TPEval: A Novel Truth-Preserving Evaluation Method for Probing LLMs Professional Factual Knowledge Mastery

Anonymous ACL submission

Abstract

Using large language models (LLMs) to solve problems in professional fields (e.g., medicine) is emerging as a research hotspot, requiring LLMs to master sufficient domain-specific factual knowledge. Recently, several LLMs achieved notable performance on multiple professional-field evaluation benchmarks. However, current benchmarks generally leverage common and fixed question formulations, allowing LLMs to provide correct answers based on surface-level patterns in questions without mastering the underlying knowledge. In this paper, we focus on this problem. We propose a general **t**ruth-**p**reserving **eval**uation framework (**TPEval**) to precisely probe LLMs' mastery of factual knowledge in professional fields through distinct representations of the same knowledge. Specifically, for each piece of knowledge, we convert its original expression into multiple truth-preserving statements with logical transformations, presenting the knowledge in diverse ways. By leveraging these statements, the proposed framework can more precisely estimate LLMs' mastery of the specified knowledge. Given the wealth of factual knowledge in medicine, we validate the effectiveness of our framework in the medical domain. We curate 6,000+ clinical facts and generate eight statements for each fact using the proposed method, evaluating the mastery of LLMs. Experimental results indicate a notable decline in LLMs' performance as the number of statements per fact increases, suggesting insufficient knowledge mastery of LLMs. Our method can serve as an effective solution for probing LLMs' knowledge mastery in professional fields.

1 Introduction

Recent years have witnessed the rapid advancement of large language models (LLMs), which have achieved considerable performance in various downstream applications (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; OpenAI, 2023; Madani et al., 2023; Boiko et al., 2023) and exhibited potential in several professional fields (e.g., medicine, finance, law). Solving problems in professional fields typically requires a comprehensive and in-depth mastery of the extensive factual knowledge within the specific domain. Recently, several studies (Petroni et al., 2019; Singh et al., 2023; Nori et al., 2023b; Wu et al., 2023) have reported that some LLMs (e.g., GPT-3.5-turbo) are capable of encoding domain-specific knowledge and largely surpass previous state-of-the-art models across various benchmark datasets within professional fields. Despite the considerable performance on existing evaluation benchmarks, it has also been observed that these LLMs are not practically applicable in real-world scenarios (Thirunavukarasu et al., 2023; Wornow et al., 2023; Li et al., 2023b), resulting in a gap between the evaluation and application. This paper aims to narrow this gap by more accurately evaluating current LLMs' proficiency in mastering domain-specific knowledge.

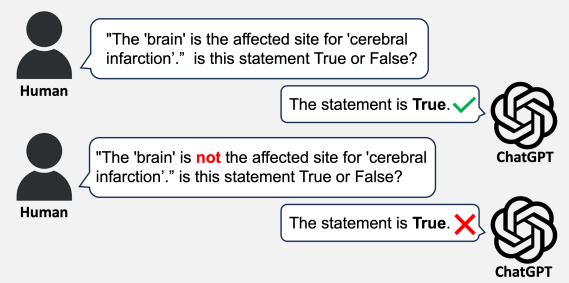


Figure 1: An example of GPT-3.5-turbo verifying two contradictory medical factual statements.

Several evaluation benchmark datasets have been proposed to evaluate LLMs' knowledge mastery in professional fields. Most of existing benchmark datasets (Hendrycks et al., 2020; Chen et al., 2021, 2022; Jin et al., 2021; Pal et al., 2022; Ben Abacha et al., 2017) leverage QA questions to evaluate LLMs' ability to answer questions using domain knowledge, while others also utilize tradi-

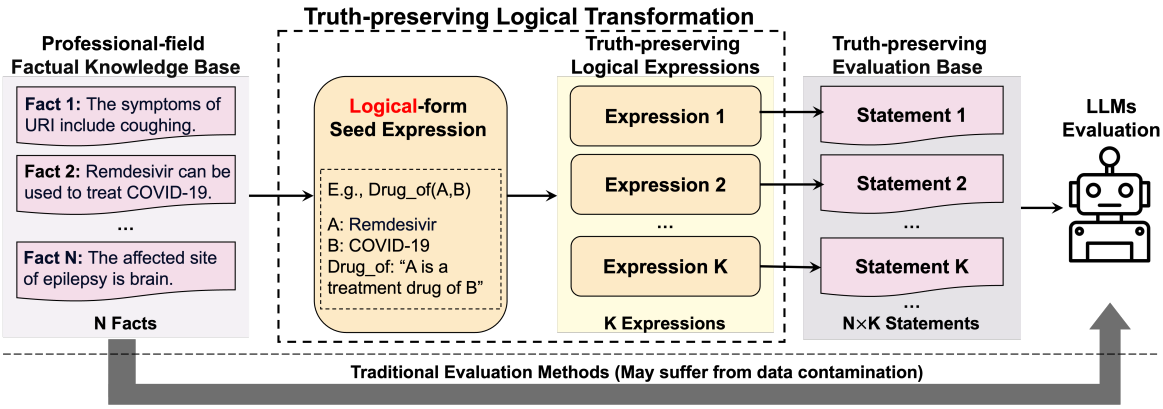


Figure 2: Framework of the proposed truth-preserving evaluation approach (**TPEval**) that probes LLMs' mastery of factual knowledge using multiple truth-preserving statements describing the knowledge in diverse ways.

tional NLP tasks (e.g., NER, sentiment analysis) (Zhang et al., 2022; Maia et al., 2018; Alvarado et al., 2015). However, these benchmark datasets typically evaluate LLMs' mastery of each fact with a fixed question formulation, which LLMs may have seen in pre-training. **As a result, LLMs may directly derive answers based on surface-level patterns in questions without mastering the underlying factual knowledge.** This issue is also known as **data contamination** (Sainz et al., 2023; Zhou et al., 2023). Figure 1 illustrates an example: GPT-3.5-turbo successfully verifies a medical factual statement but fails to check its negated version, suggesting the insufficiency of evaluating LLMs with uniform question formulations.

If an LLM masters a fact, it should understand various expressions of that fact. Motivated by this, we propose in this paper a novel **t**ruth-**p**reserving **eval**uation approach (**TPEval**) to precisely probe LLMs' mastery of factual knowledge in professional fields. Figure 2 presents the framework of our proposed method. Specifically, for each piece of knowledge, we transform it into a seed logical expression and deduce a series of expressions based on this seed expression. The truth-preserving nature of deductive reasoning guarantees the correctness of the generated expressions. Finally, all the expressions are transformed back into statements, where LLMs are asked to determine the truthfulness of these statements. **Compared to existing benchmarks, the proposed method evaluates the same knowledge through diverse expressions, thereby reducing the influence of LLMs in memorizing superficial patterns and leading to more accurate evaluation outcomes.**

Our proposed method transcends specific domains and is **adaptable** across diverse professional fields. Given the wealth of factual knowledge in the medical domain, we validate the effectiveness of our proposed method within this domain. Specifically, we select >6,000 pieces of clinical factual knowledge and generate eight statements for each of them employing the proposed method. Utilizing the generated statements, we evaluate a total of 14 LLMs, some of which (e.g., Gemini-pro) have achieved outstanding performance on existing medical benchmarks. Experimental results demonstrate that, though several LLMs perform well when evaluated with only one statement for each piece of knowledge, their performance sharply declines with the increasing number of statements. The results indicate that current LLMs have not mastered medical knowledge to the extent reflected by existing benchmarks. Moreover, we find that current LLMs generally perform worse when dealing with negative statements, suggesting they only have a surface-level mastery of medical knowledge. Our contributions are summarized as follows:

- We introduce a novel truth-preserving evaluation framework (**TPEval**) to evaluate LLMs' mastery of factual knowledge within professional domains. By evaluating LLMs with a series of truth-preserving statements, our method mitigates the impact caused by memorizing shallow cues and data contamination.
- Applying the proposed framework, we take the medical domain as an example and evaluate the mastery of LLMs on over 6,000 pieces of clinical medical knowledge.
- Furthermore, we compare LLMs' performance on different groups of statements along

three dimensions (knowledge type, statement polarity, and expression form), shedding light on developing domain-specific LLMs.

2 Related Work

LLMs in Professional Fields Recently, several famous LLMs, such as GPT-4, are reported to have achieved considerable performance on evaluation benchmarks across various professional fields. For example, in the medical domain, well-known LLMs Gemini-pro, Flan-PaLM, and GPT-4 achieve accuracies of 67.0, 67.6, and 90.2 on a medical exam benchmark MedQA (Pal and Sankarasubbu, 2024; Singhal et al., 2023; Nori et al., 2023b), largely surpassing previous SOTA performance (Liévin et al., 2023). In the financial domain, GPT-4 achieves notable performance on two financial QA datasets FinQA (Chen et al., 2021) and ConvFinQA (Chen et al., 2022) by 78.0 and 76.5, respectively (Li et al., 2023a), outperforming SOTA models by around ten percents. However, several studies (Thirunavukarasu et al., 2023; Wornow et al., 2023; Li et al., 2023b) demonstrate that these LLMs are not yet applicable in real-world scenarios. Therefore, we aim to study in this paper the causes of the gap between LLMs' benchmark performance and their insufficient practical effectiveness.

Evaluation Benchmarks for LLMs Various evaluation benchmarks have been developed in recent years to examine LLMs' mastery and application of domain-specific knowledge. Current evaluation benchmarks can be categorized into two types: (1) QA-based benchmarks that assess LLMs with multiple-choice questions, such as MMLU (Hendrycks et al., 2020), FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and LiveQA (Ben Abacha et al., 2017); (2) benchmarks that combine traditional NLP tasks with domain-specific corpora, such as CBLUE (Zhang et al., 2022), FiQA (Maia et al., 2018), and FIN (Alvarado et al., 2015). However, these benchmarks test LLMs' knowledge mastery with common questions, allowing LLMs to answer solely based on surface-level patterns. This paper tackles this issue by introducing a truth-preserving evaluation method, which generates multiple statements describing the same fact in various forms.

3 Truth-preserving Evaluation Method

3.1 Principle of the Method

In this section, we introduce the principle of our proposed truth-preserving evaluation approach (TPEval), which systematically probes LLMs' mastery of domain-specific factual knowledge based on truth-preserving statement generation. We denote the language-form expression of a piece of factual knowledge as S. The existing evaluation methods generally examine whether an LLM $\mathcal{M}$ has mastered the fact as follows:

$$m'_{\mathrm{S}} = f_{\mathrm{S}}(\mathcal{M}) \tag{1}$$

where f_{S} refers to the evaluation question generated based on S, and $m'_{\mathrm{S}} \in \{0, 1\}$ denotes the evaluation result: $m'_{\mathrm{S}} = 1$ when $\mathcal{M}$ answers the question correctly, otherwise $m'_{\mathrm{S}} = 0$. In contrast, the proposed TPEval method evaluate $\mathcal{M}$ by generating K truth-preserving statements $\{\mathrm{S}_i\}_{i=1}^{K}$ based on the same piece of factual knowledge:

$$\mathrm{p} = g(\mathrm{S}) \tag{2}$$

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

$$\mathrm{S}_i = g^{-1}(\mathrm{q}_i), 1 \leq i \leq K \tag{4}$$

$$\mathbf{m}_{\mathrm{S}} = [f_{\mathrm{S}_1}(\mathcal{M}), f_{\mathrm{S}_2}(\mathcal{M}), \cdots, f_{\mathrm{S}_K}(\mathcal{M})] \tag{5}$$

where g denotes a mapping that projects the original statement S into the associated logical form p (seed expression), and g^{-1} is its inverse operation. Deduce refers to the deductive reasoning process, and $\{\mathrm{q}_i\}_{i=1}^{K}$ are truth-preserving logical expressions deduced from the seed expression p. Compared with the former method, the proposed method mitigates the impact of the model memorizing specific patterns in the original expression, where the properties of deductive reasoning guarantee the validity of generated questions.

3.2 Truth-preserving Evaluation Framework

Built on the truth-preserving principle, we design a novel evaluation framework to evaluate LLMs' mastery of factual knowledge in professional fields more precisely.

Statement Generation We primarily consider factual knowledge that can be expressed by a triplet: $(\mathrm{A}, R, \mathrm{B})$, where A and B are entities, and R is the relation between them. Such type of factual knowledge is usually expressed as "A is/has the [attribute/relation] of/with B" in the language form

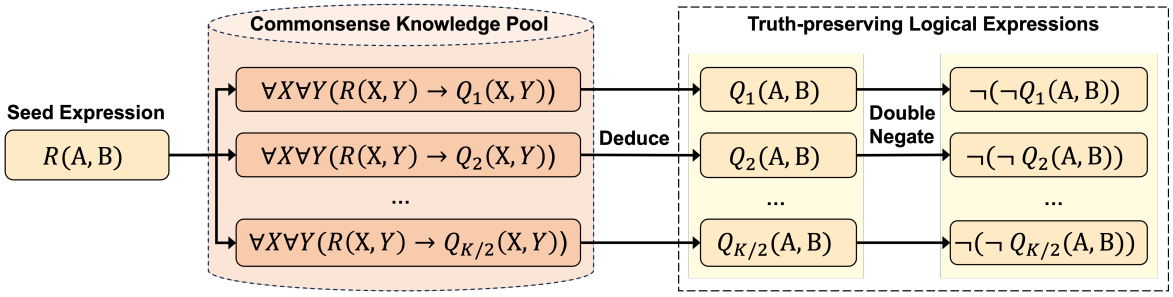


Figure 3: An overview of the truth-preserving logical transformation module in the proposed TPEval framework.

(S). The logical form of this expression is formulated as $\mathrm{p} = R(\mathrm{A},\mathrm{B})$, where $R(\mathrm{A},\mathrm{B})$ is a predicate that denotes "A has the relation R with B". For example, $\mathrm{Drug\_of}(\mathrm{A},\mathrm{B})$ presents a medical fact "A is the therapeutic medication of disease B."

We leverage deductive reasoning to transform the seed expression p into multiple truth-preserving expressions. Figure 3 depicts the whole transformation process. Specifically, we combine p with each piece of commonsense knowledge c_i to obtain a conclusion q_i based on syllogism. The commonsense knowledge encompasses the principles and rules in the specific domain, serving as the foundation for solving problems in this domain. In our framework, we focus on a type of commonsense knowledge represented as $c_i = \forall X\forall Y(R(\mathrm{X},\mathrm{Y}) \Rightarrow Q_i(\mathrm{X},\mathrm{Y}))$. To illustrate, consider a medical commonsense knowledge instance: "If a drug is a therapeutic drug of a disease, then the drug can be used to treat a person who suffers from the disease." Here, the phrases "a drug ... for a disease" and "the drug ... from the disease" can be denoted as $R(\mathrm{X},\mathrm{Y})$ and $Q_i(\mathrm{X},\mathrm{Y})$, respectively. Combining the factual knowledge p with each c_i, we have:

$$\begin{array}{ll} & \forall X\forall Y(R(\mathrm{X},\mathrm{Y}) \Rightarrow Q_i(\mathrm{X},\mathrm{Y})) \quad (c_i) \\ & R(A,B) \quad (p) \\ \hline \therefore & Q_i(A,B) \quad (\mathrm{q}_i) \end{array}$$

It is worth noting that if $\mathcal{M}$ incorrectly predicts the truth value of $Q_i(\mathrm{X},\mathrm{Y}))$, it indicates that the LLM either lacks the corresponding factual knowledge or commonsense knowledge. **In our framework, we choose commonsense knowledge that is clear, common, and simple as much as possible to ensure that the generated expressions are easy to understand.** We generate $K/2$ expressions through this process and also create another $K/2$ expressions based on the double negation rule: $\mathrm{q}_{i+K/2} = \neg(\neg \mathrm{q}_i), 1 \leq i \leq K/2$. This process yields a total of K truth-preserving expressions.

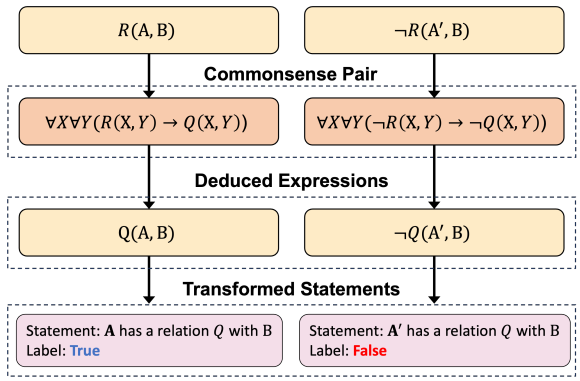


Figure 4: A comparison between the statement generation procedures based on positive and negative triplets.

Subsequently, each logical expression q_i is transformed into a statement S_i along with a label $l_i \in \{\mathbf{T},\mathbf{F}\}$. For q_i where $i \leq K/2$, we transform it using the predicate Q_i and set l_i as the true value of p. For $\mathrm{q}_{i+K/2}$ (double negation), it is transformed using the negated predicate $\neg Q_i$, with l_i set to the opposite value of p. It's worth noting that $\mathrm{S}_{i+K/2}$ can be derived from S_i by incorporating negative words (e.g., "not"). These statements are applied to evaluate $\mathcal{M}$'s mastery of the same factual knowledge.

Generation from Negative Triplets The statement generation method above is designed to generate statements based on positive knowledge triplets. As a result, p always holds in this scenario, causing statements generated by positive predicates to always be true, while those generated by negative predicates are consistently false. Therefore, generating statements exclusively from positive triplets could introduce bias, as LLMs may predict outcomes solely based on the presence of negation cues. Moreover, recognizing the absence of certain relations between entities is crucial for LLMs (e.g., a drug cannot be used to treat a specific disease). Therefore, we also generate statements from nega-

tive triplets, where these statements share the same formats as those generated from the corresponding positive triplets but have opposite labels.

Figure 4 compares the generation procedure based on positive triplets with negative ones. Specifically, for each $p = R(\mathrm{A}, \mathrm{B})$, we sample an A' that satisfies $\neg R(\mathrm{A}', \mathrm{B})$. Then, we treat $\neg R$ as a new predicate and employ the same method to generate truth-preserving statements. To ensure consistency in the formats of the generated statements with those from positive triplets, we only choose commonsense knowledge where the inverse proposition $(\forall X \forall Y(\neg R(\mathrm{X}, \mathrm{Y}) \Rightarrow \neg Q_i(\mathrm{X}, \mathrm{Y})))$ also holds. Consequently, for every $q_i = Q_i(\mathrm{A}, \mathrm{B})$, we have $q_i' = \neg Q_i(\mathrm{A}', \mathrm{B})$ holding true as well. Thus, S_i' can be derived by replacing A with A$'$ in S_i, and the corresponding label is exactly opposite to the original label: $l_i' = \neg l_i$. By generating statements from both positive and negative triplets, we can effectively prevent LLMs from verifying statements solely based on surface-level patterns and guarantee the completeness of the proposed evaluation framework.

LLM Evaluation The proposed truth-preserving evaluation principle does not restrict the types of evaluation questions. In our framework, we evaluate LLMs with statement verification questions, asking LLMs to determine whether the given statement S_i is true or false:

$$f_{\mathrm{S}_i}(\mathcal{M}) = \mathbb{1}(\mathcal{M}(\mathrm{S}_i) = l_i), 1 \leq i \leq K \quad (6)$$

Where $\mathcal{M}(\mathrm{S}_i) \in \{\mathbf{T}, \mathbf{F}\}$ denotes the LLM's prediction of S_i's truth value, $\mathbb{1}(\cdot)$ represents the characteristic function that equals 1 when the enclosed expression is true, and 0 otherwise.

We measure $\mathcal{M}$'s performance on a dataset that includes N pieces of knowledge by two modes: multi-statement **average** evaluation and multi-statement **joint** evaluation. These two modes calculate accuracies in different granularities:

$$a_{\text{avg}} = \frac{1}{N}\frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{K} f_{\mathrm{S}_j^i}(\mathcal{M}) \quad (7)$$

$$a_{\text{joint}} = \frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{K} f_{\mathrm{S}_j^i}(\mathcal{M}) \quad (8)$$

Here, S_j^i denotes the j^{th} statement derived from the i^{th} piece of knowledge. The former calculates the accuracy of $\mathcal{M}$ across all statements. In contrast, the latter calculates the accuracy across knowledge, considering a piece of knowledge being correctly answered if and only if **all the related statements** are judged correctly.

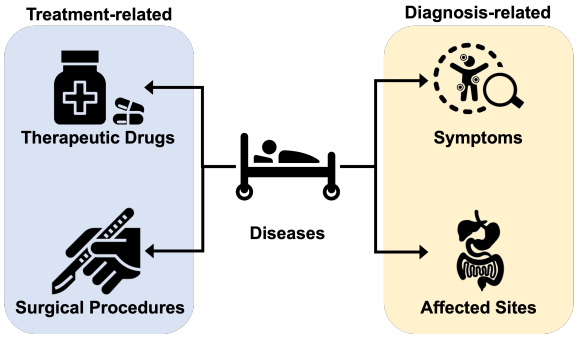


Figure 5: The knowledge structure of the proposed disease-centric medical knowledge base DiseK.

4 Experiments

4.1 Experiment Setup

Dataset Generation We select the medical domain to validate the effectiveness of the proposed method because it encompasses a wide range of factual knowledge. Moreover, several existing LLMs have been reported to achieve impressive performance on various medical benchmarks (Nori et al., 2023a; Singhal et al., 2023; Nori et al., 2023b). One of the primary goals in the medical domain is to diagnose and treat diseases. Motivated by this, we construct a large-scale disease-centric medical knowledge base, **DiseK**, covering 1,000 high-frequency diseases across four crucial knowledge aspects closely related to disease diagnosis and treatment. LLMs must master this fundamental medical knowledge to assist doctors in diagnosing and treating corresponding diseases (Wu et al., 2018; Liang et al., 2019). Figure 5 depicts the knowledge structure of the proposed knowledge base. Detailed statistics and annotation details of DiseK are provided in Appendix A.

DiseK contains more than 24k pieces of factual knowledge, each of them can be represented as a triplet $(\mathrm{A}, R, \mathrm{D})$, denoting "A is the R of disease D." Here, R corresponds to one of four knowledge aspects: symptoms, affected sites, therapeutic drugs, and surgical procedures, and A is an entity of R. To reduce computational cost in evaluation, we select a positive triplet $(\mathrm{A}, R, \mathrm{D})$ and a negative triplet $(\mathrm{A}', \neg R, \mathrm{D})$ for each pair of (R, D), resulting in 3,167 positive triplets and 3,167 negative triplets[1]. Subsequently, we generate a truth-

[1]Some diseases may not have certain aspects of knowledge,

preserving evaluation dataset **TPDiseK** using the proposed method, where each fact is evaluated by $K = 8$ statements. We initially transform a piece of knowledge into four truth-preserving statements based on commonsense knowledge. Two of the generated statements are derived through both trivial reasoning (S_1) and reverse reasoning (S_2), exhibiting minimal deviation from the original representation. The remaining two statements (S_3 and S_4) are generated by applying the factual knowledge to specific medical cases, thereby evaluating LLMs' capability of handling specific problems with the knowledge acquired. Another four statements with opposite labels (S_5 to S_8) are generated by **negating** these four statements. **The statement templates are meticulously designed to ensure they are easily understandable and faithfully express the meaning of corresponding logical expressions.** To summarize, TPDiseK consists of 6,334 knowledge triplets (positive and negative), each comprising 8 statements for evaluation. More details of TPDiseK are provided in Appendix B.

Evaluation Setting We primarily assess LLMs with the **five-shot in-context learning** strategy (Brown et al., 2020), where five demonstrative question-answer pairs are presented before the test question, guiding LLMs to produce answers consistent with the provided examples. We also examined the zero-shot performance of LLMs and found that the trend is similar to that observed in the five-shot setting. Therefore, we provide the zero-shot results for complement in Appendix D. We report the accuracies measured by the average and joint evaluation modes introduced in Sec 3.2. Note that **the joint accuracy can be regarded as the proportion of factual knowledge truly mastered by LLMs**. We present more details in Appendix C.

Evaluated Models In our study, we evaluate a total of 14 well-known general and medical-domain-specific LLMs on the proposed TPDiseK dataset: (1) general LLMs: ChatGLM (6B) (Du et al., 2022), Bloomz-mt (7.1B) (Muennighoff et al., 2023), Llama2 (7B,70B) (Touvron et al., 2023), Vicuna (7B,13B) (Zheng et al., 2023), GPT-3.5-turbo (Ouyang et al., 2022), ERNIE-Bot-turbo (Sun et al., 2021) and Gemini-pro (Team et al., 2023); (2) medical-domain-specific LLMs: Pulse (7B) (Zhang et al., 2023), ClinicalCamel (70B) (Toma et al., 2023), Meditron (7B,70B) (Chen et al., 2023) and Med42 (70B) (Christophe et al., 2023). We have not assess LLMs that are either too expensive (e.g., GPT-4 (OpenAI, 2023)) or not publicly available (e.g., MedPaLM (Singhal et al., 2023)).

such as therapeutic medication or surgeries.

Models	DiseK (K=1)	TPDiseK (K=8)
Random	50.0	50.0
ChatGLM-6B	52.6	48.7
Llama2-7B	52.8	51.7
Pulse-7B	54.9	51.4
Bloomz-mt-7.1B	52.3	48.8
Vicuna-7B	59.0	52.0
Meditron-7B	50.0	50.0
Vicuna-13B	61.2	54.0
Llama2-70B	65.9	56.5
ClinicalCamel-70B	74.4	64.4
Meditron-70B	70.3	57.0
Med42-70B	71.7	64.1
ERNIE-Bot-turbo	69.8	56.7
GPT-3.5-turbo	73.5	60.5
Gemini-pro	**78.0**	**73.1**

Table 1: Comparison of LLMs' average accuracy on the original medical evaluation dataset (**DiseK**) with that on the truth-preserving dataset (**TPDiseK**) generated by the proposed framework TPEval. K: the number of evaluated statements per knowledge triplet.

4.2 Results

4.2.1 Overall Performance

We initially compare the performance of LLMs on the original dataset DiseK with that on the truth-preserving dataset TPDiseK. LLMs' performance on DiseK is measured by their performance on the first type of statements (S_1), which expresses the knowledge in a trivial way (A is the R of B). The experimental results are provided in Table 1 and measured by average accuracy. We observe that LLMs $\leq$13B generally perform poorly on both datasets, while several 70B LLMs and commercial LLMs (Gemini-pro, ERNIE-Bot-turbo, GPT-3.5-turbo) achieve notable performance on the trivial representation of DiseK. **However, their performance significantly declines when using the proposed truth-preserving evaluation method.** LLMs like GPT-3.5-turbo, Meditron-70B, and ClinicalCamel-70B exhibit a performance decline of over 10% on TPDiseK compared to their performance on DiseK. Even the best-performing Gemini-pro demonstrates a performance gap of 4.9% between DiseK and TPDiseK.

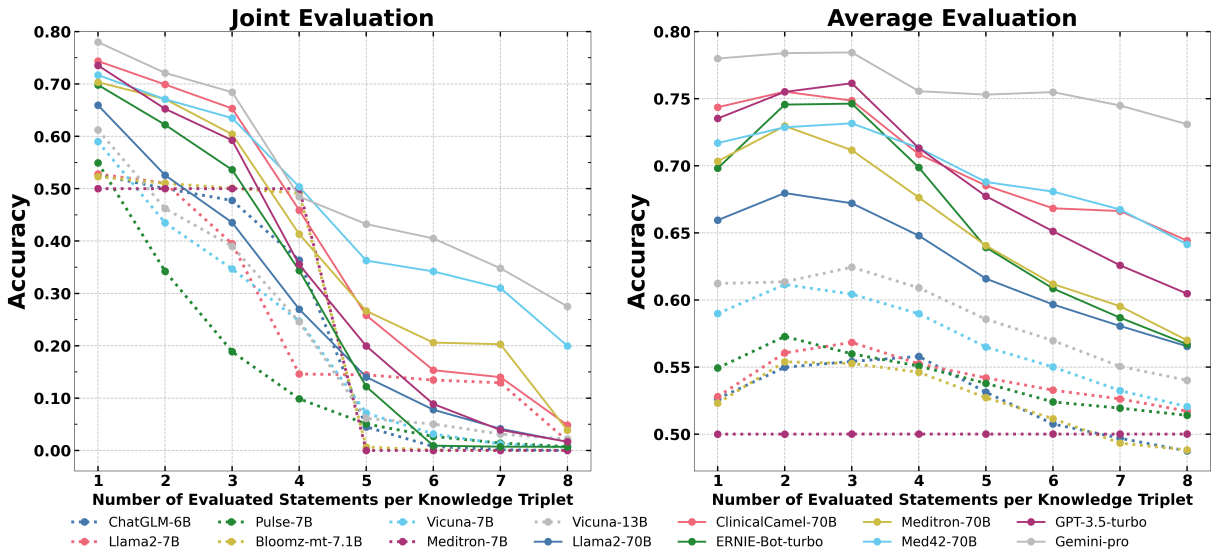


Figure 6: Performance of LLMs evaluated by increasing number of statements per triplet. **Left: joint accuracy** (proportion of triplets that all the related statements are correctly predicted); **Right: average accuracy.** Dotted lines: LLMs $\leq$13B; Solid lines: LLMs $>$13B.

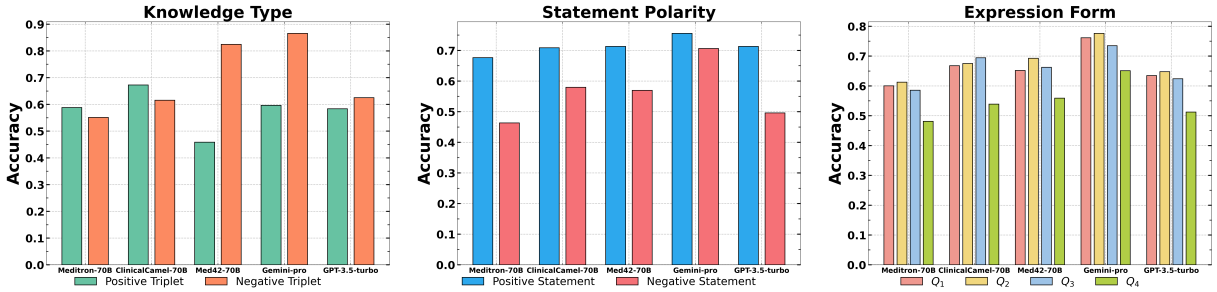
Subsequently, we study the performance of LLMs by gradually increasing the number of truth-preserving statements per knowledge triplet from 1 (DiseK) to 8. Statements are gradually added to the evaluation in the order of S_1 to S_8 defined in Sec 4.1. The experimental results are presented in Figure 6, showing accuracies for both joint evaluation (proportion of triplets where all related statements are correctly predicted) and average evaluation. The experimental results indicate a significant decrease in the joint accuracy of LLMs as the number of truth-preserving statements increases, declining much faster than their average accuracy. Surprisingly, some famous LLMs like GPT-3.5-turbo, ERNIE-Bot-turbo, and Llama2-70B even achieve comparable joint accuracies with LLMs $\leq$13B (dotted lines) when evaluated by all the statements. **The results suggest that while these LLMs may memorize more surface-level patterns of knowledge expressions than smaller models, they do not truly master a broader range of knowledge.** Gemini-pro and Med42-70B significantly outperform other LLMs regarding joint accuracy, suggesting a relatively comprehensive mastery of medical knowledge than others.

4.2.2 Performance across Statement Types

To further investigate current LLMs' performance on handling different types of statements, we categorize statements based on three criteria: (1) the type of knowledge triplets (positive/negative); (2) statement polarity (positive/negative); (3) the expression forms of statements determined by predicates $Q_i, 1 \leq i \leq 4$. We only present the top-5 performing LLMs here for convenience and provide the results of other LLMs in Appendix D. The performance is measured by average accuracy.

Knowledge Type Figure 7a presents LLMs' performance on statements generated from different types of knowledge triplets. We observe that LLMs exhibit varying degrees of proficiency in different types of factual knowledge (positive/negative). Gemini-pro and Med42-70B perform significantly better on negative triplets, indicating that they are more accurate in determining the absence of a relationship between two medical entities. It is worth noting that Med42 performs slightly worse than random guessing on positive triplets, suggesting a tendency to indicate no given relationship between two medical entities. In contrast, the other three LLMs achieve more balanced performance between different triplet types.

Statement Polarity Figure 7b examines how LLMs handle statements of different polarities: positive (S_1 to S_4) and negative (S_5 to S_8). The results show that **current LLMs generally perform significantly worse on negative statements**. Some LLMs' performance (Meditron-70B, GPT-3.5-turbo) even approaches or is inferior to random guessing. Gemini-pro significantly outperforms others on negative statements, achieving the most balanced performance on the two polarities.



(a) Performance on statements from different types of knowledge triplets.

(b) Performance on statements with different polarities.

(c) Performance on statements with varied expression forms (denoted by Q_i).

Figure 7: LLMs performance (average accuracy) grouped by different statement types according to three criteria. Only the top 5 performing LLMs are selected for convenience.

Expression Form We further investigate LLMs' performance on various expression forms in Figure 7c, where each Q_i denotes a specific expression form. Generally, LLMs perform notably worse on the last expression form than other forms. It may be because the expression form combines factual knowledge with specific cases and involves a contrapositive transform from the third expression. Therefore, this expression form occurs **less frequently** in existing medical corpora, making it less likely to be predicted by surface-level cues. Gemini-pro is the only LLM that achieves over 60% accuracy on every single expression form, suggesting its comprehensive mastery of medical factual knowledge compared to other LLMs.

5 Discussion

Effectiveness of Truth-preserving Evaluation The experimental results reveal that the LLMs' mastery of medical knowledge, as assessed by the proposed truth-preserving method, is notably lower than that evaluated by the traditional methods. Moreover, their performance declines sharply as the number of statements per knowledge triplet increases. These findings suggest that **current LLMs generate responses by memorizing surface-level patterns of knowledge expressions without genuinely mastering the underlying knowledge.** Furthermore, the examination across various statement types indicates that these LLMs struggle to manipulate factual knowledge through **basic transformations (e.g., negation)**, instead relying on specific forms of knowledge presentation. Some LLMs also prefer to memorize specific types of knowledge. All these findings demonstrate that the proposed truth-preserving evaluation method can serve as an effective solution to evaluate LLMs' mastery of factual knowledge in professional fields.

Insights into Developing Domain-specific LLMs The experimental results demonstrate that current LLMs lack an in-depth mastery of medical factual knowledge. Therefore, training LLMs to genuinely acquire domain knowledge is crucial rather than merely memorizing surface-level patterns in corresponding expressions. **Expanding the training data by expressing factual knowledge in diverse ways may potentially enhance LLMs' knowledge mastery in professional fields.**

6 Conclusion

Understanding LLMs' mastery of domain knowledge is crucial for their application in real-world scenarios. In this paper, we introduce a novel truth-preserving evaluation method (TPEval) to systematically evaluate LLMs' proficiency in factual knowledge of professional fields. The proposed method evaluates the same knowledge using multiple statements that present it in diverse ways, leading to a more precise estimation of LLMs' knowledge mastery. We investigate the proposed method in the medical domain based on >6,000 knowledge triplets from a medical knowledge base. The results reveal a significant drop in the performance of existing LLMs when assessed using the proposed TPEval method compared to traditional evaluation methods, suggesting that they lack an in-depth mastery of medical factual knowledge. Our method can serve as an effective solution for evaluating the knowledge mastery of LLMs in professional fields, shedding light on developing domain-specific LLMs. In the future, we aim to refine this method further by integrating it with more question formats, such as question answering, and applying it to more professional domains.

Limitations

One of the major limitation of our study is that we only verify the effectiveness of the proposed method in the medical field due to space constraints. However, the principle of our method is decoupled with specific professional fields, and the experimental results in this paper is already sufficient to demonstrate the effectiveness of our method. In future, we will utilize the proposed method to build more truth-preserving datasets in other professional fields to promote related researches.

Moreover, while our study evaluated several well-known general and medical-domain-specific LLMs, some other notable models like GPT-4 and MedPaLM were excluded. This was due to either their high costs (it would require $800 to evaluate GPT-4 on TPDiseK) or their unavailability for public access (e.g., MedPaLM). We will keep evaluating other LLMs in future if feasible.

References

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*.

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.

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*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023a. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Huiying Liang, Brian Y Tsui, Hao Ni, Carolina CS Valentim, Sally L Baxter, Guangjian Liu, Wenjia Cai, Daniel S Kermany, Xin Sun, Jiancong Chen, et al. 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine*, 25(3):433–438.

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. 2023. Variational open-domain question answering. In *International Conference on Machine Learning*, pages 20950–20977. PMLR.

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.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023b. 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.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Gagandeep Singh, Yue Pan, Jesus Andres-Ferrer, Miguel Del-Agua, Frank Diehl, Joel Pinto, and Paul Vozila. 2023. Large scale sequence-to-sequence models for clinical note generation from patient-doctor conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 138–143, Toronto, Canada. Association for Computational Linguistics.

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.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

Ji Wu, Xien Liu, Xiao Zhang, Zhiyang He, and Ping Lv. 2018. Master clinical medical knowledge at certificated-doctor-level with deep learning model. *Nature communications*, 9(1):4352.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Xiaofan Zhang, Kui Xue, and Shaoting Zhang. 2023. Pulse: Pretrained and unified language service engine.

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*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Details of DiseK

Knowledge Aspects	#Uniq	Avg.
#Symptoms	4,163	12.37
#Affected Sites	387	0.75
#Therapeutic Drugs	1,782	7.89
#Surgical Procedures	1,932	3.40

Table 2: Statistics of the proposed knowledge base **DiseK**. #Uniq: the number of unique entities that appear in the knowledge bases. Avg: the average number of entities associated to each disease. Note that some diseases may affect multiple organs and do not have specific affected sites.

We propose in this paper a large-scale disease-centric medical knowledge base **DiseK**, which involves a total of 1,000 high-frequency dieases along with 4 knowledge aspects that are essential to the diagnosis and treatment of diseases. The four knowledge aspects in DiseK are listed below:

- **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.

The diseases in DiseK are selected by filtering out top 1,000 most frequently occurring diseases among approximately four million medical records gathered from over 100 hospitals across five cities. After that, we ask 20 medical experts to annotate the related knowledge of these diseases from the four aspects introduced above. To ensure the quality of annotation, we first leverage a retrieval module to automatically retrieve the related knowledge from medical books, literature, and the Internet. Then the experts are asked to recheck the retrieved content, and revise the incorrect parts. We find that the utilized retrieve-and-check annotation framework can alleviate the burden of annotators while ensuring consistency in the annotations.

The proposed medical knowledge base is annotation by 20 medical experts over a period of approximately 3 months. The medical experts are employees in our company and have obtained medical practitioner licenses. The data collection process is approved by an ethics review board, and we have obtained approval from the person in charge to use the annotated data for research.

B Details of TPDiseK

Triplet	Type	Commonsense Expression
Pos	Triv.	$R(\mathrm{X},\mathrm{Y}) \rightarrow R(\mathrm{X},\mathrm{Y})$
	Rev.	$R(\mathrm{X},\mathrm{Y}) \rightarrow R^{-1}(\mathrm{Y},\mathrm{X})$
	Spec.	$R(\mathrm{X},\mathrm{Y}) \rightarrow (Has(\mathrm{Y}) \rightarrow P(\mathrm{Y}))$
	Contr.	$R(\mathrm{X},\mathrm{Y}) \rightarrow (\neg P(\mathrm{Y}) \rightarrow \neg Has(\mathrm{Y}))$
Neg	Triv.	$\neg R(\mathrm{X},\mathrm{Y}) \rightarrow \neg R(\mathrm{X},\mathrm{Y})$
	Rev.	$\neg R(\mathrm{X},\mathrm{Y}) \rightarrow \neg R^{-1}(\mathrm{Y},\mathrm{X})$
	Spec.	$\neg R(\mathrm{X},\mathrm{Y}) \rightarrow \neg(Has(\mathrm{Y}) \rightarrow P(\mathrm{Y}))$
	Contr.	$\neg R(\mathrm{X},\mathrm{Y}) \rightarrow \neg(\neg P(\mathrm{Y}) \rightarrow \neg Has(\mathrm{Y}))$

Table 3: The expressions of commonsense knowledge leveraged in our experiments. All the expressions hold for all X and all Y. Pos: positive triplets; Neg: negative triplets. R^{-1}: The R of the disease Y include X. $Has(\mathrm{Y})$: A patient **only** suffers from the disease Y.

DiseK contains 24,413 disease-related knowledge triplets in total. Assessing LLMs using all of these triplets would lead to a large computational cost. To balance the scale of evaluation and the computation efficiency, we select a single knowledge triplet $(\mathrm{A}, R, \mathrm{D})$ and a negative knowledge triplet $(\mathrm{A}', \neg R, \mathrm{D})$ for each pair of (R, D), resulting in a total of 6,334 knowledge triplets. Each positive triplet can be directly expressed by the statement "A is a R of the disease D".

Following the principle of truth-preserving evaluation, we choose four pieces of commonsense knowledge for deductive reasoning: (1) **Trivial** reasoning: the generated statement is exactly the same as the seed statement; (2) **Reverse** reasoning: we reverse the original expression, resulting in "The R of D include A"; (3) **Specialization**: we combine the triplet with a specific case, such as "If a patient only suffers from disease Y, he/she (has the symptom of $P(\mathrm{Y})$/has lesions in $P(\mathrm{Y})$/$P(\mathrm{Y})$ can be used to treat the patient)."; (4) **Contrapositive**: we conduct contrapositive transformation based on the third to derive this piece of commonsense. We also design similar deductive rules for negative triplets.

All of the commonsense knowledge used in our framework are listed in Table 3. Based on these pieces of commonsense knowledge and double negation, we generate a total of 8 statements for each triplet. Finally, the generated TPDiseK contains a total of 50,672 statements. We list all the utilized statement templates in Table 4.

Knowledge Aspect	Statement ID	Statement Template
Symptom	S_1	A is a common symptom of B.
	S_2	The common symptoms of B include A.
	S_3	If a patient only suffers from B, then he/she is likely to have symptoms of A.
	S_4	If a patient does not have the symptoms of A, then it is unlikely that he/she only suffer from B.
	S_5	A is not a common symptom of B.
	S_6	The common symptoms of B do not include A.
	S_7	If a patient only suffers from B, then he/she is unlikely to have the symptoms of A.
	S_8	If a patient have the symptoms of A, then it is unlikely that he/she only suffer from B.
Affected Site	S_1	A' is the affected site of B.
	S_2	The affected sites of B include A.
	S_3	If a patient only suffers from B, then he/she may have lesions in A.
	S_4	If a patient does not have lesions in A, then it is unlikely that he/she only suffers from B.
	S_5	A is not the affected site of B.
	S_6	The affected sites of B do not include A.
	S_7	If a patient only suffers from B, then he/she is unlikely to have lessions in A.
	S_8	If a patient have lesions in A, then it is unlikely that he/she only suffers from B.
Therapeutic Drug	S_1	A is a common therapeutic drug for B.
	S_2	The common therapeutic drugs used to treat B include A.
	S_3	If a patient only suffers from B, then A can be used to treat his/her condition.
	S_4	If A cannot be used to treat a patient's condition, then it is unlikely that he/she only suffers from B.
	S_5	A is not a common therapeutic drug for B.
	S_6	The common therapeutic drugs used to treat B do not include A.
	S_7	If a patient only suffers from B, then it is unlikely that A can be used to treat his/her condition.
	S_8	If A can be used to treat a patient's condition, then it is unlikely that he/she only suffers from B.
Surgical Procedure	S_1	A is a common surgical procedure for B.
	S_2	The common surgical procedures used to treat B include A.
	S_3	If a patient only suffers from B, then A can be used to treat his/her condition.
	S_4	If A cannot be used to treat a patient's condition, then it is unlikely that he/she only suffers from B.
	S_5	A is not a common surgical procedure for B.
	S_6	The common surgical procedures used to treat B do not include A.
	S_7	If a patient only suffers from B, then it is unlikely that A can be used to treat his/her condition.
	S_8	If A can be used to treat a patient's condition, then it is unlikely that he/she only suffers from B.

Table 4: The statement templates we use in our evaluation framework. A: an entity from the specified knowledge aspect that has/does not have the relation with the disease. B: the name of the given disease.

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

Table 5: The keywords we utilize to extract answers from LLMs' responses.

C More Details of Evaluation Setting

In our implementation, we form the statement verification question based on the template: "[Statement], is the statement above true or false? Please answer True or False." For the five-shot setting, we randomly choose another five diseases, and follow the similar method applied in constructing TPDiseK to form demonstrative examples. It is worth noting that we always leverage example statements in the same format of the test statement to achieve the best performance. For the zero-shot setting, we directly examine LLMs with the generated verification question. 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 the need for complex logical reasoning. In the inference process, we use greedy search for most of LLMs. However, some commercial LLMs (e.g, 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 recognize the answer from models' response based on keyword recognition since the words/phrases used to express True and False are limited. We listed all of the keywords we applied to recognize answers in Table 5.

D Complementary Experiments

D.1 Performance of all LLMs across Statement Types

We provide the detailed performance of all LLMs across different types of triplets, polarities, and expression forms in Figure 8, 9, and 10, respectively. The experimental results support the conclusions made in our paper: the evaluated LLMs generally perform worse on the negative and contrapositive types statements. Smaller LLMs perform significantly worse than larger LLMs on positive triplets, positive statements, and the first three expression forms.

D.2 Zero-shot Performance of LLMs

As introduced in our paper, we also study the zero-shot performance of LLMs on the proposed TPDiseK dataset. The experimental results are displayed in Figure 11, 12, 13, and 14, respectively. The experimental results show that LLMs generally achieve lower performance under the zero-shot setting ($< 10\%$), and LLMs' performance under the zero-shot setting declines faster than that under the five-shot setting. Some 70B LLMs (LLama2-70B and Meditron-70B) even achieve performance close to random guessing, indicating that they have a poor medical knowledge manipulation performance under the zero-shot setting. Nevertheless, the results under the zero-shot setting exhibit the similar trend compared to the five-shot setting, demonstrating the correctness of our conclusions based on the five-shot setting.

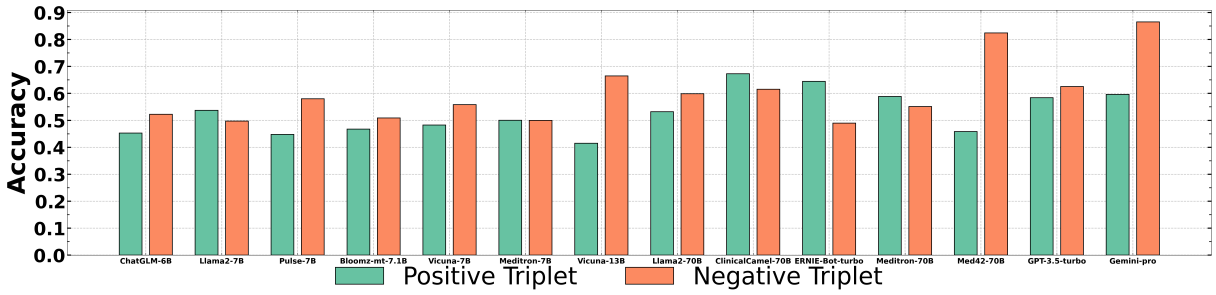


Figure 8: Average accuracy on statements generated from different types of knowledge triplets.

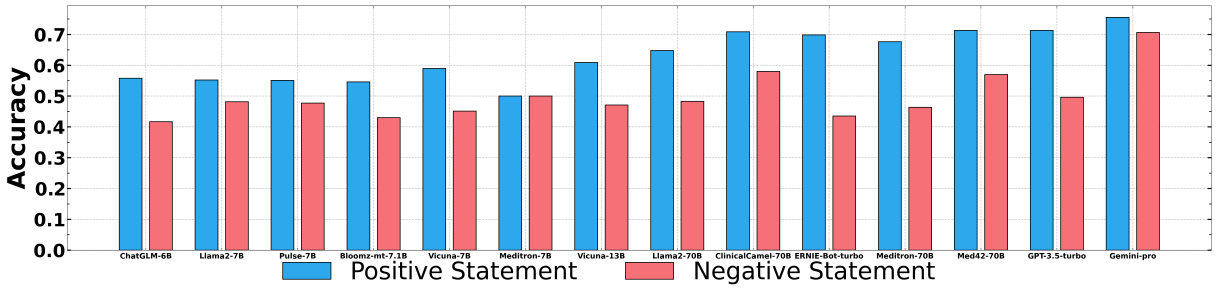


Figure 9: Average accuracy on statements with varied expression polarities. positive statements: S_1 to S_4; negative statements: S_5 to S_8.

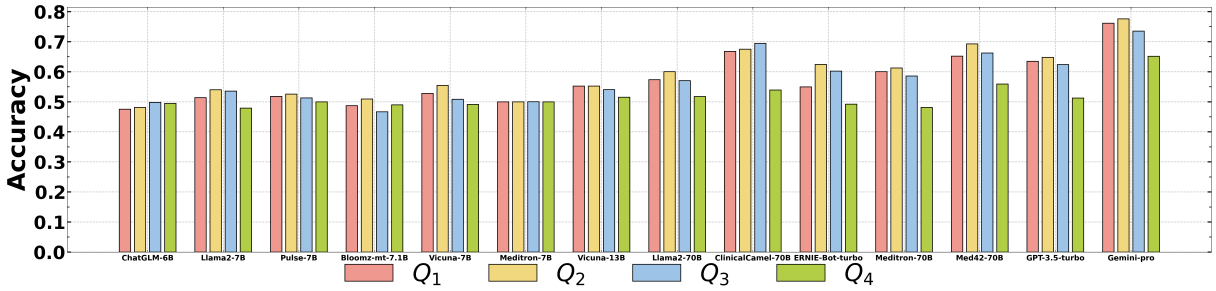


Figure 10: Average accuracy on statements with varied expression forms. Each expression form is determined by a specific predicate Q_i.

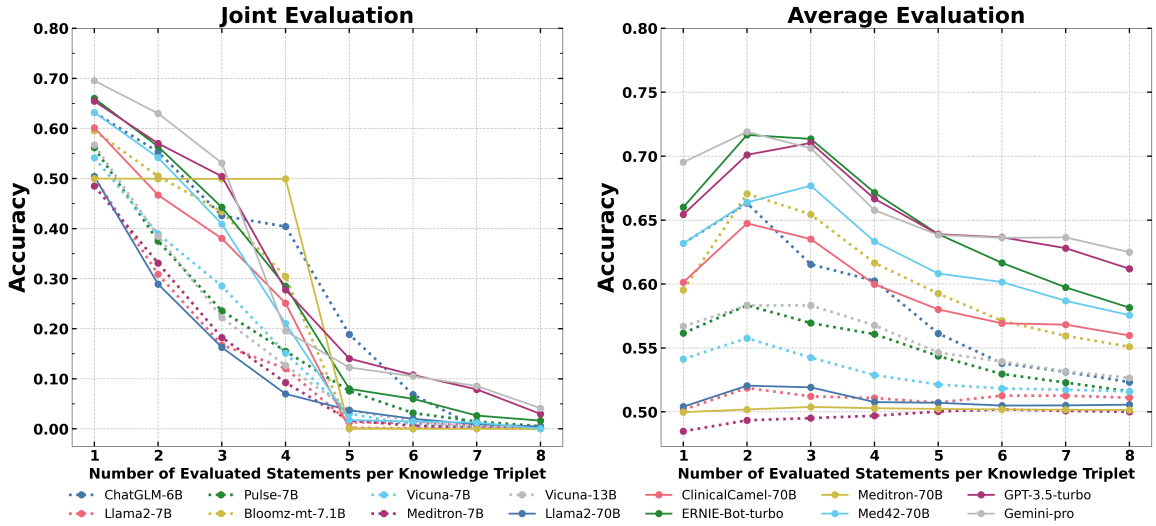


Figure 11: **Zero-shot** performance of LLMs evaluated by increasing number of truth-preserving statements. Left: joint accuracy; Right: average accuracy. Dotted lines: LLMs $\leq$13B; Solid lines: LLMs $>$13B.

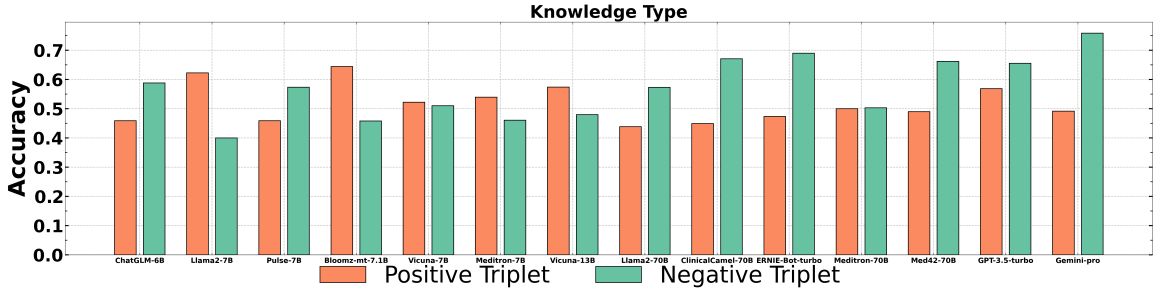


Figure 12: **Zero-shot** average accuracy on statements generated from different types of knowledge triplets.

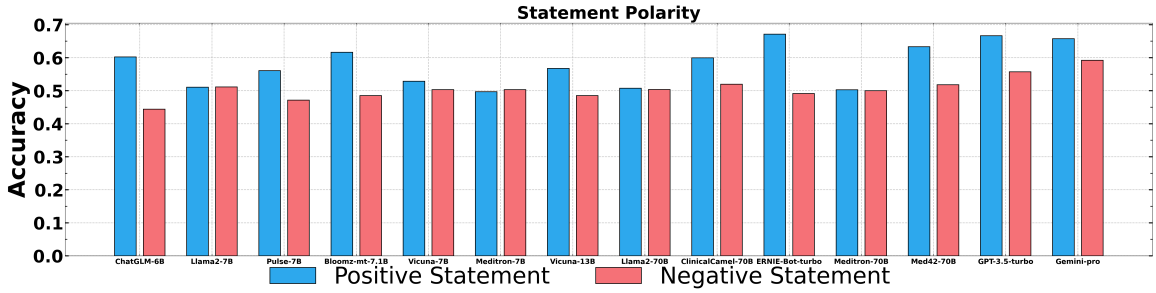


Figure 13: **Zero-shot** average accuracy on statements with varied expression polarities. positive statements: S_1 to S_4; negative statements: S_5 to S_8.

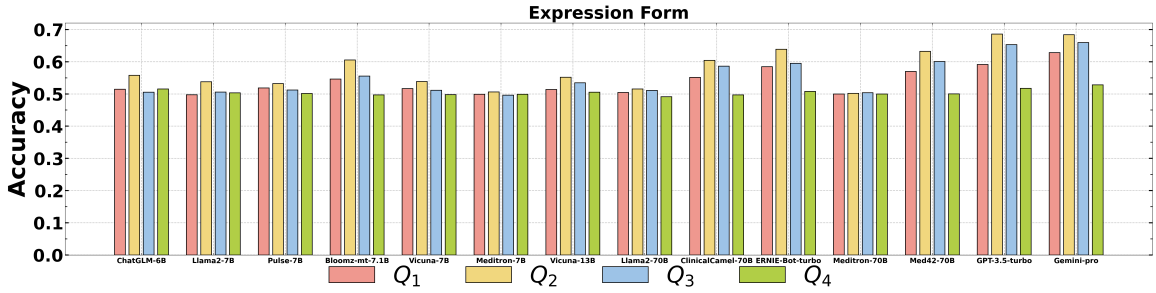


Figure 14: **Zero-shot** average accuracy on statements with varied expression forms. Each expression form is determined by a specific predicate Q_i.