SAGA-CDP: LARGE LANGUAGE MODELS CAN LEARN GUIDELINE-ADHERENT CLINICAL DECISION-MAKING FROM SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable clinical decision-making with large language models requires models to adhere to specified clinical guidelines throughout the decision-making process. In this work, we systematically investigate whether guideline adherence can be learned as a transferable capability from both real-world and synthetic guideline data. These two data sources present distinct trade-offs in guideline-adherence training. To enable structured synthetic data generation, we introduce SAGA-CDP, a data generation pipeline that produces fictional clinical guidelines, clinical decision pathways (CDPs), and case vignettes from shared underlying decision logic, yielding 2,000 fictional guidelines and 76,584 corresponding clinical cases. Extensive experiments demonstrate that real-world and fictional guideline data are both effective training sources, and that combining them achieves the strongest overall performance. Scaling analyses reveal distinct behaviors across the two data sources, with real-world guideline data continuing to benefit from more training cases, while further gains from fictional data depend more strongly on broader guideline coverage. Cross-regional real-world studies and controlled synthetic interventions provide evidence of context-memory conflict in guideline-adherent clinical decision-making. These findings establish real-world guideline data as a clinically grounded and viable resource for guideline-adherence training, while highlighting synthetic data as a scalable complement that increases the amount of training data and reduces reliance on real-world guideline knowledge.

1 INTRODUCTION

Large language models (LLMs) are increasingly expanding from traditional application domains Jiang et al. (2026); Song et al. (2026) into clinical decision support Liu et al. (2024); Tao et al. (2026). In this setting, models need to make decisions based on patient information and clinical evidence while following specific clinical practice guidelines rather than relying solely on previously acquired knowledge Hager et al. (2024). This makes the task demanding, because guideline-based decisions often involve conditional and sequential processes that require models to apply the appropriate rules in the correct order. Recent work Chen et al. (2026) has shown that models may deviate from guideline-defined clinical decision pathways (CDPs) while still arriving at correct final decisions, limiting their reliability in clinical decision support.

Existing studies on guideline adherence have primarily focused on evaluating whether LLMs can produce decisions consistent with clinical guidelines Li et al. (2025); Tan et al. (2026). Some prior work has further incorporated clinical guidelines into LLM-based decision support by providing guideline-derived prompts or executable guideline representations as context during inference Oniani et al. (2024); Deng et al. (2026), demonstrating the value of guidelines as an external source of decision knowledge. Nevertheless, systematic studies of guideline-adherence training remain limited. We therefore investigate whether guideline adherence can be learned through training as a transferable capability that generalizes beyond the guidelines seen during training.

Supervision derived from real-world clinical guidelines provides strong clinical grounding because it reflects authentic medical knowledge and decision rules used in practice Graham (2014). However, the amount and diversity of accessible guideline data are inherently constrained, and some

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.

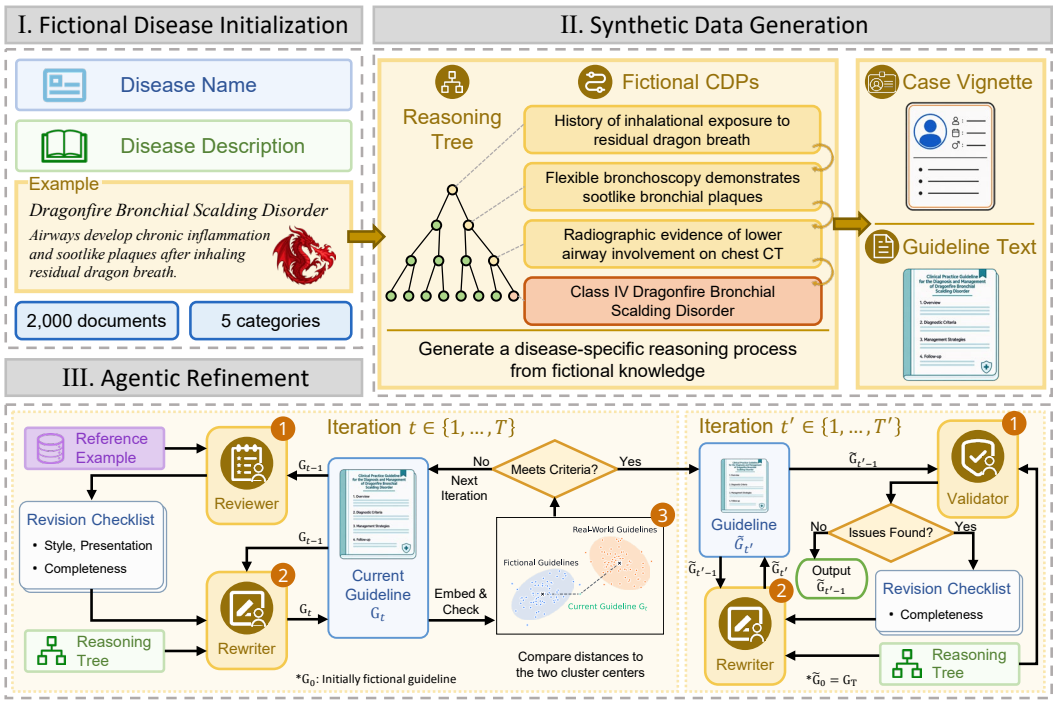


Figure 1: **Overview of SAGA-CDP**. The framework initializes fictional diseases, constructs disease-specific reasoning trees, fictional CDPs, case vignettes, and guideline texts, and further applies an agentic refinement process to improve guideline style, presentation, and completeness.

guidelines may be subject to copyright, licensing, institutional policies, or access restrictions that limit their use for model training. Training on real-world guidelines may lead models to internalize guideline-specific knowledge that does not necessarily generalize across clinical settings or remain applicable as guidelines evolve, raising the risk of context-memory conflict Kim et al. (2026); Sun et al. (2026). Synthetic guideline data therefore offer a natural complement, with greater flexibility in scale, style, and content. At the same time, synthetic data introduce their own challenges, as the quality of generated guidelines and decision processes is difficult to guarantee, and they may still exhibit systematic differences from real-world clinical guidelines. This trade-off motivates us to systematically compare real-world and synthetic supervision for guideline-adherence training.

To construct reliable synthetic data for guideline-adherence training, we introduce the **S**ynthetic **A**ugmentation for **G**uideline-**A**dherent **C**linical **D**ecision **P**athways (SAGA-CDP) pipeline, depicted in Figure 1. Starting from entirely fictional clinical knowledge, SAGA-CDP constructs a corresponding reasoning tree and uses the resulting decision logic to generate a fictional guideline, reference CDPs, and clinical cases. To reduce differences in style and presentation between fictional and real-world guidelines while ensuring complete coverage of the underlying decision logic, we further apply an agentic refinement workflow that iteratively reviews and revises the generated guidelines. Using this pipeline, we produce 2,000 fictional guidelines and 76,584 corresponding clinical cases, which serve as the synthetic training data in our subsequent experiments.

We systematically explore guideline-adherence training through the following research questions:

RQ1. *How effective are real-world and fictional guideline data for guideline-adherence training?*

We compare the effectiveness of training with real-world guideline data, fictional guideline data, and their combination, evaluating generalization to unseen real-world guidelines in terms of pathway consistency and outcome accuracy. The results show that both data sources substantially improve guideline adherence, with mixed training yielding further gains. (Section 4.2).

RQ2. *How does performance scale with the amount and diversity of guideline training data?*

We evaluate scaling across increasing numbers of training cases and varying guideline coverage for both real-world and fictional guideline data. The results show that training with real-world

guideline data continues to benefit from increased case volume, whereas further gains from training with fictional guideline data depend more strongly on broader guideline coverage (Section 4.3).

RQ3. *Does guideline-adherence training introduce context-memory conflict?*

We conduct a comprehensive set of cross-regional real-world and synthetic intervention experiments to test whether internalized guideline knowledge interferes with reasoning. The results show consistent performance degradation when training-time and test-time decision logic diverge, providing evidence of context-memory conflict in guideline-adherence training (Sections 4.4 and 4.5).

2 Related Work

2.1 Clinical Decision-Making

Recent work has moved beyond medical question answering Jin et al. (2021); Pal et al. (2022); Jin et al. (2019) toward more realistic clinical decision-making, including diagnostic reasoning Qiu et al. (2025); Gao et al. (2023), multi-stage decision processes Wu et al. (2024); Xie et al. (2026), and interactive information acquisition Werthaim et al. (2026); Liu et al. (2025). A growing subset of these studies further incorporates clinical guidelines as explicit references for assessing whether model decisions follow guideline-defined recommendations and clinical decision pathways. Representative benchmarks Hager et al. (2024); Tan et al. (2026); Li et al. (2025) have highlighted the importance of evaluating not only final outcomes but also adherence to the intermediate decision process. In particular, MEGA-CDP Chen et al. (2026) formulates guideline adherence as a pathway-level evaluation problem and evaluates pathway consistency using classical sequence-alignment algorithms. In contrast to prior work centered primarily on evaluation, we investigate whether guideline adherence can be learned through training and generalized to unseen real-world guidelines.

2.2 Synthetic Data Generation

Recent studies have shown that language models may continue to rely on parametric knowledge when it conflicts with information provided in context, raising concerns about their ability to follow newly supplied rules or evidence Kim et al. (2026); Wang et al. (2025); Sun et al. (2026); Zhao et al. (2025). This issue is particularly relevant to clinical guideline adherence, where decision rules may differ across regions, institutions, and guideline versions, and guideline-specific knowledge internalized during training may interfere with the guideline provided at inference time Harrison et al. (2010); Xu et al. (2024). This motivates synthetic data as a scalable way to diversify training beyond limited real-world rules, as explored in instruction tuning Wang et al. (2023); Li et al. (2024), reasoning Lee et al. (2025); Yu et al. (2025), and domain adaptation Corbeil et al. (2025); Kweon et al. (2024). In the medical domain, prior work has used synthetic data to generate clinical questions Kell et al. (2025); Bai et al. (2024), patient cases Bakkum et al. (2024); Berbenyuk et al. (2024), and reasoning trajectories Wu et al. (2025); Sun et al. (2025). These studies demonstrate the value of synthetic supervision for medical training, but its role in improving guideline adherence remains unexplored. Different from prior medical synthetic data efforts, we study synthetic supervision for guideline adherence by constructing fictional reasoning trees, corresponding guideline documents, reference CDPs, and case vignettes that share the same decision logic.

3 SAGA-CDP

A key challenge in synthesizing guideline-adherence training data is to maintain consistency among the disease concept, decision logic, and generated components while preserving the structure and presentation of real-world clinical guidelines. As illustrated in Figure 1, we introduce SAGA-CDP, a unified pipeline for constructing such data.

3.1 Fictional Disease Initialization

The pipeline begins by initializing a diverse collection of fictional disease concepts. As shown in Stage I of Figure 1, we use GPT-5.4 OpenAI (2026) to generate 1,000 fictional diseases, each represented by a disease name and a brief description. To reduce conceptual redundancy and promote

diversity in narrative settings, we generate the diseases across five categories, comprising *Fantasy*, *Science Fiction*, *Xuanhuan*, *Famous Fictional Works*, and *Realistic Medical Variants*, with 200 diseases in each category. All generated disease concepts are fictional and do not correspond to existing real-world disease entities. Detailed descriptions and examples of these categories are provided in Appendix B. For each of the 1,000 concepts, we create corresponding English and Chinese versions, yielding 2,000 disease entries in total. These entries anchor all subsequent stages of synthetic data generation.

3.2 Synthetic Data Generation

To match the training data formulation derived from real-world clinical guidelines, each fictional disease concept must be further instantiated into a clinical guideline, a collection of reference CDPs, and their corresponding case vignettes. Generating these components independently, however, makes it difficult to control their decision logic and may introduce inconsistencies across different parts of the same instance. We therefore introduce an explicit reasoning tree as an intermediate representation that organizes the disease-specific decision logic and constrains subsequent synthesis, as illustrated in Stage II of Figure 1.

Constructing the reasoning tree is nontrivial because fictional diseases do not come with pre-existing clinical rules that can be directly inherited from real-world medicine. The tree must achieve a careful balance between fictional content and plausible clinical reasoning, establishing decision conditions that remain compatible with the internal logic of the fictional setting and retain the rigor expected of clinical decision-making. During pipeline development, we iteratively refined the generation prompts by inspecting sampled trees for quality and internal consistency. Each generated tree defines a self-consistent set of decision conditions, branch transitions, and terminal outcomes, with its root-to-leaf paths extracted as reference CDPs.

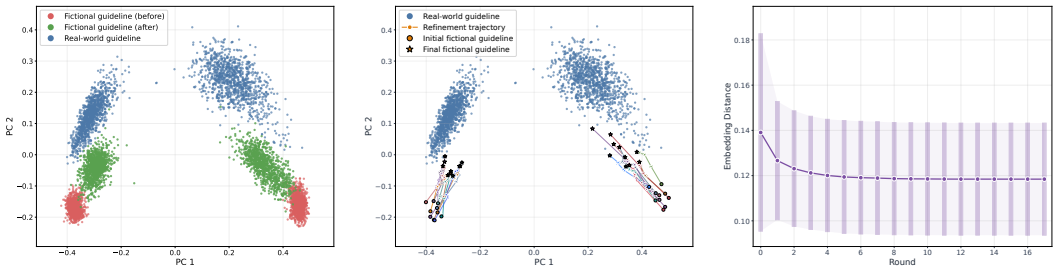
Unlike the explicit structure of the reasoning tree, the corresponding guideline must express the same decision logic as a coherent clinical document. Using GPT-5.4, we transform each reasoning tree into a complete guideline, organizing its narrative, recommendations, and decision conditions around the structured reasoning process. This avoids reducing the guideline to a direct node-by-node transcription of the tree while maintaining a coherent representation of the intended decision process. We then construct case vignettes from the corresponding reference CDPs, following the case construction procedure of MEGA-CDP. Each vignette contains the information required to follow the designated CDP and reach its final outcome, yielding a total of 76,584 cases.

3.3 Agentic Refinement

Although the synthetic generation process produces complete guideline documents from structured reasoning trees, the generated guidelines may still differ from real-world clinical guidelines in formatting, writing style, and presentation. These differences can introduce superficial cues that are specific to the synthetic generation process rather than to guideline adherence itself. As shown in Stage III of Figure 1, we therefore introduce an agentic refinement stage to improve the overall quality of fictional guidelines and reduce superficial differences from real-world clinical documents.

The refinement workflow begins with iterative review and rewriting. At each round, the reviewer examines the current guideline against real-world reference guidelines and the reasoning tree, producing a revision checklist that identifies discrepancies in guideline presentation and style, as well as any part of the reasoning tree whose decision logic is not inferable from the guideline text. The rewriter updates the guideline according to this checklist, and the revised document is passed to the next round. As an empirical stopping criterion, guideline representations are encoded with Qwen3-Embedding-8B Zhang et al. (2025), and their cosine distances to the real-world and initial fictional guideline centroids are used to monitor refinement progress. The iteration stops once the current guideline is closer to the real-world centroid and falls within a predefined distance threshold. This criterion is kept separate from the revision process, as the embedding signal is not provided to the agents and does not guide how the guideline text is revised. It is used solely to determine when the iterative refinement stage should terminate.

Once the stopping condition is satisfied, we perform a separate completeness check with a validator that focuses exclusively on whether all decision logic in the reasoning tree remains inferable from



(a) Representation distribution (b) Refinement trajectories (c) Mean embedding distance to the real-world guideline centroid

Figure 2: **Effect of agentic refinement** on reducing the representational gap between fictional and real-world guidelines.

the guideline text. Unlike the reviewer, the validator does not assess presentation or style. Any missing or insufficiently expressed logic is compiled into a revision checklist and returned to the rewriter for correction. This validation and revision process continues until the guideline passes the completeness check, after which it is retained as the final output.

Qwen3.5-122B-A10B Qwen Team (2026) is used as the backbone model for all agents in the refinement workflow. We analyze how guideline representations evolve throughout this process in Figure 2. Figure 2a presents a PCA visualization of the guideline embeddings and shows that fictional guidelines exhibit a noticeable distributional shift after refinement, with their representations moving closer to those of real-world guidelines. Figure 2b further provides a finer-grained view through sampled refinement trajectories, illustrating how individual fictional guidelines progressively move through the embedding space across successive rounds. Consistent with these observations, Figure 2c shows the mean cosine distance to the real-world guideline centroid with one standard deviation across guidelines, and the mean gradually decreases as refinement proceeds. These results provide empirical evidence that the refinement process reduces the representational gap between fictional and real-world guidelines.

4 Experiments

4.1 Experimental Setup

Data Preparation. For real-world guideline data, we used the original MEGA-CDP test set and constructed the training and validation sets from the remaining guidelines. To avoid information leakage, we further excluded guidelines corresponding to the same diseases as those appearing in the test set. To maintain a manageable computational budget, we retained guidelines whose lengths were estimated to be below 16K tokens using the Qwen tokenizer. We then split the filtered data into a training set of 452 guidelines with 8,458 cases and a validation set of 100 guidelines with 1,761 cases. For fictional guideline data, we used the full set of outputs generated by SAGA-CDP as the training candidate pool, consisting of 2,000 guidelines and 76,584 corresponding cases.

Evaluation Setup. We used Qwen3.5-9B Qwen Team (2026) as the base model for fine-tuning and compared the resulting models against 16 representative comparison LLMs. Detailed training configurations are provided in the Appendix D. We evaluate models in the multi-turn setting of MEGA-CDP using Qwen3-8B Qwen Team (2025) as the judge model. Our primary evaluation metrics are DTW-based pathway consistency cost (PCC) and outcome accuracy (Acc), which respectively measure pathway-level adherence to the reference CDP and the correctness of the final outcome. We additionally report success rate (SR) as a complementary metric to assess whether the model follows the required interaction format and successfully completes the multi-turn dialogue. Lower PCC indicates better pathway consistency, while higher values are better for Acc and SR.

4.2 Performance Comparison

In this section, we aim to answer **RQ1**: How effective are real-world and fictional guideline data for guideline-adherence training? We fine-tuned the base model using real-world guideline data, fictional guideline data, and their combination, and compared the resulting models with the baseline

Table 1: **Performance comparison** on MEGA-CDP test set. Arrows indicate whether higher (↑) or lower (↓) values are better. Best results among **fine-tuned** models and comparison models are indicated by **bold** and underline, respectively.

Model	PCC (%) ↓	Acc (%) ↑	SR (%) ↑
HuatuoGPT-o1-7B Chen et al. (2024)	91.44	4.41	24.05
HuatuoGPT-o1-72B Chen et al. (2024)	63.12	17.45	74.68
Baichuan-M3-235B Baichuan-M3 Team (2026)	54.12	35.58	98.87
Qwen3-8B Qwen Team (2025)	53.63	27.55	97.65
Qwen3-32B Qwen Team (2025)	47.91	33.24	99.24
Qwen3-235B-A22B Qwen Team (2025)	54.69	28.07	95.44
Qwen3-30B-A3B-Instruct-2507 Qwen Team (2025)	48.81	17.30	97.44
Qwen3-30B-A3B-Thinking-2507 Qwen Team (2025)	65.50	22.41	76.21
Qwen3-Next-80B-A3B-Instruct Qwen Team (2025)	47.30	22.37	98.71
Qwen3-Next-80B-A3B-Thinking Qwen Team (2025)	59.45	25.59	80.01
DeepSeek-R1-Distill-Qwen-32B DeepSeek-AI (2025)	78.80	11.89	58.69
gpt-oss-120b OpenAI (2025)	56.81	35.00	91.85
LLaMA-3.3-70B-Instruct Grattafiori et al. (2024)	46.84	26.23	99.77
Gemini-3.1-Pro-Preview Google (2026)	58.93	29.19	78.08
GPT-5.4 OpenAI (2026)	61.42	24.14	74.63
Claude-Sonnet-4.6 Anthropic (2026)	46.15	44.38	99.14
Qwen3.5-9B Fine-Tuning			
Baseline Qwen Team (2026)	53.82	24.77	89.96
Fine-Tuned on Real-World Guidelines	40.48	51.34	99.81
Fine-Tuned on Fictional Guidelines	43.65	46.16	**99.97**
Fine-Tuned on Real-World + Fictional Guidelines	**39.55**	**52.64**	99.65

Table 2: **Data scaling results** for training with real-world, fictional, and mixed guideline data. Optimal/suboptimal results are indicated by **bold**/underline, respectively.

Scale	Partial Real-World (8,458 cases)		Partial Fictional (76,584 cases)		Full Real-World (8,458 cases) + Partial Fictional (76,584 cases)	
	PCC (%)	Acc (%)	PCC (%)	Acc (%)	PCC (%)	Acc (%)
1%	52.36	20.11	56.94	13.08	39.56	50.06
3%	51.54	17.14	54.43	32.49	**39.45**	51.14
5%	51.31	17.55	48.51	40.76	39.98	**51.15**
10%	51.44	26.04	46.60	45.12	40.31	49.17
20%	48.53	41.17	**45.07**	44.60	40.00	50.41
30%	45.66	43.92	46.05	44.41	41.09	49.01
40%	45.80	46.56	46.15	43.19	43.80	48.46
50%	43.70	47.01	46.28	45.06	40.50	48.15
60%	43.41	47.35	46.33	**45.29**	42.30	47.05
70%	42.95	47.89	48.07	43.84	42.16	47.42
80%	41.26	50.36	47.40	44.95	42.52	47.46
90%	41.03	49.91	50.69	43.39	42.99	46.94
100%	**40.72**	**51.12**	49.40	43.48	42.85	47.33

and 16 additional LLMs, with the results reported in Table 1. The training scales and mixing ratios used for these fine-tuned models were selected based on the validation results in Section 4.3.

Fine-Tuned on Real-World Guidelines. Training on real-world guideline data substantially improved performance, reducing PCC from 53.82% to 40.48% and increasing Acc from 24.77% to 51.34%. This demonstrates that real-world guideline data can effectively strengthen guideline-adherent decision-making at both the pathway and outcome levels.

Fine-Tuned on Fictional Guidelines. Training on fictional guideline data also produced substantial gains, reducing PCC to 43.65% and increasing Acc to 46.16%. Despite being trained on decision processes constructed from entirely fictional clinical concepts, the resulting model outperformed all evaluated comparison models on both PCC and Acc. It provides evidence that guideline-adherence capability learned from fictional guideline data can transfer to unseen real-world clinical guidelines.

Fine-Tuned on Mixed Guidelines. Combining real-world and fictional guideline data achieved the strongest overall performance, with a PCC of 39.55% and an Acc of 52.64%. The result indicated

Table 3: **Effect of guideline diversity** under fixed training case budgets.

Guideline Coverage	Real-World PCC (%)	Real-World Acc (%)	Fictional PCC (%)	Fictional Acc (%)
20%	49.25	39.70	53.55	38.77
40%	48.97	39.49	48.09	42.45
60%	48.71	41.05	47.88	42.40
80%	48.82	41.08	47.01	43.11
100%	**48.53**	**41.17**	**45.07**	**44.60**

that fictional guideline data could complement real-world guideline data and provide additional gains, supporting its role as an augmentation source for guideline adherence training.

4.3 Data Scaling Analysis

In this section, we aim to answer **RQ2**: How does performance scale with the amount and diversity of guideline training data? We conduct scaling experiments across multiple training scales for each data source and further examine the effect of guideline diversity under fixed case budgets. All analyses in this section are conducted on the validation set. For each scale, we sampled cases evenly across the available guidelines.

Data Scaling Behavior. According to Table 2, real-world guideline data exhibited a clear scaling benefit, with PCC generally decreasing from 52.36% to 40.72% at the full scale and Acc increasing to 51.12%. Training with fictional guideline data achieved its lowest PCC at 20%, corresponding to approximately 15.3K cases, after which performance no longer improved with additional cases. Figure 3 further indicates that, at this scale, the best performance achieved by training with fictional guideline data was comparable to that obtained with approximately 3K-4K real-world training cases. We further evaluated fictional data as an augmentation to the full real-world training set. Adding 3% of the fictional data, corresponding to 2.3K cases combined with 8.5K real-world cases, achieved the lowest PCC of 39.45%, corresponding to an approximately 4:1 ratio of real-world to fictional cases.

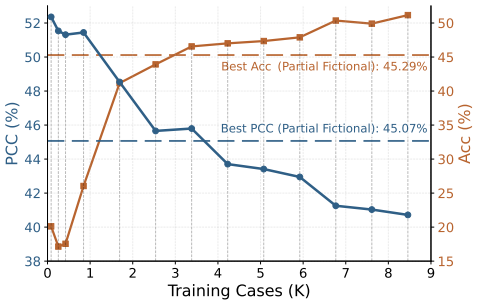
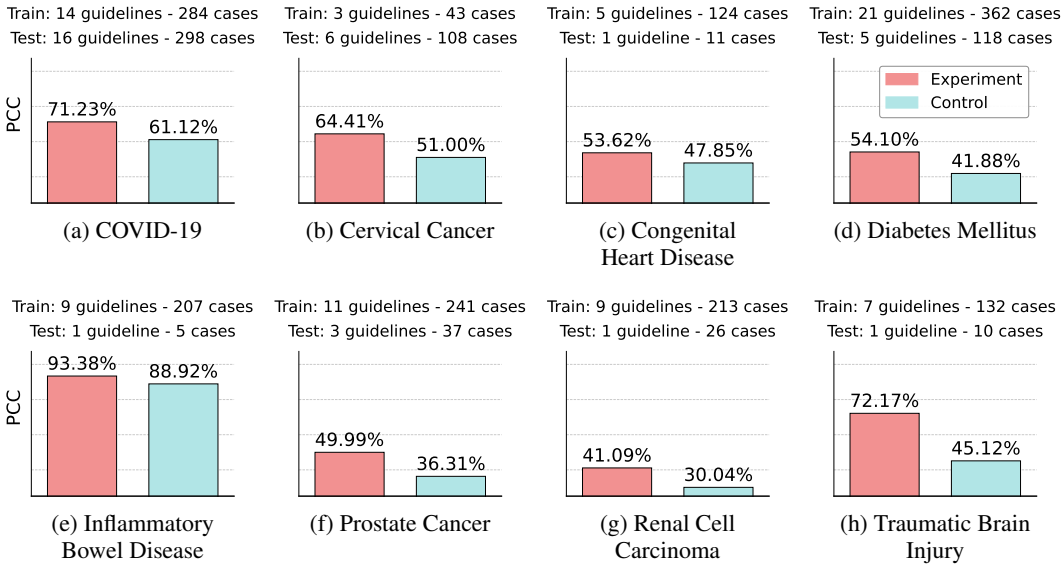


Figure 3: **Scaling behavior** across different guideline data sources.

Guideline Diversity in Scaling. At larger scales, validation performance with fictional guideline data gradually declined. We hypothesized that this degradation may arise from insufficient guideline diversity relative to the increasing number of generated cases, potentially leading to overfitting. To examine this hypothesis, we conducted controlled guideline coverage experiments for both fictional and real-world data. For each data source, we fixed the number of training cases at the 20% scale and varied the proportion of available guidelines while keeping the case budget unchanged. Guideline subsets were constructed cumulatively, such that each higher-coverage setting contained all guidelines from the preceding lower-coverage setting, thereby ensuring a strictly increasing level of guideline diversity. As shown in Table 3, broader guideline coverage substantially improved performance when training on fictional guideline data, with PCC decreasing from 53.55% to 45.07% and Acc increasing from 38.77% to 44.60%. In comparison, performance with real-world guideline data changed modestly, with PCC decreasing from 49.25% to 48.53% and Acc increasing from 39.70% to 41.17%. Given the more limited scale of the real-world guideline pool, the effect of guideline diversity is less pronounced in this setting, suggesting that training case volume remains the primary scaling factor for real-world data, while further scaling of fictional data may depend more strongly on increasing guideline diversity.

4.4 Cross-Regional Real-World Guideline Conflict

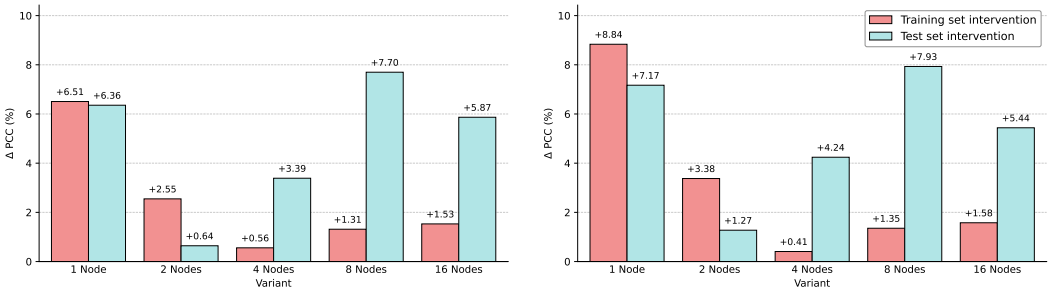
In the following two sections, we aim to answer **RQ3**: Does guideline-adherence training introduce context-memory conflict? A direct way is to evaluate models when the guidelines used for training



(a) COVID-19 (b) Cervical Cancer (c) Congenital Heart Disease (d) Diabetes Mellitus

(e) Inflammatory Bowel Disease (f) Prostate Cancer (g) Renal Cell Carcinoma (h) Traumatic Brain Injury

Figure 4: **PCC comparison** in cross-regional real-world guideline conflict experiments.



(a) Overall PCC degradation (b) PCC degradation on affected CDPs

Figure 5: **PCC degradation under controlled synthetic guideline conflicts** at the overall and affected-CDP levels. ΔPCC denotes the PCC difference between each conflict variant and the unchanged condition, where larger values indicate greater degradation.

and evaluation may differ in their decision rules. We therefore grouped guidelines for the same disease by region, using guidelines from one region for training and those from another region for evaluation. As a control, we sampled training cases from different disease groups with a matched training data scale.

Figure 4 shows performance degradation across eight disease groups, including COVID-19, cervical cancer, congenital heart disease, diabetes mellitus, inflammatory bowel disease, prostate cancer, renal cell carcinoma, and traumatic brain injury. Across all eight groups, PCC was higher in the cross-regional experimental setting than in the matched control setting. This pattern suggests that prior training on guidelines from one region can reduce adherence to guidelines for the same disease from another region, providing real-world evidence of context-memory conflict.

4.5 Controlled Synthetic Guideline Conflict

The cross-regional analysis provides evidence of context-memory conflict under real-world guideline variation, but regional guidelines may differ along many uncontrolled dimensions. To isolate the effect of conflicting decision rules directly, we conducted a controlled synthetic intervention study. Starting from a reasoning tree, we constructed an unchanged version and several conflict variants by modifying different numbers of decision nodes. The interventions were cumulative, with each

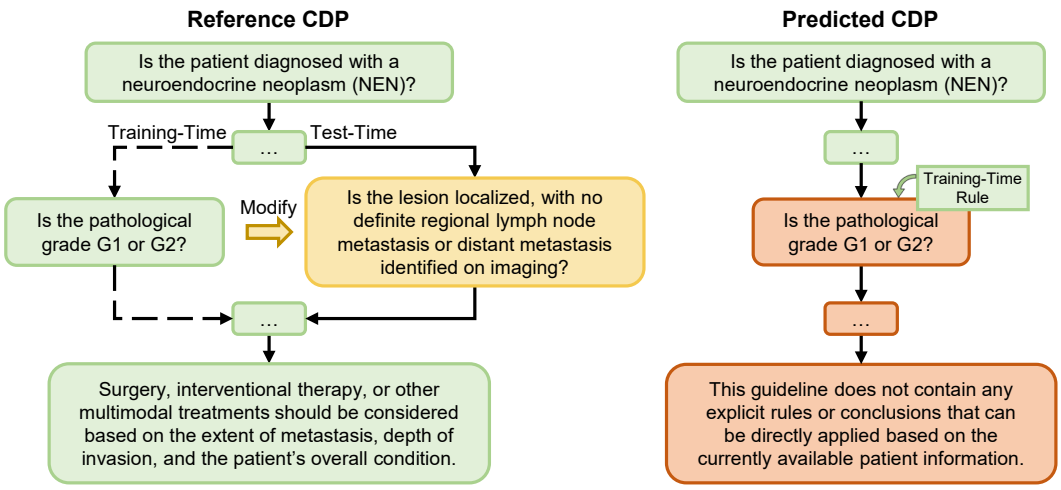


Figure 6: **Qualitative example** of context-memory conflict in a synthetic intervention.

larger variant retaining the modifications introduced in the preceding variants. Using five real-world guidelines as generation references, we generated five guidelines for the unchanged tree and each conflict variant. We then generated another five guidelines from the unchanged tree using five different real-world reference guidelines. This design enabled two complementary intervention settings. In the training set intervention, separate models were trained on each condition from the first set and evaluated on the second set of unchanged guidelines. In the test set intervention, the model was trained on the second set of unchanged guidelines and evaluated on the first set.

Performance Degradation under Controlled Conflict. As shown in Figure 5a, the conflict variants produced varying degrees of PCC degradation under both training set and test set interventions. Since each intervention affected only a subset of CDPs, we further evaluated the PCC change exclusively on the affected CDPs. Figure 5b shows consistent degradation across the evaluated variants in both settings, providing more direct evidence that the introduced conflicts impaired adherence to the modified decision logic.

Qualitative Evidence of Context-Memory Conflict. Figure 6 presents a qualitative example of this conflict. In this case, the test-time guideline replaced the training-time decision rule on pathological grade with a different criterion based on lesion localization and metastasis status. However, the predicted CDP continued to query the previously learned rule, resulting in a pathway that was no longer supported by the provided guideline. This example illustrates how internalized guideline-specific knowledge can steer the predicted CDP toward the training-time rule, consistent with the degradation observed in the controlled intervention experiments. More broadly, these results highlight a potential risk of real-world guideline data, as internalized guideline-specific knowledge may become misaligned with the decision logic required in different clinical settings or updated guideline versions.

5 Conclusion

In this work, we introduced SAGA-CDP to generate synthetic guideline data and systematically studied guideline-adherence training using real-world and fictional guideline data. Our experiments show that both real-world and fictional guideline data substantially improve guideline adherence, with their combination yielding further gains. Scaling analyses reveal different bottlenecks for the two data sources. Real-world guideline data continue to benefit from increasing case volume, whereas fictional guideline data eventually plateau as the training set grows, with further gains depending more on broader guideline coverage. Cross-regional and controlled synthetic experiments further show that prior guideline-adherence training can impair adherence when newly provided decision rules conflict with those learned during training, indicating context-memory conflict. Overall, our results support a complementary role for the two data sources in which real-world guideline data serve as the clinical foundation, while fictional guideline data provide scalable augmentation for expanding the training set.

AI USE STATEMENT

We clarify that LLMs were used solely for polishing writing and synthetic data generation. All research design, experiments, analyses, and conclusions were conducted by the authors.

REFERENCES

Anthropic. System Card: Claude Sonnet 4.6. https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf, 2026. Accessed: 2026-05-19.

Fan Bai, Keith Harrigian, Joel Stremmel, Hamid Hassanzadeh, Ardavan Saeedi, and Mark Dredze. Give me Some Hard Questions: Synthetic Data Generation for Clinical QA, 2024. URL https://arxiv.org/abs/2412.04573.

Baichuan-M3 Team. Baichuan-M3: Modeling Clinical Inquiry for Reliable Medical Decision-Making. 2026. URL https://arxiv.org/abs/2602.06570.

Michiel J Bakkum, Mariëlle G Hartjes, Joost D Piët, Erik M Donker, Robert Likic, Emilio Sanz, Fabrizio De Ponti, Petra Verdonk, Milan C Richir, Michiel A Van Agtmael, et al. Using Artificial Intelligence to Create Diverse and Inclusive Medical Case Vignettes for Education. British Journal of Clinical Pharmacology, 90(3):640–648, 2024.

Anna Berbenyuk, Leigh Powell, and Nabil Zary. Feasibility and Educational Value of Clinical Cases Generated Using Large Language Models. In Digital Health and Informatics Innovations for Sustainable Health Care Systems: Proceedings of MIE 2024, pp. 1524–1528. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs, 2024. URL https://arxiv.org/abs/2412.18925.

Nuo Chen, Xinyang Jiang, Zilong Wang, Zhifei Zhang, Xiaoye Qu, Jiajun Deng, Yulan Guo, and Cairong Zhao. Benchmarking Clinical Decision Pathway Adherence in Large Language Models, 2026. URL https://arxiv.org/abs/2608.26592.

Jean-Philippe Corbeil, Amin Dada, Jean-Michel Attendu, Asma Ben Abacha, Alessandro Sordoni, Lucas Caccia, François Beaulieu, Thomas Lin, Jens Kleesiek, and Paul Vozila. A Modular Approach for Clinical SLMs Driven by Synthetic Data with Pre-Instruction Tuning, Model Merging, and Clinical-Tasks Alignment. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 19352–19374, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL https://arxiv.org/abs/2501.12948.

Ruiqi Deng, Geoffrey Martin, Tony Wang, Gongbo Zhang, Yi Liu, Chunhua Weng, Yanshan Wang, Justin F Rousseau, and Yifan Peng. CPGPrompt: Translating Clinical Guidelines into Large Language Model-Executable Decision Support. Journal of the American Medical Informatics Association, 33(4):855–862, 2026.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. DR.BENCH: Diagnostic Reasoning Benchmark for Clinical Natural Language Processing. Journal of Biomedical Informatics, 138:104286, 2023.

Google. Gemini 3.1 Pro Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf, 2026. Accessed: 2026-05-19.

Brent Graham. Clinical Practice Guidelines: What Are They and How Should They Be Disseminated? Hand Clinics, 30(3):361–365, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making. Nature Medicine, 30(9):2613–2622, 2024.

Margaret B Harrison, France Légaré, Ian D Graham, and Béatrice Fervers. Adapting Clinical Practice Guidelines to Local Context and Assessing Barriers to Their Use. Canadian Medical Association Journal, 182(2):E78–E84, 2010.

Changhao Jiang, Ming Zhang, Yifei Cao, Junjie Ye, Xiaoran Fan, Shihan Dou, Zhiheng Xi, Jiajun Sun, Yi Dong, Yujiong Shen, et al. Beyond Scaling: Measuring and Predicting the Upper Bound of Knowledge Retention in Language Model Pre-Training. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 26377–26396, 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences, 11(14):6421, 2021.

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.

Gregory Kell, Angus Roberts, Serge Umansky, Yuti Khare, Najma Ahmed, Nikhil Patel, Chloe Simela, Jack Coumbe, Julian Rozario, Ryan-Rhys Griffiths, et al. RealMedQA: A Pilot Biomedical Question Answering Dataset Containing Realistic Clinical Questions. In AMIA Annual Symposium Proceedings, volume 2024, pp. 590, 2025.

Minsung Kim, Dong-Kyum Kim, Jea Kwon, Nakyeong Yang, Kyomin Jung, and Meeyoung Cha. How Training Data Shapes the Use of Parametric and In-Context Knowledge in Language Models. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 23242–23257, 2026.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes. In Findings of the Association for Computational Linguistics: ACL 2024, pp. 5148–5168, 2024.

Sung-Min Lee, Siyoon Lee, Juyeon Kim, and Kyoungmin Roh. ControlMed: Adding Reasoning Control to Medical Language Model. 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, pp. 2786–2799, 2025.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. Synthetic Data (Almost) from Scratch: Generalized Instruction Tuning for Language Models, 2024. URL https://arxiv.org/abs/2402.13064.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. MedGUIDE: Benchmarking Clinical Decision-Making in Large Language Models, 2025. URL https://arxiv.org/abs/2505.11613.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. Large Language Models in the Clinic: A Comprehensive Benchmark, 2024. URL https://arxiv.org/abs/2405.00716.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. Interactive Evaluation for Medical LLMs via Task-Oriented Dialogue System. In Proceedings of the 31st International Conference on Computational Linguistics, pp. 4871–4896, 2025.

David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. Enhancing Large Language Models for Clinical Decision Support by Incorporating Clinical Practice Guidelines. In 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), pp. 694–702. IEEE, 2024.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card, 2025. URL https://arxiv.org/abs/2508.10925.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026. Accessed: 2026-05-19.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A Large-Scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering. In Proceedings of the Conference on Health, Inference, and Learning, pp. 248–260. PMLR, 2022.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, et al. Quantifying the Reasoning Abilities of LLMs on Clinical Cases. Nature Communications, 16(1):9799, 2025.

Qwen Team. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Zifan Song, Kaitao Song, Guosheng Hu, Ding Qi, Junyao Gao, Xiaohua Wang, Dongsheng Li, and Cairong Zhao. Trade in Minutes! Rationality-Driven Agentic System for Quantitative Financial Trading. In International Conference on Learning Representations, volume 2026, pp. 27140–27156, 2026.

Kaiser Sun, Fan Bai, and Mark Dredze. Task Matters: Knowledge Requirements Shape LLM Responses to Context-Memory Conflict. In Findings of the Association for Computational Linguistics: ACL 2026, pp. 4154–4176, 2026.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, et al. ReasonMed: A 370K Multi-Agent Generated Dataset for Advancing Medical Reasoning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 26457–26478, 2025.

Andong Tan, Shuyu Dai, Jinglu Wang, Fengtao Zhou, Yan Lu, Xi Wang, Yingcong Chen, Can Yang, Shujie Liu, and Hao Chen. A Decade-Scale Benchmark Evaluating LLMs' Clinical Practice Guidelines Detection and Adherence in Multi-Turn Conversations, 2026. URL https://arxiv.org/abs/2603.25196.

Yuechen Tao, Xiushi Lin, He Sun, Wenrui Bao, Xiaobo Zhu, Shiwei Wu, Yixiao Liu, Ziang Li, Jialong Wu, Zhenkun Zhao, et al. Augmenting Clinical Decision-Making with Large Language Models: Evaluation across General and Specialty Tasks. Journal of the American Medical Informatics Association, pp. ocag053, 2026.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. AdaCAD: Adaptively Decoding to Balance Conflicts between Contextual and Parametric Knowledge. 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. 11636–11652, 2025.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), pp. 13484–13508, 2023.

Mai Werthaim, Maya Kimhi, Alexander Apartsin, and Yehudit Aperstein. A Benchmark for Evaluating Diagnostic Questioning Efficiency of LLMs in Patient Conversations. Scientific Reports, 16(1):6121, 2026.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs, 2025. URL https://arxiv.org/abs/2504.00993.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, et al. MedJourney: Benchmark and Evaluation of Large Language Models over Patient Clinical Journey. Advances in Neural Information Processing Systems, 37:87621–87646, 2024.

Yuzhang Xie, Keqi Han, Yunpeng Xiao, Hejie Cui, Guanchen Wu, Ziyang Zhang, Kai Shu, Jiaying Lu, Xiao Hu, and Carl Yang. EHRbench: An Automated and Reliable EHR-Based Benchmark for Clinical Decision Making with LLMs. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 10068–10079, 2026.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge Conflicts for LLMs: A Survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 8541–8565, 2024.

Ping Yu, Jack Lanchantin, Tianlu Wang, Weizhe Yuan, Olga Golovneva, Ilia Kulikov, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. CoT-Self-Instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks, 2025. URL https://arxiv.org/abs/2507.23751.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv preprint arXiv:2506.05176, 2025.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering. 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. 5117–5136, 2025.

A Task Formulation and Evaluation

We formulate each guideline-adherence instance as

$$x_i = (G_i, V_i, P_i, O_i),$$

where G_i denotes a clinical guideline, V_i the corresponding case vignette, P_i the reference clinical decision pathway (CDP), and O_i the final decision outcome. The reference CDP specifies the ordered sequence of guideline-defined decision steps leading to the final outcome and serves as the reference trajectory for pathway-level evaluation.

Multi-Turn Evaluation. We evaluate models in the multi-turn interactive setting of MEGA-CDP. The model is provided with the clinical guideline and dialogue history while interacting with a patient-side environment that responds according to the case vignette. At each turn, the model asks one question to acquire additional patient information. After the interaction terminates, the sequence of generated questions forms the predicted CDP

$$\hat{P}_i = (\hat{p}_{i,1}, \hat{p}_{i,2}, \ldots, \hat{p}_{i,n}),$$

which is compared with the reference CDP

$$P_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,m}),$$

where m and n denote the numbers of steps in the reference and predicted CDPs, respectively.

Pathway Consistency Cost. We use a Qwen-based judge model Qwen Team (2025) to measure step-level consistency between reference and predicted CDPs. For a reference CDP P_i with m steps and a predicted CDP $\hat{P}_i$ with n steps, we construct a consistency matrix

$$C_i \in [0,1]^{m \times n},$$

where

$$C_{i,a,b} = \mathcal{J}(p_{i,a}, \hat{p}_{i,b}), \quad 1 \le a \le m,\ 1 \le b \le n.$$

Here, $\mathcal{J}$ denotes the normalized scoring function implemented by the judge model, and $C_{i,a,b}$ measures the consistency between the a-th reference step and the b-th predicted step. We convert the consistency matrix into a distance matrix $D_i \in [0,1]^{m \times n}$ by

$$D_{i,a,b} = 1 - C_{i,a,b}.$$

We then align the reference and predicted CDPs using Dynamic Time Warping (DTW). Let $\mathcal{A}_i$ denote an admissible alignment path consisting of ordered index pairs (a, b) that satisfy the standard boundary, monotonicity, and continuity constraints. The optimal alignment is

$$\mathcal{A}_i^* = \arg\min_{\mathcal{A}_i} \sum_{(a,b)\in\mathcal{A}_i} D_{i,a,b}.$$

The Pathway Consistency Cost (PCC) is computed as

$$\mathrm{PCC}(P_i, \hat{P}_i) = \frac{1}{|\mathcal{A}_i^*|} \sum_{(a,b)\in\mathcal{A}_i^*} D_{i,a,b},$$

where $|\mathcal{A}_i^*|$ denotes the number of aligned step pairs along the optimal DTW path. Lower PCC indicates stronger consistency with the reference CDP.

B Fictional Disease Initialization Details

This section provides additional descriptions and representative examples of the five fictional disease categories used in the initialization stage, as detailed below:

Fantasy. Diseases in this category are inspired by magical, arcane, or mythological settings. Their causes may involve supernatural forces, enchanted substances, mythical creatures, curses, or other non-real phenomena.

Example. *Dragonfire Bronchial Scalding Disorder* describes chronic airway inflammation and soot-like bronchial plaques following inhalation of residual drake breath.

Science Fiction. Diseases in this category are based on speculative science and futuristic settings, including advanced technologies, artificial intelligence, space exploration, alien biology, and experimental scientific systems.

Example. *Quantum Nanite Autoimmunity Syndrome* describes a condition in which self-replicating maintenance nanites misclassify host proteins as foreign and trigger cascading tissue inflammation.

Xuanhuan. Diseases in this category are grounded in Chinese cultivation fantasy, with pathological concepts involving qi circulation, meridian disruption, internal cultivation, spiritual energy, soul instability, and related metaphysical systems.

Example. *Meridian Qi Reversal Syndrome* describes episodic organ pain and dizziness caused by abrupt reversal of qi circulation through the primary meridians.

Famous Fictional Works. Diseases in this category are derived from established fictional universes in films, television series, novels, and video games. Their biological, technological, or supernatural properties follow the internal rules of the corresponding fictional setting.

Example. *The T-Virus Infection Complex*, from the *Resident Evil* universe, describes an engineered pathogen associated with fever, necrosis, and severe mutagenic changes in advanced cases.

Realistic Medical Variants. Diseases in this category are entirely fictional but are designed to resemble authentic clinical terminology and disease descriptions. Their names follow common biomedical naming conventions by combining plausible anatomical locations, pathological processes, syndrome structures, and diagnostic phrasing.

Example. *Juvenile Corticostriatal Microangiopathic Degeneration* describes a childhood-onset disorder involving small-vessel injury in the basal ganglia, progressive dystonia, and executive dysfunction.

C Statistics of CDP Lengths

We analyze the length characteristics of the CDPs generated by SAGA-CDP. Table 4 compares the summary statistics of CDP lengths between SAGA-CDP and MEGA-CDP. The average CDP length in SAGA-CDP is 7.63 steps, compared with 6.51 steps in MEGA-CDP, while the corresponding medians are 7 and 6. Across most of the distribution, SAGA-CDP contains slightly longer pathways, as reflected by its higher P5, P25, median, P75, and P95 values. However, SAGA-CDP provides less coverage of the extreme long tail, with a maximum CDP length of 34 compared with 62 in MEGA-CDP. This suggests that the performance gains obtained from fictional guideline data are unlikely to be driven by enhanced coverage of unusually long CDPs.

Table 4: **Summary statistics** of CDP lengths in MEGA-CDP and SAGA-CDP.

Statistic	MEGA-CDP	SAGA-CDP
Mean	6.51	7.63
Std. Dev.	3.74	3.46
Min	1	1
P5	2	3
P25	4	5
Median	6	7
P75	8	9
P95	13	14
Max	62	34

Figure 7 further presents the CDP length distribution of SAGA-CDP for the full dataset and its English and Chinese subsets. The three distributions exhibit similar overall shapes, with most pathways concentrated around moderate lengths and progressively fewer examples at larger lengths. The English and Chinese subsets also follow closely aligned distributions, suggesting that the bilingual construction process produces comparable pathway-length characteristics across the two languages.

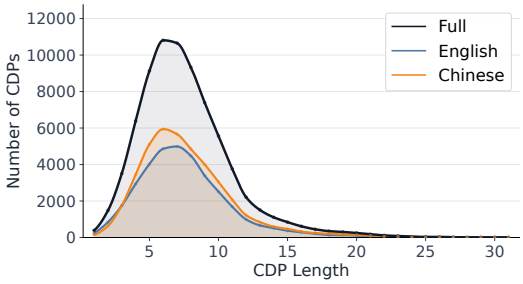


Figure 7: **Distribution** of reference CDP lengths.

Table 5: **Performance differences on the intersection of successful cases** from the MEGA-CDP test set. For each model, PCC and Acc are computed on the intersection of its successful test cases and those of the Qwen3.5-9B baseline. ΔPCC and ΔAcc denote differences relative to the baseline on the subset. Lower ΔPCC and higher ΔAcc indicate larger performance improvements relative to the baseline.

Model	Successful-Case Intersection	ΔPCC (%) ↓	ΔAcc (%) ↑
HuatuoGPT-o1-7B	1,616	+16.56	-8.33
HuatuoGPT-o1-72B	5,034	+2.32	-4.91
Baichuan-M3-235B	6,634	+4.47	+9.34
Qwen3-8B	6,575	+3.89	+1.66
Qwen3-32B	6,660	-1.28	+7.33
Qwen3-235B-A22B	6,402	+3.85	+2.74
Qwen3-30B-A3B-Instruct-2507	6,542	-1.39	-8.76
Qwen3-30B-A3B-Thinking-2507	5,275	+7.39	+1.69
Qwen3-Next-80B-A3B-Instruct	6,628	-2.44	-3.58
Qwen3-Next-80B-A3B-Thinking	5,505	+1.32	+4.38
DeepSeek-R1-Distill-Qwen-32B	3,875	+14.73	-2.70
gpt-oss-120b	6,229	+4.87	+10.86
LLaMA-3.3-70B-Instruct	6,685	-2.22	-0.27
Gemini-3.1-Pro-Preview	5,362	-0.08	+9.84
GPT-5.4	5,093	+1.05	+5.59
Claude-Sonnet-4.6	6,662	-3.04	+18.43
Qwen3.5-9B Fine-Tuning			
Fine-Tuned on Real-World Guidelines	6,689	-8.74	+24.67
Fine-Tuned on Fictional Guidelines	6,696	-5.29	+18.91
Fine-Tuned on Real-World + Fictional Guidelines	6,677	-9.78	+25.93

D Training Details

We fine-tuned Qwen3.5-9B using supervised fine-tuning with Low-Rank Adaptation (LoRA), with rank $r = 8$, scaling factor $\alpha = 16$, and dropout of 0.05. We used AdamW with a learning rate of 1×10^{-4}, weight decay of 0.01, and a warmup-decay scheduler with a warmup ratio of 0.03. All models were trained for one epoch with an effective batch size of 32 on 4 × NVIDIA H100 80GB GPUs using BF16 precision and DeepSpeed ZeRO-3. All experiments used a random seed of 42. Code is available at `https://anonymous.4open.science/r/SAGA-CDP`.

E Guideline-Adherence Performance on the Intersection of Successful Cases

To determine whether the improvements in Table 1 persist after accounting for differences in interaction success, we conduct an additional paired analysis on the intersection of successful cases. For each evaluated model, we identify the intersection between its successful test cases and those of the Qwen3.5-9B baseline, following the success definition used in SR. We then recompute PCC and Acc for both models on this shared subset and report the differences relative to the baseline in Table 5. This comparison removes cases in which either model fails to complete the required interaction, allowing the remaining differences to more directly reflect guideline-adherence performance.

The fine-tuned models retain substantial improvements under this comparison. Fine-tuning on real-world guidelines reduces PCC by 8.74 percentage points and increases Acc by 24.67 percentage points, while fine-tuning on fictional guidelines yields a 5.29-point reduction in PCC and an 18.91-point increase in Acc. Combining real-world and fictional guidelines produces the largest improvements, reducing PCC by 9.78 points and increasing Acc by 25.93 points. In contrast, the comparison models show more mixed changes across PCC and Acc. These results indicate that the gains from guideline-adherence training persist even when evaluation is restricted to cases successfully completed by both models, suggesting that the improvements cannot be explained solely by increased interaction success.

F PROMPT TEMPLATES

This section collects the English prompt templates used in the SAGA-CDP data generation process. Placeholders such as {`guideline`} denote instance-specific inputs. The complete templates are provided in Tables 6–12.

Table 6: Prompt template for fictional disease name generation.

Prompt 1: Fictional Disease Name Generation

You are generating a dataset of fictional disease names in English.

TARGET CATEGORY: {Category_Name}
{Category_Guidance}

TASK:
Generate exactly {Count} NEW items.

CONSTRAINTS:
- All disease names must be fictional. Do NOT output any real-world disease names.
- No duplicates. Do NOT repeat or closely resemble any previously generated name (including minor spelling changes, spacing, hyphenation, or only changing the suffix like Syndrome/Disorder/Virus).
- Each note must be exactly ONE sentence. Do not use semicolons.
- Names should sound medically plausible as a term, matching the category style.
- Follow the output format exactly:
<Index Number>. <Disease Name>: <One-sentence explanation>
- The output must not contain any parentheses or bracket characters of any kind.

PREVIOUSLY GENERATED NAMES (DO NOT REUSE OR CLOSELY RESEMBLE):
{Prev_Block}

Your Output:

Table 7: Prompt template for reasoning tree construction.

Prompt 2: Reasoning Tree Construction

You are an expert medical clinical guideline writer with extensive experience in evidence-based medicine and guideline development. Your task is to construct a clinical decision tree for a disease based on the information provided. You will receive a disease name, a short description of the disease, and a primary clinical focus. The disease should be treated as if it were a legitimate medical condition described in a professional clinical guideline. Do not invent a different disease or replace the provided name. Instead, expand the provided information into a structured clinical decision pathway centered on the specified primary clinical focus.

Although the disease and its related concepts may be fictional, the language and structure used in your reasoning should resemble those used in real medical clinical guidelines. If the provided disease description references or originates from an existing creative work such as literature, film, games, or other media, you may use relevant background information from that source to inform the medical reasoning and clinical descriptions. However, you must not explicitly mention the original work or indicate that the disease originates from a fictional narrative. All descriptions should be presented as if the disease were a real medical condition, and the tone should remain professional, clinical, and academically serious.

The primary clinical focus defines the central clinical objective addressed by the decision tree. It will be one of the following categories: diagnosis, screening, classification, or management. The overall structure of the decision tree should remain consistent with this focus, and all conclusions within the tree should correspond to it.

Each node in the decision tree must follow a consistent JSON structure containing the fields "type", "description", "yes", and "no". The "type" field must be either "condition" or "conclusion". The "description" field must contain a clear textual description corresponding to the node type. When the node type is "condition", the description must represent a specific clinical condition-check about patient-specific data or findings. When the node type is "conclusion", the description must represent the final outcome reached by the decision pathway.

When "type" is "condition", the node represents a clinical criterion evaluated using patient-specific information. The description should state the condition clearly and objectively, based on observable or measurable evidence such as symptoms, physical findings, laboratory results, imaging findings, physiological measurements, or clearly defined clinical events. Each condition node must branch into two outcomes labeled "yes" and "no" indicating whether the condition is satisfied.

Each condition node should evaluate only one independent clinical variable or criterion. Avoid combining multiple criteria into a single condition. If several criteria are required for a determination, represent them as a sequence of hierarchical condition nodes rather than a compound logical expression. Condition nodes must evaluate patient-level clinical evidence only. They must not ask whether the conclusion corresponding to the primary clinical focus has already been reached.

Avoid creating redundant or logically equivalent nodes. If a condition has already been fully determined by a previous branch, do not restate the complementary condition as a new node. The decision tree should remain logically consistent and free of unnecessary repetition.

Nodes must also avoid posing questions about clinical goals, tasks, or intentions. In particular, nodes must not ask about the clinical objective itself. The primary clinical focus defines the purpose of the tree but must never appear as a condition node.

When "type" is "conclusion", the node represents the final outcome reached by following the decision pathway. The description should clearly state the resulting clinical conclusion. The form of the conclusion must be consistent with the specified primary clinical focus. Conclusion nodes mark the end of a branch in the decision tree and therefore do not continue to additional nodes.

The decision tree should represent a coherent and clinically meaningful pathway with multiple hierarchical levels. As the pathway progresses, each level should refine the clinical evaluation by applying increasingly specific criteria that guide the decision process toward the final conclusions. Most branches of the decision tree should pass through several sequential condition nodes before reaching a conclusion node, reflecting a stepwise clinical reasoning process rather than ending immediately after a single condition.

The decision tree should contain around 40 nodes in total, and its maximum depth should reach approximately 10 levels, ensuring that the pathway captures a sufficiently detailed multi-step clinical reasoning process while remaining logically coherent.

Prompt 2: Reasoning Tree Construction (continued)

The final output must be a single valid JSON object containing only the complete decision tree. Do not include explanations, comments, or any additional text outside the JSON object.

Disease Name:
{Disease_Name}

Description:
{Disease_Description}

Primary Clinical Focus:
{Primary_Clinical_Focus}

Your Output:

Table 8: Prompt template for case vignette synthesis.

Prompt 3: Case Vignette Synthesis

You are a medical expert. I will provide you with a clinical decision path consisting of multiple steps. Each step includes a "condition" representing the clinical condition or decision criterion, and a "decision_outcome" indicating whether the condition is satisfied. In addition, I will specify the primary clinical focus of this path (e.g., diagnosis, classification, screening, or management). Your task is to convert this clinical decision path into a single, coherent, and natural-sounding clinical case description using medically styled language, appropriate to the specified primary clinical focus. The output should be written as a realistic clinical case vignette, resembling the narrative style commonly used in medical textbooks or case reports, rather than a summary or explanation of a decision pathway.

Note the following constraints:
1. The tone of each step's rewritten content must reflect its decision_outcome value. Use an affirmative sentence if the decision_outcome is "yes", and use a negative sentence if the decision_outcome is "no".
2. For any medical indicator involving a numerical range, you must provide a specific numeric value. The value must fall within the range if the decision_outcome indicates the condition is satisfied, and outside the range if the decision_outcome indicates the condition is not satisfied.
3. If the same medical indicator appears in multiple steps, ensure that the chosen numeric value is consistent and falls within the intersection of all applicable ranges.
4. All numeric values chosen for medical indicators must be clinically reasonable and plausible within real-world medical contexts.
5. If a step requires multiple conditions to be met simultaneously (logical AND), the description must fully and accurately reflect all such conditions.
6. If a step represents partially satisfied conditions (logical OR), you may select one or more logically valid conditions to include in the description, as long as they remain clinically consistent.
7. The description should integrate information from all steps naturally into a single narrative, without explicitly referencing decision steps, pathway logic, or the existence of a clinical pathway. Do not include step numbers or labels.
8. Write in the past tense as if describing a patient's clinical course.
9. Avoid gendered pronouns (e.g., "he", "she") when referring to the patient, unless the clinical condition explicitly involves sex-specific characteristics. If necessary, refer to the patient using neutral terms such as "the patient" or omit pronouns entirely.
10. The final sentence of the output must explicitly and neutrally state the primary clinical focus of this pathway (i.e., diagnosis, classification, screening, or management).
11. Do not use wording that implies definitive diagnosis, risk stratification, or treatment decisions beyond what is appropriate for the specified primary clinical focus. Do not include meta-commentary or self-referential statements (e.g., "this description reflects…", "this pathway focuses on…"), except for the required final sentence explicitly stating the primary clinical focus.
12. Do not add any extra information beyond what is provided in the clinical decision path. This includes avoiding the use of external medical knowledge, assumptions, or typical clinical patterns that are not explicitly represented in the input.
13. The clinical decision path may be derived from a fictional or simulated clinical guideline. You must treat all provided conditions, variables, and relationships as internally valid within this scenario, even if they do not correspond to real-world medical knowledge.
14. Do not attempt to correct, reinterpret, or replace any condition, variable, or threshold based on real-world medical knowledge. The vignette must strictly reflect the information provided in the clinical decision path.
15. Do not introduce any additional symptoms, signs, laboratory values, risk factors, or clinical interpretations that are not explicitly or implicitly specified in the input path.
16. If a clinical concept or measurement appears unusual, unfamiliar, or not aligned with real-world medicine, it should still be presented in a neutral, professional clinical tone without questioning its validity.
17. The disease context should be treated as a valid clinical entity. Do not indicate, imply, or suggest that the condition is fictional, hypothetical, or derived from non-medical sources.

Primary Clinical Focus:
{Primary_Clinical_Focus}

Clinical Decision Path:
{Path}

Your Output:

Table 9: Prompt template for fictional clinical guideline generation.

Prompt 4: Fictional Clinical Guideline Generation

You are an expert medical clinical guideline writer with extensive experience in evidence-based medicine and guideline development. Your task is to generate a comprehensive clinical guideline for a disease using the information provided. You will receive four inputs: a disease name, a short disease description, a primary clinical focus, and a decision tree in JSON format. Using this information, you must produce a complete guideline that describes the fictional disease as if it were a real and established medical condition.

Although the disease is fictional, the guideline must be written exactly in the style of a real clinical guideline. The language should be formal, academically rigorous, and consistent with documents produced by major medical organizations. If the disease description originates from a creative work such as literature, film, or games, you may use background knowledge from that source to inform the clinical and scientific content. However, the guideline must never mention the original work or indicate that the disease is fictional. All descriptions should treat the disease as a legitimate clinical entity.

The decision tree defines the core clinical decision framework for the disease under the specified primary clinical focus. This framework reflects how the condition is approached in clinical practice. When writing the guideline, the logic represented in the decision tree must be integrated naturally into the narrative. The guideline should read like a standard clinical document rather than a technical description of the decision tree. The guideline must not explicitly refer to the decision tree or describe any clinical condition or conclusion using tree-based language such as branches, nodes, or paths. The primary clinical focus is provided only as a writing constraint and must not be explicitly mentioned or explained in the guideline text. It may be reflected implicitly in the structure or section titles of the guideline, but it should not be highlighted, emphasized, or identified as the "primary clinical focus". Section titles should follow conventional clinical guideline structures rather than highlighting the concept of a "primary clinical focus".

The decision tree is provided in JSON format and follows a structured schema. Each node in the tree contains the following attributes:
- index: a unique identifier representing the position of the node within the clinical decision framework.
- type: specifies whether the node is a condition (non-leaf node) or a conclusion (leaf node).
- description: a concise statement describing the clinical condition, assessment criterion, or clinical conclusion represented by the node.

In addition, non-leaf nodes (i.e., nodes of type condition) include two further attributes:
- yes: the subsequent node to follow if the condition is satisfied.
- no: the subsequent node to follow if the condition is not satisfied.

Leaf nodes (i.e., nodes of type conclusion) do not contain further branching attributes and instead represent final clinical interpretations, classifications, or recommended actions.

Whenever the guideline text describes a clinical condition or conclusion corresponding to one or more nodes, the sentence containing that description must end with the annotation marker <Node:index>, where index must exactly match a real node index present in the provided decision tree. The model must use only indices that explicitly appear in the input JSON and must never invent, renumber, normalize, or infer new indices. All node indices must be treated as a fixed set of positive integers starting from 1 as defined in the input decision tree. The model must not generate indices outside this set, including zero or negative values. Every node index present in the decision tree must appear at least once somewhere in the final guideline text, but the same node marker should not be used excessively. Omission of any node is not allowed. These markers serve only as internal annotations indicating the correspondence between the decision tree and the guideline text and must remain embedded within the narrative without explanation or interpretation. Node annotations should appear naturally within the narrative, should not occur in long uninterrupted sequences, and should be distributed across the guideline in locations where the corresponding clinical conditions or conclusions are explained.

The guideline should be extensive and detailed, resembling a full professional guideline rather than a short article. The guideline should primarily use continuous narrative paragraphs to explain clinical concepts and recommendations. Bullet points or numbered lists may be used occasionally when summarizing criteria or key elements, but they should not dominate the document. The completed guideline must contain at least 8000 words. It should include substantial clinical discussion, such as background context, disease mechanisms, clinical presentation, and detailed discussion of clinical assessment and management approaches. The provided focus should guide the emphasis of the guideline, but the document should still read as a comprehensive clinical reference. To improve realism, the guideline may refer to hypothetical clinical studies, observational cohorts, expert consensus statements, or trial results. These references may be fictional but must be presented in a manner consistent with real medical literature.

Prompt 4: Fictional Clinical Guideline Generation (continued)

Before finalizing the guideline, perform an internal consistency and coverage verification against the supplied decision tree. The model must ensure that every node index from the input JSON has been explicitly cited at least once in the text. This includes all indices from the smallest (typically 1) through to the largest index. If any node index is missing, the model must revise the text and incorporate the missing nodes before producing the final output. The output must not be considered complete until full node coverage is achieved.

Your final output must contain only the completed guideline text. Do not output explanations, commentary, JSON structures, or process descriptions.

Disease Name:
{Disease_Name}

Description:
{Disease_Description}

Primary Clinical Focus:
{Primary_Clinical_Focus}

Decision Tree:
{Decision_Tree}

Your Output:

Table 10: Prompt template for agentic guideline review.

Prompt 5: Agentic Guideline Review

Role
You are a clinical guideline reviewer with expertise in evidence-based medicine and guideline development. Your responsibility is to evaluate synthetic guideline drafts and provide structured feedback for a subsequent revision step.

Objective
Your responsibility is to evaluate whether the synthetic clinical guideline can be revised to better match real-world clinical guideline characteristics while preserving the underlying clinical decision logic.

The guideline should satisfy two main objectives:

1. Improve Similarity to Real Clinical Guidelines
A real clinical guideline reference is provided as an example.

Evaluate whether the current synthetic guideline appropriately matches the reference guideline in terms of:
- document organization and section structure
- recommendation presentation style
- clinical question and evidence discussion format
- wording and tone of recommendations
- expression of uncertainty and evidence limitations
- use of tables, algorithms, implementation guidance, and supporting sections

The goal is not to copy the reference guideline content, but to identify structural and stylistic differences that prevent the synthetic guideline from resembling a professional clinical guideline.

2. Preserve the Underlying Clinical Decision Logic
The synthetic guideline is generated based on a clinical decision tree provided separately.

Evaluate whether the guideline correctly preserves the decision logic represented by the tree, including:
- diagnostic criteria
- eligibility conditions
- exclusion criteria
- contraindications
- alternative diagnoses
- management recommendations
- final clinical conclusions

The decision logic does not need to be explicitly written as a sequence of "if-then" rules.
Instead, the guideline should naturally contain enough clinical information and recommendations for a reader to infer the same decision process.

Identify cases where:
- important decision logic is missing
- the logic is incorrectly represented
- conclusions are changed
- the decision tree has overly influenced the document structure and made the guideline appear algorithmic

The purpose of this review is to provide actionable feedback that helps transform the synthetic guideline into a document that is both structurally similar to real-world clinical guidelines and logically consistent with the intended clinical decision framework.

Inputs
The workflow provides three inputs.

Real Guideline Reference
This contains one real-world clinical guideline provided as a reference example.

Use this guideline only to evaluate and learn:
- guideline document structure
- section organization
- recommendation presentation style

Prompt 5: Agentic Guideline Review (continued)

- evidence discussion patterns
- clinical writing conventions

<real_guideline_reference>
{real_guideline_reference}
</real_guideline_reference>

Decision Tree
This defines the intended clinical decision logic that the guideline should preserve.

Use it to evaluate whether the guideline preserves:
- decision conditions
- branches
- conclusions

<decision_tree>
{decision_tree}
</decision_tree>

Current Fictional Guideline
This is the guideline draft to evaluate.

<current_guideline>
{current_guideline}
</current_guideline>

Instructions
- Use the provided real guideline as a reference for evaluating guideline structure, writing style, and evidence presentation patterns.
- Do not evaluate whether the fictional guideline matches the medical content of the reference guideline.
- Use the decision tree as the source of truth for evaluating whether the clinical decision logic is preserved.
- Focus on identifying actionable issues that can be addressed during guideline revision.
- Distinguish between:
  - problems that affect guideline realism
  - problems that affect decision-tree fidelity
- When identifying decision logic problems, refer to the relevant decision tree node index when possible.
- Do not rewrite the guideline or provide replacement content. Only provide structured feedback for the revision step.

Output Format
Return a JSON object with the following structure:

```
{
  "guideline_realism_issues": [
    {
      "problem": "The identified issue affecting similarity to real-world clinical guidelines.",
      "evidence": "Where or how the issue appears in the current guideline.",
      "revision_guidance": "How the issue should be addressed during revision."
    }
  ],
  "decision_tree_fidelity_issues": [
    {
      "related_nodes": ["Relevant node indices if applicable"],
      "problem": "The identified issue related to decision logic preservation.",
      "evidence": "How the current guideline differs from the decision tree.",
      "revision_guidance": "How to preserve the logic while maintaining guideline style."
    }
  ]
}
```

Your Output:

Table 11: Prompt template for agentic guideline revision.

Prompt 6: Agentic Guideline Revision

Role
You are a clinical guideline editor with expertise in evidence-based medicine and guideline development.
Your responsibility is to revise a synthetic clinical guideline draft based on structured review feedback.
You perform revision, not complete regeneration.

Objective
Revise the current fictional guideline so that it better resembles a real-world clinical practice guideline while preserving the underlying clinical decision logic.

The revision should improve:
- guideline structure
- recommendation presentation
- evidence discussion style
- clinical writing style
- integration of decision logic
while maintaining the original fictional disease identity and clinical conclusions.

Inputs

Disease Name
<disease_name>
{disease_name}
</disease_name>

Disease Description
<disease_description>
{disease_description}
</disease_description>

Decision Tree
The decision tree defines the clinical decision logic that must be preserved.

<decision_tree>
{decision_tree}
</decision_tree>

Current Fictional Guideline
This is the guideline draft to be revised.

<current_guideline>
{current_guideline}
</current_guideline>

Review Report
The review report contains issues identified by the Guideline Review Agent.
Use it as the primary source for revision.

<review_report>
{review_report}
</review_report>

Instructions
- Revise the current guideline according to the review report.
- Preserve:
  - fictional disease identity
  - clinical setting
  - diagnostic logic
  - management recommendations
  - final conclusions
- Improve guideline realism by addressing issues identified in the review report.
- Preserve decision-tree logic without rewriting the guideline into an explicit decision tree or a sequence of "if-then" rules.

Prompt 6: Agentic Guideline Revision (continued)

- Integrate decision logic naturally through:
  - recommendations
  - clinical reasoning
  - diagnostic pathways
  - management sections
  - tables or algorithms when appropriate
- Make targeted revisions rather than rewriting unrelated sections.
- Do not introduce new clinical conclusions that are not supported by the original guideline and decision tree.
- Do not mention:
  - the review process
  - review feedback
  - embeddings
  - similarity metrics
  - synthetic data generation

Output Format
Output only the revised guideline in Markdown format.

Your Output:

Table 12: Prompt template for decision logic validation.

Prompt 7: Decision Logic Validation

Role
You are a clinical decision logic validator with expertise in medical reasoning and clinical guideline evaluation.
Your responsibility is to verify whether a clinical guideline preserves the decision logic represented by a decision tree.
You do not revise the guideline or provide suggestions for modification.

Objective
Map each decision tree node to the corresponding sentences in the clinical guideline.
The goal is to determine whether each decision tree node is represented in the guideline text.
A node is considered represented when the guideline contains sufficient clinical information for a reader to infer the same decision logic.
The guideline does not need to explicitly describe the decision process using "if-then" statements.

Inputs
Decision Tree
The decision tree defines the underlying clinical decision logic that the guideline should preserve.

<decision_tree>
{decision_tree}
</decision_tree>

Clinical Guideline
This is the final guideline to be validated.

<guideline>
{guideline}
</guideline>

Instructions
- Treat the decision tree as the source of truth for clinical logic.
- Match each node based on semantic meaning rather than exact wording.
- Identify guideline sentences that correspond to each decision tree node.
- Return only sentences that provide evidence that the node is represented.
- Select the minimal sufficient sentences needed to demonstrate node coverage.
- Do not include unrelated background sentences.
- Do not require the guideline to follow the order of the decision tree.
- If no corresponding guideline sentence exists, return an empty array.

Output Format
Return JSON only.
The output should directly map each decision tree node ID to the corresponding guideline sentences.

Format:
```
{
  "<node_id>": [
    "Guideline sentence corresponding to this node."
  ]
}
```

Requirements:
- Each key must be a node ID from the provided decision tree.
- Each value must be an array of guideline sentences.
- Use an empty array when no corresponding sentence is found.
- Do not include explanations, judgments, revision suggestions and any text outside the JSON object.

Your Output: