CONTEXT0: ZERO-LABEL SELF-EVOLUTION FOR CONTEXT-AWARE CLINICAL DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical dialogue is inherently context-dependent: an appropriate response must account for patient-specific constraints and healthcare settings, recognize when consequential information is missing, and avoid unnecessary clarification when the available context is sufficient. Developing these behaviors requires training experiences that capture not only medical knowledge, but also when and how that knowledge should be applied. We introduce **Context0**, a zero-label self-evolution framework that develops context-aware clinical dialogue capabilities without a pre-existing clinical dialogue dataset or example-level human annotations. Starting from zero curated training examples, a Teacher constructs context-sensitive dialogue tasks, while a Doctor learns from case-specific rubric feedback. A medical knowledge graph provides knowledge anchors for task construction, and feedback from the evolving Doctor guides the Teacher toward new learning opportunities. Alternating Teacher and Doctor updates jointly evolve the training curriculum and the response policy. On the Global Health subset of HealthBench, Context0 achieves a context-awareness score of 62.3%, exceeding GPT-5 by 3.2 percentage points in our evaluation. It also achieves full-rubric scores of 57.9% and 62.7% on the unclear- and clear-context subsets, respectively. These results highlight the potential of zero-label self-evolution to develop context-aware medical assistants through self-generated curricula and feedback. We will release the code and curriculum at `https://anonymous.4open.science/r/Context0-2878`.

1 INTRODUCTION

Context determines whether medical knowledge translates into an appropriate response. The same clinical question may require different guidance depending on a patient's history, current medications, available healthcare resources, or local practice conditions. An assistant must therefore do more than recall medical facts: it must identify which contextual factors matter and respond according to what the conversation establishes (Arora et al., 2025). When consequential information is missing, it should seek clarification or provide appropriately conditional guidance rather than silently assume an answer (Li et al., 2024b). Yet not every missing detail warrants another question. When sufficient information is already available, unnecessary clarification can distract from useful guidance rather than improve it (Arora et al., 2025).

Learning this context-sensitive behavior requires more than training on medically correct answers. A response can be factually accurate while overlooking a disclosed constraint, asking for information already provided, or making an unsupported assumption about the user. The appropriate next response depends on the relationship between medical knowledge and the information available in the dialogue. Training must therefore distinguish among three situations: relevant context is already established, consequential context is missing, or additional context would not materially change the answer. A central challenge is how to construct learning experiences and feedback that capture these distinctions without relying on a pre-existing collection of annotated clinical dialogues.

Self-evolving training offers a promising starting point by allowing models to generate learning tasks and adapt them to the learner's changing capabilities (Huang et al., 2026; Xia et al., 2025). For context-aware clinical dialogue, however, making tasks harder is not sufficient. An evolving curriculum must expose failures in how the learner uses context, rather than merely increase the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

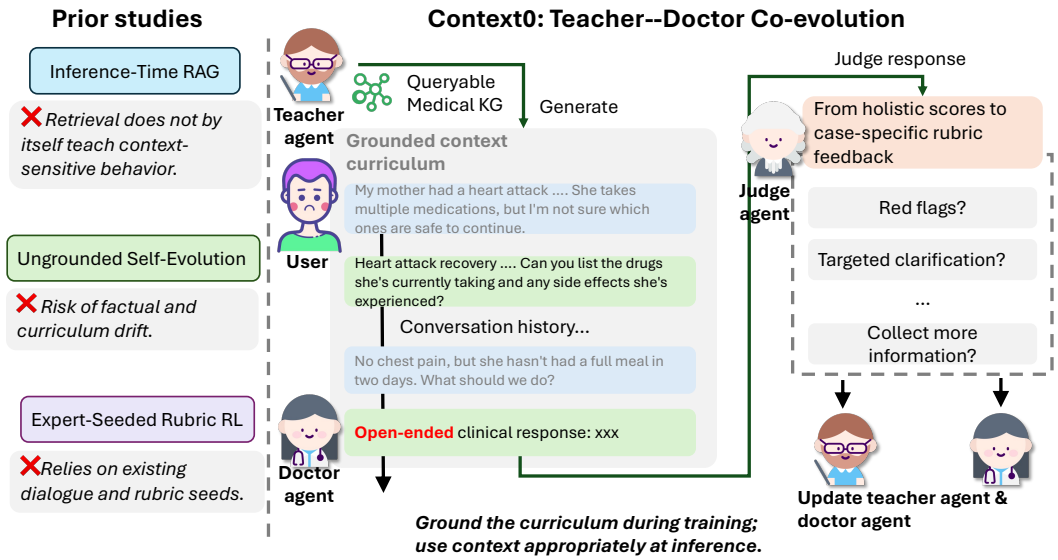

Figure 1: **Context0: zero-label self-evolution for context-aware clinical dialogue.** Starting from zero curated dialogue examples, a Teacher constructs context-sensitive clinical tasks using knowledge-graph anchors, and a Doctor learns from case-specific rubric feedback. Feedback from the evolving Doctor guides subsequent curriculum construction, forming a Teacher–Doctor co-evolution loop.

difficulty of medical questions. Its feedback must also distinguish appropriate clarification, effective use of disclosed information, and unnecessary information seeking. These requirements motivate a self-evolution process in which curriculum construction and contextual feedback develop together.

We introduce **Context0**, a zero-label Teacher–Doctor self-evolution framework for context-aware clinical dialogue. Starting from zero curated training examples, Context0 generates its own dialogue curriculum and case-specific evaluation criteria without requiring example-level human annotations. The Teacher constructs clinical scenarios using a medical knowledge graph as a source of knowledge anchors, while the Doctor learns from rubric-based feedback on its responses. Feedback from the current Doctor, in turn, guides the Teacher toward new contextual learning opportunities. Through alternating updates, the framework jointly evolves what the Doctor learns and how it responds.

We evaluate Context0 on the Global Health subset of HealthBench (Arora et al., 2025), where adapting responses to healthcare context is particularly important. Context0 achieves 62.3% in context awareness, compared with 59.1% for GPT-5 in our evaluation. It also achieves full-rubric scores of 57.9% and 62.7% on the unclear- and clear-context subsets, exceeding GPT-5 by 2.8 and 6.3 percentage points, respectively. These results demonstrate the potential of self-generated curricula and feedback for developing context-aware clinical response behavior.

Our contributions are threefold:

- **Zero-label self-evolution for clinical dialogue.** We introduce a Teacher–Doctor framework that constructs and evolves its own training curriculum from zero curated clinical dialogue examples, without example-level human annotations.
- **Context-aware curriculum and feedback.** We couple knowledge-anchored task construction with case-specific rubrics and learner-dependent curriculum selection to train responses that account for available and missing context.
- **Strong context-aware performance in global health.** Context0 achieves 62.3% context awareness on the Global Health subset of HealthBench, exceeding GPT-5 by 3.2 percentage points, and obtains higher full-rubric scores in both the unclear- and clear-context strata.

2 Related Work

2.1 Self-Evolution with Reduced Curated Supervision

Recent work has explored self-evolving or self-improving language model frameworks that reduce reliance on curated supervision by generating training tasks, reasoning trajectories, or feedback autonomously (Zelikman et al., 2022; Madaan et al., 2023; Shinn et al., 2023). A prominent setting is verifiable reasoning, including coding and mathematics (Guo et al., 2025; Shao et al., 2024), where execution feedback and deterministic answers can provide training signals. Other approaches use confidence or consistency as proxy signals (Yuan et al., 2024). More recent work investigates systems that generate and solve their own problems, further reducing dependence on externally supplied training tasks (Xia et al., 2025; Zhao et al., 2025; Huang et al., 2026; Wang et al., 2025b; Lu et al., 2025).

For clinical dialogue, increasing task difficulty or answer correctness alone does not ensure appropriate use of context: a learner may still ignore disclosed constraints, assume missing information, or ask redundant questions. Context0 targets this gap through Teacher–Doctor co-evolution, starting from zero curated clinical dialogue examples. The Teacher constructs knowledge-anchored contextual tasks, while case-specific rubrics guide Doctor training and Doctor feedback informs subsequent curriculum construction. The focus is thus on evolving contextual response behavior, rather than medical question difficulty alone.

2.2 Context-Aware Medical Dialogue and Rubric-Based Learning

Medical language models are commonly evaluated using static question-answering benchmarks (Singhal et al., 2025; Jin et al., 2019; Pal et al., 2022; Chen et al., 2025; Zuo et al., 2025). Moving beyond this setting, MediQ (Li et al., 2024b) and KnowGuard (Dang et al., 2025) study question asking for interactive clinical reasoning. AgentClinic (Schmidgall et al., 2024) evaluates agents in simulated clinical environments, while CRAFT-MD (Johri et al., 2025) assesses conversational diagnosis through doctor–patient interaction. Complementary work explores multi-agent collaboration through MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024), and doctor-agent evolution through simulated patient interactions in Agent Hospital (Li et al., 2024a). Rubric-based learning further supports explicit specification of desirable response behavior: InfiMed-ORBIT (Wang et al., 2025a) combines synthetic medical dialogue with dynamically constructed rubrics and rubric-guided reinforcement learning.

These directions advance medical reasoning, interaction, and feedback, but information acquisition or diagnostic accuracy alone does not capture whether the next response appropriately uses the conversation history. The learning objective must also distinguish consequential missing information from disclosed context and details that would not materially change the answer. Building on rubric-guided response learning, Context0 couples case-specific feedback with a learner-adaptive, self-generated curriculum. Its contribution is not rubric feedback alone, but its integration with zero-label Teacher–Doctor co-evolution to learn when to use available context, seek clarification, or answer without further questioning.

3 Methodology

3.1 Framework Overview

Context-aware clinical assistance requires more than producing a medically plausible answer to the latest user message. A response may need to incorporate information disclosed earlier in the conversation, seek clarification when missing information could change the recommendation, or answer directly when additional context would not affect the core guidance. Our framework, Context0, targets these distinct behaviors through iterative Teacher–Doctor co-evolution.

As shown in Figure 2, the framework connects knowledge-grounded task construction, learner-dependent feedback, offline curriculum curation, and Doctor optimization. The Teacher constructs clinical dialogues and case-specific rubrics using knowledge-graph evidence. A frozen snapshot of the current Doctor then responds to these dialogues, providing feedback about how the learner

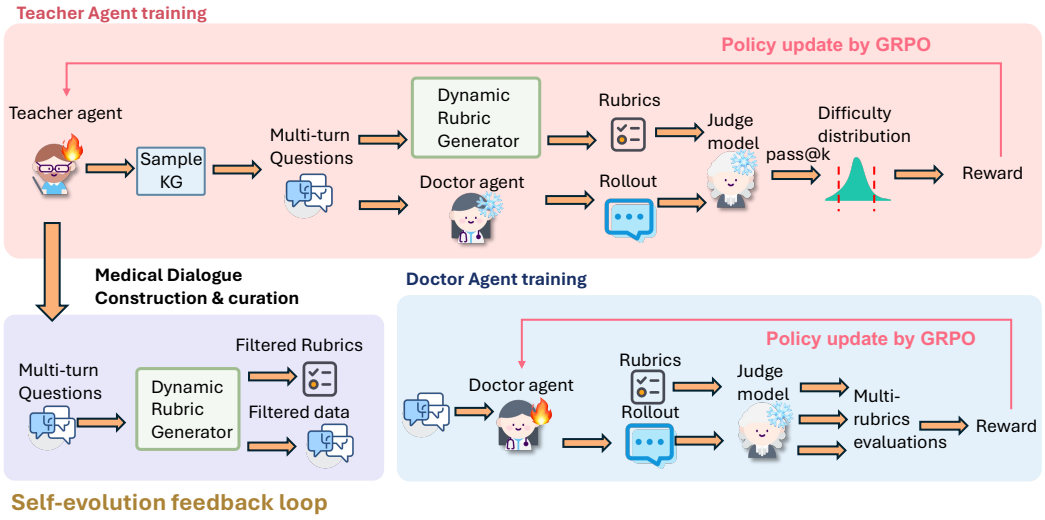

Figure 2: **Overview of Context0.** The Teacher constructs clinical dialogues and case-specific rubrics from biomedical KG evidence. A frozen Doctor provides feedback for Teacher optimization using complete rubrics and for offline curation that selects dialogues and prunes criteria. The Doctor is then trained with signed rubric rewards, and the updated Doctor informs the next round. Only the Doctor is required at inference time.

handles the generated tasks. This feedback serves two separate purposes: it guides Teacher optimization and determines which dialogue–rubric pairs are retained for Doctor training. After the Doctor is updated, the next round uses the new snapshot to reassess generated tasks.

Given a user-visible conversation history h, the Doctor generates a response $y \sim \pi_\theta(\cdot \mid h)$. Its objective is to use relevant available context, recognize consequential information gaps, and avoid unnecessary clarification. KG evidence, oracle context, and scoring rubrics are auxiliary resources for task construction and evaluation; they are not included in the Doctor's response-generation input. This separation keeps the generation task focused on information available within the conversation.

A central design choice is to distinguish *Teacher optimization*, *offline curation*, and *Doctor optimization*. Teacher optimization uses soft feedback computed from complete rubrics to improve task construction. Offline curation makes discrete decisions about which dialogues and criteria enter the training curriculum. Doctor optimization then uses signed rewards computed from the retained rubrics. Keeping these operations separate allows learner feedback to influence both generation and selection without conflating reward shaping with data filtering.

3.2 Knowledge-Grounded Task Construction

Knowledge source and motivation. Clinical task construction requires both biomedical content and a concrete contextual decision for the Doctor to make. We use a knowledge graph (KG) to provide explicit evidence around which the Teacher can organize a case, rather than asking the Teacher to generate all biomedical content solely from its parametric knowledge.

Specifically, we use OptimusKG (Vittor et al., 2026), a multimodal biomedical knowledge graph covering entities such as diseases, drugs, phenotypes, anatomical structures, and exposures. Its broad biomedical coverage supports the construction of varied clinical scenarios, while its typed relations provide structured connections that can be used to specify case-relevant evidence. Although OptimusKG is multimodal, our construction pipeline uses retrieved relational triplets as textual inputs to the Teacher. These triplets provide traceable source information; however, provenance alone does not establish that every statement in a generated dialogue is supported (See Appendix A.5).

Contextual task design. Conditioned on retrieved evidence g, the Teacher π_ϕ generates a conversation history h, a context stratum s, and a preliminary rubric $\mathcal{R}^{\text{draft}}$. We distinguish three contex-

tual targets because appropriate context use does not always correspond to asking more questions or giving a more personalized answer.

For *context matters but is unclear*, the conversation leaves consequential information unresolved, and the desired response seeks relevant clarification or provides appropriately conditional guidance. For *context matters and is clear*, the necessary information is already available, and the response should adapt to it rather than request it again. For *context does not matter*, the core response should remain stable without unnecessary clarification. Together, these strata distinguish whether context is needed from whether the needed context is available.

The generated histories may contain multiple turns, with relevant information appearing before the final user message. Accordingly, each rubric evaluates the next response against the entire conversation rather than against the latest message in isolation. Appendix A.4 summarizes the contextual behavior targeted by each synthetic task stratum.

Case-specific rubric construction. A single holistic score can obscure why a response succeeds or fails: a clinically plausible answer may still ignore disclosed information, miss a consequential uncertainty, or ask for information that is already available. We therefore represent supervision using explicit, case-specific behavioral criteria.

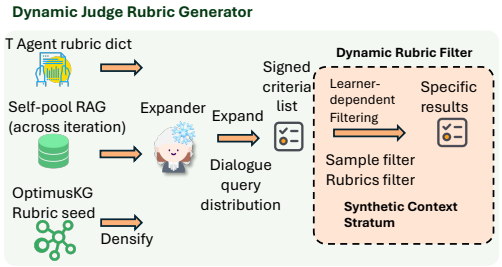

Figure 3: Case-specific rubric construction in Context0. The Teacher's preliminary rubric is augmented with KG-native templates and examples from a self-pool of previously accepted dialogues and rubrics. A frozen expansion model adapts these inputs to the conversation, producing criteria that reward desired behaviors and penalize errors through signed weights. The complete rubric supports Teacher feedback, while retained criteria guide Doctor optimization.

As shown in Figure 3, rubric construction combines the Teacher's preliminary draft with KG-native templates and a self-pool of previously accepted synthetic dialogue–rubric pairs. The templates supply reusable criterion patterns instantiated using the retrieved evidence, while the self-pool supplies examples accepted in earlier rounds. A frozen expansion model adapts the augmented material to the current case, producing a rubric that evaluates the intended response behaviors in the context of the generated history.

The resulting rubric is $\mathcal{R}(h) = \{(c_j, w_j)\}_{j=1}^{M_h}$, where c_j describes a response behavior and w_j is its signed weight. Positive weights reward desired behaviors, whereas negative weights penalize specified errors. This representation makes desirable actions and undesirable behaviors explicit within the same scoring interface.

The self-pool is initially empty and is updated with accepted dialogue–rubric pairs after each round. It therefore provides a mechanism for reusing accepted construction patterns without requiring the entire rubric to be generated from scratch in every round. Appendix A.3 distinguishes the rubric states used throughout the pipeline.

3.3 Learner Feedback and Offline Curation

The usefulness of a generated task depends partly on the Doctor's current behavior. A fixed collection of synthetic tasks cannot reflect changes in what the learner already handles reliably or continues to miss. We therefore evaluate candidate tasks using the current Doctor and use the resulting feedback both to shape subsequent generation and to curate the training data.

At round t, we freeze the current Doctor as $\bar{\pi}_t$ and sample multiple responses to each candidate history. Holding this snapshot fixed ensures that feedback within the round refers to the same learner. Teacher feedback uses K_C responses per history, whereas offline curation uses K_D responses; these sampling counts serve different operations and need not be identical.

For a criterion c_j with nonzero signed weight w_j, let $J_\omega(h, y, c_j) \in \{0, 1\}$ indicate whether the specified behavior is present: satisfaction of the desired behavior for a positive criterion, or occurrence

of the specified error for a negative criterion. We define the signed-criterion satisfaction indicator as

$$S(h, y, c_j) = \begin{cases} J_\omega(h, y, c_j), & w_j > 0, \\ 1 - J_\omega(h, y, c_j), & w_j < 0. \end{cases}$$

Thus, $S = 1$ consistently denotes a successful outcome: exhibiting a desired behavior or avoiding a specified error.

For K sampled responses, the empirical satisfaction rate is $\hat{p}_j^{(K)}(h) = K^{-1} \sum_{i=1}^{K} S(h, y_i, c_j)$, and its average across the complete rubric is $\bar{s}^{(K)}(h) = M_h^{-1} \sum_{j=1}^{M_h} \hat{p}_j^{(K)}(h)$. These statistics summarize the frozen Judge's assessments of sampled responses, rather than independently validated clinical correctness or learner mastery.

Soft feedback for Teacher optimization. The Teacher needs feedback about the tasks it generates, but this feedback need not take the form of immediate sample rejection. We therefore use the complete rubric to compute soft reward shaping during Teacher optimization. The feedback combines rubric-level judgment statistics with contextual probes that assess the behavior targeted by each context stratum.

For unclear cases, the probes assess relevant clarification and unsupported certainty. For clear cases, they assess whether the response uses information already disclosed. For context-invariant cases, they assess consistency under clinically meaning-preserving paraphrases. These probes are intended to complement rubric judgments by examining how responses relate to the contextual structure of the task. They do not append oracle information to the Doctor's response-generation input.

No criterion is removed during this stage. The complete rubric remains available when computing Teacher feedback, so the signal used to improve task generation is distinct from the reduced supervision subsequently used to train the Doctor.

Offline curriculum curation. Soft reward shaping influences the Teacher's generation policy, but it does not explicitly determine which generated examples the Doctor will train on. We therefore apply a separate offline curation step, as illustrated in Figure 2.

First, we retain dialogues whose average judgment statistic lies within $[\tau_{\text{low}}, \tau_{\text{high}}]$. This defines the target judgment band for curriculum inclusion. Second, within each retained dialogue, we remove criteria whose empirical judgment rates are at or above τ_r:

$$\mathcal{R}_{\text{filt}}(h) = \{(c_j, w_j) \in \mathcal{R}(h) : \widehat{p}_j^{(K_D)}(h) < \tau_r\}. \tag{1}$$

The dialogue-level rule controls sample inclusion, whereas the criterion-level rule controls which behaviors remain in the training reward.

Finally, we discard dialogues with fewer than $M_{\min}$ remaining criteria. The resulting $\mathcal{R}_{\text{filt}}(h)$ is the immediate post-filter rubric. In the Full configuration, offline rubric densification is subsequently applied to produce the final training rubric, denoted by $\mathcal{R}_{\text{train}}(h)$. The exported histories and their final training rubrics form the curriculum $\mathcal{D}_t$ for the current Doctor update.

3.4 Teacher–Doctor Optimization

The Teacher and Doctor optimize different objectives. The Teacher is trained to construct grounded, contextually meaningful tasks that provide useful learner feedback. The Doctor is trained to produce responses that satisfy the final case-specific training criteria. We alternate these updates so that task construction can respond to the evolving learner while each Doctor update uses a fixed curriculum.

Teacher update. With the Doctor snapshot fixed, we optimize the Teacher using a composite reward:

$$r_T(x) = R_{\text{quality}}(x) + R_{\text{feedback}}(x) - \lambda_{\text{red}} P_{\text{red}}(x), \tag{2}$$

where x denotes a generated candidate. The quality term groups format validity, KG grounding, and scenario consistency. The feedback term groups judgment-band shaping, rubric informativeness, and contextual-probe feedback. The redundancy penalty discourages repetitive generation.

These components address complementary requirements. Learner feedback alone does not establish that a task is well formed or grounded, while construction quality alone does not indicate how the current Doctor responds to it. The composite reward therefore combines task-quality checks with learner-dependent signals. All feedback is computed using complete rubrics before offline criterion pruning.

Doctor update. After curation, we hold $\mathcal{D}_t$ fixed and optimize the Doctor using a frozen criterion-level Judge. For each generated response, the Judge evaluates the behaviors described by the final training rubric. Let $J_\omega(h, y, c_j) \in \{0, 1\}$ indicate whether the behavior described by c_j is present in response y. The Doctor's signed-rubric reward is

$$r_D(h, y) = \sum_{(c_j, w_j) \in \mathcal{R}_{\text{train}}(h)} w_j J_\omega(h, y, c_j). \tag{3}$$

Matching a positive criterion increases the reward, whereas matching a negative criterion decreases it. Importantly, the Judge detects the specified behavior; the sign of w_j determines whether that behavior is desirable. This separates criterion detection from reward assignment and supports both positive and negative supervision within the same rubric.

The Doctor receives only the conversation history when generating its response. The final training rubric is used for scoring rather than exposed as an additional instruction to the response policy. Appendix A.4 specifies the signed criterion-reward convention, and Appendix A.4 lists the recorded context-axis and rubric-source weights.

Policy optimization and round transition. Both policies are optimized with GRPO (Shao et al., 2024). For Doctor optimization, we sample a group of G_D responses for each history and compute group-relative advantages $\widehat{A}_i = (r_i - \bar{r})/(\sigma_r + \epsilon)$, where $\bar{r}$ and σ_r are the mean and standard deviation of the group rewards. The optimization group size G_D is distinct from the Teacher-feedback count K_C and the curation count K_D: the former supports policy optimization, while the latter two support task assessment.

At the end of each round, the updated Doctor becomes the frozen feedback model for the next round, and accepted dialogue–rubric pairs are added to the self-pool. As shown by the feedback loop in Figure 2, subsequent task construction and selection are therefore conditioned on the Doctor's evolving behavior rather than on a permanently fixed learner. The Teacher, KG retrieval pipeline, rubric expansion model, and Judge are training-time components; only the Doctor is required at inference time.

4 Experiments

4.1 Experimental Setup

Benchmark and metrics. We use the 1,097-conversation Global Health subset of HealthBench (Arora et al., 2025), partitioned into a fixed, stratified development set of 200 conversations and a disjoint test set of 897 conversations. The development set is used for pipeline development, component ablations, and progression gates. The test set is excluded from these decisions and used only for final evaluation. For each conversation, we generate one next assistant response from the supplied history.

We report the context-awareness score using its corresponding physician-written criteria and full-rubric scores within two physician-defined strata: Context unclear and Context is clear. These strata distinguish whether relevant healthcare context is insufficiently specified or already available; they are separate from synthetic training labels. All main results are computed on the test set. All scores are percentages, with higher values indicating better performance.

Comparison systems. We compare against GPT-5 (OpenAI, 2025), GPT-5 mini (OpenAI, 2025), DeepSeek V3 (Liu et al., 2024), Qwen3-4B-Base (Yang et al., 2025), Qwen3-8B-Base (Yang et al., 2025), Qwen3-4B-Instruct-2507 (Yang et al., 2025), and the self-evolution baselines Agent0 (Xia et al., 2025) and R-Zero (Huang et al., 2026).

Table 1: Context-aware response performance on the held-out 897-conversation test split of the Global Health subset of HealthBench. Context awareness is scored using context-awareness criteria; Context unclear and Context is clear use complete rubrics within their respective physician-defined strata. All values are percentages; higher is better.

Model / Method	Context awareness ↑	Context unclear ↑	Context is clear ↑
GPT-5	59.1	55.1	56.4
GPT-5 mini	55.8	52.4	51.6
DeepSeek V3	28.6	32.7	37.4
Qwen3-4B-Base	3.9	2.5	8.9
Qwen3-8B-Base	10.0	10.3	15.2
Qwen3-4B-Instruct-2507	24.8	26.3	26.6
Agent0	24.4	28.7	25.7
R-Zero	24.9	28.4	28.3
Context0 (Ours)	**62.3**	**57.9**	**62.7**

Evaluation judge. GPT-5 grades responses against HealthBench rubrics and also serves as a response-generating baseline. This grader is separate from the frozen instruction-tuned Qwen 4B model used for training-time judgments.

4.2 Main Results

Context-awareness performance. As shown in Table 1, Context0 achieves 62.3% context awareness, the highest score among the evaluated systems. It exceeds GPT-5 by 3.2 percentage points and GPT-5 mini by 6.5 points. Compared with Qwen3-4B-Instruct-2507, Context0 improves context awareness by 37.5 points. It also exceeds Agent0 and R-Zero by 37.9 and 37.4 points, respectively, highlighting its strong performance on context-sensitive clinical response generation.

Performance across contextual strata. Context0 achieves full-rubric scores of 57.9% on Context unclear and 62.7% on Context is clear. These scores exceed GPT-5 by 2.8 and 6.3 percentage points, respectively. Relative to Qwen3-4B-Instruct-2507, the gains are 31.6 and 36.1 points. Context0 therefore obtains higher response-quality scores both when relevant healthcare context is insufficiently specified and when that context is already available.

5 Analysis

We examine component contributions and sensitivity to the knowledge source. A descriptive analysis of 77 generated unclear-context cases is provided in Appendix A.1.

5.1 Component Ablations

We ablate multi-turn context, selected clinical safety gates, and offline rubric densification. All variants start from Qwen3-4B-Instruct-2507 and are evaluated on the same fixed 200-conversation HealthBench development subset. This split is disjoint from the 897-conversation test split used for the main results.

As shown in Figure 4, removing multi-turn context lowers context awareness by 12.1 percentage points and clear-context performance by 18.9 points, while reducing unclear-context performance by only 1.6 points. Removing the red-flag and medication-blacklist gates reduces context awareness by 16.1 points and clear-context performance by 14.2 points. Removing rubric densification produces corresponding declines of 11.8 and 8.9 points.

Across all three ablations, unclear-context scores range from 48.6% to 50.4%, compared with 50.3% for the full model. The larger differences in context awareness and clear-context performance suggest that these components are particularly beneficial when relevant context is already available.

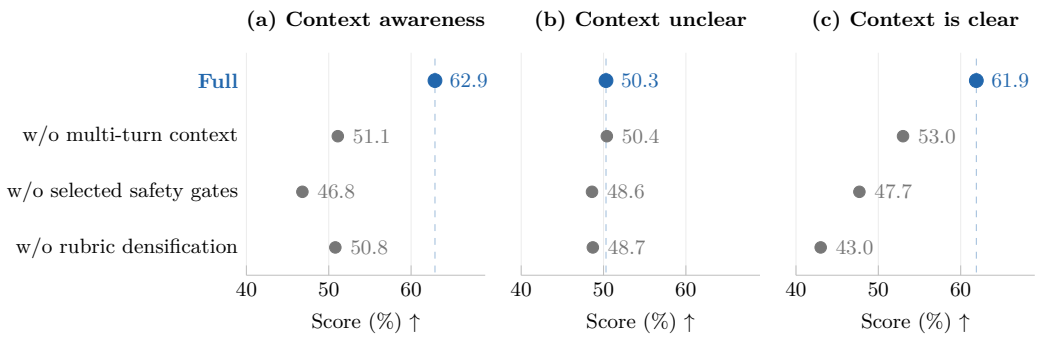

Figure 4: Component ablations on the fixed 200-conversation HealthBench development split. Context awareness uses axis-specific criteria; the other panels use full rubrics within physician-defined strata. Dashed lines mark the full configuration. All scores are percentages; higher is better.

Table 2: Sensitivity to the knowledge source. Context awareness uses context-awareness criteria; the remaining columns use complete rubrics within the corresponding physician-defined strata. All values are percentages; higher is better.

Configuration	Context awareness ↑	Context unclear ↑	Context is clear ↑	Context does not matter ↑
Reference KG	**62.7**	**60.0**	67.6	**64.0**
No KG	56.5	57.4	67.3	63.8
Incomplete KG (50%)	58.6	58.4	65.7	60.0
Out-of-domain KG (ConceptNet)	56.4	52.4	**69.5**	62.8

However, these pipeline-level comparisons do not isolate the mechanisms underlying the improvements or independently establish clinical safety.

5.2 Sensitivity to the Knowledge Source

We compare the reference medical KG with no KG, a 50% incomplete KG, and ConceptNet on the development data. Appendix A.5 specifies the triplet-removal, triplet-subsampling, and ConceptNet replacement operations.

The reference KG achieves the highest context-awareness and unclear-context scores, exceeding the no-KG variant by 6.2 and 2.6 percentage points, respectively (Table 2). The incomplete KG falls between these configurations on both metrics. However, ConceptNet obtains the highest clear-context score (69.5%), and the reference and no-KG variants perform similarly when context does not matter (64.0% versus 63.8%). Thus, medically relevant knowledge anchors support context awareness, but their advantages are not uniform across strata.

6 Conclusion

We presented Context0, a framework for training context-aware clinical assistants through multi-turn synthetic curricula, safety-oriented filtering, and rubric-based supervision. HealthBench evaluations and component ablations support the value of multi-turn training, clinical safety gates, and rubric densification for contextual response quality. Knowledge-source comparisons show that the reference medical KG yields the strongest context awareness, although its advantages vary across contextual strata. Future work should assess broader clinical generalization and whether improved context awareness translates into appropriate clarification and safer clinical responses.

AI USE STATEMENT

In this work, we used generative AI tools for paper writing, We have not used generative AI tools for surveys or for forming research ideas. Additionally, we used generative AI tools for implementation of research ideas. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by the author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a research framework for open-ended medical reasoning and evaluates it only on academic benchmarks. It is not intended for direct clinical use. Any real-world deployment would require prospective clinical validation across diverse patient populations, regulatory review, continuous monitoring for bias and safety, and mandatory oversight by licensed medical professionals. In particular, errors in reasoning or abstention behavior could cause disproportionate harm in high-stakes settings. Accordingly, **Context0** should not be used for medical diagnosis or treatment decisions, and any future clinical application must comply with applicable medical, legal, and ethical requirements.

REPRODUCIBILITY STATEMENT

Section 3 describes the Context0 framework. Appendix A.2 documents the training scope and model roles, and Appendix A.3 describes the curriculum-processing stages and rubric states. Appendix A.4 summarizes the contextual feedback targets, signed rubric rewards, and recorded weighting configuration. Appendix A.5 specifies the knowledge-source interventions. Appendix A.1 provides a descriptive analysis of a sample of generated cases and rubrics. We will release code and curriculum at `https://anonymous.4open.science/r/Context0-2878`.

REFERENCES

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3563–3599, 2025.

Xilin Dang, Kexin Chen, Xiaorui Su, Ayush Noori, Iñaki Arango, Lucas Vittor, Xinyi Long, Yuyang Du, Marinka Zitnik, and Pheng Ann Heng. Knowguard: Knowledge-driven abstention for multi-round clinical reasoning. *arXiv preprint arXiv:2509.24816*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. In *International Conference on Learning Representations*, volume 2026, pp. 130770–130790, 2026.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86, 2025.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024a.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024b.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Hongliang Lu, Yuhang Wen, Pengyu Cheng, Ruijin Ding, Haotian Xu, Jiaqi Guo, Chutian Wang, Haonan Chen, Xiaoxi Jiang, and Guanjun Jiang. Search self-play: Pushing the frontier of agent capability without supervision. *arXiv preprint arXiv:2510.18821*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

OpenAI. Openai deep research system card, 2025. URL `https://openai.com/index/introducing-deep-research/`.

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

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Lucas Vittor, Ayush Noori, Iñaki Arango, Joaquín Polonuer, Sam Rodriques, Andrew White, David A Clifton, and Marinka Zitnik. Optimuskg: Unifying biomedical knowledge in a modern multimodal graph. *arXiv preprint arXiv:2604.27269*, 2026.

Pengkai Wang, Pengwei Liu, Qi Zuo, Zhijie Sang, Congkai Xie, and Hongxia Yang. Infimed-orbit: Aligning llms on open-ended complex tasks via rubric-based incremental training. *arXiv preprint arXiv:2510.15859*, 2025a.

Shaobo Wang, Zhengbo Jiao, Zifan Zhang, Yilang Peng, Xu Ze, Boyu Yang, Wei Wang, Hu Wei, and Linfeng Zhang. Socratic-zero: Bootstrapping reasoning via data-free agent co-evolution. *arXiv preprint arXiv:2509.24726*, 2025b.

Peng Xia, Kaide Zeng, Jiaqi Liu, Can Qin, Fang Wu, Yiyang Zhou, Caiming Xiong, and Huaxiu Yao. Agent0: Unleashing self-evolving agents from zero data via tool-integrated reasoning. *arXiv preprint arXiv:2511.16043*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data, 2025. URL https://arxiv.org/abs/2505.03335.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Appendix

A.1 Generated Curriculum and Rubric Composition

For this descriptive analysis, we present cases labeled as unclear-context in the synthetic curriculum. The sample contains 77 generated cases spanning 17 disease anchors; it does not characterize the full curriculum or other contextual strata. Gout is the most frequent anchor ($n = 14$), followed by coronary artery disease and hypertension ($n = 8$ each). Figure 5 summarizes the disease coverage and rubric composition of these cases.

The sampled rubrics contain 487 criterion instances: 238 KG-derived criteria, 18 generic safety criteria, and 231 fixed-context criteria. Each case includes three fixed-context criteria, accounting for 37.5% of an eight-criterion rubric to 75.0% of a four-criterion rubric. Within this unclear-context sample, rubric construction therefore combines shared contextual criteria with knowledge-derived clinical criteria. These counts describe criterion composition, not the relative contribution of each criterion type to the training reward.

A.2 Method Scope and Model Roles

Scope of Zero-Label Training Context0 is designed for knowledge-graph-grounded training without human-annotated task examples in the main training pipeline. Here, "zero-label" refers to this intended absence of example-level human annotations; it does not imply the absence of pretrained knowledge, external knowledge resources, or manually specified templates and rules.

The main rubric-construction path uses KG-native templates and a self-pool of generated examples. HealthBench cross-theme rubric seeds are specified as an optional ablation rather than a component of the main path. GPT-5 is not used as a training-data teacher or data generator for Context0. It is used as a downstream evaluation grader and separately as an answer-generation baseline.

Model Roles The experiment records identify `Qwen/Qwen3-4B-Instruct-2507` as the model used for the Teacher, the frozen initial Doctor E_0, and the training Judge. These are distinct functional roles even when they share the same base model family. The downstream evaluation grader is recorded as Azure `gpt-5`.

A.3 Curriculum Construction and Training Stages

Separation of Feedback, Curation, and Training The method distinguishes three stages:

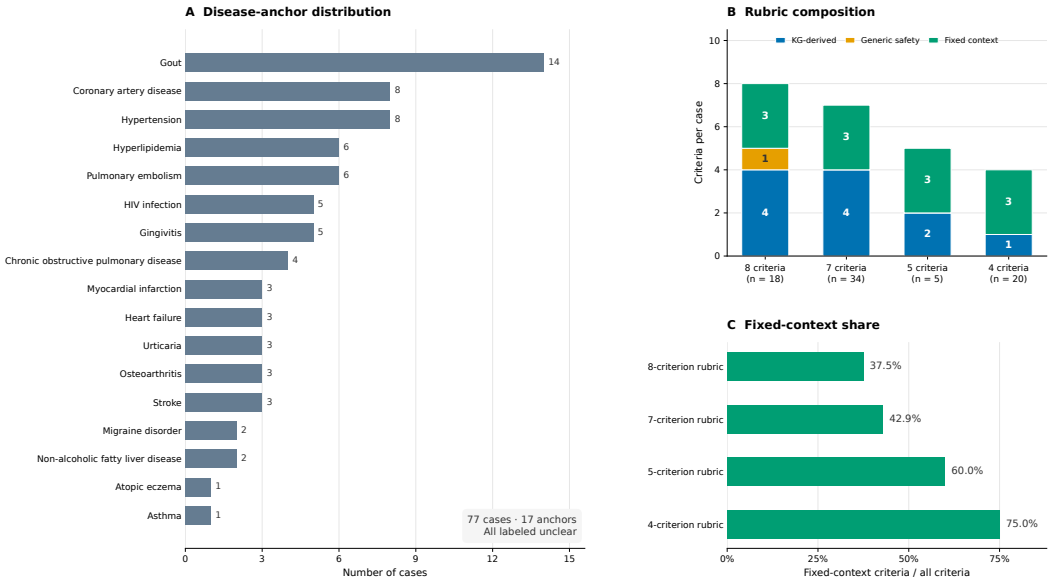

Figure 5: Disease coverage and rubric composition of 77 generated unclear-context cases, spanning 17 disease anchors and 487 criterion instances. Only cases with the synthetic unclear-context label are shown. The figure describes sample composition rather than independently validated clinical or rubric quality.

Role	Function
Teacher	Constructs candidate tasks and rubrics and is optimized using feedback from a frozen Doctor.
Frozen Doctor	Produces responses used to assess candidate tasks during curriculum optimization and curation.
Training Judge	Provides criterion-level judgments for feedback, curation, and rubric-based training rewards.
Trainable Doctor	Learns from the exported curriculum using rubric-based GRPO.
Evaluation grader	Scores downstream benchmark responses; this role is separate from the training Judge.

Table 3: Functional roles in Context0. The Teacher, Doctor, and training Judge share `Qwen/Qwen3-4B-Instruct-2507` as their base model. The downstream grader is Azure `gpt-5`.

1. **Teacher optimization.** Candidate tasks are assessed using responses from a frozen Doctor. The complete rubric available at this stage is used to form difficulty and contextual feedback. Criteria are not physically pruned during this feedback computation.
2. **Data curation.** Candidate conversations are selected and their rubrics are pruned using learner-dependent assessment. This stage changes the data exported for subsequent processing.
3. **Doctor optimization.** The Doctor is optimized using rubric-based GRPO, with rewards computed from the rubric stored in each training example.

This separation is important: evaluating a criterion during Teacher optimization does not imply that the criterion will remain in the exported training data. Conversely, rubrics can undergo additional processing after filtering.

Rubric States To distinguish the different points in this pipeline, we use the following terminology:

- **Preliminary rubric:** the initial rubric associated with a generated task.

- **Filtering-stage rubric:** the rubric presented to the multi-response assessment and filtering procedure.
- **Post-filter rubric:** the rubric immediately following criterion-level filtering.
- **Final training rubric:** the rubric stored in the exported training example after subsequent processing, including offline densification in the final Full configuration.

The "complete rubric" used during Teacher optimization means the unpruned rubric available at that stage. It does not imply that no later rubric expansion occurs. In particular, the post-filter rubric and the final training rubric should not be treated as interchangeable.

A.4 Contextual Feedback and Rubric-Based Rewards

Synthetic Context Strata The method design specifies contextual probes for three synthetic task strata:

Context matters but is unclear. The probe targets whether the Doctor requests relevant missing information, such as location or available resources, while avoiding unsupported certainty.

Context matters and is clear. The probe targets whether the Doctor uses information already disclosed in the conversation to adapt its recommendations.

Context does not matter. The probe targets whether the core recommendation remains consistent under reformulations that preserve the clinically relevant facts.

These are generation-time task categories in the method design. They should be distinguished from the physician-defined strata used in downstream benchmark evaluation.

Visible Context and KG Evidence The main input to the frozen Doctor is the user-visible conversation, stored as `prompt[]`. The oracle KG block is not used in the primary context-awareness comparison. The Doctor-training design likewise does not supply raw KG triplets or an oracle block as Doctor inputs.

Thus, KG grounding concerns the construction of tasks and supervision signals. It does not imply that the Doctor receives retrieved KG evidence during standard downstream evaluation.

Signed Criterion Rewards Let h denote the conversation history, y a Doctor response, and $\mathcal{R}_h = \{c_j\}_{j=1}^{M_h}$ the rubric associated with the training example. Let

$$J_\omega(h, y, c_j) \in \{0, 1\}$$

indicate whether the training Judge determines that the behavior described by criterion c_j is present.

Using signed criterion weights w_j, the base rubric reward is

$$r_D(h, y) = \sum_{j=1}^{M_h} w_j J_\omega(h, y, c_j). \tag{4}$$

Positive weights reward the occurrence of the specified behavior, whereas negative weights penalize it. This notation absorbs the criterion sign into w_j. It is equivalent to a magnitude-times-sign representation, but the sign must not be applied a second time.

Learner-Dependent Filtering Curriculum curation uses multiple responses from a frozen Doctor and criterion-level Judge assessments to select candidate conversations and prune rubrics. The design describes an aggregate assessment statistic of the form

$$\bar{s}_h = \frac{1}{K|\mathcal{R}_h|} \sum_{i=1}^{K} \sum_{j=1}^{|\mathcal{R}_h|} S(y_i, c_j), \tag{5}$$

where y_i is the i-th sampled response and S is the assessment indicator used by the filtering procedure.

Weighting Configuration The configuration contains:

- a context-axis weight of 2.5;
- an expansion-source weight of 1.5;
- a generic-source weight of 0.5.

The latter two are rubric-source weights, not capability-axis weights.

A.5 KNOWLEDGE-SOURCE ROBUSTNESS EXPERIMENTS

We describe the knowledge-source variants by their recorded operations rather than assuming that their names imply broader changes to the pipeline.

No-KG. This intervention clears the triplets in the training anchor subgraphs. The available records do not establish that all KG-derived rubric templates were also disabled. Consequently, this is a triplet-evidence removal intervention, not a verified removal of every KG-derived component.

Incomplete-50. This intervention randomly drops 50% of the triplets, using seed 20260813. The fraction refers to triplets, not nodes. It is not specified as a connectivity-preserving subgraph-sampling procedure.

ConceptNet. This intervention replaces the original graph evidence with an English-language sample of $5{,}000$ edges from ConceptNet 5.7.