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

**Anonymous ACL submission**


## Abstract

Large language models (LLMs) show promise in medical applications, but their ability to *detect and correct errors in clinical texts* remains under-evaluated, particularly beyond English. We introduce MEDRECT, a bilingual benchmark (Japanese/English) that formulates medical error handling as three subtasks: error detection, error sentence extraction, and error correction. MEDRECT is built with a scalable, automated pipeline from the Japanese Medical Licensing Examinations (JMLE) and an English counterpart curated from MEDEC, yielding MEDRECT-ja (663 texts) and MEDRECT-en (458 texts). We evaluate 11 LLMs spanning proprietary, open-weight, medical-domain, 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) general-purpose reasoning models outperform medical-domain models on detection and error sentence extraction, suggesting that reasoning capability matters more than domain-specific training; (iii) most models rank similarly within each language but typically score lower on the Japanese benchmark than on the English one, with the magnitude of the within-model difference varying substantially across models; (iv) LoRA fine-tuning yields asymmetric bilingual improvements on the error correction average score (Japanese: +16.8%, English: +19.6% relative gain) while preserving reasoning capabilities. MEDRECT is the first comprehensive bilingual benchmark for medical error correction, providing resources for developing safer medical LLMs across languages.


## 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 that evaluate clinically relevant capabilities such as error detection and correction, not just question-answering accuracy.

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., 2025), and the English-only MEDEC benchmark (Ben Abacha et al., 2024, 2025) relies on resource-intensive manual annotation that does not scale to other 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 bilingual benchmark for medical error detection and correction (Figure 1). 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

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

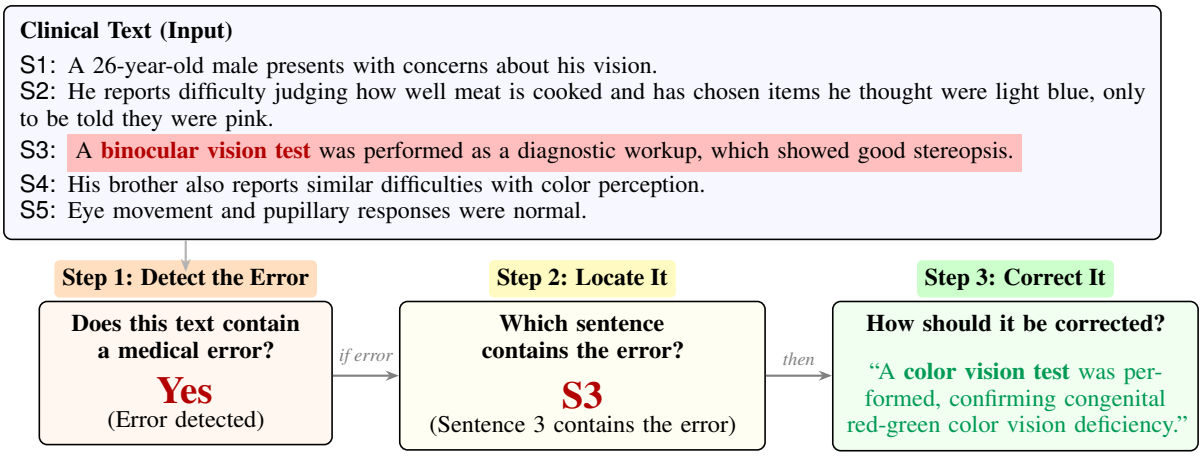


Figure 1: Overview of the MEDRECT benchmark, illustrated with a MEDRECT-ja sample (translated to English). Given a clinical text, models must (1) detect whether it contains a medical error, (2) identify the erroneous sentence, and (3) generate a corrected version. Here, the patient's symptoms (difficulty distinguishing colors) and family history suggest congenital color vision deficiency, but a binocular vision test was ordered instead of the appropriate color vision test.

bilingual benchmarks. **(2) Bilingual Evaluation**: MEDRECT-ja (663 samples) and MEDRECT-en (458 samples) with diverse error types, evaluated on 11 LLMs including medical-domain models, revealing substantial bilingual gaps and the critical role of reasoning over domain-specific training. **(3) Parameter-Efficient Fine-Tuning**: LoRA (Low-Rank Adaptation) fine-tuning (Hu et al., 2022) 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., 2021), 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).

### 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 Multilingual 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., 2025). For multilingual evaluation, prior work explored medical QA performance variations across languages (Jin et al., 2024; Alonso et al., 2024) and biomedical entity linking (Liu et al., 2021). MEDRECT is the first to provide systematic bilingual evaluation for medical error correction, a complex unstructured task requiring clinical reasoning across languages.

## 3 MEDRECT Benchmark

### 3.1 Task Definition

Following MEDEC, we decompose medical error detection and correction into three progressive subtasks (Figure 1): (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 2 illustrates concrete examples across different error types in both MEDRECT-ja and MEDRECT-en datasets.

**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 2: 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 (colored background) with the key error in **red** and the correction in **green**.

### 3.2 Dataset Construction

Both MEDRECT-ja and MEDRECT-en consist of structured clinical vignettes (chief complaints, examination findings, laboratory results, and clinical decision points) derived from medical licensing examinations, rather than free-form clinical notes or EHR records. This scope isolates medical reasoning ability from noisy-text processing challenges.

For MEDRECT-ja, we extracted 287 clinical case questions from JMLE (2024 and 2025) and applied four stages (see Figure 3 in Appendix for the complete pipeline): (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 bilingual 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 breadth of clinical scenarios covered by the JMLE. 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., 2021) and similarly contains no PII.

## 4 Experimental Setup

### 4.1 Evaluated Models

We evaluated 11 contemporary LLMs[2]: **Reasoning models** (GPT-5 (OpenAI, 2025a), o3 (OpenAI, 2025d), Claude Sonnet 4 (Anthropic, 2025), gpt-oss-120b/20b (OpenAI, 2025b), Qwen3-32B (think) (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 (no-think)) do not; **Medical-domain models** (HuatuoGPT-o1-72B (Chen et al., 2025), MedResearcher-R1-32B (Li et al., 2025), MedGemma 27B (Google DeepMind, 2025)) are specifically trained or fine-tuned for medical tasks. Qwen3-32B supports both think and no-think modes within the same architecture, enabling controlled comparison of reasoning impact; we list it in both categories accordingly. 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.

### 4.2 Evaluation Metrics

We report Error Detection F1, Sentence Extraction Accuracy, and Error Correction scores using ROUGE-1 (Lin, 2004), BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and their arithmetic average. ROUGE-1 uses MeCab (Kudo et al., 2004) for Japanese and whitespace tokenization for English; BERTScore F1 uses microsoft/mdeberta-v3-base (no baseline rescaling) so that the same multilingual encoder scores both languages; BLEURT uses BLEURT-20.

Sentence extraction and correction are computed on samples with a ground-truth error. Correction scores are averaged over all such samples, with misses (ground-truth = error, prediction = NA) assigned 0; this avoids inflating scores for low-recall models that would otherwise be evaluated only on the subset of errors they happened to flag. The Appendix MEDEC comparison (Table 6) instead follows the original MEDIQA-CORR 2024 evaluation protocol (Ben Abacha et al., 2024) for comparability with reported physician baselines.

[2]Proprietary models (GPT-5, GPT-4.1, o3, Claude Sonnet 4) and DeepSeek-V3-0324 were accessed via OpenRouter API, while open-weight models (gpt-oss, Qwen3-32B) and medical-domain models were evaluated using local inference infrastructure.

### 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).

## 5 Results

### 5.1 Impact of Reasoning Capability

Table 2 presents bilingual evaluation results for 11 models on MEDRECT-ja and MEDRECT-en. All numbers in this subsection refer to MEDRECT-ja unless otherwise noted. Claude Sonnet 4 achieves the highest average (0.620) with strong error detection (0.795 F1), followed by GPT-5 (0.577), gpt-oss-120b at high reasoning effort (0.572), and Qwen3-32B + LoRA (0.563). Error detection F1 ranges from 0.558 (MedResearcher-R1) to 0.795 (Claude Sonnet 4), while sentence extraction shows the largest variance (35.4% to 83.7%).

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

| Model | MEDRECT-ja | | | | | | MEDRECT-en | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Detection | Extraction | Correction | | | | Detection | Extraction | Correction | | | |
| | F1 | Acc. (%) | ROUGE | BERT | BLEURT | Avg. | F1 | Acc. (%) | ROUGE | BERT | BLEURT | Avg. |
| *Reasoning models* | | | | | | | | | | | | |
| GPT-5 | .758 | **83.7** | .502 | .712 | .518 | .577 | .818 | **96.3** | **.689** | **.848** | **.697** | **.745** |
| o3 | .764 | 71.4 | .454 | .636 | .458 | .516 | **.852** | 87.7 | .631 | .781 | .644 | .685 |
| Claude Sonnet 4 | **.795** | 82.3 | **.559** | **.755** | **.547** | **.620** | .784 | 84.0 | .642 | .800 | .643 | .695 |
| gpt-oss-120b (high) | .731 | 79.6 | .476 | .733 | .509 | .572 | .759 | 92.6 | .648 | .841 | .670 | .719 |
| gpt-oss-120b (medium) | .721 | 77.4 | .424 | .690 | .469 | .528 | .777 | 88.1 | .593 | .796 | .608 | .666 |
| gpt-oss-120b (low) | .704 | 68.9 | .390 | .667 | .434 | .497 | .775 | 79.4 | .552 | .751 | .568 | .624 |
| gpt-oss-20b (high) | .729 | 71.4 | .429 | .690 | .461 | .527 | .757 | 86.0 | .569 | .782 | .593 | .648 |
| gpt-oss-20b (medium) | .718 | 64.3 | .371 | .651 | .413 | .478 | .762 | 87.2 | .527 | .754 | .558 | .613 |
| gpt-oss-20b (low) | .678 | 46.9 | .281 | .584 | .335 | .400 | .723 | 71.2 | .394 | .633 | .434 | .487 |
| Qwen3-32B+LoRA (think) | .743 | 81.5 | .494 | .715 | .479 | .563 | .728 | 90.9 | .670 | .825 | .648 | .714 |
| Qwen3-32B (think) | .723 | 72.5 | .368 | .646 | .431 | .482 | .740 | 83.5 | .495 | .748 | .549 | .597 |
| *Non-reasoning models* | | | | | | | | | | | | |
| GPT-4.1 | .658 | 52.6 | .304 | .429 | .317 | .350 | .789 | 72.8 | .527 | .646 | .521 | .565 |
| DeepSeek-V3-0324 | .688 | 42.2 | .324 | .624 | .361 | .437 | .684 | 42.0 | .387 | .667 | .437 | .497 |
| Qwen3-32B (no-think) | .637 | 48.0 | .249 | .532 | .300 | .360 | .704 | 73.3 | .442 | .706 | .488 | .545 |
| *Medical-domain models* | | | | | | | | | | | | |
| HuatuoGPT-o1-72B | .702 | 62.1 | .356 | .547 | .378 | .427 | .721 | 68.3 | .504 | .674 | .511 | .563 |
| MedResearcher-R1-32B | .558 | 35.4 | .219 | .364 | .235 | .272 | .663 | 61.7 | .435 | .577 | .417 | .476 |
| MedGemma 27B | .603 | 39.2 | .273 | .481 | .282 | .346 | .658 | 64.6 | .494 | .642 | .466 | .534 |

The critical role of reasoning is demonstrated by two lines of evidence. First, the gpt-oss models show consistent degradation with reduced reasoning effort: gpt-oss-120b achieves 0.572 average score (high), 0.528 (medium), and 0.497 (low). Second, the Qwen3-32B think vs. no-think comparison, which isolates the effect of reasoning within a single architecture, reveals 0.723 vs. 0.637 error detection F1 and 72.5% vs. 48.0% sentence extraction accuracy, representing 13.5% and 51.0% relative improvements respectively.

Medical-domain models lag behind general-purpose reasoning models on error detection and sentence extraction despite being specifically trained for medical tasks. HuatuoGPT-o1-72B (Chen et al., 2025), the strongest medical-domain model, achieves 62.1% sentence extraction accuracy, below the 72.5% of the smaller Qwen3-32B (think). MedGemma 27B (Google DeepMind, 2025) achieves only 39.2% extraction accuracy, below even non-reasoning models. This suggests that for multi-step clinical error correction, the quality of the underlying reasoning engine matters more than medical domain-specific training.

### 5.2 Bilingual Comparison and Fine-tuning

Because MEDRECT-ja and MEDRECT-en are constructed from different source materials (JMLE vs. MedQA-derived MEDEC) and have different error type distributions (Table 1), absolute scores across the two benchmarks are not directly comparable. We therefore focus on within-language rankings and on how the within-model difference between the two scores varies across models; Table 2 shows that most models score lower on MEDRECT-ja than on MEDRECT-en, but the magnitude of this within-model difference varies substantially. For sentence extraction, for example, GPT-4.1 shows a sizeable within-model difference (52.6% on ja, 72.8% on en), whereas some models remain close on error detection F1 (Claude Sonnet 4: 0.795 on ja, 0.784 on en; o3: 0.764 vs. 0.852). Medical-domain models show a distinct within-language pattern: on MEDRECT-en they trail the leading reasoning models on error correction average (0.476 to 0.563 vs. around 0.69 to 0.75 for reasoning leaders), and their sentence extraction is comparatively weak within each language (61.7 to 68.3% on MEDRECT-en, 35.4 to 62.1% on MEDRECT-ja).

To control for differences in error type distribution between MEDRECT-ja and MEDRECT-en (Table 1), we also report macro-averaged sentence extraction accuracy with equal weighting across error types. For models with uneven per-type performance, macro-averaging reveals notable

Table 3: Sentence extraction accuracy (%) by error type on MEDRECT-ja and MEDRECT-en.

| | Diagnosis | | Monitoring/ management | | Physical findings | | Procedures/ intervention | | Test interpretation | | Medication selection | | History taking | | Medication dosage | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ja | en | ja | en | ja | en | ja | en | ja | en | ja | en | ja | en | ja | en |
| *#samples* | *77* | *98* | *79* | *17* | *72* | *2* | *40* | *38* | *37* | *12* | *30* | *70* | *22* | *1* | *8* | *3* |
| *Reasoning models* | | | | | | | | | | | | | | | | |
| GPT-5 | 90.9 | **99.0** | **89.9** | **100.0** | 72.2 | **100.0** | **97.5** | **89.5** | 70.3 | **100.0** | **100.0** | 94.3 | 45.5 | **100.0** | **100.0** | **100.0** |
| o3 | 80.5 | 94.9 | 73.4 | 58.8 | 56.9 | **100.0** | 90.0 | 76.3 | 56.8 | 91.7 | **100.0** | 88.6 | 27.3 | **100.0** | **100.0** | **100.0** |
| Claude Sonnet 4 | 92.2 | 89.8 | 84.8 | 64.7 | 75.0 | **100.0** | 92.5 | 68.4 | 67.6 | 75.0 | 83.3 | 88.6 | 63.6 | **100.0** | **100.0** | **100.0** |
| gpt-oss-120b (medium) | 81.8 | 91.8 | 69.6 | 82.4 | 72.2 | 50.0 | 87.5 | 84.2 | **81.1** | 75.0 | **100.0** | 88.6 | 45.5 | **100.0** | **100.0** | **100.0** |
| gpt-oss-20b (medium) | 76.6 | 84.7 | 59.5 | 76.5 | 50.0 | **100.0** | 77.5 | 78.9 | 62.2 | 91.7 | 83.3 | **95.7** | 27.3 | **100.0** | 87.5 | **100.0** |
| Qwen3-32B+LoRA (think) | **94.8** | 92.9 | 75.9 | **100.0** | **83.3** | **100.0** | 87.5 | 86.8 | 70.3 | 91.7 | 80.0 | 90.0 | **81.8** | 0.0 | 12.5 | 66.7 |
| Qwen3-32B (think) | 72.7 | 85.7 | 74.7 | 76.5 | 58.3 | **100.0** | 82.5 | 73.7 | 78.4 | 83.3 | 93.3 | 87.1 | 59.1 | **100.0** | 62.5 | 66.7 |
| *Non-reasoning models* | | | | | | | | | | | | | | | | |
| GPT-4.1 | 54.5 | 76.5 | 64.6 | 64.7 | 51.4 | **100.0** | 60.0 | 63.2 | 37.8 | 83.3 | 70.0 | 71.4 | 4.5 | **100.0** | 37.5 | 66.7 |
| DeepSeek-V3-0324 | 45.5 | 48.0 | 57.0 | 29.4 | 44.4 | 50.0 | 40.0 | 31.6 | 18.9 | 16.7 | 30.0 | 47.1 | 40.9 | 0.0 | 25.0 | 66.7 |
| Qwen3-32B (no-think) | 58.4 | 81.6 | 45.6 | 64.7 | 45.8 | 50.0 | 47.5 | 57.9 | 51.4 | 75.0 | 43.3 | 72.9 | 40.9 | **100.0** | 12.5 | 66.7 |
| *Medical-domain models* | | | | | | | | | | | | | | | | |
| HuatuoGPT-o1-72B | 74.0 | 80.6 | 64.6 | 64.7 | 41.7 | 50.0 | 72.5 | 52.6 | 54.1 | 58.3 | 83.3 | 62.9 | 40.9 | 0.0 | 87.5 | 66.7 |
| MedResearcher-R1-32B | 51.9 | 63.3 | 31.6 | 47.1 | 26.4 | 50.0 | 40.0 | 47.4 | 27.0 | 50.0 | 56.7 | 74.3 | 13.6 | **100.0** | 0.0 | 33.3 |
| MedGemma 27B | 44.2 | 70.4 | 36.7 | 58.8 | 34.7 | 0.0 | 30.0 | 55.3 | 43.2 | 83.3 | 46.7 | 64.3 | 36.4 | 0.0 | 62.5 | 33.3 |

shifts: Qwen3-32B + LoRA drops from 81.5% (micro) to 73.3% (macro) on MEDRECT-ja, primarily due to 12.5% accuracy on medication dosage (n=8), while GPT-5 remains stable (83.7% micro vs. 83.3% macro). Under macro-averaging, most models still score lower on ja than on en (e.g., GPT-5: 83.3% ja vs. 97.8% en; Qwen3-32B + LoRA: 73.3% ja vs. 78.5% en), indicating that the direction of within-model differences is not merely an artifact of the error type distribution, though macro-averaged MEDRECT-en scores should be interpreted cautiously given small sample sizes in several error categories (Table 1).

Fine-tuning with LoRA demonstrates substantial improvements across both languages, with asymmetric gains favoring English. On MEDRECT-ja, the fine-tuned Qwen3-32B + LoRA (think)[3] achieves 0.563 average score compared to 0.482 for the base model (16.8% relative improvement), with error detection F1 improving from 0.723 to 0.743 and sentence extraction accuracy advancing from 72.5% to 81.5%.

On MEDRECT-en, the within-model improvement is even more pronounced, with average score increasing from 0.597 to 0.714 (19.6% relative improvement) and English sentence extraction accuracy reaching 90.9% post fine-tuning. Fine-tuning thus yields larger within-model gains on English than on Japanese, even though the Japanese portion of the training data is more than twice as large (Section B).

The fine-tuned model also generalizes to the original MEDEC benchmark (Appendix Table 6), achieving 90.6% sentence extraction accuracy compared to 76.7% and 64.6% for two physicians, and 0.714 average correction score compared to their 0.491 and 0.678, while achieving lower error detection accuracy (62.0% vs. 81.3% and 68.9%) due to more false positives. These results should be interpreted cautiously given the small human baseline (two physicians) and the restriction to structured error correction tasks.

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

### 5.3 Performance by Error Type

Performance breakdown across different medical error categories reveals substantial variation in task difficulty and model behavior patterns (Table 3). All numbers in this subsection refer to MEDRECT-ja sentence extraction accuracy unless otherwise noted.

**Error Type Difficulty** *History taking* is the most challenging category, with accuracy ranging from 4.5% to 81.8% across models and the lowest average among all error types. This indicates that contextual understanding of patient interactions remains a fundamental challenge for current LLMs. *Physical findings* and *test interpretation* are also notably difficult. In contrast, *medication selection* is the most tractable category, with most reasoning models achieving 80% to 100%. *Diagnosis* and *procedures/intervention* also yield relatively higher performance, suggesting these structured clinical reasoning tasks align well with current LLM capabilities. *Medication dosage* shows high variance (0.0% to 100.0%), though its small sample size (n=8) limits generalizability.

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. |

**Red** in Clinical Text highlights the key error; **green** in Gold Standard highlights the correction.
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

**Reasoning and Fine-tuning Effects** The Qwen3-32B think vs. no-think comparison reveals particularly large gaps in *medication selection* (93.3% vs. 43.3%, +50.0 percentage points) and *procedures/intervention* (82.5% vs. 47.5%, +35.0 pp), indicating that explicit reasoning is especially beneficial for pharmacological judgment and procedural decision-making. LoRA fine-tuning produces the largest gains in *history taking* (+22.7 pp over the base Qwen3-32B think) and *physical findings* (+25.0 pp), targeting areas where the base model was weakest. However, fine-tuning dramatically reduces *medication dosage* performance (62.5% to 12.5%), likely reflecting overfitting given the limited training samples in this category. Model size does not always predict performance: while gpt-oss-120b outperforms gpt-oss-20b in *test interpretation* (81.1% vs. 62.2%), the smaller Qwen3-32B (think) achieves superior *history taking* accuracy over gpt-oss-120b (59.1% vs. 45.5%), suggesting that reasoning architecture may be more critical than raw capacity for certain clinical domains.

## 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: 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

On MEDRECT-ja, sentence extraction accuracy ranges from 35.4% to 83.7% across models, indicating that error sentence extraction represents the primary bottleneck in clinical error correction. Reasoning models broadly outperform both non-reasoning models (sentence extraction 42.2% to 52.6%) and medical-domain models (35.4% to 62.1%), as confirmed across gpt-oss models (varying reasoning effort), Qwen3-32B (think vs. no-think), and comparisons with HuatuoGPT-o1, MedResearcher-R1, and MedGemma; the only exceptions are low-effort reasoning configurations (e.g., gpt-oss-20b at low effort: 46.9%), which can fall within the non-reasoning range. This indicates that reasoning architecture, when given sufficient inference compute, is the primary driver of performance on multi-step clinical error correction, outweighing domain-specific medical training.

LoRA fine-tuning produces asymmetric within-model gains on the two benchmarks: the English correction average rises by +0.117 (19.6% relative) while the Japanese average rises by +0.081 (16.8% relative), despite Japanese training data being more than twice as large (5,538 vs. 2,439 samples). These two improvements are measured on different benchmarks and are not directly comparable in absolute terms, but the fact that the relative within-model gain is larger on the smaller English split is consistent with the base model's stronger English language modeling or with differences in task difficulty between the two languages. Fine-tuning can also degrade performance on underrepresented error categories (e.g., medication dosage sentence extraction drops from 62.5% to 12.5% on MEDRECT-ja), indicating a need for balanced training data across clinical domains.

Beyond the quantitative gains, the qualitative analysis (Section 5.4) shows that fine-tuning enhances clinically relevant capabilities not directly captured by reference-based metrics, including empathetic patient communication and appropriate restraint from overcorrecting already-accurate text.

## 7 Conclusion

We introduce MEDRECT, the first bilingual 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 11 contemporary LLMs including medical-domain models, we establish that reasoning capabilities are fundamental for medical error correction, with general-purpose reasoning models outperforming domain-specific alternatives on detection and sentence extraction subtasks. Bilingual 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 highlight considerations for the safe, equitable deployment of medical AI systems across languages.

## 8 Limitations

**Dataset Construction Without Clinician Validation** MEDRECT was constructed without explicit clinician-in-the-loop validation. To mitigate this, the first author (a medical student in clinical clerkship) reviewed sample quality throughout all construction stages and established objective filtering criteria (Appendix A.1); these criteria removed problematic samples such as internal contradictions in clinical values and ambiguous error labels (Appendix A.4), but the completeness of this filtering cannot be verified without expert review of all samples. The medical content itself is grounded in JMLE, which is developed by medical expert committees, and our pipeline performs format conversion rather than medical content generation. Systematic clinician validation remains a priority for future work.

**Dataset and Scope** The dataset size is constrained by the availability of suitable JMLE questions (287 of 800 across two examination years yielded 663 samples after synthesis and filtering), and our synthetic error generation may not fully cover the diversity of real clinical errors. MEDRECT also focuses on text-based scenarios with single-sentence errors; real-world errors may span multiple sentences, depend on document-level context (e.g., contradictory lab values across sections), or involve multimodal documents.

**Model-Specific Biases** The construction pipeline uses specific models at multiple steps. To mitigate this, models used in data synthesis (DeepSeek-R1-0528) and quality screening (Gemini 2.5 Pro) are excluded from evaluation, and the difficulty-based filtering in Step 2 uses multi-model consensus (11 configurations across 8 models) rather than single-model judgments. GPT-4.1 and Qwen3-32B variants are used in both consensus filtering and evaluation; while the multi-model design reduces single-model bias, their benchmark results should be interpreted with this consideration.

**Cross-Language Comparability** MEDRECT-ja and MEDRECT-en are constructed from different source materials (JMLE vs. MedQA-derived MEDEC) and have different error type distributions (Table 1). Absolute scores on the two benchmarks therefore reflect differences in source material and item difficulty as well as language, and should not be read as a direct measure of cross-lingual transfer. Throughout the paper we use ja / en scores for within-language rankings and for within-model comparisons (e.g., pre- vs. post-fine-tuning), and we avoid strong claims about the absolute magnitude of cross-language performance differences.

**Potential Risks** Performance varies substantially across error types, with particularly low accuracy in medication dosage errors (as low as 0% and 12.5% sentence extraction in Table 3), highlighting risks in high-stakes clinical domains. Most models also score lower on MEDRECT-ja than on MEDRECT-en (Table 2), which, even without claiming that the two scores are directly comparable, raises fairness concerns for non-English medical contexts. MEDRECT is designed for research evaluation, not direct clinical deployment; any real-world application should maintain human-in-the-loop validation of all automated predictions.

## Use of Generative AI

AI coding assistants (e.g., Claude Code) were used by the authors to support the implementation of data processing pipelines, evaluation scripts, and analysis tooling; all AI-generated code was reviewed and tested by the authors before inclusion.

## 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22539–22550, Vienna, Austria. Association for Computational Linguistics.

Junying Chen, Zhenyang Gui, Anningzhe Ma, Xidong Zhang, Ke Wang, Xiang Yan, Haiyang Ji, Jiatong

Qian, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

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.

Google DeepMind. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Junfeng Jiang, Jiahao Huang, and Akiko Aizawa. 2025. JMedBench: A benchmark for evaluating Japanese biomedical large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5918–5935, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

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.

Mingzhe Li, Peng Guo, Jiajun Ji, Yongxin Zhao, and Zhiyong Li. 2025. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574, Online. Association for Computational Linguistics.

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 Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

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 MEDRECT Dataset Construction

## A.1 Pipeline Details

This section provides comprehensive technical details for the data construction pipeline described in Section 3.2. Figure 3 shows the complete pipeline for both MEDRECT-ja and MEDRECT-en.

### A.1.1 Automated Synthesis

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)

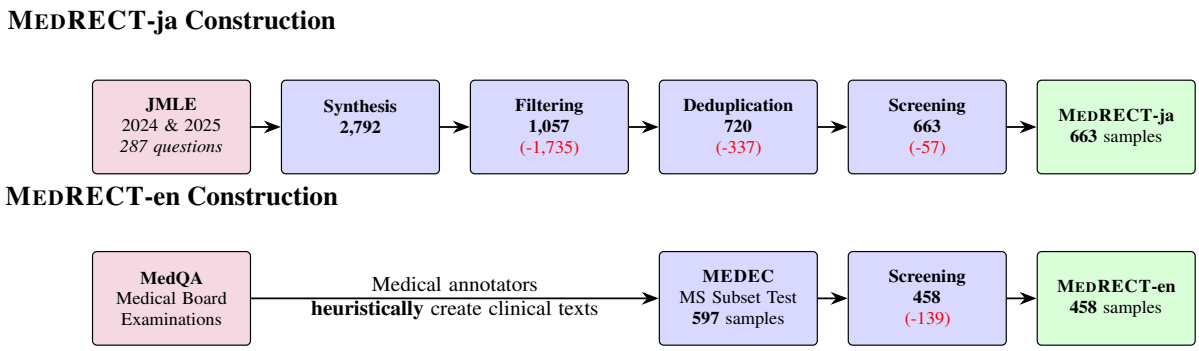


Figure 3: Data construction pipeline for MEDRECT benchmark creation. MEDRECT-ja (top) transforms JMLE questions through automated synthesis, quality filtering, model deduplication, and LLM screening to produce 663 high-quality samples. MEDRECT-en (bottom) applies identical LLM screening to the existing MEDEC MS Subset Test, yielding 458 samples. Red numbers indicate samples removed at each quality control step.

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 Consensus-Based Quality Filtering

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 $1/11 \leq \text{accuracy} \leq 7/11$.
- For ERROR samples, we applied a stricter standard, retaining samples only if their sentence extraction accuracy was within $1/11 \leq \text{accuracy} \leq 7/11$ *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 Model Deduplication

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 LLM-as-a-Judge Quality Screening

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 existing exam-derived vignettes 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).

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 |

For MEDRECT-en, the screening removed 139 of 597 samples (23.3%), primarily due to ambiguous errors (98 samples). This filtering targets objective data quality issues, not task difficulty. Two illustrative examples from the excluded samples: (1) A sample labeled CORRECT lists the patient's height as "166 cm (5 ft 40 in)." Since 1 ft = 12 in, "40 in" is impossible; the same sample also splits "blood pressure, 120/80 mm Hg" across three sentences, making sentence extraction evaluation unreliable. (2) A sample labeled CORRECT has lipid values where LDL is listed as 130 mg/dL, but the Friedewald equation (LDL = TC − HDL − TG/5) yields 208 from the other reported values, an internal contradiction. A model correctly flagging these issues would be penalized under the original ground truth. Results on the unfiltered MEDEC MS Subset (597 samples) are reported in Table 6 for direct comparison.

## B 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, i.e., 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 Data Composition

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 bilingual training coverage.

# C Additional Results

Table 6 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. To enable direct comparison with the physician baselines from the original MEDEC paper (Ben Abacha et al., 2025) and the MEDIQA-CORR 2024 shared task (Ben Abacha et al., 2024), both the BERTScore configuration and the correction-metric aggregation in this table follow the original protocol rather than the setup used in Table 2; correction scores are therefore not directly comparable across the two tables (see table note).

# D Computational Resources and Licensing

## D.1 Computational Resources

LoRA fine-tuning on Qwen3-32B and inference for the open-weight and medical-domain models in Section 4.1 were performed on single NVIDIA A100 80GB GPU nodes. Models accessed via the OpenRouter API (GPT-5, GPT-4.1, o3, Claude Sonnet 4, DeepSeek-V3-0324) were not run on local hardware. Parameter counts follow the original releases cited in Section 4.1.

## D.2 Dataset and Model Licensing

**Source materials** MEDRECT-ja is derived from the Japanese Medical Licensing Examinations (JMLE 2024–2025), published by the Ministry of Health, Labour and Welfare of Japan under the Public Data License 1.0 (PDL-1.0), which is compatible with CC BY 4.0. MEDRECT-en is derived from the MS Subset of MEDEC (Ben Abacha et al., 2025), released by its authors at `https://github.com/abachaa/MEDEC` under CC BY 4.0; we do not use the UW Subset, which requires a separate data usage agreement. LLM-as-a-Judge filtering applied to both subsets is documented in Appendix A.4.

**Released artifacts** The MEDRECT benchmark data will be released under CC BY 4.0 (compatible with both source licenses), the accompanying code under the MIT License, and the LoRA fine-tuned model under Apache 2.0 (inheriting from Qwen3-32B). For review, an anonymized version of the code and data is provided via the footnote URL in Section 1.

Table 6: 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 |
| 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 |

† To enable direct comparison with the physician baselines reported in the original MEDEC paper, this table follows the original MEDIQA-CORR 2024 protocol in two respects that differ from Table 2: (a) BERTScore uses microsoft/deberta-xlarge-mnli with English baseline rescaling rather than microsoft/mdeberta-v3-base without rescaling; and (b) the four correction metrics are averaged only over the (ground-truth = error, prediction = error) intersection, rather than over all ground-truth-error samples with misses scored as 0. ROUGE-1, BERT, BLEURT, and Avg. are therefore not directly comparable to Table 2.