MedErrBench: A Fine-Grained Multilingual Benchmark for Medical Error Detection and Correction with Clinical Expert Annotations

Congbo Ma[1], Yichun Zhang[2], Yousef Al-Jazzazi[1], Ahamed Foisal[1], Laasya Sharma[3], Yousra Sadqi[4], Khaled Saleh[4], Jihad Mallat[4], Farah E. Shamout[1]
[1]New York University Abu Dhabi, [2]New York University,
[3]University of Birmingham, [4]Cleveland Clinic Abu Dhabi

Abstract

Inaccuracies in existing or generated clinical text may lead to serious adverse consequences, especially if it is a misdiagnosis or incorrect treatment suggestion. With Large Language Models (LLMs) increasingly being used across diverse healthcare applications, comprehensive evaluation through dedicated benchmarks is crucial. However, such datasets remain scarce, especially across diverse languages and contexts. In this paper, we introduce MedErrBench, the first multilingual benchmark for error detection, localization, and correction, developed under the guidance of experienced clinicians. Based on an expanded taxonomy of ten common error types, MedErrBench covers English, Arabic and Chinese, with natural medical cases annotated and reviewed by domain experts. We assessed the performance of a range of general-purpose, language-specific, and medical-domain language models across all three tasks. Our results reveal notable performance gaps, particularly in non-English settings, highlighting the need for clinically grounded, language-aware systems. By making MedErrBench and our evaluation protocols publicly-available, we aim to advance multilingual clinical NLP to promote safer and more equitable AI-based healthcare globally. The dataset is publicly available at: https://github.com/congboma/MedErrBench.

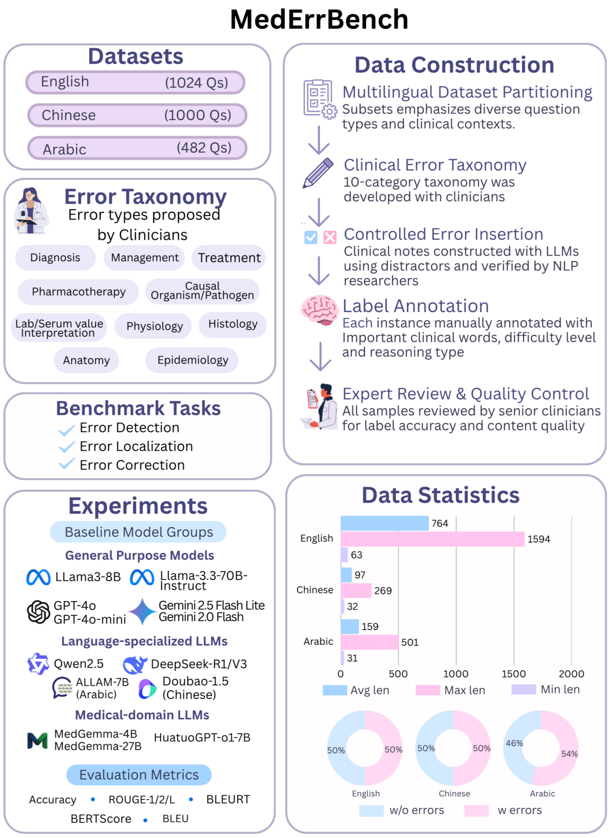


Figure 1: Overview of MedErrBench.

1 Introduction

Medical error detection and correction are essential for ensuring patient safety and healthcare quality (Ahsani-Estahbanati et al., 2022; Anjum et al., 2024; Iwase et al., 2025). Some errors, such as misdiagnoses, can lead to severe adverse outcomes, such as morbidity and mortality, and high economic costs (Newman-Toker et al., 2021; Soori, 2024). This need is especially critical in the era of generative AI and Large Language Models (LLMs) (Magnini et al., 2025). Despite their importance, automated methods for detecting and correcting medical errors remain underexplored due to three main challenges. First, there is a significant shortage of publicly available datasets for medical error correction and detection. To date, MEDEC (Abacha et al., 2025) is one of the main datasets specifically designed for this purpose, limiting the development and evaluation of robust models. Second, the diversity of error taxonomies is often insufficient, failing to reflect the complexity of real-world medical errors. The lack of standardized expert-informed annotation protocols further results in fragmented and incomplete error representations. Third, the existing datasets focus primarily on English, limiting model development for multi-

Findings of the Association for Computational Linguistics: ACL 2026, pages 11802–11827
July 2-7, 2026 ©2026 Association for Computational Linguistics

lingual medical contexts. Given the global nature of healthcare, the lack of multilingual resources hinders progress toward robust systems.

To address these limitations, we propose MedErrBench, a multilingual benchmark dataset for medical error detection and correction, grounded in expert-defined error categories and supported by high-quality annotated data. An overview is provided in Figure 1. In collaboration with clinicians, we distilled potential error types into ten representative categories: Diagnosis, Management, Treatment, Pharmacotherapy, Causal Organism/-Pathogen, Lab/Serum Value Interpretation, Physiology, Histology, Anatomy, and Epidemiology (see Table 1 for definitions). These categories provide comprehensive coverage of clinically relevant errors and serve as a practical guideline for data annotation and system evaluation. Based on this typology, we construct MedErrBench in English, Arabic, and Chinese. For English and Chinese, we adapt samples from MedQA (Jin et al., 2021) by introducing expert-verified errors into clinical notes. Each instance is manually labeled with the corresponding error type. Additional erroneous examples were created by clinicians for underrepresented error types. For Arabic, we adapt samples from MedArabiQ (Daoud et al., 2025) and MedAraBench (Mouath et al., 2026). Under the proposed typology, error types not present in the original datasets were supplemented with real-world medical error cases contributed by practicing clinicians. Beyond error categorization, we also label the difficulty level and reasoning type, enabling models to learn fine-grained, error-focused reasoning not supported by existing datasets. All data across the three languages are independently reviewed by two clinicians, who corrected any issues introduced during the transformation process and validated both content accuracy and annotation quality. This rigorous pipeline ensures that MedErrBench is both clinically valid and well-suited for training and evaluating medical error detection systems.

The new proposed MedErrBench dataset supports three key clinical NLP tasks: error detection, localization, correction. To explore these tasks, we evaluate three representative classes: general-purpose LLMs, language-specific LLMs, and domain-specific medical LLMs, across English, Arabic, and Chinese. Beyond overall performance benchmarking, we conduct a series of in-depth analyses: (1) investigation of the effects of providing error-type definitions and exemplar cases; (2) analysis of example difficulty in few-shot learning settings; (3) performance differences between knowledge-based and description-based clinical notes, and (4) cross-lingual generalization. These comprehensive experiments provide a detailed understanding of current LLM capabilities and highlight the need for clinically grounded, language-aware models in high-stakes medical applications. Our main contributions are:

Table 1: Classification of medical error types with definitions.

Error Type	Definition
Diagnosis	Failure to correctly identify the underlying condition based on medical presentation
Management	Inappropriate non-pharmacologic, non-surgical medical decision such as observation, monitoring, or disposition
Treatment	Inappropriate definitive intervention (surgical, procedural, or pharmacologic); distinct from general management
Pharmacotherapy	Incorrect drug selection, dosage, route, timing, interaction, or duration
Causal Organism / Pathogen	Misidentification of the causative microorganism in infectious disease
Lab Value Interpretation	Misreading or misapplying diagnostic thresholds, reference ranges, or derived values
Physiology	Misconception or misinterpretation of physiological principles (e.g., ECG, PFT, etc.)
Histology	Misinterpretation of tissue morphology, cellular structures, or microscopic patterns
Anatomy	Errors in anatomical structure, relation, or spatial understanding
Epidemiology	Misuse of statistical tools or misstatement of incidence, prevalence, or risk factors

- We establish a clinician-informed taxonomy of ten medical error types. This typology reflects real-world challenges, and provides a foundational schema for future dataset construction and evaluation in clinical NLP.

- We introduce MedErrBench, the first fine-grained multilingual benchmark for medical error detection and correction in English, Chinese, and Arabic. It supports three tasks: error detection, localization, correction, and is rigorously validated by clinicians to ensure medical plausibility.

- We systematically benchmark a range of LLMs across multiple languages and model families, and conduct in-depth analyses to understand their capabilities and limitations in medical error understanding. Our findings highlight the challenges faced by current models and motivate future research in building more robust and clinically aware systems.

2 Related Work

2.1 General Error Detection and Correction

Error detection and correction have been applied across a range of domains, including grammatical error correction (Peng et al., 2025; Ye et al., 2025; Kaneko et al., 2022), code debugging and repair (Tsai et al., 2024; Tian et al., 2024), data cleaning (Reis et al., 2024), and fact verification (Setty, 2024; Ni et al., 2024). To improve performance, especially under limited annotated data, researchers have proposed techniques such as synthetic error generation (Stahlberg and Kumar, 2024) and auxiliary linguistic or contextual signals (Fei et al., 2023). More recently, LLMs have been applied to error detection and correction (Kamoi et al., 2024a) through direct correction generation (Loem et al., 2023) and instruction tuning (Fan et al., 2023). Additionally, some studies leverage LLM feedback loops to refine model outputs (Kamoi et al., 2024b; Pan et al., 2024; Gou et al., 2024). Although LLMs occasionally exhibit over-correction and misalignment with user intent (Vasselli and Watanabe, 2023), human evaluations often find their corrections more fluent and acceptable compared to task-specific models (Zeng et al., 2024).

2.2 Error Detection and Correction in Healthcare

Error detection and correction are critical in healthcare due to their impact on medical decisions and patient safety. The MEDIQA-CORR 2024 Shared Task introduced MEDEC (Abacha et al., 2025), the first public dataset for evaluating errors in clinical notes based on five main error types. Existing methods can be broadly categorized into two types: (1) prompting-based LLM strategies and (2) hybrid or traditional approaches. Prompt-based systems employ few-shot in-context learning and chain-of-thought reasoning (Wu et al., 2024; Gundabathula and Kolar, 2024), with some leveraging retrieval-augmented generation to incorporate external knowledge (Rajwal et al., 2024; Corbeil, 2024). Strategies include structured prompt templates, error-type hints, and self-consistency sampling. Some systems adopt in-prompt ensembling by combining outputs from multiple expert prompts, weighted by trust scores (Valiev and Tutubalina, 2024), while others rely on manual error-type categorization to guide the reasoning process. Others train models to generate rationales before proposing corrections (Wu et al., 2024). Hybrid methods combine traditional classifiers, such as support vector machines, with QA-based correction modules (Saeed, 2024). These approaches emphasize interpretability and efficiency by incorporating domain-specific features like TF-IDF scores, medical terminology patterns, and handcrafted rules. Nevertheless, current research predominantly depends on MEDEC, which is monolingual and lacks a broader coverage of other medical error types, constraining the generalizability and medical applicability of proposed methods.

3 Methodology

3.1 Multilingual Dataset Partitioning

To support robust and generalizable research in medical error detection and correction, we constructed a trilingual dataset in English, Chinese, and Arabic, reflecting linguistic and regional diversity in medical education systems: English as the global scientific lingua franca, Chinese representing the world's most spoken language, and Arabic capturing the Middle East and North Africa region. We note that the datasets are not translations and are collected from multiple native-language sources, ensuring multilingual fidelity. Multi-source design introduces cross-site variability for more robust evaluation.

The English and Chinese subsets were initially sampled from MedQA (Jin et al., 2021), which includes medical cases sourced from medical licensing examination questions used in the US and China. We applied filtering criteria to ensure contextual richness, removing short factoid-style questions (typically 1–2 sentences) and retaining longer, multi-sentence clinical notes that provide realistic and meaningful diagnostic or therapeutic contexts. The Arabic subset was sampled from MedArabiQ (Daoud et al., 2025) and MedAraBench (Mouath et al., 2026). Each Arabic question was tagged with a question style label: either scenario-based or knowledge-based. Scenario-based questions typically include a brief medical case and require interpretation in context, whereas knowledge-based questions assess general medical facts. We used keyword heuristics (e.g., “patient", “age", “child") to extract scenario-based questions, and then hand-picked more challenging knowledge-based questions with sufficient word count and specialty coverage.

3.2 Building the Taxonomy of medical Error Types

To construct a medically grounded error typology, we collaborated with experienced clinicians to identify and refine ten representative error categories, building on MEDEC and adding five new error types: Lab/Serum Value Interpretation, Physiology, Histology, Anatomy, and Epidemiology, alongside previously established ones including Diagnosis, Management, Treatment, Pharmacotherapy, and Causal Organism/Pathogen. These categories are designed to comprehensively capture the range of factual errors commonly encountered in medical practice and serve as a practical framework for both data annotation and system evaluation. For each error type, we provide a clear definition along with an example scenario to illustrate its typical manifestations. Table 1 provides detailed descriptions, while in the appendix Figures S9–S11 illustrate the ten error types in MedErrBench.

3.3 Error Injection & Dataset Construction

To develop a clinically grounded dataset for medical error detection, we introduced the errors into the partitioned multilingual medical cases.

For each question, we preserved the correct answer and randomly selected one plausible but incorrect alternative, while discarding the remaining distractors. Using these selected answers, we constructed two versions of a clinical note: one in which the correct answer was naturally integrated into the context, and another in which the incorrect answer was inserted in its place.

The original datasets lacked certain error types. To address this gap and ensure comprehensive coverage, we collaborated with experienced clinicians who contributed real-world medical cases for the missing categories. Specifically, the English dataset originally lacked Physiology, Histology, Anatomy, and Epidemiology examples, while the Arabic dataset lacked Lab/Serum Value Interpretation examples. To operationalize this construction process, we employed LLMs to assist in transforming the original samples into full-length medical narratives in all three languages. Supplementary Figures S2–S4 provide the prompts for reproducibility, while Figures S6-S8 illustrate error insertion examples from MedErrBench.

3.4 Important medical Words, Difficulty-Level and Reasoning-Type Annotation

To better analyze task complexity and enable more fine-grained evaluation of model capabilities, we additionally manually annotate each instance with three auxiliary attributes: important medical words, difficulty level, and reasoning type. Important medical words capture the most salient concept or decision point in each case (e.g., a diagnosis, therapeutic action, or critical finding), highlighting the key linguistic cues that both humans and models must attend to during error detection. Each instance was assigned one of three difficulty levels: Easy, Medium, or Hard. The annotation process followed a set of predefined medical reasoning guidelines that considered multiple factors, including the clarity of medical cues, the rarity or complexity of the underlying condition, the number of reasoning steps required to reach a correct conclusion, and the overall length and ambiguity of the question text. Additionally, each item was annotated according to the type of reasoning required to answer it, using a three-level classification scheme: Factual Recall, Single-hop Reasoning, and Multi-hop Reasoning. These categories reflect increasing levels of inferential complexity and are critical for evaluating the diagnostic reasoning abilities of LLMs.

3.5 Expert Review and Quality Control

We employed a rigorous two-stage human review and quality control process. In the first stage, three native-speaking NLP researchers (English, Chinese, and Arabic), after studying the clinician-defined taxonomy of ten medical error types, performed initial annotation and manual verification. This process involved identifying hallucinated or unnatural content, removing incorrect or extraneous information introduced by LLMs, and segmenting overly long sentences, particularly in the Chinese dataset, where punctuation such as commas often failed to separate clauses properly.

In the second stage, two clinicians were assigned per language reviewed all instances to ensure medical validity [1]. They corrected transformation errors, validated error labels, and verified key medical terms, difficulty levels, and reasoning types.

[1]For the English and Arabic datasets, the clinicians are licensed practitioners currently practicing in the Middle East at a multi-specialty American hospital. For the Chinese dataset, the clinicians are licensed practitioners currently practicing in China.

Disagreements were categorized as logical errors, misclassifications, or typos; typos were corrected directly, while other issues were revised if flagged by either clinician. No conflicting corrections were proposed.

4 Experiment and Results

4.1 Evaluation Metrics

We evaluated model performance across three subtasks: error detection, error localization, sentence correction. We use Accuracy, ROUGE (Lin, 2004) (including ROUGE-1, ROUGE-2, ROUGE-L (R-L)), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020) as the main evaluation metrics[2].

4.2 Baseline Models

We evaluate a diverse set of recent language models, grouped by their design objectives:
Group 1: General-purpose LLMs. This group includes models developed for broad, cross-domain language tasks. We evaluate GPT-4o (OpenAI, 2024b), GPT-4o-mini (OpenAI, 2024a), Gemini 2.5 Flash Lite (Google Research, 2025a), Gemini 2.0 Flash (Google Research, 2025a), LLaMA3-8B (Meta AI, 2024) and Llama-3.3-70B-Instruct[3].
Group 2: Language-specialized LLMs. These models are primarily optimized for specific languages. Our selection includes Chinese models (Qwen2.5-7B-Instruct (Alibaba DAMO Academy, 2024), DeepSeek-R1 (DeepSeek AI, 2024a), DeepSeek-V3 (DeepSeek AI, 2024b), Doubao-1.5-Thinking-Pro (Doubao Team, Volcano Engine, 2024) and Arabic model[4] ALLAM-7B (Bari et al., 2025).
Group 3: Medical-domain LLMs. This group comprises models specifically designed for medical and biomedical applications. We include MedGemma-4B (Google Research, 2025a,b), MedGemma-27B (Google Research, 2025a,b), and HuatuoGPT-o1-7B (Chen et al., 2024).

[2]Due to page limitations, we report Accuracy (Acc), ROUGE-1 (R-1), BERTScore (BS), and BLEURT (BRT) in the main paper, and present the remaining evaluation results in the appendix.

[3]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[4]We evaluated Falcon and Jais; however, both exhibited issues for error detection and correction task. For example, Falcon-Arabic consistently returned `<text id> 0 -1 NA` for all cases. Therefore, we do not report results for these two models.

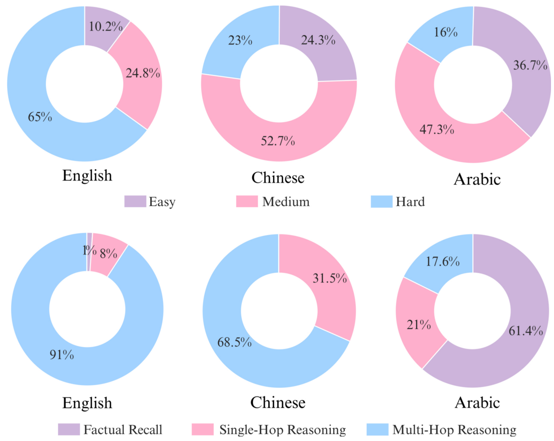


Figure 2: Distribution of difficulty level and reasoning type.

4.3 Overview of Dataset

Figure 2 shows the distribution of difficulty level for all three languages. The English dataset is skewed toward higher-difficulty content, with 65% of questions labeled as Hard, reflecting a greater emphasis on abstraction and multi-step integration. The Chinese dataset shows a more balanced distribution, with 52.7% of questions categorized as Medium, and roughly equal proportions of Easy (24.3%) and Hard (23%) items. In contrast, the Arabic dataset contains a higher proportion of Medium questions (47.3%) and fewer Hard items (16%), suggesting simpler question formats and a stronger emphasis on factual recall.

Figure 2 also shows the distribution of reasoning type for all three languages. The English dataset is overwhelmingly multi-hop in nature, with 91% of questions requiring the integration of multiple medical elements, consistent with case-based diagnostic reasoning. Chinese questions show a more moderate distribution, with 68.5% classified as Multi-hop and 31.5% as Single-hop, indicating a balance between direct inference and integrative tasks. Arabic questions, by contrast, are predominantly based on Factual Recall (61.4%), with lower proportions of Single-hop (21%) and Multi-hop (17.6%) reasoning. This variation in reasoning types across datasets aligns closely with the observed difficulty distributions and reinforces the need for language-specific modeling and evaluation strategies.

In the English dataset, 7.5% of the test set instances were found to be mislabeled. The most frequent issue involved confusion between the Management and Treatment categories, accounting for approximately 60% of all labeling errors. Additional misclassifications included examples such as Diagnosis incorrectly labeled as Management, as well as errors involving Pharmacotherapy, Epidemiology, and Lab/Serum Value Interpretation. In the Chinese dataset, two clinicians independently reviewed the data and identified issues in 5% and 6.5% of instances, respectively. The most common concern was the expression of medical measurement units, which often did not align with standard medical usage. In the Arabic subset, medical review identified issues in approximately 12% of the data. These included outdated medical procedures, corrections that were inapplicable or medically inappropriate, and other medical inaccuracies. Additionally, clinicians flagged a further 5% of the entries for spelling mistakes. Eight error-type misclassifications were also identified, including three physiology-related questions that had been incorrectly categorized under other types. Detailed data splits, basic statistics, and the distribution of error types are provided in Table S2 and Figure S1 in the appendix.

Table 2: Results on MedErrBench-EN.

Models	Det Acc	Loc Acc	Error Correction R-1	BS	BRT
General-purpose LLMs					
gpt-4o	0.596	0.346	0.415	0.428	0.407
gpt-4o-mini	0.664	0.524	0.487	0.498	0.472
Gemini 2.5 Flash Lite	0.567	0.264	0.349	0.362	0.346
Gemini 2.0 Flash	0.514	0.168	0.281	0.294	0.288
Llama3-8b	0.519	0.361	0.266	0.261	0.282
Llama-3.3-70B-Instruct	0.582	0.255	0.369	0.369	0.385
Language-specialized LLMs					
Qwen2.5-7B-Instruct	0.563	0.490	0.372	0.450	0.371
Deepseek-R1 †	0.582	0.577	0.700	0.716	0.681
Deepseek-V3	0.587	0.582	0.703	0.732	0.693
Doubao-1.5 †	**0.779**	**0.774**	**0.766**	**0.783**	**0.773**
ALLAM-7B	0.029	0.014	0.015	0.020	0.014
Medical-domain LLMs					
MedGemma-4b	0.505	0.438	0.511	0.518	0.513
MedGemma-27b	0.543	0.245	0.377	0.390	0.349
HuatuoGPT-o1-7b †	0.574	0.530	0.486	0.475	0.475

4.4 Overall Performance

We evaluated the models on three core tasks across the English, Chinese, and Arabic datasets (Tables 2-4). Detailed results for all evaluation metrics are reported in Tables S3-S5 in the appendix. Overall, Doubao-1.5-thinking-pro, Deepseek-R1, Deepseek-V3 perform better than other models across the languages. Despite being trained on domain-specific data, medical LLMs like MedGemma and HuatuoGPT do not consistently outperform general-purpose models. MedGemma models are trained on medical text, medical QA, EHR, and medical images, while HuatuoGPT leverages medical exam questions. These models focus on domain knowledge and factual recall rather than error detection and correction, which requires broader linguistic reasoning and robustness to noisy clinical text. The Arabic LLM underperformed significantly perhaps since it lacks medical domain adaptation. We tested three different prompts for ALLAM-7B, yielding an average detection accuracy of only 0.083, suggesting that prompt sensitivity alone does not explain the failure.

Table 3: Results on MedErrBench-CN.

Models	Det Acc	Loc Acc	Error Correction R-1	BS	BRT
General-purpose LLMs					
gpt-4o	0.630	0.205	0.265	0.365	0.266
gpt-4o-mini	0.505	0.115	0.244	0.390	0.257
Gemini 2.5 Flash Lite	0.600	0.375	0.448	0.533	0.455
Gemini 2.0 Flash	0.705	0.455	0.569	0.659	0.577
Llama3-8b	0.500	0.320	0.416	0.532	0.483
Llama-3.3-70B-Instruct	0.675	0.380	0.506	0.606	0.509
Language-specialized LLMs					
Qwen2.5-7B-Instruct	0.625	0.570	0.493	0.576	0.462
Deepseek-R1	0.735	0.705	0.802	0.851	0.781
Deepseek-V3	0.650	0.640	**0.833**	**0.873**	**0.806**
Doubao-1.5	**0.750**	**0.735**	0.788	0.835	0.777
ALLAM-7B	0.395	0.340	0.284	0.360	0.286
Medical-domain LLMs					
MedGemma-4b	0.525	0.500	0.549	0.581	0.547
MedGemma-27b	0.605	0.285	0.441	0.537	0.438
HuatuoGPT-o1-7b	0.525	0.275	0.167	0.545	0.530

Table 4: Results on MedErrBench-Ara.

Models	Det Acc	Loc Acc	Error Correction R-1	BS	BRT
General-purpose LLMs					
gpt-4o	0.680	0.320	0.399	0.592	0.414
gpt-4o-mini	0.577	0.175	0.260	0.469	0.292
Gemini 2.5 Flash Lite	0.495	0.268	0.303	0.432	0.318
Gemini 2.0 Flash	0.598	0.299	0.315	0.503	0.332
Llama3-8b	0.371	0.309	0.311	0.324	0.313
Llama-3.3-70B-Instruct	0.557	0.381	0.412	0.454	0.405
Language-specialized LLMs					
Qwen2.5-7B-Instruct	0.536	0.381	0.329	0.473	0.353
Deepseek-R1	**0.711**	**0.505**	0.568	0.756	0.610
Deepseek-V3	0.608	**0.505**	**0.677**	**0.814**	**0.699**
Doubao-1.5	0.670	**0.505**	0.582	0.736	0.583
ALLAM-7B	0.072	0.021	0.045	0.049	0.046
Medical-domain LLMs					
MedGemma-4b	0.454	0.433	0.438	0.450	0.439
MedGemma-27b	0.552	0.240	0.266	0.456	0.286
HuatuoGPT-o1-7b	0.371	0.397	0.302	0.450	0.420

The drop in accuracy from detection to localization is expected, as detection is a binary task, whereas localization is a more complex multi-class task requiring precise error-span identification; localization failures typically arise when an error is detected but its span is misidentified. Overall, localization and correction remain more challenging than detection across all models and languages, as reflected by their consistently lower scores. We did not observe any error type that consistently chal-

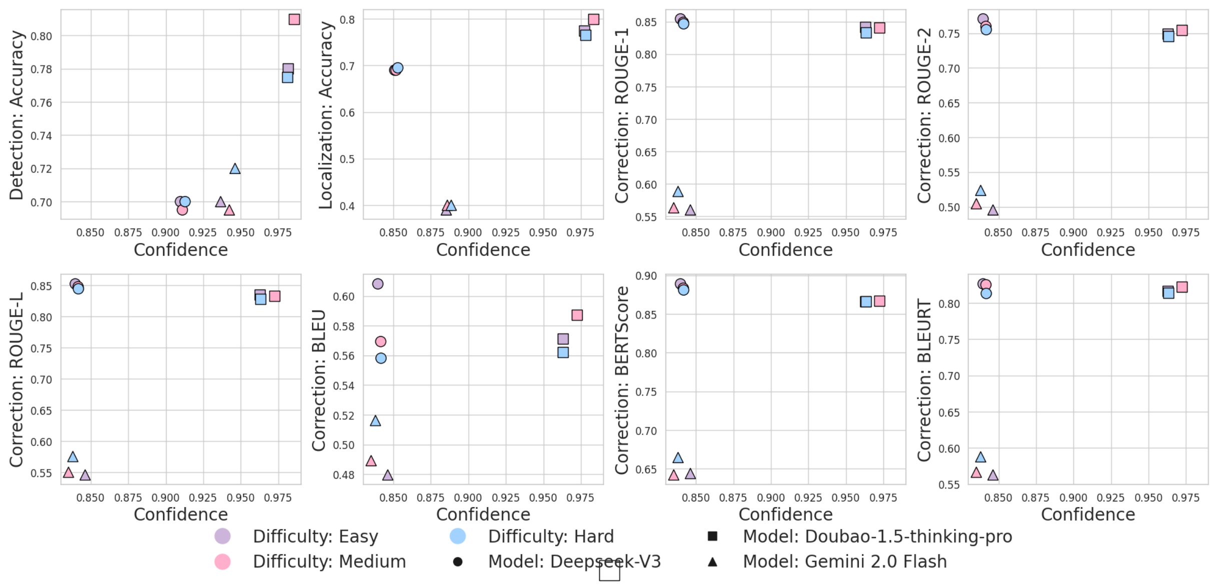


Figure 3: Performance comparison across easy, medium, and hard examples in few-shot learning.

lenged all models. However, Llama3-8B showed weaker performance on Management errors in Arabic, while ALLAM-7B and Qwen2.5-7B-Instruct struggled primarily with Anatomy, Causal Organism/Pathogen, and Diagnosis errors. Llama3-8B exhibits signs of overcorrection in Arabic: frequent NA predictions, combined with the metric's assignment of a score of 1 when both predicted and true labels are NA, artificially inflate performance by increasing true negatives. Meanwhile, models like ALLAM-7B fail on out-of-domain or cross-lingual tasks, highlighting the importance of robust multilingual evaluation. A clinician performed a human evaluation of 50 Chinese samples, yielding average scores of 0.52 for Gemini 2.0 Flash and 0.17 for GPT-4o-mini.

We included three reasoning-oriented models (DeepSeek-R1, Doubao-1.5-Thinking-Pro, HuatuoGPT-o1-7B). Reasoning models achieve a total average score 0.518, compared to 0.451 for non-reasoning models (we selected the top-3 non-reasoning models for comparison). This indicates a consistent advantage for reasoning models under the same evaluation protocol.

4.5 Impact of Providing Examples and Error Type Definitions

Table 5 presents the performance of models under varying configurations of Error Type (ET) and Definition (DEF) availability. Detailed evaluation results are provided in Table S6 in the appendix. Overall, we observe that providing error type definitions consistently improves performance, particularly in the zero-shot setting. For example, in the Deepseek-V3 (Zero-shot) setup, adding definitions (*w DEF* or *w ET & DEF*) boosts error correction metrics compared to the baseline (*w/o ET & DEF*). This highlights that semantic clarity from definitions is beneficial even without structural labels like error types.

Additionally, few-shot configurations consistently outperform their zero-shot counterparts across all models and settings, indicating that in-context examples provide strong guidance for both detection and correction tasks. Interestingly, while providing both ET and DEF is generally helpful,

Table 5: Performance comparison of models under different error type Conditions. "ET" and "DEF" indicate error types and definitions, respectively.

	Det Acc	Loc Acc	Error Correction R-1	BS	BRT
Deepseek-V3 (Zero-shot)					
w/o ET & DEF	0.690	0.645	0.660	0.735	0.599
w/o DEF	0.625	0.610	0.731	0.795	0.685
w ET & DEF	0.650	0.640	0.730	0.794	0.684
Deepseek-V3 (Few-shot)					
w/o ET & DEF	0.720	0.690	0.695	0.767	0.639
w/o DEF	0.710	0.705	0.736	0.796	0.684
w ET & DEF	0.715	0.715	0.763	0.821	0.705
Doubao-1.5-thinking-pro (Zero-shot)					
w/o ET & DEF	0.695	0.640	0.637	0.727	0.607
w/o DEF	0.730	0.710	0.673	0.751	0.646
w ET & DEF	0.750	0.725	0.669	0.728	0.636
Doubao-1.5-thinking-pro (Few-shot)					
w/o ET & DEF	0.735	0.695	0.707	0.765	0.651
w/o DEF	0.765	0.750	0.729	0.777	0.671
w ET & DEF	0.775	0.765	0.699	0.753	0.665

the isolated impact of error types alone (i.e., *w/o DEF*) can be inconsistent. In some few-shot settings (e.g., Doubao-1.5-thinking-pro), adding ETs without definitions slightly improves detection but may reduce correction performance, suggesting potential cognitive overload or prompt misalignment.

The performance divergence between Deepseek-V3 and Doubao under different prompt settings may stem from their architectural and training differences. Deepseek-V3 appears more responsive to structured definitions and error-type annotations, possibly due to multilingual and multitask training, which enhances its generalization across abstract prompt forms. In contrast, Doubao demonstrates stronger alignment with example-driven prompts, but exhibits sensitivity or degradation when additional structured elements (e.g., both ET and DEF) are introduced, especially in few-shot scenarios. This suggests that prompt design must consider model-specific alignment and interpretability characteristics, as misaligned guidance may counterintuitively hinder performance.

4.6 Impact of Providing Example Difficulty Levels in Few-shot Learning Settings

Figure 3 illustrates the relationship between example difficulty levels and model performance in few-shot learning across three baseline models with good and stable performance on Chinese dataset: Deepseek-V3 (circle), Doubao-1.5-thinking-pro (square), and Gemini 2.0 Flash (triangle). For Doubao-1.5-thinking-pro, the performance trend across difficulty levels follows a consistent pattern of Medium > Easy > Hard across nearly all metrics. In contrast, Deepseek-V3 shows a different behavior. For correction tasks, the model clearly follows Easy > Medium > Hard. For localization, Deepseek-V3 performs similarly across all difficulty levels, with Hard examples even slightly outperforming the others. Gemini 2.0 Flash, however, displays an opposite trend to Doubao-1.5-thinking-pro and Deepseek-V3. For most metrics, the order is Hard > Medium > Easy, with overall lower performance. This may imply that the task is relatively more difficult for Gemini 2.0 Flash, and that harder in-context examples are more informative and helpful for its learning. In terms of confidence scores, Doubao-1.5-thinking-pro exhibits the highest confidence across all difficulty levels, noticeably exceeding both Deepseek-V3 and Gemini 2.0 Flash. This suggests that Doubao-1.5-thinking-pro is more certain in its predictions, although confidence does not always correlate perfectly with correctness, especially in challenging scenarios.

4.7 Analysis of Knowledge vs. Scenario-based Data

Figure 4 shows the evaluation results of six LLMs on an Arabic dataset, segmented by knowledge-based and scenario-based clinical notes. The goal is to assess whether LLMs can capture language-independent conceptions. Most models perform better on scenario-based tasks, suggesting a reliance on contextual pattern recognition rather than robust internal medical knowledge in Arabic. An exception is Llama3-8b, which slightly performs better on knowledge-based tasks. This may indicate that its learned representations are more tightly coupled with language-independent factual knowledge, allowing it to resist some misconceptions when direct medical facts are queried. Qwen2.5-7B-Instruct exhibits the largest performance gap which suggests that the model is over-reliant on surface-level patterns and instruction-following heuristics, making it more vulnerable to reproducing misconceptions in structured factual queries, especially in low-resource languages like Arabic. We also evaluate the cross-lingual generalization, please refer to the appendix section K.1.

5 Conclusion and Discussion

In this work, we present a novel multilingual benchmark dataset for medical error detection and correction, grounded in expert-defined error categories and validated across English, Chinese, and Arabic. The work addresses key limitations in existing resources by providing diverse, high-quality annotations that reflect the complexity of real-world medical errors in multiple languages. Through extensive evaluation of various LLMs, we reveal the current challenges in automated medical error understanding and emphasize the importance of medically informed and language-specific approaches. Our dataset and analysis lay a solid foundation for advancing research in clinical NLP focused on improving patient safety. In addition to detection, localization, and correction, our dataset supports tasks such as error classification, key concept extraction, difficulty assessment, and reasoning type classification, enabling new avenues for fine-grained clinical reasoning in NLP models. Our dataset in this component focuses on errors in professional medical knowledge. In future work, we

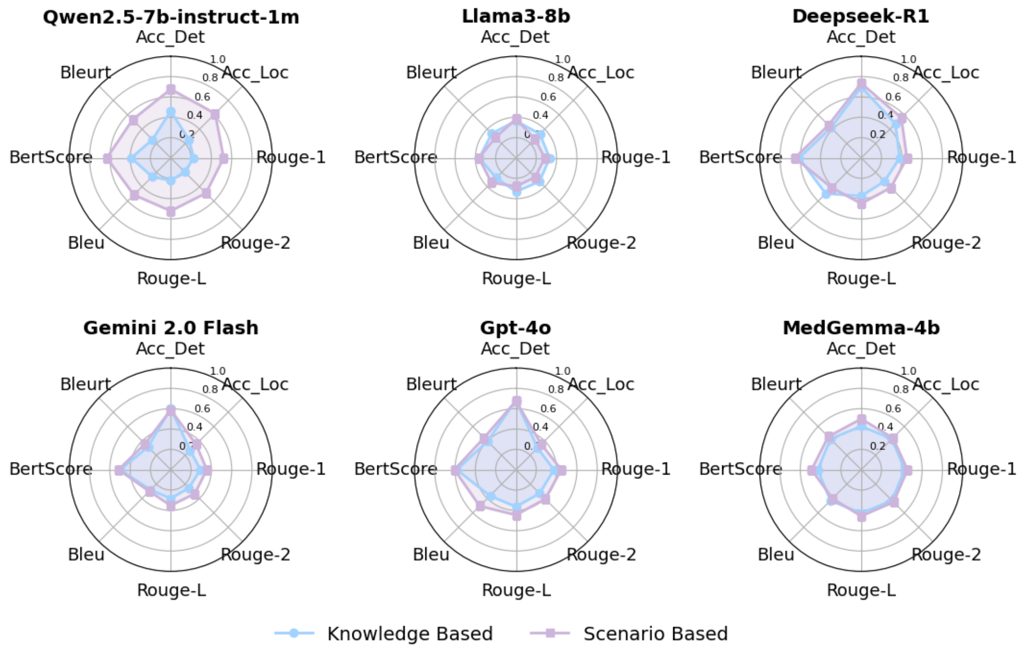


Figure 4: Comparison of models based on knowledge-based and description-based evaluation.

plan to expand it with real clinical data to further enhance its coverage and completeness. Our future efforts will also focus on (1) increasing the overall scale and diversity of the corpus; (2) developing multi-agent systems to improve LLM performance on clinical tasks; (3) advancing evaluation methodologies for medical error detection and correction; (4) integrating severity stratification and assessing harm-reduction strategies.

6 Ethical Issues

This work introduces a benchmark for detecting and correcting medical errors in clinical text. As this task involves high-stakes medical information, several ethical concerns must be considered. First, all data used in this benchmark are de-identified and derived from publicly available sources, ensuring no patient-identifiable information is included. Second, while our benchmark supports research progress, we caution against deploying automatic correction systems in real clinical practice without professional supervision, as incorrect corrections may lead to harmful outcomes. Third, biases in medical datasets may impact model performance across populations. Future work should validate robustness and fairness across diverse medical settings. Finally, we advocate for human-in-the-loop approaches and transparent reporting when developing medical NLP systems based on this benchmark.

Limitations

We note three limitations of the present study. First, the proposed dataset does not include explicit annotations for severity levels or equity-related dimensions. Our primary objective is to establish a robust multilingual foundation for medical error detection and correction, rather than to exhaustively characterize downstream medical risk or fairness properties. In this sense, the dataset represents an initial step toward such analyses, and is, to our knowledge, only the second publicly available resource after MEDEC that addresses medical errors in a multilingual setting. Future work can build upon this foundation by incorporating severity stratification and equity-aware annotations. Second, the Arabic portion of the dataset remains under active expansion. Due to the limited availability of high-quality publicly accessible medical data in this low-resource language, the current Arabic subset is smaller and less diverse than those of higher-resource languages. We plan to continue data collection and curation efforts to further expand and balance the dataset across languages, which may improve both coverage and robustness in future iterations. Third, measuring human expert performance (and inter-annotator agreement) would help contextualize the difficulty of MedErrBench. However, establishing a reliable human baseline for MedErrBench is non-trivial: the benchmark covers 10 diverse error types, and a fair comparison would require a properly designed clinical annotation study including recruiting qualified clinicians, potentially those with auditing backgrounds, standardized training/calibration across annotators, and clear adjudication protocols. That being said, medical error detection and correction is not necessarily a “routine task” that clinicians do. The process usually involves extensive manual review and the use of automated systems to ensure patient safety, due to the high burden nature of clinical documentation. The future work should focus on the design of appropriate reader studies to measure the human expert baseline for error detection and correction.

Acknowledgments

This work was supported by the Meem Foundation, the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010, the Center for Brain and Health, funded by Tamkeen under NYU Abu Dhabi Research Institute grant CG012, and the Center for Interdisciplinary Data Science & AI (CIDSAI), funded by Tamkeen under the NYUAD Research Institute Award CG016. The research was carried out on the High Performance Computing resources at New York University Abu Dhabi (Jubail). We are grateful to Dr. Xinwei Hou and Dr. Zijie Fang for their support and contributions to data annotation.

References

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen-Yildiz. 2024. Overview of the mediqa-corr 2024 shared task on medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 596–603.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. MEDEC: A benchmark for medical error detection and correction in clinical notes. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22539–22550.

Ehsan Ahsani-Estahbanati, Vladimir Sergeevich Gordeev, and Leila Doshmangir. 2022. Interventions to reduce the incidence of medical error and its financial burden in health care systems: A systematic review of systematic reviews. *Frontiers in medicine*, 9:875426.

Alibaba DAMO Academy. 2024. Qwen2.5-7b-instruct-1m. `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`. Accessed: 2025-08-01.

Renad Alzghoul, Ayaabdelhaq Ayaabdelhaq, Abdulrahman Tabaza, and Ahmad Altamimi. 2024. Cld-mec at mediqa-corr 2024 task: Gpt-4 multi-stage clinical chain of thought prompting for medical errors detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 537–556.

Fauzia Anjum, Brigadier Raffi Ud Din, and Saira Ashraf. 2024. Patient safety and quality improvement: Reducing medical errors in healthcare. *Multidisciplinary Journal of Healthcare (MJH)*, 1(2):13–23.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, and 1 others. 2025. Allam: Large language models for Arabic and English. In *ICLR*.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *Preprint*, arXiv:2412.18925.

Jean-Philippe Corbeil. 2024. Iryonlp at MEDIQA-CORR 2024: Tackling the medical error detection & correction task on the shoulders of medical agents. In *Proceedings of the 6th Clinical Natural Language Processing Workshop, ClinicalNLP@NAACL 2024, Mexico City, Mexico, June 21, 2024*, pages 570–580.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E Shamout. 2025. Medarabiq: Benchmarking large language models on Arabic medical tasks. *arXiv preprint arXiv:2505.03427*.

DeepSeek AI. 2024a. Deepseek-r1. `https://huggingface.co/deepseek-ai/deepseek-llm-7b-base`. Accessed: 2025-08-01.

DeepSeek AI. 2024b. Deepseek-v3. `https://huggingface.co/deepseek-ai/deepseek-vl-7b`. Accessed: 2025-08-01.

Doubao Team, Volcano Engine. 2024. Doubao-1.5-thinking-pro. `https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-1-5-thinking-pro`. Accessed: 2025-08-01.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. In *Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part III*, volume 14304, pages 69–80.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7489–7501.

Google Research. 2025a. Medgemma: Open multimodal models for medical ai. https://research.google/blog/medgemma-our-most-capable-open-models-for\-health-ai-development. Accessed: 2025-08-01.

Google Research. 2025b. Medgemma: our most capable open models for health ai. https://research.google/blog/medgemma-our-most-capable-open-models-for\-health-ai-development. Accessed: 2025-08-01.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Satya Kesav Gundabathula and Sriram R. Kolar. 2024. Promptmind team at MEDIQA-CORR 2024: Improving clinical text correction with error categorization and LLM ensembles. In *Proceedings of the 6th Clinical Natural Language Processing Workshop, ClinicalNLP@NAACL 2024, Mexico City, Mexico, June 21, 2024*, pages 367–373.

Naoto Iwase, Hiroki Okuyama, and Junichiro Iwasawa. 2025. Medrect: A medical reasoning benchmark for error correction in clinical texts. *arXiv preprint arXiv:2511.00421*.

Suramya Jadhav, Abhay Shanbhag, Sumedh Joshi, Atharva Date, and Sheetal Sonawane. 2024. Maven at mediqa-corr 2024: Leveraging rag and medical llm for error detection and correction in medical notes. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 374–381.

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*, page 6421.

Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Ranran Haoran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. 2024a. Evaluating llms at detecting errors in llm responses. *2024 Conference on Language Modeling*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024b. When can llms *Actually* correct their own mistakes? A critical survey of self-correction of llms. *Trans. Assoc. Comput. Linguistics*, 12:1417–1440.

Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for language learners using example-based grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7176–7187.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*, pages 205–219.

Bernardo Magnini, Marco Madeddu, Michele Resta, Roberto Zanoli, Martin Cimmino, Paolo Albano, and Viviana Patti. 2025. A leaderboard for benchmarking llms on italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 636–646.

Meta AI. 2024. Llama 3: Open foundation and instruction models. https://github.com/meta-llama/llama3. Accessed: 2025-08-01.

Abu-Daoud Mouath, Kharouf Leen, El Hajj Omar, El Samad Dana, Al-Omari Mariam, Mallat Jihad, Saleh Khaled, Habash Nizar, and E. Shamout Farah. 2026. Medarabench: Large-scale Arabic medical question answering dataset and benchmark. *arXiv preprint arXiv:2602.01714*.

David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. 2021. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the “big three”. *Diagnosis*, 8(1):67–84.

Jingwei Ni, Minjing Shi, Dominik Stammbach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable LLM annotators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1890–1912.

OpenAI. 2024a. Gpt-4o-mini model card. https://research.google/blog/medgemma-our-most-capable-open-models-for\-health-ai-development. Accessed: 2025-08-01.

OpenAI. 2024b. Gpt-4o technical overview. https://platform.openai.com/docs/models#gpt-4o. Accessed: 2025-08-01.

Juan Pajaro, Edwin Puertas, David Villate, Laura Estrada, and Laura Tinjaca. 2024. Verbanexai at mediqa-corr: Efficacy of gru with biowordvec and clinicalbert in error correction in clinical notes. In *Proceedings of the 6th clinical natural language processing workshop*, pages 461–469.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: *Surveying the Landscape of Diverse Automated Correction Strategies*. *Trans. Assoc. Comput. Linguistics*, 12:484–506.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Guangyue Peng, Wei Li, Wen Luo, and Houfeng Wang. 2025. Encode errors: Representational retrieval of in-context demonstrations for multilingual grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21166–21180.

Swati Rajwal, Eugene Agichtein, and Abeed Sarker. 2024. Em_mixers at mediqa-corr 2024: Knowledge-enhanced few-shot in-context learning for medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*.

Eduardo Reis, Mohamed Abdelaal, and Carsten Binnig. 2024. Generalizable data cleaning of tabular data in latent space. *Proceedings of the VLDB Endowment*, 17(13):4786–4798.

Nadia Saeed. 2024. Medifact at MEDIQA-CORR 2024: Why AI needs a human touch. In *Proceedings of the 6th Clinical Natural Language Processing Workshop, ClinicalNLP@NAACL 2024, Mexico City, Mexico, June 21, 2024*, pages 346–352.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892.

Vinay Setty. 2024. Factcheck editor: Multilingual text editor with end-to-end fact-checking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2744–2748.

Hamid Soori. 2024. Errors in medical procedures. In *Errors in Medical Science Investigations*, pages 205–224.

Felix Stahlberg and Shankar Kumar. 2024. Synthetic data generation for low-resource grammatical error correction with tagged corruption models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2024, Mexico City, Mexico, June 20, 2024*, pages 11–16.

Runchu Tian, Yining Ye, Yujia Qin, Xin Cong, Yankai Lin, Yinxu Pan, Yesai Wu, Haotian Hui, Weichuan Liu, Zhiyuan Liu, and Maosong Sun. 2024. Debugbench: Evaluating debugging capability of large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4173–4198.

YunDa Tsai, Mingjie Liu, and Haoxing Ren. 2024. Rtlfixer: Automatically fixing rtl syntax errors with large language model. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pages 1–6.

Airat Valiev and Elena Tutubalina. 2024. Hse nlp team at mediqa-corr 2024 task: In-prompt ensemble with entities and knowledge graph for medical error correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*.

Justin Vasselli and Taro Watanabe. 2023. A closer look at k-nearest neighbors grammatical error correction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*, pages 220–231.

Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason Cheung, Teng Zhang, and Honghan Wu. 2024. Knowlab_aimed at mediqa-corr 2024: Chain-of-though (cot) prompting strategies for medical error detection and correction. In *proceedings of the 6th clinical natural language processing workshop*, pages 353–359.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Ying Shen, Peng Xing, Zishan Xu, Guo Cheng, and 1 others. 2025. Excgec: A benchmark for edit-wise explainable chinese grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25678–25686.

Min Zeng, Jiexin Kuang, Mengyang Qiu, Jayoung Song, and Jungyeul Park. 2024. Evaluating prompting strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 6426–6430.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

A Taxonomy of Medical Error Types

Table S1 shows the classification of medical error types, the definitions of error types and representative examples.

Table S1: Classification of medical error types, including definitions and representative examples.

Error Type	Definition	Example Scenario
Diagnosis	Failure to correctly identify the underlying condition based on medical presentation	Interpreting myocardial infarction as GERD despite ECG abnormalities and chest pain
Management	Inappropriate non-pharmacologic, non-surgical clinical decision such as observation, monitoring, or disposition	Advising "observation" in a patient with acute ST-elevation myocardial infarction
Treatment	Inappropriate definitive intervention (surgical, procedural, or pharmacologic); distinct from general management	Recommending testicular biopsy in suspected torsion, delaying emergency surgery
Pharmacotherapy	Incorrect drug selection, dosage, route, timing, interaction, or duration	Prescribing heparin in heparin-induced thrombocytopenia (HIT)
Causal Organism / Pathogen	Misidentification of the causative microorganism in infectious disease	Attributing syphilis to *Pseudomonas* instead of *Treponema pallidum*
Lab Value Interpretation	Misreading or misapplying diagnostic thresholds, reference ranges, or derived values	Interpreting HbA1c of 5.2% as diagnostic for diabetes
Physiology	Misconception or misinterpretation of physiological principles (e.g., ECG, PFT, etc.)	Reading an irregular rhythm without P waves as sinus rhythm rather than atrial fibrillation
Histology	Misinterpretation of tissue morphology, cellular structures, or microscopic patterns	Identifying psammoma bodies from papillary thyroid carcinoma as colon adenocarcinoma features
Anatomy	Errors in anatomical structure, relation, or spatial understanding	Describing the pancreas as an intraperitoneal organ
Epidemiology	Misuse of statistical tools or misstatement of incidence, prevalence, or risk factors	Claiming colorectal cancer is more prevalent than breast cancer among women globally

B Data Statistics

B.1 Language-wise data split and statistics

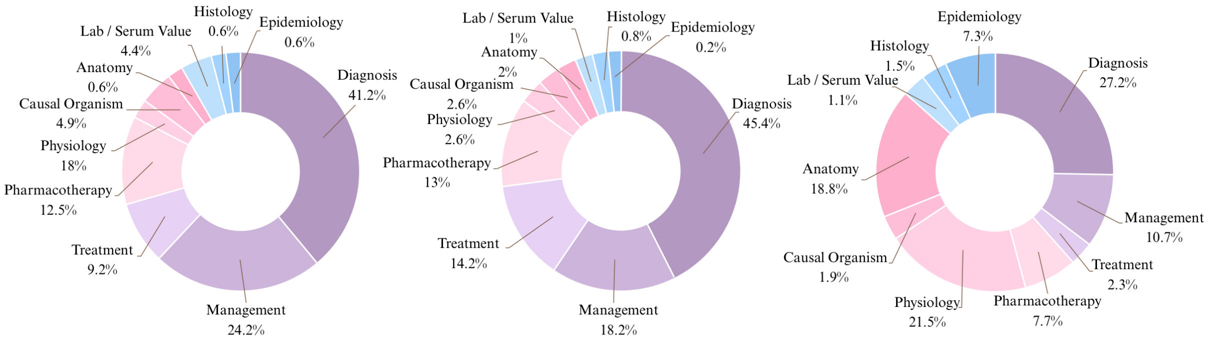
Table S2 presents the official data splits and basic statistics of the dataset, a multilingual clinical benchmark covering English, Chinese, and Arabic. For each language, we report the number of instances in the training, validation, and test sets, along with the total number of samples. The English dataset contains 1,024 instances, the Chinese dataset includes 1,000, and the Arabic dataset consists of 482. We also provide average, maximum, and minimum input lengths, as well as the number of samples with and without factual errors in each split.

Table S2: Summary Statistics of the Dataset.

Language	Metric	Train	Validation	Test	Total
English	Num.	708	108	208	1024
	Avg len	755.7	604.3	872.1	763.9
	Max len	1594	1250	1396	1594
	Min len	220	101	63	63
	w errors	354	54	104	512
	w/o errors	354	54	104	512
Chinese	Num.	700	100	200	1000
	Avg len	97.4	96.3	98	97.3
	Max len	269	262	191	269
	Min len	32	41	40	32
	w errors	350	50	100	500
	w/o errors	350	50	100	500
Arabic	Num.	334	51	97	482
	Avg len	156	168.2	165.6	159.2
	Max len	457	429	501	501
	Min len	31	48	37	31
	w errors	179	29	53	261
	w/o errors	155	22	44	221

B.2 Distribution of medical error types

Figure S1 illustrates the distribution of ten common medical error types in the dataset across three languages: English, Chinese, and Arabic. Each donut chart represents the relative proportions of error categories, including Diagnosis, Management, Treatment, Pharmacotherapy, Physiology, Causal Organism, Anatomy, Lab/Serum Value, Histology, and Epidemiology. In the English subset (a), Diagnosis (41.2%) and Management (24.2%) are the most prevalent error types, followed by Physiology (18.0%) and Pharmacotherapy (12.5%). The



(a) English (b) Chinese (c) Arabic

Figure S1: Distribution of Error Types by Language

Chinese subset (b) shows a stronger concentration in Diagnosis errors (45.4%), with moderate representations of Mangement (18.2%), Treatment (14.2%), Pharmacotherapy (13.0%), and notably fewer errors related to Physiology and Anatomy (2.6% each). In contrast, the Arabic subset (c) presents a more balanced distribution, where Diagnosis (27.2%) remains the largest category, but Physiology (21.5%) and Anatomy (18.8%) are more prominent, while categories such as Treatment (2.3%) and Causal Organism (1.9%) are less frequent. These distributions highlight linguistic and potentially systemic differences in error typologies across multilingual medical corpora.

C Benchmark Tasks

The proposed MedErrBench dataset supports three core tasks for benchmarking medical error understanding:

- Error Detection: Determine whether a given clinical note contains an error. This is formulated as a binary classification task distinguishing between error-free and erroneous notes.

- Error Localization: Identify which specific sentence within the clinical note contains the error. This task focuses on sentence-level localization rather than token-level span extraction, aligning with how clinicians typically review clinical documentation.

- Error Correction: Generate a revised version of the clinical note with the error corrected. This task requires contextual understanding and medical knowledge to produce plausible and medically valid corrections.

Beyond the three primary tasks supported by MedErrBench, the dataset includes annotations of difficulty level, reasoning type, and important medical terms, which enable flexible task customization. For instance, MedErrBench can be used for error classification by assigning each clinical note to a predefined error category curated by expert clinicians. In addition, annotated key clinical terms support token-level localization, facilitating the development of alternative task formulations.

D Experimental Settings

During dataset construction, LLMs used for the English, Chinese, and Arabic datasets were Gemini 2.0 Flash, DeepSeek-V3-0324, and Gemini-1.5-pro, respectively. In regard to the models, Qwen2.5-7B-Instruct was used and all the hyperparameters were set to the default values[5]. ALLAM-7B was run with hyperparameters: max_new_tokens=256, top_k=50, temperature=0.7, top_p=0.9, and do_sample=True. JAIS-adapted-13b-chat was run wiht max_new_tokens=256, do_sample=True, temperature=0.7, top_k = 50 and top_p=0.95. Doubao-1.5-Thinking-Pro was run with max_tokens=512, with all other parameters (including temperature and top_p) set to default values[6]. DeepSeek-V3 and DeepSeek-R-1 were run with hyperparameters: max_tokens=512, while all other parameters were set to their default values[7]. Meta-Llama-3.1-8B-Instruct was accessed via the Novita InferenceClient with hyperparameters: max_tokens=256, temperature=0.2, do_sample=False. HuatuoGPT-o1-8B was run locally with PyTorch with hy-

[5] https://www.alibabacloud.com/help/en/model-studio/use-qwen-by-calling-api
[6] https://www.volcengine.com/docs/82379/1494384
[7] api-docs.deepseek.com/api/create-chat-completion

perparameters: max_new_tokens=1024, temperature=0.2, do_sample=False, and pad_token_id set to the tokenizer's eos_token_id. GPT-4o, GPT-4o-mini, Gemini 2.0 Flash, Gemini 2.5 Flash Lite, and MedGemma-4B, MedGemma-27B are all used with default parameters. We use BLEURT with BLEURT-20 model[8] and BERTScore with the deberta-xlarge-mnli checkpoint[9]. For Chinese data, ROUGE evaluation is performed using the ROUGE-chinese package[10], which provides tokenization and evaluation methods specifically designed for Chinese text.

E Evaluation Metrics

ROUGE-1 (Lin, 2004) measures unigram overlap between the generated and reference texts, while ROUGE-2 extends this to bigram overlap. ROUGE-L (Lin, 2004) captures structural similarity based on the longest common subsequence and ROUGE-SU (Lin, 2004) incorporates both unigrams and skip-bigrams with a maximum skip distance, offering a balance between flexibility and structure. BERTScore (Zhang et al., 2020) evaluates semantic similarity using contextual embeddings from a pre-trained BERT model. BLEURT (Sellam et al., 2020) combines pre-trained language models with human-annotated data to assess the fluency and adequacy of the generated text.

F Detailed Overall Performance Across Languages

We report the complete results on all evaluation metrics, including detection accuracy, localization accuracy, ROUGE-1, ROUGE-2, ROUGE-L, BLEU, BERTScore, and BLEURT, evaluated on MedErrBench-EN (Table S3), MedErrBench-CN (Table S4), and MedErrBench-Ara (Table S5).

G Detailed Experimental Results on the Impact of Providing Example Difficulty Levels in Few-Shot Learning Settings

Table S6 presents the complete results across all evaluation metrics, including detection accuracy, localization accuracy, ROUGE-1/2/L, BLEU, BERTScore, and BLEURT, for analyzing the impact of providing example difficulty levels in few-shot learning settings.

[8] https://github.com/google-research/bleurt

[9] https://huggingface.co/microsoft/deberta-xlarge-mnli

[10] https://pypi.org/project/rouge-chinese/

H Prompts of Data Construction

We utilized large language models to help convert medical examination materials into comprehensive medical stories across English, Chinese, and Arabic. The language-specific prompts used for generation are illustrated in Figure S2, Figure S3 and S4.

I Examples for Error Insertion in MedErrBench Dataset

Figure S6, S7 and S8 are examples for error insertion in the proposed MedErrBench dataset.

J Illustrative Examples of the Ten Error Types Defined in the MedErrBench Dataset

Figure S9, S10 and S11 are the illustrative examples of the ten error types defined in the MedErrBench dataset.

K Performance of Existing Error Detection and Correction Methods on MedErrBench

The MEDIQA-CORR 2024 shared task (Abacha et al., 2024) focused on detecting and correcting multiple types of medical errors in clinical texts and attracted participation from seventeen teams. We attempted to benchmark representative systems from this shared task on MedErrBench, our trilingual dataset. However, due to the unavailability or inactivity of most public GitHub repositories, only a limited subset of methods could be successfully reproduced and evaluated. Specifically, we benchmarked CLD-MEC (Alzghoul et al., 2024), Maven (Jadhav et al., 2024), and VerbaNexAI (Pajaro et al., 2024), including both VerbaNexAI-GRU and VerbaNexAI-ClinicalBERT.

The detailed results are reported in Table S7. Overall, these task-specific error detection and correction models underperform compared to state-of-the-art language-specialized LLMs, suggesting that large pretrained models with strong multilingual and generative capabilities may generalize more effectively across diverse error types and languages than systems optimized for narrow task formulations.

Maven relies on a Retrieval-Augmented Generation (RAG) pipeline that is explicitly designed for English medical text. Its retrieval component uses an English-only knowledge base. When applied

Table S3: Results on MedErrBench-EN.

Models	Detection Accuracy	Localization Accuracy	Error Correction					
			ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
General-purpose LLMs								
gpt-4o	0.596	0.346	0.415	0.365	0.403	0.329	0.428	0.407
gpt-4o-mini	0.664	0.524	0.487	0.441	0.477	0.409	0.498	0.472
Gemini 2.5 Flash Lite	0.567	0.264	0.349	0.307	0.340	0.279	0.362	0.346
Gemini 2.0 Flash	0.514	0.168	0.281	0.231	0.268	0.186	0.294	0.288
Llama3-8b	0.519	0.361	0.266	0.225	0.261	0.219	0.261	0.283
Llama-3.3-70B-Instruct	0.582	0.255	0.369	0.328	0.356	0.292	0.369	0.385
Language-specialized LLMs								
Qwen2.5-7B-Instruct	0.563	0.490	0.372	0.334	0.345	0.381	0.450	0.371
Deepseek-R1	0.582	0.577	0.700	0.612	0.682	0.551	0.716	0.681
Deepseek-V3	0.587	0.582	0.703	0.626	0.687	0.569	0.732	0.693
Doubao-1.5	**0.779**	**0.774**	**0.766**	**0.707**	**0.752**	**0.662**	**0.783**	**0.773**
ALLAM-7B	0.029	0.014	0.015	0.014	0.015	0.286	0.020	0.014
Medical-domain LLMs								
MedGemma-4b	0.505	0.438	0.511	0.499	0.508	0.489	0.518	0.513
MedGemma-27b	0.543	0.245	0.377	0.337	0.369	0.305	0.390	0.349
HuatuoGPT-o1-7b	0.574	0.530	0.486	0.466	0.485	0.446	0.475	0.475

Table S4: Results on MedErrBench-CN.

Models	Detection Accuracy	Localization Accuracy	Error Correction					
			ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
General-purpose LLMs								
gpt-4o	0.630	0.205	0.265	0.169	0.247	0.148	0.365	0.266
gpt-4o-mini	0.505	0.115	0.244	0.145	0.223	0.130	0.390	0.257
Gemini 2.5 Flash Lite	0.600	0.375	0.448	0.401	0.439	0.310	0.533	0.455
Gemini 2.0 Flash	0.705	0.455	0.569	0.510	0.557	0.405	0.659	0.577
Llama3-8b	0.500	0.320	0.416	0.251	0.377	0.168	0.532	0.483
Llama-3.3-70B-Instruct	0.675	0.380	0.506	0.459	0.500	0.438	0.606	0.509
Language-specialized LLMs								
Qwen2.5-7B-Instruct	0.625	0.570	0.493	0.429	0.492	0.331	0.576	0.462
Deepseek-R1	0.735	0.705	0.802	0.708	0.799	0.491	0.851	0.781
Deepseek-V3	0.650	0.640	**0.833**	**0.741**	**0.830**	**0.540**	**0.873**	**0.806**
Doubao-1.5	**0.750**	**0.735**	0.788	0.709	0.784	0.508	0.835	0.777
ALLAM-7B	0.395	0.340	0.284	0.260	0.283	0.080	0.360	0.286
Medical-domain LLMs								
MedGemma-4b	0.525	0.500	0.549	0.528	0.545	0.500	0.581	0.547
MedGemma-27b	0.605	0.285	0.441	0.386	0.430	0.270	0.537	0.438
HuatuoGPT-o1-7b	0.525	0.275	0.167	0.056	0.167	0.169	0.545	0.530

to Arabic or Chinese clinical texts, the retrieval mechanism fails to identify semantically relevant context, as embeddings derived from non-English inputs do not align well with English documents in the knowledge base. As a result, Maven produces irrelevant or incorrect disease or pathogen predictions and cannot be reliably evaluated on the Chinese or Arabic subsets of MedErrBench.

In addition, both VerbaNexAI-GRU and VerbaNexAI-ClinicalBERT do not produce localization or correction outputs, as they are fundamentally classification-based models rather than generative systems. Consequently, they are only partially comparable to generative LLM-based approaches in the full error detection and correction setting.

K.1 Cross-lingual Generalization

To assess the effectiveness of machine-translated multilingual data for error detection, localization, and correction, we translated the Chinese dataset into English and Arabic. Chinese was selected arbitrarily as the source language and not for any specific preference, but to test whether linguistic representation differences across languages and cross-lingual variation alone affect model performance. The results are shown in Figure S5. We observe that across all three tasks, most models experience a performance drop to varying degrees on the translated English and Arabic datasets, with the decline being particularly pronounced on the Arabic data, especially with larger gaps observed in localization accuracy and average correction scores. This is mainly due to the substantial structural differences of the language and insufficient training data. In particular, translated data fails to capture the unique expressions and error patterns specific to Arabic, making it difficult for models to effectively transfer learning. Therefore, relying solely on translated

Table S5: Results on MedErrBench-Ara.

Models	Detection Accuracy	Localization Accuracy	Error Correction ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
General-purpose LLMs								
gpt-4o	<u>0.680</u>	0.320	0.399	0.357	0.393	0.321	0.592	0.414
gpt-4o-mini	0.577	0.175	0.260	0.212	0.250	0.164	0.469	0.292
Gemini 2.5 Flash Lite	0.495	0.268	0.303	0.280	0.299	0.253	0.432	0.318
Gemini 2.0 Flash	0.598	0.299	0.315	0.281	0.308	0.251	0.503	0.332
Llama3-8b	0.371	0.309	0.311	0.304	0.310	0.296	0.324	0.313
Llama-3.3-70B-Instruct	0.557	0.381	0.412	0.385	0.406	0.364	0.454	0.405
Language-specialized LLMs								
Qwen2.5-7B-Instruct	0.536	0.381	0.329	0.298	0.327	0.348	0.473	0.353
Deepseek-R1	**0.711**	**0.505**	0.568	0.500	0.564	<u>0.457</u>	<u>0.756</u>	<u>0.610</u>
Deepseek-V3	0.608	**0.505**	**0.677**	**0.628**	**0.675**	**0.592**	**0.814**	**0.699**
Doubao-1.5	0.670	**0.505**	<u>0.582</u>	<u>0.510</u>	<u>0.574</u>	0.452	0.736	0.583
ALLAM-7B	0.072	0.021	0.045	0.044	0.045	0.219	0.049	0.046
Medical-domain LLMs								
MedGemma-4b	0.454	0.433	0.438	0.436	0.438	0.436	0.450	0.439
MedGemma-27b	0.552	0.240	0.266	0.220	0.257	0.185	0.456	0.286
HuatuoGPT-o1-7b	0.371	0.397	0.351	0.260	0.329	0.158	0.450	0.420

Table S6: Performance comparison of models under different error type Conditions. “ET" and “DEF" indicate error types and definitions, respectively.

	Detection Accuracy	Localization Accuracy	Error Correction ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
Deepseek-V3 (Zero-shot)								
w/o ET & DEF	0.690	0.645	0.660	0.490	0.659	0.192	0.735	0.599
w/o DEF	0.625	0.610	0.731	0.564	0.726	0.227	0.795	0.685
w ET & DEF	0.650	0.640	0.730	0.574	0.724	0.228	0.794	0.684
Deepseek-V3 (Few-shot)								
w/o ET & DEF	0.720	0.690	0.695	0.535	0.695	0.249	0.767	0.639
w/o DEF	0.710	0.705	0.736	0.561	0.728	0.215	0.796	0.684
w ET & DEF	0.715	0.715	0.763	0.596	0.758	0.234	0.821	0.705
Doubao-1.5-thinking-pro (Zero-shot)								
w/o ET & DEF	0.695	0.640	0.637	0.503	0.632	0.166	0.727	0.607
w/o DEF	0.730	0.710	0.673	0.521	0.665	0.164	0.751	0.646
w ET & DEF	0.750	0.725	0.669	0.531	0.660	0.171	0.728	0.636
Doubao-1.5-thinking-pro (Few-shot)								
w/o ET & DEF	0.735	0.695	0.707	0.553	0.695	0.248	0.765	0.651
w/o DEF	0.765	0.750	0.729	0.568	0.716	0.260	0.777	0.671
w ET & DEF	0.775	0.765	0.699	0.535	0.687	0.209	0.753	0.665

data cannot meet the demands of high-quality multilingual models. It is crucial to construct authentic, diverse, and high-quality native multilingual datasets, so that the models deeply understand the characteristics of different languages, improve their capabilities in fine-grained tasks like localization and correction, and thereby enhance cross-lingual generalization performance.

Prompt Used in English Dataset Construction

The following is a medical narrative about a patient. You are a skilled medical doctor reviewing the clinical text. The text is either correct or contains one error. The text has one sentence per line. Each line starts with the sentence ID, followed by a pipe character then the sentence to check.

This text may contain the following types of errors: Failure of identification of the disease or clinical condition based on the case presentation (Diagnosis). Incorrect or inappropriate immediate clinical decision-making step (non-pharmacologic, non-surgical) based on the scenario. This includes next steps, monitoring, disposition, or supportive interventions. This excludes pharmacotherapy and procedural treatments (Management). Errors in recommending or describing the definitive intervention, typically surgical, procedural, or medication. This is distinct from general management (Treatment). Wrong drug choice, dose, route, timing, interactions or duration (Pharmacotherapy). Errors in identifying the microorganism responsible for an infection or disease state (Causal organism / pathogen). Errors in lab reference ranges, thresholds, or interpretations (e.g., diagnostic cutoffs, ABG interpretation) (Lab/serum value interpretation cause). Misinterpretation of physiological concepts, including ECG, PFT, Capnography or Jugular waveform, this excludes ABG which interpreted via serum readings (Physiology), errors in describing microscopic appearance, tissue structures, or classic cellular features (Histology). Errors in location, relation, or function of anatomical structures (Anatomy). False or misleading data regarding incidence, prevalence, risk factors, or statistical analysis. This includes misuse or misinterpretation of biostatistical tools commonly used in epidemiology : misuse of sensitivity /specificity fomulas , P value , matching study design (Epidemiology).

You need to determine whether the text is correct and output a single line result consisting of the following four parts:

<Text ID>: A unique identifier for the text

<Error Flag>: Indicates whether there is an error; enter 0 if correct, or 1 if there is an error

<Error Sentence ID>: The ID of the erroneous sentence; enter -1 if correct, or the sentence ID if there is an error

<Corrected sentence>: The corrected sentence; enter NA if correct, or the corrected sentence enclosed in double quotes if there is an error

Do not output any explanations, additional text, or other formats.
Output strictly one line, in one of the following two formats:

If the text is correct, output:
<Text ID> 0 -1 NA

If the text has an error, output:
<Text ID> 1 <Error Sentence ID> "<Corrected sentence>"

Now, please review the following text:
Text ID: {text_id}
{text}

Figure S2: Prompt Used in MedErrBench-En Construction.

Table S7: Results on Error Detection and Correction Models.

Models	Detection Accuracy	Localization Accuracy	Error Correction ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
				MedErrBench-EN				
CLD-MEC	0.639	0.615	0.502	0.477	0.497	0.462	0.505	0.534
Maven	0.264	0.192	0.035	0.021	0.027	0.013	0.128	0.081
VerbaNexAI-GRU	0.510	-	-	-	-	-	-	-
VerbaNexAI-ClinicalBERT	0.601	-	-	-	-	-	-	-
				MedErrBench-CN				
CLD-MEC	0.655	0.630	0.479	0.430	0.472	0.414	0.581	0.503
				MedErrBench-Ara				
CLD-MEC	0.464	0.381	0.302	0.276	0.296	0.260	0.355	0.315

Prompt Used in Chinese Dataset Construction

以下是一段关于患者的医学叙述。
你是一位经验丰富的医生，正在审阅这段临床文本。
文本要么完全正确，要么最多包含一个错误。
每行是一句句子，格式为：句子 ID + 句子内容。
这段文本中包含以下类型的错误：未能根据病例表现正确识别疾病或临床状态（诊断错误）；在当前情境下选择了不正确或不恰当的即时临床决策步骤（疾病管理错误），包括下一步措施、监测、处置或支持性干预，但不包括药物或手术治疗；在推荐或描述最终干预措施时出错，通常指手术、操作或药物干预（治疗错误）；选择了错误的药物、剂量、途径、时间、相互作用或疗程（药物治疗错误）；未能正确识别引起感染或疾病状态的微生物（致病生物识别错误）；对实验室参考范围、阈值或解释有错误（实验室/血清值解读错误）；对生理学概念的误解，包括心电图、肺功能、呼气末二氧化碳或颈静脉波形的解读（生理学误解）；在描述显微镜下外观、组织结构或典型细胞特征时出错（组织学描述错误）；在描述解剖结构的位置、关系或功能时出错（解剖学描述错误）；以及关于发病率、患病率、危险因素或统计分析的数据错误或误导性解释，包括生物统计工具的误用或误解，如敏感性/特异性公式、P 值、匹配研究设计的误用（流行病学分析错误）。

你需要判断文本是否正确，并输出一行结果，结果由以下四个部分组成：
- <Text ID>：文本的唯一标识符
- <Error Flag>：是否有错误；如果正确填 0，如果有错误填 1
- <Error Sentence ID>：出错的句子 ID；如果正确填 -1，如果有错误填出错的句子 ID
- <Corrected sentence>：更正后的句子；如果正确填 NA，如果有错误填更正后的句子（需要用双引号括起来）

不要输出任何解释、附加文字或其他格式。
只输出严格一行，格式必须为以下两种之一：

如果文本正确，输出：
<Text ID> 0 -1 NA

如果文本有错误，输出：
<Text ID> 1 <Error Sentence ID> "<Corrected sentence>"

现在，请审阅以下文本：
Text ID: {text_id}
{text}

Figure S3: Prompt Used in MedErrBench-CN Construction.

Prompt Used in Arabic Dataset Construction

الفقرة التالية هي سرد طبي لحالة مريض. أنت طبيب ماهر تقوم بمراجعة النص السريري.
النص إما أن يكون صحيحًا أو يحتوي على خطأ واحد. النص يتكوّن من جملة واحدة في كل سطر.
كل سطر يبدأ بمعرّف الجملة (sentence ID)، ثم الجملة المطلوب التحقق منها.

قد يحتوي النص على الأنواع التالية من الأخطاء:

* فشل في تحديد المرض أو الحالة السريرية بناءً على العرض (التشخيص).
* قرار سريري غير صحيح أو غير مناسب في الخطوة الفورية (غير دوائي، غير جراحي) بناءً على السيناريو. يشمل ذلك الخطوات التالية، المراقبة، الخطة العلاجية، أو التدخلات الداعمة (الإدارة العامة)، ولا يشمل ذلك المعالجة الدوائية أو الإجرائية.
* أخطاء في التوصية أو وصف التدخل النهائي، والذي يكون عادةً جراحيًا أو إجرائيًا أو دوائيًا. وهذا يختلف عن الإدارة العامة (العلاج).
* اختيار دواء خاطئ، أو جرعة، أو طريق إعطاء، أو توقيت، أو تداخلات دوائية، أو مدة العلاج (المعالجة الدوائية).
* أخطاء في تحديد الكائن الممرض المسؤول عن عدوى أو حالة مرضية (العامل المسبب / الممرض).
* أخطاء في القيم المرجعية المخبرية، العتبات التشخيصية، أو التفسيرات (مثل حدود التشخيص، تفسير غازات الدم الشرياني) (تفسير القيم المخبرية/المصلية).
* تفسير خاطئ للمفاهيم الفسيولوجية، مثل تخطيط القلب، اختبارات وظائف الرئة، قياس ثاني أكسيد الكربون، أو موجات الوريد الوداجي، باستثناء غازات الدم الشرياني التي تُفسر من خلال القراءات المخبرية (الفيزيولوجيا).
* أخطاء في وصف البنية المجهرية، أو مظهر الأنسجة، أو السمات الخلوية النموذجية (النسج).
* أخطاء في تحديد موقع أو علاقة أو وظيفة البُنى التشريحية (التشريح).
* بيانات كاذبة أو مضللة تتعلق بالحدوث، الانتشار، عوامل الخطر، أو التحليل الإحصائي. يشمل ذلك سوء استخدام أدوات الإحصاء الحيوي الشائعة في علم الوبائيات، مثل الحساسية / النوعية، القيم الاحتمالية (P value)، أو تصميم الدراسة (الوبائيات).

مهمتك هي تحديد ما إذا كان النص صحيحًا أو يحتوي على خطأ، وأن تُرجع سطرًا واحدًا فقط يحتوي على الأجزاء الأربعة التالية:

* <Text ID>: معرّف فريد للنص
* <Error Flag>: يحدد ما إذا كان هناك خطأ؛ أدخل 0 إذا كان النص صحيحًا، أو 1 إذا كان يحتوي على خطأ
* <Error Sentence ID>: معرّف الجملة التي تحتوي على الخطأ؛ أدخل -1 إذا كان النص صحيحًا، أو رقم الجملة إذا كان هناك خطأ
* <Corrected sentence>: الجملة المصححة؛ أدخل NA إذا كان النص صحيحًا، أو أدخل الجملة المصححة بين علامتي تنصيص إذا كان هناك خطأ

لا تُرجع أي تفسيرات أو نص إضافي أو أي تنسيق آخر.

يجب أن يكون الإخراج في سطر واحد فقط، بصيغة من الصيغتين التاليتين:

إذا كان النص صحيحًا، أرجع: <Text ID> 0 -1 NA

إذا كان النص يحتوي على خطأ، أرجع: <Text ID> 1 <Error Sentence ID> <Corrected sentence>

الآن، يرجى مراجعة النص التالي:
Text ID: {text_id}
{text}

Figure S4: Prompt Used in MedErrBench-Ara Construction.

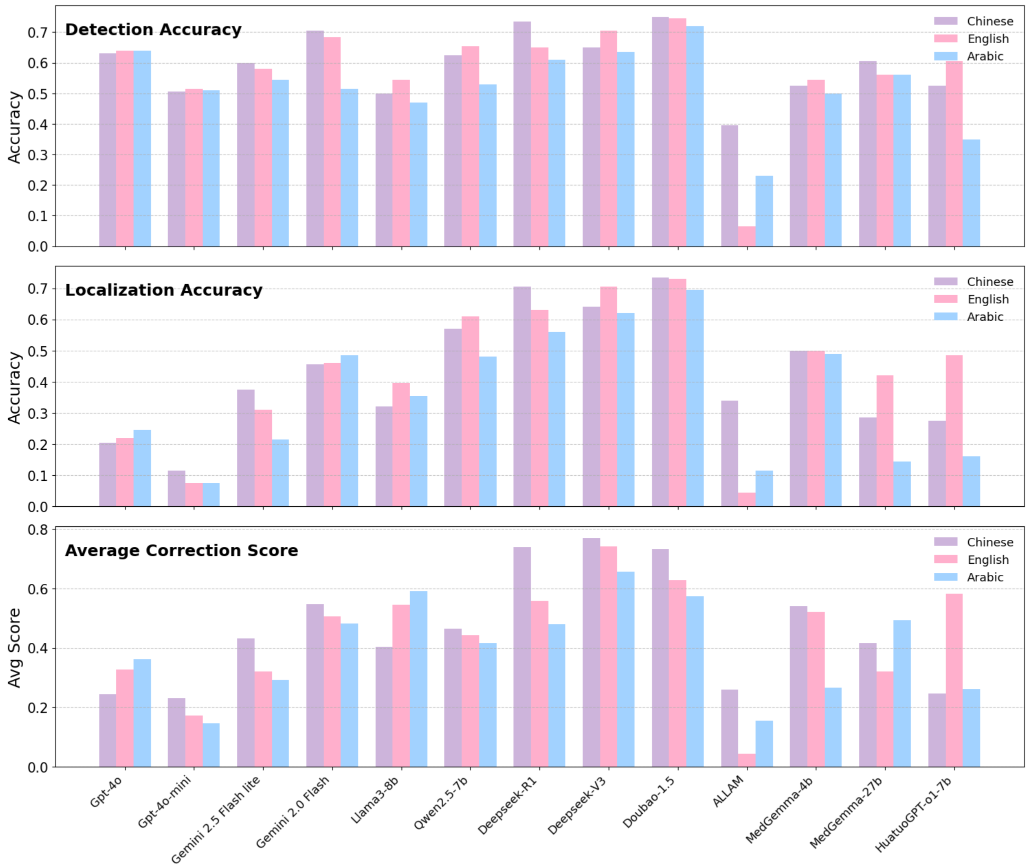


Figure S5: Cross-lingual performance comparison of LLMs on Chinese-origin tasks.

English: Initial Question and Correct/Incorrect Answers	
Question	A 36-year-old woman is brought to the emergency department because of loss of consciousness. Her husband says that she complained of vomiting and dizziness for 2 days, but did not seek medical care. She has a history of acral vitiligo. Her blood pressure is 90/60 mm Hg, temperature is 35.9°C (96.6°F), and pulse is 90/min. On physical examination, she is obtunded, skin turgor is reduced, and her oral mucosa is dry with hyperpigmented spots on the inner side of both cheeks. Lab results are shown:\nRandom plasma glucose 40 mg/dL\nSerum osmolarity 275 mOsm/kg H_2O\nSerum sodium 130 mEq/L\nSerum potassium 6 mEq/L\nUrine sodium level 30 mEq/L\nECG shows normal sinus rhythm without ST-T wave changes. Dextrose 5% is given and random plasma glucose becomes 60 mg/dL, but her blood pressure is still the same. Which of the following is the best next step in the management of this patient condition?
Options	{’A’: ‘Intravenous calcium gluconate’, ’B’: ‘ACTH stimulation test’, ’C’: ‘Subcutaneous octreotide’, ’D’: ‘Oral dehydroepiandrosterone (DHEA)’, ’E’: ‘Normal saline and intravenous dexamethasone’}
Answer	E

English: Generated Scenario with Correct Answer	
Scenario with Answer at the End	A 36-year-old woman is brought to the emergency department because of loss of consciousness. Her husband says that she complained of vomiting and dizziness for 2 days, but did not seek medical care. She has a history of acral vitiligo. Her blood pressure is 90/60 mm Hg, temperature is 35.9°C (96.6°F), and pulse is 90/min. On physical examination, she is obtunded, skin turgor is reduced, and her oral mucosa is dry with hyperpigmented spots on the inner side of both cheeks. Lab results are shown: Random plasma glucose 40 mg/dL Serum osmolarity 275 mOsm/kg H_2O Serum sodium 130 mEq/L Serum potassium 6 mEq/L Urine sodium level 30 mEq/L ECG shows normal sinus rhythm without ST-T wave changes. Dextrose 5% is given and random plasma glucose becomes 60 mg/dL, but her blood pressure is still the same. The next best step in management is to administer normal saline and intravenous dexamethasone.
Type	Management
Important Words	Normal saline and intravenous dexamethasone

English: Initial Question and Correct/Incorrect Answers	
Question	A 59-year-old woman presents to her primary care provider with a 6-month history of progressive left-arm swelling. Two years ago she had a partial mastectomy and axillary lymph node dissection for left breast cancer. She was also treated with radiotherapy at the time. Upon further questioning, she denies fever, pain, or skin changes, but reports difficulty with daily tasks because her hand feels heavy and weak. She is bothered by the appearance of her enlarged extremity and has stopped playing tennis. On physical examination, nonpitting edema of the left arm is noted with hyperkeratosis, papillomatosis, and induration of the skin. Limb elevation, exercise, and static compression bandaging are started. If the patient has no improvement, which of the following will be the best next step?
Options	{ ’A’: ‘Diethylcarbamazine’, ’B’: ‘Low molecular weight heparin’, ’C’: ‘Endovascular stenting’, ’D’: ‘Vascularized lymph node transfer’, ’E’: ‘Antibiotics’ }
Answer	D

English: Generated Scenario with Correct Answer	
Scenario with Answer at the End	A 59-year-old woman presents to her primary care provider with a 6-month history of progressive left-arm swelling. Two years ago she had a partial mastectomy and axillary lymph node dissection for left breast cancer. She was also treated with radiotherapy at the time. Upon further questioning, she denies fever, pain, or skin changes, but reports difficulty with daily tasks because her hand feels heavy and weak. She is bothered by the appearance of her enlarged extremity and has stopped playing tennis. On physical examination, nonpitting edema of the left arm is noted with hyperkeratosis, papillomatosis, and induration of the skin. Limb elevation, exercise, and static compression bandaging are started. If the patient has no improvement, the best next step is vascularized lymph node transfer.
Type	Treatment
Important Words	Vascularized lymph node transfer

Figure S6: Examples for Error Insertion in MedErrBench-EN.

Chinese: Initial Question and Correct/Incorrect Answers	
Question	患者，男，37岁，高热、皮肤瘙痒半月，右颈及锁骨上淋巴结肿大，无压痛，互相粘连。血红蛋白90g/L，白细胞10×10^9/L，中性粒细胞66%，淋巴细胞24%。骨髓涂片找到里-斯细胞。哪项为常用的化疗方案?（ ）
Options	{'A': 'MOPP', 'B': 'VDP', 'C': '羟基脲', 'D': '苯丁酸氮芥', 'E': 'HA＋DA'}
Answer	A

Chinese: Generated Scenario with Correct Answer	
Scenario with Answer at the End	患者，男，37岁，高热、皮肤瘙痒半月，右颈及锁骨上淋巴结肿大，无压痛，互相粘连。血红蛋白90g/L，白细胞10×10^9/L，中性粒细胞66%，淋巴细胞24%。骨髓涂片找到里-斯细胞。考虑给予MOPP方案进行化疗。
Scenario with Answer in the Middle	患者，男，37岁，高热、皮肤瘙痒半月，右颈及锁骨上淋巴结肿大，无压痛，互相粘连。考虑给予MOPP方案进行化疗。血红蛋白90g/L，白细胞10×10^9/L，中性粒细胞66%，淋巴细胞24%。骨髓涂片找到里-斯细胞。
Type	Pharmacotherapy
Important Words	MOPP

Chinese: Initial Question and Correct/Incorrect Answers	
Question	患儿3岁，因发热，头痛，呕吐3天伴强直性抽搐2次入院。体检：意识模糊，营养差，颈抵抗阳性，右侧鼻唇沟较浅，右眼闭合不全，心肺腹未见异常，Kernig征、Barbinski征阳性，脑脊液蛋白0.8g/L，糖2.24mmol/L（400mg/100ml），氯化物100mmol/L（585mg/dl），白细胞60×106/L（60/mm3），多核细胞45%，单核细胞55%。为进一步明确诊断，最有价值的检查是（ ）。
Options	{'A': '做结核菌素试验', 'B': '做血培养', 'C': '脑脊液乳酸盐测定', 'D': '脑脊液免疫球蛋白测定', 'E': '脑脊液找菌'}
Answer	E

Chinese: Generated Scenario with Correct Answer	
Scenario with Answer at the End	患儿3岁，因发热，头痛，呕吐3天伴强直性抽搐2次入院。体检：意识模糊，营养差，颈抵抗阳性，右侧鼻唇沟较浅，右眼闭合不全，心肺腹未见异常，Kernig征、Barbinski征阳性。脑脊液蛋白0.8g/L，糖2.24mmol/L（400mg/100ml），氯化物100mmol/L（585mg/dl），白细胞60×10^6/L（60/mm3），多核细胞45%，单核细胞55%。为明确诊断，进行脑脊液找菌。
Scenario with Answer in the Middle	患儿3岁，因发热，头痛，呕吐3天伴强直性抽搐2次入院。体检：意识模糊，营养差，颈抵抗阳性，右侧鼻唇沟较浅，右眼闭合不全，心肺腹未见异常，Kernig征、Barbinski征阳性。为明确诊断，进行脑脊液找菌。脑脊液蛋白0.8g/L，糖2.24mmol/L（400mg/100ml），氯化物100mmol/L（585mg/dl），白细胞60×10^6/L（60/mm3），多核细胞45%，单核细胞55%。
Type	Management
Important Words	脑脊液找菌

Figure S7: Examples for Error Insertion in MedErrBench-CN.

Arabic: Initial Question and Correct/Incorrect Answers	
Question	العضلة المنحرفة البطنية الظاهرة (كل ما يلي صحيح عدا):
Options	{'A': 'صفيحة عضلية صفاقية', 'B': 'تسير أليافها بشكل مائل للأمام والأسفل والأنسي', 'C': 'ترتكز على الخط الأبيض من الرهابة إلى العانة', 'D': 'تتعصب من القطني الثاني', 'E': 'تتعصب من القطني الأول'}
Answer	D

Arabic: Generated Text with Correct Answer	
Text with Answer at the End	العضلة المنحرفة البطنية الظاهرة عبارة عن صفيحة عضلية صفاقية، تسير أليافها بشكل مائل للأمام والأسفل والأنسي، وترتكز على الخط الأبيض من الرهابة إلى العانة، وتتعصب من القطني الأول.
Text with Answer in the Middle	العضلة المنحرفة البطنية الظاهرة عبارة عن صفيحة عضلية صفاقية، تتعصب من القطني الأول، وتسير أليافها بشكل مائل للأمام والأسفل والأنسي، وترتكز على الخط الأبيض من الرهابة إلى العانة..
Type	Anatomy
Important Words	القطني الأول

Arabic: Initial Question and Correct/Incorrect Answers	
Question	مريض عمره 60 سنة كان يعمل بمنجم للفحم لمدة 20 سنة، راجع بسعال وزلة تنفسية، صورة الصدر أظهرت ارتشاحات خلالية، فحص القشع كان إيجابيًا لعصية كوخ، أي من العوامل الآتية قد يكون عاملًا مؤهبًا لإصابته بالتدرن:
Options	{'A': 'بيريليوم', 'B': 'كادميوم', 'C': 'الفحم', 'D': 'السيليكا', 'E': 'غبار المنزل'}
Answer	D

Arabic: Generated Scenario with Correct Answer	
Scenario with Answer at the End	مريض عمره 60 سنة، عمل بمنجم فحم لمدة 20 سنة، يعاني من سعال وضيق في التنفس. أظهرت صورة الصدر ارتشاحات خلالية، وكان فحص القشع إيجابياً لعصية كوخ. يُعتبر تعرضه للسليكا عاملاً مُؤهباً لإصابته بالتدرن.
Scenario with Answer in the Middle	مريض عمره 60 سنة، عمل بمنجم فحم لمدة 20 سنة، يُعتبر تعرضه للسليكا عاملاً مُؤهباً لإصابته بالتدرن، يعاني من سعال وضيق في التنفس. أظهرت صورة الصدر ارتشاحات خلالية، وكان فحص القشع إيجابياً لعصية كوخ.
Type	Causal Organism
Important Words	تعرضه للسليكا

Figure S8: Examples for Error Insertion in MedErrBench-Ara.

	ERROR	CORRECTION
DIAGNOSIS	A 6-year-old boy presents to your office with hematuria. Two weeks ago the patient had symptoms of a sore throat and fever. Although physical exam is unremarkable, laboratory results show a decreased serum C3 level and an elevated anti-DNAse B titer. **Renal biopsy would most likely show Polyclonal IgA deposition on immunofluorescence.**	**Renal biopsy would most likely show large, hypercellular glomeruli on light microscopy.**
CAUSAL ORGANISM	A 43-year-old patient with a history of esophageal cancer presents with sudden onset slurred speech. On CT scan, she is found to have an acute intraparenchymal hemorrhage, likely from a metastasis, with interventricular extension. Due to concern for hydrocephalus, an external ventricular drain is placed in order to control the intracerebral pressure. Two weeks later, she undergoes a lumbar puncture due to altered mental status, and the gram stain of the fluid is depicted in Figure A. **The endoplasmic reticulum is responsible for the blue color on the stain.**	**The part of the bacteria responsible for the blue color on the stain is the Cell wall.**
MANAGEMENT	A 45-year-old man in respiratory distress presents to the emergency department. He sustained a stab to his left chest and was escorted to the nearest hospital. The patient appears pale and has moderate difficulty with breathing. His O2 saturation is 94%. The left lung is dull to percussion. **The next best step in management is to Thoracotomy**. CXRs are ordered and confirm the likely diagnosis. His blood pressure is 95/57 mm Hg, the respirations are 22/min, the pulse is 87/min, and the temperature is 36.7¬∞C (98.0¬∞F). His chest X-ray is shown.	**The next best step in management is to Chest tube insertion.**
TREATMENT	A 21-year-old female is brought by her brother to the emergency department after having a generalized tonic-clonic seizure one hour ago. She is slightly confused and has no recollection of her seizure. Her brother relayed that the patient has a history of severe anxiety for which she takes medication. For the past several days, he noticed that his sister exhibited body tremors, appeared to be agitated with quick mood changes, and, at times, was delirious. He states his sister recently ran out of her medications while visiting from out of town. **Naloxone would best treat the patient's condition.**	**Diazepam would best treat the patient's condition.**
PHARMACOTHERAPY	A 52-year-old man awakens in the middle of the night with excruciating pain in his right great toe. He reports that even the touch of the bed sheet was unbearably painful. His right foot is shown in figure A. He is treated with colchicine. **The mechanism of action of colchicine is inhibition of reabsorption of uric acid in proximal convoluted tubule.**	**The mechanism of colchicine is to decrease microtubule polymerization.**
PHYSIOLOGY	A 20-year-old woman is brought to the emergency department with a puncture wound on the right side of her chest. She was walking to her apartment when she was assaulted. As she resisted to give up her purse, the assailant stabbed her in the chest with a knife and ran away. She is in severe respiratory distress. Her heart rate is 140/min, respiratory rate is 28/min, and blood pressure is 145/65 mm Hg. The pulse oximetry shows an oxygen saturation of 84%. An oval puncture wound is seen on the right lateral aspect of her chest and she is stuporous. The heart sounds are normal and no jugular venous distension is seen. Distant breath sounds are present on the right. **the decreased intrapleural pressure during inspiration explains her breathing difficulty**	**The equal intrapleural and atmospheric pressures during inspiration explains her breathing difficulty.**
ANATOMY	An 85 year old ventilator dependent male was endotracheally intubated 10 days ago. He remains unresponsive and he is not a candidate for early extubation . The ICU attending elects to perform tracheotomy at the bedside. During the procedure copious amount of dark blood is encountered. **this is most likely due to the transection of inferior thyroid vein.**	**This is most likely due to the transection of anterior jugular vein**
LAB / SERUM VALUE	A 2-month-old boy is brought to his pediatrician for a routine visit. His mother is concerned because he developed a rash one month ago that has not resolved (Figure A). Furthermore, she states that he has seemed to constantly be sick ever since his birth. On physical exam, the pediatrician notes the findings demonstrated in Figure B. **The expected lab/serum value is Increased IgM; Decreased IgG, IgA, IgE**. The pediatrician pursues further workup and orders a number of lab tests	**The expected lab/serum value is Increased IgE, IgA; Decreased IgM.**
HISTOLOGY	A new drug is developed that prevents the demyelinization occurring in the progress of multiple sclerosis. The drug protects the cells responsible for the synthesis and maintenance of myelin in the central nervous system. **These cells are most likely astrocytes.**	**These cells are most likely oligodendrocyte.**
EPIDEMIOLOGY	Investigators wish to compare the baseline age of participants in a randomized controlled trial. There are 50 patients in each group. The researchers find that in both arms, distribution of age is bimodal. **The most appropriate statistical test that can be used for this comparison is Chi square test .**	**The most appropriate statistical test that can be used for this comparison is Mann-Whitney testU.**

Figure S9: Illustrative Examples of the Ten Error Types Defined in the MedErrBench-EN.

	ERROR	CORRECTION
DIAGNOSIS	4岁患儿，明显浮肿，血压12/8kPa。**尿蛋白（+++），尿红细胞0～2/HP，考虑IgA肾病。**血浆白蛋白22g/L。	尿蛋白（+++），尿红细胞0～2/HP，考虑肾炎型肾病。
CAUSAL ORGANISM	某男，24岁，有不洁性接触史，因近2日尿急、尿频、排尿刺痛而来院就诊。查体见尿道口有白色脓性分泌物。分泌物涂片染色，镜下见到G-成双排列的球菌。**考虑肺炎球菌感染**。	考虑淋球菌感染。
MANAGEMENT	女性，18岁，1型糖尿病病史3年，因“肺部感染”诱发糖尿病酮症酸中毒。**给予积极大量补碱**，尽快纠正酸中毒治疗。	给予积极治疗肺部感染。
TREATMENT	男性，19岁，9个月前开始右上臂肿胀，疼痛。**左肩关节离断术被考虑为治疗方案**。诊断为右肱骨上端骨肉瘤。	术前化疗加根治性手术后放疗被考虑为治疗方案。
PHARMACOTHERAPY	患者，男，42岁，2个月来左颈部淋巴结进行性肿大、无痛，周期性发热，消瘦。**近1周上胸部水肿，颈粗，考虑给予阿霉素（多柔比星）治疗**。淋巴结活检有里-斯细胞，胸片示纵隔有肿块。	近1周上胸部水肿，颈粗，考虑给予MOPP方案治疗。
PHYSIOLOGY	男，3个月，生长发育良好，体重为5kg。**每日所需水分是450ml**。因母亲患慢性疾病，需停用母乳而改牛奶喂养。	每日所需水分是750ml。
ANATOMY	54岁女性，脑动脉硬化症病史3年，突感眩晕、呕吐、言语不清。查体：声音嘶哑、吞咽困难、言语含混，左眼裂小、瞳孔小、水平眼震、左面部右半身痛觉减退，左侧指鼻试验不准。**结合临床表现及查体结果，考虑为右侧小脑后下动脉血栓形成。**	结合临床表现及查体结果，考虑为左侧小脑后下动脉血栓形成。
LAB / SERUM VALUE	女性，29岁，贫血病史1年，浅表淋巴结不肿大，肝脾未触及。**网织红细胞减少。**血象呈现全血细胞减少。	骨髓增生低下，造血细胞减少。
HISTOLOGY	男，46岁，湖北农民。肝功能反复异常10余年。1个月来出现腹胀，尿黄。查体：面色晦暗，巩膜黄染，见肝掌及蜘蛛痣，腹水征（+）。**肝细胞水肿，假小叶形成。**实验室检查：ALT 180U/L，TBil 37μmol/L，PTA 60%，HBsAg（—），抗HCV（—）。	肝脏呈干线状纤维化，肝脏表面有大小不等的结节。
EPIDEMIOLOGY	某幼儿园大班11名6岁儿童接受百白破疫苗注射后，其抗体滴度分别是1:20，1:20，1:20，1:40，1:40，1:80，1:80，1:160，1:160，1:320，1:640。**为了描述其抗体滴度的集中趋势，计算了四分位间距。**	为了描述其抗体滴度的集中趋势，计算了几何平均数。

Figure S10: Illustrative Examples of the Ten Error Types Defined in the MedErrBench-CN.

	ERROR	CORRECTION
DIAGNOSIS	أثناء إجراء تصوير بالموجات فوق الصوتية للبطن لسيدة حامل تبلغ من العمر 30 عامًا، تبين وجود كتلة متكلسة في الطحال قطرها 1 سم مع سماع لغط في نفس المنطقة، ولم تظهر عليها أي أعراض. **يُشير ذلك إلى احتمالية وجود احتشاء طحال**.	يُشير ذلك إلى احتمالية وجود أم دم شريان طحالي
CAUSAL ORGANISM	يُعد التهاب الجيوب الأنفية، وأمراض المناعة الذاتية مثل التصلب المتعدد، **بالإضافة إلى البراريق**، من أسباب التهاب العصب خلف المقلة	بالإضافة إلى الانسمامات الكحولية والتبغية
MANAGEMENT	حضر شاب يبلغ من العمر 32 عامًا إلى قسم الطوارئ يُعاني من تقيؤ دماء حمراء قانية بكميات كبيرة، **وكانت الخطوة الأولى والأهم في تدبير حالته هي تحديد عيار الخضاب**.	وكانت الخطوة الأولى والأهم في تدبير حالته هي الإنعاش بالسوائل
TREATMENT	في حالة تقرير إجراء تشعيع وقائي للقحف بجرعة ٢ غري لكل جلسة، **فإن الجرعة الكلية الموصى بها هي ١٨ غري**.	فإن الجرعة الكلية الموصى بها هي ٣٠ غري
PHARMACOTHERAPY	**لا تُعتبر المضادات الحيوية ريفامبيسين**، وحمض الديسكيك، وسبيروفلوكساسين من الأدوية التي تمنع تشكل الأحماض النووية في البكتيريا.	تُعتبر المضادات الحيوية ريفامبيسين
PHYSIOLOGY	ضغط معصرة المريء السفلية يخفض من التدخين، والكافيين، والأطعمة الدسمة، **وزيادة حموضة المعدة**.	بينما زيادة حموضة المعدة لا تؤثر على هذا الضغط
ANATOMY	**لا يحتوي البنكرياس على الأقنية المخططة**، كما أنه يُصنّف غدةً أنبوبيةً مركبة	يحتوي البنكرياس على الأقنية المخططة
LAB / SERUM VALUE	طُلِب منك مراجعة حالة مريضة تبلغ من العمر 59 عامًا تم إدخالها إلى الجناح الطبي الحاد في المستشفى. أبلغتك الممرضة بأنها تبدو في حالة ضِيق في التنفس رغم أنها تتلقى حاليًا 3 لترات من الأوكسجين عبر قنيّات أنفية. أجريتَ لها تحليل غازات الدم الشرياني وأظهرت النتائج ما يلي: pH = 7.30، $PaCO_2$ = 8.4 كيلوباسكال، PaO_2 = 9.1 كيلوباسكال، HCO_3^- = 29 مليمول/لتر، Base Excess = +4، **ويُظهر تحليل الغازات وجود قلاء تنفسي مع معاوضة استقلابية**.	وتشير هذه النتائج إلى وجود حُماض تنفسي مع معاوضة استقلابية.
HISTOLOGY	**الغدد العرقية هي غدد متشعبة ملتفة**، يحاط الجزء المفرز منها بخلايا ظهارية عضلية.	الغدد العرقية هي غدد بسيطة أنبوبية ملتفة
EPIDEMIOLOGY	في حالات انسداد الأمعاء المصاحبة للحمل، **غالبًا ما يكون السبب هو حدوثه خلال الثلث الأول من الحمل**.	غالبًا ما يكون السبب هو الالتصاقات الناتجة عن جراحات سابقة

Figure S11: Illustrative Examples of the Ten Error Types Defined in the MedErrBench-Ara.