

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

Naoto Iwase ^{*}
Nagoya University
Preferred Networks, Inc.
naoto@iwase.dev

Hiroki Okuyama
Preferred Networks, Inc.
hokuyama@preferred.jp

Junichiro Iwasawa 
Preferred Networks, Inc.
iwasawa@preferred.jp

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 for Japanese and English that formulates medical error handling as three subtasks: error detection, error sentence extraction, and error correction. MEDRECT-ja contains 663 samples derived from the Japanese Medical Licensing Examinations, while the separately sourced MEDRECT-en contains 458 samples curated from MEDEC. We evaluate 11 LLMs across 17 configurations that cover proprietary and open-weight models, medical-domain specialization, and multiple reasoning settings. Qwen3-32B scores higher in its thinking mode than in its non-thinking mode on error detection F1 and sentence extraction accuracy in both subsets, with sentence extraction accuracy higher by 24.5 percentage points on MEDRECT-ja and 10.3 on MEDRECT-en. Several leading general-purpose reasoning models outperform all three evaluated medical-domain models on these two subtasks. Most models have lower point estimates on the Japanese subset, although absolute scores are not directly comparable because the subsets differ in source material and error distributions. LoRA fine-tuning yields higher sentence extraction accuracy and higher point estimates on all three reference-based correction similarity metrics in both languages. MEDRECT provides an open, reusable evaluation resource for studying medical error correction and reasoning across Japanese and English.

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 MEDRECT¹ (A Bilingual **M**edical **R**easoning Benchmark for **E**rror **C**orrection in **C**linical **T**exts), to our knowledge the first benchmark for medical error detection and correction in Japanese, paired with English evaluation under a shared task definition (Figure 1). Our contributions are: **(1) Open Benchmark Construction:** A reproducible pipeline that creates error correction data from the Japanese

^{*}This work was done when N. Iwase worked at Preferred Networks, Inc. as a part-time engineer.

¹Dataset and code: <https://github.com/pfnet-research/medrect>

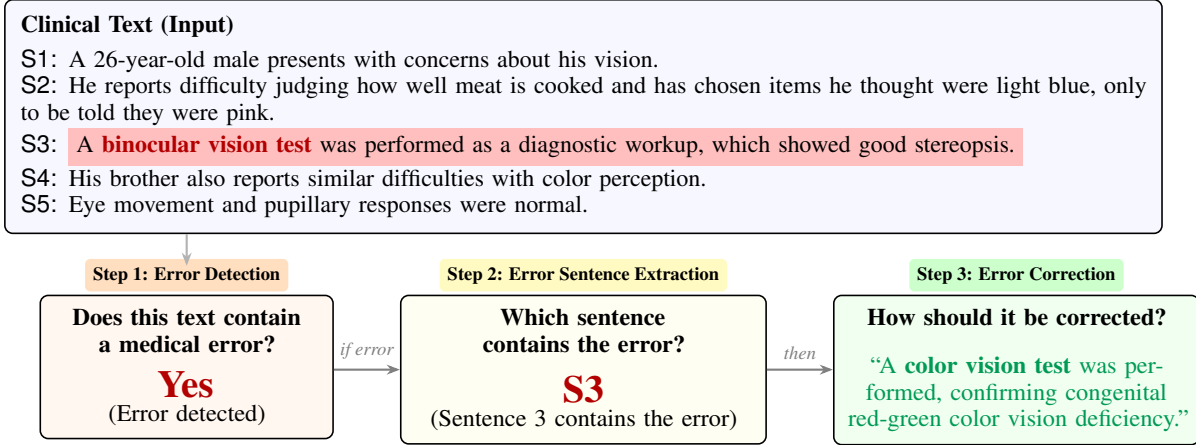


Figure 1: Overview of the MEDRECT benchmark, illustrated with a MEDRECT-ja sample (translated to English). Given a clinical text, models must (1) detect whether it contains a medical error, (2) identify the erroneous sentence, and (3) generate a corrected version. Here, the patient’s symptoms (difficulty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

Medical Licensing Examinations (JMLE), accompanied by benchmark data and code. **(2) Bilingual Evaluation:** We evaluate 11 LLMs across 17 configurations on MEDRECT-ja (663 samples) and MEDRECT-en (458 samples) under a shared task definition. MEDRECT-ja is derived from JMLE, whereas MEDRECT-en is drawn from the MedQA-based MEDEC MS subset of U.S. medical licensing exam questions. Neither subset is a translation of the other. **(3) Parameter-Efficient Fine-Tuning:** LoRA (Low-Rank Adaptation) fine-tuning (Hu et al., 2022) improves sentence extraction accuracy and yields higher point estimates on all three reference-based correction similarity metrics in both languages.

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). Beyond MCQA, clinical natural language inference performance varies across prompt structures, and LoRA effects differ by reasoning type (Jullien et al., 2026).

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. The closest multilingual benchmark is MEDERRBENCH (Ma et al., 2026), which contains 2,506 English, Chinese, and Arabic samples for the same three tasks, uses ten clinician-informed error types, and has all samples independently reviewed by two clinicians. MEDRECT complements these resources with Japanese coverage, a reusable automated construction pipeline based on JMLE, and analyses of reasoning settings and bilingual fine-tuning. Its English subset provides evaluation under the same task definition using quality-screened MEDEC data.

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). To our knowledge, MEDRECT is the first benchmark for medical error detection and correction in Japanese and pairs

this new coverage with English evaluation under a shared task definition.

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**, identifying the specific sentence with the error, and (3) **Error Correction**, generating a corrected version of the erroneous sentence. This decomposition enables fine-grained evaluation of model capabilities and helps identify weaknesses in the three-stage evaluation. The latter two subtasks apply only to clinical texts that contain an error. Figure 2 illustrates examples across different error types in both MEDRECT-ja and MEDRECT-en.

Example 1 (MEDRECT-ja):

Diagnosis

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*.

Reference correction: The laboratory findings strongly suggested meningitis caused by *GBS (Streptococcus agalactiae)*.

Example 2 (MEDRECT-ja):

Procedures/intervention

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.

Reference correction: The nutrition support team decided to prioritize consideration of non-invasive nutritional management methods.

Example 3 (MEDRECT-en):

Medication selection

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.

Reference correction: Itraconazole is prescribed after physical exam.

Figure 2: Examples from the MEDRECT dataset showing different error types. Examples 1 and 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), the key error in red, and the reference 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. MEDRECT-ja is constructed directly from JMLE rather than translated English data, grounding evaluation in Japanese-language vignettes and the institutional context of Japanese medical education and licensure. The two subsets share a task definition but are neither matched nor translated counterparts, so they support language-specific evaluation rather than direct estimates of language or cultural effects. Both subsets underwent automated screening but not systematic clinician adjudication.

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

During development of the Japanese pipeline, the first author, a medical student in clinical clerkship, manually inspected sample quality throughout the process, including the final retained set, and used these inspections to iteratively adjust parameters throughout the pipeline, including prompts, filtering thresholds, and screening criteria.

For MEDRECT-en, we applied the same LLM-as-a-Judge screening framework with source-adapted criteria to the MEDEC MS test subset (597 samples) (Ben Abacha et al., 2024, 2025), producing 458 samples.

The automated Japanese construction pipeline is designed for reuse with other exam sources. In this work, the English subset is instead curated from existing manually annotated MEDEC data. Complete procedures are detailed in Appendix A.1.

3.3 Dataset Statistics

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 two subsets have similar error-to-correct ratios of approximately 55:45.

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

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%)
<i>Error Type Distribution</i>		
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%)
Other	2 (0.5%)	2 (0.8%)

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.

For analysis, we use the eight error types specified in the Japanese synthesis prompt. Gemini 2.5 Pro (Comanici et al., 2025) reassigned the original MEDEC labels to this taxonomy. Two residual Japanese labels and two unmapped English outputs are reported as *Other*.

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 the MEDEC MS test subset (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 across 17 configurations². **Reasoning models** (GPT-5 (OpenAI, 2025a), o3 (OpenAI, 2025d), Claude Sonnet 4 (Anthropic, 2025), gpt-oss-120b and gpt-oss-20b (OpenAI, 2025b), and Qwen3-32B in thinking mode (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), and Qwen3-32B in non-thinking mode) do not. **Medical-**

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

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 two inference modes, thinking and non-thinking (labeled think and no-think in the tables), within the same architecture, enabling a within-model comparison of reasoning modes. 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, error sentence extraction accuracy, and three reference-based correction similarity metrics: ROUGE-1 (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020). We treat ERROR as the positive class for error detection F1. An always-ERROR predictor yields F1 scores of 0.713 and 0.693 on MEDRECT-ja and MEDRECT-en, respectively. We report the arithmetic average of the three correction metrics as a compact descriptive summary while retaining all three component scores. 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, and BLEURT uses BLEURT-20.

Error sentence extraction accuracy and reference-based correction similarity scores are computed on samples with a ground-truth error. The similarity scores are averaged over all such samples, with missed detections (the model outputs CORRECT for an error-containing sample) assigned 0. This avoids inflating scores for low-recall models that would otherwise be evaluated only on the errors they happened to flag. The Appendix MEDEC comparison (Table 7) instead follows the original MEDIQA-CORR 2024 evaluation protocol (Ben Abacha et al., 2024) for comparability with reported physician baselines.

For comparisons between Qwen3-32B in thinking and non-thinking modes and between Qwen3-32B in thinking mode with and without LoRA,

we estimate source-cluster uncertainty with 20,000 paired nonparametric bootstrap resamples (seed 20260715) and report 95% percentile confidence intervals (CIs). Predictions are aligned by sample ID. For MEDRECT-ja, samples are clustered by source JMLE question. For MEDRECT-en, each MEDEC Text ID is a separate source record. Error detection F1 is resampled over all source clusters, and error sentence extraction accuracy over source clusters containing at least one ground-truth error. These intervals quantify source-record variation for fixed predictions, not run-to-run generation variability.

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 directly translated the prompt (0_shot_ja), preserving the same task specification and output format.

4.4 Fine-tuning Configuration

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

Qwen3-32B was fine-tuned on combined Japanese (5,538 samples) and English (2,439 samples) training data with reasoning traces generated by DeepSeek-R1-0528 (see Appendix B for details).

5 Results

5.1 Performance across Reasoning Settings

Table 2 presents bilingual results for 17 configurations of 11 models on MEDRECT-ja and MEDRECT-en. All numbers in this subsection

refer to MEDRECT-ja unless otherwise noted. Claude Sonnet 4 achieves both the highest average reference-based correction similarity (0.620) and the highest error detection F1 (0.795). The next-highest averages are GPT-5 (0.577), gpt-oss-120b at high reasoning effort (0.572), and Qwen3-32B with LoRA in thinking mode (0.563). Error detection F1 ranges from 0.558 (MedResearcher-R1-32B) to 0.795 (Claude Sonnet 4), while sentence extraction accuracy ranges from 35.4% to 83.7%.

Two within-model comparisons show higher scores under higher-reasoning settings. First, average reference-based correction similarity falls as reasoning effort decreases for both gpt-oss model sizes. For gpt-oss-120b, it is 0.572 at high effort, 0.528 at medium effort, and 0.497 at low effort. Second, comparing Qwen3-32B in thinking and non-thinking modes holds model weights and architecture fixed, although the decoding settings also differ. On MEDRECT-ja, the thinking and non-thinking modes yield 0.723 vs. 0.637 error detection F1 and 72.5% vs. 48.0% sentence extraction accuracy. The paired differences are +0.086 (95% CI [0.051, 0.124]) and +24.5 percentage points (95% CI [17.5, 31.7]). On MEDRECT-en, the corresponding differences are +0.037 (95% CI [0.008, 0.066]) and +10.3 percentage points (95% CI [4.9, 15.6]) (Appendix Table 6).

Several leading general-purpose reasoning models score above all three evaluated medical-domain models on error detection and sentence extraction. 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 in thinking mode. MedGemma 27B (Google DeepMind, 2025) achieves 39.2%. These cross-family comparisons do not isolate reasoning from model size or training data. Together with the within-model results, they show that medical-domain specialization alone does not ensure stronger performance on this benchmark.

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 language subsets are not directly comparable. We therefore focus on within-language rankings and within-model differences. Most models score lower on MEDRECT-ja, but

Table 2: Bilingual performance for 17 configurations of 11 LLMs on MEDRECT-ja and MEDRECT-en. Parentheses indicate reasoning effort levels for gpt-oss (high, medium, or low) and reasoning modes for Qwen3-32B (think or no-think).

Model	MEDRECT-ja						MEDRECT-en					
	Error Detection	Sentence Extraction	Reference-based correction similarity			Avg.	Error Detection	Sentence Extraction	Reference-based correction similarity			Avg.
			F1	Acc. (%)	ROUGE-1				BERTScore	BLEURT	ROUGE-1	
<i>Reasoning models</i>												
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
<i>Non-reasoning models</i>												
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
<i>Medical-domain models</i>												
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

Table 3: Error sentence extraction accuracy (%) by error type on MEDRECT-ja and MEDRECT-en. For gpt-oss, only medium reasoning effort is shown. The four samples categorized as *Other* are excluded. Results based on fewer than 10 samples should be interpreted cautiously. Best values are bolded only for columns with at least 10 samples.

	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
	<i>#samples</i>	77	98	79	17	72	2	40	38	37	12	30	70	22	1	8
<i>Reasoning models</i>																
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
<i>Non-reasoning models</i>																
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
<i>Medical-domain models</i>																
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

the magnitude varies substantially. For sentence extraction, GPT-4.1 scores 52.6% on MEDRECT-ja and 72.8% on MEDRECT-en, while error detection F1 is similar for Claude Sonnet 4 (0.795 vs. 0.784). Within each language, the evaluated medical-domain models also trail the leading reasoning models on sentence extraction and average reference-based correction similarity.

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 the eight named error types in Table 3, excluding *Other*. For models with uneven per-type performance, macro-averaging reveals notable shifts: Qwen3-32B with LoRA in thinking mode 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 MEDRECT-ja than on MEDRECT-en (e.g., GPT-5: 83.3% vs. 97.8%, and Qwen3-32B with LoRA in thinking mode: 73.3% vs. 78.5%), indicating

that the direction of within-model differences is not merely an artifact of the error type distribution. Macro-averaged MEDRECT-en scores should be interpreted cautiously given small sample sizes in several error categories.

LoRA fine-tuning yields higher sentence extraction accuracy and higher point estimates on all three reference-based correction similarity metrics in both languages. On MEDRECT-ja, Qwen3-32B with LoRA in thinking mode³ achieves 0.563 average reference-based correction similarity compared with 0.482 without LoRA in the same mode (16.8% relative improvement). Sentence extraction accuracy increases by 9.0 percentage points (95% CI [2.8, 15.0]). The error detection F1 difference is +0.020 (95% CI [-0.011, 0.051]).

On MEDRECT-en, average reference-based correction similarity increases from 0.597 to 0.714 (19.6% relative improvement), and sentence extraction accuracy increases by 7.4 percentage points (95% CI [2.5, 12.3]) to 90.9%. The error detection F1 difference is -0.012 (95% CI [-0.040, 0.016]). Thus, the sentence extraction differences exclude zero in both languages, while the error detection differences do not.

Under the original MEDEC evaluation protocol, the fine-tuned model achieves 90.6% sentence extraction accuracy compared with 76.7% and 64.6% for two physicians, and 0.714 average reference-based correction similarity compared with 0.491 and 0.678 (Appendix Table 7). Its error detection accuracy is lower (62.0% vs. 81.3% and 68.9%) due to more false positives. These results should be interpreted cautiously given the two-person human baseline and the restricted task setting.

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* has the lowest average accuracy across models on MEDRECT-ja, with values ranging from 4.5% to 81.8%. This suggests that contextual understanding of patient interactions is difficult under the evaluated settings. *Physical findings* and *test interpre-*

tation also have relatively low scores. In contrast, most reasoning models achieve 80% to 100% on *medication selection*. *Medication dosage* varies from 0.0% to 100.0%, but this category contains only eight Japanese samples and three English samples.

Reasoning and Fine-tuning Comparisons

Comparing Qwen3-32B in thinking and non-thinking modes shows large gaps in *medication selection* (93.3% vs. 43.3%, +50.0 percentage points) and *procedures/intervention* (82.5% vs. 47.5%, +35.0 percentage points) under the evaluated settings. LoRA fine-tuning in thinking mode produces the largest gains in *history taking* (+22.7 percentage points) and *physical findings* (+25.0 percentage points). The observed drop in *medication dosage* (62.5% to 12.5%) is based on eight samples, so it cannot establish overfitting or a systematic fine-tuning effect. The gpt-oss model comparison also shows that parameter count alone does not order category-level performance in these settings.

5.4 Qualitative Analysis

Table 4 presents three cases and their model outputs, covering palliative care, communication with a distressed patient, and false positive correction. These examples describe individual outputs and are not a systematic clinical evaluation.

In Sample 1, most models identify gastrostomy placement as inappropriate for a terminally ill patient with limited prognosis. The fine-tuned model and proprietary systems propose non-invasive alternatives that prioritize comfort care.

In Sample 2, GPT-5 proposes direct assessment of suicidal ideation and safety, Claude Sonnet 4 invites further disclosure, and the LoRA fine-tuned model offers an empathic reflection. The JMLE-derived reference specifies the keyed best initial reply rather than a complete management plan. This case illustrates that reference-based similarity does not fully capture clinical adequacy.

In Sample 3, GPT-5, Claude Sonnet 4, and Qwen3-32B in thinking mode without LoRA incorrectly flag an accurate text and propose unnecessary additions. The LoRA fine-tuned model returns no correction. This example illustrates how a false positive could create unnecessary review work in a clinical workflow.

³Fine-tuned model: <https://huggingface.co/pfnet/Preferred-MedRECT-32B>

Table 4: Model outputs on three selected 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	<i>Procedures/intervention</i>	<i>History taking</i>	<i>(No error)</i>
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.
Reference Output	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 have been carrying this alone and must be very tired. "	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 Reference Output highlights the correction where applicable.

MEDRECT-ja samples (119B36_a_deepseek-r1-0528 and 118E37_c_qwen3-235b-a22b-thinking-2507) are translated to English for readability.

For Sample 2, the reference reproduces the JMLe-keyed best initial reply. It does not represent a complete suicide-risk assessment. Outputs are shown without clinical correctness ratings.

6 Discussion

On MEDRECT-ja, sentence extraction accuracy varies from 35.4% to 83.7% across settings. Higher-reasoning settings generally outperform lower-reasoning settings within the gpt-oss and Qwen3 families, although low-effort configurations can overlap the non-reasoning range. Leading general-purpose reasoning models also outperform the medical-domain models, indicating that domain specialization alone does not ensure strong performance.

For Qwen3-32B, sentence extraction accuracy is 24.5 percentage points higher in thinking mode on MEDRECT-ja and 10.3 higher on MEDRECT-en. LoRA raises average reference-based correction similarity by 0.081 (16.8% relative) and 0.117 (19.6% relative), respectively. Contrasting within-model patterns across MEDRECT-ja and

MEDRECT-en highlight the value of evaluation across language, source, and medical-education contexts. Because the subsets are not item-aligned, the patterns indicate context-specific behavior rather than isolated language or cultural effects. MEDRECT-ja provides a non-translated benchmark grounded in Japanese medical education and licensure.

The examples in Section 5.4 illustrate clinically relevant behaviors beyond reference-based correction similarity, but do not establish general clinical capability.

7 Conclusion

We introduce MEDRECT, an open bilingual benchmark for medical error detection and correction. Across 11 LLMs, higher-reasoning settings are associated with stronger error detection and sentence

extraction, while leading general-purpose reasoning models outperform the medical-domain alternatives. LoRA fine-tuning improves sentence extraction accuracy and all three reference-based correction similarity point estimates in both subsets.

8 Limitations

Dataset Validation MEDRECT was constructed without systematic clinician adjudication. During development, manual quality checks by the first author, a medical student in clinical clerkship, included the final retained set and informed iterative refinement of the Japanese pipeline (Section 3.2). Automated screening applied the resulting criteria (Appendix A.1) to remove problems such as internal contradictions in clinical values and ambiguous error labels (Appendix A.4). This medically informed oversight does not establish clinical correctness for every sample or replace independent clinician review. The pipeline uses constrained synthetic transformations grounded in exam materials developed by medical expert committees. Systematic clinician validation remains 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.

Reference-Based Evaluation The correction metrics measure similarity to a single reference correction. They do not directly assess clinical correctness, sufficiency, or safety. As Sample 2 in Table 4 illustrates, a clinically defensible response may add safety assessment beyond the keyed reference and receive lower similarity.

Model-Specific Biases The construction pipeline uses specific models at multiple steps. The two synthesis models (DeepSeek-R1-0528 and Qwen3-235B-A22B-Thinking-2507) and the quality-screening model (Gemini 2.5 Pro) are excluded from evaluation. The consensus-based difficulty filtering in Step 2 uses 11 configurations across 8 models rather than single-model judgments. GPT-4.1 and Qwen3-32B variants are used in both con-

sensus filtering and evaluation. Appendix C reports a conservative sensitivity check, but the unavailable item-level historical votes prevent exact reconstruction of a fully disjoint filtered benchmark. Results for reused models should therefore be interpreted with this limitation. Some released English training traces retain meta-references to provided answer information despite prompt instructions to avoid them, so these traces should not be interpreted as independently elicited rationales.

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 language subsets 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 MEDRECT-ja and MEDRECT-en scores for within-language rankings and 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 sentence extraction accuracy for medication dosage errors (as low as 0.0% and 12.5% in Table 3), highlighting risks in high-stakes clinical domains. Most models also score lower on MEDRECT-ja than on MEDRECT-en (Table 2). Without attributing this gap to language alone, the pattern motivates separate evaluation and development 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.

Acknowledgments

We are grateful to the Preferred Networks cluster team for the computational infrastructure support.

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, et al. 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, et al. 2025a. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, et al. 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.
- Mael Jullien, Andre Freitas, Marco Valentino, and Leonardo Ranaldi. 2026. Dissecting clinical reasoning in natural language inference for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 26229–26250, San Diego, California, United States. Association for Computational Linguistics.
- 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.
- Congbo Ma, Yichun Zhang, Yousef Al-Jazzazi, Ahamed Foisal, Laasya Sharma, Yousra Sadqi, Khaled Saleh, Jihad Mallat, and Farah E. Shamout. 2026. MedErrBench: A fine-grained multilingual benchmark for medical error detection and correction with clinical expert annotations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 11802–11827, San Diego, California, United States. 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 et al. 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, et al. 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, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Wen-wai Yim, Yajuan 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, et al. 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 (2024–2025), focusing on 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) and Qwen3-235B-A22B-Thinking-2507 (Yang et al., 2025).

The synthesis prompt requested one of eight error types: history taking, physical findings, test interpretation, diagnosis, monitoring/management, medication selection, medication dosage, and procedures/intervention. Each model was asked for five records per question (2,870 requested). Parsing the available complete and partial responses yielded 2,792 candidate samples.

A.1.2 Consensus-Based Difficulty Filtering

We evaluated each synthesized sample on error detection and error sentence extraction using eight validation models: Gemini 2.5 Pro (Comanici et al., 2025), GPT-4.1 (OpenAI, 2025c), PLaMo 2.0

Prime (Preferred Networks et al., 2025), Qwen3-8B, Qwen3-14B, Qwen3-32B, Qwen3-30B-A3B-Thinking-2507 (Yang et al., 2025), and QwQ-32B (Qwen Team, 2025; Yang et al., 2024). Qwen3-8B, Qwen3-14B, and Qwen3-32B were evaluated in both thinking and non-thinking modes, yielding 11 configurations.

We counted a complete response as correct when it returned CORRECT for a no-error sample or identified the gold error sentence for an ERROR sample. Across all samples, we retained cases with complete-response accuracy between 1/11 and 7/11 and a gap of at most 3/11 between error detection accuracy and complete-response accuracy. This excluded cases that were uniformly easy or difficult under the validation configurations and cases where models detected an error without reliably locating it.

This process filtered the dataset from 2,792 to 1,057 samples with intermediate difficulty under the validation configurations.

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. For each duplicate pair, we retained one sample and alternated the retained source model across pairs to balance representation. This reduced the dataset from 1,057 to 720 samples.

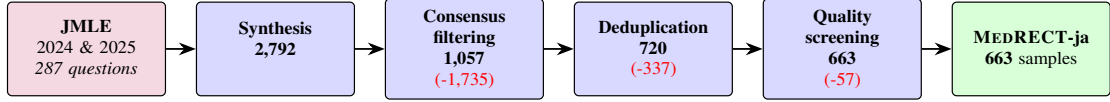
A.1.4 LLM-as-a-Judge Quality Screening

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

MEDRECT-ja Construction



MEDRECT-en Construction

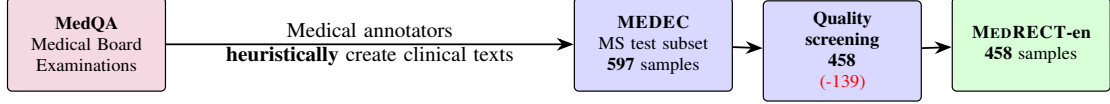


Figure 3: Data construction pipeline for MEDRECT. MEDRECT-ja (top) transforms JMLE questions through automated synthesis, consensus-based difficulty filtering, model deduplication, and LLM-as-a-Judge quality screening. MEDRECT-en (bottom) applies the same screening framework with source-adapted criteria to the existing MEDEC MS test subset. Red numbers indicate samples removed at each step.

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 screening produced 663 samples for MEDRECT-ja (92.1% retention) and 458 samples for MEDRECT-en (76.7% retention).

A.2 Data Synthesis Prompt

An abridged English translation of the synthesis prompt used to transform JMLE questions into clinical texts is shown below. The synthesis example is omitted for brevity.

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, medication
selection, medication dosage, procedures/intervention,
monitoring/management]
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]

(Synthesis example omitted for brevity.)
  
```

A.3 Quality Screening Prompt

An abridged English translation of the no-demonstration quality-screening prompt used with Gemini 2.5 Pro is shown below. We also used a four-demonstration variant with the same quality dimensions.

Quality Screening Prompt (No Demonstrations, Abridged 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
{sences}

### 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

We applied both the no-demonstration and four-demonstration prompting configurations to each sample. Any sample that scored 1 (problematic) on any quality dimension in either configuration was excluded from the final dataset, yielding different retention rates for the two language subsets (Table 5).

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 come 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,

Table 5: Quality screening results and exclusion reasons. Reasons are non-mutually exclusive, so their counts can exceed the number of excluded samples.

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

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 test subset (597 samples) are reported in Table 7 for direct comparison.

B Training Dataset Construction

B.1 Reasoning Synthesis

To construct fine-tuning data with explicit reasoning traces, we used DeepSeek-R1-0528 with 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.

The prompt instructed the synthesis model to reason from the clinical text without mentioning provided answer information. However, the released English traces retain residual meta-references to that information. We therefore treat these traces as synthetic supervision conditioned on the optional hints, not as independently elicited or faithful rationales.

B.2 Training Data Composition

We constructed bilingual fine-tuning data using DeepSeek-R1-0528 for reasoning synthesis. We retained only samples where the model produced correct responses and used 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 using JMLE questions from 2018–2023, covering diverse clinical domains beyond those used in the benchmark construction. Following the synthesis procedure 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 process yielded a final training dataset of **5,538 samples** with a resulting distribution of 34.8% CORRECT and 65.2% ERROR cases.

English Training Dataset For English training data, we used the MEDEC MS training and validation subsets, containing 2,763 expert-annotated clinical texts. These samples underwent reasoning synthesis using DeepSeek-R1-0528 with the same prompts used for the Japanese dataset. The resulting English training dataset comprised **2,439 samples**, with 49.0% CORRECT and 51.0% ERROR cases.

C Additional Results

Table 6 reports the full paired source-cluster bootstrap results for the within-model comparisons dis-

cussed in Sections 5.1 and 5.2. The Japanese benchmark contains 663 samples from 281 retained JMLE source questions, including 188 questions represented by multiple samples. The English benchmark contains 458 singleton source records.

Difficulty-Filter Sensitivity Three of the 11 validation configurations used for difficulty filtering (GPT-4.1 and Qwen3-32B in thinking and non-thinking modes) also appear in the final evaluation. The individual historical votes were not retained, so we cannot exactly refilter all 2,792 candidates after removing these configurations. We instead used the retained aggregate counts to identify a conservative core guaranteed to remain within proportionally scaled eight-configuration thresholds for any allocation of the three overlapping votes. Specifically, the retained count must contain four or five complete-response votes and a detection–complete-response gap of at most two. This core contains 121 samples (37 ERROR) from 102 JMLE questions. On this core, the difference between Qwen3-32B in thinking and non-thinking modes is +0.156 error detection F1 (95% source-cluster CI [0.004, 0.307]) and +27.0 percentage points in sentence extraction accuracy (95% CI [5.3, 48.7]). The difference between Qwen3-32B with and without LoRA in thinking mode is +0.065 error detection F1 (95% CI [-0.043, 0.181]) and +27.0 percentage points in sentence extraction accuracy (95% CI [5.1, 50.0]). The intervals are wide because the core contains only 37 error samples. This one-sided check cannot recover rejected candidates that might become eligible after refiltering, so it reduces but does not eliminate the selection concern.

Table 7 reports performance on the unfiltered MEDEC MS test subset (597 samples) before quality screening (see Appendix A.4). 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 reference-based correction similarity aggregation in this table follow the original protocol rather than the setup used in Table 2. Reference-based correction similarity scores are therefore not directly comparable across the two tables (see table note).

Table 6: Paired source-cluster bootstrap differences for focused within-model configuration changes. Each difference is the right-hand configuration minus the left-hand configuration. Brackets show 95% percentile confidence intervals from 20,000 resamples. Sentence extraction differences are percentage points.

Configuration change	Benchmark	Error detection F1 difference	Sentence extraction accuracy difference
Qwen3-32B: (no-think) → (think)	MEDRECT-ja	+0.086 [0.051, 0.124]	+24.5 [17.5, 31.7]
Qwen3-32B: (no-think) → (think)	MEDRECT-en	+0.037 [0.008, 0.066]	+10.3 [4.9, 15.6]
Qwen3-32B (think): without LoRA → with LoRA	MEDRECT-ja	+0.020 [-0.011, 0.051]	+9.0 [2.8, 15.0]
Qwen3-32B (think): without LoRA → with LoRA	MEDRECT-en	-0.012 [-0.040, 0.016]	+7.4 [2.5, 12.3]

Table 7: Performance on the unfiltered MEDEC MS test subset. Parentheses give gpt-oss reasoning effort or Qwen3-32B reasoning mode.

Model	Error Detection		Sentence Extraction	Reference-based correction similarity [†]			
	F1	Acc. (%)	Acc. (%)	ROUGE-1	BERTScore	BLEURT	Avg.
<i>MEDEC Paper Results</i>							
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
<i>Reasoning models</i>							
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	0.626	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	0.579	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
<i>Non-reasoning models</i>							
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	0.470	0.432
Qwen3-32B (no-think)	0.688	57.6	69.8	0.461	0.486	0.517	0.488

[†] For direct comparison with the published physician baselines, this table follows the original MEDIQA-CORR 2024 protocol: BERTScore uses microsoft/deberta-xlarge-mnli with English baseline rescaling, and the three reference-based correction similarity metrics are averaged over the (ground-truth = error, prediction = error) intersection. Table 2 instead uses microsoft/mdeberta-v3-base without rescaling and assigns 0 to missed detections over all samples with a ground-truth error. Values are not directly comparable across the two tables.

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 test 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 quality screening applied to both language subsets is documented in Appendix A.4.

Released artifacts The MEDRECT benchmark data are 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). The public code and data are available at the URL in Section 1, and the fine-tuned model is linked in Section 5.2.