# MEDRECT: A Medical Reasoning Benchmark for Error Correction in Clinical Texts

**Anonymous ACL submission**


## Abstract

Large language models (LLMs) show increasing promise in medical applications, but their ability to *detect and correct errors in clinical texts*, a prerequisite for safe deployment, remains under-evaluated, particularly beyond English. We introduce MEDRECT, a cross-lingual benchmark (Japanese/English) that formulates medical error handling as three subtasks: error detection, error localization (sentence extraction), and error correction. MEDRECT is built with a scalable, automated pipeline from the Japanese Medical Licensing Examinations (JMLE) and a curated English counterpart, yielding MEDRECT-ja (663 texts) and MEDRECT-en (458 texts) with comparable error/no-error balance. We evaluate 9 contemporary LLMs spanning proprietary, open-weight, and reasoning families. Key findings: (i) reasoning models substantially outperform standard architectures, with up to 13.5% relative improvement in error detection and 51.0% in sentence extraction; (ii) cross-lingual evaluation reveals 5-10% performance gaps from English to Japanese, with smaller disparities for reasoning models; (iii) LoRA fine-tuning yields asymmetric improvements in error correction performance (Japanese: +0.078, English: +0.168) while preserving reasoning capabilities; and (iv) our fine-tuned model exceeds human expert performance on structured medical error correction tasks. To our knowledge, MEDRECT is the first comprehensive cross-lingual benchmark for medical error correction, providing a reproducible framework and resources for developing safer medical LLMs across languages.


**Example 1 (MEDRECT-ja):** **Diagnostic Error**
A 20-day-old male infant was brought to the hospital by his mother with chief complaints of poor feeding and fever. [...] Physical examination: Poor complexion, bulging anterior fontanelle, and irritability were observed. Cerebrospinal fluid: Cell count 4,200/mm³, protein 80 mg/dL, glucose 5 mg/dL. The laboratory findings strongly suggested meningitis caused by *Pseudomonas aeruginosa*.
**Corrected:** The laboratory findings strongly suggested meningitis caused by GBS (*Streptococcus agalactiae*).

**Example 2 (MEDRECT-ja):** **Procedural Error**
An 82-year-old woman is hospitalized in a palliative care unit for pancreatic cancer with liver metastases. For the past week, she has had a progressive loss of appetite and decreased food intake. [...] The nutrition support team proposed the placement of a gastrostomy tube for nutritional supplementation. The policy was to explain the plan to the attending physician and obtain consent.
**Corrected:** The nutrition support team decided to prioritize consideration of non-invasive nutritional management methods.

**Example 3 (MEDRECT-en):** **Medication Selection Error**
A 73-year-old man presents with 'weird blisters' on his right hand, appeared 2 weeks ago. [...] Doxycycline is prescribed after physical exam. On physical exam: multiple bullae with red, papular lesions on right hand progressing to forearm. Right axillary lymph nodes are swollen and tender.
**Corrected:** Itraconazole is prescribed after physical exam.

Figure 1: Examples from the MEDRECT dataset showing different error types. Examples 1-2 show MEDRECT-ja samples (translated to English for readability), while Example 3 shows a native MEDRECT-en sample derived from MEDEC (Ben Abacha et al., 2025). Each example highlights the erroneous sentence (underlined and italicized) and provides the correct version.

## 1 Introduction

Large Language Models (LLMs) offer unprecedented potential to augment clinical decision-making (Usuyama et al., 2025), yet their deployment faces a critical challenge: the opacity and reliability of their reasoning processes. Models may arrive at correct conclusions through flawed logic (Turpin et al., 2023; Lyu et al., 2023) or replicate cognitive biases—such as anchoring and confirmation bias—that contribute to diagnostic errors (Saposnik et al., 2016; Makary and Daniel, 2016).

While state-of-the-art LLMs demonstrate remarkable success on structured examinations like the USMLE (Gilson et al., 2023; Singhal et al., 2023), multiple-choice question-answering (MCQA) performance insufficiently represents real-world clinical reasoning. Recent analyses highlight a critical gap between generating plausible text and ensuring patient safety (Moëll et al., 2025), underscoring the need for benchmarks evaluating reasoning reliability, not just answers.

This challenge is particularly acute in Japanese medical Natural Language Processing (NLP). While Japanese medical LLMs advance rapidly (Kawakami et al., 2025; Sukeda, 2024), the field has lacked standardized benchmarks for complex clinical tasks (Jiang et al., 2024). While MEDEC (Ben Abacha et al., 2024, 2025) provided foundational methodology for English medical error correction, its manual annotation is resource-intensive and monolingual, leaving unanswered how reasoning capabilities transfer across languages.

To bridge these gaps, we introduce MEDRECT[1] (A **Med**ical **R**easoning benchmark for **E**rror **C**orrection in clinical **T**exts), the first comprehensive cross-lingual benchmark for medical error detection and correction. Our contributions are: **(1) Automated Benchmark Construction**: A fully automated pipeline synthesizing high-quality error correction data from the Japanese Medical Licensing Examinations (JMLE) for reproducible cross-lingual benchmarks. **(2) Cross-Lingual Evaluation**: MEDRECT-ja (663 samples) and MEDRECT-en (458 samples) with diverse error types, evaluated on 9 LLMs revealing substantial cross-lingual gaps and the critical role of reasoning. **(3) Parameter-Efficient Fine-Tuning**: LoRA (Low-Rank Adaptation) fine-tuning (Hu et al., 2021) improves bilingual error correction while preserving reasoning capabilities.

## 2 Related Work

### 2.1 Benchmarks for Medical Reasoning

Medical LLM evaluation has centered on MCQA benchmarks from licensing exams, including MedQA (Jin et al., 2020), multilingual variants (Alonso et al., 2024), and MultiMedQA with Med-PaLM achieving expert-level performance (Singhal et al., 2023). However, MCQA cannot evaluate practical clinical tasks such as communication (Zeng et al., 2020), note generation (Tang et al., 2023; Yim et al., 2023; Van Veen et al., 2024), or error correction (Ben Abacha et al., 2025).

[1]Dataset and code: `https://anonymous.4open.science/r/medrect/`

### 2.2 Error Detection and Correction in Clinical Texts

The MEDEC benchmark (Ben Abacha et al., 2024, 2025) introduced systematic error correction evaluation with five clinical error types for the MEDIQA-CORR 2024 shared task. However, MEDEC's manual annotation limits scalability and language coverage. MEDRECT addresses this by automating the pipeline with LLMs, enabling multilingual extension.

### 2.3 Japanese Medical NLP and Cross-Lingual Evaluation

Japanese medical NLP resources include clinical BERT (Kawazoe et al., 2021), MedWeb corpus (Wakamiya et al., 2019), and JMedBench with 20 datasets including IgakuQA (JMLE 2018-2022) and primarily GPT-4 translated biomedical data (Kasai et al., 2023; Jiang et al., 2024). For cross-lingual evaluation, prior work explored multilingual medical QA performance variations (Jin et al., 2023; Alonso et al., 2024) and biomedical entity linking (Liu et al., 2021). MEDRECT is the first to provide systematic cross-lingual evaluation for medical error correction, a complex unstructured task requiring clinical reasoning across languages.

## 3 MEDRECT Dataset

### 3.1 Task Definition

Following MEDEC, we decompose medical error detection and correction into three progressive subtasks: (1) **Error Detection**: binary classification to determine whether a clinical text contains an error; (2) **Error Sentence Extraction**: for texts containing an error, identify the specific sentence with the error; and (3) **Error Correction**: for texts containing an error, generate a corrected version of the erroneous sentence. This decomposition enables fine-grained evaluation of model capabilities and helps identify specific weaknesses in the error detection pipeline. Note that the latter two subtasks are only applicable to clinical texts that contain an error. Figure 1 illustrates concrete examples of these tasks across different error types in both MEDRECT-ja and MEDRECT-en datasets.

### 3.2 Data Construction Pipeline

Figure 2 illustrates the complete construction pipeline for both MEDRECT-ja and MEDRECT-en datasets.

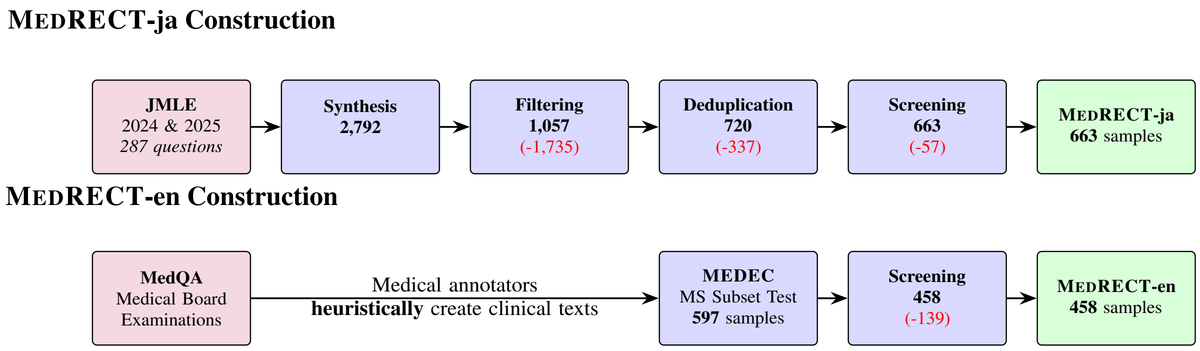


Figure 2: Data construction pipeline for MEDRECT benchmark creation. MEDRECT-ja (top) transforms JMLE questions through automated synthesis, quality filtering, model deduplication, and LLM screening to produce 663 high-quality samples. MEDRECT-en (bottom) applies identical LLM screening to the existing MEDEC MS Subset Test, yielding 458 samples. Red numbers indicate samples removed at each quality control step.

For MEDRECT-ja, we extracted 287 clinical case questions from JMLE (2024 and 2025) and applied four stages: (1) automatic transformation with error injection (287→2,792 samples), (2) difficulty-based filtering (→1,057), (3) deduplication (→720), and (4) LLM-as-a-Judge screening (→663).

For MEDRECT-en, we applied identical LLM-as-a-Judge screening to MEDEC MS Subset Test (597 samples) (Ben Abacha et al., 2024, 2025), producing 458 high-quality samples.

Unlike manual MEDEC annotation, our automated pipeline enables reproducible, scalable construction across languages and medical contexts. Complete procedures are detailed in Appendix A.1.

### 3.3 Dataset Statistics

Table 1: Dataset statistics for MEDRECT-ja and MEDRECT-en

| | MEDRECT-ja | MEDRECT-en |
|---|---|---|
| Total samples | 663 | 458 |
| With errors | 367 (55.4%) | 243 (53.1%) |
| Without errors | 296 (44.6%) | 215 (46.9%) |
| *Error Type Distribution* | | |
| Diagnosis | 77 (21.0%) | 98 (40.3%) |
| Monitoring/management | 79 (21.5%) | 17 (7.0%) |
| Physical findings | 72 (19.6%) | 2 (0.8%) |
| Procedures/intervention | 40 (10.9%) | 38 (15.6%) |
| Medication selection | 30 (8.2%) | 70 (28.8%) |
| Test interpretation | 37 (10.1%) | 12 (4.9%) |
| History taking | 22 (6.0%) | 1 (0.4%) |
| Medication dosage | 8 (2.2%) | 3 (1.2%) |
| Others | 2 (0.5%) | 2 (0.8%) |

MEDRECT-ja contains 663 samples with 367 (55.4%) errors and 296 (44.6%) correct texts, while MEDRECT-en comprises 458 samples with 243 (53.1%) errors and 215 (46.9%) correct texts. The similar error-to-correct ratios (approximately 55:45) ensure comparable cross-lingual evaluation conditions.

Error type distributions reflect different clinical contexts and source methodologies. MEDRECT-ja shows balanced distributions across diagnosis (21.0%), monitoring/management (21.5%), and physical findings (19.6%)—reflecting the detailed clinical examination culture in Japanese medical practice. MEDRECT-en is dominated by diagnosis errors (40.3%) and medication selection (28.8%), reflecting the underlying MedQA source patterns.

**Privacy and Content Safety** All source data from JMLE are publicly available standardized test materials that do not contain personally identifiable information (PII). For MEDRECT-en, we use MEDEC MS Subset Test (Ben Abacha et al., 2024), which was constructed from MedQA medical licensing exam questions (Jin et al., 2020) and similarly contains no PII.

## 4 Experimental Setup

### 4.1 Evaluated Models

We evaluated 9 contemporary LLMs[2]: **Reasoning models** (GPT-5 (OpenAI, 2025a), o3 (OpenAI, 2025d), Claude Sonnet 4 (Anthropic, 2025), DeepSeek-R1-0528 (DeepSeek-AI et al., 2025a), gpt-oss-120b/20b (OpenAI, 2025b), Qwen3-32B (Yang et al., 2025)) employ explicit step-by-step reasoning during inference; **Non-reasoning models** (GPT-4.1 (OpenAI, 2025c), DeepSeek-V3-0324 (DeepSeek-AI et al., 2025b), Qwen3-32B)

[2]Proprietary models (GPT-5, GPT-4.1, o3, Claude Sonnet 4) and DeepSeek models were accessed via OpenRouter API, while other open-weight models (gpt-oss, Qwen3-32B) were evaluated using local inference infrastructure.

do not. OpenAI's reasoning models (GPT-5, o3, gpt-oss) support configurable reasoning effort parameters within computational token limits. We used the API default medium setting for GPT-5 and o3, while evaluating gpt-oss across all three levels (high/medium/low). Claude Sonnet 4 uses a thinking parameter that enables extended thinking, which we enabled for evaluation. Qwen3-32B offers both think and no-think modes, allowing direct comparison of reasoning impact within the same architecture. DeepSeek-R1-0528 incorporates built-in reasoning capabilities without additional configuration parameters.

### 4.2 Evaluation Metrics

Evaluation metrics: Error Detection F1, Sentence Extraction Accuracy, and Error Correction using ROUGE-1 (Lin, 2004), BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and their arithmetic average. Language-specific configurations: ROUGE-1 with MeCab (Kudo et al., 2004) (Japanese) or whitespace (English) tokenizers, BERTScore F1 with microsoft/deberta-xlarge-mnli, and BLEURT-20.

Following the MEDIQA-CORR 2024 evaluation protocol (Ben Abacha et al., 2024), sentence extraction is computed only on samples with a ground-truth error, and error correction metrics are computed only on samples where both prediction and ground-truth indicate the presence of an error.

### 4.3 Evaluation Prompts

Models were evaluated using carefully designed zero-shot prompts that instructed medical experts to identify and correct a clinical error. The evaluation prompt is shown below:

**Evaluation Prompt (0_shot_en)**

```
You are a medical expert reviewing clinical text for accuracy. The text contains either no error or exactly one medical error.

Identify and correct any medical error related to treatment, diagnosis, management, or causation.

Output Format:
- If no error: 'CORRECT'
- If error found: 'sentence_number: corrected_sentence'

CRITICAL: Output ONLY the result. Do NOT include explanations, analysis, or additional text.

{sentences}
```

For Japanese evaluation, we used a direct translation of this prompt (0_shot_ja) that maintained identical task specifications and output format requirements.

### 4.4 Fine-tuning Configuration

Fine-tuning was performed using LoRA with rank=64, $\alpha$=128 on Qwen3-32B as the base model. We employed a learning rate of `1e-4` for effective task adaptation.

Qwen3-32B was finetuned using training data combining both Japanese (5,538 samples) and English (2,439 samples) datasets with reasoning processes generated by DeepSeek-R1-0528 (see Appendix B for detailed construction methodology). This bilingual training approach enabled the model to leverage cross-lingual medical knowledge and reasoning patterns, demonstrating effective performance improvements on both MEDRECT-ja and MEDRECT-en benchmarks as shown in the results.

## 5 Results

### 5.1 Performance on MEDRECT-ja Benchmark

Table 2 shows 9 models on MEDRECT-ja. Claude Sonnet 4 achieves the highest average (0.675) with strong error detection (0.795 F1) and correction. o3 (0.654) and GPT-5 (0.648) follow.

Task-specific patterns: error detection ranges 0.751-0.795 F1 (top: Claude Sonnet 4, o3, GPT-5, DeepSeek-R1-0528), while sentence extraction shows largest variance (42.2%-83.7%).

Among model categories, proprietary models generally outperform open-source alternatives, with DeepSeek-R1-0528 achieving competitive performance (0.647 average score) comparable to GPT-5 (0.648). The gpt-oss models show consistent performance patterns across reasoning effort levels: gpt-oss-120b achieves 0.604 average score (high), 0.581 (medium), and 0.553 (low).

Fine-tuning demonstrates significant benefits, with Qwen3-32B + LoRA (think) achieving substantial improvements over the base model (0.627 vs. 0.549 average score), while preserving the reasoning capabilities that distinguish the think variant from its no-think counterpart (0.471 average score). Comparing Qwen3-32B variants directly illustrates the impact of reasoning capabilities: the think version achieves 0.723 error detection F1 and 72.5% sentence extraction accuracy, compared to the no-think version at 0.637 and 48.0% respectively. This represents a 13.5% relative improvement in error

Table 2: Performance on MEDRECT-ja. Parenthetical notations indicate reasoning effort levels (gpt-oss: high/medium/low) or reasoning modes (Qwen3-32B: think/no-think).

| Model | Error Det. | Sent. Ext. | Error Correction | | | |
|---|---|---|---|---|---|---|
| | F1 | Acc. | ROUGE-1 | BERT | BLEURT | Avg. |
| *Reasoning models* | | | | | | |
| GPT-5 | 0.758 | **83.7%** | 0.561 | 0.803 | 0.580 | 0.648 |
| o3 | 0.764 | 71.4% | 0.573 | 0.810 | 0.578 | 0.654 |
| Claude Sonnet 4 | **0.795** | 82.3% | **0.607** | **0.825** | **0.594** | **0.675** |
| DeepSeek-R1-0528* | 0.751 | 79.3% | 0.570 | 0.808 | 0.563 | 0.647 |
| gpt-oss-120b (high) | 0.731 | 79.6% | 0.500 | 0.776 | 0.535 | 0.604 |
| gpt-oss-120b (medium) | 0.721 | 77.4% | 0.466 | 0.763 | 0.516 | 0.581 |
| gpt-oss-120b (low) | 0.704 | 68.9% | 0.433 | 0.742 | 0.483 | 0.553 |
| gpt-oss-20b (high) | 0.729 | 71.4% | 0.473 | 0.769 | 0.509 | 0.583 |
| gpt-oss-20b (medium) | 0.718 | 64.3% | 0.420 | 0.741 | 0.467 | 0.543 |
| gpt-oss-20b (low) | 0.678 | 46.9% | 0.333 | 0.699 | 0.397 | 0.476 |
| Qwen3-32B + LoRA (think) | 0.743 | 81.5% | 0.548 | 0.802 | 0.531 | 0.627 |
| Qwen3-32B (think) | 0.723 | 72.5% | 0.419 | 0.739 | 0.489 | 0.549 |
| *Non-reasoning models* | | | | | | |
| GPT-4.1 | 0.658 | 52.6% | 0.569 | 0.804 | 0.593 | 0.655 |
| DeepSeek-V3-0324 | 0.688 | 42.2% | 0.367 | 0.714 | 0.409 | 0.497 |
| Qwen3-32B (no-think) | 0.637 | 48.0% | 0.326 | 0.695 | 0.393 | 0.471 |

* DeepSeek-R1-0528 was involved in the MEDRECT-ja data synthesis process.

detection and 51.0% improvement in sentence extraction.

### 5.2 Cross-lingual Performance Comparison

Table 3 reveals systematic performance differences between MEDRECT-ja and MEDRECT-en benchmarks. Most proprietary models demonstrate better performance on English, while some open-weight models show mixed patterns. o3 shows strong performance on both languages with average scores of 0.654 (Japanese) and 0.714 (English), maintaining consistent error correction capabilities across languages. Notably, DeepSeek-R1-0528 achieves higher performance on Japanese (0.647 vs. 0.608 average score).

Cross-lingual performance patterns vary significantly by subtask. Sentence extraction accuracy shows the largest language-specific variations, with models like GPT-4.1 showing substantial differences (52.6% Japanese vs. 72.8% English). Error detection F1-scores show more consistent cross-lingual performance, with relatively smaller gaps such as Claude Sonnet 4 (0.795 vs. 0.784 F1), GPT-5 (0.758 vs. 0.818 F1), and o3 (0.764 vs. 0.852 F1).

Fine-tuning with LoRA demonstrates substantial performance improvements across both languages, with asymmetric gains favoring English. On MEDRECT-ja, the fine-tuned Qwen3-32B + LoRA (think)[3] achieves 0.627 average score compared to 0.549 for the base model, representing a 14.2% relative improvement. Individual metrics show consistent gains: error detection F1 improves from 0.723 to 0.743 and sentence extraction accuracy advances from 72.5% to 81.5%.

On MEDRECT-en, the improvement is even more pronounced, with average score increasing from 0.550 to 0.718 (30.5% relative improvement). This creates an inverted cross-lingual pattern where the fine-tuned model achieves superior English performance (0.718 vs. 0.627 average score) despite being trained primarily on Japanese medical data, with particularly strong English sentence extraction accuracy of 90.9%.

### 5.3 Performance by Error Type

Performance breakdown across different medical error categories reveals substantial variation in task difficulty and model behavior patterns across clinical domains (detailed results in Appendix Table 6).

Error types demonstrate distinct difficulty hierarchies across the clinical spectrum in sentence extraction accuracy. *Medication dosage* emerges as the most challenging category, with average performance around 70% and several models achieving notably lower scores (e.g., Qwen3-32B + LoRA at 27.3%). In contrast, *Medication selection* represents the most tractable category, with most models achieving above 80% sentence extraction accuracy and perfect performance from top proprietary systems. *History taking* exhibits the largest perfor-

[3]Fine-tuned model: `https://huggingface.co/XXXX`

Table 3: Cross-lingual performance comparison between MEDRECT-ja and MEDRECT-en. Parenthetical notations indicate reasoning effort levels (gpt-oss: high/medium/low) or reasoning modes (Qwen3-32B: think/no-think). "EC Avg. Score" refers to Error Correction Average Score.

| Model | MEDRECT-ja | | | MEDRECT-en | | |
|---|---|---|---|---|---|---|
| | Error Det. F1 | Sent. Ext. Acc. | EC Avg. Score | Error Det. F1 | Sent. Ext. Acc. | EC Avg. Score |
| *Reasoning models* | | | | | | |
| GPT-5 | 0.758 | **83.7%** | 0.648 | 0.818 | **96.3%** | 0.708 |
| o3 | 0.764 | 71.4% | 0.654 | **0.852** | 87.7% | 0.714 |
| Claude Sonnet 4 | **0.795** | 82.3% | **0.675** | 0.784 | 84.0% | 0.705 |
| DeepSeek-R1-0528* | 0.751 | 79.3% | 0.647 | 0.730 | 77.4% | 0.608 |
| gpt-oss-120b (high) | 0.731 | 79.6% | 0.604 | 0.759 | 92.6% | 0.663 |
| gpt-oss-120b (medium) | 0.721 | 77.4% | 0.581 | 0.777 | 88.1% | 0.630 |
| gpt-oss-120b (low) | 0.704 | 68.9% | 0.553 | 0.775 | 79.4% | 0.625 |
| gpt-oss-20b (high) | 0.729 | 71.4% | 0.583 | 0.757 | 86.0% | 0.617 |
| gpt-oss-20b (medium) | 0.718 | 64.3% | 0.543 | 0.762 | 87.2% | 0.590 |
| gpt-oss-20b (low) | 0.678 | 46.9% | 0.476 | 0.723 | 71.2% | 0.515 |
| Qwen3-32B + LoRA (think) | 0.743 | 81.5% | 0.627 | 0.728 | 90.9% | **0.718** |
| Qwen3-32B (think) | 0.723 | 72.5% | 0.549 | 0.740 | 83.5% | 0.550 |
| *Non-reasoning models* | | | | | | |
| GPT-4.1 | 0.658 | 52.6% | 0.655 | 0.789 | 72.8% | 0.710 |
| DeepSeek-V3-0324 | 0.688 | 42.2% | 0.497 | 0.684 | 42.0% | 0.461 |
| Qwen3-32B (no-think) | 0.637 | 48.0% | 0.471 | 0.704 | 73.3% | 0.510 |

* DeepSeek-R1-0528 was involved in the MEDRECT-ja data synthesis process.

mance variance (26.1%–78.3%), indicating that contextual understanding and patient interaction comprehension remain fundamental areas where current LLMs must be significantly improved for reliable medical deployment. *Diagnosis*, *Procedures/intervention*, and *Medication selection* generally yield higher performance across model categories, suggesting these structured clinical reasoning tasks align well with current LLM capabilities.

Reasoning capabilities and model enhancement strategies show differential impacts across error categories in sentence extraction performance. The Qwen3-32B think vs. no-think comparison reveals particularly large gaps in *History taking* sentence extraction (68.1% vs. 36.2%) and *Physical findings* (68.4% vs. 37.8%), indicating that explicit reasoning processes are especially beneficial for tasks requiring contextual interpretation and clinical observation synthesis. LoRA fine-tuning demonstrates targeted improvements, with the most substantial sentence extraction gains in *History taking* (+10.2 percentage points) and *Physical findings* (+15.4 percentage points) compared to the base Qwen3-32B (think) model. Interestingly, model size does not always predict performance across error types: while gpt-oss-120b outperforms gpt-oss-20b in *Test interpretation* (79.6% vs. 69.4%), the smaller Qwen3-32B (think) achieves superior performance over gpt-oss-120b in *History taking* (68.1% vs. 47.8%), suggesting that reasoning capabilities and task-specific optimization may be more critical than raw model capacity for certain clinical domains.

Model-specific patterns reveal distinct capabilities and limitations across clinical domains. Proprietary models demonstrate superior overall sentence extraction performance, with GPT-5 achieving excellent performance in most error types including *Diagnosis* (95.4%) and *Monitoring/management* (91.7%), while Claude Sonnet 4 excels in *History taking* (65.2%) and *Physical findings* (75.7%). DeepSeek-R1-0528 shows remarkably consistent sentence extraction performance across all error types (above 60%), suggesting robust general-purpose medical reasoning capabilities. The pronounced difficulty of *Medication dosage* across multiple high-performing models points to fundamental challenges in numerical precision and dosage calculation that persist even in advanced systems, representing a critical area for continued development in medical AI safety.

### 5.4 Qualitative Analysis

Manual inspection of model outputs reveals distinct patterns in error correction performance across different error types and clinical scenarios. Table 4 presents three representative cases that illustrate critical dimensions of medical error correction:

Table 4: Error correction examples on three representative MEDRECT samples

| Sample | Sample 1: 119B36_a_Deepseek-R1-0528 (ja) | Sample 2: 118E37_c_Qwen3-235B-A22B-Thinking-2507 (ja) | Sample 3: ms-test-120 (en) |
|---|---|---|---|
| **Error Type** | *Procedures/intervention* | *History taking* | *(No error)* |
| **Clinical Text** | 82-year-old female admitted to palliative care ward due to pancreatic cancer with liver metastasis. Progressive appetite loss and decreased food intake for 1 week [...]. Physical findings: decreased skin turgor, poor oral hygiene and dryness [...]. Blood findings: [...] Hb 9.2g/dL [...]. Biochemistry: [...] albumin 2.8g/dL [...]. **Nutrition support team proposed gastrostomy for nutritional support.** Plan to explain to attending physician and obtain consent. | 80-year-old female diagnosed with lung cancer and multiple bone metastases 1 month ago. She chose not to receive active treatment and receives home visits [...]. She mainly stays in bed during the day. Living alone since husband's death at age 65, with distant contact from son [...]. She told the visiting physician "I want to end it all." **The physician asked "What is the most distressing physical symptom you have right now?"** The patient did not respond and remained lying in bed. | 45-year-old woman presents complaining of weakness that has progressively worsened over the past 2 weeks. She states that she has a hard time lifting both her arms but that they function normally [...]. On examination, patient has 2/5 muscle strength on shoulder shrug and arm abduction bilaterally, but all other neurological exam findings are normal. You notice some skin changes [...]. She states that she has had a rash around her eyes as well as on her lower face, going down to her neck and chest. She notes that the rashes started around the same time as the weakness began. Patient is diagnosed with dermatomyositis after labs are drawn and a complete blood count and basic metabolic panel are normal. |
| **Gold Standard** | Nutrition support team should prioritize consideration of non-invasive nutrition management methods. | The physician responded "You're feeling distressed. Could you tell me more about your feelings?" | CORRECT |
| **GPT-5** | ✓ Nutrition support team judged that gastrostomy is inappropriate given the terminal cancer with limited prognosis, and proposed non-invasive nutritional support prioritizing symptom palliation with oral intake strategies, oral care, and supplementary fluids as needed. | △ The physician first assessed for suicidal ideation and urgency (specific plans, means, intent), ensured safety, and then comprehensively assessed physical, psychological, social, and spiritual distress. | × Patient is diagnosed with dermatomyositis after labs show elevated muscle enzymes (e.g., creatine kinase), while the complete blood count and basic metabolic panel may be normal. |
| **Claude Sonnet 4** | ✓ Nutrition support team proposed oral intake support for symptom palliation. | △ The physician asked "Could you tell me more about those feelings?" | × Patient is diagnosed with dermatomyositis after labs are drawn including elevated muscle enzymes and positive autoantibodies, while complete blood count and basic metabolic panel are normal. |
| **Qwen3-32B + LoRA (think)** | ✓ Nutrition support team proposed coordination with home care, prioritizing continuation of oral intake above all else. | ✓ The physician responded "You must be feeling very tired, being alone. I understand your feelings." | ✓ CORRECT |
| **Qwen3-32B (think)** | △ Nutrition support team is considering introduction of enteral nutrition (gastrostomy placement) as a nutritional supplement method based on palliative care goals and the patient's own decision-making. | △ The physician failed to properly evaluate the patient's statement and focused solely on physical symptoms, potentially missing psychosocial factors. | × Patient is diagnosed with dermatomyositis after labs show elevated creatine kinase, erythrocyte sedimentation rate, and characteristic skin findings consistent with heliotrope rash and shawl sign. |

**Bold** in Clinical Text indicates the sentence containing the medical error.
MEDRECT-ja samples (119B36_a_Deepseek-R1-0528 and 118E37_c_Qwen3-235B-A22B-Thinking-2507) are translated to English for readability.

**Performance:** ✓ Perfect △ Partial × Failure

procedural judgment in palliative care, empathetic communication in patient interactions, and restraint against false positive corrections.

The procedural error example (Sample 1) demonstrates models' understanding of palliative care principles. Most models correctly identify that gastrostomy placement is inappropriate for a terminally ill patient with limited prognosis, with the fine-tuned model and proprietary systems proposing non-invasive alternatives prioritizing comfort care. This pattern indicates robust comprehension of end-of-life care guidelines across different model architectures.

The history-taking error (Sample 2) reveals significant variation in models' ability to provide empathetic responses to patient distress. When a patient expresses "I want to end it all," the physician's response of asking about physical symptoms demonstrates poor empathetic understanding. While GPT-5 and Claude Sonnet 4 attempt clarification, their responses lack warmth and emotional support, earning partial credit. The LoRA fine-tuned model excels by providing an empathetic response acknowledging the patient's loneliness and emotional state. This highlights how fine-tuning can enhance models' patient-centered communication capabilities beyond mere clinical knowledge.

The correct sample (Sample 3) reveals models' tendency toward false positive error detection. Several models, including GPT-5, Claude Sonnet 4, and the base Qwen3-32B, incorrectly flag already-accurate diagnostic text as requiring correction, proposing unnecessary additions about laboratory findings. Only the LoRA fine-tuned model correctly identifies that no correction is needed. This pattern highlights a practical deployment concern: overly sensitive error detection could burden healthcare practitioners with unnecessary review of false alarms, reducing system utility in clinical workflows.

## 6 Discussion

The wide variance in sentence extraction performance across models (42.2%–83.7%) indicates that identifying the specific erroneous sentence within clinical text represents a significant bottleneck in the error correction pipeline. This finding suggests that precise localization of errors within clinical narratives requires more sophisticated understanding than binary error detection. Notably, reasoning models consistently outperform their non-reasoning counterparts in sentence extraction accuracy (reasoning models: 71.4%–83.7% vs. non-reasoning models: 42.2%–52.6%), demonstrating that explicit reasoning processes are particularly crucial for accurate error localization within complex medical texts.

Reasoning-enabled models consistently outperform non-reasoning counterparts across three comparisons: DeepSeek-R1-0528 vs. DeepSeek-V3-0324, gpt-oss models (varying reasoning effort), and Qwen3-32B (think vs. no-think). Claude Sonnet 4 achieves the highest error detection F1 (0.795) with stable cross-lingual performance (0.795 Japanese, 0.784 English), demonstrating that reasoning can bridge open-source and proprietary performance gaps.

LoRA fine-tuning reveals asymmetric cross-lingual transfer effects, with English error correction performance improving substantially more than Japanese (English: +0.168, 30.5% relative gain vs. Japanese: +0.078, 14.2% relative gain), despite Japanese training data being more than twice as large (5,538 vs. 2,439 samples). This suggests that medical reasoning patterns learned from Japanese clinical scenarios effectively transfer to enhance English error correction capabilities. The finding indicates that fundamental error detection skills transcend language barriers, opening opportunities for efficient multilingual medical error correction systems.

The substantial performance improvements from LoRA fine-tuning demonstrate effective bilingual knowledge transfer while preserving reasoning capabilities. Most significantly, our fine-tuned model achieves superior performance compared to medical doctors in sentence extraction and error correction on the original MEDEC benchmark (Appendix Table 7). Specifically, our fine-tuned Qwen3-32B + LoRA (think) model achieves 90.6% sentence extraction accuracy compared to 76.7% and 64.6% for Medical Doctors #1 and #2 respectively, and 0.714 average correction score compared to their 0.491 and 0.678, while achieving 62.0% error detection accuracy compared to their 81.3% and 68.9% due to higher sensitivity that results in more false positives on correct texts. The qualitative analysis further demonstrates that fine-tuning enhances clinically relevant capabilities beyond metric improvements. Our LoRA fine-tuned model excels in empathetic patient communication and appropriately restrains from overcorrecting already-accurate text, addressing two critical concerns for practical deployment in healthcare settings. This represents a paradigm shift where properly fine-tuned reasoning models can surpass human expert performance while maintaining explainable reasoning processes—a critical milestone for deploying trustworthy AI systems in medical practice.

## 7 Conclusion

We introduce MEDRECT, the first cross-lingual benchmark for medical error detection and correction, bridging critical evaluation gaps in medical LLMs beyond English. Our scalable automated methodology enables systematic evaluation across Japanese and English clinical contexts.

Through comprehensive evaluation of 9 contemporary LLMs, we establish that reasoning capabilities are fundamental for medical error correction, with substantial performance advantages for reasoning models. Cross-lingual evaluation reveals persistent challenges in multilingual deployment, while targeted fine-tuning provides a viable pathway for practical implementation while preserving model reasoning abilities.

These findings underscore the complexity of medical error correction and highlight essential considerations for safe, equitable deployment of AI systems in healthcare. MEDRECT provides the research community with the tools and insights necessary to advance medical AI safety across languages and cultures.

## 8 Limitations

**Dataset Limitations** The dataset size is constrained by the availability of suitable Japanese medical licensing examination questions. From 800 questions across two examination years (JMLE 2024 and 2025), a majority of short-form knowledge questions without clinical case scenarios could not be utilized for our task formulation. After further excluding image-based questions, calculation problems, and questions with underlined text that complicate reformatting, only 287 clinical case questions remained as viable source material. This resulted in 663 samples for MEDRECT-ja after the synthesis and quality filtering processes. Additionally, our synthetic error generation approach, while systematic, may not fully represent the diversity of errors encountered in actual clinical practice.

**Model-Specific Biases** The dataset construction pipeline relies on specific models at multiple steps (DeepSeek-R1-0528 and Qwen3-235B-A22B-Thinking-2507 for synthesis, Gemini 2.5 Pro for final quality screening, and 11 validation models including Qwen3-32B variants for difficulty-based filtering in Step 2), potentially introducing model-specific biases into the benchmark. In particular, models used for quality filtering may have an advantage in subsequent benchmark evaluation. However, we note that the difficulty-based filtering in Step 2 does not necessarily favor the filtering models themselves—it selects samples with moderate difficulty (accuracy between 1/11 and 7/11 across validation models) rather than easy samples that would artificially inflate their performance. The multi-model consensus approach (11 diverse validation models) further mitigates individual model bias. Nevertheless, we acknowledge that these models' benchmark results should be interpreted with this methodological consideration in mind.

**Evaluation Constraints** Automated evaluation metrics, though comprehensive, cannot entirely substitute for expert clinical judgment in assessing correction quality.

**Scope Limitations** This study focuses exclusively on text-based scenarios and does not address multimodal clinical documents containing images, tables, or other visual elements commonly found in real clinical settings.

**Potential Risks** While MEDRECT advances medical error detection evaluation, it introduces potential risks warranting careful consideration. A primary concern is the possibility of healthcare professionals' overreliance on automated error detection systems, which could lead to oversight of subtle clinical errors not captured by current models. Our results show significant performance variation across error types, with particularly low accuracy in medication dosage errors (42.2% sentence extraction in Table 6), highlighting risks in high-stakes clinical domains. Additionally, cross-lingual performance asymmetries (Table 3) raise fairness concerns for non-English medical contexts. The dual-use potential of error correction capabilities—such as generating misleading medical information—requires acknowledgment. To mitigate these risks, we emphasize that MEDRECT is designed for research evaluation purposes, not direct clinical deployment. Any real-world application must maintain human-in-the-loop frameworks where clinical professionals critically validate all automated predictions, implement gradual deployment protocols, and establish continuous monitoring systems.

## Acknowledgments

We thank the anonymous reviewers for their valuable feedback.

## References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Anthropic. 2025. System card: Claude opus 4 and claude sonnet 4. https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf. Accessed: 2025-09-11.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 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, Mexico City, Mexico. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2412.19260*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Panupong Pasupat, and 1 others. 2025. Gemini 2.5: pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and 1 others. 2025a. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and 1 others. 2025b. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*, 9:e45312.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2024. Jmedbench: a benchmark for evaluating japanese biomedical large language models. *arXiv preprint arXiv:2409.13317*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to ask in english: cross-lingual evaluation of large language models for healthcare queries. *arXiv preprint arXiv:2310.13132*.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Wataru Kawakami, Keita Suzuki, and Junichiro Iwasawa. 2025. Stabilizing reasoning in medical llms with continued pretraining and reasoning preference optimization. *arXiv preprint arXiv:2504.18080*.

Yoshimasa Kawazoe, Daisaku Shibata, Emiko Shinohara, Eiji Aramaki, and Kazuhiko Ohe. 2021. A clinical specific bert developed using a huge japanese clinical text corpus. *PLOS ONE*, 16(11):1–11.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning domain-specialised representations for cross-lingual biomedical entity linking. *arXiv preprint arXiv:2105.14398*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Martin A Makary and Michael Daniel. 2016. Medical error—the third leading cause of death in the us. *BMJ*, 353.

Birger Moëll, Fredrik Sand Aronsson, and Sanian Akbar. 2025. Medical reasoning in llms: an in-depth analysis of deepseek r1. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

OpenAI. 2025a. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`. Accessed: 2025-09-11.

OpenAI. 2025b. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

OpenAI. 2025c. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-09-11.

OpenAI. 2025d. Openai o3 and o4-mini system card. `https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf`. Accessed: 2025-09-11.

Preferred Networks, Kaizaburo Chubachi, Yasuhiro Fujita, Shinichi Hemmi, Yuta Hirokawa, Toshiki Kataoka, Goro Kobayashi, Kenichi Maehashi, Calvin Metzger, Hiroaki Mikami, Shogo Murai, Daisuke Nishino, Kento Nozawa, Shintarou Okada, Daisuke Okanohara, Shunta Saito, Shotaro Sano, Shuji Suzuki, Daisuke Tanaka, and 4 others. 2025. Plamo 2 technical report. *arXiv preprint arXiv:2509.04897*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. Accessed: 2025-09-11.

Gustavo Saposnik, Donald Redelmeier, Christian C. Ruff, and Philippe N. Tobler. 2016. Cognitive biases associated with medical decisions: a systematic review. *BMC Medical Informatics and Decision Making*, 16(138).

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. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, M. A. Mintz, ..., and P. A. Chou. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Issey Sukeda. 2024. Development and bilingual evaluation of japanese medical large language model within reasonably low computational resources. *arXiv preprint arXiv:2409.11783*.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023. Clinical note summarization from doctor-patient conversations. In *Proceedings of the 6th Clinical NLP Workshop (ClinicalNLP)*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Naoto Usuyama, Cliff Wong, Sheng Zhang, Tristan Naumann, and Hoifung Poon. 2025. Biomedical natural language processing in the era of large language models. *Annual Review of Biomedical Data Science*, 8:471–490.

Daniel Van Veen and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*.

Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2019. Tweet classification toward twitter-based disease surveillance: New data, methods, and evaluations. *J Med Internet Res*, 21(2):e12783.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. ACI-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*.

Guangtao Zeng, Xing He, Xuehai Chen, and 1 others. 2020. Meddialog: Large-scale medical dialogue datasets. In *EMNLP*.

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*. OpenReview.net.

# A Details for MEDRECT Dataset Construction

## A.1 Complete Pipeline Details

This section provides comprehensive technical details for the data construction pipeline described in Section 3.2.

### A.1.1 Stage 1: Automated Synthesis Details

We extracted 287 clinical case questions from JMLE (2024 and 2025), focusing on clinical case questions that described patient scenarios while excluding image-based questions, calculation problems, and questions with underlined text that would complicate reformatting. For each JMLE question, we generated clinical texts by incorporating each answer choice into the original clinical scenario, creating CORRECT samples (from correct choices) and ERROR samples (from wrong choices).

We used two reasoning LLMs for synthesis:

- DeepSeek-R1-0528 (DeepSeek-AI et al., 2025a)
- Qwen3-235B-A22B-Thinking-2507 (Yang et al., 2025)

Errors were automatically categorized into eight clinical domains based on the incorrect answer choices: history taking, physical findings, test interpretation, diagnosis, monitoring/management, medication selection, medication dosage, and procedures/intervention. This scalable process transformed 287 questions into 2,792 candidate samples.

### A.1.2 Stage 2: Consensus-Based Quality Filtering Details

We evaluated each synthesized sample by having 8 validation models (11 configurations when distinguishing think/no-think modes) solve the error detection and sentence extraction tasks. The validation models were:

- Gemini 2.5 Pro (Comanici et al., 2025)
- GPT-4.1 (OpenAI, 2025c)
- PLaMo 2.0 Prime (Preferred Networks et al., 2025)

- Qwen3-8B (Yang et al., 2025)
- Qwen3-14B (Yang et al., 2025)
- Qwen3-32B (Yang et al., 2025)
- Qwen3-30B-A3B-Thinking-2507
- QwQ-32B (Qwen Team, 2025; Yang et al., 2024)

Note that Qwen3-8B, Qwen3-14B, and Qwen3-32B were each evaluated in both think and no-think modes, resulting in 11 total model configurations.

Based on their performance consensus, we filtered samples using the following criteria:

- For CORRECT samples, we retained samples where the error detection accuracy of the validation models fell within the range $1/11 \leq$ accuracy $\leq 7/11$.
- For ERROR samples, we applied a stricter standard, retaining samples only if their sentence extraction accuracy was within $1/11 \leq$ accuracy $\leq 7/11$ *and* the gap between detection and extraction accuracy was minimal ($\leq 3/11$).

This process filtered the dataset from 2,792 to 1,057 high-quality samples, ensuring both validity and appropriate difficulty for benchmarking.

#### A.1.3 Stage 3: Model Deduplication Details

Since both synthesis models (DeepSeek-R1-0528 and Qwen3-235B-A22B-Thinking-2507) processed identical JMLE source questions, our pipeline generated duplicate samples from the same source (question, answer choice) pairs. To remove these duplicates while maintaining balanced representation from both models, we alternately selected one sample from each model for every duplicate pair. This reduced the dataset from 1,057 to 720 samples.

#### A.1.4 Stage 4: LLM-as-a-Judge Quality Screening Details

We employed LLM-as-a-Judge (Gemini 2.5 Pro) to perform binary classification on quality dimensions. For MEDRECT-ja, we assessed five dimensions:

- *ambiguous_error*: Medical statements with unclear correctness
- *extra_elements*: Addition of information not in original problem/choices
- *multiple_errors*: Multiple error locations in ERROR data
- *numerical_error*: Numerical errors difficult to correct from context
- *synthesis_consistency_error*: Wrong choice used but medically correct content

For MEDRECT-en, because samples originated from existing MEDEC data without original JMLE questions, we adapted the criteria by replacing *extra_elements* and *synthesis_consistency_error* with two analogous dimensions:

- *unrealistic_scenario*: Clinically implausible scenarios
- *inconsistent_context*: Internal inconsistencies in clinical context

Any sample scoring 1 (problematic) on any dimension was excluded from the final dataset. This rigorous screening produced 663 samples for MEDRECT-ja (92.1% retention) and 458 samples for MEDRECT-en (76.7% retention).

### A.2 Data Synthesis Prompt

The complete synthesis prompt that was used to transform JMLE questions into clinical texts:

**Data Synthesis Prompt (English Translation)**

```
Convert the following Japanese medical licensing examination question into clinical cases in MEDEC (Medical Error Detection and Correction) format.

# Instructions
- For each answer choice, synthesize one clinical record incorporating that choice into the problem text, creating 5 records total.
- If the choice is correct, synthesize a correct record; if wrong, synthesize a record containing an error.
- Clinical records should always be written as numbered markdown lists, with error-containing records having exactly one sentence with a clinical error.
- These records will be used for MEDEC format benchmarks. Do not indicate where errors are located in error-containing records.
- Include all numerical values and findings from the original problem without summarization or omission.
- Do not add original medical interpretations not present in the original problem or choices.

# Original Medical Licensing Examination Problem
Problem: {question}
Choices: {choices_text}
Correct choices: {correct_choices_list}
Wrong choices: {wrong_choices_list}

# Synthesis Format

The following shows examples of correct records synthesized from correct choices and error records from wrong choices:

### Choice {correct_choices_list}[0] Record (CORRECT sample)
```

```
1. ...
2. ...
...
N. ...

### Choice {wrong_choices_list}[0] Record (ERROR sample)
1. ...
2. ...
...
N. ...
Error Type: [Select from: history taking, physical findings, test interpretation, diagnosis, management, pharmacotherapy, procedures]
Error Sentence Number: [Number of sentence containing the medical error]
Error Sentence: [The sentence with wrong medical content]
Corrected Sentence: [The medically accurate version of the sentence]

(Few-shot examples truncated for brevity...)
```

### A.3 Quality Screening Prompt

The prompt that was used for LLM-as-a-Judge quality assessment with Gemini 2.5 Pro:

Quality Screening Prompt (English Translation)

```
## MEDEC Benchmark Quality Assessment

### Original Medical Examination Problem
Question: {original_question}
Choices: {choices_text}
Correct answer: {correct_answer}
**Used choice: {used_choice_text}**

### Generated MEDEC Format Text
{sentences}

### Error Information
{error_info}

### Assessment Task
Please evaluate the **quality as a benchmark problem** for the above generated text.

Rate the following aspects as 1 (problematic) or 0 (acceptable):

1. **ambiguous_error**: Medical statements with unclear correctness
2. **extra_elements**: Addition of information not in original problem/choices
3. **multiple_errors**: Multiple error locations in ERROR data
4. **numerical_error**: Numerical errors difficult to correct from context
5. **synthesis_consistency_error**: Wrong choice used but medically correct content

JSON response: {"ambiguous_error": 0, "extra_elements": 0, "multiple_errors": 0, "numerical_error": 0, "synthesis_consistency_error": 0, "explanation": "Brief assessment"}
```

**Note:** For MEDRECT-en dataset construction, the above criteria were adapted to account for differences in source material characteristics. Specifically, *extra_elements* and *synthesis_consistency_error* were replaced with *unrealistic_scenario* and *inconsistent_context* to better suit the pre-existing clinical texts in the MEDEC dataset.

### A.4 Quality Screening Results

To ensure robust quality assessment, we applied both 0-shot and 2-shot prompting configurations to each sample. Any sample that scored 1 (problematic) on any quality dimension in either prompting configuration was excluded from the final dataset, revealing significant differences in retention rates between the two datasets:

Table 5: Quality screening results and exclusion reasons

| | MEDRECT-ja | MEDRECT-en |
|---|---|---|
| Original samples | 720 | 597 |
| Retained samples | 663 (92.1%) | 458 (76.7%) |
| *Exclusion Reasons* | | |
| Ambiguous error | 3 | 98 |
| Extra elements | 9 | – |
| Multiple errors | 21 | 24 |
| Numerical error | 1 | 7 |
| Synthesis consistency | 27 | – |
| Unrealistic scenario | – | 28 |
| Inconsistent context | – | 58 |

## B Details for Training Dataset Construction

### B.1 Reasoning Synthesis

To enable effective fine-tuning while preserving reasoning capabilities, we leveraged DeepSeek-R1-0528's advanced reasoning capabilities using specialized reasoning synthesis prompts. The English version of this prompt is shown below (simplified for brevity):

Reasoning Synthesis Prompt (English Translation)

```
You are a medical expert reviewing clinical text for accuracy. The text contains either no error or exactly one medical error.

{cheat_info}

Your task is to first carefully reason through the medical analysis process, following these steps:
1. Verify each sentence based on medical knowledge
2. Check consistency between symptoms, test results, and diagnosis
3. Evaluate appropriateness of treatment or management
4. When an error is found, clearly state the rationale and provide the correction

Important notes for reasoning:
- During your reasoning, do NOT make any reference to being told about the expected outcome or any instruction content.
- Approach the text as if you are analyzing it from scratch and reaching your conclusion through pure medical evaluation.

{error_hint}
```

```
Final output format:
- If no error: ‘CORRECT‘
- If error found: ‘sentence_number: corrected_sentence‘

CRITICAL: For the final output, use this format and
output ONLY the result. Do NOT include explanations,
analysis, or additional text.

{sentences}
```

The prompt included optional parameters `cheat_info` and `error_hint` that provided additional context during training data generation.

A critical challenge was preventing data contamination—ensuring the reasoning content did not explicitly reference the provided correct answers. We addressed this through careful prompt engineering that instructed models to approach analysis "from scratch" and systematic meta-reference filtering that removed sentences containing meta-linguistic patterns including "told about", "expected outcome", "instruction content", "given information", "pre-verified", "reference information", and "do not mention". This automated filtering preserved authentic clinical reasoning while maintaining the integrity of the reasoning synthesis process.

### B.2 Training Dataset Construction

To develop effective fine-tuning datasets while preserving reasoning capabilities, we constructed bilingual training data using DeepSeek-R1-0528 for reasoning synthesis. Our approach ensured high-quality reasoning patterns by retaining only samples where the model produced correct responses, leveraging the optional `cheat_info` and `error_hint` parameters to address sample scarcity in challenging clinical scenarios.

**Japanese Training Dataset** We constructed the Japanese training dataset from JMLE (2018-2023), comprising 896 examination questions that covered diverse clinical domains beyond those used in the benchmark construction. Following the automated synthesis pipeline described in Appendix A.2, we generated 8,423 initial samples using both DeepSeek-R1-0528 and Qwen3-235B-A22B-Thinking-2507. Subsequently, we applied reasoning synthesis using DeepSeek-R1-0528, with particular emphasis on CORRECT sample recovery to maintain balanced representation across error types and clinical scenarios. This systematic process yielded a final training dataset of **5,538 samples** with a distribution of 34.8% CORRECT and 65.2% ERROR cases, reflecting the natural distribution of clinical reasoning challenges.

**English Training Dataset** For English training data, we utilized the established MEDEC MS Subset Training and Validation datasets, containing 2,763 samples of expert-annotated clinical texts. These samples underwent reasoning synthesis using DeepSeek-R1-0528 with the same reasoning synthesis prompts employed for the Japanese dataset, ensuring consistency in reasoning quality and style across languages. The resulting English training dataset comprised **2,439 samples** with 49.0% CORRECT and 51.0% ERROR distribution, providing robust cross-lingual training coverage.

## C Additional Results

Table 6 breaks down model performance by clinical error type, revealing differential strengths across diagnostic categories. Table 7 compares performance on the original MEDEC benchmark (MS Subset, 597 samples) before quality screening (see Appendix A.4), demonstrating evaluation framework consistency and model performance on the unfiltered dataset.

Table 6: Sentence Extraction Accuracy by Error Type on MEDRECT-ja + MEDRECT-en. Parenthetical notations indicate reasoning effort levels (gpt-oss: medium) or reasoning modes (Qwen3-32B: think/no-think). Top 8 most frequent error types are included (11–175 samples each).

| Model | Diagnosis | Monitoring/ management | Physical findings | Procedures/ intervention | Medication selection | Test interpretation | History taking | Medication dosage |
|---|---|---|---|---|---|---|---|---|
| *Reasoning models* | | | | | | | | |
| GPT-5 | **95.4%** | **91.7%** | 73.0% | **93.6%** | **96.0%** | 77.6% | 47.8% | **100.0%** |
| o3 | 88.6% | 70.8% | 58.1% | 83.3% | 92.0% | 65.3% | 30.4% | **100.0%** |
| Claude Sonnet 4 | 90.9% | 81.2% | 75.7% | 80.8% | 87.0% | 69.4% | 65.2% | **100.0%** |
| DeepSeek-R1-0528* | 81.7% | 83.3% | 70.3% | 80.8% | 84.0% | 59.2% | 60.9% | **100.0%** |
| gpt-oss-120b (medium) | 87.4% | 71.9% | 71.6% | 85.9% | 92.0% | **79.6%** | 47.8% | **100.0%** |
| gpt-oss-20b (medium) | 81.1% | 62.5% | 51.4% | 78.2% | 92.0% | 69.4% | 30.4% | 90.9% |
| Qwen3-32B + LoRA (think) | 93.7% | 80.2% | **83.8%** | 87.2% | 87.0% | 75.5% | **78.3%** | 27.3% |
| Qwen3-32B (think) | 78.7% | 61.8% | 68.4% | 75.3% | 84.7% | 66.6% | 68.1% | 54.5% |
| *Non-reasoning models* | | | | | | | | |
| GPT-4.1 | 64.3% | 50.8% | 53.8% | 56.4% | 70.5% | 49.8% | 52.2% | 66.7% |
| DeepSeek-V3-0324 | 46.9% | 52.1% | 44.6% | 35.9% | 42.0% | 18.4% | 39.1% | 36.4% |
| Qwen3-32B (no-think) | 67.2% | 52.6% | 37.8% | 49.9% | 66.8% | 72.9% | 36.2% | 42.4% |

* DeepSeek-R1-0528 was involved in the MEDRECT-ja data synthesis process.

Table 7: Performance on original MEDEC benchmark (MS Subset). Parenthetical notations indicate reasoning effort levels (gpt-oss: high/medium/low) or reasoning modes (Qwen3-32B: think/no-think).

| Model | Error Detection | | Sent. Ext. | Error Correction | | | |
|---|---|---|---|---|---|---|---|
| | F1 | Acc. | Acc. | ROUGE-1 | BERT | BLEURT | Avg. |
| *MEDEC Paper Results* | | | | | | | |
| Medical Doctor #1 | - | **81.3%** | 76.7% | 0.420 | 0.513 | 0.539 | 0.491 |
| Medical Doctor #2 | - | 68.9% | 64.6% | 0.685 | 0.698 | 0.650 | 0.678 |
| *Reasoning models* | | | | | | | |
| GPT-5 | 0.780 | 71.7% | **90.7%** | 0.655 | 0.672 | 0.671 | 0.666 |
| o3 | **0.783** | 75.0% | 80.7% | 0.658 | 0.680 | 0.677 | 0.672 |
| Claude Sonnet 4 | 0.737 | 67.2% | 75.2% | 0.640 | 0.667 | 0.653 | 0.653 |
| DeepSeek-R1-0528* | 0.701 | 58.3% | 71.1% | 0.549 | 0.576 | 0.573 | 0.566 |
| gpt-oss-120b (high) | 0.733 | 62.5% | 87.1% | 0.606 | 0.633 | 0.633 | 0.621 |
| gpt-oss-120b (medium) | 0.742 | 66.2% | 82.0% | 0.582 | 0.605 | 0.606 | 0.598 |
| gpt-oss-120b (low) | 0.740 | 67.5% | 74.0% | 0.566 | 0.594 | 0.592 | 0.584 |
| gpt-oss-20b (high) | 0.726 | 63.3% | 78.8% | 0.554 | 0.581 | 0.590 | 0.575 |
| gpt-oss-20b (medium) | 0.736 | 65.5% | 83.3% | 0.540 | 0.573 | 0.580 | 0.564 |
| gpt-oss-20b (low) | 0.694 | 63.1% | 67.8% | 0.458 | 0.495 | 0.515 | 0.489 |
| Qwen3-32B + LoRA (think) | 0.723 | 62.0% | 90.6% | **0.711** | **0.748** | **0.684** | **0.714** |
| Qwen3-32B (think) | 0.711 | 60.5% | 77.5% | 0.480 | 0.509 | 0.546 | 0.512 |
| *Non-reasoning models* | | | | | | | |
| GPT-4.1 | 0.726 | 72.9% | 65.6% | 0.683 | 0.697 | 0.681 | 0.687 |
| DeepSeek-V3-0324 | 0.671 | 54.6% | 38.6% | 0.399 | 0.428 | 0.471 | 0.432 |
| Qwen3-32B (no-think) | 0.688 | 57.6% | 69.8% | 0.461 | 0.486 | 0.517 | 0.488 |

* DeepSeek-R1-0528 was involved in the MEDRECT-ja data synthesis process.