Distilling Expert-level Planning for Real-world Diabetes Prescribing

Yein Park [1 2 †] **Sihyeon Park** [1 †] **Taeyun Roh** [1] **Mujeen Sung** [3] **Jaehoon Yun** [4] **Sooyeon Lee** [2]
Sanghoon Lee [2 *] **Jaewoo Kang** [1 2 *]

Abstract

Real-world prescribing is constrained not only by clinical appropriateness but also by reimbursement policy, prior treatment history, and patient preference, which remain under-modeled in exam-style medical QA training and evaluation. We present a practical framework for expert-usable clinical agents for diabetes prescribing that combines expert planning distillation—distilling an expert prescribing workflow from patient assessment and safety checks to candidate regimen selection and reimbursement verification—with dynamic search over guidelines, reimbursement criteria, drug information, and up-to-date references. To evaluate real-world prescribability, we construct an EMR-derived diabetes prescribing benchmark from 170 de-identified real-world patient cases and evaluate on 340 test instances with a clinician-aligned, reimbursement-aware rubric. On this benchmark, our expert planning distillation and dynamic search improve reimbursement-aware prescribing quality over general and biomedical baselines and approach the performance of frontier proprietary models.

1. Introduction

Large language models (LLMs) have shown strong performance on medical question answering, clinical reasoning, and open-ended health evaluation, raising interest in their use as clinician-facing assistants (Jin et al., 2021; Abbasian et al., 2024; Arora et al., 2025; Singhal et al., 2025). At the same time, medical evaluation is moving beyond static exam-style QA toward more realistic, workflow-aware settings such as rubric-based assessment, interactive EHR environments, and agentic clinical tasks (Schmidgall et al., 2024; Jiang et al., 2025; Lee et al., 2025; Bedi et al., 2025a). However, strong benchmark performance does not directly translate to clinically usable prescribing (Hager et al., 2024). Real prescribing requires interpreting patient-specific EMR evidence, prior prescriptions, and longitudinal treatment context while also accounting for reimbursement policy and patient preference. This is especially important in diabetes, where treatment decisions are shaped not only by glycemic control and comorbidities, but also by insurance eligibility, access, and modality preference.

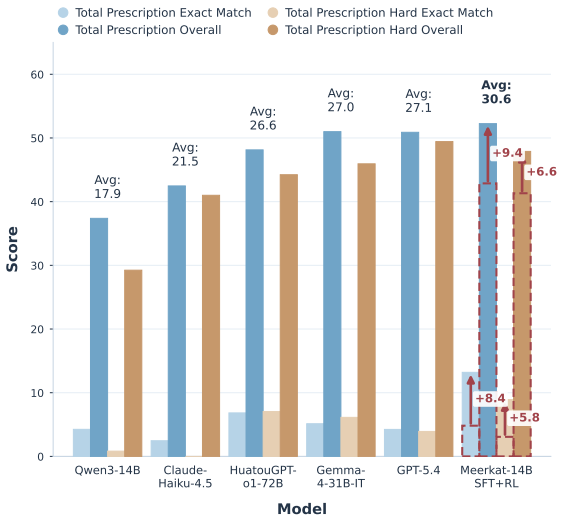

Figure 1. Average performance across selected models on Total Prescription tasks. Meerkat-14B SFT+RL achieves the highest average score, with dashed outlines indicating the original Meerkat-14B baseline and arrows showing the improvement after expert-level planning distillation.

We study this problem in de-identified real-world Korean EMR data and build an expert-usable, reimbursement-aware diabetes prescribing agent. Starting from 170 real patient cases, we construct 340 test instances for evaluation and train the model with synthetic EMR-based supervision, expert planning-and-answer distillation, and search-augmented reinforcement learning. Our approach combines learning an expert prescribing workflow—from patient assessment and safety checking to candidate regimen selection and reimbursement verification—with dynamic search over

[1]Department of Computer Science, Korea University, Seoul, Republic of Korea [2]AIGEN Sciences Inc, Seoul, Republic of Korea [3]School of Computing, Kyung Hee University, Yongin, Republic of Korea [4]College of Medicine, Hanyang University, Seoul, Republic of Korea. Correspondence to: Sanghoon Lee <sanghoon.lee@aigensciences.com>, Jaewoo Kang <kangj@korea.ac.kr>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

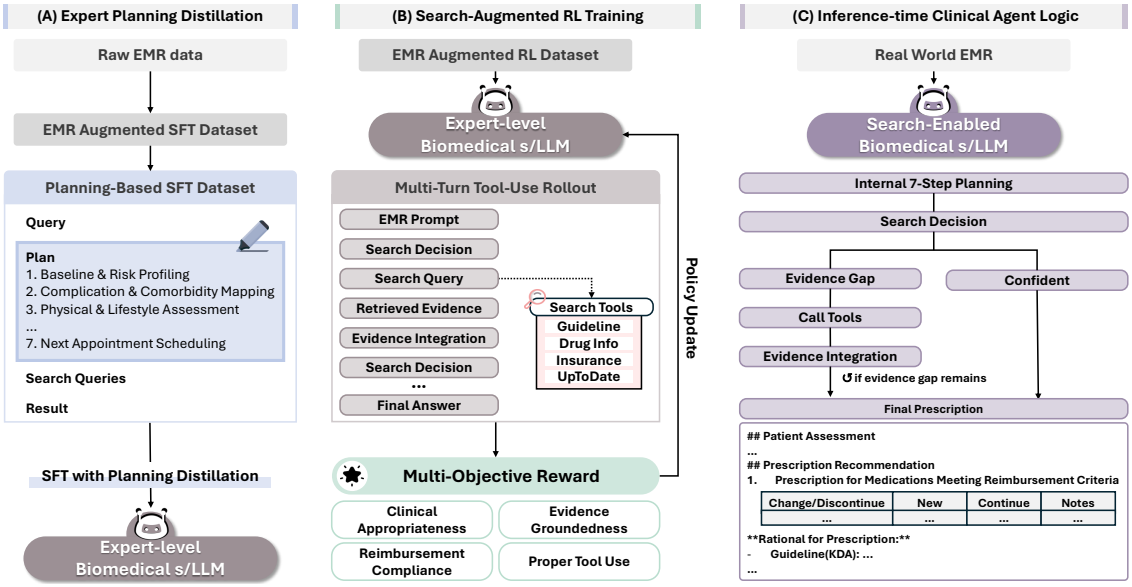

Figure 2. Overview of the planning-distilled and search-augmented pipeline. (A) A planning-based SFT dataset is constructed from EMR-augmented cases and used for SFT with planning distillation. (B) Search-augmented RL optimizes multi-turn tool-use with search tools and rewards. (C) During inference, the model applies planning, retrieves evidence when needed, and outputs the final prescription.

guidelines[1], reimbursement criteria[2], and UpToDate[3], since reimbursement-aware prescribing depends on both insurance rules and clinical prescribing constraints. We further introduce an EMR-derived diabetes prescribing benchmark with reimbursement-aware evaluation. Together, our results suggest that clinically deployable prescribing agents should be trained on expert process traces and evaluated on real prescribing constraints, not only on medical answer correctness.

2. Related Work

Healthcare benchmark for LLM Medical LLM evaluation has progressed from static exam-style QA toward more realistic, rubric-based, and workflow-aware settings. While early benchmarks mainly measured factual recall and short-form reasoning (Jin et al., 2021), recent work evaluates broader clinical usefulness, including conversational quality, physician-defined criteria, and task diversity (Schmidgall et al., 2024; Abbasian et al., 2024; Arora et al., 2025; Bedi et al., 2025a; Wang et al., 2026). MedAgentBench and FHIR-AgentBench further extend this line by evaluating interactive EHR-based agent behavior, emphasizing sequential decision-making and tool use rather than answer accuracy alone (Jiang et al., 2025; Lee et al., 2025). Our benchmark follows this shift, focusing specifically on real-world prescribing under practical clinical constraints.

Real-world aware medical agent Recent medical AI has moved beyond standalone diagnosis or reasoning toward agents that act within realistic clinical workflows. Prior work has explored interactive EHR agents, system-level failure modes in clinical multi-agent settings, and trustworthy decision-making under uncertainty (Jiang et al., 2025; Bedi et al., 2025b; Song et al., 2025). In parallel, medical reasoning models such as HuatuoGPT-o1 and Clinical-R1 have strengthened clinical reasoning itself (Chen et al., 2024; Gu et al., 2025). Our work builds on these directions but targets a more specific setting: real-world-aware metabolic disorder prescribing, where planning, active retrieval, and prescription generation must be aligned with reimbursement rules, patient context, and clinically usable treatment decisions.

3. Method

3.1. Task Setup and Training Overview

We study clinician-assistive diabetes prescribing from de-identified Korean EMR snapshots, where the model must generate an actionable medication plan that is clinically appropriate and also feasible under real-world constraints such as contraindications, prior prescriptions, patient preference, and reimbursement eligibility. Unlike medical QA settings that mainly reward answer correctness, our target is *real-world prescribability*: producing a structured regimen that remains usable under local clinical and policy constraints. Starting from 170 real patient cases, we construct 340 test instances for evaluation, while training is performed on three synthetic resources: a 3K EMR-augmented dataset for task specialization, a 3K dataset with expert planning-and-answer traces, and a separate 3K synthetic dataset for search-augmented reinforcement learning. The

[1] 2023 Clinical Practice Guidelines for Diabetes
[2] MOHW, Reimbursement Criteria for Pharmaceuticals
[3] https://www.uptodate.com/

Table 1. Model performance under two evaluation scopes: **Total Prescription** (reimbursed, non-reimbursed, and additional medications) and **Reimbursed Only** (medications subject to reimbursement criteria only). **Exact Match** is the strict identical-match score against the gold regimen, **Insurance** evaluates reimbursement compliance of model-prescribed reimbursed medications without using gold labels, and **Overall** is the insurance-aware auxiliary score with partial credit for clinically equivalent alternatives and reimbursement-based penalties. Values are reported as mean ± standard deviation across patient cases. Judge model is Qwen3.5-122B.

	Total Prescription (340)			Reimbursed Only (340)		
Model	**Exact Match**	**Insurance**	**Overall**	**Exact Match**	**Insurance**	**Overall**
Proprietary LLMs						
GPT-5	3.37 ± 13.9	70.89 ± 24.3	38.95 ± 22.5	3.45 ± 14.2	75.00 ± 23.7	41.14 ± 23.4
GPT-5.4	4.24 ± 16.1	73.49 ± 17.1	50.89 ± 16.8	4.59 ± 17.7	**81.33** ± 15.9	**53.96** ± 17.3
Cluade-Haiku-4.5	2.46 ± 12.0	68.70 ± 25.5	42.47 ± 20.7	2.56 ± 12.3	74.07 ± 26.2	45.92 ± 21.9
Open-source LLMs						
Qwen3-14B	4.25 ± 16.4	60.29 ± 33.1	37.38 ± 25.1	4.35 ± 17.0	61.85 ± 33.6	37.61 ± 26.0
Gemma-4-31B-IT	5.13 ± 17.7	**76.85** ± 16.5	50.99 ± 18.5	5.23 ± 18.1	78.83 ± 16.6	51.72 ± 18.8
GLM-4.7-Flash (31B)	4.90 ± 17.8	55.58 ± 36.7	36.26 ± 27.7	4.42 ± 17.1	56.09 ± 38.5	34.24 ± 28.3
Kimi K2.5 (1.1T)	0.85 ± 6.5	42.92 ± 40.5	27.98 ± 27.7	0.79 ± 6.5	44.49 ± 41.8	29.11 ± 28.8
Medical LLMs						
MedResearcher-R1-32B	1.76 ± 11.4	25.23 ± 37.4	16.05 ± 26.3	1.79 ± 11.6	25.19 ± 37.9	15.87 ± 26.4
HuatouGPT-o1-72B	6.83 ± 20.1	75.20 ± 19.7	48.13 ± 19.0	6.90 ± 20.4	76.16 ± 19.3	48.42 ± 19.6
Meerkat-14B	4.82 ± 17.1	67.56 ± 24.4	42.87 ± 19.8	4.21 ± 15.4	69.61 ± 23.3	41.69 ± 20.6
SFT without Planning	4.88 ± 16.7	72.78 ± 22.9	42.19 ± 22.0	3.96 ± 15.2	72.43 ± 24.2	40.22 ± 22.6
SFT with Planning	4.82 ± 16.5	70.91 ± 23.4	43.96 ± 20.8	4.48 ± 16.2	71.83 ± 23.0	43.01 ± 21.4
SFT with Planning + RL	**13.2** ± 27.6	72.65 ± 20.9	**52.25** ± 21.0	**13.3** ± 28.0	73.40 ± 23.5	52.21 ± 21.5

overall pipeline follows three stages—clinical specialization, expert planning distillation, and search-augmented RL—combining medical-domain fine-tuning (Kim et al., 2025), checklist- or rubric-based supervision (Shao et al., 2025; Viswanathan et al., 2025; Li et al., 2026), and tool-use RL (Jin et al., 2025; Chai et al., 2025; Gu et al., 2025; Liu et al., 2026) in a prescribing-centered setup.

3.2. Expert Planning Distillation

We represent prescribing as an expert workflow (ame, 2026) that summarizes the patient state, checks missing information and contraindications, proposes candidate regimens, verifies reimbursement feasibility, and produces a final prescription and follow-up plan, as in §A.1. We distill both the intermediate plan and final answer into the model, so that supervision reflects not only *what* to prescribe but also *how* clinicians reach a prescribing decision. Unlike prior rubric-based approaches that rely on generic or automatically generated checklists (Shao et al., 2025; Viswanathan et al., 2025; Li et al., 2026), our plan is a clinician-authored procedural rubric grounded in real diabetes prescribing practice. It is designed to internalize expert workflow more directly than prompt-only planning or answer-only supervision.

3.3. Reimbursement-Aware Search-Augmented RL

Because reimbursement rules and local prescribing constraints cannot be fully captured by parametric knowledge alone, we further train the model in a multi-turn tool-use environment over guideline, reimbursement, Up-To-Date, and drug-information corpora. This stage is inspired by search-enabled RL and tool-use training (Jin et al., 2025; Chai et al., 2025), but our setting differs from general-domain or data-free search agents (Yue et al., 2026) by optimizing directly for clinically grounded prescribing quality. We use rewards for clinical appropriateness, evidence-groundedness, reimbursement compliance, and proper tool use, following recent clinical multi-objective RL directions (Gu et al., 2025; Liu et al., 2026). We also adopt FP16 training for more stable RL optimization (Qi et al., 2025); additional implementation details are provided in Appendix B.

4. Benchmark and Evaluation

4.1. EMR-Derived Prescribing Benchmark

We evaluate our model on a diabetes prescribing benchmark constructed from 170 de-identified real-world patient cases collected from Korean clinical practice, and use 340 derived instances as the test set. Each instance is built from EMR evidence necessary for prescribing, including clinical history, laboratory findings, and prior medications, and the task is to predict a new prescription decision for the current visit. Our goal is not to test abstract medical QA, but to evaluate whether a model can produce prescribing outputs that are usable for assisting physicians in real hospital settings.

For each case, we use a professor-level medical expert response as the gold reference. Because multiple treatment decisions can be clinically reasonable, the benchmark is designed as a realistic prescribing task rather than a single-label medication recommendation problem. Model outputs therefore include diagnosis and assessment, reimbursed pre-

Table 2. Inference-time prompt ablation with original 140 cases. We compare base prompting, planning-only, tool-use-only, and all combined settings for GPT-5 and Meerkat-14B, reporting Total Prescription performance in Exact Match, Insurance, and Overall. Planning gives modest gains for GPT-5, whereas tool use is critical for Meerkat-14B and the combined setting performs best overall.

Model	EM	Insurance	Overall
GPT-5	1.59	73.93	43.28
+ Planning Prompt	2.40	74.71	44.10
+ Tool-Use Prompt	1.94	67.04	39.43
+ Planning/Tool-Use	4.31	69.69	41.72
Meerkat-14B	2.02	25.41	15.90
+ Planning Prompt	0.00	3.375	2.087
+ Tool-Use Prompt	5.50	67.53	42.87
+ Planning/Tool-Use	6.83	67.41	42.97

scribing decisions, non-reimbursed prescribing decisions, and additional medications, so that evaluation reflects both clinical appropriateness and practical usability. Data formatting and prompting details are provided in Appendix A.

4.2. Expert-Curated Hard Subset

To further assess model behavior in clinically difficult prescribing scenarios, we construct an expert-curated hard subset from the EMR-derived benchmark. Candidate hard cases are first identified using predefined diabetes-prescribing difficulty rules, including advanced CKD with severe hypoglycemia history, frailty with HFrEF, decompensated cirrhosis, LADA, hypoglycemia unawareness, gastroparesis with ASCVD, steroid-induced glycemic variability, complex drug–drug interactions, severe injection phobia, and economic constraints.

We use GPT-5 (Singh et al., 2025) and Gemini 3.1 Pro (DeepMind, 2025) for complementary first-pass screening: the former provides conservative high-specificity labels, while the latter captures broader contextual signals with higher sensitivity. All candidate labels are then reviewed, verified, and reclassified by two professor-level diabetes experts, allowing a case to receive multiple difficulty labels when clinically appropriate. After de-duplicating patient cases and applying the same instance construction protocol as the main benchmark, the resulting hard subset contains 40 derived test instances. Detailed criteria, review procedure, and category statistics are provided in Appendix C.

4.3. Evaluation Metrics

We evaluate outputs under two complementary scopes: **Total Prescription**, which considers the full final regimen including reimbursed, non-reimbursed, and supportive medications, and **Reimbursed Only**, which restricts evaluation to core diabetes medications relevant to reimbursement review. For each scope, we report three metrics: **Exact Match**, which measures strict regimen alignment with the expert reference; **Insurance**, which measures reimbursement compliance of score-able diabetes medications; and **Overall**, which combines alignment with reimbursement awareness.

Table 3. Performance difference across baseline variants. Alignment values are reported same as mean of Table 15 with original 140 cases, and the result is about Total Prescription. Score of HealthBench-Diabetes are also reported same as Table 4.

	Total Prescription			HB-Diabetes	
Model	EM	Insurance	Overall	Full	Hard
Qwen3-14B	7.51	61.23	40.65	51.1	19.9
SFT without Planning	4.21	71.28	37.48	54.0	18.6
SFT with Planning	3.25	70.03	39.54	51.1	17.4
SFT with Planning + RL	5.26	74.07	45.29	50.5	25.4
Meerkat-14B	6.83	67.41	42.97	49.4	12.9
SFT without Planning	5.40	73.92	43.87	47.6	17.1
SFT with Planning	6.53	71.70	46.39	48.6	16.4
SFT with Planning + RL	14.7	73.14	53.12	50.6	15.4

This design emphasizes the two criteria that matter most in our setting: whether the predicted regimen remains clinically aligned with expert prescribing, and whether it satisfies reimbursement constraints required in practice. Exact matching alone is too narrow for real prescribing, since clinically acceptable substitutions may exist, while reimbursement compliance alone does not capture agreement with the expert regimen. We therefore report both alignment-centered and insurance-centered metrics, and leave the detailed parsing rules, matching categories, and scoring formulas to §D.

5. Results and Discussion

We compare both closed- and open-source LLMs on our diabetes prescribing benchmark: three proprietary LLMs; GPT-5 (Singh et al., 2025), GPT-5.4 (OpenAI, 2026), Claude-Haiku-4.5 (Anthropic, 2025), four open-source LLMs; Qwen3-14B (Yang et al., 2025), Gemma-4-31B-IT (Farabet & Lacombe, 2026), GLM-4.7-Flash (Zeng et al., 2025), Kimi K2.5 (Team et al., 2026), and three medical knowledge fine-tuned LLMs; MedResearcher-R1-32B (Yu et al., 2025), HuatuoGPT-o1-72B (Chen et al., 2024), Meerkat-14B (Kim et al., 2025). All models are evaluated under a unified agentic setup with the same prompt, tool-use interface, parsing pipeline, and scoring protocol, enabling a controlled comparison across model families. Judge model is Qwen3.5-122B (Qwen Team, 2026) for rubric based LLM-as-a-judge (Zheng et al., 2023). Detailed decoding, retrieval, and judge settings are provided in Appendix E.

5.1. Main Results

Table 1 shows that our post-training pipeline improves reimbursement-aware prescribing on the EMR-derived benchmark. Because Exact Match requires identical reconstruction of the final regimen, scores remain low across all models, making Insurance and Overall more informative indicators of practical prescribing quality. Under *Total Prescription*, *SFT with Planning + RL* achieves the strongest

Table 4. Performance on HealthBench-Diabetes. Values are reported as scores on the Full and Hard subsets.

Model	HB-Diabetes Full	HB-Diabetes Hard
GPT-5	61.5	35.8
GPT-5.4	52.5	24.5
Claude-Haiku-4.5	43.2	8.53
Qwen3-14B	51.1	19.9
Gemma-4-31B-IT	47.9	14.7
GLM-4.7-Flash (31B)	42.1	11.8
Kimi K2.5 (1.1T)	57.9	29.2
MedResearcher-R1-32B	33.3	3.68
HuatuoGPT-o1-72B	28.3	0.00
Meerkat-14B	49.4	12.9
SFT without Planning	47.6	17.1
SFT with Planning	48.6	16.4
SFT with Planning + RL	50.6	15.4

overall performance, obtaining the highest Exact Match and Overall scores among all models, while *Gemma-4-31B-IT* attains the highest Insurance score. Under *Reimbursed Only*, however, GPT-5.4 achieves the highest Insurance and Overall scores, whereas *SFT with Planning + RL* still obtains the best Exact Match score and remains competitive in Overall. Among general-purpose baselines, *Gemma-4-31B-IT* performs best under *Total Prescription*. These results suggest that reimbursement compliance, clinically useful partial-credit prescribing, and strict gold-regimen reconstruction are related but distinct capabilities; detailed case studies are provided in Appendix F. The result of original 140 cases are provided in Table 15.

5.2. Results on Hard-subset

Table 16 further evaluates whether these trends hold on the expert-curated hard subset. On this more challenging setting, *SFT with Planning + RL* achieves the highest Exact Match under both *Total Prescription Hard* and *Reimbursed Only Hard* (8.92 and 9.07, respectively), outperforming even GPT-5.4 in strict regimen reconstruction. Although GPT-5.4 obtains the best Insurance and Overall scores, our RL-trained model remains highly competitive, reaching the second-best Overall score under *Total Prescription Hard* (47.89 vs. 49.43) and a comparable Overall score under *Reimbursed Only Hard* (49.15 vs. 51.81). This indicates that the proposed post-training pipeline does not merely improve average-case performance, but also remains competitive with the latest proprietary SOTA model on difficult expert-defined prescribing cases.

5.3. Ablation Study

Table 2 studies inference-time prompt design by varying planning and tool-use instructions. Judging is done with same Qwen3.5-122B. For GPT-5, adding planning yields modest gains, and combining planning with tool use substantially improves Exact Match, although tool-enabled settings slightly reduce Insurance, suggesting that retrieval can sometimes conflict with strong parametric knowledge. For the base Meerkat-14B model, tool use is the main driver of improvement, while planning alone is ineffective and planning plus tool use gives the best overall inference-time performance. Overall, although the magnitude differs by model, the combination of planning and tool use is the most effective setting for open-source models, consistent with the strongest results of the trained Meerkat variants in Table 1.

Table 3 examines how backbone choice affects post-training gains. For both Qwen3-14B and Meerkat-14B, domain adaptation improves reimbursement compliance, but the two backbones show different trajectories: under Qwen, SFT lowers Exact Match and Overall relative to the base model and RL is required to recover and surpass the baseline. In contrast, the Meerkat backbone shows a more stable pattern, with Overall improving step by step from the base model to SFT and then to RL while maintaining strong Insurance performance. These results suggest that planning- and RL-based adaptation is more effective on a medically specialized backbone. We also analyze search behavior and corpus effects in Appendix G.

5.4. Additional Results

We additionally evaluate all models on HealthBench-Diabetes, a filtered subset of HealthBench designed to measure broader diabetes-related medical reasoning; subset construction and evaluation details are provided in Appendix H. We run the standard HealthBench generation and judge pipeline and report the overall score on this filtered subset, using a larger generation budget of 8,196 tokens. Judge model is same as original: GPT-4.1 (OpenAI, 2025).

As shown in Table 4, GPT-5 achieves the highest scores on both the Full and Hard subsets, indicating strong general diabetes knowledge and clinical competence. However, this advantage does not translate directly to the EMR-derived prescribing benchmark, where *SFT with Planning + RL* performs best on Total Prescription Exact Match and Overall. This contrast suggests that broad diabetes reasoning alone is not sufficient for real-world prescribing, which additionally requires case-grounded regimen construction and reimbursement-aware decision making.

6. Limitations and Future Work

Our current study is limited to diabetes prescribing and a Korea-specific reimbursement setting, and the benchmark remains relatively small and centered on a single institution reference rather than multi-institution clinician review. Future work will expand evaluation through blinded clinician assessment, comparisons against machine-generated planning, and broader external validation on public benchmarks and case-based challenges. We are also planning to expand backbone and LLM evaluator into other model families.

Impact Statement

This work aims to advance clinically usable language agents by moving beyond exam-style medical QA toward real-world prescribing support grounded in electronic medical records, expert prescribing workflows, and reimbursement-aware decision making. A potential positive impact of this research is to improve physician-facing decision support in chronic disease management by helping models generate treatment plans that are not only medically plausible but also feasible under practical clinical constraints such as prior treatment history, insurance eligibility, and patient preference.

At the same time, this work should not be interpreted as supporting autonomous prescribing. Because our study is limited to diabetes, a Korea-specific reimbursement setting, and retrospective benchmark-based evaluation, inappropriate deployment could lead to over-reliance on model outputs, failure under distribution shift, or unfair decisions if reimbursement constraints are treated as overriding clinical judgment. Our intended use is therefore clinician-assistive rather than clinician-replacing, and future work should include broader multi-institution evaluation, blinded clinician review, and prospective validation before any real-world deployment.

References

4. comprehensive medical evaluation and assessment of comorbidities: Standards of care in diabetes—2026. *Diabetes Care*, 49(Supplement_1):S61–S88, 2026.

Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Shakeri Hossein Abad, Z., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *Npj digital medicine*, 7(1):82, 2024.

Anthropic. Introducing claude haiku 4.5. 2025.

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

Bedi, S., Cui, H., Fuentes, M., Unell, A., Wornow, M., Banda, J. M., Kotecha, N., Keyes, T., Mai, Y., Oez, M., et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025a.

Bedi, S., Mlauzi, I., Shin, D., Koyejo, S., and Shah, N. H. The optimization paradox in clinical ai multi-agent systems. *arXiv preprint arXiv:2506.06574*, 2025b.

Chai, J., Yin, G., Xu, Z., Yue, C., Jia, Y., Xia, S., Wang, X., Jiang, J., Li, X., Dong, C., He, H., and Lin, W. Rl-factory: A plug-and-play reinforcement learning post-training framework for llm multi-turn tool-use, 2025. URL https://arxiv.org/abs/2509.06980.

Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., and Wang, B. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

DeepMind, G. Gemini 3.1 pro - model card. 2025. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Farabet, C. and Lacombe, O. Gemma 4: Byte for byte, the most capable open models, April 2026. URL https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/.

Gu, B., Zhou, H., Segal, B. M., Wu, J., Cao, Z., Zhong, H., Clifton, L., Liu, F., and Clifton, D. A. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. *arXiv preprint arXiv:2512.00601*, 2025.

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Hou, X., Zhao, Y., Wang, S., and Wang, H. Model context protocol (mcp): Landscape, security threats, and future research directions. *ACM Transactions on Software Engineering and Methodology*, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jeong, M., Sohn, J., Sung, M., and Kang, J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129, 2024a.

Jeong, S., Baek, J., Cho, S., Hwang, S. J., and Park, J. C. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024b.

Jiang, Y., Black, K. C., Geng, G., Park, D., Zou, J., Ng, A. Y., and Chen, J. H. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):AIdbp2500144, 2025.

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S. O., Wang, D., Zamani, H., and Han, J. Search-r1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., Yoon, C., Sohn, J., Park, J., Reykhart, O., et al. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240, 2025.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Lee, G., Bach, E., Yang, E., Pollard, T., JOHNSON, A., Choi, E., Lee, J. H., et al. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. In *Machine Learning for Health 2025*, 2025. URL https://openreview.net/forum?id=Le1hGVQNb8.

Li, S., Zhao, J., Wei, M., Ren, H., Zhou, Y., Yang, J., Liu, S., Zhang, K., and Chen, W. Rubrichub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation. *arXiv preprint arXiv:2601.08430*, 2026.

Liu, S.-Y., Dong, X., Lu, X., Diao, S., Belcak, P., Liu, M., Chen, M.-H., Yin, H., Wang, Y.-C. F., Cheng, K.-T., et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

OpenAI. Introducing gpt-4.1 in the api. 2025.

OpenAI. Introducing gpt-5.4. 2026.

Qi, P., Liu, Z., Zhou, X., Pang, T., Du, C., Lee, W. S., and Lin, M. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., and Moor, M. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, R., Asai, A., Shen, S. Z., Ivison, H., Kishore, V., Zhuo, J., Zhao, X., Park, M., Finlayson, S. G., Sontag, D., et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S. R., Cole-Lewis, H., et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3): 943–950, 2025.

Sohn, J., Park, Y., Yoon, C., Park, S., Hwang, H., Sung, M., Kim, H., and Kang, J. Rationale-guided retrieval augmented generation for medical question answering. 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. 12739–12753, 2025.

Song, Y., Jeong, M., and Sung, M. Trustworthy agents for electronic health records through confidence estimation. *arXiv preprint arXiv:2508.19096*, 2025.

Team, K., Bai, T., Bai, Y., Bao, Y., Cai, S., Cao, Y., Charles, Y., Che, H., Chen, C., Chen, G., et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Viswanathan, V., Sun, Y., Kong, X., Cao, M., Neubig, G., and Wu, T. Checklists are better than reward models for aligning language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RPRqKhjrr6.

Wang, X., shuqi, G., Shen, Y., Chen, J., Wang, J., Gu, J., Zhang, P., Liu, L., and Wang, B. Liveclin: A live clinical benchmark without leakage. In *The Fourteenth*

International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=E0WSAugJ0j.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Yu, A., Yao, L., Liu, J., Chen, Z., Yin, J., Wang, Y., Liao, X., Ye, Z., Li, J., Yue, Y., et al. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*, 2025.

Yue, Z., Upasani, K., Yang, X., Ge, S., Nie, S., Mao, Y., Liu, Z., and Wang, D. Dr. zero: Self-evolving search agents without training data. *arXiv preprint arXiv:2601.07055*, 2026.

Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

A. Detail of Planning and Tool-Use

A.1. Expert-level Planning Prompt

The planning prompt, shown in Table 5, was constructed to reflect the real clinical workflow used by diabetes specialists in outpatient prescribing (ame, 2026). We organized it as a seven-step procedural rubric covering patient profiling, complication and comorbidity review, physical and lifestyle assessment, biochemical target evaluation, prescription optimization with reimbursement review, monitoring, and follow-up scheduling. This structure was intended to support both planning distillation during training and structured reasoning during inference.

A.2. SFT Training Dataset Generation Prompt

The system prompt shown in Table 6 was used to generate planning-based SFT examples from EMR-derived synthetic diabetes cases. For each case, the prompt produces a structured `Query-Plan-Search_Queries-Result` tuple, where the query represents a plausible clinician request, the `Plan` component follows the seven-step workflow introduced in Table 5, the search queries reflect external evidence needs, and the result is restricted to case-supported decisions. We used GPT-5 (Singh et al., 2025) for generating SFT dataset, which is renowned for its description following ability.

A.3. Inference Prompt

During inference, we used a composite setup with multiple components. Representative components of this setup are shown in Tables 7 to 10, corresponding respectively to planning guidance, tool-use instruction prompts, a search-and-reasoning policy, and a fixed final output template. The planning guidance and tool-use instructions correspond to the inference-time prompt components ablated in Table 2, while the remaining excerpts illustrate how these components were situated within the broader inference procedure. The planning guidance reused the seven-step workflow shown in Table 5, while the tool-use instructions defined the available retrieval tools and their calling format. The search-and-reasoning policy guided the model to decide when external retrieval was necessary, how to prioritize tools, and when to perform iterative follow-up searches. The final template specified the required clinician-facing answer structure.

B. Detailed Model Training Pipeline

B.1. Task Definition and Input Schema

Our input is a de-identified EMR snapshot containing structured and semi-structured clinical information, including demographics, glycemic status, renal function, comorbidities, current medications, adverse-event history, and patient preferences when available. The target output is an actionable medication plan with drug or drug class, dose or titration strategy, rationale, and follow-up considerations. The system is intended for clinician assistance, and final decisions remain with human providers. For all synthetic dataset generation, we used GPT-5 (Singh et al., 2025).

B.2. Stage 1: Biomedical Expert Base Model

We initialize from an open-source instruction-tuned LLM and first adapt it to the biomedical domain through supervised fine-tuning on medical knowledge and reasoning data. This stage is designed to provide strong parametric medical competence before task-specific prescribing adaptation. We use this model as the backbone for subsequent planning distillation and search-based RL. Following Kim et al., we used a newly-trained Qwen3-14B (Yang et al., 2025), which is a baseline LLM having both reasoning ability and tool-call availability.

B.3. Stage 2: Expert Planning Distillation

For each case, we construct a clinician-authored intermediate plan paired with a final prescription answer. The plan follows a structured workflow: (1) summarize the patient state and treatment goal; (2) identify required information and missing data; (3) check contraindications and safety issues; (4) propose candidate regimens with rationale; (5) verify reimbursement eligibility; and (6) produce a final prescription and monitoring plan. Given an input case x, expert plan p, and final answer y, we train the model to generate (p, y) conditioned on x. We compare the effect and the advantage of planning against ablation, which is *SFT without Planning* and *SFT with Planning* in Table 1.

Training Details. We train the Stage 2 model with supervised fine-tuning using LoRA (Hu et al., 2022) on 3K synthetic plan-answer pairs by MS-SWIFT (Zhao et al., 2025). Training is run on 4 NVIDIA B200 180GB GPUs for approximately 3 hours, using bfloat16 precision and DeepSpeed ZeRO-3. We train for 600 steps with a learning rate of 1×10^{-5}, per-device batch size 1, and gradient accumulation of 8. We use a maximum sequence length of 16,384 with packing enabled. For LoRA, we set rank to 16 and alpha to 32, applying adapters to all linear layers. We used 400 steps checkpoints as they show the best performance.

B.4. Stage 3: Search-Augmented Reinforcement Learning

We further train the model in a multi-turn tool-use environment. At each step, the model may issue a search query to one of several corpora, including clinical guidelines, reimbursement criteria, up-to-date references, and drug information resources. Retrieved observations are appended to the context, and the model continues reasoning and prescribing conditioned on both the original EMR and retrieved evidence.

Implementation Details. Stage 3 reinforcement learning is conducted with RL Factory (Chai et al., 2025) using the PPO (Schulman et al., 2017) trainer and GDPO (Liu et al., 2026) as the advantage estimator. We train on 3,000 synthetic cases, with separate training and validation parquet files. Training uses same 4 GPUs (NVIDIA B200 180GB) for 5 epochs with batch size 64, actor learning rate 1×10^{-6}, PPO minibatch size 16, and micro-batch size 1 per GPU. We disable KL reward shaping in the reward term and instead apply KL control with coefficient 0.001. For GDPO, we use reward weights $[0.35, 0.25, 0.30, 0.10]$, batch normalization, and reward scaling. Training is performed for approximately 8 hours.

For rollout, we use vLLM (Kwon et al., 2023) with 4 sampled rollouts per prompt, a maximum of 2 tool-use turns according to the context limit, and stopping criteria based on `</answer>`, `</tool_call>`, and `<|im_end|>`. The maximum prompt length is 16,384 tokens and the maximum response length is 8,192 tokens. The environment is `search_v4`, with thinking enabled and tool calls executed through the Qwen3-style tool manager over the configured search resources. Tool responses are truncated to 128 tokens when necessary. Also, we implement FP16 training to reduce training–inference mismatch and improve optimization stability (Qi et al., 2025). We used 100 steps checkpoints as they show the best performance.

B.5. Reward Design

Our RL objective combines multiple verifiable signals for clinical prescribing quality, including final-answer quality, evidence-grounded reasoning, reimbursement-aware decision making, search behavior, and output validity. In implementation, these signals are operationalized as four reward components: tool use, format, content, and weak answer accuracy. To avoid instability from naively mixing heterogeneous rewards, we adopt a decoupled multi-reward optimization scheme.

Operationalization of Rewards. We implement the reward with four operational components: **tool**, **format**, **content**, and **acc**. These jointly instantiate the conceptual objectives described above: tool-use quality is captured by **tool**; output validity and completion by **format**; clinical appropriateness, evidence-groundedness, and reimbursement-aware reasoning primarily by **content**; and weak answer correctness by **acc**.

To avoid instability from naively mixing heterogeneous rewards, we use a GDPO-style decoupled normalization scheme. For rollout j of case i, let $r_{i,j}^{(m)}$ denote the raw reward for objective $m \in \{\text{tool}, \text{format}, \text{content}, \text{acc}\}$. We first compute a within-case normalized score

$$z_{i,j}^{(m)} = \frac{r_{i,j}^{(m)} - \mu_i^{(m)}}{\sigma_i^{(m)} + \epsilon},$$

where $\mu_i^{(m)}$ and $\sigma_i^{(m)}$ are the mean and standard deviation of objective m across sampled rollouts for the same case. However, if the number of sampled rollouts for a given case is less than a minimum group size ($N < 3$), calculating intra-case statistics becomes highly unstable. In such instances, we fall back to using the pooled mean and standard deviation across the entire batch to ensure stable normalization. We then form the pre-normalized advantage

$$a_{i,j} = \sum_m w_m z_{i,j}^{(m)},$$

with training-time weights

$$(w_{\text{tool}}, w_{\text{format}}, w_{\text{content}}, w_{\text{acc}}) = (0.35, 0.25, 0.30, 0.10).$$

Finally, we apply batch-wise normalization to $a_{i,j}$ before policy optimization. This procedure improves stability by aligning each objective to a comparable relative scale before aggregation.

The detailed evaluation logic for each operational component is structured as follows:

- **Tool Use Reward (tool, $w_{\text{tool}} = 0.35$)**: encourages the appropriate and proactive utilization of external search tools. We implement a Dynamic Query Quality Assessment, which verifies the inclusion of essential keywords tailored to the specific characteristics of each invoked search tool. Positive rewards are granted for contextually appropriate invocations of valid tools, while penalties are applied to constrain hallucinations, such as the use of undefined tools (unknown tool) or incorrectly formatted invocations (invalid call).

- **Format Reward (format, $w_{\text{format}} = 0.25$)**: evaluates whether the model's output strictly adheres to the requested Markdown format and the designated tag structures (e.g., `<answer>`, `</answer>`). To ensure the structural stability and consistency of the final response, penalties are applied in cases where stray tables are generated without the requisite textual analysis structure, or when unnecessary trailing text is produced.

- **Content Reward (content, $w_{\text{content}} = 0.30$)**: assesses the clinical validity and the thoroughness of the supporting evidence within the response. We independently evaluate the appropriateness of the evidence for each distinct clinical section (e.g., patient condition analysis, reimbursement verification, and consideration of comorbidities). A maximum reward of +0.08 is granted per section if sufficient evidence is provided. Furthermore, the score reflects whether critical patient risk factors and keywords related to insurance reimbursement are logically addressed.

- **Accuracy Reward (acc, $w_{\text{acc}} = 0.10$)**: evaluates the extent to which the model's generated response incorporates the core keywords required by the ground truth structure. Instead of relying on a strict hard-matching approach, we employ a section-weighted F1 score-based soft-matching mechanism. This allows for flexibility in measuring clinical contextual alignment while accommodating natural variations in expression.

B.6. MCP-Based Search Interface

During RL training and inference, retrieval is provided through an Model Context Protocol (MCP) (Hou et al., 2025)-based tool interface that allows the model to issue explicit search calls during multi-turn reasoning. We expose separate tools for the main evidence sources used in reimbursement-aware prescribing: `search_uptodate`, `search_guideline`, `search_insurance`, and `search_drug_info`. Each tool takes a natural-language query and returns the top-k retrieved passages from its corresponding corpus, which are then appended to the model context as observations for subsequent reasoning and answer generation.

This design encourages the model to decompose evidence gathering by source rather than relying on a single undifferentiated retriever. In practice, the model can selectively consult clinical references for treatment recommendations, insurance resources for reimbursement rules, and drug information resources for medication-specific prescribing constraints. We additionally retain a general `query_rag` interface for backward compatibility, although our main experiments use the source-specific tools above.

All main experiments in this paper use BM25-based sparse retrieval for search. We use this setup consistently throughout training and evaluation so that the policy learns when to search, which source to query, and how to incorporate retrieved evidence into the final prescription.

C. Construction of the Expert-Curated Hard Subset

We additionally construct an expert-curated hard subset to evaluate whether models can handle prescribing situations that are difficult even in real clinical practice. Unlike the main benchmark, which covers general EMR-derived diabetes prescribing cases, this subset focuses on cases where standard treatment algorithms are insufficient because of contraindications, competing risks, atypical disease mechanisms, drug–drug interactions, patient preference, or cost constraints.

Difficulty criteria. We define ten clinically motivated difficulty categories in consultation with diabetes prescribing experts. These categories are not intended to be mutually exclusive, because a single patient may present with multiple sources of prescribing difficulty. The criteria are as follows:

1. **Advanced CKD with severe hypoglycemia history**: patients with CKD stage 4–5 and a history of severe hypoglycemia, where many glucose-lowering agents are contraindicated or require substantial caution due to renal impairment and hypoglycemia risk.

2. **Frailty with HFrEF**: elderly frail patients with heart failure with reduced ejection fraction, where SGLT2 inhibitors may be clinically beneficial for heart failure but can be difficult to prescribe because of concerns about weight loss, dehydration, and worsening frailty.

3. **Decompensated cirrhosis**: patients with decompensated liver cirrhosis, where the use of medications metabolized in the liver is restricted and treatment choices require careful risk assessment.

4. **LADA**: patients with latent autoimmune diabetes in adults, who may be easily misclassified as having type 2 diabetes but require insulin because of impaired insulin secretory capacity.

5. **Hypoglycemia unawareness**: patients whose average glucose level may appear high but whose glycemic target should be relaxed because of a high risk of severe hypoglycemia.

6. **Obesity with gastroparesis and ASCVD**: obese patients with atherosclerotic cardiovascular disease for whom GLP-1 receptor agonists may be clinically indicated, but whose gastroparesis or gastrointestinal symptoms limit their use.

7. **Periodic long-term high-dose steroid use**: patients receiving repeated or long-term high-dose steroid treatment, where glucose variability is large and postprandial hyperglycemia is prominent, making ordinary insulin dose adjustment strategies less reliable.

8. **HIV with complex concomitant medications**: patients with HIV who are taking multiple drugs such as antivirals, anti-tuberculosis medications, or anticonvulsants, requiring careful consideration of drug–drug interactions.

9. **Severe injection phobia**: patients with marked hyperglycemia who clinically require insulin but remain on oral agents because of strong fear of injections.

10. **Economic constraints**: economically vulnerable patients for whom medically optimal agents may be difficult to prescribe because of treatment cost.

LLM-assisted candidate screening. To reduce the risk of missing difficult cases while maintaining precision, we use two general-purpose LLMs for complementary first-pass screening. GPT-5 (Singh et al., 2025) is used as a conservative classifier that applies explicit numerical and textual criteria strictly. This strategy provides relatively high reliability and specificity, but may have lower sensitivity because cases whose difficulty depends on implicit clinical context can be missed. In contrast, Gemini 3.1 Pro (DeepMind, 2025) is used as a broader contextual screener that considers the full patient context, including prescription patterns and patient statements. This improves sensitivity and helps identify potentially difficult cases missed by the conservative classifier, but can introduce false positives due to broader interpretation. We therefore treat the LLM outputs only as candidate labels rather than final annotations.

Expert verification and reclassification. All candidate hard-case labels are reviewed by two professor-level diabetes experts. The experts verify whether each candidate case truly satisfies the intended clinical difficulty criterion, remove unsupported labels, and reclassify cases when the primary difficulty source is better explained by another category. Because real patients often involve multiple simultaneous prescribing challenges, we allow multi-label assignment at the patient-case level. For the final hard subset, we retain only conservative, expert-verified labels.

Final hard-subset composition. Table 11 summarizes the final category-level statistics. The counts sum to 22 because duplicate category assignments are allowed across difficulty types. After de-duplicating patient cases and applying the same two-instance construction protocol used for the main benchmark, the final hard subset contains 40 derived test instances.

D. Detailed Metric Definitions

This appendix provides the full definitions of the evaluation metrics reported in Table 1.

D.1. Evaluation Scopes

We evaluate model outputs under two prescription scopes.

Total Prescription. This scope considers the full final regimen, including (i) reimbursed diabetes medications, (ii) non-reimbursed diabetes medications, and (iii) additional supportive or adjunctive medications proposed by the model.

Reimbursed Only. This scope restricts evaluation to core diabetes medications that are directly relevant to reimbursement review.

Alignment metrics are computed on the medication set defined by the chosen scope. For reimbursement-related metrics, however, only medications that are actually governed by the diabetes reimbursement document are scored. Medications outside that scope are marked as *excluded* and do not contribute to the insurance denominator or to regimen-level reimbursement penalties.

D.2. Action Parsing and Matching Categories

Model-generated prescription updates are parsed into medication-level actions: `keep`, `stop`, `add`, `modify`, and `replace`. Here, `modify` denotes a change in dose, strength, frequency, or formulation of the same medication, while `replace` denotes substitution of a baseline drug by a related alternative.

After parsing, the final predicted regimen is compared against the gold final regimen using greedy exact matching followed by relaxed equivalence matching. Each predicted medication is then assigned to one of four comparison categories:

- **identical**: exact match to a gold medication,
- **equivalent**: not strictly identical, but matched as the same core ingredient or a closely corresponding alternative,
- **additional**: present in the prediction but unmatched to the gold regimen,
- **missing**: present in the gold regimen but absent from the prediction.

D.3. Exact Match (Primary Alignment)

The Exact Match score measures how faithfully the model reproduces the gold final prescription using strict identical matching only.

Let

$$G = \text{number of medications in the gold final regimen},$$

$$P = \text{number of medications in the predicted final regimen},$$

$$I = \text{number of identical matches}.$$

We define

$$\text{ExactMatch} = 100 \times \left(0.7 \times \frac{I}{G} + 0.3 \times \frac{I}{P}\right). \tag{1}$$

The first term is a recall-like gold coverage term, and the second term is a precision-like prediction term. By assigning a larger weight to coverage, the metric penalizes missing necessary medications more strongly than adding extra ones.

D.4. Auxiliary Alignment with Partial Credit

Strict exact matching may under-credit clinically reasonable substitutions. We therefore define an auxiliary alignment score that gives partial credit to equivalent matches.

Let

$$E = \text{number of equivalent matches}.$$

We assign weight $\alpha = 0.7$ to each equivalent match. The auxiliary alignment score is

$$\text{Auxiliary} = 100 \times \left(0.7 \times \frac{I + \alpha E}{G} + 0.3 \times \frac{I + \alpha E}{P} \right). \tag{2}$$

This score is not reported as a separate column in the main table, but it serves as the basis for the reimbursement-aware **Overall** metric below.

D.5. Insurance Review Labels

To avoid conflating true reimbursement violations with lack of document coverage, each predicted medication is first assigned a *scope status*:

- **scoreable**: the medication is directly governed by the diabetes reimbursement document and should be evaluated,
- **excluded**: the medication is outside the scope of the document and should not affect reimbursement scoring.

Typical excluded items include supportive medications, complication-related drugs, and other adjunctive prescriptions that are not directly adjudicated by the diabetes reimbursement criteria.

For *scoreable* medications only, we assign one of three reimbursement judgments:

- **eligible**: the prescription is supported by the reimbursement document,
- **uncertain**: the medication is within scope, but the available case evidence is insufficient to determine eligibility,
- **not_eligible**: the prescription explicitly violates or fails to satisfy the reimbursement criteria.

Thus, **uncertain** is reserved for in-scope but unresolved cases, whereas out-of-scope medications are handled separately as **excluded**.

D.6. Insurance Metric

The **Insurance** metric evaluates reimbursement compliance of the model output itself rather than agreement with the gold regimen.

Let $\mathcal{S}$ be the set of *scoreable* predicted medications and let

$$Q = |\mathcal{S}|.$$

If $Q = 0$, the insurance score is not reported for that case.

For each scoreable medication $m \in \mathcal{S}$, we define an item credit

$$c(m) = \begin{cases} 1.00, & \text{if } m \text{ is eligible,} \\ 0.75, & \text{if } m \text{ is uncertain,} \\ 0.00, & \text{if } m \text{ is not_eligible.} \end{cases} \tag{3}$$

The item-level insurance score is

$$\text{InsuranceItem} = 100 \times \frac{1}{Q} \sum_{m \in \mathcal{S}} c(m). \tag{4}$$

We then apply a regimen-level penalty based only on the *scoreable core diabetes regimen*. Excluded medications do not by themselves make the regimen uncertain or not eligible. The regimen penalty is

$$\pi_{\text{ins}} = \begin{cases} 0, & \text{eligible,} \\ 2.5, & \text{uncertain,} \\ 12, & \text{not_eligible.} \end{cases} \tag{5}$$

The final insurance score is

$$\text{Insurance} = \max\left(0,\ \text{InsuranceItem} - \pi_{\text{ins}}\right). \tag{6}$$

This formulation makes the metric sensitive to explicit reimbursement violations while preventing out-of-scope supportive medications from disproportionately lowering the score.

D.7. Overall: Reimbursement-Aware Alignment

The **Overall** metric augments auxiliary alignment with reimbursement status.

For each matched predicted medication, we assign a reimbursement-aware item credit according to both match type and reimbursement status:

$$\omega = \begin{cases} 1.00, & \text{identical} \ \& \ \text{eligible}, \\ 0.90, & \text{identical} \ \& \ \text{uncertain}, \\ 0.60, & \text{identical} \ \& \ \text{not_eligible}, \\ 1.00, & \text{identical} \ \& \ \text{excluded}, \\ 0.70, & \text{equivalent} \ \& \ \text{eligible}, \\ 0.60, & \text{equivalent} \ \& \ \text{uncertain}, \\ 0.30, & \text{equivalent} \ \& \ \text{not_eligible}, \\ 0.70, & \text{equivalent} \ \& \ \text{excluded}, \\ 0, & \text{additional}. \end{cases} \tag{7}$$

The key design choice is that **excluded** medications retain their original alignment credit: an identical excluded medication still receives 1.00, and an equivalent excluded medication still receives 0.70. Therefore, excluded medications are evaluated for alignment, but they do not incur reimbursement-specific penalties.

Let Ω denote the sum of these item credits over the predicted regimen:

$$\Omega = \sum \omega. \tag{8}$$

We then define reimbursement-aware coverage and precision as

$$\text{coverage}^* = \frac{\Omega}{G}, \qquad \text{precision}^* = \frac{\Omega}{P}. \tag{9}$$

The reimbursement-aware auxiliary base is

$$\text{OverallBase} = 100 \times \left(0.7 \times \text{coverage}^* + 0.3 \times \text{precision}^*\right). \tag{10}$$

Finally, we apply a regimen-level reimbursement penalty, again computed only from the scoreable core diabetes regimen:

$$\pi_{\text{overall}} = \begin{cases} 0, & \text{eligible}, \\ 2.5, & \text{uncertain}, \\ 8, & \text{not_eligible}. \end{cases} \tag{11}$$

The final overall score is

$$\text{Overall} = \max\left(0,\ \text{OverallBase} - \pi_{\text{overall}}\right). \tag{12}$$

This metric preserves the structure of alignment-based evaluation while selectively penalizing reimbursement problems only when they arise in scoreable core diabetes prescribing.

D.8. Relation to Main Table Columns

Under both **Total Prescription** and **Reimbursed Only**, the reported columns correspond to:

- **Exact Match**: the strict alignment score in Eq. (1),
- **Insurance**: the reimbursement-compliance score in Eq. (6),
- **Overall**: the reimbursement-aware alignment score in Eq. (12).

In summary, **Exact Match** measures fidelity to the physician reference regimen, **Insurance** measures reimbursement compliance for scoreable diabetes medications only, and **Overall** integrates the two while preventing out-of-scope adjunctive drugs from dominating the final score.

E. Detailed Experimental Setup

We evaluate model performance on the EMR-derived diabetes prescribing benchmark using a unified agentic evaluation pipeline. The evaluated systems span general-purpose frontier models, open-weight reasoning models, and domain-specialized medical models: GPT-5 (Singh et al., 2025), GPT-5.4 (OpenAI, 2026), Claude-Haiku-4.5 (Anthropic, 2025), Qwen3-14B (Yang et al., 2025), Gemma-4-31B-IT (Farabet & Lacombe, 2026), GLM-4.7-Flash (Zeng et al., 2025), Kimi K2.5 (Team et al., 2026), MedResearcher-R1-32B (Yu et al., 2025), HuatuoGPT-o1-72B (Chen et al., 2024), and Meerkat-14B (Kim et al., 2025). This selection allows comparison across model scale, openness, and degree of medical specialization.

All models are prompted with the same system and user template and are evaluated with the same tool interface, output post-processing, medication parsing, and scoring functions. This controlled setup minimizes evaluation variance from prompt format or parser differences and isolates model behavior on the prescribing task itself.

For generation, we adopt a best-performance decoding strategy for open-source models, using temperature 0.6 and top_p 0.95, following the recommendations in their technical reports when applicable. GPT-5 and GPT-5.4 are evaluated with temperature 1.0 due to API constraints. For agent execution, we set the maximum number of turns to 10 to avoid excessive rollout length and timeout-related failures. For retrieval-based evidence access, we use top_k=10 documents for each search step.

F. Case Study

We highlight failure modes of strong baselines that ignore patient preferences (e.g., injection aversion) or reimbursement feasibility, despite being clinically plausible. Table 12 to 14 shows one-to-one comparison among various models. Our agent better aligns to real-world constraints by integrating expert planning and retrieved evidence. We additionally provide the original Korean snippets alongside English translations for multilingual cases, allowing readers to verify both the translated content and the source-language wording.

Case A: Avoiding Over-expansion (Table 12). For a patient with well-controlled HbA1c (6.1%), Meerkat Base proposes unnecessary additions of a GLP-1 receptor agonist and rapid-acting insulin. In contrast, GPT-5 maintains the baseline regimen with minimal drift, avoiding excessive treatment when clinical targets are already met. This case exemplifies the "avoid over-expansion" principle rewarded by the judge model for structural fidelity.

Case B: Minimizing Regimen Drift (Table 13). When faced with uncontrolled glucose (HbA1c 8.2%), GPT-5 introduces excessive simultaneous changes across multiple medication axes, including drug discontinuation and escalation, leading to high regimen drift. However, Meerkat SFT+RL preserves the core treatment skeleton and performs limited, targeted adjustments. This structure-preserving approach significantly reduces the mismatch burden compared to the expert gold standard.

Case C: Conservative vs. Radical Adjustment (Table 14). For a nearly-controlled patient (HbA1c 6.5%), the base model suggests radical drug replacement, whereas Meerkat SFT+RL opts for conservative de-intensification through SU (gliclazide) dose reduction. This learned clinical intuition allows the model to remain structurally close to expert practice while effectively managing overtreatment risks.

G. Additional Analysis with Search and Corpus

To better understand the role of retrieval, we analyze search traces along three axes: whether a model invokes search, how intensively it searches once triggered, and how search calls are distributed across corpora (Table 18, Table 19, Figure 3). Four observations emerge.

First, search volume alone is a poor proxy for prescribing quality. Models such as MedResearcher-R1-32B and Qwen3-14B search in most cases (89.3% and 88.6% usage rate, respectively), yet they do not achieve the strongest Insurance or Overall scores. In contrast, GPT-5 and Meerkat-14B use search more selectively (26.4% and 7.1%), while remaining competitive in Insurance and substantially stronger in Overall. This suggests that effective retrieval depends more on calibrated triggering and evidence use than on raw search frequency (Asai et al., 2023; Jeong et al., 2024b; Sohn et al., 2025).

Second, within the Meerkat family, planning and RL improve search efficiency rather than simply increasing search frequency. Search usage remains low across *SFT without Planning*, *SFT with Planning*, and *SFT with Planning + RL* (15.0%, 12.9%, and 11.4%), but Exact Match and Overall improve steadily, with *SFT with Planning + RL* achieving the best performance under both evaluation scopes. Thus, the main benefit of planning and RL appears to be better search allocation and integration, not more search (Yao et al., 2022; Schick et al., 2023).

Third, the corpus distribution shifts in a clinically meaningful way. *SFT without Planning* is dominated by Insurance retrieval (79%), which is consistent with its strong reimbursement compliance but weaker Overall score. With planning, retrieval becomes more balanced across Insurance, Guideline, and Drug Info (47%, 28%, and 22%). With the improved RL reward, however, retrieval becomes more targeted again, with Insurance regaining the largest share (65%) while Guideline and Drug Info remain non-trivial (19% and 16%). Together with the lower overall search usage of the RL model, this suggests that RL makes retrieval more selective and more focused on reimbursement-critical cases rather than simply broadening search across all corpora (Jeong et al., 2024a; Xiong et al., 2024).

Fourth, corpus utility is driven more by task relevance than by corpus size. UpToDate is by far the largest corpus in our setup ($9{,}681$ chunks), yet it accounts for only a small fraction of search calls across models. By contrast, Insurance contains only 10 chunks but occupies a large share of retrieval, especially for the Meerkat variants. Together, these results indicate that retrieval in this benchmark is governed primarily by the decision-critical information required for reimbursement-aware prescribing rather than by corpus scale alone (Jeong et al., 2024a; Xiong et al., 2024).

A complementary single-corpus ablation in Table 17 supports the corpus-allocation analysis, although it should be interpreted cautiously because the ablated runs use a specialized prompt and are not strictly matched to the full-corpus baseline. Even so, several patterns are clear: removing Insurance most directly hurts reimbursement-related performance, removing Drug Info lowers Exact Match and Overall in the RL model, and removing UpToDate degrades Meerkat-14B despite its small retrieval share, suggesting that a small number of UpToDate lookups can provide high-value evidence. By contrast, Guideline ablation has limited impact after RL, which may indicate partial internalization of guideline knowledge or reduced retrieval noise. These ablation trends are broadly consistent with the updated retrieval pattern of the RL model, which places greater weight on Insurance while still relying on complementary external evidence.

Overall, these findings suggest that the gains from planning and RL come less from making the model search more often and more from making it search more selectively, across more appropriate evidence sources, and with better integration of retrieved evidence into final prescribing decisions. The corpus-ablation results further indicate that these gains are supported by complementary evidence sources with different roles: Insurance for reimbursement verification, Drug Info for medication-level precision and safety, and a small number of high-value UpToDate retrievals for difficult cases.

H. HealthBench-Diabetes Subset Construction and Evaluation

To assess diabetes-specific medical capability in a broader open-ended benchmark setting, we constructed a diabetes-focused subset from the original HealthBench benchmark (Arora et al., 2025).

Subset construction. We intentionally based subset construction on the original `prompt` field only, without using the rubric or ideal completion, in order to avoid selecting examples using information from the reference side. We first applied a broad keyword-based filter to collect candidate prompts related to diabetes. Keywords included terms such as `diabetes`, `diabetic`, `type 1 diabetes`, `type 2 diabetes`, `T1D`, `T2D`, `diabetes mellitus`, `hyperglycemia`, `hypoglycemia`, `A1c`, `HbA1c`, `blood sugar`, `glucose`, `insulin`, `metformin`, and `CGM`.

This first-stage filter was designed to maximize recall and produced 235 candidate prompts from the original 5,000-example benchmark.

In the second stage, we used GPT-4.1 (OpenAI, 2025) to classify each candidate into one of three categories: `primary_diabetes`, `diabetes_comorbidity_only`, or `not_diabetes`. Here, `primary_diabetes` indicates that the main user request is directly about diabetes diagnosis, glucose management, diabetes medications, insulin use, HbA1c control, hypoglycemia/hyperglycemia, or major diabetes complications. `diabetes_comorbidity_only` indicates that diabetes appears only as background history or a secondary condition while the main question is about another topic. `not_diabetes` indicates that diabetes is not a primary topic of the request. Borderline cases were manually reviewed, and only `primary_diabetes` examples were retained in the final subset.

Inclusion and exclusion criteria. We included prompts where diabetes itself was the central topic, including diagnosis, glycemic control, insulin adjustment, oral antidiabetic drugs such as metformin, diet and exercise guidance, hypoglycemia management, and diabetes-related complications such as diabetic ketoacidosis, diabetic retinopathy, diabetic neuropathy, and diabetic foot. We excluded prompts where diabetes was mentioned only briefly as past medical history or comorbidity, as well as broad metabolic-disease questions in which diabetes was not the core issue. Borderline categories such as gestational diabetes were reviewed explicitly and handled consistently.

Final subset. After classification and manual review, the final HealthBench-Diabetes subset consisted of 135 prompts in the full subset and 31 prompts in the hard subset.

Evaluation procedure. We saved the selected prompt IDs as a separate subset definition, ran the standard HealthBench generation and judge pipeline on the benchmark, and then computed subset performance by filtering judged outputs to the selected prompt IDs only. We report the overall benchmark score for this subset, consistent with the original HealthBench evaluation protocol. Unless otherwise noted, the judge model follows the original setup and uses GPT-4.1.

Generation length. Because medical responses in this subset are often longer and more structured than general benchmark answers, we increased the generation budget from 2,048 to 8,196 tokens during HealthBench evaluation.

This subset score is intended as a supplementary analysis alongside the original HealthBench full-benchmark results, rather than as a replacement for the official benchmark score.

Expert-level Planning: Step-by-step clinical reasoning and workflow following 7 tasks

1. Baseline & Risk Profiling
Core Indicators: Current age, age at onset, and DM duration.

Analysis Logic:
- Phenotype Classification: Differentiate between insulin deficiency-dominant vs. resistance-dominant profiles based on duration and onset age.
- Beta-cell Function Estimation: Calculate the likelihood of endogenous insulin secretory capacity decline (beta-cell exhaustion) according to disease progression.
- Risk Assessment: Predict the difficulty of glycemic control and the future risk of complications.

2. Complication & Comorbidity Mapping
Microvascular: Check for nephropathy (proteinuria), retinopathy, neuropathy, and diabetic foot.
Macrovascular: Review history of cardiovascular (CAD) and cerebrovascular (Stroke) diseases.
Special Contexts: Identify CKD, Liver disease (Cirrhosis/NAFLD), Malignancy, and conditions requiring steroids (as they cause glucose variability).

3. Physical & Lifestyle Assessment
Clinical Metrics: Blood pressure (BP), Body Weight, and BMI.
Trend Analysis: Compare weight changes against the previous visit and evaluate progress toward target weight.
Lifestyle Screening: Assess smoking/alcohol status and adherence to exercise/dietary regimens.

4. Biochemical Target Evaluation
Lab Items: HbA1c, FPG, LFT (Liver function), eGFR/Cr (Renal function), and Lipid profile.

Analysis Logic:
- Target vs. Actual: Check if each metric is within the individualized target range.
- Trend Analysis: Determine improvement or worsening by comparing with previous lab results.

5. Prescription Optimization & Regulatory Review
Adequacy Check: Assess if current medications are sufficient to reach the HbA1c goal.
Insulin Fine-tuning: Adjust insulin type, frequency, and dosage for precision.
Safety & Multi-drug Review: Monitor side effects and drug-drug interactions (especially for 3+ agents).
Benefit Assessment: Evaluate agents for weight gain, hypoglycemia risk, and ASCVD/HF/CKD benefits.
Regulatory Compliance: Filter for K-NHIS (Korean insurance) reimbursement criteria and drug pricing.

6. Monitoring & Follow-up Planning
Cycle Management:
- Ensure routine tests (HbA1c, Admission panel, Lipid panel) at every visit.
- Annual Screening: Check for missing annual complications screens (UACR, Fundoscopy).
Additional Actions: Reinforce preventive measures (Anti-platelets, Statins) and refer for specialty exams if needed.

7. Next Appointment Scheduling
Stable Group (4+ months): Targets met (A1c, BP, weight) with stable low-dose oral medications.
Unstable Group ($\leq$3 months): Targets not met, high glucose variability, or frequent medication adjustments.

Only include steps relevant to the specified subtask.

Table 5. Expert-level seven-step planning prompt. This prompt operationalizes a clinician-authored diabetes prescribing workflow derived from real outpatient practice and was used to support both planning distillation during training and structured reasoning during inference.

System Prompt Used for SFT Training Dataset Generation

```
You are an expert endocrinologist specializing in diabetes care in Korea. Your task
is to generate realistic and technically grounded **task-oriented requests** that
practicing diabetes clinicians (or clinical pharmacist collaborators) would plausibly
issue when working with the given patient case bundle (EMR-like structured data + lab
trends + medication history + comorbidities + insurance rules).

---

<instructions>
- The ultimate goal is to build a **colleague clinician LLM agent system**. Therefore,
we require queries that diabetes experts would naturally issue to a peer clinician
agent when seeking assistance during **prescription optimization, safety checks,
regulatory/insurance compliance, and follow-up planning**.
- Based on the provided '<case>' content (all data/tools are assumed available), your
role is to synthesize **query-plan-search_queries-result** quadruplets that reflect how
a diabetes expert would operationally proceed in real clinic.
- The type of subtask relevant to each query is specified in the '<task>' section
below. You must generate only queries and results that fall strictly within the
defined task scope. If no such query can be generated from the case, output only:
``NOT FOUND''.
- Each query must be an explicit instruction intended for execution by an agent system
(not a vague question and not a request ``about the case narrative'').
- For each generated query, describe the planning process the clinician would follow
to resolve it. The plan should be written as methodological guidelines (step-by-step),
incorporating practical clinical knowledge: what patient data is used, which criteria
are checked (e.g., A1c goal, CKD/ASCVD/HF benefit, hypoglycemia risk), what medication
changes are considered, and how monitoring/follow-up is determined.
- Directly following the plan, you must identify clinical or regulatory uncertainties
and resolve them by generating precise search queries. These queries serve as
a mandatory bridge to the final 'Result', prioritizing search_guideline (KDA)
and search_insurance (HIRA) to ensure gold-standard compliance, followed by
search_drug_info for safety audits and search_uptodate for specialized evidence.
Every query must be professional and highly specific (e.g., ``HIRA reimbursement for
SGLT2i/DPP4i dual therapy''), ensuring that the final recommendation is grounded in
verified medical and regulatory facts. Crucially, while all other sections must be in
English, the 'search_queries' string must be generated ONLY in Korean. When generating
these queries, do NOT mention specific years or specific organization names like
``KDA'' (e.g., use ``당뇨병 진료 지침'' or ``당뇨병 약제 가이드라인'' instead of ``2023 KDA
가이드라인'') to maintain general and up-to-date applicability.
- A corresponding ground-truth answer or outcome must be derivable from the provided
'<case>' content (labs, diagnoses, medication list, events, insurance eligibility,
etc.). In other words, the query must be answerable using information that is actually
present in the case bundle and would have been charted or decided in the scenario.
- Do not introduce new clinical facts not contained in '<case>'. If uncertainty
remains due to missing data, the proper result is to explicitly flag the missing
variables and output the conservative/standard-of-care safe action consistent with
the task constraints.
</instructions>

<output_format>
### Query
[Describe the instruction that a clinician would issue to the agent]

### Plan
[Step-by-step clinical reasoning and workflow required to address the query,
addressing the following 7 tasks:
...

### Search Queries
[Generate appropriate search queries in KOREAN by selecting the necessary tools based
on the following descriptions. Do not include specific years or specific organization
names like ``KDA'' in the queries. Each query must be wrapped in the designated
format:
```

```
- search_guideline: Use this tool to retrieve the latest clinical practice guidelines
 from the Korean Diabetes Association (KDA). Focus on treatment algorithms, target
 HbA1c levels, and therapy escalation/de-escalation rules.
- search_insurance: Use this tool to verify the Health Insurance Review and Assessment
 Service (HIRA) reimbursement criteria. Essential for checking drug combination
 restrictions, A1c thresholds for specific medications, and required documentation for
 coverage.
- search_drug_info: Use this tool to check pharmacological details such as standard
 dosage, contraindications based on renal/hepatic function (e.g., eGFR cutoffs),
 drug-drug interactions, and common side effects.
- search_uptodate: Use this tool to find the most recent evidence-based clinical
 summaries, international trial results, or expert opinions for complex cases that are
 not fully covered by standard guidelines.]

 ### Result
 [Ground-truth decision/outcome strictly derivable from '<case>':
- finalized medication adjustment (drug/class, dose, frequency, start/stop/switch)
- safety constraints (contraindications, hypoglycemia risk mitigation)
- insurance/coverage eligibility conclusion and required documentation
- monitoring items and follow-up interval
- concise justification anchored to case data]
 </output_format>

 <task>
 ### Task
 Validation

 ### Subtask
 Prescribing & Safety Verification (Final Clinical QA)

 ### Goal
 Perform one last clinician-grade verification of the proposed diabetes treatment
 plan to ensure the regimen is clinically coherent, safe, evidence-aligned, and
 reimbursement-compliant, and that the follow-up plan matches patient risk.

 ### Description
 Task Definition: Manual forensic examination of the entire prescribing decision
 underlying key clinical outcomes. This task acts as a final quality assurance step
 to ensure the recommended therapy is supported by solid patient-specific evidence,
 avoiding unsafe escalation, guideline-incoherent combinations, or insurance-ineligible
 prescriptions.

 Core Components:
- Guideline Coherence Audit: Verify that the regimen aligns with individualized
 glycemic targets and evidence-based priorities (ASCVD/HF/CKD benefit, weight,
 hypoglycemia avoidance), and that drug-class selection is appropriate for patient
 phenotype (insulin deficiency vs resistance proxy, diabetes duration, age, frailty).
- Safety & Contraindication Check: Review kidney/liver function, hypoglycemia risk,
 drug interactions, dosing appropriateness (eGFR-based), and special situations
 (steroid use, malignancy, cirrhosis, pregnancy possibility if present).
- Therapeutic Duplication & Sequencing: Detect duplicate mechanisms (e.g., DPP-4 +
 GLP-1 RA redundancy), inappropriate triple/quad therapy without rationale, insulin
 titration logic errors, and missed cardiometabolic opportunities.
- Regulatory / Insurance Compliance: Confirm Korean reimbursement criteria are met
 (required prior therapies, A1c thresholds, eGFR cutoffs, combination restrictions),
 and identify required documentation/codes.
- Monitoring & Follow-up Consistency: Ensure monitoring schedule (A1c, CMP, renal
 panel, lipids, albuminuria, retinal exam, foot exam) and revisit interval match
 stability/instability status and recent medication changes.
 </task>
 <query_examples>
 The draft regimen escalates to GLP-1 RA + basal insulin. Perform a final safety and
```

```
duplication audit: confirm there is no contraindication (eGFR, pancreatitis history if
provided), check that no redundant DPP-4 inhibitor remains, and output the corrected
final medication list with monitoring plan.
Validate the SGLT2 inhibitor initiation: verify eGFR eligibility, evaluate DKA
risk factors if present, confirm HF/CKD/ASCVD benefit fit, and generate the exact
insurance-justification statement supported by the case data.
The plan increases sulfonylurea dose despite recurrent hypoglycemia in logs. Audit
hypoglycemia risk and propose a safer alternative adjustment consistent with the
patient's targets and comorbidities.
</query_examples>
```

```
Using the '<instructions>', '<output_format>', '<task>', and '<query_examples>'
above as guidance, synthesize query-answer pairs based on the '<case>' provided in
the user message.
```

Table 6. System prompt used for planning-based SFT data generation from diabetes cases. For each case, it produces a structured `Query-Plan-Search Queries-Result` tuple, where the `Plan` component follows the seven-step workflow shown in Table 5.

Planning System Prompt Used for Inference

[Original]

```
## Internal Planning Guidance (SILENT)
```

아래 단계를 내부적으로 짧게 점검하십시오. 이 계획을 장황하게 출력할 필요는 없습니다.

```
1. Baseline & Risk Profiling
Core Indicators: Current age, age at onset, and DM duration.
...
```

[English Translation]

```
## Internal Planning Guidance (SILENT)

Briefly review the following steps internally. There is no need to output this plan in
a verbose manner.

1. Baseline & Risk Profiling
Core Indicators: Current age, age at onset, and DM duration.
...
```

Table 7. Representative excerpt of the inference-time planning guidance. This component reuses the same seven-step clinician-authored diabetes prescribing workflow shown in Table 5, but provides it as silent internal guidance for structured reasoning during inference.

Tool-Use System Prompt Used for Inference

```
# Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools></tools> XML tags:
<tools>
{"name":"search-search_guideline", ...}
{"name":"search-search_insurance", ...}
...
</tools>

For each tool call, return a json object with its name and arguments within
<tool_call></tool_call> XML tags:
<tool_call>
{"name": "search-search_guideline", "arguments": {"query": "...", "topk": 4}}
</tool_call>
```

Table 8. Representative excerpt of the inference-time tool-use instructions. This component specifies the available retrieval tools and the required XML-wrapped JSON format for structured tool invocation.

Search and Reasoning Policy Used for Inference

[Original]

```
## Search & Reasoning Policy (STRICT)

### STEP 1: 검색 필요성 판단 (Search Decision) -- ALWAYS DO THIS FIRST
매 환자 질의를 받으면, 반드시 먼저 <think> 태그 안에서 다음을 판단하십시오:

검색 없이 답변 가능한 경우 (아래 조건을 모두 만족):
- (1) 환자의 현재 약물이 명확하고 금기사항이 없음
- (2) 처방 변경이 표준 임상 원칙으로 충분히 판단 가능함
- (3) 급여기준/보험 확인이 불필요하거나 명확함
- (4) 특수한 신기능/간기능 용량조절이 필요하지 않음
→ 이 경우 </think> 후 바로 <answer> 작성

검색이 필요한 경우 (아래 중 하나라도 해당):
- 급여/보험 기준 확인이 필요한 경우
- 약물 허가사항/신기능 용량조절 확인이 필요한 경우
- 최신 가이드라인 우선순위 확인이 필요한 경우
- 본인의 전문 지식에 확신이 낮거나 불확실한 경우
- 환자의 복잡한 동반질환/합병증으로 인해 추가 근거가 필요한 경우
→ 이 경우 </think> 후 <tool_call> 출력

### STEP 2: Tool 사용 및 반복 검색
- Evidence Quality: 검색을 수행한 경우, 근거 공백이 남으면 단순 추측으로 메우지 말고 검색 전략을 고도화하십시오.
- Iterative Search: 1차 검색 후 환자의 특정 동반 질환이나 상충하는 지표가 발견되면, 필요 시 2차 심층 검색을 통해 해당 상황에서의 적정성을 재검증하십시오.
- Tool Priority: 권장 우선순위는 search-search_guideline → search-search_insurance → search-search_drug_info → search-search_uptodate 이다. search-query_rag는 환자 EMR 또는 내부 문서 재확인이 필요할 때 선택적으로 사용한다.

### STEP 3: 답변 작성
- Reasoned Extrapolation: 모든 반복 검색 시도 후에도 직접적인 근거가 부족한 경우, 전문의로서 약리학적 기전과 임상 원칙에 기반하여 논리적으로 추론하십시오. 단, 추론 과정에서는 근거의 한계와 불확실성을 명확히 인식하고, 최종 답변에서도 필요한 경우 이를 분명히 드러내십시오.
- Self-Check: 최종 답변 생성 전, 결론이 환자의 현재 상태 및 주요 안전성 이슈와 충돌하지 않는지 비판적으로 재검토하십시오.
```

[English Translation]

```
## Search & Reasoning Policy (STRICT)

### STEP 1: Determine Whether Search Is Needed (Search Decision) -- ALWAYS DO THIS
FIRST
For every patient query, first assess the following within the <think> tag:

Cases where an answer can be provided without retrieval (all of the following must
hold):
- (1) The patient's current medications are clearly identified and there are no
contraindications.
- (2) The prescription change can be adequately determined based on standard clinical
principles.
- (3) Reimbursement/insurance verification is unnecessary or already clear.
- (4) No special renal or hepatic dose adjustment is required.
→ In this case, write <answer> immediately after </think>.

Cases where retrieval is required (if any of the following applies):
- Reimbursement or insurance criteria must be verified.
- Drug label information or renal dose adjustment must be checked.
- The priority or preference in the latest guideline must be confirmed.
- Confidence in one's own domain knowledge is low or uncertainty remains.
```

```
- Additional evidence is needed due to complex comorbidities or complications.
 → In this case, output <tool_call> after </think>.

 ### STEP 2: Tool Use and Iterative Search
- Evidence Quality: If search is performed and evidence gaps remain, do not fill them
 with unsupported guesses; instead, refine the search strategy.
- Iterative Search: If a first-round search reveals a specific comorbidity or
 conflicting clinical indicators, conduct a second-round targeted search when needed
 to re-evaluate appropriateness under that condition.
- Tool Priority: The recommended priority is search-search_guideline →
 search-search_insurance → search-search_drug_info → search-search_uptodate. Use
 search-query_rag selectively when re-checking the patient EMR or internal documents
 is necessary.

 ### STEP 3: Write the Final Answer
- Reasoned Extrapolation: If direct evidence remains insufficient even after iterative
 search attempts, reason logically as a specialist based on pharmacologic mechanisms
 and clinical principles. However, clearly recognize the limitations and uncertainty
 of the evidence during reasoning, and explicitly state them in the final answer when
 appropriate.
- Self-Check: Before generating the final answer, critically review whether the
 conclusion conflicts with the patient's current condition or major safety issues.
```

Table 9. Representative excerpt of the inference-time search-and-reasoning policy. This component instructs the model to determine whether retrieval is necessary, prioritize tools by evidence type, and perform iterative follow-up searches when evidence is needed.

Final Inference Output Template

[Original]

```
<answer>
## 환자 상태 분석
[Long-form patient assessment in Korean prose]

## 처방 제안

### 1. 급여 기준 충족 약물 처방
| 변경/중단 | 신규 | 유지 | 비고 |
|---|---|---|---|
| ... | ... | ... | ... |

**처방 근거:**
- 가이드라인(KDA): ...
- 급여기준(보험): ...
- 약물정보(Drug info): ...
- UpToDate: ...

### 2. 비급여 기준 고려 약물 처방
| 변경/중단 | 신규 | 유지 | 비고 |
|---|---|---|---|
| ... | ... | ... | ... |

**처방 근거:**
- 가이드라인(KDA): ...
- 급여기준(보험): ...
- 약물정보(Drug info): ...
- UpToDate: ...

### 3. 기타 추가적인 약물 처방 제안
| 변경/중단 | 신규 | 유지 | 비고 |
|---|---|---|---|
| ... | ... | ... | ... |

**처방 근거:**
- 가이드라인(KDA): ...
- 급여기준(보험): ...
- 약물정보(Drug info): ...
- UpToDate: ...
</answer>
```

[English Translation]

```
<answer>
## Patient Assessment
[Long-form patient assessment in Korean prose]

## Prescription Recommendation

### 1. Prescription for Medications Meeting Reimbursement Criteria
| Change/Discontinue | New | Continue | Notes |
|---|---|---|---|
| ... | ... | ... | ... |

**Rationale for Prescription:**
- Guideline (KDA): ...
- Reimbursement Criteria (Insurance): ...
- Drug Information: ...
- UpToDate: ...
```

```
### 2. Prescription for Medications Considered Outside Reimbursement Criteria
| Change/Discontinue | New | Continue | Notes |
|---|---|---|---|
| ... | ... | ... | ... |

**Rationale for Prescription:**
- Guideline (KDA): ...
- Reimbursement Criteria (Insurance): ...
- Drug Information: ...
- UpToDate: ...

### 3. Additional Medication Suggestions
| Change/Discontinue | New | Continue | Notes |
|---|---|---|---|
| ... | ... | ... | ... |

**Rationale for Prescription:**
- Guideline (KDA): ...
- Reimbursement Criteria (Insurance): ...
- Drug Information: ...
- UpToDate: ...
</answer>
```

Table 10. Final inference output template specifies the required clinician-facing answer structure, including patient assessment, prescription sections, and evidence-based justification fields.

Table 11. Composition of the expert-curated hard subset. Counts are category-level expert-verified labels, and one patient case may be assigned to multiple difficulty categories. The final hard subset contains 40 derived test instances after de-duplicating patient cases and applying the same instance construction protocol as the main benchmark.

Difficulty category	**# of cases**
Advanced CKD with severe hypoglycemia history	5
Frailty with HFrEF	1
Decompensated cirrhosis	2
LADA	2
Hypoglycemia unawareness	4
Obesity with gastroparesis and ASCVD	3
Periodic long-term high-dose steroid use	1
HIV with complex concomitant medications	1
Severe injection phobia	2
Economic constraints	1
Total multi-label assignments	22
Final derived hard-set instances after de-duplication	40

Case Study A : GPT-5 vs Meerkat Base

EMR Snapshot

Item	Details
Disease history	Long-standing diabetes (diagnosed in 1991; approximately 34 years).
Recent status	HbA1c 6.1%; fasting glucose around 105 mg/dL; postprandial excursions reported up to 230 mg/dL.
Renal / risk profile	CKD stage 3a context (eGFR around 45.7), albuminuria noted in model analysis context; low body weight/BMI profile.
Baseline regimen	Diabex XR 500 mg BID, Tenelia 20 mg QD, Mevalotin 20 mg (0.5 tab) QD.

Group	GPT-5	Meerkat Base	Δ (GPT−Base)
GPT higher	67.75	22.12	+45.63

Highlight legend: better-aligned medication handling , mismatch / unnecessary expansion , major regimen change .

Winner	GPT-5 Output (EN/KO)	Meerkat Base Output (EN/KO)
GPT-5	*EN: " Maintain Metformin XR 500 mg BID , Teneligliptin 20 mg QD , and continue Pravastatin ; no add-on expansion."* *KO: " Metformin XR 500 mg BID 유지 , Teneligliptin 20 mg QD 유지 , Pravastatin 지속 ; 추가 약제 확대 없음."*	*EN: " Maintain Metformin XR 500 mg BID and Teneligliptin 20 mg QD , but additionally propose a GLP-1 receptor agonist and rapid-acting insulin despite an already controlled A1c profile."* *KO: " Metformin XR 500 mg BID 유지 및 Teneligliptin 20 mg QD 유지 , 이미 조절된 A1c 프로파일에도 불구하고 GLP-1 receptor agonist 추가와 rapid-acting insulin 추가 를 제안."*

Medication-centric difference	GPT-5 preserves the baseline 3-drug backbone with minimal drift. In contrast, Meerkat Base introduces extra treatment branches (GLP-1 and rapid insulin) and ends with higher mismatch burden, which is less consistent with the gold target structure for this EMR.
Interpretation of score gap	Because the judge rewards structural fidelity and penalizes unnecessary additions, GPT-5 gains a large margin (+45.63). This case is representative of "avoid over-expansion when baseline control is already acceptable."

Table 12. Patient case where GPT-5 outperforms Meerkat Base by preserving the baseline regimen and minimizing unnecessary expansion.

Case Study B : GPT-5 vs Meerkat SFT+RL

EMR Snapshot

Item	Details
Disease history	Long diabetes duration (diagnosed in 2003; approximately 22 years).
Recent status	HbA1c 8.2%, fasting glucose around 175 mg/dL (control not at target).
Renal / risk profile	eGFR around 63.2 (CKD 3a context in model analysis), with cardiometabolic risk requiring careful titration rather than broad expansion.
Baseline regimen	Lantus 40U QD, Metformin 850 mg BID, Amaryl 4 mg QD, Atorvastatin 10 mg QD, plus GI meds (Lanston, Gasmotin).

Group	GPT-5	Meerkat SFT+RL	Δ (GPT$-$SFT+RL)
SFT+RL higher	20.62	74.22	-53.60

Highlight legend: *better-aligned medication handling* , *mismatch / unnecessary expansion* , *major regimen change* .

Winner	GPT-5 Output (EN/KO)	Meerkat SFT+RL Output (EN/KO)
Meerkat SFT+RL	*EN: " Discontinue glimepiride , increase insulin glargine to 46 IU , add empagliflozin , and further add losartan / escalate atorvastatin ."* *KO: " Glimepiride 중단 , insulin glargine 46 IU로 증량 , empagliflozin 추가 , losartan 추가 , atorvastatin 증량 "*	*EN: " Maintain the core oral plus basal insulin regimen (metformin, glimepiride, basal insulin, atorvastatin) with a limited add-on adjustment rather than broad multi-axis expansion."* *KO: " 기존 경구혈당강하제 및 기저인슐린 기반 치료 구조 유지 (metformin, glimepiride, basal insulin, atorvastatin)하되, 제한적 추가 조정 으로 대응하며 광범위한 다축 확장은 지양."*

Medication-centric difference	In this case, GPT-5 introduces many simultaneous regimen changes (higher drift), while SFT+RL remains closer to the target treatment skeleton and introduces fewer mismatched additions. Judge-side mismatch counts reflect this contrast: GPT-5 shows very high omission/additional burden, whereas SFT+RL remains substantially lower.
Interpretation of score gap	The large negative delta for GPT-5 (-53.60) is primarily explained by over-expansion versus structure-preserving adjustment. This is a typical failure mode when aggressive optimization introduces too many extra axes at once.

Table 13. Patient case where Meerkat SFT+RL outperforms GPT-5 through lower regimen drift and better structure preservation.

Case Study C : Meerkat SFT+RL vs Meerkat Base

EMR Snapshot

Item	Details
Disease history	Type 2 diabetes with long disease duration (approximately 25 years).
Recent status	HbA1c around 6.5% with mild upward trend; fasting SMBG generally in the 80s–90s; no strong hypoglycemia signal in narrative.
Renal / metabolic profile	eGFR reported as normal range in model analysis context; relatively lean body habitus (BMI around low 20s).
Baseline regimen	Glupa 1 g BID, Diamicron MR 60 mg QD, Mevalotin 20 mg QD.

Group	Meerkat SFT+RL	Meerkat Base	Δ (SFT+RL−Base)
SFT+RL higher	79.00	32.07	+46.93

Highlight legend: *better-aligned medication handling* , *mismatch / unnecessary expansion* , *major regimen change* .

Winner	Meerkat SFT+RL Output (EN/KO)	Meerkat Base Output (EN/KO)
Meerkat SFT+RL	*EN: " Maintain metformin and pravastatin , and de-intensify gliclazide (from 60 mg to half dose) to reduce overtreatment risk while preserving the regimen structure."* *KO: " metformin 유지 , pravastatin 유지 , gliclazide는 60 mg에서 1/2 용량으로 감량 하여 과치료 위험을 낮추면서 기존 치료 구조를 유지."*	*EN: " Discontinue gliclazide and add semaglutide , while maintaining metformin and pravastatin."* *KO: " gliclazide 중단 , semaglutide 추가 , metformin 및 pravastatin 유지 ."*

Medication-centric difference	SFT+RL preserves the 3-drug baseline backbone and performs a conservative SU de-intensification, which remains structurally close to the target. In contrast, Base removes a core SU component and introduces a new GLP-1 branch, increasing mismatch.
Interpretation of score gap	The +46.93 gap shows the benefit of conservative, structure-preserving adjustment in a relatively controlled patient profile. This case illustrates improvement from base Meerkat to SFT+RL after training.

Table 14. Patient case where Meerkat SFT+RL outperforms Meerkat Base via conservative de-intensification and stronger structure fidelity.

Table 15. Model performance under two evaluation scopes: **Total Prescription** (reimbursed, non-reimbursed, and additional medications) and **Reimbursed Only** (medications subject to reimbursement criteria only). Target dataset is original 140 patient cases. Score metric is same as Table 1. Values are reported as mean ± standard deviation across patient cases. Judge model is Qwen3.5-122B.

	Total Prescription Original (140)			Reimbursed Only Original (140)		
Model	**Exact Match**	**Insurance**	**Overall**	**Exact Match**	**Insurance**	**Overall**
Proprietary LLMs						
GPT-5	4.31 ± 15.6	69.69 ± 25.0	41.72 ± 22.4	4.54 ± 16.4	73.28 ± 25.2	44.29 ± 23.6
GPT-5.4	5.09 ± 17.7	71.07 ± 16.6	51.90 ± 15.7	5.53 ± 19.9	**81.73** ± 15.1	**55.55** ± 15.6
Claude-Haiku-4.5	3.81 ± 14.5	68.15 ± 25.6	43.90 ± 21.6	3.81 ± 14.0	72.37 ± 26.1	46.53 ± 22.0
Open-source LLMs						
Qwen3-14B	7.51 ± 21.4	61.23 ± 31.2	40.65 ± 25.0	7.54 ± 22.1	62.98 ± 31.8	41.12 ± 26.4
Gemma-4-31B-IT	7.22 ± 21.4	**78.35** ± 18.7	**53.24** ± 18.9	7.22 ± 21.5	78.63 ± 17.9	53.69 ± 19.0
GLM-4.7-Flash (31B)	7.30 ± 21.3	54.64 ± 37.5	38.09 ± 28.9	7.15 ± 21.7	55.49 ± 38.3	36.28 ± 29.3
Kimi K2.5 (1.1T)	2.07 ± 10.0	51.27 ± 39.4	33.98 ± 28.1	1.94 ± 10.1	53.48 ± 40.9	35.52 ± 29.1
Medical LLMs						
MedResearcher-R1-32B	1.63 ± 10.2	26.22 ± 38.3	17.64 ± 26.9	1.63 ± 10.2	26.53 ± 38.7	17.75 ± 27.1
HuatouGPT-o1-72B	8.64 ± 23.0	75.18 ± 21.4	49.53 ± 20.1	8.81 ± 23.4	76.92 ± 19.6	49.58 ± 20.6
Meerkat-14B	6.83 ± 20.4	67.41 ± 25.6	42.97 ± 21.4	6.29 ± 19.1	69.37 ± 23.3	41.48 ± 22.5
SFT without Planning	5.40 ± 17.9	73.92 ± 21.5	43.87 ± 22.0	4.79 ± 16.8	73.47 ± 22.1	41.79 ± 22.7
SFT with Planning	6.53 ± 18.8	71.70 ± 23.1	46.39 ± 21.2	6.42 ± 19.0	72.85 ± 23.5	45.41 ± 21.8
SFT with Planning + RL	**14.7** ± 28.5	73.14 ± 21.0	53.12 ± 22.2	**14.9** ± 29.1	73.72 ± 23.3	52.92 ± 23.1

Table 16. Model performance under two evaluation scopes: **Total Prescription Hard** (reimbursed, non-reimbursed, and additional medications) and **Reimbursed Only Hard** (medications subject to reimbursement criteria only). 40 hard subset is specifically collected following the rule of diabetes prescribing experts. Score metric is same as Table 1. Values are reported as mean ± standard deviation across patient cases. Judge model is Qwen3.5-122B.

	Total Prescription Hard (40)			Reimbursed Only Hard (40)		
Model	**Exact Match**	**Insurance**	**Overall**	**Exact Match**	**Insurance**	**Overall**
Proprietary LLMs						
GPT-5	3.48 ± 12.8	67.12 ± 28.8	31.91 ± 20.3	3.52 ± 12.9	68.65 ± 28.2	33.50 ± 20.4
GPT-5.4	3.92 ± 12.6	**79.13** ± 15.1	**49.43** ± 16.0	4.23 ± 13.6	**80.93** ± 11.8	**51.81** ± 15.9
Cluade-Haiku-4.5	0.00 ± 0.0	66.81 ± 24.8	41.00 ± 17.3	0.00 ± 0.0	69.88 ± 23.4	45.34 ± 18.9
Open-source LLMs						
Qwen3-14B	0.83 ± 5.3	52.75 ± 37.3	29.24 ± 24.2	1.33 ± 8.4	53.48 ± 38.8	29.12 ± 25.1
Gemma-4-31B-IT	6.13 ± 17.1	73.93 ± 10.8	45.94 ± 18.2	6.82 ± 19.2	80.53 ± 14.3	49.26 ± 20.2
GLM-4.7-Flash (31B)	2.20 ± 10.0	59.94 ± 33.9	35.79 ± 24.4	2.26 ± 10.3	59.53 ± 36.1	33.55 ± 26.2
Kimi K2.5 (1.1T)	0.00 ± 0.0	35.45 ± 40.2	18.57 ± 22.9	0.00 ± 0.0	35.64 ± 41.2	18.78 ± 22.9
Medical LLMs						
MedResearcher-R1-32B	0.83 ± 5.3	20.31 ± 33.5	10.87 ± 21.5	0.83 ± 5.3	20.39 ± 34.5	9.39 ± 20.2
HuatouGPT-o1-72B	7.04 ± 17.8	74.53 ± 17.8	44.23 ± 18.0	7.29 ± 18.4	74.06 ± 18.3	45.75 ± 18.3
Meerkat-14B	3.07 ± 11.5	70.10 ± 19.0	41.32 ± 19.6	3.22 ± 12.2	70.00 ± 15.9	41.65 ± 18.5
SFT without Planning	5.81 ± 16.6	71.42 ± 24.5	41.44 ± 21.7	4.87 ± 15.8	70.92 ± 25.3	39.49 ± 23.7
SFT with Planning	3.67 ± 14.0	63.32 ± 27.9	41.75 ± 21.0	2.21 ± 10.1	65.98 ± 28.2	41.35 ± 22.7
SFT with Planning + RL	**8.92** ± 19.1	74.75 ± 12.3	47.89 ± 16.0	**9.07** ± 19.4	75.39 ± 11.7	49.15 ± 14.5

Table 17. Performance difference across corpus ablation. Specialized prompt is used for ablating single corpus source. Alignment values are reported same as mean of Table 15 with original 140 cases. Judge model is Qwen3.5-122B.

	Total Prescription			**Reimbursed Only**		
Model	**Exact Match**	**Insurance**	**Overall**	**Exact Match**	**Insurance**	**Overall**
Meerkat-14B	6.83	67.41	42.97	6.29	69.37	41.48
- Guideline	6.37	71.41	44.17	6.09	69.91	42.10
- Insurance	5.58	68.29	44.39	5.50	65.54	42.12
- Drug Info	6.52	71.20	44.46	6.20	73.37	43.29
- UpToDate	5.34	70.76	41.86	5.01	68.35	39.97
Meerkat-14B SFT+RL	14.7	73.14	53.12	14.9	73.72	52.92
- Guideline	16.6	74.07	54.13	16.3	74.77	53.45
- Insurance	14.2	72.05	52.96	13.4	71.94	51.96
- Drug Info	13.3	71.71	52.01	13.0	73.62	51.78
- UpToDate	12.6	72.40	52.77	12.3	73.61	52.17

Table 18. Search analysis across models on 140 evaluation records. **Search-Used/Total** denotes the number of records with at least one search call over the total number of records, **Usage Rate** is its percentage, **# of Calls** is the total number of search calls, **Avg. Calls / Total** is the average number of search calls per record, and **Avg. Calls / Search** is the average number of search calls among records that used search.

Model	Search-Used/Total	Usage Rate (%)	# of Calls	Avg. Calls / Total	Avg. Calls / Search
GPT-5	37/140	26.4	85	0.607	2.297
GPT-5.4	124/140	88.6	410	2.929	3.306
Claude-Haiku-4.5	8/140	5.7	30	0.214	3.750
Qwen3-14B	124/140	88.6	564	4.029	4.548
Gemma-4-31B-IT	80/140	57.1	137	0.979	1.712
GLM-4.7-Flash (31B)	57/140	40.7	198	1.414	3.474
Kimi K2.5 (1.1T)	87/140	62.1	693	4.950	7.966
MedResearcher-R1-32B	125/140	89.3	1073	7.664	8.584
HuatuoGPT-o1-72B	21/140	15.0	41	0.293	1.952
Meerkat-14B	10/140	7.1	16	0.114	1.600
SFT without Planning	21/140	15.0	29	0.207	1.381
SFT with Planning	18/140	12.9	32	0.229	1.778
SFT with Planning + RL	16/140	11.4	31	0.221	1.938

Table 19. Composition of the search corpora used for retrieval. **# of Chunks** counts chunked entries, and **Avg. Tokens / Chunk** reports the mean chunk length measured with the `cl100k_base` tokenizer.

Source	# of Chunks	Avg. Tokens / Chunk
Guideline	1593	415.5
Insurance	10	397.8
Drug Info	4403	473.9
UpToDate	9681	235.6

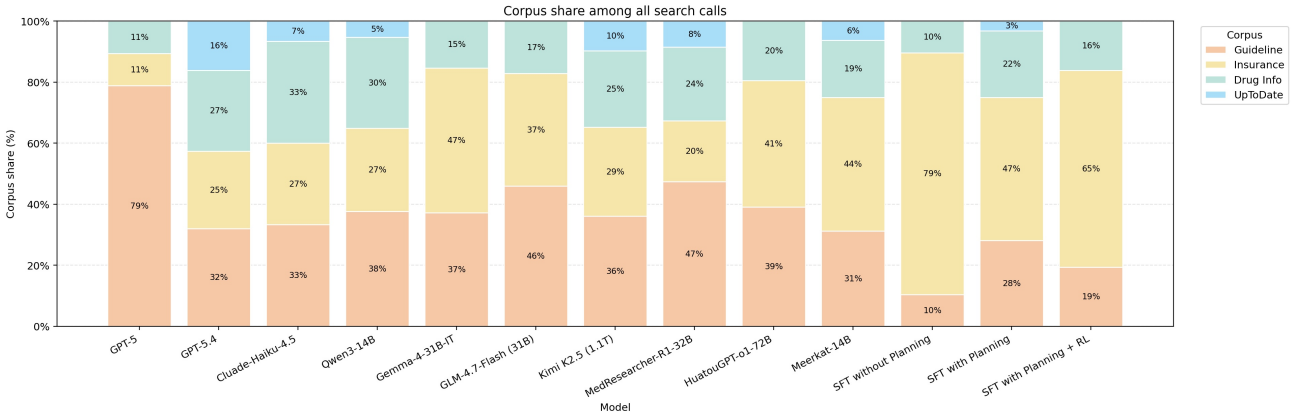

Figure 3. Corpus share across models. Each stacked bar represents the percentage breakdown of total search calls across corpora for a given model, with segment labels indicating the corresponding share (%).