From Guidelines to Guarantees: A Graph-Based Evaluation Harness for Domain-Specific Evaluation of LLMs

Jessica M. Lundin
Usman Nasir Nakakana
Guillaume Chabot-Couture
Gates Foundation

Abstract

Rigorous evaluation of domain-specific language models requires benchmarks that are comprehensive, contamination-resistant, and maintainable. Static, manually curated datasets do not satisfy these properties. We present a graph-based evaluation harness that transforms structured clinical guidelines into a queryable knowledge graph and dynamically instantiates evaluation queries via graph traversal. The framework provides three guarantees: (1) complete coverage of guideline relationships; (2) surface-form contamination resistance through combinatorial variation; and (3) validity inherited from expert-authored graph structure. Applied to the WHO IMCI guidelines, the harness generates clinically grounded multiple-choice questions spanning symptom recognition, treatment, severity classification, and follow-up care. Evaluation across five language models reveals systematic capability gaps. Models perform well on symptom recognition but show lower accuracy on treatment protocols and clinical management decisions. The framework supports continuous regeneration of evaluation data as guidelines evolve and generalizes to domains with structured decision logic. This provides a scalable foundation for evaluation infrastructure.

Data and Code Availability The WHO IMCI handbook is publicly available (WHO, 2014). Our graph construction, question generation code, and generated question dataset are available at `https://github.com/jessicalundin/graph_testing_harness`.

1 Introduction

1.1 The Evaluation Coverage Problem

Rigorous evaluation of language models faces a critical challenge: the distribution gap between application-specific text and existing benchmark datasets. This gap encompasses both context (domain, localization, complexity) and coverage (tasks, content). Current medical benchmarks rely on human curation, which is resource-intensive and results in incomplete coverage of specific medical guidelines.

MCQA benchmark datasets serve dual purposes: training new models and evaluating across models. The test split has widespread utility as a yardstick for comparison across models. While vignettes and multi-turn conversation with evaluation rubrics (Tu et al., 2024; Nori et al., 2025; Arora et al., 2025) more closely resemble real-world scenarios, MCQA remains an important evaluation format because it is less ambiguous, easy to grade, and scalable.

Despite advances in model architectures and training paradigms, MCQA benchmarks remain central for both evaluation and post-training. In health-domain models, supervised finetuning continues to be useful. Within alignment, MCQA also provides naturally ranked outputs for methods such as GRPO, where correct answers serve as high-reward samples and incorrect options serve as progressively lower-reward samples without requiring expensive human ranking.

WHO guidelines are an appropriate use case for this setting because there is substantial need for AI systems that support scarce healthcare workers in low- and middle-income countries (LMICs). These guidelines are often country-specific, which makes custom evaluation necessary for accurate measurement of model performance.

1.2 Limitations of Existing Medical Benchmarks

Medical benchmarks exist in multiple languages, and rely on questions from licensing exams, textbooks, journals, and crowdsourcing (Jin et al., 2021; Pal et al., 2022; Vilares and Gómez-Rodríguez, 2019; Labrak et al., 2022; Kasai et al., 2023; Jin et al., 2019; Zhang et al., 2017; Olatunji

et al., 2024; Hendrycks et al., 2021; Alonso et al., 2024). Synthetic medical QA datasets employ diverse generation strategies: template-based approaches as in emrQA (Pampari et al., 2018) and RadQA (Soni et al., 2022), generation using ontology concepts (Dong et al., 2023), and LLM-based generation for hallucination detection (Pal et al., 2023).

Existing MCQA benchmarks differ from our approach in three important ways. First, they rely on static question sets drawn from licensing exams, textbooks, and crowdsourcing, which are vulnerable to contamination as models are trained on increasingly broad corpora. Second, they provide aggregate scores that obscure performance on specific clinical relationships: a model may score well overall while systematically failing on treatment protocols or follow-up schedules. Third, they do not provide coverage guarantees relative to any specific guideline, making it impossible to know which relationships have and have not been tested. Non-MCQA evaluation formats such as patient vignettes (Tu et al., 2024) and multi-turn conversations with evaluation rubrics (Nori et al., 2025; Arora et al., 2025) more closely approximate real clinical reasoning but are expensive to construct, difficult to grade consistently, and cannot be regenerated as guidelines evolve. Graph-based MCQA occupies a complementary position: it provides the discrete gradability and scalability of MCQA with coverage guarantees and contamination resistance that static benchmarks lack, while serving as a structured precursor to higher-stakes evaluation in more realistic formats.

1.3 Contributions

Our main contributions are as follows:

1. We introduce a graph-based evaluation harness that provides explicit guarantees of coverage, contamination resistance, and validity.
2. We present a method for transforming structured clinical guidelines into a knowledge graph that supports systematic evaluation.
3. We demonstrate dynamic evaluation through on-demand query instantiation rather than static datasets.
4. We empirically show that this framework reveals systematic weaknesses in clinical reasoning that are not captured by aggregate benchmarks.

2 Method

2.1 Graph Construction from Clinical Guidelines

We transform the WHO IMCI handbook (WHO, 2014) into a directed graph structure. The handbook, an 80-page document containing flowcharts and checklists for childhood illness management, is parsed to extract medical entities and their relationships. The resulting graph contains 200+ nodes and 300+ edges spanning respiratory, gastrointestinal, nutritional, and infectious diseases.

The graph schema consists of five node types:

- **Condition** (31 nodes): Medical conditions with age range attributes (0–2 months for young infants, 2–60 months for children)
- **Symptom** (79 nodes): Observable clinical indicators (e.g., “fast breathing”, “convulsions”)
- **Treatment** (84 nodes): Medical interventions (e.g., “give oral Amoxicillin for 5 days”)
- **FollowUp** (15 nodes): Monitoring schedules (e.g., “3 days”, “7 days”)
- **Severity** (4 nodes): Triage classifications (severe, moderate, mild, none)

Four edge types connect these nodes:

- **INDICATES**: Symptom → Condition
- **TREAT**: Condition → Treatment
- **FOLLOW**: Condition → FollowUp
- **TRIAGE**: Condition → Severity

Automated extraction via PDF parsers and LLMs failed to reliably capture the conditional logic embedded in IMCI flowcharts. Relationships expressed visually through color-coded triage paths and nested decision branches cannot be faithfully reconstructed as directed edges by current PDF and LLM pipelines. The knowledge graph was therefore manually curated by a co-author with over 15 years of clinical practice, specialized pediatric training, and extensive experience implementing WHO IMCI guidelines in sub-Saharan Africa. Curation proceeded in three stages: (1) the clinical expert parsed each flowchart and checklist page to identify entity mentions and candidate relationships; (2) candidate edges were encoded in a structured schema and reviewed against the source document

for completeness; and (3) ambiguous cases, where visual triage paths implied conditional logic not expressible as a single directed edge, were resolved by the expert and annotated with explanatory notes. This clinical authorship of the graph establishes validity at the source: all generated questions inherit their accuracy from expert-constructed relationships rather than requiring post-hoc review of generated outputs.

2.2 Evaluation Query Instantiation

We employ graph traversal to automatically instantiate MCQA evaluation queries that ensure complete coverage of medical relationships. For each condition node, we traverse its connected nodes to instantiate the five question types shown in Table 1.

The framework dynamically instantiates evaluation queries using four templates for each of five question types while maintaining clinical relevance and variability. Random age generation is constrained to the condition's valid range (e.g., 0–8 weeks for young infants, 2–60 months for children).

The distractor sampling algorithm prioritizes clinical validity through age-stratified selection. For each question requiring $k = 3$ distractors, the system first identifies all conditions sharing the same age range as the target condition, creating an age-appropriate candidate pool.

For a question with correct answer v_{corr} of type τ and target condition with age range α, we construct an age-appropriate distractor pool by selecting candidate nodes that (i) match the required type and (ii) are compatible with the target age range. Distractors are then sampled uniformly without replacement from this pool.

This construction ensures that all distractors are clinically plausible within the relevant age group while maintaining variability across generated questions. A formal specification of the distractor construction is provided in Appendix A.

The dynamic generation process creates novel evaluation instances through variation in templates, ages, and distractors while maintaining consistent difficulty and clinical relevance. This mitigates a key limitation of static benchmarks, in which models may have seen evaluation questions during training, while enabling substantial variation for robust statistical analysis.

2.3 Contamination Resistance

The harness addresses two distinct contamination risks that static benchmarks cannot mitigate.

Surface-form contamination occurs when evaluation questions appear verbatim in training data. By generating questions at evaluation time with randomized ages, distractor sampling, and template selection, the probability of repeated surface forms is reduced relative to static benchmarks; the valid combinatorial space is bounded in practice by clinical constraints on age–condition–distractor compatibility, as discussed in Section 4.5.

Relationship-level contamination occurs when a model has learned the underlying clinical relationships from source documents, such that it can answer questions correctly regardless of surface form. Unlike surface-form contamination, this cannot be mitigated through variation in phrasing alone.

Rather than attempting to eliminate this form of contamination, the proposed harness enables a complementary evaluation strategy. Because evaluation queries are generated dynamically from a structured representation of the guidelines, the same framework can be applied to updated or modified guidelines that postdate model training. This allows evaluation to probe whether models have genuinely acquired generalizable clinical reasoning or are relying on memorized relationships from specific guideline versions.

In this sense, the harness supports temporal and versioned evaluation, making it possible to identify knowledge gaps as clinical guidelines evolve. This shifts evaluation from static benchmarking to continuously refreshable assessment aligned with evolving domain knowledge.

Graph-level errors represent a third risk, where inaccuracies in the knowledge graph propagate to all generated questions. Expert authorship of the graph (Section 3.1) directly addresses this by establishing the graph as a clinically verified source of evaluation truth.

3 Case Study: WHO IMCI

3.1 Clinical Expert Authorship and Validation

The knowledge graph underlying all generated questions was manually curated by a co-author who is a board-certified physician with over 15 years of clinical practice, specialized pediatric training, and extensive experience implementing WHO IMCI guidelines in clinical settings in sub-Saharan Africa. This authorship model, where domain expertise is

Table 1: Examples of auto-generated questions by relationship type.

Type	Example
Condition → Symptom	**Q:** A 2 year old child with Very Severe Disease would most likely present with which symptom? **Options:** A: convulsions, B: chest indrawing, C: pus draining from the eye, D: WFH/L 2 z-scores or more **Answer:** A
Symptom → Condition	**Q:** A 21 month old child presenting with convulsions is most likely to have: **Options:** A: Cough or Cold, B: Very Severe Disease, C: Severe Pneumonia or Very Severe Disease, D: Very Severe Febrile Disease with no Malaria Risk **Answer:** B
Condition → Treatment	**Q:** Which treatment is recommended for a 21 month old child with Very Severe Disease? **Options:** A: assess or refer for TB assessment and INH preventive therapy, B: if mouth ulcers treat with gentian violet, C: do virological test at age 4–6 weeks or repeat 6 weeks after the child stops breastfeeding, D: give first dose of intramuscular antibiotics **Answer:** D
Condition → FollowUp	**Q:** What is the appropriate follow-up schedule for a 3 year old child with Some Dehydration? **Options:** A: follow-up in 14 days, B: follow-up in 5 days, C: follow-up in 2 days if not improving, D: follow-up in 7 days **Answer:** C
Condition → Severity	**Q:** A 13 month old child with Very Severe Disease should be classified as: **Options:** A: moderate, B: mild, C: none, D: severe **Answer:** D

embedded at the graph construction stage rather than applied as post-hoc review, provides stronger validity guarantees than question-level annotation alone: every generated question inherits its clinical accuracy from expert-constructed graph relationships.

To further validate the generated question set, the same expert reviewed the 432 auto-generated questions across the five relationship types: Condition → Treatment (130), Symptom → Condition (118), Condition → Symptom (118), Condition → Severity (37), and Condition → FollowUp (29). For each question, the review assessed: (1) clinical accuracy of the correct answer, (2) appropriateness of distractors for the specified age range, and (3) clarity and unambiguity of question phrasing. Given that questions are derived from an expert-curated graph, this review serves primarily to verify that the generation pipeline correctly traverses and formats the underlying relationships rather than to establish clinical accuracy de novo.

The graph was curated by a single clinical expert, which precludes inter-rater reliability assessment. The underlying guidelines provide deterministic decision rules, which partially mitigates subjectivity in annotation. Independent validation by additional clinicians with IMCI expertise remains important future work for establishing the rigor required of a production evaluation instrument.

3.2 LLM Inference Results

We conduct baseline inference evaluation to assess out-of-the-box model performance for the closed-source models Claude Sonnet 4.6, o4-mini, and GPT-5.2, the open-weights model GPT-OSS-20B, and the domain fine-tuned model MedGemma-4B. Models are compared across size and training regime, including closed-source frontier, open-weights, and domain fine-tuned, to characterize the performance landscape broadly; within-class comparisons are deferred to future work. Models receive questions in a standardized format with explicit instructions to respond with only the letter (A, B, C, or D) corresponding to the correct answer. We measure accuracy per question type with uncertainty over the template variations.

Figure 1 and Table 2 present model performance across question types.

Figure 2 presents model performance variations across clinical question types, measured as the delta between question-specific accuracy and overall model accuracy.

3.3 Key Findings

1. The three frontier closed-source models, Claude Sonnet 4.6, GPT-5.2, and o4-mini, achieve statistically indistinguishable overall accuracy (71–72%), as their confidence intervals overlap substantially. The smaller models GPT-OSS-20B (61%) and MedGemma-4B

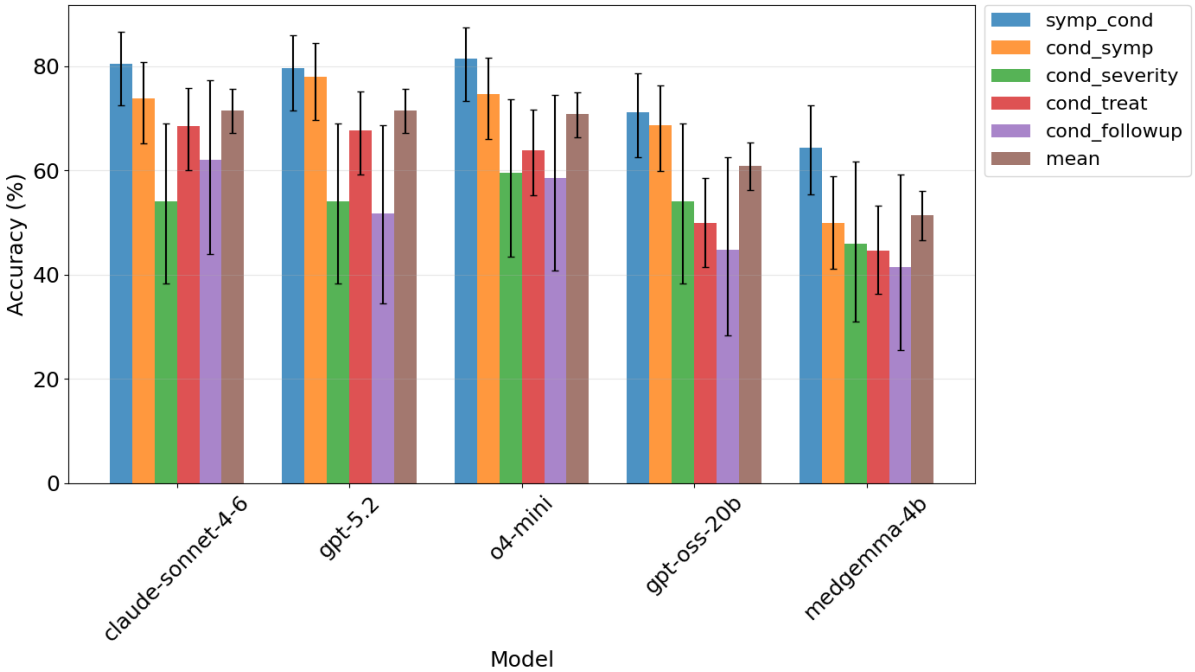

Figure 1: Model accuracy across five clinical question categories: condition-symptom (C→S), symptom-condition (S→C), condition-treatment (C→T), condition-severity (C→Sv), and condition-followup (C→F), along with overall mean accuracy across all categories. Error bars represent 95% Wilson score confidence intervals computed at the question level, treating each question as an independent Bernoulli trial.

Table 2: Model accuracy (%) on the IMCI knowledge graph evaluation across five clinical question categories: condition-symptom (C→S), symptom-condition (S→C), condition-treatment (C→T), condition-severity (C→Sv), and condition-followup (C→F). Values are reported as accuracy ± the half-width of the 95% Wilson score confidence interval, computed at the question level. Overall accuracy is pooled across all questions (question-weighted). Bold indicates the highest accuracy in each column.

Model	Overall	C→S	S→C	C→T	C→Sv	C→F
Claude Sonnet 4.6	**72.0±4.2**	73.7±7.8	80.5±7.1	**68.5±7.9**	54.0±15.3	**62.1±16.6**
GPT-5.2	**72.0±4.2**	**78.0±7.4**	79.7±7.2	67.7±7.9	54.0±15.3	51.7±17.1
o4-mini	71.0±4.3	74.6±7.8	**81.4±7.0**	63.9±8.2	**59.5±15.1**	58.6±16.9
GPT-OSS-20B	61.0±4.6	68.6±8.3	71.2±8.1	50.0±8.5	54.0±15.3	44.8±17.0
MedGemma-4B	51.0±4.7	50.0±8.9	64.4±8.5	44.6±8.4	46.0±15.3	41.4±16.9

(51%) perform well above random (25%).

2. Symptom → Condition questions show the highest performance across all models (64–81%), indicating that models better recognize symptoms than prescribe treatments or protocols.

3. Within-model performance varies substantially across question types, underscoring that aggregate accuracy obscures meaningful capability differences.

4. MedGemma-4B has lower performance than larger models across all question types, indicating that model scale and general reasoning capacity may dominate performance in this setting.

Unlike human-curated benchmarks, our dynamic graph-based method ensures complete coverage of all guideline relationships, consistent terminology from source documents, reduced data contamination through automated generation, and scalability to other medical guidelines.

3.4 Template Ablation Study

Figure 3 reveals substantial within-type variance across question templates, demonstrating that phrasing significantly affects model performance independently of the underlying clinical relationship being tested. The `cond_followup_t1` tem-

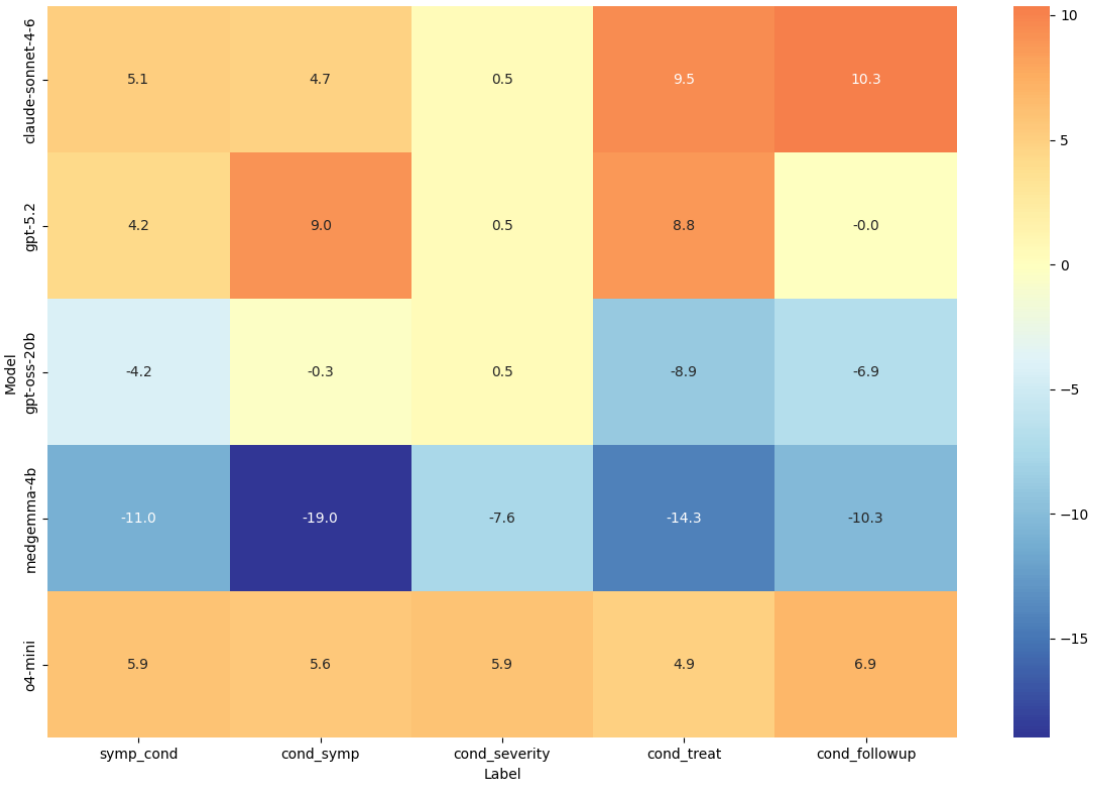

Figure 2: Accuracy delta heatmap showing the difference between question-type-specific accuracy and overall model accuracy for each model. Positive values (red/orange) indicate above-average performance for that question type, while negative values (blue) indicate below-average performance. Values are expressed as percentage points.

plate ("When should a {age} old child with {cond} return for follow-up?") consistently produces the lowest accuracy across all models (14–57%), while `cond_symp_t3` produces some of the highest (50–90%). This variance has direct implications for evaluation harness design: using multiple templates per question type, as our harness does, provides more robust estimates of model capability than single-template approaches, and averaging over template variants reduces the influence of phrasing artifacts on reported accuracy.

4 Evaluation Considerations

4.1 Operationalization

A key question for operationalization is how performance on this benchmark translates to real-world deployment. We argue that grounding evaluation in WHO guidelines provides a meaningful bridge: because the guidelines represent human-reviewed, authoritative clinical decision logic, high performance on graph-derived questions indicates alignment with expert-validated protocols. This supports a unit and integration testing analogy: unit tests verify that a model correctly handles individual clinical relationships (e.g., symptom → condition), while integration tests verify coherent reasoning across chains of relationships (e.g., symptom → condition → treatment → follow-up). While MCQA cannot capture the full complexity of patient vignettes or multi-turn clinical conversations, its discrete, unambiguous structure makes it well-suited for unit testing: each question has a single correct answer that requires no rubric to grade. The dynamic nature of the harness further strengthens this analogy, because questions are instantiated at evaluation time from the graph rather than drawn from a fixed set, models cannot memorize the test suite, preserving the integrity of repeated evaluation as guidelines and models evolve. In practice, this enables two concrete deployment decisions: models that fall below acceptable performance thresholds on clinically critical question types can be replaced by better-performing alternatives, and if a frontier model's guardrails change, a known risk in health domains where medically valid questions can trigger content filters, the harness provides a reproducible basis for selecting a replacement model with documented clinical pro-

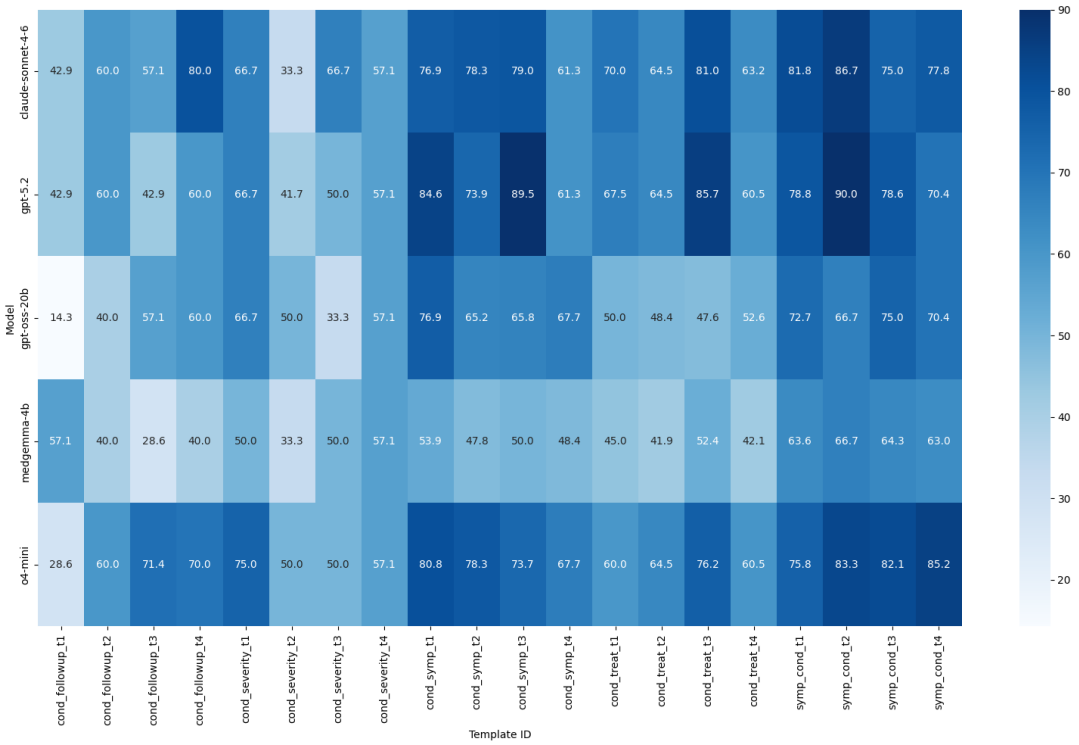

Figure 3: Accuracy by template and model. Each cell shows the accuracy (%) for a given model and question template. Darker blue indicates higher accuracy. Substantial within-type variance across templates demonstrates that question phrasing affects model performance independently of the underlying clinical relationship.

tocol alignment.

4.2 Cost and Scalability Relative to Manual Curation

Manual benchmark curation requires domain experts to author, review, and validate each question individually, a process that does not scale and produces static artifacts vulnerable to contamination. Our harness shifts the labor from question authorship to graph construction: a one-time cost that yields a large and refreshable space of evaluation instances for practical evaluation. For IMCI, manual curation of the knowledge graph by a domain clinical expert required significant upfront investment, after which 432 questions were generated automatically with validity inherited from the graph structure. Expanding across the combinatorial space induced by templates, ages, and distractors requires no additional expert labor beyond graph maintenance as guidelines are updated.

The primary scaling bottleneck is graph construction itself. Automated extraction via PDF parsers and LLMs missed critical relationships because the conditional logic in IMCI flowcharts is expressed visually through color-coded triage paths and nested decision branches that current pipelines cannot faithfully reconstruct as directed edges. Future work could reduce this bottleneck through semi-automated graph construction with expert review, particularly for guidelines with consistent structure such as WHO protocols.

4.3 Stakeholder Roles

The harness separates evaluation into three distinct stakeholder roles with different expertise requirements. *Graph constructors* require deep domain expertise to accurately encode guideline relationships; in our case, a pediatrician with IMCI implementation experience in sub-Saharan Africa. *Harness operators* require technical expertise to run generation and evaluation pipelines but not medical knowledge. *Model developers* can consume evaluation results without access to the underlying graph, enabling third-party evaluation with separation between evaluators and developers, a property the EvalEval community has identified as important for accountability (Reuel et al., 2025).

This separation also clarifies accountability: errors in evaluation results can be traced to graph inaccuracies (domain expert responsibility), generation bugs (harness operator responsibility), or model failures (developer responsibility).

4.4 Extensibility to Other Guidelines

The graph schema, including conditions, symptoms, treatments, follow-ups, severities, and their directed relationships, is not specific to IMCI. Any clinical guideline with structured decision logic is a candidate. WHO produces guidelines across malaria, tuberculosis, HIV, and maternal health that share the same flowchart structure as IMCI. Beyond healthcare, structured regulatory guidelines, legal compliance frameworks, and technical standards with explicit relationship structures could support the same approach. The primary requirement is that the source document encodes relationships explicitly enough to support graph construction, a property common to clinical and regulatory guidelines by design.

4.5 Limitations

Question quality depends entirely on graph accuracy: any errors in manual annotation propagate to all generated questions. The graph was curated by a single clinical expert, which precludes inter-rater reliability assessment; independent validation by additional clinicians with IMCI expertise remains important future work for establishing the rigor required of a production evaluation instrument. We evaluate only MCQA format, which cannot capture the complexity of real clinical reasoning involving differential diagnosis and incomplete information. Our text-only approach excludes visual diagnostic elements present in the original IMCI handbook. While question generation is automated, initial graph construction remains manual, limiting scalability. Our evaluation on IMCI guidelines may not generalize to other medical domains. Although the framework admits a large combinatorial space of possible instances, the practically valid subset is smaller because clinical constraints introduce dependencies among age, condition, and distractor choices, and we have not exhaustively verified all such variants. Finally, the absence of a human expert baseline makes it difficult to interpret absolute model accuracy; frontier models scoring 71–72% may represent strong or weak performance depending on task difficulty, and establishing a human ceiling is an important direction for future work.

4.6 Potential Risks

This work presents evaluation tools for medical AI systems. Models performing well on MCQA may still fail in actual clinical scenarios requiring differential diagnosis and incomplete information. Any errors in manual graph annotation propagate to evaluation, potentially validating incorrect medical knowledge. Our focus on WHO IMCI guidelines may not generalize to other healthcare contexts. This evaluation harness is intended for research purposes only and is not suitable for clinical decision-making.

5 Conclusion

This work introduces a graph-based evaluation harness for systematically instantiating evaluation queries from clinical guidelines, demonstrated on the WHO IMCI handbook. By transforming medical guidelines into queryable graphs, the framework achieves complete coverage of encoded relationships, which is not feasible through manual curation alone. Its dynamic design allows new evaluation instances with different ages and distractors to be sampled continuously, including as guidelines are updated. While baseline inference provides initial scores, the main value lies in granular performance across relationship types, which reveals systematic strengths and weaknesses in clinical protocol understanding.

The clinical validity of the generated questions rests on expert authorship of the underlying graph rather than post-hoc sampling, a design choice that both strengthens the validity claim and clarifies the role of domain expertise in evaluation infrastructure. The graph-based approach is extensible beyond IMCI, addressing the gap between general-purpose benchmarks and real-world domain-specific applications.

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.

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.

Hang Dong, Jiaoyan Chen, Yuan He, and Ian Horrocks. 2023. Ontology enrichment from texts: A biomedical dataset for concept discovery and placement. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*,

CIKM '23, page 5316–5320, New York, NY, USA. Association for Computing Machinery.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of ICLR*.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of EMNLP-IJCNLP*, pages 2567–2577.

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.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2022. Frenchmedmcqa: A french multiple-choice question answering dataset for medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 41–46, Abu Dhabi, United Arab Emirates.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. 2025. Sequential diagnosis with language models. *Preprint*, arXiv:2506.22405.

Tobi Olatunji, Abraham Owodunni, Tassallah Abdullahi, Ayokunmi Ilesanmi, Olalekan Obadun, Aimérou Ndiaye Etori, Ifeoma Okoh, Evans Doe Ocansey, Wendy Kinara, Michael Best, and 1 others. 2024. Afrimedqa: A pan-african, multi-specialty, medical question-answering benchmark dataset. *arXiv preprint arXiv:2411.15640*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *Preprint*, arXiv:2307.15343.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of EMNLP*, pages 2357–2368.

Anka Reuel, Avijit Ghosh, Jenny Chim, Andrew Tran, Yanan Long, Jennifer Mickel, Usman Gohar, Srishti Yadav, Pawan Sasanka Ammanamanchi, Mowafak Allaham, Hossein A. Rahmani, Mubashara Akhtar, Felix Friedrich, Robert Scholz, Michael Alexander Riegler, Jan Batzner, and 1 others. 2025. Who evaluates AI's social impacts? Mapping coverage and gaps in first and third party evaluations. *arXiv preprint arXiv:2511.05613*.

Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. 2022. Radqa: A question answering dataset to improve comprehension of radiology reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6250–6259, Marseille, France. European Language Resources Association.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards conversational diagnostic ai. *Nature*, 629(8010):331–338.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

WHO. 2014. Integrated management of childhood illness - chart booklet. Technical report, World Health Organization. Technical document.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8).

A Distractor Pool Construction

We formalize distractor construction for completeness. Let $G = (V, E)$ denote the IMCI knowledge graph.

For a question with correct answer v_{corr} of type τ and age range α, the distractor pool is defined as

$$P_{\tau,\alpha} = \begin{cases} C_\alpha \setminus \{v_{\text{corr}}\}, & \tau = \text{Cond}, \\ \mathcal{N}_{\tau,\alpha} \setminus \{v_{\text{corr}}\}, & \tau \in \mathcal{T}, \\ S \setminus \{v_{\text{corr}}\}, & \tau = \text{Sev}. \end{cases} \quad (1)$$

where $\mathcal{T} = \{\text{Sym}, \text{Treat}, \text{FollowUp}\}$.

The condition set is

$$C_\alpha = \{c \in V : \text{type}(c) = \text{Condition},\ \text{age_range}(c) = \alpha\}, \quad (2)$$

and the aggregated neighborhood is

$$\mathcal{N}_{\tau,\alpha} = \bigcup_{c \in C_\alpha} N_\tau(c). \tag{3}$$

The neighborhood function is

$$N_\tau(c) = \begin{cases} \{u : (u,c) \in E,\ \mathrm{type}(u) = \tau\}, & \tau = \mathrm{Sym}, \\ \{u : (c,u) \in E,\ \mathrm{type}(u) = \tau\}, & \tau \in \{\mathrm{Treat}, \mathrm{FollowUp}\}. \end{cases} \tag{4}$$

For severity classification,

$$S = \{u \in V : \mathrm{type}(u) = \mathrm{Severity}\}. \tag{5}$$

The final distractor set is

$$D = \mathrm{sample}(P_{\tau,\alpha}, k), \tag{6}$$

where $k = 3$.