GPT-3.5-turbo do not support greedy search, and we use their default generation setting to make a relative fair comparison across LLMs. We extract the prediction from models' response based on keywords since the words/phrases used to express True and False are limited. We listed all of the keywords applied to recognize answers in Table 7.

D Details of Baselines

We implement the LLMEval method by directly generating diverse statements using Llama3-70B-Instruct. Specifically, we prompt the LLM with the following instruction: "*Based on the given knowledge triplet, generate 8 statement to express the underlying knowledge in different ways. Output one statement per line. Directly output the statements without other text. Knowledge triplet: [triplet].*" To ensure the quality of generated samples, we use the greedy search for the decoding process. We find that Llama3-70B-Instruct can follow the instruction, generating samples in separated lines.

E Complementary Experiments

E.1 Joint accuracy

We illustrate the joint accuracy of all LLMs evaluated by PretextTrans and LLMEval in Figure 9 and 10, respectively. The experimental results support our conclusions: the evaluated LLMs generally perform worse on datasets generated by PretextTrans. Moreover, LLMs' performance decline faster when evaluated by PretextTrans compared with evaluated by LLMEval, indicating that current LLMs lack consistency in understanding medical knowledge presented in various structures.

E.2 Ablation Study

We also presents the ablation results of all evaluated LLMs regarding key components and logical implication types in Table 13 and 14, respectively. These results are consistent with our findings in the paper, demonstrating the effectiveness of our framework.

Relation Type	Description
Symptoms	Physical or mental feature that indicates the presence of the disease.
Affected sites	Specific parts of the body that are impacted or harmed by the disease.
Therapeutic Drugs	Pharmaceutical substances prescribed to manage, alleviate, or cure the symptoms and effects of the disease.
Surgical Procedures	Medical procedures that treat the disease, involving the cutting, repairing, or removal of body parts.

Table 9: Relation types involved in the DiseK dataset.

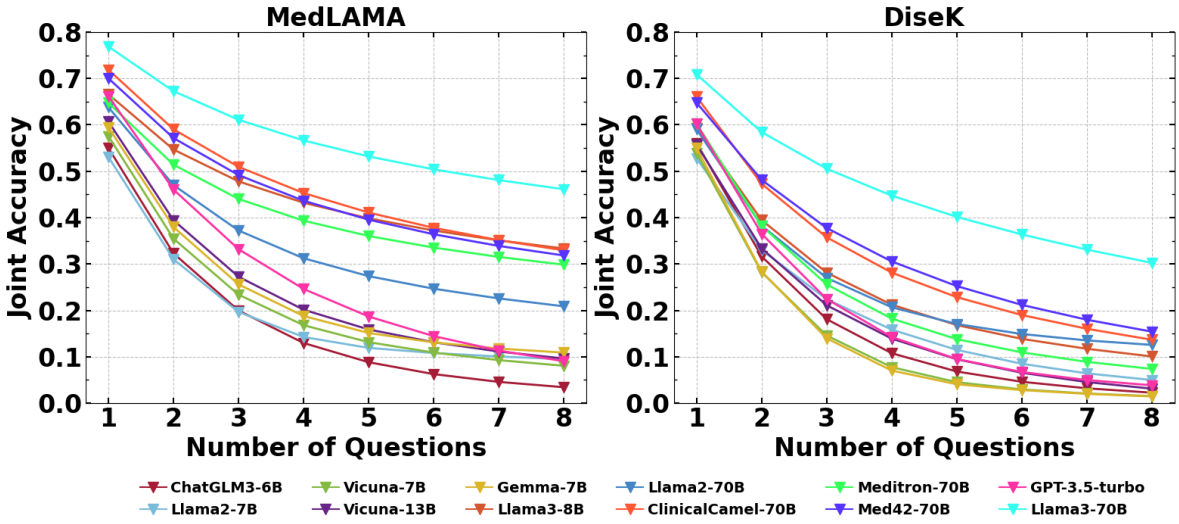

Figure 9: Performance (**joint accuracy**) of all LLMs evaluated by the proposed **PretextTrans** framework.

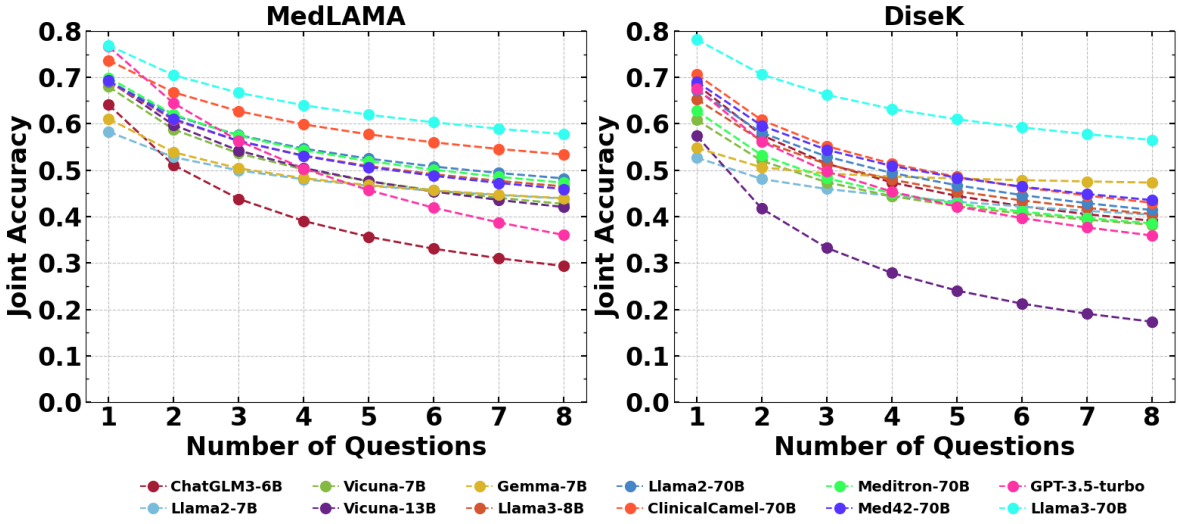

Figure 10: Performance (**joint accuracy**) of all LLMs evaluated by the **LLMEval** method.

Relation Type	Logical Implication Type			
	None	Inv	Ins	Inv+Ins
associated morphology of	[X] is the associated morphology of [Y] .	[Y] is often accompanied by the morphology of [X].	If a patient exhibits a morphological change of [X], then he/she may suffer from [Y].	If a patient suffers from [Y], then he/she is exhibiting a morphological change of [X].
disease has abnormal cell	[X] has the abnormal cell [Y] .	The abnormal cell type [Y] is detected within [X].	If a patient suffers from [X], then he/she has the abnormal cell [Y].	If a patient has the abnormal cell [Y], then he/she may suffer from [X].
disease has associated anatomic site	The disease [X] can stem from the associated anatomic site [Y] .	Anatomical site [Y] is associated with the development of disease [X].	If a patient suffers from [X], then he/she has lesions in [Y].	If a patient has lesions in [Y], then he/she may suffer from [X].
disease has normal cell origin	The disease [X] stems from the normal cell [Y] .	Normal cell [Y] is associaated with the development of disease [X].	If a patient suffers from [X], then he/she has lesions in [Y].	If a patient has lesions in [Y], then he/she may suffer from [X].
disease has normal tissue origin	The disease [X] stems from the normal tissue [Y] .	Normal tissue [Y] is associated with the development of disease [X].	If a patient suffers from [X], then he/she has lesions in [Y].	If a patient has lesions in [Y], then he/she may suffer from [X].
disease mapped to gene	The disease [X] is mapped to gene [Y] .	Gene [Y] is associated with the disease [X].	If a patient suffers from [X], then he/she has lesions in [Y].	If a patient has lesions in [Y], then he/she may suffer from [X].
disease may have associated disease	The disease [X] might have the associated disease [Y] .	The disease [Y] might have the associated disease [X] .	If a patient suffers from [X], then the likelihood of he/she suffering from [Y] is higher.	If a patient suffers from [Y], then the likelihood of he/she suffering from [X] is higher.
disease may have finding	[X] may have [Y] .	[Y] may be associate with [X]	If a patient suffers from [X], then he/she has [Y].	If a patient has [Y], then he/she may suffer from [X].
disease may have molecular abnormality	The disease [X] may have molecular abnormality [Y] .	Molecular abnormality [Y] may be associated with the disease [X].	If a patient suffers from [X], then he/she may has molecular abnormality [Y].	If a patient has molecular abnormality [Y], then he/she may suffer from [X].
gene encodes gene product	The gene [X] encodes gene product [Y] .	The gene product [Y] is encoded by the gene [X].	If the expression level of [X] decreases, it may lead to a reduction in the production or activity of [Y].	If the production or activity of [Y] decreases, it may caused by the reduction in the expression level of [X].
gene product has associated anatomy	The gene product [X] has the associated anatomy [Y] .	The anatomy [Y] is associated with the gene product [X].	The gene product [X] plays a role in anatomy [Y].	Anatomy [Y] is where [X] functions.
gene product has biochemical function	[X] has biochemical function [Y] .	[Y] is a biochemical function of [X].	If the production of [X] decreases, the functionality of [Y] may decrease.	If the functionality of [Y] decreases, it may caused by the reduction in the production of [X].
gene product plays role in biological process	The gene product [X] plays a role in biological process [Y] .	Biological process [Y] is associated with the gene product [X]	If the production of [X] decreases, the process of [Y] may be influenced.	If [Y] is affected, it may caused by the reduction in the production of [X].
has physiologic effect	[X] has physiologic effect of [Y] .	[Y] can be caused by [X].	If a patient takes [X], he/she may have physiologic effect of [Y] .	If a patient has physiologic effect of [Y], he/she may have taken [X].
may prevent	[X] may be able to prevent [Y] .	[Y] may be prevented by [X]	If a patient takes [X], he/she can prevent [Y].	If a patient wishes to prevent [Y], he/she should take [X].
may treat	[X] might treat [Y] .	[Y] may be treated by [X]	If a patient takes [X], he/she can treat [Y].	If a patient suffers from [Y], he/she should take [X].
occurs after	[X] occurs after [Y] .	[Y] may occur before [X].	If a patient occurs [X], he/she may occur [Y] before.	If a patient occurs [Y], he/she may occur [X] afterwards.

Table 10: Prototypes crafted for the MedLAMA dataset (1/2). Inv: inversion; Ins: instantiation.

Relation Type	Logical Implication Type			
	DN	Inv+DN	Ins+DN	Inv+DN
associated morphology of	[X] is not the associated morphology of [Y].	[Y] is not accompanied by the morphology of [X].	A patient that exhibits a morphological change of [X] does not suffer from [Y].	A patient that suffers from [Y] does not exhibit a morphological change of [X].
disease has abnormal cell	[X] does not has the abnormal cell [Y].	The abnormal cell type [Y] is not detected within [X].	A patient that suffers from [X] does not have the abnormal cell [Y].	A patient that has the abnormal cell [Y] does not suffer from [X].
disease has associated anatomic site	The disease [X] is not stem from the associated anatomic site [Y].	Anatomical site [Y] is not associated with the development of disease [X].	A patient that suffers from [X] does not have lesions in [Y].	A patient that has lesions in [Y] does not suffer from [X].
disease has normal cell origin	The disease [X] does not stem from the normal cell [Y].	Normal cell [Y] is not associaated with the development of disease [X].	A patient that suffers from [X] does not have lesions in [Y].	A patient that has lesions in [Y] does not suffer from [X].
disease has normal tissue origin	The disease [X] is not stem from the normal tissue [Y].	Normal tissue [Y] is not associated with the development of disease [X].	A patient that suffers from [X] does not have lesions in [Y].	A patient that has lesions in [Y] does not suffer from [X].
disease mapped to gene	The disease [X] is not mapped to the gene [Y].	Gene [Y] is not associated with the disease [X].	A patient that suffers from [X] does not have lesions in [Y].	A patient that has lesions in [Y] does not suffer from [X].
disease may have associated disease	The disease [X] is not associated with disease [Y] .	The disease [Y] is not associated with disease [X] .	If a patient suffers from [X], then the likelihood of he/she suffering from [Y] is not higher.	If a patient suffers from [Y], then the likelihood of he/she suffering from [X] is not higher.
disease may have finding	[X] does not have [Y] .	[Y] is not associated with [X]	A patient that suffers from [X] does not have [Y].	A patient that has [Y] does not suffer from [X].
disease may have molecular abnormality	The disease [X] does not have molecular abnormality [Y] .	Molecular abnormality [Y] is not associated with the disease [X].	A patient that suffers from [X] does not have molecular abnormality [Y].	A patient that has molecular abnormality [Y] does not suffer from [X].
gene encodes gene product	The gene [X] does not encode gene product [Y] .	The gene product [Y] is not encoded by the gene [X]	A decrease in the expression level of [X] does not affect the production and activity of [Y].	A decrease in the production or activity of [Y] is not caused by the reduction in the expression level of [X].
gene product has associated anatomy	The gene product [X] does not have the associated anatomy [Y] .	The anatomy [Y] is not associated with the gene product [X].	The gene product [X] does not play a role in anatomy [Y].	Anatomy [Y] is not where [X] functions.
gene product has biochemical function	[X] does not have biochemical function [Y] .	[Y] is not a biochemical function of [X].	A decrease in the production of [X] does not affect the functionality of [Y].	A decrease in the functionality of [Y] is not caused by the reduction in the production of [X].
gene product plays role in biological process	The gene product [X] does not play a role in biological process [Y] .	Biological process [Y] is not associated with the gene product [X]	A decrease in the production of [X] does not affect the process of [Y].	A change of [Y] is not caused by the reduction in the production of [X].
has physiologic effect	[X] does not have physiologic effect of [Y] .	[Y] cannot be caused by [X].	A patient that takes [X] does not have physiologic effect of [Y] .	A patient that has physiologic effect of [Y] has not taken [X].
may prevent	[X] is not able to prevent [Y] .	[Y] cannot be prevented by [X]	Taking [X] have no effect on preventing [Y].	A patient wishes to prevent [Y] has no need to take [X].
may treat	[X] is not able to treat [Y] .	[Y] cannot be treated by [X]	Taking [X] have no effect on treating [Y].	A patient that suffers from [Y] has no need to take [X].
occurs after	[X] does not occur after [Y] .	[Y] cannot occur before [X].	A patient occurs [X] will not occur [Y] before.	A patient occurs [X] will not occur [Y] afterwards.

Table 11: Prototypes crafted for the MedLAMA dataset (2/2). Inv: inversion; Ins: instantiation; DN: double negation.

Implication Type	Relation Type			
	Symptoms	Affected Sites	Therapeutic Drugs	Surgical Procedures
None	[Y] is a common symptom of [X].	[Y] is the affected site for [X].	[Y] is a common medication for [X].	[Y] is a common procedure for [X].
Inv	Common symptoms of [X] include [Y].	Affected sites for [X] include [Y].	Common medications for treating [X] include [Y].	Common procedures for treating [X] include [Y].
Ins	If a patient has [X], they are very likely to have symptoms of [Y].	If a patient has [X], their [Y] site is very likely to show lesions.	If a patient has [X], [Y] can be used to treat their condition.	If a patient has [X], [Y] can be used to treat their condition.
Inv+Ins	If a patient has symptoms of [Y], they are very likely to have [X].	If a patient shows lesions in their [Y] site, they are very likely to have [X].	If [Y] can be used to treat a patient's condition, they may have [X].	If [Y] can be used to treat a patient's condition, they may have [X].
DN	[Y] is not a common symptom of [X].	[Y] is not the affected site for [X].	[Y] is not a common medication for [X].	[Y] is not a common procedure for [X].
Inv+DN	Common symptoms of [X] do not include [Y].	Affected sites for [X] do not include [Y].	Common medications for treating [X] do not include [Y].	Common procedures for treating [X] do not include [Y].
Ins+DN	Patients with [X] are unlikely to have symptoms of [Y].	Patients with [X] are unlikely to show lesions in their [Y] site.	Patients with [X] do not commonly use [Y] for treatment.	Patients with [X] do not commonly use [Y] for treatment.
Inv+DN	Patients with symptoms of [Y] are unlikely to have [X].	Patients showing lesions in their [Y] site are unlikely to have [X].	Patients who can be treated with [Y] are unlikely to have [X].	Patients who can be treated with [Y] are unlikely to have [X].

Table 12: Prototypes crafted for the DiseK dataset. Inv: inversion; Ins: instantiation; DN: double negation.

Model	MedLAMA			DiseK		
	PretextTrans	-LogImp	-LMReph	PretextTrans	-LogImp	-LMReph
ChatGLM3-6B	55.0	$67.4_{\uparrow 12.4}$	$54.8_{\downarrow 0.2}$	56.1	$71.8_{\uparrow 15.7}$	$55.6_{\downarrow 0.5}$
Llama2-7B	53.1	$57.4_{\uparrow 4.4}$	$51.9_{\downarrow 1.1}$	52.8	$57.5_{\uparrow 4.7}$	$52.6_{\downarrow 0.2}$
Vicuna-7B	57.5	$72.1_{\uparrow 14.5}$	$55.7_{\downarrow 1.8}$	53.9	$59.5_{\uparrow 5.6}$	$52.5_{\downarrow 1.4}$
Vicuna-13B	60.7	$70.3_{\uparrow 9.6}$	$61.0_{\uparrow 0.4}$	55.7	$59.2_{\uparrow 3.5}$	$55.9_{\uparrow 0.2}$
Gemma-7B	59.4	$66.2_{\uparrow 6.8}$	$62.8_{\uparrow 3.4}$	55.0	$57.2_{\uparrow 2.2}$	$56.9_{\uparrow 2.0}$
Llama3-8B	66.6	$74.1_{\uparrow 7.5}$	$68.5_{\uparrow 2.0}$	59.3	$68.9_{\uparrow 9.7}$	$60.2_{\uparrow 0.9}$
Llama2-70B	63.8	$78.2_{\uparrow 14.4}$	$64.6_{\uparrow 0.8}$	59.0	$68.4_{\uparrow 9.3}$	$57.8_{\downarrow 1.3}$
ClinicalCamel-70B	71.9	$80.6_{\uparrow 8.8}$	$72.8_{\uparrow 1.0}$	66.1	$73.1_{\uparrow 7.1}$	$68.0_{\uparrow 1.9}$
Meditron-70B	64.7	$75.7_{\uparrow 11.0}$	$65.8_{\uparrow 1.1}$	60.2	$68.1_{\uparrow 7.9}$	$61.5_{\uparrow 1.3}$
Med42-70B	70.0	$78.2_{\uparrow 8.1}$	$70.4_{\uparrow 0.4}$	64.8	$70.4_{\uparrow 5.7}$	$67.9_{\uparrow 3.1}$
GPT-3.5-turbo	66.2	$78.3_{\uparrow 12.1}$	$67.9_{\uparrow 1.8}$	60.3	$67.1_{\uparrow 6.8}$	$61.8_{\uparrow 1.6}$
Llama3-70B	76.9	$83.0_{\uparrow 6.1}$	$80.4_{\uparrow 3.6}$	70.9	$77.8_{\uparrow 7.0}$	$74.0_{\uparrow 3.1}$

Table 13: Ablation results of all evaluated LLMs for key components of the proposed PretextTrans framework.

Model	MedLAMA				DiseK			
	None	+DN	+DN+Inv	+All	Origin	+DN	+DN+Inv	+All
ChatGLM3-6B	67.4	$55.7_{\downarrow 11.6}$	$55.9_{\downarrow 11.5}$	$55.0_{\downarrow 12.4}$	71.8	$56.0_{\downarrow 15.8}$	$57.1_{\downarrow 14.7}$	$56.1_{\downarrow 15.7}$
Llama2-7B	57.4	$53.6_{\downarrow 3.9}$	$53.6_{\downarrow 3.9}$	$53.1_{\downarrow 4.4}$	57.5	$54.3_{\downarrow 3.2}$	$53.9_{\downarrow 3.6}$	$52.8_{\downarrow 4.7}$
Vicuna-7B	72.1	$57.8_{\downarrow 14.3}$	$58.2_{\downarrow 13.9}$	$57.5_{\downarrow 14.5}$	59.5	$54.0_{\downarrow 5.5}$	$54.7_{\downarrow 4.8}$	$53.9_{\downarrow 5.6}$
Vicuna-13B	70.3	$62.0_{\downarrow 8.3}$	$61.6_{\downarrow 8.7}$	$60.7_{\downarrow 9.6}$	59.2	$53.8_{\downarrow 5.4}$	$55.8_{\downarrow 3.4}$	$55.7_{\downarrow 3.5}$
Gemma-7B	66.2	$61.5_{\downarrow 4.7}$	$60.8_{\downarrow 5.4}$	$59.4_{\downarrow 6.8}$	57.2	$53.6_{\downarrow 3.6}$	$55.2_{\downarrow 2.0}$	$55.0_{\downarrow 2.2}$
Llama3-8B	74.1	$69.0_{\downarrow 5.1}$	$68.5_{\downarrow 5.6}$	$66.6_{\downarrow 7.5}$	68.9	$60.9_{\downarrow 8.0}$	$60.1_{\downarrow 8.8}$	$59.3_{\downarrow 9.7}$
Llama2-70B	78.2	$66.6_{\downarrow 11.6}$	$65.8_{\downarrow 12.4}$	$63.8_{\downarrow 14.4}$	68.4	$61.0_{\downarrow 7.4}$	$59.7_{\downarrow 8.7}$	$59.0_{\downarrow 9.3}$
ClinicalCamel-70B	80.6	$73.8_{\downarrow 6.9}$	$73.2_{\downarrow 7.4}$	$71.9_{\downarrow 8.8}$	73.1	$68.9_{\downarrow 4.2}$	$67.9_{\downarrow 5.2}$	$66.1_{\downarrow 7.1}$
Meditron-70B	75.7	$66.8_{\downarrow 8.9}$	$65.8_{\downarrow 9.9}$	$64.7_{\downarrow 11.0}$	68.1	$60.2_{\downarrow 7.9}$	$61.1_{\downarrow 7.1}$	$60.2_{\downarrow 7.9}$
Med42-70B	78.2	$72.4_{\downarrow 5.8}$	$71.9_{\downarrow 6.3}$	$70.0_{\downarrow 8.1}$	70.4	$64.1_{\downarrow 6.3}$	$65.7_{\downarrow 4.7}$	$64.8_{\downarrow 5.7}$
GPT-3.5-turbo	78.3	$68.1_{\downarrow 10.2}$	$67.6_{\downarrow 10.7}$	$66.2_{\downarrow 12.1}$	67.1	$59.0_{\downarrow 8.1}$	$59.6_{\downarrow 7.5}$	$60.3_{\downarrow 6.8}$
Llama3-70B	83.0	$80.6_{\downarrow 2.4}$	$78.6_{\downarrow 4.3}$	$76.9_{\downarrow 6.1}$	77.8	$72.3_{\downarrow 5.6}$	$72.3_{\downarrow 5.5}$	$70.9_{\downarrow 7.0}$

Table 14: Ablation results of all evaluated LLMs for types of logical implication in the proposed framework.