PHARMEVAL: A DIAGNOSTIC BENCHMARK FOR PHARMACEUTICAL KNOWLEDGE AND REASONING OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing pharmaceutical LLM benchmarks primarily focus on accuracy, making it difficult to determine whether model failures arise from insufficient knowledge, ineffective information use, or unstable reasoning. We introduce `PharmEval`, a pharmaceutical benchmark comprising 6,174 questions across 9 task families and 47 subcategories. Beyond conventional accuracy-based evaluation, `PharmEval` assesses explanation quality and incorporates four diagnostics of pharmaceutical knowledge use, including web-based knowledge retrieval, compositional knowledge use, counterfactual sensitivity, and response modality variation. These studies examine how external information and explicit premises support knowledge use, and whether models maintain or correctly update their judgments as conditions and response formats change. We evaluate a diverse set of LLMs and conduct an in-depth analysis of representative models. Our results show that benchmark accuracy alone does not fully reflect how reliably models use pharmaceutical knowledge. High accuracy can coexist with weak open-ended performance and failures to update judgments correctly when key conditions change. These findings suggest that a key challenge for current LLMs lies not only in acquiring relevant knowledge, but also in applying it consistently and reliably in context. By evaluating models from both performance and behavioral perspectives, `PharmEval` provides a more nuanced characterization of their capabilities and limitations in pharmaceutical reasoning.

1 INTRODUCTION

Pharmaceutical science is a highly rigorous discipline, where the reliability of knowledge application directly bears on medication safety and the quality of clinical decision-making (Stegmann et al., 2023). In recent years, LLMs have made significant progress in pharmaceutical question answering and reasoning (Singhal et al., 2023), with growing applications in clinical medication use, drug development (Lu et al., 2025), and regulatory compliance (Xiong et al., 2026). To assess models' pharmaceutical capabilities, a range of benchmarks have emerged, covering medication knowledge (Wu et al., 2025), drug reasoning (Liévin et al., 2026), prescription review (Yang et al., 2025), drug label understanding (Xiong et al., 2026), and multi-hop reasoning (Kim et al., 2025). However, existing evaluations still focus mainly on final answer correctness, leaving open how well models support their answers (Turpin et al., 2023) and whether their success extends beyond the conditions under which they are evaluated (Singh et al., 2025).

Model outputs depend on both stored knowledge and the context in which that knowledge is used (Mallen et al., 2023; Xiong et al., 2024). Additional information may supply missing facts (Lewis et al., 2020), while organizing or providing relevant premises may help models apply them (Press et al., 2023). Success with such support raises a further question of whether the resulting judgments remain reliable as the context changes. Models should update their judgments when key conditions change (Gardner et al., 2020), maintain them under equivalent wording (Elazar et al., 2021), and remain stable when options are reordered (Zheng et al., 2024). Removing candidate answers further tests how much performance depends on answer cues (Singh et al., 2025). Controlled comparisons can therefore connect the conditions that support correct answers with the ability to use pharmaceutical knowledge reliably across contexts (Ribeiro et al., 2020).

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

To address these limitations, we propose PharmEval, a diagnostic benchmark for evaluating the pharmaceutical knowledge and reasoning capabilities of LLMs. It contains 6,174 questions across 9 task families and 47 subcategories, covering core pharmaceutical knowledge and clinical reasoning scenarios. PharmEval begins with answer correctness and explanation quality, then examines pharmaceutical knowledge use through four diagnostic studies, including web-based knowledge retrieval, compositional knowledge use, counterfactual sensitivity, and response modality variation. Together, these studies examine how information and premise presentation support successful knowledge use, and whether models preserve or appropriately revise their judgments when conditions and response formats change.

We evaluated a range LLMs on PharmEval and selected representative models for in-depth analysis. The results show that similar answer accuracy can accompany different explanation quality, while high explanation scores can accompany incorrect conclusions. Retrieval improves overall accuracy for most models but consistently reduces performance on certain task families. Organizing or supplying premises can improve target accuracy, yet the benefits differ across models and may coexist with newly introduced errors. Even when models answer correctly under the original conditions, they may fail to update their judgments after an intervention. Strong multiple-choice performance can also coexist with weak open-ended performance and sensitivity to option position. These findings suggest that the challenge lies not only in acquiring relevant knowledge, but also in applying it correctly and consistently as the context changes.

Our contributions are threefold.

1. We construct PharmEval, a pharmaceutical benchmark that links structured pharmaceutical records to 6,174 questions across 9 task families and 47 subcategories covering a broad range of pharmaceutical knowledge and its application to supply a comprehensive testbed for domain evaluation.
2. We develop a diagnostic evaluation framework that connects answer accuracy and explanation quality with four controlled studies. By varying information access, premise presentation, key conditions, and response format, it examines what supports successful knowledge use and whether judgments remain stable or update correctly when the context changes.
3. We show that higher accuracy does not ensure reliable knowledge use across conditions. Models can benefit from retrieval or explicit premises while still introducing new errors, failing to update judgments correctly, or relying heavily on candidate answers.

2 Related Work

Table 1: Comparison of representative benchmarks in pharmaceutical knowledge and diagnostic evaluation. ✓ denotes explicit benchmark-level evaluation, △ partial or indirect coverage, and ✗ no dedicated evaluation identified.

Benchmark	**Knowledge Coverage**				**Diagnostic Evaluation**			
	Medication	Safety	Relations	Physchem	Composition	Access	Counterfactual	Format
ChiDrug	✓	✓	△	✗	△	✗	✗	✗
RxQA	✓	△	✗	✗	△	✓	✗	✗
RxBench	✓	✓	△	✗	△	✗	✗	△
FDARxBench	✓	△	✗	✗	✓	✓	✗	✗
BioHopR	△	✗	✓	✗	✓	✗	✗	✗
MIRAGE	△	△	✗	✗	△	✓	✗	✗
FreeMedQA	△	△	✗	✗	✗	✗	✗	✓
MediEval	△	△	✓	✗	△	✗	✓	✗
PharmEval	✓	✓	✓	✓	✓	✓	✓	✓

2.1 Pharmaceutical Knowledge Benchmarks

Existing pharmaceutical LLM benchmarks primarily focus on clinical medication knowledge, prescription safety, drug-label understanding, and structured biomedical reasoning. ChiDrug evaluates medication knowledge and reasoning across indications, dosage and administration, contraindications, and drug interactions (Wu et al., 2025), while RxQA further compares drug reasoning under closed-book and open-book conditions (Liévin et al., 2026). More specialized work includes RxBench (Yang et al., 2025), which focuses on prescription review and medication error identification, FDARxBench (Xiong et al., 2026), which is based on FDA drug labels, and BioHopR (Kim et al., 2025), which examines multi-hop reasoning over structured biomedical knowledge. As shown in Table 1, existing benchmarks typically focus on specific pharmaceutical scenarios or forms of reasoning. In comparison, `PharmEval` covers multiple types of pharmaceutical knowledge and reasoning tasks within a unified framework, including drug combinations, drug relationships, drug repurposing, indications, metabolism and clinical reasoning, entity normalization, physicochemical properties, rule-based reasoning, and toxicity and safety.

2.2 Diagnostic Evaluation of LLMs

In addition to expanding task coverage, recent studies have begun to examine the specific behaviors underlying model predictions (Cocchieri et al., 2026). MIRAGE and BioHopR analyze how knowledge acquisition and organization affect medical question answering from the perspectives of retrieval-augmented generation and knowledge-graph reasoning, respectively (Xiong et al., 2024; Kim et al., 2025). FreeMedQA compares performance on the same medical questions in multiple-choice and open-ended formats (Singh et al., 2025), while MediEval analyzes models' knowledge and contextual reasoning using factual and counterfactual patient information (Qu & Färber, 2026). As shown in Table 1, diagnostic factors related to information access, knowledge organization, counterfactual changes, and response format are typically examined across separate benchmarks and medical settings. `PharmEval` evaluates these factors through controlled diagnostics within a unified pharmaceutical knowledge framework, examining both the conditions that support knowledge use and its reliability as the context changes.

3 PharmEval

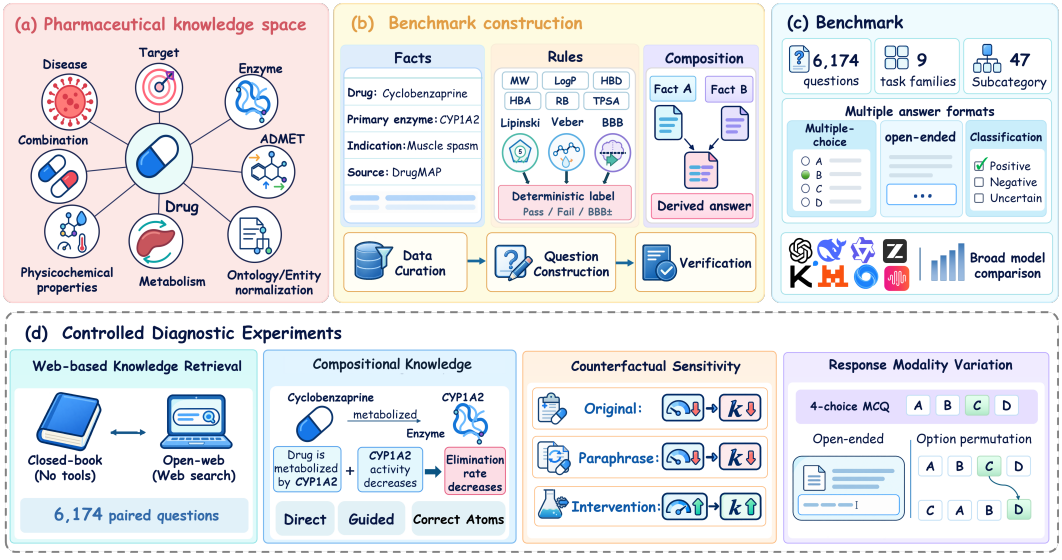


Figure 1: Construction and diagnostics of `PharmEval`. Source entities and relations support Fact, Rule, and Composition questions in three response formats. Four diagnostics examine knowledge use by varying information access, premises, causal conditions, and response format.

`PharmEval` evaluates pharmaceutical knowledge and reasoning through answer accuracy, explanation quality, and four diagnostic studies. It contains 6,174 questions across 9 task families and 47 subcategories. The benchmark links each question to source knowledge and uses controlled vari-

ants to examine how models respond to changes in information, premises, conditions, and answer format. Figure 1 summarizes the benchmark construction process and diagnostic framework.

3.1 Instance Construction

We extract relevant entities and relations from the source database, identify the facts and operations needed to answer each question, and construct the question and reference answer. Fact questions retrieve recorded information, while Rule and Composition questions require applying stated conditions or combining premises. For these questions, the prompt specifies the assumptions that connect the source facts to the target judgment.

Checks cover required fields and reference answers, together with consistency between the options and the reference for multiple-choice questions. Appendix A describes the source mappings, question generation, and validation procedures.

3.2 Instance Structure

Each instance records the question, reference answer, source entities and relations, pharmaceutical domain, knowledge operation, and response format. The Source record field in the examples identifies the relevant drugs, diseases, enzymes, or targets. These links let us vary information access, premise presentation, and answer format while tracking the same target knowledge.

3.3 Benchmark Composition

The benchmark contains 2,877 multiple-choice questions, 2,067 open-ended questions, and 1,230 classification questions. Figure 2 shows their distribution across response formats and pharmaceutical task families. The nine families describe the subject of a question, while Fact, Rule, and Composition describe the operation needed to answer it. Keeping these dimensions separate supports analyses of both domain coverage and knowledge use. Appendix B provides the full taxonomy, subcategory counts, and representative examples.

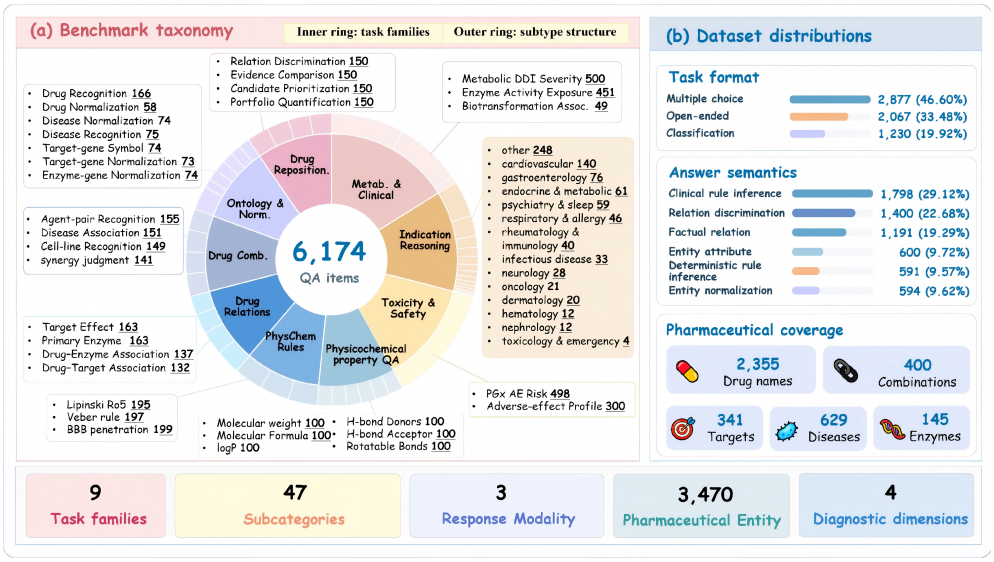


Figure 2: Task coverage and response-format composition of `PharmEval`. The distributions show the relative representation of pharmaceutical task families and the balance between multiple-choice, open-ended, and classification questions, providing context for the overall and task-family results.

3.4 Explanation Quality Evaluation

Answer accuracy measures whether a model reaches the correct conclusion. We assess how well it supports that conclusion through factual support, logical coherence, clinical relevance, and preci-

sion of expression. Factual support concerns the premises used, while logical coherence concerns whether they support the conclusion. Relevance and precision assess how directly the explanation addresses the question. Comparing explanation scores for correct and incorrect answers shows how these two assessments relate. Both assess responses under fixed input conditions. The diagnostic studies examine whether this performance extends to knowledge use under changed conditions.

3.5 Diagnostic Studies

The four diagnostics examine what supports pharmaceutical knowledge use and whether it remains reliable across contexts. Web-based knowledge retrieval tests whether models can use external information to improve their answers. Compositional knowledge use examines whether organizing or supplying relevant premises helps models reach the correct conclusion. We then test whether successful judgments persist or are appropriately revised as conditions change. Counterfactual sensitivity distinguishes stability under rewording from correct updating under intervention. Response modality variation tests whether knowledge use depends on candidate answers or option position. Appendix C details the construction of these comparisons.

3.5.1 Web-based knowledge retrieval

We compare closed-book and search-enabled responses to the same questions (Mallen et al., 2023; Lewis et al., 2020). The reference answer stays fixed, allowing us to track which existing errors are corrected and which correct answers are lost after search. Accuracy changes measure the overall benefit, while corrections and newly introduced errors show when additional information helps or disrupts a judgment (Yoran et al., 2024).

3.5.2 Compositional knowledge use

We define each atomic knowledge unit as an atom and compare direct answering, guided prompting thact asks models to state the relevant facts first, and correct atom prompting that supplies reference premises. Separate atomic probes test the componentd facts (Press et al., 2023; Dziri et al., 2023). Target answers test whether models can use these facts to reach the required conclusion. Comparing Guided with Correct atom further shows what reference premises add beyond asking models to organize the relevant facts themselves.

3.5.3 Counterfactual sensitivity

We compare original questions with paraphrases that preserve the reference answer and interventions that change an answer-determining condition (Gardner et al., 2020; Elazar et al., 2021). Such conditions can concern drug properties, metabolic relationships, or rule parameters. Paraphrase consistency measures answer agreement, while intervention answer-change rate measures how often answers change. Joint correctness tests whether both original and intervention answers are correct. Scoring both variants distinguishes a successful update from an answer change that still leads to an incorrect conclusion.

3.5.4 Response modality variation

We compare multiple-choice questions with open-ended versions and reorder the same options (Singh et al., 2025; Cocchieri et al., 2026). Removing options tests dependence on answer cues, while reordering tests stability when content is unchanged (Zheng et al., 2024). We report both accuracy across positions and correctness across all permutations of each item, distinguishing stable average performance from consistently correct answers. These comparisons show how knowledge use depends on answer cues and presentation.

4 Experiments

4.1 Experimental Setup

We evaluate 13 models on all 6,174 questions under closed-book and search-enabled conditions. GPT-5.6 Terra, DeepSeek-V4.1-flash, Qwen-3.8-max, Gemma-4-E4B-it, and Hulu-Med-7B also

undergo the otherg diagnostics. We abbreviate them as GPT, DeepSeek, Qwen, Gemma, and Hulu-Med. Composition uses 451 targets under direct, guided, and correct atom prompting. Counterfactual evaluation uses original, paraphrase, and enzyme-activity intervention variants of these targets. Response modality uses 2,877 multiple-choice questions, their open-ended versions, and four option-position conditions. Within each study, models answer the samec items and comparisons pair variants of each source question.

Multiple-choice and classification answers use deterministic matching. GPT-4o-mini (OpenAI, 2024) judges open-ended correctness and separately scores 1,600 explanations per model on the four dimensions in Section 3.4. Each dimension uses a 1–5 scale, and their mean gives the overall explanation score. Appendices D and F detail scoring and the rubric.

4.2 Main Results

4.2.1 Answer Accuracy

Table 2 shows closely matched overall accuracy for GPT, DeepSeek, and GLM at 77.42%, 77.36%, and 77.15%, with different task strengths. GPT leads Metab. & Clinical, DeepSeek leads four families, and GLM leads indication reasoning. Mistral has the highest drug-combination accuracy despite lower overall performance. These differences motivate examining both which questions models answer correctly and how well they explain their answers.

Table 2: Accuracy (%) across the nine task families in `PharmEval`, ordered by overall accuracy. Throughout the paper, bold and underlined values denote the best and second-best results in each ranked metric column, including ties. Family abbreviations are defined in Appendix B.1.

Model	**Task Family**									**Overall**
	Phys.	Rules	Rel.	Metab.	Indic.	Safety	Comb.	Entity	Repurp.	
GPT-5.6 Terra (OpenAI, 2026)	68.33	<u>84.43</u>	**86.89**	**68.80**	77.50	72.18	<u>78.52</u>	80.30	87.50	**77.42**
DeepSeek-V4.1-flash (DeepSeek-AI, 2026)	**71.33**	**87.82**	83.36	62.00	<u>81.00</u>	**75.19**	74.83	**82.66**	88.00	<u>77.36</u>
GLM-5.3 (GLM-5-Team, 2026)	67.50	82.06	**86.89**	63.80	**81.88**	<u>74.56</u>	75.84	80.98	89.17	77.15
MiniMax-M3 (MiniMax, 2026)	<u>71.00</u>	83.76	82.52	<u>64.30</u>	78.12	73.43	70.81	<u>82.15</u>	88.67	76.26
Kimi-k3 (Kimi Team, 2026)	68.00	80.54	<u>84.37</u>	61.70	78.88	70.93	72.32	79.80	**89.83**	75.22
Qwen-3.8-max (Alibaba Cloud, 2026)	68.83	74.79	83.36	61.60	79.12	72.56	71.48	78.11	<u>89.67</u>	74.62
Gemma-4-E4B-it (Gemma Team, 2026)	36.83	50.93	82.02	55.60	65.62	66.67	61.24	72.39	76.33	62.78
Qwen3.5-9B (Qwen Team, 2026)	32.50	48.22	82.02	28.90	67.38	51.25	62.75	72.73	75.83	56.14
Hulu-Med-7B (Jiang et al., 2025)	28.17	52.96	71.26	25.50	63.75	51.38	61.07	71.21	61.00	52.38
Mistral-7B-Instruct-v0.3 (Mistral AI, 2024)	31.17	38.41	63.53	22.60	53.87	57.02	**84.06**	64.31	44.17	49.43
MedGemma-4B-it (Sellergren et al., 2025)	28.00	38.41	68.07	30.20	53.12	48.12	68.79	70.37	44.83	48.72
ATHENA-R1-Qwen3-8B (Gao et al., 2026)	17.83	56.51	65.38	20.40	52.38	47.74	66.95	60.94	66.00	48.45
Pharma Assistant v3 (ragpalgit, n.d.)	46.33	60.07	22.35	14.30	23.25	22.68	56.38	25.59	20.00	30.52

4.2.2 Explanation Quality

Answer accuracy and explanation quality assess different aspects of a response. We examine both the four explanation dimensions and their relationship to final-answer correctness. Appendix F provides the scoring rubric and a recorded example.

Dimension-wise scores. We generated 1,600 explanation texts for each model and used GPT-4o-mini as an LLM judge to score the four dimensions on a 1–5 scale. The overall score is the mean of the four dimension scores. Table 3 reports model-level means across the 1,600 explanations.

Table 3 shows why explanation quality adds to answer accuracy. GPT and DeepSeek have nearly identical benchmark accuracy at 77.42% and 77.36%, but explanation means of 3.02 and 3.51. Kimi has the highest explanation mean at 3.54 despite lower benchmark accuracy. DeepSeek exceeds GPT in every explanation dimension, including factual support at 3.48 versus 2.97 and logical coherence at 3.45 versus 2.96. Thus, the difference extends beyond wording quality to the scored support and logic of the explanations.

Table 3: Mean explanation-quality scores by model on a 1–5 scale.

Model	Overall	Factual	Logical	Clinical	Expression
GPT-5.6 Terra	3.02	2.97	2.96	3.18	2.99
DeepSeek-V4.1-flash	3.51	3.48	3.45	3.64	**3.49**
GLM-5.3	3.31	3.29	3.27	3.44	3.26
MiniMax-M3	3.45	3.42	3.42	3.58	3.38
Kimi-k3	**3.54**	**3.53**	**3.49**	**3.68**	3.46
Qwen-3.8-max	3.39	3.35	3.35	3.54	3.32
Gemma-4-E4B-it	3.07	3.05	3.03	3.16	3.06
Qwen3.5-9B	3.04	2.98	3.05	3.16	2.95
Hulu-Med-7B	2.58	2.56	2.47	2.70	2.60
Mistral-7B-Instruct-v0.3	2.57	2.54	2.46	2.72	2.56
ATHENA-R1-Qwen3-8B	2.70	2.66	2.61	2.86	2.69
MedGemma-4B-it	2.53	2.52	2.41	2.61	2.58
Pharma Assistant v3	1.06	1.06	1.05	1.08	1.06

A common pattern also emerges across models. Clinical relevance scores consistently exceed factual support and logical coherence scores, suggesting that explanations address the question more successfully than they support the conclusion. Together with the differences between accuracy and explanation quality, this finding motivates examining how models use relevant facts and whether their judgments remain reliable when the input conditions change.

Appendix F.2 further examines explanation scores for correct and incorrect answers. We examine these questions through controlled changes in information access, premise presentation, causal conditions, and response format.

4.3 Diagnostic Studies

We first examine how external information and premise presentation support knowledge use. We then test whether models maintain or correctly revise their judgments when causal conditions and response formats change. Figures 3–6 summarize the results. Appendices G–J provide detailed analyses.

4.3.1 Web-based knowledge retrieval

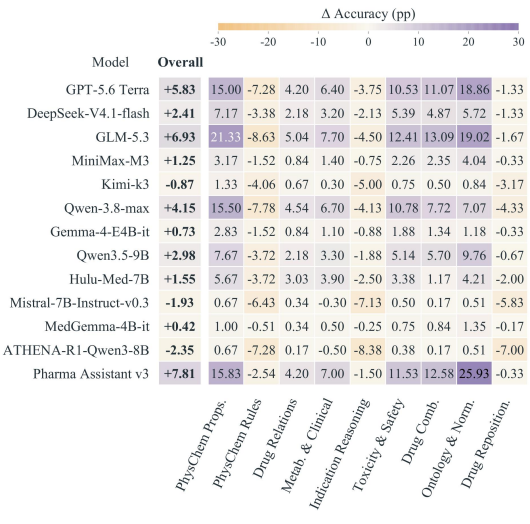


Figure 3: Retrieval effects. Accuracy change with retrieval (pp). Purple indicates gains, orange losses.

Figure 3 shows task-dependent retrieval effects. Overall accuracy improves for ten of 13 models. PhysChem Props., Drug Relations, Toxicity & Safety, Drug Comb., and Ontology & Norm. improve for every model, while PhysChem Rules, Indication Reasoning, and Drug Repositioning decline throughout. Metab. & Clinical improves for 11 models and declines for two. Retrieval gains also differ from final performance. Pharma Assistant v3 gains the most at 7.81 percentage points but reaches 38.32% accuracy, while GLM achieves the highest search-enabled accuracy at 84.08% (Table G1).

The improving families contain many questions about recorded attributes, entity mappings, and relations. The declining families require applying rules or matching evidence to specific relations and constraints. GPT gains 15.00 percentage points on physicochemical properties but loses 7.28 on physicochemical rules. This contrast suggests that retrieving relevant attributes

and applying them under the required criteria pose different demands. Retrieval gains therefore depend on how a task uses information. Gains also coexist with new errors. GPT corrects 480 answers but reverses 120 correct ones (Table G1). The compositional study therefore examines whether models use relevant premises more effectively when those premises are organized or directly supplied.

4.3.2 COMPOSITIONAL KNOWLEDGE USE

Figure 4 shows unequal benefits from reference premises. Correct atom raises GPT, DeepSeek, and Qwen above 99% target accuracy. Gemma improves from 68.96% to 93.57%, while Hulu-Med changes from 47.89% to 48.78%. GPT, DeepSeek, and Qwen already exceed 98% under Guided prompting (Table H1), showing that much of the improvement occurs when models first state the premises themselves.

Correct premises alone do not ensure success. Gemma corrects 112 Direct errors and introduces one under Correct atom, whereas Hulu-Med makes 31 corrections but introduces 27 errors under Correct atom (Table H3). Correct atomic-probe answers also do not guarantee a correct target answer. Gemma answers 102 of 124 targets correctly when all probes are correct, while Hulu-Med answers 17 of 42 (Table H2). Some targets already state the enzyme identity and activity change, so separate fact recall and target answering test different uses of the information. Information and premise organization can support correct answers, but their value depends on how models use them. We next examine whether successful knowledge use persists when models must update their judgments after a key condition changes.

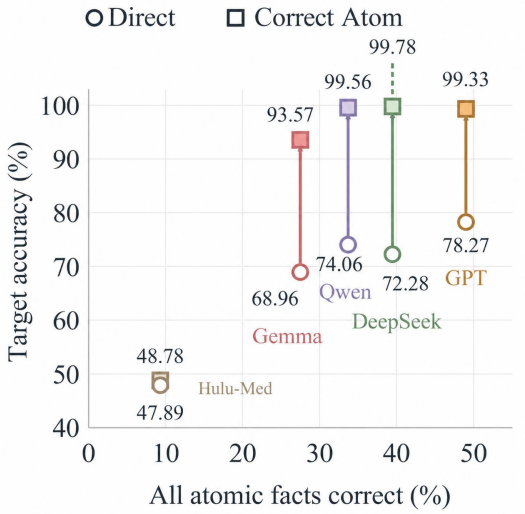


Figure 4: Compositional knowledge use. Direct (circles) and Correct atom (squares) target accuracy on 451 instances, plotted against the measured all-atomic-correct rate. Each model retains the same horizontal coordinate across conditions.

4.3.3 COUNTERFACTUAL SENSITIVITY

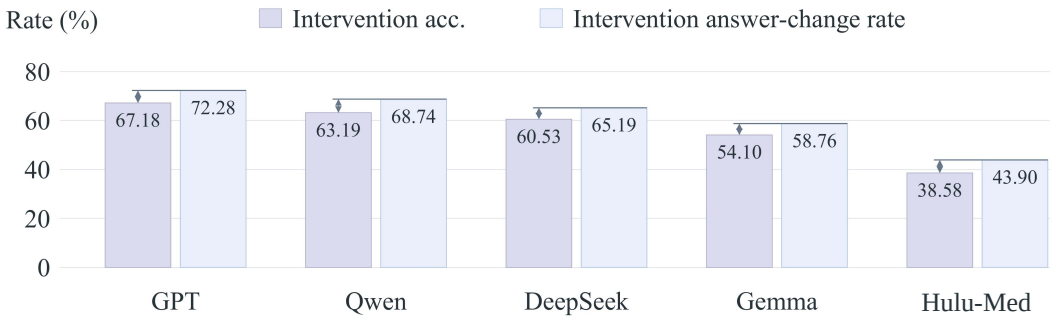


Figure 5: Counterfactual sensitivity. Intervention accuracy and answer-change rate on 451 triplets. Answer-change rate is the percentage of original–intervention pairs with different answers.

Across all five models, paraphrase accuracy stays close to original accuracy, while intervention accuracy falls (Table I1). The condition change therefore exposes difficulties that rewording alone does not reveal. Changing an answer does not ensure a correct update. GPT changes its answer on 72.28% of original–intervention pairs but achieves 67.18% intervention accuracy (Figure 5). Its original accuracy is 78.27% and paraphrase consistency is 85.81%, yet both original and intervention

answers are correct on 61.20% of pairs (Table I1). Of its 353 correct original answers, 77 fail after intervention, representing 21.81% of its original successes. Reliable knowledge use requires both stability under equivalent wording and correct updating when conditions change.

4.3.4 Response modality variation

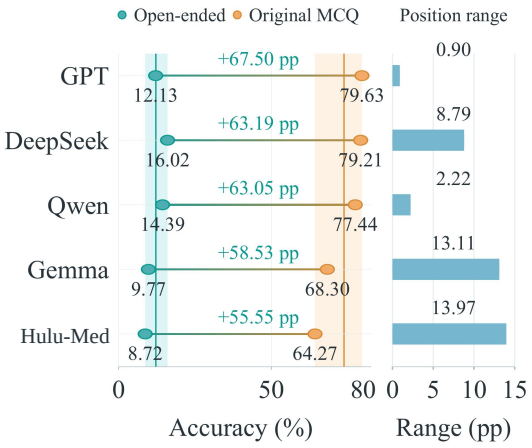


Figure 6: Response format variation. Open-ended and original MCQ accuracy on 2,877 questions (left), with their gaps in percentage points. Option-position accuracy ranges are shown on the right.

We next examine knowledge use when the target knowledge and reference answer remain fixed but answer cues and option position change. Open-ended accuracy is 8.72%–16.02%, compared with 64.27%–79.63% for original MCQ, giving within-model gaps of 55.55–67.50 percentage points (Figure 6). The ranking also changes, with GPT leading original MCQ accuracy and DeepSeek leading open-ended accuracy. High MCQ performance therefore does not carry over to answering without candidates.

Accuracy varies by 0.90 points across option positions for GPT, compared with 13.11 for Gemma and 13.97 for Hulu-Med. Even GPT answers only 61.45% of questions correctly in all four conditions (Table J1). Errors can shift between questions while average accuracy across positions remains similar. The all-four-correct measure reveals whether the same questions remain correct throughout. Together with the counterfactual results, this shows why reliable knowledge use requires correct responses to meaningful changes and stability under changes in presentation.

4.4 Discussion

The results distinguish performance under a fixed setting from reliable pharmaceutical knowledge use across contexts. Accuracy identifies task strengths, while explanation quality assesses the support models provide for their conclusions. Retrieval and premise guidance reveal conditions that help models use knowledge, but improvements under these conditions do not ensure correct updating after an intervention or successful answering without candidates. Read together, the diagnostics show what model success depends on and which changes expose its limits.

5 Conclusion

We introduced `PharmEval` to connect answer accuracy and explanatory support with the conditions and reliability of pharmaceutical knowledge use. The results show that similar accuracy can accompany different explanation quality, while high explanation scores can accompany incorrect conclusions. Retrieval and compositional studies further show that the benefits of additional information and explicit premises depend on the task and the model's ability to use them. Counterfactual and response modality studies reveal that success under one condition can fail to carry over when key conditions, answer cues, or option positions change. Reliable pharmaceutical knowledge use therefore requires both effective use of relevant knowledge and appropriate responses to changes in context. `PharmEval` thus characterizes how well models answer pharmaceutical questions, what supports their success, and where that success fails to extend across contexts.

USE OF LARGE LANGUAGE MODELS (LLMS)

We used a large language model (LLM) to improve the language of the manuscript and assist with benchmark construction. Its uses are outlined below.

1. **Grammar and spell checking.** Checking and correcting grammatical errors, spelling, and punctuation.
2. **Wording and phrasing refinement.** Improving word choice and sentence structure for clearer, more fluent academic writing.
3. **Data construction.** Assisting with benchmark construction through an AI agent workflow.

REFERENCES

Alibaba Cloud. Alibaba unveils Qwen3.8-Max: Its largest and most capable flagship model to date, 2026. URL https://www.alibabacloud.com/en/press-room/alibaba-unveils-qwen3-8-max. Official model identifier: Qwen3.8-Max.

David E. Clark. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. prediction of blood–brain barrier penetration. *Journal of Pharmaceutical Sciences*, 1999. doi: 10.1021/js980402t. URL https://pubmed.ncbi.nlm.nih.gov/10430548/.

Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, and Gianluca Moro. ReMedQA: Are we done with medical multiple-choice benchmarks? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2706–2738, 2026. doi: 10.18653/v1/2026.eacl-long.124. URL https://aclanthology.org/2026.eacl-long.124/.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Nouha Dziri, Ximing Lu, Melanie Sclar, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/deb3c28192f979302c157cb653c15e90-Abstract.html.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Shanghua Gao, Ayush Noori, Richard Zhu, et al. An AI agent for treatment reasoning over a biomedical tool universe, 2026. URL https://arxiv.org/abs/2606.28692. Exact ATHENA-R1-Qwen3-8B model card: https://huggingface.co/mims-harvard/ATHENA-R1-Qwen3-8B.

Matt Gardner, Yoav Artzi, Victoria Basmov, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770. Exact E4B-it model card: https://huggingface.co/google/gemma-4-E4B-it.

GLM-5-Team. GLM-5.3 model card, 2026. URL https://huggingface.co/zai-org/GLM-5.3. The model card requests citation of the GLM-5 family report at https://arxiv.org/abs/2602.15763.

Aleksandr Ianevski, Anil K. Giri, and Tero Aittokallio. SynergyFinder 2.0: visual analytics of multi-drug combination synergies. *Nucleic Acids Research*, 2020. doi: 10.1093/nar/gkaa216. URL https://pubmed.ncbi.nlm.nih.gov/32246720/.

Songtao Jiang, Yuan Wang, Sibo Song, et al. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding, 2025. URL https://arxiv.org/abs/2510.08668. Exact 7B model card: https://huggingface.co/ZJU-AI4H/Hulu-Med-7B.

Yunsoo Kim, Yusuf Abdulle, and Honghan Wu. BioHopR: A benchmark for multi-hop, multi-answer reasoning in biomedical domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12894–12908, 2025. doi: 10.18653/v1/2025.findings-acl.668. URL https://aclanthology.org/2025.findings-acl.668/.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Fengcheng Li, Minjie Mou, Xiaoyi Li, et al. DrugMAP 2.0: molecular atlas and pharma-information of all drugs. *Nucleic Acids Research*, 53(D1):D1372–D1382, 2025. doi: 10.1093/nar/gkae791. URL https://pubmed.ncbi.nlm.nih.gov/39271119/. Open full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC11701670/.

Valentin Liévin, Anil Palepu, Wei-Hung Weng, et al. Towards conversational artificial intelligence for disease management. *Nature*, 2026. doi: 10.1038/s41586-026-10764-5. URL https://pubmed.ncbi.nlm.nih.gov/42310463/. Introduces the RxQA medication-reasoning benchmark; open full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC13421318/.

Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 1997. doi: 10.1016/S0169-409X(96)00423-1. URL https://pubmed.ncbi.nlm.nih.gov/11259830/.

Yang Liu, Dan Iter, Yichong Xu, et al. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

James Lu, Keunwoo Choi, Maksim Eremeev, et al. Large language models and their applications in drug discovery and development: A primer. *Clinical and Translational Science*, 18(4): e70205, 2025. doi: 10.1111/cts.70205. URL https://pubmed.ncbi.nlm.nih.gov/40208836/. Open full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC11984503/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

MiniMax. MiniMax M3: Frontier coding, 1m context, native multimodality, 2026. URL https://www.minimax.io/blog/minimax-m3. Exact model card: https://huggingface.co/MiniMaxAI/MiniMax-M3.

Mistral AI. Mistral 7B v0.3 model documentation, 2024. URL https://docs.mistral.ai/models/mistral-7b-0-3. Official page dated 22 May 2024; exact Instruct-v0.3 weights: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

OpenAI. GPT-4o mini: advancing cost-efficient intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/. Includes GPT-5.6 Terra; exact identifier: https://developers.openai.com/api/docs/models/gpt-5.6-terra.

Ofir Press, Muru Zhang, Sewon Min, et al. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Zhan Qu and Michael Färber. MediEval: A unified medical benchmark for patient-contextual and knowledge-grounded reasoning in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. doi: 10.18653/v1/2026.acl-long.734. URL https://aclanthology.org/2026.acl-long.734/.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. The exact Qwen3.5-9B model card supplies this author, year, and citation title.

ragpalgit. Pharma Assistant v3 model card, n.d. URL https://huggingface.co/ragpalgit/pharma-assistant-v3. Model source confirmed by the study authors; accessed 26 September 2026.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, et al. MedGemma technical report, 2025. URL https://arxiv.org/abs/2507.05201. Exact medgemma-4b-it checkpoint: https://huggingface.co/google/medgemma-4b-it.

Shrutika Singh, Anton Alyakin, Daniel Alexander Alber, et al. The pitfalls of multiple-choice questions in generative AI and medical education. *Scientific Reports*, 15:42096, 2025. doi: 10.1038/s41598-025-26036-7. URL https://pubmed.ncbi.nlm.nih.gov/41298584/. Open full text: https://pmc.ncbi.nlm.nih.gov/articles/PMC12658246/.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 2023. doi: 10.1038/s41586-023-06291-2. URL https://pubmed.ncbi.nlm.nih.gov/37438534/.

Jens-Ulrich Stegmann, Rory Littlebury, Markus Trengove, Lea Goetz, Andrew Bate, and Kim M. Branson. Trustworthy AI for safe medicines. *Nature Reviews Drug Discovery*, 22:855–856, 2023. doi: 10.1038/s41573-023-00769-4. URL https://pubmed.ncbi.nlm.nih.gov/37550364/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html.

Daniel F. Veber, Stephen R. Johnson, Hung-Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45:2615–2623, 2002. doi: 10.1021/jm020017n. URL https://pubmed.ncbi.nlm.nih.gov/12036371/.

Michelle Whirl-Carrillo, Rachel Huddart, Li Gong, Katrin Sangkuhl, Caroline F. Thorn, Ryan Whaley, and Teri E. Klein. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics*, 110:563–572, 2021. doi: 10.1002/cpt.2350. URL https://pubmed.ncbi.nlm.nih.gov/34216021/.

Yue Wu, Yangmin Huang, Qianyun Du, Lixian Lai, Zhiyang He, Jiaxue Hu, and Xiaodong Tao. Should I believe in what medical AI says? a chinese benchmark for medication based on knowledge and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1155–1164, 2025. doi: 10.18653/v1/2025.acl-short.91. URL https://aclanthology.org/2025.acl-short.91/.

Betty Xiong, Jillian Fisher, Benjamin Newman, et al. FDARxBench: Benchmarking regulatory and clinical reasoning on FDA generic drug assessment, 2026. URL https://arxiv.org/abs/2603.19539.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

Yan Yang, Mouxiao Bian, Peiling Li, et al. Human-level and beyond: Benchmarking large language models against clinical pharmacists in prescription review, 2025. URL https://arxiv.org/abs/2512.02024.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8011b23e1dc3f57e1b6211ccad498919-Abstract-Conference.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/54dd9e0cff6d9214e20d97eb2a3bae49-Abstract-Conference.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Detailed Benchmark Construction

This appendix explains how the 6,174 questions connect to source records, reference answers, and scoring rules. The instance schema defines what is stored, the source mappings locate the underlying records, and the examples show how those records support different question formats.

A.1 Complete Instance Structure

These fields serve different roles in evaluation. Evidence and verification fields support answer validation, grouping identifiers enable analyses over shared entities, knowledge units, and templates, and `diagnostic_eligibility` specifies which controlled studies can be constructed from each instance.

Table A1: Fields and roles in a complete benchmark instance.

Field	Purpose
id / type / subtype	Question ID, task family, and subcategory
question / task_format	Question and response format
answer / scoring	Reference answer, accepted answers, and scoring rule
evidence / verification	Source, supporting facts, and verification record
metadata / group ids	Entities, attributes, and knowledge-unit/entity/template groups
diagnostic eligibility	Eligibility for each diagnostic study
item history / temporal scope	Version history and temporal scope

Table A1 links each evaluated answer to its source and scoring metadata. Keeping these fields separate supports controlled comparisons. Retrieval changes available information, composition changes premise presentation, and response modality changes answer cues while retaining the target answer. Counterfactual intervention changes a key condition and assigns a corresponding reference answer. The stored records make these relationships traceable across variants.

Example A1 · Complete benchmark instance · `MCR_001`

Task family Metab. & Clinical

Subcategory Enzyme Activity Exposure

Response format / answer semantics open-ended / clinical rule inference

Question
For Cyclobenzaprine, metabolism through the primary enzyme Cytochrome P450 1A2 (CYP1A2) becomes much slower. If the apparent volume of distribution remains stable, indicate the expected direction of change in the first-order elimination rate constant.

Reference answer **The elimination rate constant decreases**

Recorded explanation
The source database records Cytochrome P450 1A2 (CYP1A2) as a primary metabolizing enzyme for Cyclobenzaprine. Reduced enzyme activity lowers metabolic clearance; with apparent volume of distribution unchanged, the elimination rate constant therefore decreases.

Supporting fact
Cyclobenzaprine is linked to the primary metabolizing enzyme Cytochrome P450 1A2 (CYP1A2).

Source record
Drug Name: Cyclobenzaprine; Enzyme Name: Cytochrome P450 1A2 (CYP1A2); Drug ID: DM1YBRM; DME ID: DEJGDUW

Source fields DrugID, Drug_Name, DME_ID, DME Name, Primary_enzyme

Recorded reference matching normalized exact match

Example A1 (continued)

Verification status schema and answer checks passed; manual review completed

Evaluation scoring
Semantic correctness judged by GPT-4o-mini (Appendix D.4)

Diagnostic eligibility information_access, compositional, counterfactual

Example A1 connects a source drug–enzyme relation to a pharmacokinetic question and its reference direction. The source record establishes the relation, while the question's stated assumptions determine how it is used. Table A1 defines the full set of fields.

A.2 Source Data and Evidence Extraction

Table A2 aligns each question group with the relevant DrugMAP 2.0 source table (Li et al., 2025) and its entity or relation fields. The mapping identifies where the corresponding drugs, diseases, proteins, or relations occur in the supplied raw data. Pharmacogenomic evidence levels used by the PGx adverse-event task follow the PharmGKB evidence framework (Whirl-Carrillo et al., 2021). The HSA, Bliss, Loewe, and ZIP fields used by the drug-combination tasks correspond to established synergy models summarized in SynergyFinder 2.0 (Ianevski et al., 2020).

Table A2: Source-table alignment for the PharmEval question groups. Numbered files follow the DrugMAP 2025 full-data download; listed fields identify the available entity or relation anchors.

Question group (n)	Raw source table	Record key and related fields
Physicochemical properties and rules (1,191)	#1, General Information of Drug	DrugID; mw, Formula, xlogp, hbonddonor, hbondacc, rotbonds, polararea.
Target/enzyme normalization (221)	#5, General Information of Drug Therapeutic Target (DTT); #8, General Information of Drug Metabolism Enzyme (DME)	TargetID or DME ID; entity name, Gene Name.
Combination records (596)	#2, General Information of Drug Combination	DrugCom ID; member drugs, Disease Entry, Cell Line, HSA/ZIP/Bliss/Loewe.
Disease recognition (75)	#4, General Information of Disease	DiseaseID; Disease Entry, Disease Synonymous.
Drug/disease ontology records (298)	#1, General Information of Drug; #4, General Information of Disease	DrugID or DiseaseID; names and synonyms.
Indications (800)	#1, General Information of Drug; #16, Drug to Indication Mapping Information; #4, General Information of Disease	DrugID + DiseaseID; Drug Name, Disease Entry, Status.
Drug–enzyme relations and exposure (800)	#1, General Information of Drug; #12, Drug to DME Mapping Information; #8, General Information of Drug Metabolism Enzyme (DME)	DrugID + DME ID; Drug Name, DME Name, Primary enzyme.
Drug–target relations (295)	#1, General Information of Drug; #10, Drug to DTT Mapping Information; #5, General Information of Drug Therapeutic Target (DTT)	DrugID + TargetID; Target Name, MOA, Activity.
Repositioning (600)	#1, General Information of Drug; #3, General Information of Drug Reposition; #4, General Information of Disease	DrugID + DiseaseID; Disease Entry, Status.

Question group (n)	Raw source table	Record key and related fields
Metabolic DDI severity (500)	#1, General Information of Drug; #12, Drug to DME Mapping Information; #8, General Information of Drug Metabolism Enzyme (DME)	Paired DrugID values and DME ID; metabolic-interaction category, severity label.
PGx risk and adverse-effect profiles (798)	#1, General Information of Drug; #12, Drug to DME Mapping Information; #8, General Information of Drug Metabolism Enzyme (DME)	DrugID + DME ID; variation, adverse-reaction relation, adverse-effect terms.

Note. Question groups are organized by their source records. Some groups combine questions from multiple task families that share a source table; for example, physicochemical properties and physicochemical rules are grouped together. The number in parentheses gives the total number of questions in each group. The nine-task-family taxonomy is detailed in Appendix B.

Table A2 shows how questions draw on entity attributes and recorded relations. Attribute questions can use a drug or disease record, whereas indication, enzyme, target, and repositioning questions may require linked entity and mapping tables. A database join identifies the relevant relation. Rule and Composition questions additionally require applying conditions or combining premises to derive the answer. This distinction connects the source mapping to the knowledge-operation taxonomy.

A.3 Question and Reference Answer Generation

Question construction specifies the source facts, target answer, and response format before forming the question. Multiple-choice items record the options and correct choice. Open-ended items record canonical and `accepted_answers`, while classification items record the allowed labels and scoring rule. Deterministic checks cover required fields and reference answers, with an additional option-consistency check for multiple-choice items.

Examples A2(a)–A2(c) illustrate clinical inference in open-ended responses, multi-constraint choice among repositioning programs, and evidence-threshold classification. Each example reports the reference answer, recorded explanation, and supporting information.

Example A2(a) · Open-ended · `MCR_193`

Task family Metab. & Clinical

Subcategory Enzyme Activity Exposure

Question
Suppose a patient at steady state on Amitriptyline hydrochloride develops substantially reduced Mephenytoin 4-hydroxylase (CYP2C19) activity. The source database identifies this enzyme as a primary metabolic route. With the same dosing rate, will average steady-state concentration rise or fall?

Reference answer **Average steady-state concentration increases**

Recorded explanation
The source database records Mephenytoin 4-hydroxylase (CYP2C19) as a primary metabolizing enzyme for Amitriptyline hydrochloride. With linear pharmacokinetics and an unchanged dosing rate, reduced enzyme activity lowers clearance and raises the average steady-state parent-drug concentration.

Supporting fact
Amitriptyline hydrochloride is linked to the primary metabolizing enzyme Mephenytoin 4-hydroxylase (CYP2C19).

Recorded reference matching
Normalized exact match after Unicode NFKC, trimming, whitespace collapse, and casefolding

Example A2(a) (continued)

Evaluation scoring
Semantic correctness judged by GPT-4o-mini against the reference and accepted answers (Appendix D.4)

Example A2(b) · Multiple choice · `REP_302`

Task family Drug Reposition.

Subcategory Candidate Prioritization

Decision rule
A development team is prioritizing a non-approved, currently active repositioning program for Olaparib. Inactive programs and already approved indications are ineligible; among eligible programs, prefer the highest development phase.

Recorded programs
Peritoneal cancer: Withdrawn; Ovarian epithelial tumor: Discontinued in Phase 1; Pancreatic cancer: Phase 3; Ovarian cancer: Approved

Source rows
Peritoneal cancer: 727; Ovarian epithelial tumor: 1030; Pancreatic cancer: 2447; Ovarian cancer: 2446

Question Which condition should be prioritized?

Options
A. Peritoneal cancer; B. Ovarian cancer; C. Ovarian epithelial tumor; D. Pancreatic cancer

Reference answer **D**

Recorded explanation
Pancreatic cancer is the eligible active, non-approved program with the highest development phase among the listed candidates.

Scoring Multiple choice

Example A2(c) · Classification · `TS_508`

Task family Toxicity & Safety

Subcategory PGx AE Risk

Allowed labels true; false

Question
Does rs9923231 (VKORC1) have high-level pharmacogenomic evidence for an association with over-anticoagulation during treatment with Warfarin?

Reference answer **true**

Recorded explanation
True. This association has evidence level 1B. Levels 1A, 1B, 2A, and 2B are treated as high-level evidence; level 3 is not.

Scoring Deterministic matching

The three examples in this subsection illustrate how answer generation follows the recorded task format. The open-ended case applies a pharmacokinetic rule, the multiple-choice case selects among repositioning programs under stated constraints, and the classification case applies an evidence threshold. Their different response formats and scoring fields motivate the format-specific evaluation procedures in Appendix D.3.

B Task Taxonomy and Benchmark Statistics

B.1 Task Families and Subcategory Definitions

Table B1 organizes the 6,174 questions into nine task families and 47 subcategories, with question counts and response formats for each subcategory.

The deterministic physicochemical-rule tasks apply established medicinal-chemistry criteria. The Lipinski task uses the Rule-of-Five properties (Lipinski et al., 1997), and the Veber task uses molecular-flexibility and polar-surface-area criteria associated with oral bioavailability (Veber et al., 2002). The BBB task uses polar surface area as a predictor of blood–brain barrier penetration (Clark, 1999). Any exact cutoff used by `PharmEval` is treated as a benchmark-specific operational threshold rather than attributed to Clark (1999).

Table B1: Task families, subcategories, question counts, and response formats in the current benchmark set.

Subcategory	Questions	Response format
PhysChem Props.		
Molecular Formula	100	Open-ended
H-bond Acceptors	100	Open-ended
H-bond Donors	100	Open-ended
LogP	100	Open-ended
Molecular Weight	100	Open-ended
Rotatable Bonds	100	Open-ended
PhysChem Rules		
BBB Penetration	199	Classification
Lipinski Rule of Five	195	Classification
Veber Rule	197	Classification
Drug Relations		
Drug–Enzyme Association	137	Multiple choice
Drug–Target Association	132	Multiple choice
Target Effect	163	Open-ended
Primary Enzyme	163	Multiple choice
Metab. & Clinical		
Biotransformation Assoc.	49	Multiple choice
Metabolic DDI Severity	500	Open-ended
Enzyme Activity Exposure	451	Open-ended
Indication Reasoning		
Cardiovascular	140	Multiple choice
Dermatology	20	Multiple choice
Endocrine & Metabolic	61	Multiple choice
Gastroenterology	76	Multiple choice
Hematology	12	Multiple choice
Infectious Disease	33	Multiple choice
Nephrology	12	Multiple choice
Neurology	28	Multiple choice
Oncology	21	Multiple choice
Other Indications	248	Multiple choice
Psychiatry & Sleep	59	Multiple choice
Respiratory & Allergy	46	Multiple choice
Rheumatology & Immunology	40	Multiple choice
Toxicology & Emergency	4	Multiple choice
Toxicity & Safety		
Adverse-effect Profile	300	Multiple choice
PGx AE Risk	498	Classification
Drug Comb.		
Disease Association	151	Multiple choice

Subcategory	Questions	Response format
Cell-line Recognition	149	Multiple choice
Agent-pair Recognition	155	Multiple choice
Synergy Judgment	141	Classification
Ontology & Norm.		
Disease Recognition	75	Multiple choice
Disease Normalization	74	Open-ended
Drug Recognition	166	Multiple choice
Drug Normalization	58	Open-ended
Enzyme-gene Normalization	74	Open-ended
Target-gene Normalization	73	Open-ended
Target-gene Symbol	74	Open-ended
Drug Reposition.		
Candidate Prioritization	150	Multiple choice
Evidence Comparison	150	Multiple choice
Portfolio Quantification	150	Multiple choice
Relation Discrimination	150	Multiple choice

Note. In Table 2, Phys., Rules, Rel., Metab., Indic., Safety, Comb., Entity, and Repurp. denote physicochemical properties, physicochemical rules, drug relationships, metabolic and clinical reasoning, indication reasoning, toxicity and safety, drug combinations, ontology and entity normalization, and drug repurposing, respectively.

Table B1 shows the distribution of questions across task families and response formats. The family counts establish the denominators used for family-level benchmark accuracy, while the response-format column shows why accuracy must be interpreted together with the scoring procedure used for each item.

B.2 Knowledge Operation Taxonomy

Task families and knowledge operations describe complementary properties of each instance. Task families specify the pharmaceutical domain, whereas knowledge operations characterize the dominant reasoning operation required to obtain the answer. An operation label can include factual retrieval as an intermediate step. Subtype and answer semantics are recorded separately for each question.

Table B2: Knowledge operations used to characterize answer derivation.

Knowledge operation	Required knowledge	Answer derivation
Fact	One entity attribute or relation	Retrieve or match the recorded fact
Rule	Known attributes and an explicit pharmaceutical rule	Apply the rule to the stated conditions
Composition	Multiple related knowledge units	Combine premises to derive the target answer

Table B2 distinguishes Fact questions that retrieve a recorded attribute or relation, Rule questions that apply stated conditions, and Composition questions that combine premises. These labels describe the operation required by a question, rather than a model's observed strategy. They complement task-family labels, which describe the pharmaceutical subject. This distinction is useful when interpreting retrieval effects because questions about the same subject can require different uses of information.

B.3 Representative Benchmark Examples

Examples B.1–B.47 provide one representative question for each task-family/subcategory combination in Table B1. Each box shows the question, response format, reference answer, and source record so that readers can connect the taxonomy to an actual benchmark item. The examples illustrate task

coverage and answer meaning. Aggregate results support comparisons between models. Questions and reference answers come from the current benchmark release.

Example B.1 · PhysChem Props. · `PP_387`

Subcategory **Molecular Formula**

Response format open-ended

Question
What is the molecular formula of Fedratinib hydrochloride? Fill in the blank: ____

Reference answer **C27H40Cl2N6O4S**

Source record Drug Name: Fedratinib hydrochloride

Example B.2 · PhysChem Props. · `PP_223`

Subcategory **H-bond Acceptors**

Response format open-ended

Question
What hydrogen-bond acceptor count best matches Naldemedine Tosylate? Fill in the blank: ____

Reference answer **8 or more**

Source record Drug Name: Naldemedine Tosylate

Example B.3 · PhysChem Props. · `PP_214`

Subcategory **H-bond Donors**

Response format open-ended

Question
What hydrogen-bond donor count best matches Pentaerythritol tetranitrate? Fill in the blank: ____

Reference answer **0**

Source record Drug Name: Pentaerythritol tetranitrate

Example B.4 · PhysChem Props. · `PP_080`

Subcategory **LogP**

Response format open-ended

Question
Which LogP range best matches Clevidipine butyrate? Fill in the blank: ____

Reference answer **$\geq$4**

Source record Drug Name: Clevidipine butyrate

Example B.5 · PhysChem Props. · `PP_167`

Subcategory **Molecular Weight**

Response format open-ended

Example B.5 (continued)

Question
What molecular-weight range best matches Lomitapide mesylate? Fill in the blank: ____

Reference answer ≥500 g/mol

Source record Drug Name: Lomitapide mesylate

Example B.6 · PhysChem Props. · `PP_329`

Subcategory **Rotatable Bonds**

Response format open-ended

Question
What rotatable-bond count best matches Olmesartan medoxomil? Fill in the blank: ____

Reference answer 10 or more

Source record Drug Name: Olmesartan medoxomil

Example B.7 · PhysChem Rules · `PRI_405`

Subcategory **BBB Penetration**

Response format classification

Question
Fill in the blank with true or false: Can Furosemide cross the blood-brain barrier? Answer: ____

Reference answer false

Source record Drug Name: Furosemide

Example B.8 · PhysChem Rules · `PRI_004`

Subcategory **Lipinski Rule of Five**

Response format classification

Question
Fill in the blank with true or false: Does Daclatasvir comply with Lipinski's Rule of Five for good oral bioavailability? Answer: ____

Reference answer false

Source record Drug Name: Daclatasvir

Example B.9 · PhysChem Rules · `PRI_334`

Subcategory **Veber Rule**

Response format classification

Question
Fill in the blank with true or false: Does Paclitaxel comply with the Veber Rule for good oral bioavailability? Answer: ____

Reference answer false

Example B.9 (continued)

Source record Drug Name: Paclitaxel

Example B.10 · Drug Relations · `DRL_474`

Subcategory **Drug–Enzyme Association**

Response format multiple choice

Question
Which enzyme has a documented metabolic, bioactivation, or drug-inactivation relationship with Tenofovir alafenamide? Options: A. CYP3A4; B. CYP3A43; C. DZB84_00060; D. 10-LAH

Reference answer **A**

Source record
Drug Name: Tenofovir alafenamide; Enzyme Name: Cytochrome P450 3A4 (CYP3A4)

Example B.11 · Drug Relations · `DRL_508`

Subcategory **Drug–Target Association**

Response format multiple choice

Question
Which target has a documented pharmacological association with Tbo-Filgrastim? Options: A. Granulocyte colony-stimulating factor (CSF3); B. Granulocyte colony-stimulating factor receptor (G-CSF-R); C. Leukemia RAW264.7 cells; D. Alpha(B)-crystallin

Reference answer **B**

Source record
Drug Name: Tbo-Filgrastim; Target Name: Granulocyte colony-stimulating factor receptor (G-CSF-R)

Example B.12 · Drug Relations · `DRL_201`

Subcategory **Target Effect**

Response format open-ended

Question
How does Avalglucosidase alfa functionally affect Lysosomal alpha-glucosidase (GAA)? Fill in the blank: ____

Reference answer
A replacement, cell-based, or direct structural mechanism

Source record
Drug Name: Avalglucosidase alfa; Target Name: Lysosomal alpha-glucosidase (GAA)

Example B.13 · Drug Relations · `DRL_579`

Subcategory **Primary Enzyme**

Response format multiple choice

Example B.13 (continued)

Question
Which enzyme is identified as a primary human drug-metabolizing enzyme for Canagliflozin? Options: A. UGT2B7; B. UGT2B4; C. aadA1a; D. 10-LAH

Reference answer B

Source record
Drug Name: Canagliflozin; Enzyme Name: UDP-glucuronosyltransferase 2B4 (UGT2B4)

Example B.14 · Metab. & Clinical · `MCR_486`

Subcategory **Biotransformation Assoc.**

Response format multiple choice

Question
Which enzyme or catalytic system has a documented metabolism, activation, or inactivation relationship with Bleomycin? Options: A. BLMH; B. bla; C. gap; D. aac

Reference answer A

Source record Drug Name: Bleomycin; Enzyme Name: Bleomycin hydrolase (BLMH)

Example B.15 · Metab. & Clinical · `MCR_877`

Subcategory **Metabolic DDI Severity**

Response format open-ended

Question
What is the clinical severity of the metabolic interaction between Interferon Alfa-2a, Recombinant and Theophylline? Fill in the blank: ____

Reference answer Moderate

Source record Drug Names: Interferon Alfa-2a, Recombinant; Theophylline

Example B.16 · Metab. & Clinical · `MCR_282`

Subcategory **Enzyme Activity Exposure**

Response format open-ended

Question
The source database identifies UDP-glucuronosyltransferase 1A1 (UGT1A1) as a primary metabolizing enzyme for Dolutegravir sodium. During repeated dosing at a fixed dose and interval, what happens to parent-drug accumulation if UDP-glucuronosyltransferase 1A1 (UGT1A1) activity becomes markedly reduced and other clearance routes do not compensate?

Reference answer Parent-drug accumulation increases

Source record
Drug Name: Dolutegravir sodium; Enzyme Name: UDP-glucuronosyltransferase 1A1 (UGT1A1)

Example B.17 · Indication Reasoning · IR_259

Subcategory **Cardiovascular**

Response format multiple choice

Question
Which of the following conditions is an indication for Eptifibatide? Options: A. Angina pectoris; B. MALT lymphoma; C. Acute cardiac ischemic events; D. 2-methylbutyryl-CoA dehydrogenase deficiency

Reference answer C

Source record Drug Name: Eptifibatide; Disease Name: Acute cardiac ischemic events

Example B.18 · Indication Reasoning · IR_607

Subcategory **Dermatology**

Response format multiple choice

Question
Which of the following conditions is an indication for Sulfacetamide? Options: A. Verruca vulgaris; B. Short-rib thoracic dysplasia 17 with or without polydactyly; C. Acne vulgaris; D. 2-methylbutyryl-CoA dehydrogenase deficiency

Reference answer C

Source record Drug Name: Sulfacetamide; Disease Name: Acne vulgaris

Example B.19 · Indication Reasoning · IR_560

Subcategory **Endocrine & Metabolic**

Response format multiple choice

Question
Which of the following conditions is an indication for Cipaglucosidase alfa? Options: A. Glycogen storage disease type IA; B. Hereditary spastic paraplegia 2; C. 2-aminoadipic 2-oxoadipic aciduria; D. Glycogen storage disease type II

Reference answer D

Source record
Drug Name: Cipaglucosidase alfa; Disease Name: Glycogen storage disease type II

Example B.20 · Indication Reasoning · IR_514

Subcategory **Gastroenterology**

Response format multiple choice

Question
Which of the following conditions is an indication for Morantel tartrate? Options: A. Gastrointestinal infection; B. Mature gastrointestinal nematode infection; C. Pediatric growth disorder; D. 2-aminoadipic 2-oxoadipic aciduria

Reference answer B

Source record
Drug Name: Morantel tartrate; Disease Name: Mature gastrointestinal nematode infection

Example B.21 · Indication Reasoning · IR_710

Subcategory **Hematology**

Response format multiple choice

Question
Which of the following conditions is an indication for Sodium ferric gluconate complex? Options: A. Folate-deficiency anemia; B. Iron-deficiency anemia; C. Central nervous system injury; D. 22q11.2 deletion syndrome

Reference answer **B**

Source record
Drug Name: Sodium ferric gluconate complex; Disease Name: Iron-deficiency anemia

Example B.22 · Indication Reasoning · IR_114

Subcategory **Infectious Disease**

Response format multiple choice

Question
Which of the following drugs is indicated for Human immunodeficiency virus infection? Options: A. Cidofovir; B. Tenofovir; C. Pyrrolidin-3-yl-acetic acid; D. ((3R,4R)-4-(o-tolyloxy)chroman-3-yl)methanamine

Reference answer **B**

Source record
Drug Name: Tenofovir; Disease Name: Human immunodeficiency virus infection

Example B.23 · Indication Reasoning · IR_716

Subcategory **Nephrology**

Response format multiple choice

Question
Which of the following drugs is indicated for Chronic kidney disease? Options: A. Secalciferol; B. 5-Bromo-8-imidazol-1-ylmethyl-chromen-4-one; C. (35)S-labeled oligonucleotide; D. Doxercalciferol

Reference answer **D**

Source record Drug Name: Doxercalciferol; Disease Name: Chronic kidney disease

Example B.24 · Indication Reasoning · IR_747

Subcategory **Neurology**

Response format multiple choice

Question
Which of the following conditions is an indication for Ethotoin? Options: A. Absence seizure; B. Hypophosphatemic nephrolithiasis/osteoporosis 1; C. Complex partial seizure; D. 2-aminoadipic 2-oxoadipic aciduria

Reference answer **C**

Source record Drug Name: Ethotoin; Disease Name: Complex partial seizure

Example B.25 · Indication Reasoning · IR_479

Subcategory **Oncology**

Response format multiple choice

Question
Which of the following conditions is an indication for Exemestane? Options: A. HER2-positive breast cancer; B. Severe combined immunodeficiency due to CTPS1 deficiency; C. Hormonally-responsive breast cancer; D. 2-aminoadipic 2-oxoadipic aciduria

Reference answer C

Source record Drug Name: Exemestane; Disease Name: Hormonally-responsive breast cancer

Example B.26 · Indication Reasoning · IR_428

Subcategory **Other Indications**

Response format multiple choice

Question
Which of the following conditions is an indication for Doxylamine? Options: A. Motion sickness; B. PURA-related severe neonatal hypotonia-seizures-encephalopathy syndrome due to a point mutation; C. 2-aminoadipic 2-oxoadipic aciduria; D. Morning sickness

Reference answer D

Source record Drug Name: Doxylamine; Condition: Morning sickness

Example B.27 · Indication Reasoning · IR_331

Subcategory **Psychiatry & Sleep**

Response format multiple choice

Question
Which of the following conditions is an indication for Escitalopram? Options: A. Manic depression disorder; B. Syndromic oculocutaneous albinism; C. Major depressive disorder; D. 2-aminoadipic 2-oxoadipic aciduria

Reference answer C

Source record Drug Name: Escitalopram; Disease Name: Major depressive disorder

Example B.28 · Indication Reasoning · IR_449

Subcategory **Respiratory & Allergy**

Response format multiple choice

Question
Which of the following conditions is an indication for Bromodiphenhydramine? Options: A. Hay fever; B. Obsolete ocular albinism with congenital sensorineural hearing loss; C. Vasomotor symptom; D. 2-aminoadipic 2-oxoadipic aciduria

Reference answer A

Source record Drug Name: Bromodiphenhydramine; Disease Name: Hay fever

Example B.29 · Indication Reasoning · IR_642

Subcategory **Rheumatology & Immunology**

Response format multiple choice

Question
Which of the following conditions is an indication for obeticholic acid? Options: A. Primary sclerosing cholangitis; B. Primary biliary cholangitis; C. Cardioacrofacial dysplasia 2; D. 2-aminoadipic 2-oxoadipic aciduria

Reference answer B

Source record Drug Name: obeticholic acid; Disease Name: Primary biliary cholangitis

Example B.30 · Indication Reasoning · IR_483

Subcategory **Toxicology & Emergency**

Response format multiple choice

Question
Which of the following drugs is indicated for Ethylene glycol or methanol poisoning? Options: A. Penicillamine; B. NV-31; C. Fomepizole; D. ((E)-3-Amino-propenyl)-methyl-phosphinic acid

Reference answer C

Source record Drug Name: Fomepizole; Condition: Ethylene glycol or methanol poisoning

Example B.31 · Toxicity & Safety · TS_059

Subcategory **Adverse-effect Profile**

Response format multiple choice

Question
Which option best matches the adverse-event profile associated with carbamazepine? Options: A. Amblyopia / Asthenia / Ataxia / Constipation / Dermatitis; B. Abnormal behaviour / Agitation / Anorexia / Anxiety / Arrhythmia; C. Abdominal pain / Amblyopia / Asthenia / Chills / Dermatitis; D. Abdominal abscess / Abdominal pain / Abdominal pain upper / Aggression / Alanine aminotransferase increased

Reference answer A

Source record
Drug Name: carbamazepine; Adverse Events: Amblyopia, Asthenia, Ataxia, Constipation, Dermatitis

Example B.32 · Toxicity & Safety · TS_954

Subcategory **PGx AE Risk**

Response format classification

Question
Does HLA-B*57:01 (HLA-B) have high-level pharmacogenomic evidence for an association with Drug Hypersensitivity and HIV Infections during treatment with Abacavir?

Reference answer true

Source record
Drug Name: Abacavir; Variant: HLA-B*57:01; Gene Name: HLA-B; Conditions: Drug Hypersensitivity, HIV Infections

Example B.33 · Drug Comb. · COMBO_003

Subcategory **Disease Association**

Response format multiple choice

Question
For Pazopanib + Moxifloxacin, which option names the disease or clinical condition linked to this combination record? Options: A. Carcinoma, Renal Cell; B. Squamous Cell Carcinoma; C. Carcinoma, Large Cell; D. Hepatocellular Carcinoma Stage IIIa

Reference answer **A**

Source record
Drug Names: Pazopanib + Moxifloxacin; Disease Name: Carcinoma, Renal Cell; DrugCom ID: DCWOC84

Example B.34 · Drug Comb. · COMBO_308

Subcategory **Cell-line Recognition**

Response format multiple choice

Question
For Ruxolitinib + ABIRATERONE in Clear cell renal cell carcinoma, which option names the experimental cell line used in this combination record? Options: A. CAKI-1; B. HL-60(TB); C. MDA-MB-435; D. SF-268

Reference answer **A**

Source record
Drug Names: Ruxolitinib + ABIRATERONE; Disease Name: Clear cell renal cell carcinoma; Cell Line: CAKI-1; DrugCom ID: DCMRNHR

Example B.35 · Drug Comb. · COMBO_541

Subcategory **Agent-pair Recognition**

Response format multiple choice

Question
For the combination record linked to Chronic myelogenous leukemia and KBM-7 cells, which option matches the pharmacological pair? Options: A. Mitomycin + Everolimus; B. Diethylcarbamazine + Olmesartan medoxomil; C. Fluorouracil + Erlotinib; D. Mitomycin + SNX-2112

Reference answer **B**

Source record
Drug Names: Diethylcarbamazine + Olmesartan medoxomil; Disease Name: Chronic myelogenous leukemia; Cell Line: KBM-7; DrugCom ID: DCKOVCO

Example B.36 · Drug Comb. · COMBO_268

Subcategory **Synergy Judgment**

Response format classification

Question
For Dactinomycin + Raloxifene in Adult T acute lymphoblastic leukemia using MOLT-4 cells, the HSA score is 5.61. Using > 10 for synergistic, -10 to 10 inclusive for additive, and < -10 for antagonistic, how should the interaction be classified?

Example B.36 (continued)

Reference answer additive

Source record
Drug Names: Dactinomycin + Raloxifene; Disease Name: Adult T acute lymphoblastic leukemia; Cell Line: MOLT-4; HSA: 5.61; DrugCom ID: DCNLRGK

Example B.37 · Ontology & Norm. · ONT_016

Subcategory Disease Recognition

Response format multiple choice

Question
Which option names a disease or clinical condition rather than a drug, enzyme, or protein target? Options: A. Plasmodium Histidine-rich protein II (Malaria HRP II); B. AS1 14-3-3 protein eta; C. Mitochondrial complex 1 deficiency, nuclear type 36; D. Sulfotransferase 1A3

Reference answer C

Source record
Disease Name: Mitochondrial complex 1 deficiency, nuclear type 36; Disease ID: DIS5WUK9

Example B.38 · Ontology & Norm. · ONT_115

Subcategory Disease Normalization

Response format open-ended

Question
What is the standard disease name corresponding to the synonym “phenylketonuria type 2”? Fill in the blank: ____

Reference answer Dihydropteridine reductase deficiency

Source record
Disease Name: Dihydropteridine reductase deficiency; Synonym: phenylketonuria type 2; Disease ID: DIS1IC3E

Example B.39 · Ontology & Norm. · ONT_017

Subcategory Drug Recognition

Response format multiple choice

Question
Which option names a drug or therapeutic product rather than a disease, enzyme, or protein target? Options: A. PSMA-11 Ga-68; B. Attention deficit hyperactivity disorder; C. Bacterial 50S ribosomal RNA (Bact 50S rRNA); D. Voltage-gated sodium channel alpha Nav1.5

Reference answer A

Source record Drug Name: PSMA-11 Ga-68; Drug ID: DMWLTU2

Example B.40 · Ontology & Norm. · ONT_073

Subcategory **Drug Normalization**

Response format open-ended

Question
What is the standard drug name corresponding to the synonym "Benzenesulfonamide, 4-amino-N-2-pyrimidinyl-, monosilver(1+) salt"? Fill in the blank: ____

Reference answer **Silver sulfadiazine**

Source record
Drug Name: Silver sulfadiazine; Synonym: Benzenesulfonamide, 4-amino-N-2-pyrimidinyl-, monosilver(1+) salt; Drug ID: DMVTHWQ

Example B.41 · Ontology & Norm. · ONT_142

Subcategory **Enzyme-gene Normalization**

Response format open-ended

Question
What is the canonical enzyme name corresponding to gene symbol CYP4F2? Fill in the blank: ____

Reference answer **Cytochrome P450 4F2**

Source record Enzyme Name: Cytochrome P450 4F2; Gene Name: CYP4F2; DME ID: DE3GT9C

Example B.42 · Ontology & Norm. · ONT_158

Subcategory **Target-gene Normalization**

Response format open-ended

Question
What is the canonical protein target corresponding to gene symbol CSF1R? Fill in the blank: ____

Reference answer **Macrophage colony-stimulating factor 1 receptor**

Source record
Target Name: Macrophage colony-stimulating factor 1 receptor; Gene Name: CSF1R; Target ID: TT7MRDV

Example B.43 · Ontology & Norm. · ONT_557

Subcategory **Target-gene Symbol**

Response format open-ended

Question
What gene or locus label corresponds to the target Chloride channel protein 3 (CLC-3)? Fill in the blank: ____

Reference answer **CLCN3**

Source record
Target Name: Chloride channel protein 3 (CLC-3); Gene Name: CLCN3; Target ID: TT8XNZ7

Example B.44 · Drug Reposition. · `REP_421`

Subcategory **Candidate Prioritization**

Response format multiple choice

Question
A development team is prioritizing a non-approved, currently active repositioning program for Hydromorphone. Inactive programs and already approved indications are ineligible; among eligible programs, prefer the highest development phase. Programs: Common cold — Approved; Pain — Approved; Chronic pain — Phase 3; Back pain — Approved. Which condition should be prioritized? Options: A. Back pain; B. Common cold; C. Chronic pain; D. Pain

Reference answer **C**

Source record
Drug Name: Hydromorphone; Programs: Common cold, Pain, Chronic pain, Back pain

Example B.45 · Drug Reposition. · `REP_266`

Subcategory **Evidence Comparison**

Response format multiple choice

Question
For Mefloquine, compare two repositioning programs on two separate dimensions: Coronavirus Disease 2019 (COVID-19) — Phase 2; Severe acute respiratory syndrome (SARS) — Preclinical. Historical maturity is the highest level reached, whereas current state distinguishes approved, active, inactive, and early-stage programs. Which interpretation is accurate? Options: A. Coronavirus Disease 2019 (COVID-19) reached the higher historical development level; currently it is active but not approved, while Severe acute respiratory syndrome (SARS) is early-stage or investigative.; B. Coronavirus Disease 2019 (COVID-19) reached the higher historical development level; currently it is early-stage or investigative, while Severe acute respiratory syndrome (SARS) is active but not approved.; C. Severe acute respiratory syndrome (SARS) reached the higher historical development level; currently it is early-stage or investigative, while Coronavirus Disease 2019 (COVID-19) is active but not approved.; D. The two programs reached the same historical development level; currently both are active but not approved.

Reference answer **A**

Source record
Drug Name: Mefloquine; Programs: Coronavirus Disease 2019 (COVID-19), Severe acute respiratory syndrome (SARS)

Example B.46 · Drug Reposition. · `REP_451`

Subcategory **Portfolio Quantification**

Response format multiple choice

Question
Audit six repositioning programs for Rituximab: Metastatic melanoma — Discontinued in Phase 2; Rheumatoid arthritis — Discontinued in Phase 3; B-cell neoplasm — Investigative; Central nervous system non-hodgkin lymphoma — Approved; MALT lymphoma — Approved; Pemphigus vulgaris — Phase 3. Classify each program once as approved, active non-approved, inactive, or early-stage/investigative. Which count profile is correct? Options: A. approved 2; active non-approved 0; inactive 3; early-stage/investigative 1; B. approved 3; active non-approved 0; inactive 2; early-stage/investigative 1; C. approved 2; active non-approved 1; inactive 1; early-stage/investigative 2; D. approved 2; active non-approved 1; inactive 2; early-stage/investigative 1

Reference answer **D**

Example B.46 (continued)

Source record
Drug Name: Rituximab; Programs: Metastatic melanoma, Rheumatoid arthritis, B-cell neoplasm, Central nervous system non-hodgkin lymphoma, MALT lymphoma, Pemphigus vulgaris

Example B.47 · Drug Reposition. · `REP_034`

Subcategory **Relation Discrimination**

Response format multiple choice

Question
Which clinical condition is associated with a repositioning-development program for Bupropion? Options: A. Nephrotic syndrome, type 6; B. Seasonal affective disorder; C. Fanconi anemia complementation group T; D. Schizoaffective disorder

Reference answer **B**

Source record Drug Name: Bupropion; Disease Name: Seasonal affective disorder

C Detailed Construction of Diagnostic Studies

The four studies examine pharmaceutical knowledge use through controlled changes to information access, premise presentation, causal conditions, and response format. Retrieval and composition test the support available for a judgment. Counterfactual and response-format comparisons test whether judgments remain reliable when the conditions or presentation change.

C.1 Web-based Knowledge Retrieval

Closed-book and search-enabled evaluations are paired at the instance level, allowing retrieval effects to be measured on identical questions and reference answers. Retrieved information is treated as model-accessible evidence rather than benchmark ground truth. This design follows medical RAG evaluation that separates access to evidence from effective use of that evidence (Xiong et al., 2024).

The web-based knowledge retrieval study uses all 6,174 benchmark questions. Each model answers under closed-book and search-enabled conditions. The records include search queries, returned snippets, final answers, and item-level scores. Each question allows up to five queries, ten pages, and six tool rounds.

Example C1 · Web retrieval / GPT-5.6 Terra · `PP_399`

Question What is the molecular formula of Clobetasol? Fill in the blank: ____

Closed-book answer C22H29ClFO4

Retrieved result
The search record for Clobetasol lists Molecular Formula: C22H28ClFO4. The same result separately lists the 17-propionate derivative as C25H32ClFO5 (DrugFuture, drugfuture.com/chemdata/clobetasol.html).

Search-enabled answer C22H28ClFO4

Reference answer **C22H28ClFO4**

Case reading
The retrieved record supplies the formula for the queried compound and distinguishes it from a derivative; the model changes from an incorrect closed-book formula to the reference formula.

Source record Drug Name: Clobetasol

Example C1 changes the information available to the model while keeping the question and reference formula fixed. The corrected search-enabled answer illustrates a beneficial change on one item. Appendix G reports overall effects, correctness transitions, and further recorded retrieval cases.

C.2 Compositional Knowledge Use

C.2.1 Compositional Instance Selection

The compositional study uses 451 instances from Enzyme Activity Exposure. Each associates a drug–enzyme relation with an enzyme-activity rule and a target pharmacokinetic judgment. Separate probes elicit the relation and rule, following prior work on atomic facts and composition (Press et al., 2023). Some target questions explicitly supply the enzyme identity and activity change. The study therefore measures target performance across premise conditions and its association with separate probe responses.

C.2.2 Atomic Knowledge Decomposition

Example C2 · Atomic premises and composition · `MCR_296`

Target question
In a patient receiving Letrozole, Cytochrome P450 2A6 (CYP2A6)-dependent metabolism is markedly reduced. Holding apparent distribution volume and other elimination mechanisms constant, predict the direction of change in the drug's elimination rate constant.

Atomic premise A1 Primary enzyme in the benchmark record: Cytochrome P450 2A6 (CYP2A6).

Atomic premise A2 Effect of reduced enzyme activity: Metabolism decreases.

Target reference answer **The elimination rate constant decreases**

Hulu-Med-7B atomic answers A1: CYP3A4 (CYP3A4); A2: Metabolic clearance decreases.

Hulu-Med-7B target answers
Direct: increases; guided: The elimination rate constant decreases; correct atom: Elimination rate constant decreases.

Case reading
The direct answer gives the opposite direction; after the atomic step or supplied premises, the target answer follows the reference direction. The example distinguishes retrieval of a premise from using premises to answer the target.

Source record Drug Name: Letrozole; Enzyme Name: Cytochrome P450 2A6 (CYP2A6)

Example C2 demonstrates the value of retaining both atomic and target responses. Hulu-Med-7B gives the wrong direction directly but the reference direction after the premise step. The comparison records a change after the premise step. Aggregate rates in Appendix H show how often such changes help or hurt.

C.2.3 Evaluation Settings

In the Direct condition, the model receives only the target question. In the Guided condition, the model first answers the associated atomic questions before producing the target answer. In the Correct atom condition, the reference atomic premises are explicitly provided before target answering.

Table C1: Information and measured behavior in the compositional evaluation conditions.

Condition	Information provided	Measured behavior
Direct	Target question only	Target judgment without additional premises
Guided	Answer atomic questions before the target question	Effect of explicit stepwise organization
Correct atom	Provide both correct atoms before the target question	Target performance with reference premises

Table C1 separates two ways of making premises explicit. Guided requires the model to produce them, while Correct atom supplies the reference premises. Their comparison shows whether reference facts add to the benefit of organizing the response. The target question remains in the input under every condition, including any enzyme identity it already names. Separate atomic-probe scores therefore measure fact answering in their own question context. Appendix H relates these scores to target accuracy and answer transitions.

C.3 Counterfactual Sensitivity

C.3.1 Counterfactual Instance Selection

The counterfactual study uses the same 451 instances. For each target question, a paraphrase control changes the wording while preserving the relevant condition, and a causal intervention changes the enzyme-activity condition that determines the answer. These minimally changed variants follow the contrast-set methodology (Gardner et al., 2020) and complement medical evaluation using factual and counterfactual patient contexts (Qu & Färber, 2026).

C.3.2 Variant Construction

Table C2: Changes and expected answer relations in the counterfactual variants.

Variant	Change	Expected answer relation
Original	Original condition and question	Original reference answer
Paraphrase	Reword the question; retain the pharmacokinetic condition	Same answer as the original
Causal-intervention	Change the key pharmacokinetic condition	Reference answer changes with the causal condition

Table C2 separates a wording control from a causal intervention. The wording control preserves the answer-bearing condition, so a changed answer indicates sensitivity to phrasing. The intervention reverses the relevant enzyme-activity condition, so the expected pharmacokinetic direction changes. Retaining the original direction then fails to update the judgment.

Example C3 · Counterfactual sensitivity · `CF_MCR_430`

Original question
A loss of Glutathione S-transferase alpha-1 (GSTA1) activity reduces the primary metabolic pathway for Melphalan. Assuming linear kinetics and constant distribution volume, what happens to the time required for the parent-drug concentration to fall by half?

Original reference answer **The elimination half-life increases.**

Paraphrase control
The loss of GSTA1 activity and the half-life endpoint are retained; only the wording is changed. Expected answer: the elimination half-life increases.

Causal intervention
GSTA1 activity rises rather than falls while the stated pharmacokinetic assumptions and endpoint are held fixed. Expected answer: the elimination half-life decreases.

Diagnostic contrast
The paraphrase control tests stability under equivalent wording; the intervention tests whether the half-life prediction updates when the causal condition reverses.

Source record
Drug Name: Melphalan; Enzyme Name: Glutathione S-transferase alpha-1 (GSTA1)

Example C3 holds the pharmacokinetic assumptions and half-life endpoint fixed while changing enzyme activity. This controlled contrast makes the expected answer reversal interpretable and motivates the paired stability and update measures reported in Appendix I.

C.3.3 Counterfactual Validation

The study records responses to the original question, its paraphrase, and the intervention. Paraphrase consistency is the percentage of the 451 original–paraphrase pairs with the same answer. The intervention answer-change rate is the percentage of the 451 original–intervention pairs with differ-

ent answers. These metrics describe response behavior. Accuracy scores each answer against its reference, and joint correctness requires both original and intervention answers to be correct.

The paraphrase preserves the entities, condition, endpoint, and reference answer. The intervention reverses enzyme activity while holding the stated assumptions and endpoint fixed, so the expected direction reverses. A consistent paraphrase pair can contain two incorrect answers. A changed intervention answer can also be incorrect. Reporting behavior and correctness together distinguishes stability from correctness and answer changes from successful updating.

C.4 Response Modality Variation

C.4.1 Instance Selection

The response modality study derives open-ended and option-permuted versions from 2,877 four-option multiple-choice questions. All source questions remain in the 6,174-question set. The paired design follows evidence that multiple-choice and free-response formats can reveal materially different medical-model behavior (Singh et al., 2025).

C.4.2 Response Modality Construction

Example C4 · Response modality · `IR_002`

Original multiple-choice question
Which of the following drugs is indicated for Hairy cell leukaemia? Options: A. Cisplatin; B. Pentostatin; C. 7-bromoindirubin-3-oxime; D. ((3R,4R)-4-(o-tolyloxy)chroman-3-yl)methanamine

Open-ended condition The model names the indicated drug without seeing candidate options.

Reference answer **Pentostatin (option B in the original question).**

Option permutations
The same four option texts are reordered so that Pentostatin appears at B, A, D, and C in P1 to P4, respectively.

Controlled contrast
The underlying drug-indication relation is unchanged; only the response form and option position vary.

Source record Drug Name: Pentostatin; Disease Name: Hairy cell leukaemia

Example C4 keeps the drug–indication relation fixed while changing the available answer cues. The open-ended condition requires the model to produce the drug name, whereas the permuted multiple-choice conditions test recognition under different option positions. Their paired results in Appendix J distinguish generation difficulty from sensitivity to presentation.

C.4.3 Option Permutation and Semantic Control

Position range measures the difference between the highest and lowest accuracy obtained when the correct option appears at different positions. Semantic consistency evaluates whether the selected answer content remains consistent across option permutations independent of the option label.

The four permutations place the correct option at A, B, C, and D while retaining the option texts and question meaning. Mean permutation accuracy averages scores over the four conditions. All-four-correct counts an instance only when every permutation is answered correctly. Semantic consistency instead concerns agreement in selected content, which can include consistently choosing the same wrong answer. These measures capture different aspects of response stability.

D Experimental Setup and Evaluation Protocol

D.1 Evaluated Models

All 13 models undergo closed-book and search-enabled evaluation on 6,174 questions. Five models additionally undergo compositional knowledge use, counterfactual sensitivity, and response modality variation studies. Table D1 identifies this five-model subset. Model IDs remain fixed across the subcategory plots.

Table D1: The 13 evaluated models. Interface identifiers come from run metadata; model licensing is not inferred from those identifiers.

ID	Model	Additional Diagnostics
M1	GPT-5.6 Terra	Yes
M2	DeepSeek-V4.1-flash	Yes
M3	GLM-5.3	No
M4	MiniMax-M3	No
M5	Kimi-k3	No
M6	Qwen-3.8-max	Yes
M7	Gemma-4-E4B-it	Yes
M8	Qwen3.5-9B	No
M9	Hulu-Med-7B	Yes
M10	Mistral-7B-Instruct-v0.3	No
M11	ATHENA-R1-Qwen3-8B	No
M12	MedGemma-4B-it	No
M13	Pharma Assistant v3	No

Table D1 specifies model coverage for the three subset studies. All models within a study answer the same eligible questions, and each comparison pairs responses derived from the same source question. Retrieval covers 6,174 questions, composition and counterfactual sensitivity share 451 targets, and response modality uses 2,877 source MCQs. These scopes determine the denominator for each study. Retrieval results for all 13 models appear in Appendix G.

D.2 Inference Settings

During inference, all evaluated models used a temperature of 0.0 and a maximum output length of 512 tokens. The retry limit was set to 5 for models with available retry control. These configurations were applied consistently across model evaluations to maintain comparable generation conditions.

D.3 Response Formats and Answer Normalization

The benchmark contains 2,877 multiple-choice questions, 2,067 open-ended questions, and 1,230 classification questions (Table B1). The classification count includes all 141 Synergy Judgment questions. Multiple-choice responses are matched to the reference option label or text, and classification responses to the allowed labels. Open-ended items retain canonical references, accepted answers, and normalization metadata. GPT-4o-mini supplies their final semantic correctness labels as described in Section D.4.

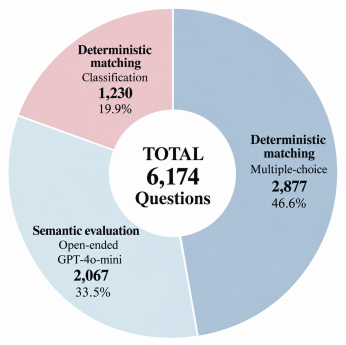


Figure D.1: Answer-scoring procedures in `PharmEval`. Multiple-choice and classification responses use deterministic matching with format-specific rules. Multiple-choice responses are matched to the reference option label or text, while classification responses are matched to the reference class within the allowed label set for each task. Open-ended responses are evaluated semantically by GPT-4o-mini.

Normalization follows each item's recorded settings. These may specify Unicode NFKC, trimming, whitespace collapse, or case-insensitive comparison. Numerical units, tolerances, and range rules apply only when encoded in the item. These fields describe reference normalization. The final open-ended judgment is semantic.

D.4 Semantic Evaluation with LLM-as-a-Judge

GPT-4o-mini judges open-ended correctness relative to the reference answer, including open-ended versions of the 2,877 questions in the response modality study. Accepted answers and normalization fields specify the reference, while the judge supplies the final correctness label. Multiple-choice and classification answers use deterministic matching. GPT-4o-mini separately scores explanation quality on four dimensions (Liu et al., 2023). Answer correctness and explanation scores are retained as separate outcomes. Model-based judgments can reflect evaluator preferences and biases (Zheng et al., 2023).

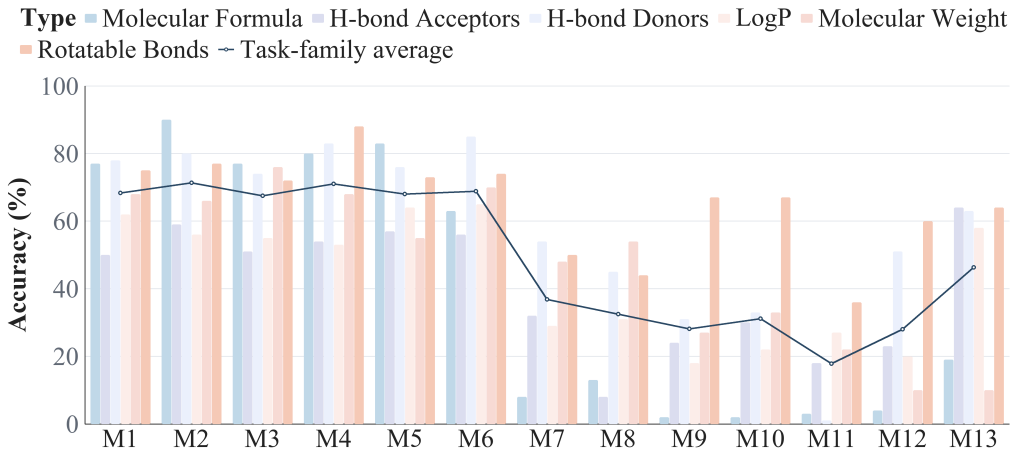
Table D2: Recorded semantic-evaluation settings.

Setting	**Recorded setting**
Judge model	GPT-4o-mini
Application	Semantic answer judgments and explanation-quality scoring, as applicable.
Explanation-quality dimensions	Factual support, logical coherence, clinical relevance, and precision of expression.
Scale	1–5 per dimension; overall score is the mean of four dimensions.

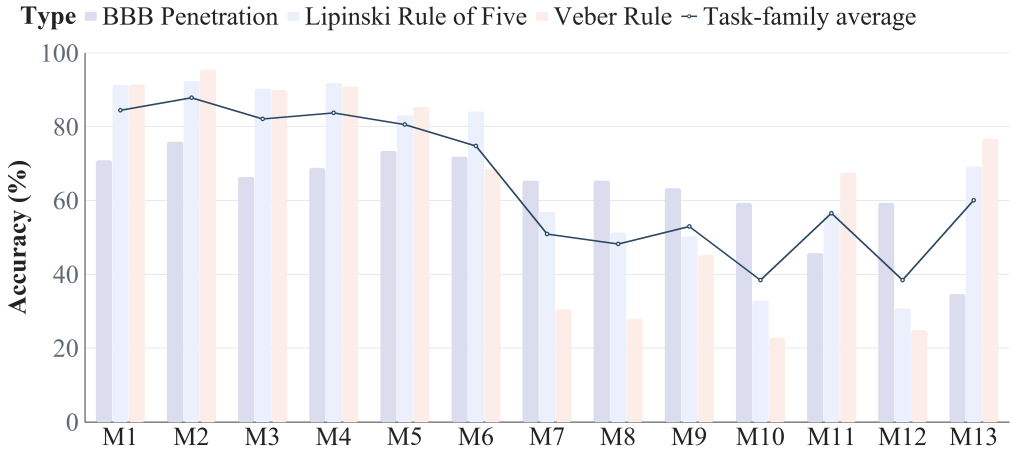
Table D2 records the judge used for semantic answer and explanation assessments and the four explanation dimensions. Reporting these settings makes the explanation scores in Appendix F interpretable alongside, but distinct from, deterministic answer accuracy.

E FINE-GRAINED PERFORMANCE ACROSS PHARMEVAL SUBCATEGORIES

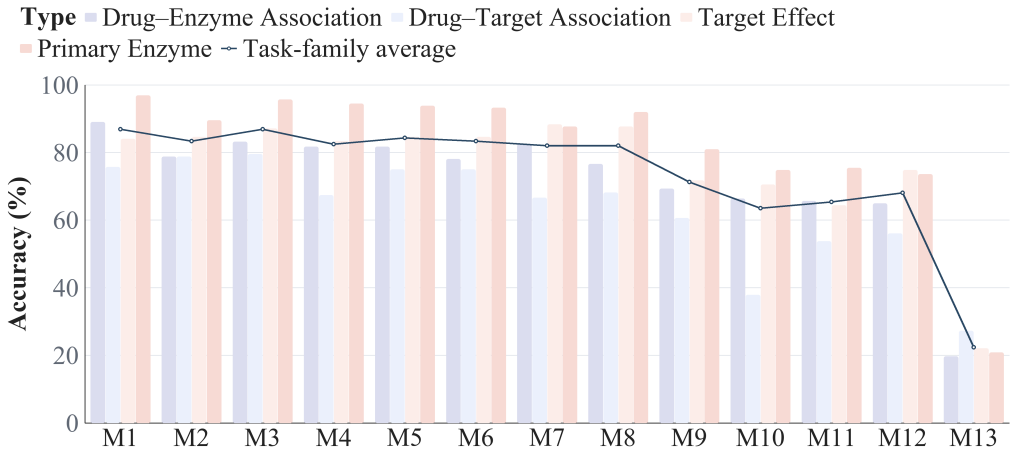
Figure E.1 reports subcategory accuracy within the nine task families. Model identifiers M1–M13 follow Table D1. The line in each panel shows the task-family average.



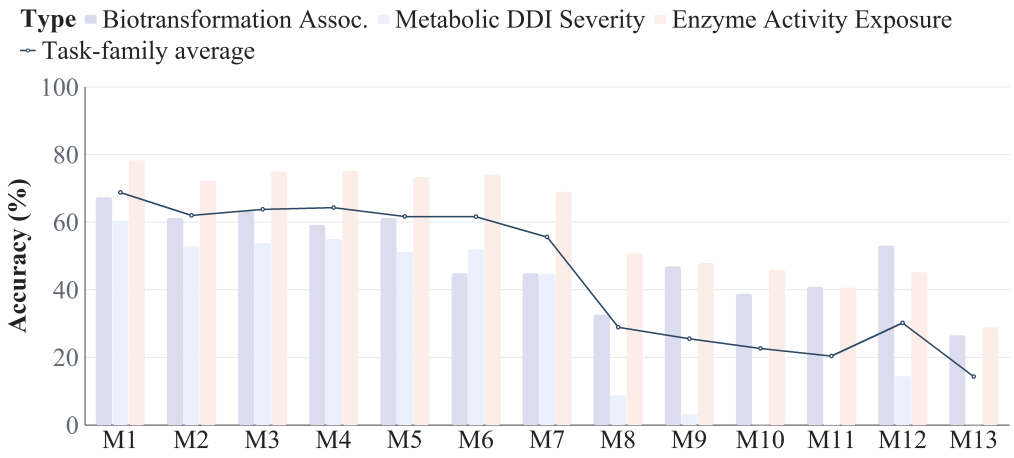
(a) PhysChem Props.



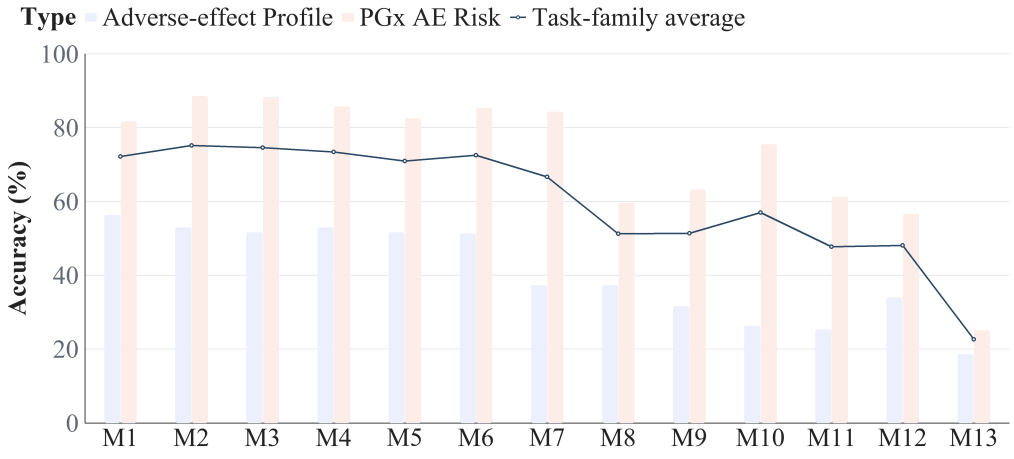
(b) PhysChem Rules



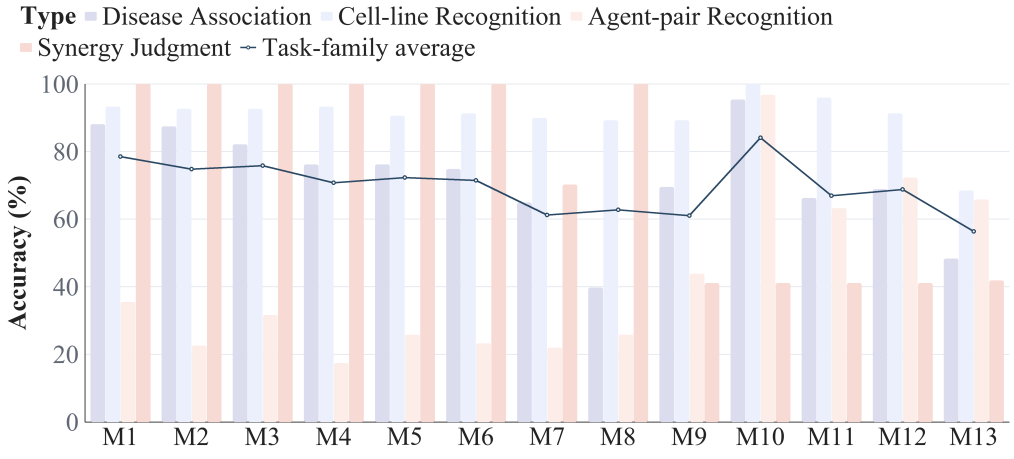
(c) Drug Relations



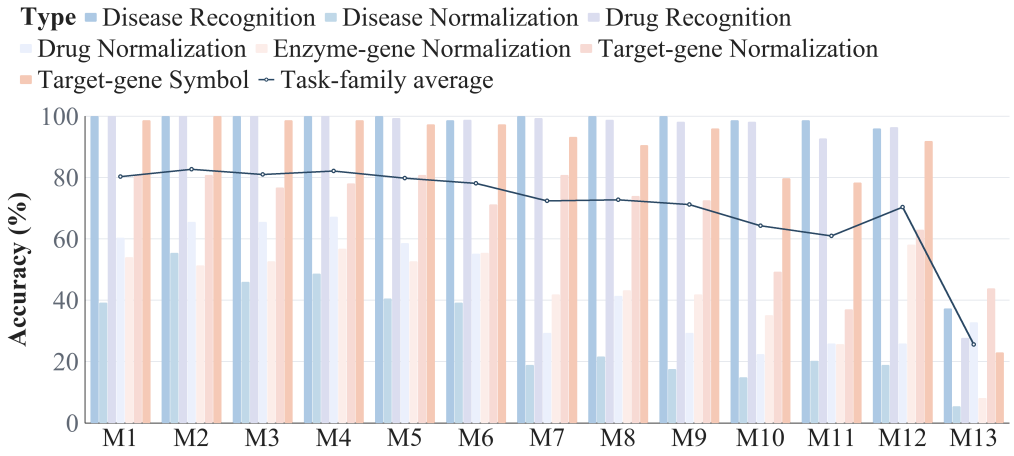
(d) Metab. & Clinical



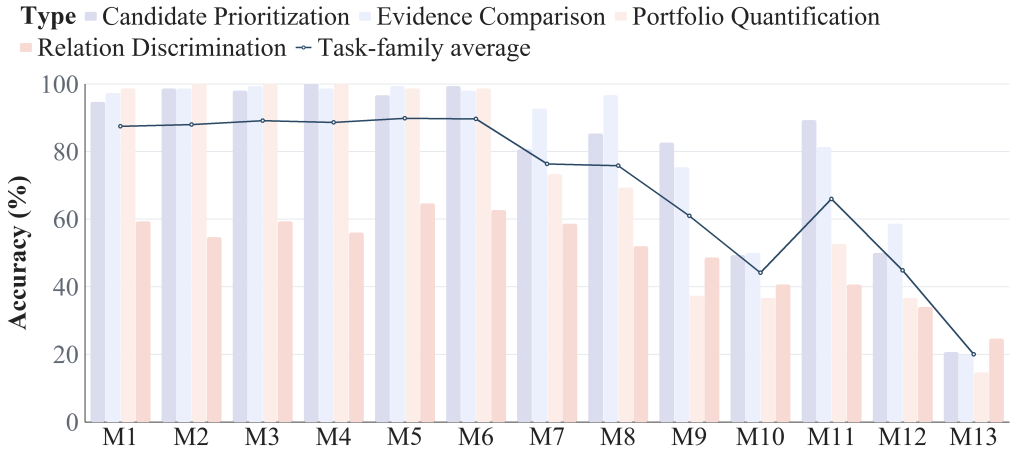
(e) Toxicity & Safety



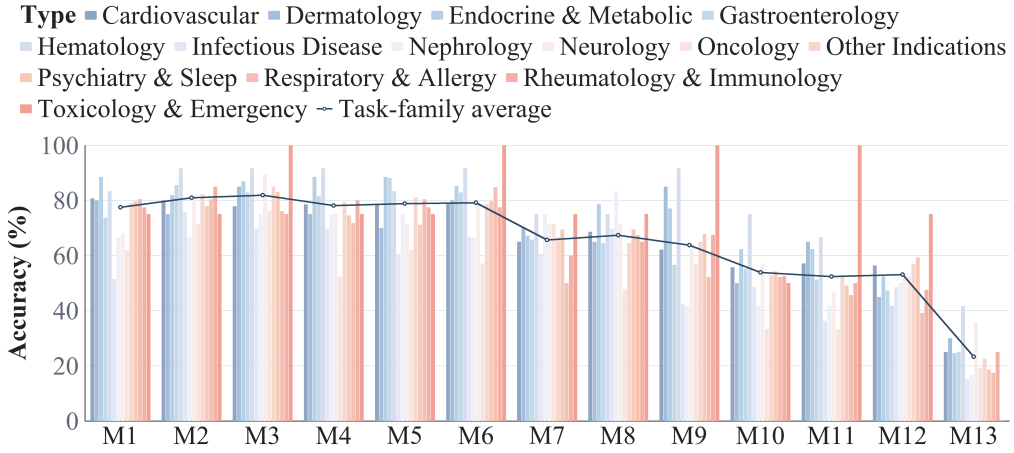
(f) Drug Comb.



(g) Ontology & Norm.



(h) Drug Reposition.



(i) Indication Reasoning

Figure E.1: Fine-grained subcategory accuracy (%) for (a) PhysChem Props., (b) PhysChem Rules, (c) Drug Relations, (d) Metab. & Clinical, (e) Toxicity & Safety, (f) Drug Comb., (g) Ontology & Norm., (h) Drug Reposition., and (i) Indication Reasoning. M1–M13 follow Table D1; the line shows the task-family average. The panels expose variation among subcategories within each family, which an overall accuracy score can conceal.

F Detailed Explanation Quality Evaluation

The model-level dimension scores are reported in Section 4.2.2. This appendix provides the evaluation rubric, the analysis by answer correctness, and a recorded explanation example.

F.1 Evaluation Dimensions and Rubric

Explanations are assessed for factual support, logical coherence, clinical relevance, and precision of expression. Each dimension uses a five-point scale, with higher scores indicating better explanations.

Table F1: Dimensions used to evaluate explanation quality.

Dimension	**What the dimension measures**
Factual support	Whether the explanation is supported by pharmaceutical facts or item evidence.
Logical coherence	Whether the premises support the conclusion coherently.
Clinical relevance	Whether the explanation addresses the target pharmaceutical or clinical question.
Precision of expression	Whether the explanation is focused, concise, and faithful to the given facts.

Table F2: Interpretation of the 1–5 scores for each evaluation dimension.

Score	**Interpretation**
1	Absent or clearly fails the dimension.
2	Major errors or substantial irrelevant content.
3	Partially satisfies the dimension, with important deficiencies.
4	Largely satisfies the dimension, with minor deficiencies.
5	Fully satisfies the dimension.

Table F1 defines the four dimensions, and Table F2 defines the score levels. Factual support concerns the premises used in an explanation. Logical coherence concerns whether those premises support its conclusion. Clinical relevance concerns whether it addresses the question, while precision concerns focused and faithful expression. Reporting all four dimensions helps identify weaknesses that their overall mean can hide.

F.2 Explanation Quality by Answer Correctness

Table F3 compares overall explanation scores for correct and incorrect final answers.

Table F3: Explanation quality by final-answer correctness on a 1–5 scale. The difference is the correct-answer mean minus the incorrect-answer mean.

Model	Correct answers	Incorrect answers	Difference
GPT-5.6 Terra	3.06	2.84	+0.22
DeepSeek-V4.1-flash	**3.57**	3.31	+0.26
GLM-5.3	3.31	3.33	-0.02
MiniMax-M3	3.49	3.28	+0.21
Kimi-k3	3.56	**3.46**	+0.10
Qwen-3.8-max	3.42	3.26	+0.16
Gemma-4-E4B-it	3.18	2.80	+0.38
Qwen3.5-9B	3.21	2.78	+0.43
Hulu-Med-7B	2.98	2.14	+0.84
Mistral-7B-Instruct-v0.3	2.74	2.42	+0.32
ATHENA-R1-Qwen3-8B	2.98	2.43	+0.55
MedGemma-4B-it	3.15	2.02	+1.13
Pharma Assistant v3	1.07	1.06	+0.01

Table F3 groups explanations by final-answer correctness. Correct-answer explanations receive higher mean scores for 12 of the 13 models, but the gap varies. It is 1.13 points for MedGemma-4B-it and 0.84 for Hulu-Med, compared with 0.10 for Kimi and 0.01 for Pharma Assistant v3. Kimi scores 3.56 on correct answers and 3.46 on incorrect answers. GLM shows a small reversal at 3.31 and 3.33.

The size of the gap must be read alongside the scores themselves. Kimi's two groups both score relatively highly, whereas Pharma Assistant v3 remains near the bottom of the scale in both groups. A small gap therefore has different meanings across models. These group averages show how explanation quality relates to correctness, but do not establish a threshold for deciding whether an individual answer is correct.

F.3 Recorded Explanation Example

The recorded Droxidopa case shows the distinction within one response. The final answer is correct, while the explanation receives 4/5 for factual support and logical coherence and 3/5 for clinical relevance and precision. Its mean of 3.50/5 records useful support together with remaining weaknesses. This is why the assessment retains both the answer label and the four explanation scores.

Example F · GPT-5.6 Terra explanation · MCR_483

Question
Which enzyme or catalytic system has a documented metabolism, activation, or inactivation relationship with Droxidopa? Options: A. acxB; B. DDO; C. DDC; D. 10-LAH.

Reference answer C. DDC

Model answer C

Model explanation (verbatim)
Droxidopa is a synthetic precursor of norepinephrine that is converted to norepinephrine by dopa decarboxylase (DDC), also called aromatic L-amino acid decarboxylase. DDC catalytic activity activates droxidopa rather than inactivating it, increasing formation of norepinephrine and producing its pressor effect in neurogenic orthostatic hypotension. Inhibition of DDC can reduce this conversion and may decrease the pharmacodynamic response, although peripheral decarboxylase inhibition does not completely prevent CNS or extra-neuronal conversion. Therefore, DDC is the documented enzyme relationship in option C.

Recorded judgment
Answer correct; overall explanation score 3.50/5. Factual support 4/5; logical coherence 4/5; clinical relevance 3/5; precision of expression 3/5.

Source record Drug Name: Droxidopa; Enzyme Name: dopa decarboxylase (DDC)

G Detailed Results for Web-based Knowledge Retrieval

G.1 Overall Effects and Answer Transitions

Table G1 compares closed-book and search-enabled answers on the same 6,174 questions. The net gain equals the number of wrong-to-correct transitions minus correct-to-wrong transitions, divided by 6,174 and expressed in percentage points. The two transition columns reveal changes that an overall difference alone would conceal.

Table G1: Paired closed-book and search-enabled results for all 13 models. Accuracies are percentages; changes are percentage points. Changes are calculated from unrounded counts over 6,174 paired questions.

Model	Closed-book accuracy	Search-enabled accuracy	Net change (pp)	Wrong → correct	Correct → wrong
GPT-5.6 Terra	**77.42**	83.25	+5.83	480	120
DeepSeek-V4.1-flash	77.36	79.77	+2.41	213	64
GLM-5.3	77.15	**84.08**	+6.93	539	111
MiniMax-M3	76.26	77.50	+1.25	150	73
Kimi-k3	75.22	74.34	-0.87	89	143
Qwen-3.8-max	74.62	78.77	+4.15	370	114
Gemma-4-E4B-it	62.78	63.51	+0.73	112	67
Qwen3.5-9B	56.14	59.12	+2.98	245	61
Hulu-Med-7B	52.38	53.94	+1.55	153	57
Mistral-7B-Instruct-v0.3	49.43	47.51	-1.93	67	186
MedGemma-4B-it	48.72	49.14	+0.42	98	72
ATHENA-R1-Qwen3-8B	48.45	46.10	-2.35	54	199
Pharma Assistant v3	30.52	38.32	**+7.81**	530	48

Table G1 shows overall improvements for ten of the 13 models and declines for Kimi-k3, Mistral-7B-Instruct-v0.3, and ATHENA-R1-Qwen3-8B. GLM achieves the highest search-enabled accuracy at 84.08%, followed by GPT at 83.25%. Pharma Assistant v3 has the largest net gain, +7.81 percentage points, but its search-enabled accuracy remains 38.32%. GPT corrects 480 initial errors and reverses 120 correct answers, yielding 360 net corrections. Hulu-Med-7B corrects 153 errors and reverses 57 correct answers, yielding 96 net corrections. Reporting both transition counts distinguishes the benefits of retrieval from the errors it introduces.

G.2 Task-dependent Retrieval Effects

Figure 3 shows five families that improve for every model and three that decline for every model. The improving families are PhysChem Props., Drug Relations, Toxicity & Safety, Drug Comb., and Ontology & Norm. They contain questions about recorded attributes, entity mappings, and documented relations that can benefit from external information. The declining families are PhysChem Rules, Indication Reasoning, and Drug Repositioning. Metab. & Clinical has mixed effects, improving for 11 models and declining for Mistral and ATHENA.

The properties–rules comparison shows why information access and information use need separate assessment. GPT improves by 15.00 percentage points on physicochemical properties but declines by 7.28 points on physicochemical rules. Both involve drug properties, yet the latter requires applying the relevant criteria to reach a judgment. Repositioning questions require evidence to meet selection and comparison constraints, while indication questions require matching a drug–disease relation to the specific question. These contrasts suggest that retrieval is most useful when the returned information directly supports the required answer. Its value also depends on selecting and applying that information under the task's conditions.

The task-family pattern concerns the use of retrieved information, rather than a simple split between factual and reasoning questions. Some improving families include classification and rule application, and some declining families include relation lookup. The operation labels in Appendix B help

interpret these differences. This motivates the compositional study, which controls how relevant premises are presented instead of leaving their selection and organization to web retrieval.

G.3 Representative Retrieval Cases

Example G · DeepSeek · wrong to right · PP_015

Question What is the molecular formula of Ridogrel? Fill in the blank: ____

Reference answer **C18H17F3N2O3**

Closed-book answer C18H17F3N2O2

Search-enabled answer C18H17F3N2O3

Search query
What is the molecular formula of Ridogrel? Fill in the blank: ____ pharmacology drug database

Retrieved result 1
Ridogrel [www.drugfuture.com]. Title: Ridogrel CAS Registry Number: 110140-89-1 CAS Name: 5-[[(E)-[3-Pyridinyl[3-(trifluoromethyl)phenyl]methylene]amino]oxy]pentanoic acid Manufacturers' Codes: R-68070 Molecular Formula: C18H17F3N2O3 Molecular Weight: 366.33 Percent Composition: C 59.02%, H 4.68%, F 15.56%, N 7.65%, O 13.10% Literature References: Combined thromboxane A2 synthase inhibitor and thromboxane A2/prostaglandin . . .

Retrieved result 2
[Non-English text in retrieval record; source: www.drugfuture.com] [www.drugfuture.com].
[Non-English text in retrieval record; source: www.drugfuture.com]
The excerpts record the information returned by retrieval; the reference answer determines the correctness of the final response.

Source record Drug Name: Ridogrel

Example G · DeepSeek · right to wrong · PP_005

Question
What molecular-weight range best matches Valganciclovir? Fill in the blank: ____

Reference answer **350 to $<$ 500 g/mol**

Closed-book answer 350–400 g/mol

Search-enabled answer 300–400 Da

Search query
What molecular-weight range best matches Valganciclovir? Fill in the blank: ____ pharmacology drug database

Retrieved result 1
Valganciclovir 175865-60-8 [www.chemicalbook.com]. Search Products: Chemical Name: Valganciclovir Synonyms
Valgancyclovir;Valganciclov;valganciclovir;Valganciclovir-d8;Valganciclovir-d8 (Mixture of Diastereomers);Valganciclovir & Valganciclovir Hydrochloride;Valganciclovir-d5 HCl (Mixture of diastetroMers);L-Valine, 2-[(2-amino-1,6-dihydro-6-oxo-9H-purin-9-yl)methoxy]-3-hydroxypropyl ester;(2S)-2-((2-AMino-

Retrieved result 2
[Non-English text in retrieval record; source: www.wuxuwang.com] [www.wuxuwang.com].
[Non-English text in retrieval record; source: www.wuxuwang.com]

Source record Drug Name: Valganciclovir

Example G · GPT · answer reformulation · PP_007

Question
What molecular-weight range best matches Voclosporin? Fill in the blank: ____

Reference answer ≥500 g/mol

Closed-book answer 1,200–1,300 g/mol

Search-enabled answer 1000-1500 g/mol

Search query
What molecular-weight range best matches Voclosporin? Fill in the blank: ____ pharmacology drug database

Retrieved result 1
[Non-English text in retrieval record; source: www.e-biochem.com] [www.e-biochem.com]. [Non-English text in retrieval record; source: www.e-biochem.com]

Retrieved result 2
Metrics other than potency reveal systematic variation in responses to cancer drugs - Dataset - HMS LINCS Database - HMS LINCS Project [lincs.hms.harvard.edu]. Metrics other than potency reveal systematic variation in responses to cancer drugs - Dataset (ID:20120) Excel (XLSX) Page 69 of 134 1701 to 1725 of 3339 records Small Molecule Cell Line log10[max. dose (M)] R^2 (goodness of fit) log10[EC50 (M)] log10[IC50 (M)] log10[GI50 (M)] HillSlope E_inf E_max AUC Doxorubicin BT-20 -6.2596 0.9277 -6.3089 -6.3089 -6.5270 2.3340 0.0000 0.4342 8.2808 . . .

Source record Drug Name: Voclosporin

The Ridogrel case shows a retrieved formula matching the reference after an incorrect closed-book answer. The molecular-weight cases illustrate why a changed response must be interpreted against the required answer category. For Voclosporin, both displayed intervals lie above the reference threshold of 500 g/mol. The response changes its numerical interval but remains within the same reference category. This case distinguishes answer reformulation from a demonstrated correction.

H Detailed Results for Compositional Knowledge Use

H.1 Performance across Knowledge Conditions

Table H1: Recorded accuracy on 451 compositional instances (%).

Model	All atomic facts correct	Direct	Guided	Correct atom
GPT-5.6 Terra	**49.00**	**78.27**	98.89	99.33
DeepSeek-V4.1-flash	39.47	72.28	**99.11**	**99.78**
Qwen-3.8-max	33.70	74.06	98.67	99.56
Gemma-4-E4B-it	27.49	68.96	89.36	93.57
Hulu-Med-7B	9.31	47.89	51.66	48.78

Table H1 compares the same 451 targets across premise conditions. Guided raises GPT, DeepSeek, and Qwen from Direct accuracies of 78.27%, 72.28%, and 74.06% to 98.89%, 99.11%, and 98.67%. Correct atom reaches 99.33%, 99.78%, and 99.56%. Much of the improvement therefore occurs when the models first state the premises themselves, before receiving reference facts. The results connect premise organization to target success as well as showing the benefit of supplied premises.

The benefits differ across models. Gemma rises from 68.96% under Direct to 89.36% under Guided and 93.57% under Correct atom. Hulu-Med rises from 47.89% to 51.66% under Guided but reaches 48.78% under Correct atom. Relative to Direct, Correct atom adds 95 correct answers for GPT, 124 for DeepSeek, 115 for Qwen, 111 for Gemma, and four for Hulu-Med. Reference premises improve aggregate accuracy for all five models, but do not provide the same benefit or produce reliable answers in every case.

H.2 Target Correctness by Atomic-Probe Performance

Table H2: Joint distribution of atomic-knowledge scores and target-question correctness under direct answering.

	All atomic facts correct			Atomic facts not all correct		
Model	Items	Target correct	Conditional accuracy (%)	Items	Target correct	Conditional accuracy (%)
GPT-5.6 Terra	221	213	96.38	230	140	60.87
DeepSeek-V4.1-flash	178	174	97.75	273	152	55.68
Qwen-3.8-max	152	149	**98.03**	299	185	61.87
Gemma-4-E4B-it	124	102	82.26	327	209	**63.91**
Hulu-Med-7B	42	17	40.48	409	199	48.66

Table H2 groups Direct responses by performance on the separate atomic probes. Each conditional accuracy divides the correct targets by the number of items in that group. For GPT, DeepSeek, Qwen, and Gemma, accuracy in the all-correct group is 96.38%, 97.75%, 98.03%, and 82.26%, compared with 60.87%, 55.68%, 61.87%, and 63.91% in the remaining group. Hulu-Med shows the reverse pattern at 40.48% and 48.66%.

Correct probe answers are associated with higher target accuracy for four models, but do not guarantee it. Gemma answers 102 of 124 targets correctly in the all-correct group, while Hulu-Med answers 17 of 42. Conversely, every model answers some targets correctly despite errors on the probes. Some targets already name the enzyme and activity change, whereas the atomic probe separately asks for the drug–enzyme relation. The two scores therefore capture performance under different information conditions, which explains why they should be reported separately.

H.3 Direct-to-Correct Atom Transition Analysis

Table H3: Recorded item-level correctness transitions from direct answering to correct atom (counts).

Model	Correct → correct	Wrong → correct	Correct → wrong	Wrong → wrong
GPT-5.6 Terra	351	97	2	1
DeepSeek-V4.1-flash	325	125	1	0
Qwen-3.8-max	332	117	2	0
Gemma-4-E4B-it	310	112	1	28
Hulu-Med-7B	189	31	27	204

Table H3 explains the net changes by tracking each Direct answer after reference premises are supplied. GPT, DeepSeek, and Qwen correct 97, 125, and 117 errors while introducing two, one, and two, respectively. Gemma corrects 112 errors, introduces one, and retains 28. Its gain of 111 correct answers reflects many corrections with few reversals.

Hulu-Med corrects 31 errors but reverses 27 correct answers and retains 204 errors. Its gain of four answers therefore hides changes in both directions. Together, the three tables distinguish improvement in overall target accuracy, the association with atomic-probe correctness, and the reliability of the premise intervention on individual items. They show why access to correct premises and successful use of those premises require separate assessment.

H.4 Representative Compositional Reasoning Cases

Example H · `MCR_179` · Five-model comparison

Target question
The source database identifies Cytochrome P450 2C9 (CYP2C9) as a primary metabolizing enzyme for Paramethadione. If Cytochrome P450 2C9 (CYP2C9) activity is markedly reduced while dose, bioavailability, and other clearance pathways remain unchanged, how does parent-drug AUC change?

Reference answer **Parent-drug AUC increases**

Atomic fact A1: primary enzyme Cytochrome P450 2C9 (CYP2C9)

Atomic fact A2: effect of reduced enzyme activity Metabolism decreases

Recorded model responses
A1 and A2 test the two atomic facts. Direct, guided, and correct atom are answers to the composition target above.

DeepSeek-V4.1-flash
A1: Cytochrome P450 2C9 (CYP2C9)
A2: Metabolism decreases
Direct: Parent-drug AUC increases.
Guided: Parent-drug AUC increases.
Correct atom: Parent-drug AUC increases.

GPT-5.6 Terra
A1: Cytochrome P450 2C9 (CYP2C9)
A2: Metabolic clearance decreases
Direct: Paramethadione AUC increases
Guided: Parent-drug AUC increases
Correct atom: Parent-drug AUC increases

Qwen-3.8-max
A1: Cytochrome P450 2C19 (CYP2C19)
A2: Metabolic clearance decreases
Direct: Parent-drug AUC increases

Example H / `MCR_179` (continued)

Guided: Parent-drug AUC increases
Correct atom: Parent-drug AUC increases

Gemma-4-E4B-it
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: Parent-drug AUC increases
Guided: Parent-drug AUC increases
Correct atom: Parent-drug AUC increases

Hulu-Med-7B
A1: CYP2C19
A2: Metabolism decreases
Direct: increases
Guided: Increases
Correct atom: increases

Case reading
This case separates identification of the primary enzyme from the final AUC judgment: several models name a different enzyme in A1 while still predicting the correct AUC direction.

Source record Drug Name: Paramethadione; Enzyme Name: Cytochrome P450 2C9 (CYP2C9)

Example H · `MCR_296` · Five-model comparison

Target question
In a patient receiving Letrozole, Cytochrome P450 2A6 (CYP2A6)-dependent metabolism is markedly reduced. Holding apparent distribution volume and other elimination mechanisms constant, predict the direction of change in the drug's elimination rate constant.

Reference answer **The elimination rate constant decreases**

Atomic fact A1: primary enzyme Cytochrome P450 2A6 (CYP2A6)

Atomic fact A2: effect of reduced enzyme activity Metabolism decreases

Recorded model responses
A1 and A2 test the two atomic facts. Direct, guided, and correct atom are answers to the composition target above.

DeepSeek-V4.1-flash
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolism decreases
Direct: The elimination rate constant decreases.
Guided: The elimination rate constant decreases.
Correct atom: The elimination rate constant decreases.

GPT-5.6 Terra
A1: Cytochrome P450 2A6 (CYP2A6)
A2: Metabolic clearance decreases
Direct: The elimination rate constant decreases.
Guided: The elimination rate constant decreases.
Correct atom: The elimination rate constant decreases.

Qwen-3.8-max
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: decreases
Guided: The elimination rate constant decreases
Correct atom: The elimination rate constant decreases

Gemma-4-E4B-it
A1: Cytochrome P450 3A4 (CYP3A4)

Example H / `MCR_296` (continued)

A2: Metabolic clearance decreases
Direct: decreases
Guided: The elimination rate constant decreases.
Correct atom: The elimination rate constant decreases.

Hulu-Med-7B
A1: CYP3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: increases
Guided: The elimination rate constant decreases
Correct atom: Elimination rate constant decreases

Case reading
Hulu-Med-7B predicts an increase when answering the target directly, but a decrease when responding with elicited or correct atoms. The five-model comparison also shows that an incorrect A1 answer does not always imply an incorrect target answer.

Source record Drug Name: Letrozole; Enzyme Name: Cytochrome P450 2A6 (CYP2A6)

Example H · `MCR_001` · Five-model comparison

Target question
For Cyclobenzaprine, metabolism through the primary enzyme Cytochrome P450 1A2 (CYP1A2) becomes much slower. If the apparent volume of distribution remains stable, indicate the expected direction of change in the first-order elimination rate constant.

Reference answer **The elimination rate constant decreases**

Atomic fact A1: primary enzyme Cytochrome P450 1A2 (CYP1A2)

Atomic fact A2: effect of reduced enzyme activity Metabolism decreases

Recorded model responses
A1 and A2 test the two atomic facts. Direct, guided, and correct atom are answers to the composition target above.

DeepSeek-V4.1-flash
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolism decreases
Direct: The elimination rate constant decreases.
Guided: The elimination rate constant decreases.
Correct atom: The elimination rate constant decreases.

GPT-5.6 Terra
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: The elimination rate constant decreases.
Guided: The elimination rate constant decreases.
Correct atom: The elimination rate constant decreases.

Qwen-3.8-max
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: The elimination rate constant decreases
Guided: The elimination rate constant decreases
Correct atom: The elimination rate constant decreases

Gemma-4-E4B-it
A1: Cytochrome P450 3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: decreases
Guided: The elimination rate constant decreases

Example H / `MCR_001` (continued)

Correct atom: The elimination rate constant decreases.

Hulu-Med-7B
A1: CYP3A4 (CYP3A4)
A2: Metabolic clearance decreases
Direct: decreases
Guided: Decreases
Correct atom: The elimination rate constant decreases.

Case reading
All five models name CYP3A4 rather than the reference enzyme CYP1A2 in A1, yet their target responses predict the expected decrease in the elimination rate constant. The atom and composition scores therefore measure distinct behavior.

Source record Drug Name: Cyclobenzaprine; Enzyme Name: Cytochrome P450 1A2 (CYP1A2)

In these three cases, the target question names the enzyme and states the change in its activity, while the separate atomic probe asks for the drug–enzyme association. A model can therefore give the expected target direction even when its separately elicited enzyme answer is wrong. The recorded responses illustrate why atomic probing and target answering should be interpreted in their respective question contexts.

I Detailed Counterfactual Sensitivity Results

I.1 Overall Counterfactual Results

Table I1: Counterfactual sensitivity across 451 question triplets (%).

Model	Original acc.	Paraphrase acc.	Intervention acc.	Both correct	Paraphrase consistency	Intervention answer change
GPT-5.6 Terra	**78.27**	**78.49**	**67.18**	**61.20**	**85.81**	72.28
DeepSeek-V4.1-flash	72.28	71.84	60.53	54.55	80.27	65.19
Qwen-3.8-max	74.06	74.06	63.19	57.43	81.82	68.74
Gemma-4-E4B-it	68.96	68.74	54.10	49.67	74.72	58.76
Hulu-Med-7B	47.89	47.23	38.58	33.92	61.42	43.90

Table I1 reports accuracy for each variant alongside joint correctness and response behavior. All six columns use the same 451 items. Both correct requires correct answers to the original and intervention questions. Paraphrase consistency counts original–paraphrase agreement, while intervention answer change counts different answers across original–intervention pairs. Agreement and change are evaluated separately from correctness.

Paraphrase accuracy is close to original accuracy for all five models, yet paraphrase consistency ranges from 61.42% to 85.81%. The paired consistency scores reveal answer changes that similar aggregate accuracies can hide. Intervention accuracy is below original accuracy for every model. Reading accuracy and pairwise behavior together shows both the overall effect of the condition change and the stability of individual judgments.

I.2 Joint Analysis of Stability and Sensitivity

GPT leads original accuracy at 78.27%, intervention accuracy at 67.18%, and joint correctness at 61.20%. Of its 353 correct original answers, 276 remain correct under intervention and 77 become incorrect. Thus, 21.81% of its original successes fail after the key condition changes. This conditional view directly tests whether an initially correct judgment carries over to the corresponding intervention.

GPT changes its answer on 326 pairs, but only 276 pairs are correct on both variants. Because the reference direction reverses, jointly correct pairs must contain different answers. The remaining changed pairs fail on at least one variant. Figure 5 similarly shows answer-change rates above intervention accuracy for all five models. For GPT the rates are 72.28% and 67.18%, and for Hulu-Med they are 43.90% and 38.58%. A change in response is therefore evidence of sensitivity, while successful updating requires checking the corresponding answers against their references.

I.3 Representative Counterfactual Cases

Example I · Counterfactual triplet · `CF_MCR_430`

Original question
A loss of Glutathione S-transferase alpha-1 (GSTA1) activity reduces the primary metabolic pathway for Melphalan. Assuming linear kinetics and constant distribution volume, what happens to the time required for the parent-drug concentration to fall by half?

Original reference answer **The elimination half-life increases.**

Causal intervention GSTA1 activity **increases** rather than **decreases**.

Expected intervention answer **The elimination half-life decreases.**

Example I (continued)

Model	Original response	Paraphrase response	Intervention response
GPT-5.6 Terra	Increases	Increases	Increases ×
DeepSeek-V4.1-flash	increased	increased	decreased ✓
Qwen-3.8-max	increases	increases	increases ×
Gemma-4-E4B-it	increases	increases	increases ×
Hulu-Med-7B	Increased half-life	increased half-life	decreased half-life ✓

Source record
Drug Name: Melphalan; Enzyme Name: Glutathione S-transferase alpha-1 (GSTA1)

All five models answer the original Melphalan question and its paraphrase correctly. After enzyme activity reverses, DeepSeek and Hulu-Med predict the expected shorter half-life, while GPT, Qwen, and Gemma retain the original direction. Holding the endpoint and assumptions fixed makes the failure specific to updating the judgment after the condition change. The case illustrates the distinction between consistency and appropriate change that the aggregate metrics measure.

J Detailed Results for Response Modality Variation

J.1 Overall Response Modality Results

Table J1: Response modality results on 2,877 paired instances (%). Lower position range is better.

Answer format

Model	Open-ended answer	Original MCQ	Difference (pp)
GPT-5.6 Terra	12.13	**79.63**	+67.50
DeepSeek-V4.1-flash	**16.02**	79.21	+63.19
Qwen-3.8-max	14.39	77.44	+63.05
Gemma-4-E4B-it	9.77	68.30	+58.53
Hulu-Med-7B	8.72	64.27	+55.55

Permutation stability

Model	Permutation mean	All four correct	Position range (pp)	Semantic consistency
GPT-5.6 Terra	**78.13**	**61.45**	**0.90**	**69.38**
DeepSeek-V4.1-flash	69.58	50.57	8.79	61.23
Qwen-3.8-max	74.28	56.03	2.22	64.87
Gemma-4-E4B-it	55.52	31.70	13.11	39.42
Hulu-Med-7B	44.28	13.28	13.97	15.51

The answer-format comparison in Table J1 shows open-ended accuracy of 8.72%–16.02%, compared with 64.27%–79.63% for original MCQ. Within-model gaps range from 55.55 to 67.50 percentage points. GPT leads original MCQ accuracy, while DeepSeek leads open-ended accuracy. Thus, both the score and the model ranking depend on how the answer is requested. Appendix D specifies deterministic MCQ matching and semantic judging of open-ended responses.

The permutation measures examine stability when the candidates remain available. GPT averages 78.13% across permutations but answers 61.45% of items correctly in all four positions. For Hulu-Med, these values are 44.28% and 13.28%. The average measures success over individual presentations, while all-four-correct identifies items answered correctly throughout. Semantic consistency ranges from 15.51% to 69.38% and measures agreement in selected content, including agreement on an incorrect option. Reading these measures together distinguishes average performance, repeated correctness, and repeated selection.

J.2 Performance by Correct Option Position

Table J2: Accuracy by position of the correct option (%).

Model	A	B	C	D
GPT-5.6 Terra	**78.59**	**77.69**	**78.14**	**78.10**
DeepSeek-V4.1-flash	69.34	65.38	74.17	69.41
Qwen-3.8-max	74.31	73.17	75.39	74.24
Gemma-4-E4B-it	55.54	48.97	62.08	55.47
Hulu-Med-7B	44.32	37.30	51.27	44.25

Table J2 locates the differences summarized by the position range. B has the lowest accuracy for every model. C has the highest accuracy for DeepSeek, Qwen, Gemma, and Hulu-Med, while A is highest for GPT. Gemma scores 48.97% at B and 62.08% at C, a range of 13.11 points. Hulu-Med scores 37.30% and 51.27%, a range of 13.97 points.

GPT's range is only 0.90 points, yet its all-four-correct rate is 61.45%. Similar averages across positions can result even when success shifts between individual questions. Position range therefore

describes aggregate sensitivity, and the paired measures show whether the same questions remain correct. These complementary views support checking format stability alongside accuracy.

J.3 Representative Option-Permutation Case

Example J · `AF_DRL_002` · Option permutation / five-model comparison

Original multiple-choice question
Which enzyme has a documented metabolic, bioactivation, or drug-inactivation relationship with Ticarcillin? Options: A. blaB; B. bla; C. DPYD; D. 10-LAH

Reference answer **blaB**

Permutation positions The correct option (blaB) appears at A, D, C, and B in P1 to P4, respectively.

DeepSeek-V4.1-flash

Open-ended beta-lactamase

Permutations
P1: B (bla; incorrect); P2: B (DPYD; incorrect); P3: C (blaB; correct); P4: B (blaB; correct)

GPT-5.6 Terra

Open-ended Beta-lactamase

Permutations
P1: A (blaB; correct); P2: D (blaB; correct); P3: B (10-LAH; incorrect); P4: B (blaB; correct)

Qwen-3.8-max

Open-ended Beta-lactamase

Permutations
P1: B (bla; incorrect); P2: C (10-LAH; incorrect); P3: D (bla; incorrect); P4: C (bla; incorrect)

Gemma-4-E4B-it

Open-ended Penicillin-binding proteins (PBPs)

Permutations
P1: C (DPYD; incorrect); P2: B (DPYD; incorrect); P3: A (DPYD; incorrect); P4: D (DPYD; incorrect)

Hulu-Med-7B

Open-ended
P-glycoprotein (P-gp) has a documented drug-inactivation relationship with Ticarcillin. Ticarcillin is a substrate for P-gp, which is involved in the efflux of drugs out of cells, thereby reducing its intracellular concentration and potentially decreasing its efficacy.

Permutations
P1: C (DPYD; incorrect); P2: B (DPYD; incorrect); P3: A (DPYD; incorrect); P4: D (DPYD; incorrect)

Source record Drug Name: Ticarcillin; Enzyme Name: blaB

The Ticarcillin case shows why selected content must be tracked across option permutations. GPT selects the correct enzyme in three permutations, while DeepSeek does so in two. Gemma and Hulu-Med repeatedly select DPYD, which is incorrect throughout. Their changing option letters still refer to the same wrong content. The example therefore separates consistency from correctness and shows how stable aggregate scores can hide unstable answers on particular items.