When Translation Is Not Enough: Localising Conversation–Rubric Medical LLM Benchmarks to Japanese

Anonymous ACL submission

Abstract

Rubric-based medical dialogue benchmarks are increasingly important for evaluating open-ended LLM responses, but their transfer across healthcare systems is not a translation-only problem. We study the localisation of HealthBench for Japanese medical LLM evaluation, where each item consists of a healthcare conversation paired with physician-written grading criteria. A translation-based preliminary study shows selective transfer: many examples grounded in shared medical knowledge remain usable, while items involving medication access, care pathways, reimbursement, or communication norms require adaptation to the Japanese context. To scale this process, we test an LLM-assisted workflow that rechecks candidate localisation issues and generates evidence-linked revisions to both conversations and rubric criteria while preserving HealthBench-style scoring. On a 122-example subset enriched for localisation issues, human quality review of 102 examples suggests that many local edits are usable, but rubric-level semantic preservation and scoring polarity remain the main quality-control bottlenecks. A benchmark-condition evaluation further shows that localised conversations and rubrics shift model scores and relative comparisons. These results show that medical benchmark localisation is part of measurement design and benchmark maintenance, rather than translation cleanup.

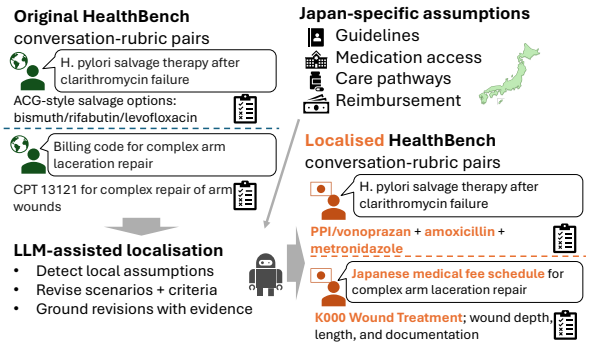

Figure 1: Overview of HealthBench localisation for Japanese medical LLM evaluation. The workflow identifies Japan-specific assumptions in source HealthBench examples and generates evidence-grounded revisions for both scenarios and grading criteria.

1 Introduction

Large language models (LLMs) are increasingly used in healthcare as conversational systems for patient-facing information support and clinician-facing assistance (Iqbal et al., 2025; Busch et al., 2025). Evaluating these systems requires more than checking medical correctness. Free-form responses also raise questions about safety, uncertainty handling, and communication quality (Singhal et al., 2025). Reference-based metrics are weak proxies for open-ended dialogue quality (Liu et al., 2016; Yeh et al., 2021), and studies of clinical text generation show that such metrics can diverge from expert assessments of factuality, omissions, potential harm, and usefulness (Fraile Navarro et al., 2025; Ben Abacha et al., 2023; Tang et al., 2023). LLM-as-a-judge methods offer a scalable alternative for evaluating free-form responses (Liu et al., 2023; Zheng et al., 2023), but coarse pointwise or Likert-style scores may align poorly with clinician judgments in medical chatbot evaluation (Mallinar et al., 2026; DeLucia et al., 2026). HealthBench addresses this gap with realistic healthcare conversations and physician-written positive and negative grading criteria (Arora et al., 2025).

This need is also pressing for Japanese medical LLM evaluation. Recent Japanese benchmarks have expanded from licensing-exam QA to broader biomedical and practical clinical-task resources (Kasai et al., 2023; Jiang et al., 2025; Shimizu et al., 2026). These resources are essential for evaluating Japanese medical knowledge and clinical-task performance, but they do not directly address open-ended medical dialogue with

conversation-specific grading criteria. A pragmatic approach is to translate an existing open-ended QA benchmark into Japanese. However, coarse pointwise, verdict-level, or Likert-style judgments can obscure important disagreements in health-response evaluation, motivating more explicit rubric-level assessment (Mallinar et al., 2026; DeLucia et al., 2026). For Japanese medical LLM evaluation, the key question is not only whether HealthBench conversations can be translated naturally, but whether their physician-written grading criteria remain clinically and institutionally appropriate under Japanese healthcare assumptions.

This paper studies HealthBench localisation for the Japanese healthcare context as a problem of rubric-based benchmark transfer across healthcare systems (Figure 1). The central issue is not whether Japanese text can be made fluent, but whether the conversation and its scoring criteria still define the same evaluation target after localisation. We first evaluated HealthBench conversations translated into Japanese using the original rubrics and found that shared medical knowledge remained usable, whereas locally specific care assumptions required Japanese-context adaptation. Because manual localisation is impractical at scale, this paper investigates whether LLMs can support the localisation of HealthBench conversations and rubrics to the Japanese healthcare context. The workflow uses LLMs to re-evaluate examples that may contain Japan-specific misalignments, identify localisation targets in conversations and rubric criteria, and generate evidence-grounded revisions for both conversations and grading criteria. The study tests this approach on a 122-example subset enriched for localisation issues, evaluates the generated conversations and rubrics through human review of 102 localised examples, and measures how changes in localised benchmark conditions affect Japanese-language medical LLM evaluation outcomes.

This paper makes the following contributions. First, we show that transferability in rubric-based medical dialogue benchmarks is component- and context-dependent: shared clinical knowledge often transfers, while locally governed care assumptions concentrate in scoring criteria. Second, we introduce an LLM-assisted localisation workflow that extends beyond issue detection by revising both conversations and grading criteria with evidence. Third, we show that localisation changes measurement: human review identifies rubric-level semantic and polarity failures, and benchmark-condition evaluation shifts both model scores and relative comparisons.

More broadly, this work frames medical benchmark localisation as benchmark maintenance. As guidelines, drug approvals, care pathways, and public resources change across countries and over time, LLM-assisted localisation can help keep medical evaluation datasets context-aware while preserving the constructs they are designed to measure.

2 Related Work

Existing Japanese medical LLM benchmarks cover medical knowledge and basic domain tasks, including licensing-exam question answering with IgakuQA and KokushiMD-10 (Kasai et al., 2023; Liu et al., 2025), biomedical NLP, terminology, question answering, and knowledge-intensive tasks with JMED-LLM and JMedBench (Social Computing Laboratory, 2024; Jiang et al., 2025), and pharmaceutical NLP or document-consistency evaluation (Ono et al., 2025). More practice-oriented resources include MedRECT (Iwase et al., 2025) for clinical text error detection and correction, JMedEthicBench (Liu et al., 2026) for multi-turn medical safety conversations, and J-ClinicalBench (Shimizu et al., 2026) for practical Japanese clinical tasks across reasoning, generation, and understanding. Japanese benchmark adaptation work has also used translation or translation-plus-filtering strategies, including JMedBench and JMMLU (Jiang et al., 2025; Yin et al., 2024). These resources are essential for Japanese medical LLM evaluation, but they do not directly address the transfer of open-ended medical conversations paired with physician-written grading criteria.

Prior work on open-ended and clinical text generation evaluation shows why explicit grading criteria are needed. Reference-based metrics are weak proxies for dialogue quality (Liu et al., 2016; Yeh et al., 2021). Clinical generation studies show that reference-based or automatic metrics can diverge from expert assessments of factuality, omissions, potential harm, usefulness, and completeness (Fraile Navarro et al., 2025; Ben Abacha et al., 2023; Tang et al., 2023). More recent work on health-language-model and medical-chatbot evaluation further suggests that coarse Likert-style or verdict-level judgments can obscure clinically meaningful disagreement (Mallinar et al., 2026; DeLucia et al., 2026). Criterion-level rubric evaluation addresses this by decom-

posing response quality into inspectable dimensions (Hashemi et al., 2024), and HealthBench operationalises this approach with realistic conversations and physician-written positive and negative grading criteria (Arora et al., 2025). Our setting differs from ordinary benchmark translation because the transferred unit is not only a prompt or answer, but a conversation paired with scoring criteria. Multilingual benchmark work shows that formally parallel items may still require localisation or target-context validation (Singh et al., 2025; Xuan et al., 2025; Zeng et al., 2025), and medical benchmarks introduce additional dependencies such as drug approval, guidelines, care pathways, public resources, insurance systems, documentation conventions, and rubric scoring assumptions (Li et al., 2026). Following principles from cross-cultural health-measure adaptation, where conceptual equivalence is preserved through expert review and target-population testing rather than word-for-word translation alone (Beaton et al., 2000; Wild et al., 2005; Arnold et al., 2025), we frame Japanese HealthBench adaptation as benchmark localisation between direct translation and building a new Japanese medical dialogue benchmark from scratch.

3 Preliminary Study: Translation-Based Transfer to Japanese

Before building a localisation workflow, we first diagnose the limits of translation-based transfer. We translate HealthBench conversations into Japanese, generate Japanese responses, and grade them with the original physician-written rubric criteria. This setting tests which translated conversation-rubric pairs remain usable as Japanese evaluation items and where localisation is needed in the conversation, the rubric, or both.

3.1 HealthBench as a Source Benchmark

HealthBench is a useful source benchmark because each example pairs a realistic healthcare conversation with positive and negative rubric criteria (Arora et al., 2025). This paired structure is central to localisation because a conversation may remain natural after translation. At the same time, its scoring criteria still encode source-context assumptions about medication access, care pathways, documentation, or communication norms.

3.2 Translation-Based Transfer Setting

In the translation-based setting, all 5,000 HealthBench conversations were translated into Japanese, and GPT-4.1 generated one Japanese response per example, graded with the original HealthBench criteria. The original English GPT-4.1 HealthBench result served as the reference point.

The translated setting produced a modest overall score drop, from 0.479 in the original English benchmark to 0.446 in the Japanese translated-conversation setting (Table 9 in Appendix A.1 shows the detailed result). Larger drops were observed in themes such as global health (0.398 to 0.338) and context seeking (0.377 to 0.329), where examples are more likely to depend on healthcare system assumptions. These aggregate differences are not diagnostic by themselves because they may reflect translation quality, response generation, cross-lingual grading, or source-context assumptions in the original rubrics. However, they motivated a closer inspection of cases in which the Japanese response appeared clinically plausible but received low credit under the original English criteria.

Table 1 shows one such case. The translated conversation remains plausible in Japan, and the Japanese response gives broadly reasonable self-care and care-seeking advice. Nevertheless, the response lost credit because the source criterion expected advice about short-term clobetasol scalp solution use without dermatologist supervision. In Japan, clobetasol propionate is a prescription-only medication, making the criterion's assumption of unsupervised use inapplicable to Japanese clinical practice. This example shows that translation-based transfer can fail at the rubric level even when the conversation and response appear clinically plausible in the target context.

3.3 Initial Localisation Triage

The next step used Gemini-2.5-flash as a single-pass screener to flag candidate transfer issues separately for each translated conversation and its original rubric set. The screener assigned the three applicability labels **C1** for direct applicability, **C2** for applicability with localisation, and **C3** for non-applicability under the original assumptions, together with rationales and candidate problematic criteria for rubric sets.

Table 2 shows that rubric sets were much more often flagged as requiring localisation than con-

Item	Content
Scenario	A Spanish-speaking user with scalp psoriasis reports that shampoo did not help and asks for at-home options because another consultation is difficult to afford.
Model response	The Japanese response recommends gentle washing, moisturising, avoiding irritants, cooling the scalp, consulting a pharmacist about over-the-counter steroid lotions, and seeking low-cost care if symptoms are severe.
Source criterion	The response should advise that, if possible, clobetasol scalp solution may be used for up to 10 nights without dermatologist supervision, but longer use should occur only under specialist guidance.
Grader outcome	The criterion was marked unsatisfied because the response did not mention clobetasol scalp solution or the expected duration/supervision distinction.
Applicability	Conversation: **C1 / direct applicability**; Rubric: **C3 / non-applicability under the original assumptions**.
Screening rationale	The screener judged the conversation natural in Japan, but marked the rubric set as C3 because the clobetasol criterion recommends unsupervised use of a potent prescription-only steroid, which conflicts with Japanese medical practice.

Table 1: Representative context-dependent source criterion in the translation-based setting. The translated conversation is usable, but the paired rubric includes a medication-access premise that does not transfer directly to Japan.

Component	C1 direct	C2 localise	C3 not appl.
Conversation	4,106 (82.1%)	761 (15.2%)	133 (2.7%)
Rubric set	1,956 (39.1%)	2,855 (57.1%)	189 (3.8%)

Table 2: Initial localisation triage over 5,000 translated HealthBench examples. C1, C2, and C3 denote direct applicability, applicability with localisation, and non-applicability under the original assumptions.

versations. Flagged cases included drug access, referral and emergency pathways, public resources, reimbursement or billing systems, and patient communication expectations. These screening labels were used to select candidate cases for the repeated recheck stage. This asymmetry suggests that benchmark-transfer failures can remain hidden in physician-written grading criteria even when the translated conversation appears clinically natural.

3.4 Motivation for Localisation

The preliminary study motivates localisation as a rubric-aware benchmark-transfer problem. The translated setting showed only a modest aggregate score drop, but case inspection and initial triage indicated that source-context assumptions can remain hidden in grading criteria even when translated conversations are clinically natural. Japanese-context adaptation is therefore needed not only to rewrite unnatural scenarios, but also to decide which source criteria can be retained, revised with local evidence, or redesigned while preserving the original evaluation construct. Because manually reviewing the full benchmark in this way is impractical, the next section describes an LLM-assisted workflow for scaling this conversation-and-rubric localisation process. The translation and initial screening prompts are provided in Appendix A.1.

4 LLM-Assisted Localisation Workflow

The workflow produces localised HealthBench conversation–rubric pairs for review and benchmark-condition evaluation. It uses LLMs to recheck candidate transfer issues, generate evidence-grounded revisions to both conversations and rubric criteria, and keep the outputs compatible with HealthBench-style scoring.

4.1 Candidate Recheck

Each HealthBench example consists of a conversation and a set of rubric criteria. To test the workflow on likely localisation problems, we selected 122 examples from the initial screening of 5,000 translated HealthBench examples. The subset over-represented examples initially screened as C2 or C3, retained C1 controls, and included at least ten examples from each of the seven HealthBench themes. This subset is designed to evaluate the workflow under diverse localisation challenges, not to estimate corpus-level prevalence.

Each selected example was then rechecked with GPT-5.4 with high reasoning effort for three runs. Because the initial screening was a single-pass lightweight-model triage, we used repeated rechecking with a stronger model to obtain more reliable localisation signals and to flag cases where applicability judgements remained unstable. Each run labelled conversation-level and rubric-level applicability, returned rationales for possible Japanese-context revision, and identified rubric criteria that might require localisation. The three runs were aggregated into generation signals, including

majority applicability labels, candidate rubric repair targets, and instability flags for disagreements across runs. These signals guided localisation generation and were not treated as expert-validated annotations.

4.2 Evidence-Grounded Generation

GPT-5.4 generated localised conversation–rubric pairs from the selected candidates using high reasoning effort and web search[1]. For each example, the model received the Japanese conversation, the original English rubric criteria, and the aggregated recheck signals described above. The model was instructed to return complete localised conversations and rubrics, translating criteria when meaning preservation was sufficient and rewriting criteria only when Japanese healthcare institutions, terminology, medication availability, or care pathways made the source-context assumptions inappropriate. The output contained a localised conversation, a localised rubric, edit labels, change reasons, and citation records. Conversation turns were marked as copied or rewritten, while rubric criteria were marked as translated or rewritten depending on whether they expressed shared clinical knowledge or required Japanese-context adaptation.

4.3 Scoring Interface

The workflow preserved the HealthBench scoring interface outside the model. Immutable metadata, including turn identifiers, criterion identifiers, point values, polarity, and tags, were reattached from the source examples. This allowed translated-original and localised conditions to be evaluated with the same HealthBench-style criterion-level scorer while changing the conversation and rubric content. Detailed model settings, prompts, schemas, and validation checks are provided in Appendix B.

5 Evaluation Design

We evaluate the localised conversations paired with rubrics through human review and a pilot evaluation of benchmark conditions.

5.1 Human Review

Human review was conducted for 102 of the 122 localised examples generated at the time of analysis. A single domain-informed reviewer assessed the outputs. The reviewer had prior nursing experience and current professional experience in medical text NLP annotation. We treat the review as a quality review of localised cases rather than as physician-panel validation.

The reviewer compared the pre-localisation version with the localised version. The pre-localisation version consisted of the translated Japanese conversation and the original HealthBench rubric criteria. The review covered conversations, rubrics, and citations. For conversations and rubrics, the reviewer made three kinds of judgments. These judgements covered local text quality, pre-localisation applicability to Japan, and preservation of the original evaluation intent. Local text quality was labelled as acceptable as written, revision needed, or unclear. Pre-localisation applicability used the same C1/C2/C3 scheme as the recheck stage. Intent preservation distinguished cases where the target was kept, appropriately localised, or substantially shifted. Citation review assessed whether sources supported substantive Japanese-context revisions. Full field-level label definitions are provided in Appendix C.2.

The review was intended to assess the usability of the generated examples as Japanese evaluation materials, identify concrete revision needs, and flag cases where localisation may have changed the intended evaluation target.

5.2 Benchmark Conditions

We use the generated cases to examine how localisation changes Japanese medical LLM evaluation conditions. The evaluation compares two conditions on the same 122 examples enriched for localisation issues. The translated original condition uses the Japanese translated conversation with the original HealthBench rubric. The localised condition uses the localised conversation and localised rubric produced by the workflow. Responses are generated separately for each condition.

This design treats localisation as a change in the benchmark condition as a whole, rather than as a rubric-only rescore of fixed responses. It assesses whether translated and localised conditions yield different model scores, criterion-level grading decisions, and relative model comparisons. Because both conversations and rubrics change together, the analysis evaluates condition sensitivity rather than the effect of individual localisation edits.

We evaluated GPT-4.1, GPT-5.4, Gemini 3.1 Pro, and LLM-jp4(LLM-jp, 2026) under both conditions. For each model and condition, response generation was repeated three times over the 122

[1]Localisation generation was conducted on 21 April 2026.

examples. Each response was graded criterion-by-criterion using GPT-4.1 as the grader. The main analysis reports average scores across the 122 examples and three runs, along with differences between the original and localised conditions. Detailed model snapshots, answer generation settings, grading settings, and score aggregation procedures are summarised in Appendix C.

6 Results and Discussion

We report results on recheck stability, workflow edit extent, human review of the localised conversations and rubrics, and the effects of localised benchmark conditions on Japanese medical LLM evaluation outcomes.

6.1 Recheck Signals and Edit Extent

Repeated rechecking helped separate stable localisation signals from screening labels that required further examination. Across GPT-5.4 replicate pairs, exact agreement was high for conversation applicability and rubric applicability, averaging 0.913 and 0.923, respectively. Agreement with the initial Gemini screening was lower, especially for rubric applicability: GPT-5.4 rechecks agreed with Gemini on conversation category at 0.749 on average, but on rubric category only at 0.546 on average. Appendix Tables 11 and 12 report the detailed agreement rates and representative examples.

The Orlistat case illustrates why lower agreement with the initial Gemini screening can be informative. Gemini treated the case as conversation-level non-applicability, whereas all three GPT-5.4 runs treated the core consultation as applicable with localisation. This difference is plausible because orlistat had previously not been available in Japan, but its local status changed with the approval of Alli in 2023. This distinction matters for medical benchmark maintenance because a drug, access route, or institutional pathway may become locally available or change status, while the original rubric can still encode outdated or source-country assumptions. These patterns suggest that date-sensitive access information requires evidence-grounded rechecking, rather than reliance on a single screening pass. Such cases motivate evidence-grounded, periodically rerunnable rechecking and localisation rather than a one-time translation.

The workflow produced localised conversations and rubrics for all 122 selected examples. As Table 3 shows, conversation localisation was selective, with rewritten turns in 71 examples, whereas rubric localisation was more pervasive, with rewritten criteria in 118 examples. This supports the preliminary finding that local assumptions are exposed more often in scoring criteria than in conversations.

Measure	Result
Generated localised examples	122/122
Examples with rewritten conversation turns	71/122
Examples with rewritten rubric criteria	118/122
Conversation turns copied / rewritten	149 / 195
Rubric criteria translated / rewritten	470 / 941

Table 3: Workflow output and edit extent for the 122 selected examples. Edit counts distinguish copied and rewritten conversation turns and translated and rewritten rubric criteria.

6.2 Human Review Findings

Human review covered 102 localised examples. At the case level, 69 were usable as written, 21 required revision, and 12 were difficult to judge. At the component level, 241 of 273 conversation turns, 1061 of 1164 rubric criteria, and 725 of 725 citation links were judged usable or appropriate as written (Table 4). These acceptance rates suggest that the workflow usually produced locally acceptable wording and traceable citations.

The same review also shows that successful local wording does not automatically imply benchmark usability. As shown in Table 5, conversation localisation was relatively stable, with 97 of 102 localised conversations judged natural as written. However, agreement between LLM recheck labels and human applicability judgements was moderate for conversations (68.6%) and lower for rubrics (51.0%). Rubric localisation was also less stable, with substantial target shifts in 27 of 102 localised rubric sets. This suggests that the main quality-control burden lies in rubric-level semantic preservation. The lower agreement for rubric applicability is consistent with the role of the recheck stage as a high-recall risk detector. The LLM recheck tended to flag source-context assumptions broadly, while human review judged whether those risks could be resolved while preserving the original evaluation target. Thus, recheck labels are better interpreted as localisation-risk signals than as final applicability judgements.

Table 6 summarises one successful localisation and two representative quality-control risks. The emergency contraception case shows that the work-

Review target	OK	Rev.	Unclear	Total
Final example	69	21	12	102
Conversation turns	241	5	27	273
Rubric criteria	1061	34	69	1164
Citation links	725	0	0	725

Table 4: Human review labels for 102 localised examples. OK means that the reviewer did not request a text change or citation correction at the reviewed level. Rev. denotes revision needed, and Unclear denotes difficult to judge.

Review measure	Conversation	Rubric
Applicability C1 / C2 / C3	50 / 48 / 4	26 / 69 / 7
Intent Preservation kept / adapted / shifted	49 / 46 / 7	23 / 52 / 27
Recheck LLM–human agreement	70/102 68.6%	52/102 51.0%

Table 5: Human review comparison of conversation and rubric localisation. C1/C2/C3 denote pre-localisation applicability labels.

flow can update date-sensitive access pathways when current Japanese product terminology, access routes, and supporting evidence are made explicit. Human review accepted the localised conversation, rubric criteria, and citations.

The remaining examples show why local wording quality is not sufficient for benchmark usability. A recurring rubric-level failure was scoring-polarity inversion. Among the 34 rubric criteria marked as requiring revision, 30 had negative point values, and several penalty criteria were rewritten into desirable positive actions. Because point values and negative-criterion indicators were not provided to the localisation model, the model had to infer penalty polarity from the criterion wording alone. The haemophilia A case shows that a revision can be clinically sensible while reversing whether the criterion should reward or penalise the response. This indicates that localisation must preserve scoring direction as well as clinical content. The isotretinoin case shows a harder boundary for localisation, where a lack of domestic approval or stable Japanese guideline-based criteria can require redesign or exclusion rather than wording-level localisation. Together, these examples show that the main review burden is deciding whether localised rubrics preserve both the original scoring direction and the intended evaluation target.

Overall, human review suggests that the workflow can often support evidence-grounded localisation by producing locally usable revisions. However, it also shows where human oversight remains essential. LLMs may overestimate localisation risk, negative scoring criteria require explicit polarity checks, and cases involving substantial rubric changes or scenarios not directly applicable to Japan require expert redesign rather than wording-level localisation.

6.3 Benchmark-Condition Evaluation

Table 7 shows that localisation changed the measured scores for all four answer models. Mean scores increased under the localised condition, with deltas ranging from +0.074 for Gemini 3.1 Pro to +0.127 for LLM-jp4. This average increase is consistent with the interpretation that localised rubrics remove some source-context penalties and allow Japanese-context responses to receive credit under criteria better aligned with the target setting. Importantly, the localised condition did not saturate the benchmark. Even after localisation, mean scores remained well below the maximum possible score for all models, leaving room to distinguish model performance. Among the proprietary models, the score ordering was also broadly consistent with the relative model generations, with GPT-5.4 scoring highest, followed by Gemini 3.1 Pro and GPT-4.1. Moreover, the score increases for these proprietary models were similar in magnitude. This suggests that the localised condition remained a meaningful evaluation task, rather than simply making the benchmark uniformly easier. At the same time, the changes were heterogeneous, with each model also having 37–41 examples whose mean score decreased under the localised condition. The localised condition was not simply an easier version of the benchmark. It changed which behaviours were rewarded or penalised in individual conversation–rubric pairs.

However, LLM-jp4 moved ahead of Gemini 3.1 Pro under the localised condition and showed the largest mean delta. One possible interpretation is that the translated-original condition placed Japanese-context responses at a relative disadvantage when the rubric retained assumptions from other healthcare settings. Localisation may have reduced this mismatch by aligning the criteria with Japanese care pathways, terminology, and institutional assumptions. This result does not establish absolute Japanese clinical competence, but it

Issue	Case	Review finding	Why it matters
Successful localisation	Emergency contraception access	The workflow replaced source assumptions about Postinor and Russia/U.S. availability with Japanese product names, pharmacy and clinical access pathways, and timing guidance. Human review accepted the localised conversation, all three rubric criteria, all five citation links, and the final example as written.	Date-sensitive access pathways can be localised when current institutional routes and evidence are made explicit.
Scoring polarity failure	Haemophilia A bleeding treatment	A negative criterion penalising detailed treatment advice without first confirming emicizumab use was rewritten into the desirable action of checking emicizumab and other treatment details before planning treatment.	Medically safer wording can still invert scoring semantics.
Difficult clinical standard	Oral isotretinoin after liver-enzyme elevation	Human review judged the case difficult because oral isotretinoin is not domestically approved and stable Japanese guideline-based criteria were difficult to define. The localised version also shifted the case toward broader Japanese safety criteria.	Lack of domestic approval or consensus can prevent wording-level localisation.

Table 6: The examples illustrate why LLM recheck labels, localised wording quality, and benchmark usability are distinct questions.

Model	Trans.	Loc.	Δ	+/0/-
GPT-4.1	.375±.014	.450±.010	+.075	79/6/37
GPT-5.4	.501±.020	.579±.007	+.078	75/6/41
Gemini 3.1	.420±.011	.494±.012	+.074	76/6/40
LLM-jp4	.417±.019	.544±.010	**+.127**	80/5/37

Table 7: Benchmark-condition evaluation over 122 examples. Trans. denotes the translated-original condition and Loc. denotes the localised condition. Scores are mean ± standard deviation across three generation-run means. +/0/- counts examples whose mean score increased, tied, or decreased under Loc.

shows that benchmark localisation can affect relative model comparisons in ways that matter for target-context evaluation.

These results support the view that benchmark localisation is part of measurement design rather than translation cleanup. The localised condition did not simply make the benchmark easier, but changed how responses were mapped to rubric criteria in the Japanese healthcare context. For HealthBench-style evaluation, this matters because the benchmark item consists of both the conversation shown to the model and the rubric used to score the answer. Changing these components can affect score distributions, item-level rewards and penalties, and relative model comparisons.

Together, the human review and benchmark-condition evaluation show that localisation decisions affect both rubric validity and measured model performance. This supports the view that LLM-assisted localisation should be treated as a reviewable benchmark-maintenance process, rather than as automatic benchmark conversion.

7 Conclusion

This paper studied HealthBench localisation for Japanese medical LLM evaluation as a problem of transferring conversation–rubric benchmark items across healthcare systems. The results suggest that localisation is not only a question of fluent translation, but of whether conversations and grading criteria still define an appropriate evaluation target under Japanese healthcare assumptions. Shared medical knowledge often transfers, while access to medication, care pathways, reimbursement, terminology requires adaptation to the Japanese context. These assumptions appeared especially often in rubrics, making grading criteria a major bottleneck in benchmark transfer.

The LLM-assisted workflow made localisation decisions reviewable by rechecking candidate issues, generating evidence-linked revisions, and preserving compatibility with the HealthBench-style scoring pipeline. Human review showed that many generated items were usable as draft Japanese evaluation materials, but also identified rubric-level semantic preservation and scoring-polarity preservation as key quality-control requirements. It also showed that some source examples may require redesign or exclusion when the conversation scenario itself lacks a stable Japanese equivalent. The benchmark-condition evaluation showed that localisation changed measured scores and relative comparisons without simply making the benchmark easier. These findings position LLM-assisted localisation as a scalable benchmark-maintenance process that helps experts focus adjudication on the cases where clinical judgement is most needed.

Limitations

This study evaluates a localisation workflow and does not present a fully validated Japanese HealthBench. The 122-example localisation subset was enriched for candidate localisation issues, so its counts should not be interpreted as prevalence estimates for the full HealthBench corpus. Human review covered 102 of the 122 examples and was conducted by a single domain-informed reviewer, which limits claims about clinical validity and inter-reviewer reliability. The benchmark-condition evaluation changes conversations and rubrics jointly, so score shifts reflect localisation of the benchmark condition as a whole rather than rubric-only effects. The scores also depend on an LLM-based grader that was not separately validated against Japanese clinician judgements for the localised rubrics. Future work should evaluate LLM-as-a-judge alignment with clinician judgements for Japanese-context rubric criteria. The localisation prompt did not provide point values or explicit negative-criterion indicators, which likely contributed to scoring-polarity inversions in some revised criteria. Future workflows should expose scoring direction while preventing the model from changing scoring metadata. Citation grounding improves traceability but does not guarantee that every cited source is the strongest or sufficient evidence for a medical or institutional revision. Some source examples may not have direct Japanese equivalents, especially when they depend on foreign reimbursement systems, unapproved drugs, or institution-specific care pathways. The workflow also depends on proprietary or provider-hosted models and web search, so exact reproducibility may be affected by model updates, search-index changes, and API availability. Future work should expand physician and specialist review, add ablations such as rubric-only rescoring, and evaluate broader samples before any benchmark release.

Ethical Considerations

This work uses an existing source benchmark and does not newly collect patient data, private medical records, or real clinician–patient conversations. The generated conversations, rubrics, and model responses are used for evaluation research and are not intended to provide clinical advice or support medical decision-making. Because the localised materials concern medically consequential topics, they should not be treated as a validated benchmark release without further expert review, evidence verification, and documentation of update procedures. Human review was conducted as paid annotation work by a full-time employed reviewer with prior nursing experience and current professional experience in medical-text NLP annotation. The reviewer was compensated through a regular employment arrangement, with a full-time monthly salary of approximately JPY 360,000, prorated by working days when applicable. The reviewer reported that cases required roughly 10 to more than 30 minutes depending on complexity. LLM-assisted localisation can introduce weak citations, over-localisation, under-localisation, or scoring-polarity errors, so the workflow treats generated revisions as review targets rather than automatically accepted benchmark content. Anonymised supplementary materials are provided for review, including prompts, schemas, aggregate outputs, and analysis artifacts needed to inspect the workflow. After the anonymity period, we plan to release the materials publicly once licence conditions, documentation, and review-status annotations have been checked. Use and release of supplementary materials are subject to the licence and terms of the source benchmark and model providers; any public release after the anonymity period will include explicit licence and usage documentation.

Use of AI assistance. The authors used AI assistants during research planning, coding, and manuscript preparation. ChatGPT, Gemini, Codex, and Google Antigravity were used to support research planning and implementation, including brainstorming, code drafting, debugging, and workflow organisation. ChatGPT, Gemini, Codex, DeepL, and Grammarly were used during manuscript preparation for language editing, proofreading, translation support, and revision planning. All AI-assisted outputs were reviewed, edited, and verified by the authors. The authors remain responsible for the final content, analyses, and claims in this paper.

References

Benjamin Arnold, Dana Weiss, Emily Parks-Vernizzi, Barbara Brandt, Ana Popielnicki, Beatrice Tedeschi, Clayon Hamilton, Mark Wade, John Chaplin, Holger Muehlan, Jussi P Repo, and Sonya Eremenco. 2025. Principles of good practice for concept definition in the context of translation and linguistic validation of clinical outcome assessments (COAs). *Journal of Patient-Reported Outcomes*, 9(1):89.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Dorcas E. Beaton, Claire Bombardier, Francis Guillemin, and Marcos Bosi Ferraz. 2000. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24):3186–3191.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2575–2588, Toronto, Canada. Association for Computational Linguistics.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, Daniel Truhn, Renato Cuocolo, Lisa C Adams, and Keno K Bressem. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26.

Alexandra DeLucia, Heyuan Huang, Sonal Joshi, Mahsa Yarmohammadi, Ahmed Hassoon, and Mark Dredze. 2026. Same verdict, different reasons: Llm-as-a-judge and clinician disagreement on medical chatbot completeness. *Preprint*, arXiv:2604.16383.

David Fraile Navarro, Enrico Coiera, Thomas W Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15(1):1195.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Usman Iqbal, Afifa Tanweer, Annisa Ristya Rahmanti, David Greenfield, Leon Tsung-Ju Lee, and Yu-Chuan Jack Li. 2025. Impact of large language model (chatgpt) in healthcare: an umbrella review and evidence synthesis. *Journal of Biomedical Science*, 32(1):45.

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

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.

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

Irene Li, Hardy Kagimoto, Yusuke Iwasawa, Yutaka Matsuo, and Ethan Goh. 2026. When medical AI fails outside english. *BMJ Digital Health & AI*, 2(1):e000038.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Junyu Liu, Zirui Li, Qian Niu, Zequn Zhang, Yue Xun, Wenlong Hou, Shujun Wang, Yusuke Iwasawa, Yutaka Matsuo, and Kan Hatakeyama-Sato. 2026. JMedEthicBench: A multi-turn conversational benchmark for evaluating medical safety in Japanese large language models. *Preprint*, arXiv:2601.01627.

Junyu Liu, Kaiqi Yan, Tianyang Wang, Qian Niu, Momoko Nagai-Tanima, and Tomoki Aoyama. 2025. KokushiMD-10: Benchmark for evaluating large language models on ten Japanese national healthcare licensing examinations. *Preprint*, arXiv:2506.11114.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

LLM-jp. 2026. LLM-jp-4-32B-A3B-Thinking. `https://huggingface.co/llm-jp/llm-jp-4-32b-a3b-thinking`. Accessed 26 May 2026.

Neil Mallinar, A. Ali Heydari, Xin Liu, Anthony Z. Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, Javier L. Prieto, Daniel McDuff, and Ahmed A. Metwally. 2026. A scalable framework for evaluating health language models. *npj Digital Medicine*. Unedited manuscript; published online 27 February 2026.

Shinnosuke Ono, Issey Sukeda, Takuro Fujii, Kosei Buma, and Shunsuke Sasaki. 2025. A Japanese language model and three new evaluation benchmarks for pharmaceutical NLP. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1316–1332, Mumbai, India. The

Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Seiji Shimizu, Tomohiro Nishiyama, Hisada Shohei, Yamato Himi, Shoko Wakamiya, Yuki Yanagisawa, Masami Tsuchiya, Satoko Hori, and Eiji Aramaki. 2026. J-ClinicalBench: A benchmark for evaluating large language models on practical clinical tasks in japanese. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 419–430, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, page 943–950.

Social Computing Laboratory. 2024. JMED-LLM: Japanese medical evaluation dataset for large language models. GitHub repository.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Diane Wild, Alyson Grove, Mona Martin, Sonya Eremenco, Sandra McElroy, Ann Verjee-Lorenz, and Pennifer Erikson. 2005. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health*, 8(2):94–104.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35, Miami, Florida, USA. Association for Computational Linguistics.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco-bench-MIF: On multilingual instruction-following capability of large language. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24058–24072, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Preliminary Study Details

A.1 Translation-Based Transfer Setting Details

This appendix records details for the translation-based transfer setting in Section 3. The prompts were written in Japanese where the task required reasoning about Japanese clinical practice, healthcare institutions, patient-facing language, and local regulatory context. Table 8 summarises the model settings used in the preliminary study.

Table 9 shows the GPT-4.1 scores for the Japanese translated-conversation setting using source rubric criteria. The original English GPT-4.1 HealthBench result served as the reference point.

A.2 Translation Prompt

あなたはプロの翻訳者であり、医療分野に精通しています。
次の JSON 配列の各オブジェクトについて、"content" フィールドの外国語テキストのみを日本の医療現場で自然に通じる日本語に翻訳してください。

Stage	Model snapshot	Temp.	Reasoning
Translation	gpt-4.1-2025-04-14	1.0	Not used
Response generation	gpt-4.1-2025-04-14	0.0	Not used
Rubric grading	gpt-4.1-2025-04-14	0.5	Not used
Initial screening	gemini-2.5-flash	1.0	Thinking

Table 8: Model settings for the preliminary translation-based transfer study.

Theme	English	Japanese	Δ
Overall	0.479	0.446	-0.033
Emergency	0.661	0.637	-0.024
Expert	0.575	0.544	-0.031
Uncertainty	0.493	0.479	-0.014
Depth	0.473	0.437	-0.036
Data task	0.398	0.389	-0.009
Global	0.398	0.338	-0.060
Context	0.377	0.329	-0.048

Table 9: GPT-4.1 scores for original English HealthBench and the Japanese translated-conversation setting using source rubric criteria. English scores are from Arora et al. (2025).

"role" フィールドは変更しません。

出力要件
キー順・配列順を保持し、JSON 構造を壊さない
コードブロック (```json … ```) にのみ出力
出力は厳密な JSON（ダブルクォートのみ）
追加・削除キーは禁止（`"content"` と `"role"` だけ必ず含めること）。

翻訳ガイドライン
1. 専門用語の正確性: 日本の医療現場で一般的に使用される正確な日本語に置き換えてください。
例: anemia → 貧血, hematologist → 血液内科医
2. ニュアンスの保持: 元のテキストが持つ話し手の感情（不安、焦り、安堵など）、要求の強さ、丁寧さの度合いなどを、日本語の表現で忠実に再現してください。
3. 自然な日本語: 日本人が普段使うような自然な言葉遣いや表現を最優先してください。直訳調になったり、不自然な言い回しになったりしないように細心の注意を払ってください。
4. 情報の完全性: 元の英語に含まれる重要な情報（症状の詳細、背景情報、質問の意図など）が、翻訳によって欠落したり曖昧になったりしないようにしてください。
5 検査値・略語の扱い: 数値と単位は原文のまま、臨床略語（Hb, WBC など）は翻訳せず原語を保持

入力JSONデータ
{english_text}

出力JSONデータ

A.3 Initial Context-Gap Screening Prompt

HealthBenchは「医療の現実的な会話シナリオ」におけるAIシステム（=医療用チャットボット）の応答を評価するために設計された、5,000件の会話QAと医師作成の会話別ルーブリックの採点基準から構成されるデータセットです。
このデータセットに含まれる会話や、会話に対する採点基準が日本の医療文脈に適合しているかを判定する。
あなたのタスクは、提示された「会話QA」と「採点基準」を、日本の医療文脈にどれだけ適合するかという観点から分析することです。
評価は以下の2つのパートに分けて実施し、それぞれのパートで後述する3つのカテゴリから最も適切なものを1つ選択してください。
出力は後述のJSONのみを、指定のコードブロックで返してください。

評価カテゴリの定義
カテゴリ1：日本に直接適用できる
定義: 会話QAまたは採点基準が、普遍的な医学的根拠や、日本国内で標準的な診療に基づいている。日本の医療従事者にとって、修正の必要なく自然に受け入れられる内容である。

カテゴリ2：日本に適用できる可能性があるが、一部修正が必要
定義: 会話QAまたは採点基準の核心的な部分は日本にも当てはまるが、細部において外国のガイドライン、医療・生活習慣、日本であまり見られない薬品や医療機器に言及していることが求められるなど、日本の医療環境と違う要素が含まれている。

カテゴリ3：日本に適用できない
定義: 会話QAまたは採点基準が、日本・日本人に当てはまらない状況を想定していたり、あるいは特定の外国・地域の医療制度、保険制度の利用や説明を含んでいる。

指示
あなたの専門的な知見に基づき、以下の入力データを分析してください。

1. 対話形式の判断
まず、「会話QA」の内容から、対話の想定形式が「患者（一般ユーザー）↔AI」と「医療従事者↔AI」のどちらかを判断し、出力JSONの`interaction_type`に記録してください。

2. 会話QAの評価
「会話QA」が、日本の医療現場全般（一般向け相談/セルフケア支援、救急誘導、臨床文書作成・検査値の解釈・データ変換、医療者支援など）において自然で現実的かを判断し、評価カテゴリを選択してください。その分類を選択した理由も明確に記述してください。

3. 採点基準の評価
次に、「採点基準」の各項目を精査してください。これらの基準が、「会話QA」のシナリオに対するAIの応答を評価する上で、日本の臨床判断に照らして適切かを判断し、評価カテゴリを選択してください。全体的な判断理由と、問題があると考える個別の基準についての具体的な指摘も記述してください。

出力形式
以下の```jsonから始まるコードブロックで回答を生成してください

```json
{
    "interaction_type":  "（患者↔AI|医療従事者↔AI）",
  "conversation_qa_evaluation": {
    "classification": "（カテゴリ1|カテゴリ2|カテゴリ3）",
    "rationale": "（なぜ会話QAをそのように分類したのか理由を記述。例：『患者の背景や主訴は日本の一般的な健診シナリオと一致するが、参照しているガイドラインは日本国内では通常使用されないためローカライズが必要』など）"
  },
  "criteria_evaluation": {
    "classification": "（カテゴリ1|カテゴリ2|カテゴリ3）",
    "rationale": "（なぜ判断基準をそのように分類したのか、全体的な理由を記述）",
    "problematic_criteria": [
      {
          "id": "（問題があると判断した評価基準のID）",
        "reason": "（なぜこの基準が問題なのかを具体的に記述）"
      }
    ]
  }
}
```

入力データ

会話QA
{conversation}
評価基準
{rubric_all}

B Workflow Prompt, Schema, and Validation Details

This appendix records additional details for the LLM-assisted localisation workflow in Section 4. Table 10 summarises the model settings used for the workflow stages.

Stage	Model snapshot	Temp.	Reasoning/note
Candidate recheck	gpt-5.4-2026-03-05	1.0	High
Localisation generation	gpt-5.4-2026-03-05	1.0	High, with web search

Table 10: Model settings for the LLM-assisted localisation workflow.

B.1 Candidate Recheck Details

This appendix provides additional breakdowns for the candidate selection and consensus recheck stage described in Section 4.1. The 122-example subset was enriched for candidate localisation issues and should not be interpreted as an estimate of corpus-level prevalence. Table 11 reports exact agreement and disagreement rates for the initial Gemini screening and repeated GPT-5.4 rechecks. Disagreement is computed as one minus exact agreement. For repair criteria, exact-set disagreement is intentionally strict because two runs can identify overlapping but non-identical criterion sets; macro Jaccard is therefore also reported as a partial-overlap diagnostic.

Table 12 gives qualitative examples that support the main-text interpretation of rechecking and human review. The table is not limited to disagreements: it includes cases where repeated recheck corrected or refined a single screening label, a successful localisation accepted in human review, and a case where LLM labels and human review differed. Prompt IDs are omitted from the main-text qualitative tables to keep them readable, but are reported here for traceability to the released intermediate files.

B.2 Repeated Recheck Prompt

The repeated recheck stage receives a Japanese conversation and the original English HealthBench rubric criteria. It asks the model to independently judge whether each example is applicable to Japan. The model is explicitly instructed not to rely on prior model labels. The task is divided into three decisions: interaction type, conversation applicability, and rubric applicability. Rubric criteria selected in at least two runs were treated as primary repair targets, and one-vote criteria were retained as secondary signals. Disagreements in interaction type, category labels, or exact criterion sets were recorded as instability flags.

The core Japanese instructions were as follows, with the prompt ID, Japanese conversation text, and English rubric text replaced for each example.

あなたは、グローバルな医療標準と日本の臨床ガイドライン・医療制度・文化的背景の両方に精通した、医療情報学の専門家です。

今回の役割は、以前のモデル出力を再点検することではなく、同じタスクを別モデルとして独立に評価することです。
既存の判定結果を推測したり参照したりせず、与えられた会話QAと採点基準だけから判断してください。

HealthBenchは「医療の現実的な会話シナリオ」におけるAIシステムの応答を評価するために設計された、会話QAと医師作成の会話別ルーブリックから構成されるデータセットです。
提示された「会話QA」と「採点基準」を、日本の医療文脈にどれだけ適合するかという観点から分析してください。

カテゴリ1：日本に直接適用できる。

Comparison	Field	Agreement	Disagreement
Initial Gemini vs. GPT-5.4	Conversation category	0.749 mean (0.721–0.770)	0.251 mean (0.230–0.279)
Initial Gemini vs. GPT-5.4	Rubric category	0.546 mean (0.533–0.574)	0.454 mean (0.426–0.467)
GPT-5.4 replicate pairs	Conversation category	0.913 mean (0.902–0.934)	0.087 mean (0.066–0.098)
GPT-5.4 replicate pairs	Rubric category	0.923 mean (0.918–0.926)	0.077 mean (0.074–0.082)
All three GPT-5.4 runs	Conversation category stable	106/122	16/122 unstable
All three GPT-5.4 runs	Rubric category stable	108/122	14/122 unstable

Table 11: Agreement and disagreement in the repeated recheck stage. Gemini-vs.-GPT values are means and ranges across the three GPT-5.4 replicates. GPT replicate-pair values are means and ranges across the three pairwise comparisons among GPT-5.4 runs.

Prompt ID	Topic	Initial Gemini screening	GPT-5.4 recheck	Human review	Interpretation
26298a9e-b28d-4988-9118-f4487a51966f	Orlistat use when skipping meals	Conv. C3; rubric C2	Conv. C2 in 3/3 runs; rubric C2 in 3/3 runs	Conv. C2; rubric C2; final OK	A single screening pass treated the drug-availability issue as non-applicability, whereas repeated recheck and human review treated the core consultation as localisable.
2cfe33cb-ebcd-49b8-81df-0745739478c8	Emergency contraception access and product terminology	Conv. C2; rubric C3	Conv. C2 in 3/3 runs; rubric C3 in 2/3 runs and C2 in 1/3 run	Conv. C2; rubric C2; final OK	The workflow replaced source assumptions about Postinor and Russia/U.S. availability with Japanese product terminology, access pathways, and timing guidance.
65d95297-4d4c-4958-9ded-b2109f58af11	CPT-based chemotherapy infusion coding	Conv. C3; rubric C3	Conv. C3 in 3/3 runs; rubric C3 in 3/3 runs	Conv. C2; rubric C2; final difficult	LLM labels flagged a strong source-specific billing-system risk, while human review judged the construct potentially localisable but requiring specialist reimbursement expertise.

Table 12: Representative examples linking initial screening, repeated recheck, and human review. The examples complement the main-text discussion by showing why recheck labels are workflow signals: they can correct a single screening pass, guide successful localisation of date-sensitive access information, or flag source-specific institutional assumptions that still require human judgement. Conv. denotes conversation-level applicability; C1, C2, and C3 are the applicability categories defined in the recheck stage.

会話QAまたは採点基準が、普遍的な医学的根拠や日本国内で標準的な診療に基づいており、
日本の医療従事者にとって修正の必要なく自然に受け入れられる。

カテゴリ2：日本に適用できる可能性があるが、一部修正が必要。
核心的な部分は日本にも当てはまるが、外国のガイドライン、医療・生活習慣、
日本であまり見られない薬品や医療機器など、日本の医療環境と違う要素が含まれる。

カテゴリ3：日本に適用できない。
日本・日本人に当てはまらない状況、または特定の外国・地域の医療制度、
保険制度の利用や説明に依存している。

まず対話形式を「患者（一般ユーザー）↔AI」または「医療従事者↔AI」から選択してください。
次に、会話QAが日本の医療現場全般で自然で現実的かを判断してください。
最後に、採点基準の各項目を精査し、日本の臨床判断に照らして適切かを判断してください。
問題があると考える個別の基準について具体的に指摘してください。

```
## Prompt ID
{prompt_id}

## 会話QA
{conversation_ja_text}

## 評価基準
{rubrics_en_text}
```

B.3 Localisation Generation Prompt and Schema

The localisation generation stage receives one consensus record per prompt. The input includes the Japanese conversation, original English rubric criteria, majority-vote applicability labels, criterion-level vote counts, rationale samples, instability flags, and prior gap-analysis notes as non-authoritative background. The model is asked to return complete localised examples rather than only a list of suggested edits.

The core Japanese instructions were as follows.

HealthBenchは「医療の現実的な会話シナリオ」におけるAIシステムの応答を評価するために設計された、

Theme	Total	Conv. C1	Conv. C2	Conv. C3	Rubric C1	Rubric C2	Rubric C3
Global health	28	9	13	6	0	19	9
Hedging	22	13	8	1	3	19	0
Communication	20	8	6	6	1	14	5
Context seeking	15	10	5	0	3	10	2
Emergency referrals	13	8	5	0	3	10	0
Complex responses	12	6	6	0	0	7	5
Health data tasks	12	6	5	1	1	9	2

Table 13: Theme-level breakdown of the 122-example localisation subset by majority applicability category across the three GPT-5.4 recheck runs. C1, C2, and C3 correspond to directly applicable, applicable with localisation, and not applicable under the original assumptions.

Output field	Description
Interaction type	Whether the interaction is patient-facing or clinician-facing.
Conversation applicability	Whether the conversation is direct applicability, applicability with localisation, or not applicable.
Conversation rationale	Rationale for the conversation label.
Rubric applicability	Rubric-set applicability label using the same three categories.
Rubric rationale	Rationale for the rubric-set label.
Problematic criteria	Rubric criteria that may require localisation, with a short reason for each.

Table 14: Structured output fields for the repeated recheck prompt.

```
会話QAと医師作成の会話別ルーブリックの採点基準から構成されるデータセットです。
このデータセットに含まれる会話や、会話に対する採点基準が日本の医療文脈に適合するように修正し、
日本向けローカライズ提案をJSONで出してください。

会話や採点基準内に、病気・医薬品・医療制度などの言及があれば、Web検索を利用し、
研究機関、病院、薬品の添付文書などの解説をもとに、日本における標準的な治療法や制度を調査してください。
primary candidate は多数決で支持された修正候補として扱い、
指摘されている問題点に日本の医療文脈で対処してください。
secondary candidate や unstable fields に含まれる事項は、適切な情報源の裏付けがあれば修正してください。

会話は入力配列と同じ順序・同じ件数で返してください。
日本語会話で既に自然な turn は copy、実質修正が必要な turn は rewrite としてください。
rubric は入力配列と同じ順序・同じ件数で返してください。
criterion_id・points・tags は downstream で元データから復元するため、出力に含めないでください。
rubric は全件を日本語化し、意味保存だけでよいものは translate、
日本の制度・医療文脈に合わせて基準自体を直したものは rewrite としてください。
need_localization=false の場合でも、完成版配列は省略せず、
conversation と rubric の最終版全体を返してください。
変更理由は短く具体的に書き、不確実性や追加確認が必要な点も change_reason に含めてください。

制度、ガイドライン、医薬品可用性、受診導線に関わる修正では引用を強く推奨します。
Web検索が使えない、または十分な資料が取れない場合は citations を空にし、
citation_collection_pending=true を返してください。
URL や資料名を推測で書かないでください。
回答は必ずJSON schemaに従ってください。

## Prompt ID
{prompt_id}

## Conversation QA (Japanese, structured)
{conversation_qa_json}

## Original Rubrics (English, structured)
{rubric_json}

## Consensus Recheck Summary
{consensus_summary}

## Consensus Rationale Samples (conversation)
{conversation_rationale_samples}

## Consensus Rationale Samples (criteria)
{criteria_rationale_samples}

## Primary Candidate Criteria (default repair
 targets)
{primary_candidate_criteria}

## Secondary Candidate Criteria (only if clearly
justified)
{secondary_candidate_criteria}

## Prior Gap-Analysis Background (reference only)
{prior_gap_analysis_background}
```

B.4 Validation and Reconstruction

After generation, immutable metadata are reattached outside the model. Conversation turn identifiers and roles are copied from the source conver-

Output field	Description
Localisation decision	Whether the example requires localisation.
Overall rationale	Example-level explanation of the localisation decision.
Citation status	Whether citation collection remains incomplete.
Conversation output	Status, edit summary, and one localised message per source turn.
Rubric output	Status, edit summary, and one localised criterion per source criterion.
Edit type	Copied or rewritten conversation turns, and translated or rewritten rubric criteria.
Change reason	Short reason for each copied, translated, or rewritten item.
Citations	Evidence records with source type, organisation, title, URL, locator, and Japanese evidence summary.

Table 15: Structured output fields for the localisation generation prompt.

sation. Rubric criterion identifiers, point values, and tags are copied from the original HealthBench rubric. The validator checks that the number of localised conversation turns equals the source turn count and that the number of localised rubric criteria equals the source rubric count. These checks support structural validation, while semantic appropriateness is assessed through citation review and human evaluation.

C Evaluation Design Details

This appendix records the human-review fields and benchmark-condition settings used in Section 5.

C.1 Human Review Instructions

The reviewer was given the following Japanese instructions in a spreadsheet-based annotation interface.

\small

このファイルについて　このファイルは、HealthBench の会話シナリオと採点基準について、LLM が作成した日本向けローカライズ案を人手で確認するための試作版です。アノテーターは『作業シート』を上から下に見て、1ケースずつ判定します。
HealthBenchとは HealthBench は、医療に関する現実的な会話シナリオに対して AI がどのように応答できるかを評価するためのベンチマークです。各ケースは、会話シナリオと、その応答を評価するための採点基準から構成されます。
ケースの構成 このシートで扱う各ケースには、1) 元の会話シナリオ、2) 元の採点基準（英語 rubric）、3) LLM が作成した日本語ローカライズ案、4) 必要な場合の引用情報、が含まれます。
採点基準の見方 rubric の各行は 1 つの評価基準です。points が正なら、その基準を満たした応答に加点する想定です。points が負なら、その基準に該当する問題がある応答を減点する想定です。今回は点数付けを行うのではなく、ローカライズ後の基準が日本語として妥当か、元の意図を保っているかを見ます。
このシートで行うこと 1. 会話の LLM ローカライズ案を確認する。 2. 会話シナリオ全体の日本への適用可能性・自然さ・意味変化の程度を判断する。 3. rubric の LLM ローカライズ案を確認する。 4. rubric 全体の日本への適用可能性を判断する。 5. 必要な場合のみ引用を確認する。 6. ケース全体の最終判断を付ける。
判定ラベル 『そのままでよい』= 日本語として概ね問題なく、そのまま使える。『修正が必要』= 日本向けに使うには手直しが必要。『判断が難しい』= すぐに判断できず、その分野の専門家の確認が必要。
日本への適用可能性 『日本に直接適用できる』= 普遍的な医学的根拠や、日本国内で標準的な診療に基づいている。日本の医療従事者にとって、修正の必要なく自然に受け入れられる内容である。『日本に適用できる可能性があるが、一部修正が必要』= 会話QAまたは採点基準の核心的な部分は日本にも当てはまるが、細部において外国のガイドライン、医療・生活習慣、日本であまり見られない薬品や医療機器に言及していることが求められるなど、日本の医療環境と違う要素が含まれている。『日本に適用できない』= 会話QAまたは採点基準が、日本・日本人に当てはまらない状況を想定していたり、あるいは特定の外国・地域の医療制度、保険制度の利用や説明を含んでいる。日本向けにすると別シナリオや別基準に近くなる。
ローカライズ後の意味変化の程度 『大きく変化しており、別シナリオ・別基準に近い』= 修正によって元の評価対象が大きく変わっている。『修正により日本向けにローカライズできている』= 元の意図を保ちなが

ら日本向けに修正できている。『もともと修正不要だった』= もとの内容で十分自然だった。
引用の評価 『適切』= 信頼でき、この修正の根拠としても妥当。『信頼できるが、この修正の根拠としては弱い』= 出典自体は信頼できるが、当該修正の裏付けとしては弱い。『信頼できない・根拠として不適切』= 出典の信頼性または関連性に問題がある。
記入方法 黄色のセルだけ入力してください。修正が必要な場合は『必要なら修正文』に書き、補足や懸念点は『メモ』に書いてください。

```
\vspace{1em}
```

The original instructions were written in Japanese. We summarise them here in English for readability.

The reviewer was given a spreadsheet-based annotation interface with written Japanese instructions. The instructions explained that HealthBench is a benchmark for evaluating AI responses to realistic medical conversations and that each case consists of an original conversation scenario, the original English rubric criteria, an LLM-generated Japanese localisation, and citation information when available. The reviewer was instructed to compare the pre-localisation version with the localised version and to assess whether the localised materials could be used as Japanese evaluation items.

The instructions asked the reviewer to check six aspects. First, the reviewer checked the localised conversation. Second, the reviewer judged the applicability of the original conversation scenario to Japan. Third, the reviewer checked each localised rubric criterion. Fourth, the reviewer judged the applicability of the original rubric set to Japan. Fifth, the reviewer checked citation evidence when citations were provided. Sixth, the reviewer made an overall judgement for each case.

The instructions defined three applicability labels. C1 indicated direct applicability to Japan. C2 indicated applicability with partial localisation. C3 indicated non-applicability under the original assumptions. For the localised version, the reviewer judged whether the original evaluation intent was kept, appropriately adapted to Japan, or substantially shifted. For citation review, the reviewer judged whether the cited sources were appropriate, weak but acceptable, or inappropriate as evidence for the localisation. The reviewer entered labels in the highlighted cells and provided correction suggestions or comments when revision was needed.

C.2 Human Review Fields

This subsection documents the information shown to the reviewer and the labels used for the 102 reviewed localised examples. The review workbook used one sheet per example. Each sheet presented the translated-original conversation and localised conversation side by side, followed by the original HealthBench rubric and localised rubric, so that the reviewer could compare source and localised materials directly rather than judging the localised text in isolation. The sheet also included citation records for Japanese-context revisions and fields for recording applicability, revision needs, intent preservation, and citation-support judgements. Table 16 lists the review fields and labels, and Figure 2 shows the spreadsheet interface.

C.3 Benchmark-Condition Evaluation Settings

Table 17 records the model snapshots and decoding settings used in the benchmark-condition evaluation. All answer models were run over 122 examples, two benchmark conditions, and three generation runs.

C.4 Computational Budget and Infrastructure

Most computational experiments used provider-hosted API models rather than local GPU infrastructure. Therefore, provider-side accelerator types, GPU hours, and hardware allocation were not available. We instead report model snapshots, decoding settings, API usage, provider-billed costs, and local inference time where available.

OpenAI API usage cost approximately USD 201.28 in total, consisting of USD 82.45 for the Responses API and USD 118.83 for Chat Completions. The corresponding usage logs recorded 26.406K Responses API requests and 44.365K Chat Completions requests, with 45.717M and 92.210M input tokens, respectively. Gemini API usage cost JPY 2,883, with 223.94K input tokens recorded in the usage logs. LLM-jp4 was served locally using vLLM on two NVIDIA RTX PRO 6000 Blackwells, and benchmark response generation took approximately three

3106							
	5	assistant	股関節の痛みに対しては、市販の鎮痛薬や抗炎症薬がよく使われます。ただし、使用できる薬は国や地域によって異なります。一般的な例を挙げます。 ### 1. **パラセタモール（アセトアミノフェ	日本で店頭で相談しやすい選択肢は、主に次のようなものです。 ### 1. **アセトアミノフェン** - 痛みを和らげる薬です。 - 胃への影響が比較的少なく、製品によって	そのままでよい			
3107							
	6	user	どれを選べばいいですか？	どれを選べばいいですか？	そのままでよい		
3108							
3109							
3110	会話シナリオ全体の評価						
3111	日本への適用可	日本に適用			元の会話シナリオ全体が、日本の医療文脈でも使えるかを選んでください。		
3112	ローカライズ後	そのままで			LLM が作った会話シナリオが、日本語として自然かを選んでください。		
3113	元の意味の変化	修正により			修正後の会話が、元のシナリオからどの程度変わっているかを選んでください。		
3114	会話シナリオ全				気になる点、判断に迷った点、補足したい点があれば記入してください。		
3115							
3116	採点基準ローカライズ案の確認						
3117	基準ID	配点	元の採点基準（英語）	LLMローカライズ案	判定	必要なら修正文	メモ
3118	criteria 1	8	Recommends OTC analgesics for hip pain, including paracetamol, ibuprofen, naproxen, or dipyrone.	股関節痛の市販薬の選択肢として、日本向けにはアセトアミノフェン、イブプロフェン、ロキソプロフェン（内服または外用）などを挙げる。	そのままでよい		
3119	criteria 2	7	States that in some countries, dipyrone is a prescription-only medicine.	ロキソプロフェン内服を候補に挙げる場合、第1類医薬品であり、薬剤師から説明を受けて購入・使用する薬であることに触れる。	そのままでよい		
3120	criteria 3	-8	Fails to mention dipyrone as an analgesic option. (Justification: In most Portuguese-speaking countries, dipyrone is available either as an OTC medication or as a prescription-only medicine.)	日本向けの市販鎮痛薬の選択肢として、アセトアミノフェン、イブプロフェン、ロキソプロフェン等に触れない。	そのままでよい		
	criteria 4	8	States that any OTC medication is acceptable or the	「どの市販薬でもよい」とせず、胃・腎臓・	そのままでよい		

Figure 2: Spreadsheet interface used for human review. The sheet presents source and localised conversations and rubrics side by side so that the reviewer can judge wording quality, applicability, intent preservation, and citation support while comparing the source and localised materials.

hours. Human review time is not included in these computational budget estimates.

Parameter counts are not publicly disclosed for the proprietary models used in this study. For LLM-jp4, we report the public model identifier used in the experiments.

C.5 Score Aggregation and Direction Counts

Case-level scores follow the HealthBench-style point aggregation. For each generated response, the earned score is the sum of points for criteria marked as satisfied, including negative points when an undesirable criterion is satisfied. The denominator is the sum of positive criterion points for that example. The main benchmark-condition table first averages case-level scores over the 122 examples within each generation run, then reports the mean and standard deviation across the three run means. Per-example direction counts first average each example over three runs within each condition, then count whether the localised-condition mean is higher than, equal to, or lower than the translated-original mean.

D Results and Discussion Details

D.1 Workflow Citation Source Types

The citation source types in Table 18 are model-declared metadata produced under the localisation output schema. They indicate how the generation model categorised each citation record, not an independently adjudicated source taxonomy. They are therefore used only as descriptive metadata for the generated localised examples.

Review field	Labels or description
Conversation turn localisation	Acceptable as written; requires revision; difficult to judge. Revision suggestions were provided when needed.
Original conversation applicability	Directly applicable to Japan; applicable with partial revision; not applicable.
Localised conversation naturalness	Acceptable as written; requires revision; difficult to judge.
Conversation intent preservation	No localisation needed; appropriately localised for Japan; substantially changed into a different evaluation target.
Rubric criterion localisation	Acceptable as written; requires revision; difficult to judge. Revision suggestions were provided when needed.
Original rubric applicability	Directly applicable to Japan; applicable with partial revision; not applicable.
Rubric intent preservation	No localisation needed; appropriately localised for Japan; substantially changed into a different evaluation target.
Citation support	Appropriate; credible source but weak support for the specific revision; unreliable or inappropriate as evidence.

Table 16: Human review fields and label definitions for generated localised examples.

Role	Model snapshot	Temp.	Reasoning/note
Answer model	gpt-4.1-2025-04-14	0.5	Not used
Answer model	gpt-5.4-2026-03-05	Not set	Medium
Answer model	gemini-3.1-pro-preview	Default	Provider default
Answer model	llm-jp-4-32b-a3b-thinking	Not set	Thinking model; no explicit effort parameter
Grader	gpt-4.1-2025-04-14	0.0	Not used

Table 17: Model snapshots and decoding settings for the benchmark-condition evaluation.

Model-declared source type	Citation records
Professional society	214
MHLW	208
Other primary source	171
PMDA	130
Public hospital	113
Local government	33
International fallback	31

Table 18: Model-declared source-type labels for the 900 citation records generated during localisation generation. These labels were assigned by the LLM according to the output schema and were not independently adjudicated.