VARIANTBENCH: BENCHMARKING LANGUAGE MODELS ON SCIENTIFIC REASONING ACROSS THE PHARMACOGENOMIC EVIDENCE PIPELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly serve as reasoning engines over scientific literature, yet it remains unclear whether they can sustain logical consistency across the multi-stage workflows required for real-world literature analysis. We introduce VARIANTBENCH, a benchmark that mirrors the full pharmacogenomic evidence curation pipeline grounded in expert-curated annotations from the ClinPGx research team. The benchmark comprises 79,592 structured single-paper questions and 394 agentic cross-document and clinical reasoning tasks spanning three tiers of complexity: factual extraction, dependent multi-turn reasoning, and CPIC guideline recreation under zero-context and evidence-provided settings. Evaluating frontier tool-use agents with the Harbor framework reveals substantial brittleness in multi-step reasoning. While per-step accuracy on chained tasks exceeds 60%, requiring all steps in a chain to be correct reduces success to 22.7%. Cross-document synthesis further degrades performance relative to single-paper comprehension. For clinical guideline recreation, providing the referenced literature improves mean reward by 20 points, indicating that models benefit substantially from explicit evidence access but remain unreliable when relying solely on parametric recall. VARIANTBENCH provides deterministic verifiers, reproducible agent infrastructure, and a large-scale expert-grounded evaluation suite for measuring progress toward robust scientific reasoning. [1]

1 INTRODUCTION

Reasoning over scientific literature demands more than single-step comprehension. Consider a researcher evaluating a pharmacogenomic study: one must first identify whether a genetic variant is reported as a significant predictor of drug response, then determine the nature of the association before extracting the supporting statistics and judging the strength of the evidence. This body of evidence must then be distilled into an actionable clinical guideline that a physician can use at the point of care. Each step depends on the last, and a competent scientific reader must maintain logical consistency when recalling specific claims from individual papers and synthesize findings across multiple studies to build a coherent picture of the evidence landscape for a given variant for effective clinical application.

Large language models (LLMs) are increasingly deployed as scientific reasoning agents, with growing interest in their use for literature review, evidence synthesis, and clinical decision support (Zhao et al., 2025). Yet it remains unclear whether these models can sustain the kind of chained, dependent reasoning that scientific analysis requires—or whether they merely approximate it through pattern matching over surface-level cues.

We introduce VARIANTBENCH, a benchmark that mirrors the full pharmacogenomic evidence curation pipeline, organised into three tiers of increasing reasoning complexity. VARIANTBENCH is grounded in the ClinPGx dataset (Whirl-Carrillo et al., 2012), a curated resource built over two decades of expert annotation of pharmacogenomic associations from scientific literature. Every question in VARIANTBENCH is derived from, and evaluated against, annotations that have been verified by domain specialists, providing a uniquely rigorous foundation for benchmark construction.

[1] Data and code will be released upon publication.

Built on this foundation, VARIANTBENCH comprises over 79,000 expert-grounded questions and nearly 400 agentic tasks spanning five task types across three tiers: *Tier 1 (single-paper comprehension)*, including multiple-choice questions that probe factual recall and self-consistency within individual studies, and chained four-step question answering that tests sustained reasoning across predictor identification, significance judgement, phenotype classification, and evidence selection; *Tier 2 (cross-document synthesis)*, requiring models to identify relevant findings about a genetic variant from a corpus containing majority-irrelevant papers; and *Tier 3 (clinical reasoning)*, evaluating the ability to reproduce prescribing guidelines created by the Clinical Pharmacogenomics Implementation Consortium (CPIC) Caudle et al. (2014) both with and without access to source literature. Together, these tiers test whether models that appear competent on isolated questions can maintain that competence as reasoning demands scale from single-paper extraction to evidence-based clinical recommendation.

We evaluate a set of frontier language models and command-line agents on VARIANTBENCH and find systematic performance degradation as reasoning complexity increases. Models achieve strong accuracy on single-paper factual extraction but degrade on chained reasoning, cross-document synthesis, and clinical guideline generation. Providing source literature for guideline recreation substantially improves scores over zero-context baselines, confirming that models can reason from evidence when supplied but that their pre-training recall of clinical knowledge is unreliable.

Our contributions are as follows:

1. We introduce VARIANTBENCH, the first benchmark to evaluate the full pipeline of scientific reasoning—from single-paper analysis through cross-document synthesis to clinical guideline generation—grounded in two decades of expert-curated pharmacogenomic annotations.

2. We propose a tiered evaluation framework comprising five task types across three levels of reasoning complexity, including a controlled comparison of evidence-based reasoning versus pre-training recall.

3. We provide an empirical evaluation of frontier language models and agents revealing systematic degradation across the pipeline: up to 93.7% on single-paper drug identification, but below 22.7% on full-chain reasoning, with clinical guideline quality improving 20 points when source literature is provided.

2 RELATED WORK

2.1 SCIENTIFIC AND BIOMEDICAL QUESTION ANSWERING.

A growing body of work evaluates language models on scientific literature understanding. **PubMedQA** (Jin et al., 2019) focuses on biomedical question answering grounded primarily in abstracts or curated corpora. More recent datasets extend to full-text reasoning: **QASPER** (Dasigi et al., 2021) requires evidence-based answers from NLP research papers, while **LitQA2-FullText**, **LitQA2-FullText-Search** (Skarlinski et al., 2024), and **ScholarQA-Bench** (Asai et al., 2026) evaluate long-context reasoning and retrieval across full scientific documents. **MultiXScience** (Lu et al., 2020) further tests cross-document scientific synthesis. These benchmarks primarily assess answer accuracy under prompted or retrieval-augmented settings.

2.2 BIOMEDICAL INFORMATION EXTRACTION.

Structured extraction from biomedical literature has been studied extensively in relation extraction benchmarks such as **BC5CDR** (Li et al., 2016) (chemical–disease relations) and the **DDIExtraction 2013 Challenge** (Chowdhury & Lavelli, 2013). BioCreative shared tasks similarly evaluate entity and relation extraction grounded in expert annotations. However, these datasets are typically sentence-level, limited in document scope, or not designed to replicate the full multi-step curation process used in clinical knowledge bases.

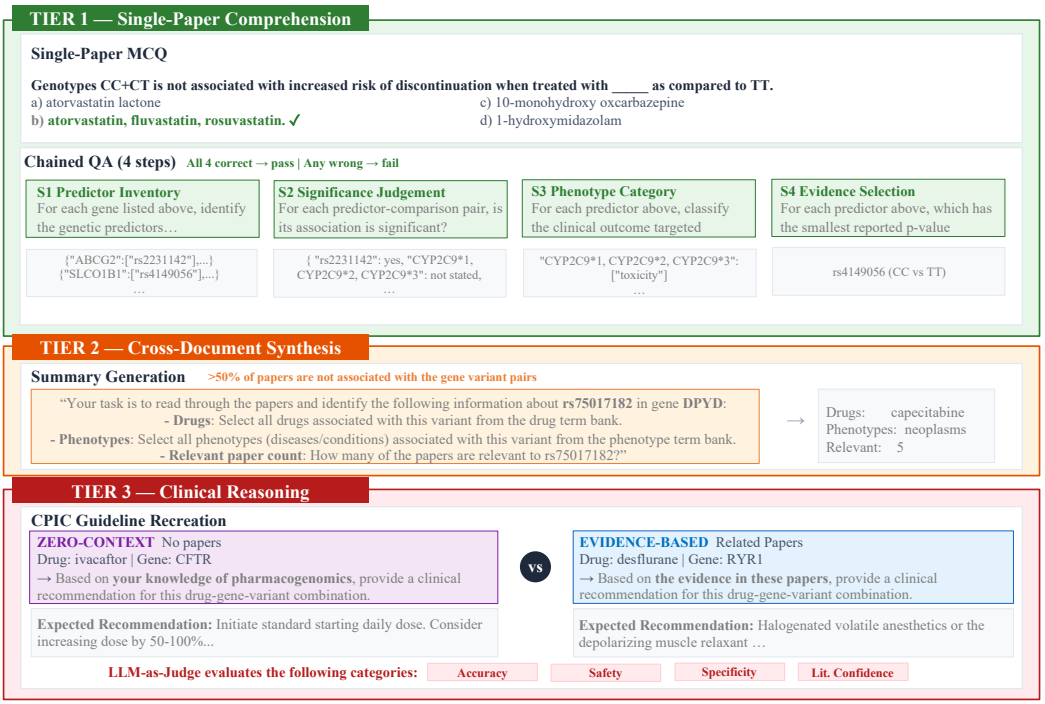

Figure 1: **Overview of the VariantBench benchmark.** VariantBench evaluates scientific reasoning across three tiers of increasing complexity, grounded in expert-curated ClinPGx annotations. Tier 1 tests single-paper comprehension through multiple-choice questions (left) and chained four-step question answering (right), where each step depends on the previous answer and chain-level credit requires all steps correct. Tier 2 tests cross-document synthesis by requiring the model to identify relevant drugs, phenotypes, and papers from a corpus containing $\geq$ 50% irrelevant documents. Tier 3 tests clinical reasoning by comparing CPIC guideline recreation under zero-context (pre-training recall only) and evidence-provided conditions, scored by an LLM judge on accuracy, safety, specificity, and literature confidence.

2.3 Distinction of Our Work

VariantBench differs from prior work in three respects. First, all questions are derived from large-scale expert-curated pharmacogenomic annotations, providing over 79,000 ground-truth comparisons—substantially larger than existing biomedical reasoning benchmarks such as PubMedQA (Jin et al., 2019) (∼1,000 questions) and QASPER (Dasigi et al., 2021) (∼5,000 questions). Second, the benchmark mirrors the real-world curation workflow end-to-end: models must identify relevant variants, judge statistical significance, classify clinical outcomes, and synthesize findings across documents, approximating the full annotation pipeline used to maintain pharmacogenomic knowledge bases. Third, we evaluate not only single-pass prompted models but also modern agentic and CLI-based tools that iteratively search, retrieve, and reason over documents, testing whether tool-augmented reasoning closes the gap that prompting alone cannot. Together, these design choices position VariantBench at the intersection of long-context scientific reasoning, structured biomedical extraction, and tool-based evaluation.

3 The VariantBench Benchmark

3.1 Data Source

All questions in VariantBench are grounded in expert-curated pharmacogenomic annotations from the ClinPGx (formerly PharmGKB) database (Whirl-Carrillo et al., 2012). This dataset com-

prises papers manually annotated into a standardized format, with each annotation capturing associations between a particular genetic variant and a studied outcome (differing responses to certain drugs, phenotype changes, etc.) along with the statistical significance of the findings. In total, the dataset includes over 9,000 manually analyzed articles, each containing on average 5 distinct annotations. Each annotation is linked to a PubMed ID (PMID) and, where available, a PubMed Central Open Access full-text article (PMC). Full-text papers are retrieved from PMC and converted to structured Markdown to preserve section headings, table structure, and inline references, enabling evidence grounding over actual study text rather than curated abstracts alone. We worked directly with the original curation team to convert the pre-existing annotation corpus into questions representative of their paper analysis process, with ground-truth values rooted in the original annotations.

3.2 Single Association QA

3.2.1 Term-Based Multiple Choice

As a preliminary step, we test each model's ability to identify the correct associations found by the curators by creating multiple-choice questions that remove a key term from the ground-truth annotation and prompt the model to fill in the blank. Distractors are drawn from three sources: Jaccard-similar terms from the broader corpus, terms that appear in the same paper but belong to a different annotation, and a "None of the above" option for cases where no correct answer is present. Since each annotation follows a standardized format, we generate questions across three entity types (drug, variant, and phenotype) by replacing the corresponding term in the template.

3.2.2 Statistical Significance (P-Value) Extraction

A critical step in paper analysis is accurately identifying the statistical significance of a reported association. In some cases this is straightforward, with authors providing p-values explicitly. However, part of this dataset's richness comes from the curation team having manually recalculated p-values from the experiment details in the paper and supplementary material. For each association, we task the model with (a) extracting the correct p-value or confidence interval and (b) determining whether the finding is statistically significant given the paper context.

3.3 Chain Structure

Each chained instance is derived from a single pharmacogenomic paper and comprises four sequentially dependent steps, with gold-standard outputs from earlier steps provided as context for later ones. The steps are: (1) *Predictor Inventory*—selecting the correct genetic predictors from a multiple-choice list with plausible distractors; (2) *Significance Judgement*—classifying each predictor–comparison pair as significant, not significant, or not stated; (3) *Phenotype Category*—assigning clinical outcome labels (`efficacy`, `toxicity`, `dosage`, `metabolism/PK`, `PD`, `other`); and (4) *Evidence Selection*—identifying the predictor–comparison pair with the smallest reported p-value. Each step uses a deterministic verifier, and chain-level credit is binary: a chain passes only if all four steps receive full credit. A complete example is in Appendix A.

3.4 Summary Generation

Beyond single-paper analysis, scientific curation requires synthesizing findings across a corpus of studies. The ClinPGx team maintains variant-level summaries that aggregate evidence across all papers reporting on a given variant–drug pair. We mirror this task by providing each agent with 10–30 research papers—at least 50% of which are irrelevant distractors—and requiring it to identify the relevant papers, extract associated drugs and phenotypes, and produce a structured summary that can be compared directly against the ground-truth annotation.

3.5 CPIC Guideline Recreation

As a final task, we evaluate whether models can synthesize clinically actionable recommendations from pharmacogenomic evidence. The Clinical Pharmacogenomics Implementation Consortium (CPIC) maintains dosing guidelines for common variant: drug pairs, each derived from a body of supporting literature. We benchmark models on reproducing these guidelines under two conditions:

1. **Zero-Context:** The model must recreate the guideline for a given variant (drug pair without access to any literature, isolating what clinical knowledge is retained from pre-training).
2. **Evidence-Provided:** The model receives all papers referenced in the original CPIC guideline and must produce a recommendation grounded in this evidence.

We compare each model's output to the original CPIC guideline using an LLM-as-a-judge (Zheng et al., 2023) approach, scoring along four axes: accuracy, safety, specificity, and literature confidence.

3.6 Dataset Statistics

Benchmark Type	Question Set	Task Type	# Questions / Tasks
Model QA	Term-Based MC	Multiple-choice	32,681 questions
Model QA	Statistical Significance	Numeric answer extraction	46,823 questions
Model QA	Chained Questions	Multi-turn reasoning	88 Tasks
Agent CLI	Summary QA	Cross-document extraction	100 tasks
Agent CLI	CPIC Zero-Context	Clinical guideline generation	100 tasks
Agent CLI	CPIC Evidence	Clinical guideline generation	106 tasks
Total (Questions)			**79,592**
Total (Agent Tasks)			**394 tasks**

Table 1: Unified scale of the VariantBench questions. VariantBench provides large-scale structured reasoning questions for preliminary understanding of model performance followed by cross-document and clinical reasoning tasks.

The **Model QA** style components of VariantBench comprises **79,504 total questions**. This includes 32,681 term-based multiple-choice questions and 46,823 statistical-significance extraction questions.

The **Agent CLI** components contain **394 agentic tasks**. These include 88 chained multi-turn reasoning tasks, 100 Summary QA tasks, 100 CPIC Zero-Context tasks, and 106 CPIC Evidence tasks. The chained tasks consist of 22 distinct paper instances (PMIDs), each expanded into a structured sequence of dependent reasoning steps (e.g., claim verification, evidence localization, statistical extraction, and objective evaluation).

Unlike the Model QA setting, Agent CLI tasks give models access to the context management and file lookup systems present in the Claude Code (Anthropic, 2026) and Codex (Chen et al., 2021) CLI tools. These tasks also typically require models to perform structured reasoning over many full-text papers and produce outputs dependent on many disparate sources of information.

Across all components, VariantBench spans multiple decades of pharmacogenomic research and covers a broad range of genes with diverse drug–phenotype associations.

Evaluation methodology. All tasks except for the CPIC analysis use deterministic verifiers: MCQ answers are compared by exact label match, structured JSON outputs (Steps 2–4) are compared field by field with controlled normalization (e.g., case-insensitive phenotype aliases, numeric tolerance for frequencies), and free-text evidence selection (Step 4) uses token-level normalization. This allows us to avoid using LLM-as-a-judge approach for the vast majority of our questions and even in the CPIC guideline case we only task the judge model to compare to a provided ground truth.

4 Experiments

4.1 Setup

We evaluate performance in two settings:

1. **Model QA:** Prompted model responses given a single paper as context, evaluated against ground-truth annotations.

Table 2: Performance summary across all three tiers of VARIANTBENCH evaluated with the Harbor framework. Chain Per-Turn reports average accuracy across the four individual steps; Chain Full requires all four steps correct. Summary Gen. evaluates cross-document synthesis. CPIC scores are LLM-as-judge ratings (0–100) averaged across accuracy, safety, specificity, and literature confidence.

Model	TIER 1: Single-Paper		TIER 2	TIER 3: Clinical	
	Chain Per-Turn	Chain Full	Summary Gen.	CPIC Zero	CPIC Evid.
Claude Code	64.7%	13.6%	**48%**	**44**	**65**
Codex	**68.2%**	**22.7%**	30%	33	55

2. **Agent CLI:** A CLI-based agent (Claude Code or Codex) operating in an isolated Docker container with access to papers and tools, evaluated using the Harbor framework.

Each Agent CLI task is executed with a 10-minute timeout, access to PubMed papers (pre-converted to Markdown), and a `variant_lookup.py` utility that normalizes variant terms via the ClinPGx API. The agent writes its answer to a designated file; a deterministic verifier then computes a reward in $[0, 1]$.

4.2 RESULTS

We present results across all three tiers of VARIANTBENCH. Table 3 reports Model QA performance on single-paper multiple-choice and statistical significance extraction. Table 2 summarizes Agent CLI performance across chained reasoning, cross-document synthesis, and clinical guideline recreation. Figure 2 visualizes the performance degradation across tiers.

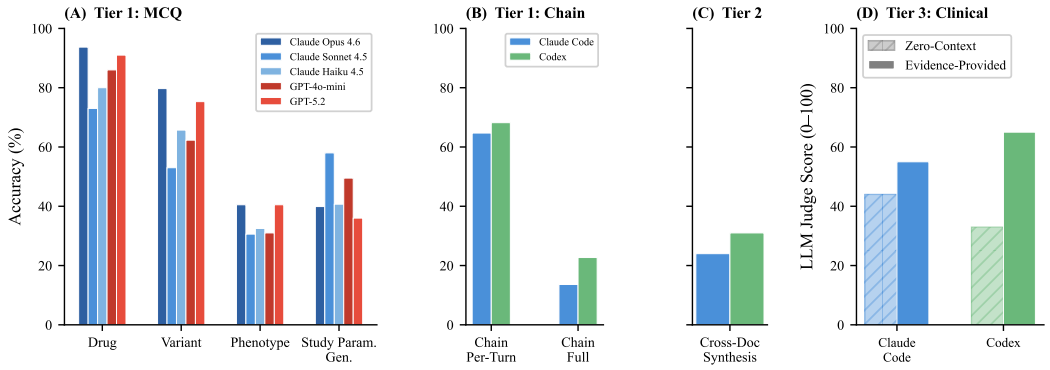

Figure 2: **Performance across the three tiers of VARIANTBENCH.** (A) Single-paper MCQ accuracy across five models and three entity types. (B) Chain per-turn and chain-level accuracy for two agent systems. (C) Cross-document synthesis accuracy. (D) Mean CPIC guideline scores by an LLM judge; hatched bars show zero-context (pre-training recall only), solid bars show evidence-provided.

4.3 ANALYSIS AND IMPLICATIONS

4.3.1 SINGLE PAPER QA (MC + STATISTICAL SIGNIFICANCE

For the Model QA evaluation, each model receives a single paper alongside a set of multiple-choice and statistical significance questions. Models perform well on drug and variant identification but substantially worse on phenotype matching and study parameter extraction (Table 3). Phenotypes in this context include both pre-existing patient conditions and medical outcomes, which use highly

specific terminologies that may differ from how the same concepts are described in the paper text. Statistical significance extraction proves similarly challenging, with no model exceeding 58% accuracy on correctly identifying both the p-value and significance determination.

Model	Drug MCQ	Variant MCQ	Phenotype MCQ	Stat. Significance Analysis
Claude Opus 4.6	**93.7%**	**79.7%**	**40.5%**	39.9%
Claude Sonnet 4.5	73.0%	53.0%	30.6%	**58.0%**
Claude Haiku 4.5	80.0%	65.7%	32.5%	40.7%
GPT-4o-mini	86.0%	62.3%	31.0%	49.5%
GPT-5.2	91.0%	75.3%	40.5%	36.0%

Table 3: Single-question multiple-choice accuracy across entity types and study parameter extraction (both p-value and significance identified correctly). Values reported as percentage accuracy. For the statistical significance analysis, results are from a randomly selected 10,000 examples from the total 46,823 available.

Contradiction Rates To test self-consistency, we created paired versions of each MC question in which all answer options are incorrect and a "None of the above" choice is added. A contradiction occurs when the model selects the correct answer on the standard version but fails to select "None of the above" on the paired version, or vice versa. Models contradict themselves on **35.3%** of paired questions on average (Table 4), with the dominant failure mode being the model selecting a plausible-sounding but incorrect answer rather than "None of the above."

Model	Correct → Wrong (NOTA missed)	Wrong → Correct (NOTA spurious)	Total Rate
Claude Haiku 4.5	32.5%	8.9%	41.4%
Claude Sonnet 4.5	41.8%	3.1%	44.9%
GPT-4o-mini	29.4%	4.2%	33.6%
GPT-5.2	21.8%	8.6%	30.4%
Claude Opus 4.6	21.1%	5.2%	26.4%

Table 4: Macro-averaged contradiction rates across Variant, Drug, and Phenotype MCQs. "Correct → Wrong" indicates the model answered correctly on the standard question but failed to select "None of the above" on the paired version. Values are averages of task-level percentages.

4.3.2 Chained Question Analysis

Chain Results. Table 5 reports per-step accuracy for both agents. Although chain-level pass rates are low (13.6% for Claude Code and 22.7% for Codex), the per-step numbers reveal that failure is neither uniform nor random and each step exposes a qualitatively distinct failure mode.

Model	Q1: Predictor	Q2: Significance	Q3: Phenotype	Q4: Evidence
Claude Code	68.2%	54.5%	63.6%	72.7%
Codex	63.6%	59.1%	**81.8%**	68.2%

Table 5: Per-step full-credit accuracy across the four evaluated chain steps

Step 1 (Predictor Inventory). Both models achieve moderate accuracy (Claude 68.2%, Codex 63.6%), with errors split between including spurious distractors and omitting relevant predictors. Models tend to anchor on the most prominently reported allele rather than enumerating the full set (see Appendix B for examples).

Step 2 (Significance Judgement). This is the weakest step (Claude 54.5%, Codex 59.1%) and the most common first-failure point, accounting for 36.4% of initial chain failures. The dominant error mode is false positives—labelling non-significant associations as significant. Notably, the papers on

which both models fail are identical, suggesting systematic ambiguity in statistical reporting rather than model-specific weaknesses.

Step 3 (Phenotype Category). Codex substantially outperforms Claude Code (81.8% vs. 63.6%). Claude's failures are almost exclusively empty-list outputs rather than incorrect categories, indicating an instruction-following failure rather than a conceptual one.

Step 4 (Evidence Selection). Accuracy is comparable (Claude 72.7%, Codex 68.2%), but errors differ qualitatively: Claude selects the correct gene but wrong genotype comparison, while Codex frequently selects predictors from entirely different genes. Both patterns suggest models match on lexical proximity rather than ranking evidence by the specific quantitative comparison requested.

Failure patterns. Per-chain error analysis reveals that failures are driven by step-specific difficulty rather than accumulated error. First, Step 4 (evidence selection) is never the *only* step to fail in any chain—when it fails, at least one earlier step also failed, suggesting that the same paper-level properties that make significance judgement and phenotype classification difficult also make evidence selection difficult. Second, Steps 3 and 4 are largely decoupled: in 4 of 6 chains where Step 3 fails completely, Step 4 still receives full credit, demonstrating that the model can rank evidence by quantitative strength without successfully classifying the clinical phenotype.

4.3.3 Summary Generation Analysis

Agent	Model	Mean Reward
Codex	GPT-5	63%
Claude Code	Opus 4.6	60%

Table 6: Mean reward across agent configurations.

Both models performed similarly in terms of overall benchmark results, however, the way both achieve their scores differs greatly. Claude Code achieves more perfect scores (31% vs. 24%) of trials but fails completely far more often (19% vs. 8%). Codex is more consistent — 68% of tasks earn partial credit compared to 44% for Claude.

Subtask	Failure Rate	Notes
Drug recall	$< 10\%$	Reliable for both agents
Phenotype recall	$\sim$ 40–50%	String-matching mismatches dominate
Paper count	$\sim$ 50–60%	Off-by-1-to-4 errors in both directions

Table 7: Failure breakdown by subtask in the Summary QA agent evaluation.

For the generated summaries, both models perform well in terms of matching the correct drug terminologies however the desired level of granularity of assigned phenotypes along with exact paper matching poses the greatest issue with model performance. Even though we provide the model with a complete list of available phenotype terms to choose from along with direction to be as specific as possible, we find the model to tend to prefer more general terms such as "mental disorders" when the ground truth answer was "major depressive disorder."

4.3.4 CPIC GUIDELINE ANALYSIS

Benchmark	Model	Mean Reward
CPIC Zero-Context	Claude Code (Opus 4.6)	44
	Codex (GPT-5)	33
CPIC Evidence	Claude Code (Opus 4.6)	65
	Codex (GPT-5)	55

Table 8: Mean reward comparison across CPIC benchmarks.

We find that providing source literature consistently improves guideline quality. However, the non-trivial zero-context scores indicate that both models retain some pharmacogenomic knowledge from pre-training, making it difficult to fully isolate evidence-based reasoning from recall.

While the models did not perform poorly, we did find common trends among the failure modes. The deterministic `test_classification` is the single most common source of point loss. In many cases, the agent identifies the correct clinical action but assigns the wrong classification strength.

Agent Output	Expected	Direction	Example
Strong	Moderate	Over-confident	CFTR/ivacaftor: agent says "strongly recommend," gold says "moderate"
Moderate	Strong	Under-confident	CYP2B6/efavirenz: agent hedges, CPIC says strong
Strong	Optional	Over-confident	CYP2C19/voriconazole: strong vs. optional
Optional	No Recommendation	Inventing a recommendation	CYP2C19/clopidogrel (ultrarapid metabolizer): no CPIC recommendation exists
No Recommendation	Strong	Missing knowledge	SLCO1B1/lovastatin: agent unsure, CPIC says strong

Table 9: Common classification strength mismatches in CPIC tasks.

This pattern reflects a fundamental challenge: CPIC classification strength encodes the *quality and quantity of evidence*, not the severity of the clinical effect. Even though this is part of the model's instruction, the model's ability to understand strength of evidence quality remains not in line with the human curation team which may also be a recurring pattern given the model's poor performance on p-value/statistical significance analysis as previously discussed.

When the classification is incorrect but the action is evaluated by an LLM judge, agents often produce clinically reasonable yet non-matching recommendations.

Task	Agent Action	Expected Action	Issue
DPYD/capecitabine (AS 1.0)	Avoid	Reduce dose by 50%	Over-cautious
CYP2B6/efavirenz (rapid)	Use alternative or reduce dose	Standard dosing (600 mg/day)	Over-cautious
CYP2C19/clopidogrel (ultrarapid)	Use standard doses	No recommendation	Provides unsolicited advice

Table 10: Representative action-level discrepancies.

A recurring pattern is **over-caution**: agents default to more conservative recommendations (avoid $>$ dose reduction $>$ standard dose), even when CPIC guidelines specify a less restrictive action.

5 LIMITATIONS AND FUTURE WORK

Domain scope. All instances in VARIANTBENCH are drawn from pharmacogenomics, a field characterised by structured variant–drug–phenotype relationships and well-defined statistical reporting conventions. While this structure enables rigorous evaluation, it remains an open question whether the reasoning failures we observe—particularly error cascades in chained questions and degraded cross-document synthesis—generalise to other scientific domains with less standardised reporting (e.g., ecology, biology, social science). We plan to extend the benchmark to additional biomedical subfields and to release the underlying annotation-to-question generation pipeline so that other curated knowledge bases can be converted into analogous evaluation suites.

End-to-end Chaining Our current evaluation provides gold-standard context from prior steps at each turn, isolating per-step reasoning ability. However, the most practically relevant setting is *end-to-end* chaining, where the model's own outputs feed into subsequent steps. This ablation would directly measure error propagation, self-consistency, and whether models can detect and recover from their own mistakes. The current Harbor framework evaluates tasks independently, making end-to-end chaining non-trivial to implement within its task structure. We are developing a multi-turn task format for Harbor that supports forward-propagating model outputs across chained steps, and plan to release this as a first-class evaluation mode alongside the benchmark.

LLM-as-judge calibration. CPIC guideline recreation is scored by an LLM judge comparing model outputs to ground-truth guidelines along four axes (accuracy, safety, specificity, literature confidence). While the judge is constrained to compare against a provided reference rather than evaluate open-endedly, we have not yet validated its scores against human expert ratings. Establishing inter-rater agreement between LLM and human judges is a priority for the full benchmark release.

Training data construction. The scale of the ClinPGx annotation corpus (over 9,000 articles, $\sim$46,000 annotations) substantially exceeds what is needed for evaluation alone. We plan to partition the corpus into training and held-out evaluation splits, enabling researchers to fine-tune models on pharmacogenomic reasoning tasks and measure whether targeted training reduces the cascading failures and cross-document synthesis gaps that our benchmark exposes.

6 CONCLUSION

We introduced VARIANTBENCH, a large-scale benchmark for evaluating multi-step scientific reasoning grounded in two decades of expert-curated pharmacogenomic annotations from ClinPGx. Spanning 79,504 structured single-paper questions and 394 agentic tasks, VARIANTBENCH organizes five task types into three tiers of increasing complexity: single-paper comprehension, cross-document synthesis, and clinical guideline generation. This structure explicitly tests whether models can maintain logical consistency across dependent reasoning steps rather than merely answer isolated questions correctly.

Our results demonstrate a persistent gap between per-step competence and end-to-end reliability. While models often perform adequately on individual subtasks, requiring consistency across chained reasoning, multi-document synthesis, or guideline-level decisions exposes substantial brittleness. These findings suggest that conventional single-question benchmarks may systematically overestimate model readiness for real-world scientific workflows, where errors propagate and evidence calibration is critical.

We release VARIANTBENCH with deterministic verifiers and reproducible agent infrastructure to enable targeted progress on robust, evidence-grounded, multi-stage scientific reasoning.

REFERENCES

Anthropic. Claude code overview. *Claude Code Documentation*, 2026. URL `https://code.claude.com/docs/en/overview`. Accessed: 2026-02-21.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Jenna Sparks, Jena D Hwang, Varsha Kishore, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S Weld, Doug Downey, Wen-Tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, February 2026.

Kelly E. Caudle, Allan E. Rettie, Michelle Whirl-Carrillo, David M. Smith, Jim E. Mintzer, Michael T. M. Lee, Teri E. Klein, James T. Callaghan, and Clinical Pharmacogenetics Implementation Consortium. Clinical pharmacogenetics implementation consortium (cpic): The driving force for consistent clinical translation of pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 95 (4):422–428, 2014. doi: 10.1038/clpt.2013.244.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Md Faisal Mahbub Chowdhury and Alberto Lavelli. Fbk-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 351–355, 2013.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. *CoRR*, abs/2105.03011, 2021. URL https://arxiv.org/abs/2105.03011.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05 2016. ISSN 1758-0463. doi: 10.1093/database/baw068. URL https://doi.org/10.1093/database/baw068.

Yao Lu, Yue Dong, and Laurent Charlin. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *CoRR*, abs/2010.14235, 2020. URL https://arxiv.org/abs/2010.14235.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge, 2024. URL https://arxiv.org/abs/2409.13740.

M Whirl-Carrillo, E M McDonagh, J M Hebert, L Gong, K Sangkuhl, C F Thorn, R B Altman, and T E Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4):414–417, 2012. doi: https://doi.org/10.1038/clpt.2012.96. URL https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1038/clpt.2012.96.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. URL https://arxiv.org/abs/2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A CHAINED QUESTION EXAMPLE

Below we present a complete chained question instance derived from a single pharmacogenomic paper, along with the model outputs from Claude Code. Each step receives the gold-standard output from the previous step as context (oracle chaining). Correct responses are marked with ✓ and incorrect responses with ×. This instance illustrates a common failure mode: the model correctly identifies the predictors and their significance but fails at phenotype classification, causing the chain to fail despite 3/4 steps correct.

Step 1: Predictor Inventory ✓

INSTRUCTION.MD

```
**Genes studied in this paper (PMID 38707740):** UGT1A1

For each gene listed above, identify the genetic predictors explicitly
evaluated in the paper (e.g., rsID/variant, star allele/haplotype,
diplotype/genotype, or metabolizer/phenotype group).

Below is a list of candidate genetic predictors (rsIDs, star alleles,
haplotypes, etc.). Some of these appear in the paper; others are
distractors.

Options:
  A) rs2228171
  B) UGT1A1*1, UGT1A1*6, UGT1A1*28, UGT1A1*60
  C) UGT1A1*1, UGT1A1*6, UGT1A1*28
  D) UGT1A1*6, UGT1A1*28, UGT1A1*37
```

Gold: C

Model: "The paper (PMID 38707740) studies Japanese patients receiving nal-IRI/FL for metastatic pancreatic cancer. For UGT1A1, it explicitly evaluates: UGT1A1*6* and UGT1A1*28 as the variant alleles associated with reduced glucuronidation and increased toxicity UGT1A1*1 (wild-type) as the reference, with patients categorized as wild-type (–/–), heterozygous (+/–), or homozygous/compound heterozygous (+/+). Options B and D are ruled out because UGT1A1*60 and UGT1A1*37 are not mentioned anywhere in the paper. Option A (rs2228171) is also absent. The answer is C."

Step 2: Significance Judgement ×

INSTRUCTION.MD

```
Predictor comparison pairs from this paper

The following predictor comparison pairs have been identified in this
    paper:
{
  "UGT1A1*1, UGT1A1*6, UGT1A1*28": [
    "*28 + *6 vs *1/*1",
    "*28/*28 + *6/*28 + *6/*6 vs *1/*1"
  ]
}

Question (Step 2 of 4): Significance Judgement

For each predictor comparison pair listed above, determine whether
the authors reported a statistically significant association.

**Options:** `"yes"`, `"no"`, `"not stated"`
```

Gold:

- `*28 + *6 vs *1/*1`: not significant
- `*28/*28 + *6/*28 + *6/*6 vs *1/*1`: not significant

Model Reasoning:

Comparison	Predicted	Model Rationale
*28 + *6 vs *1/*1	not stated	The paper never reports a specific p-value for the heterozygous vs. wild-type pairwise comparison. The Kruskal–Wallis p-values test all three groups together, and regression analyses only model the homozygous genotype as a predictor.
*28/*28 + *6/*28 + *6/*6 vs *1/*1	yes	The paper explicitly reports significant associations for the homozygous/compound-heterozygous group: nadir WBC p=0.033, nadir neutrophil p=0.043 (Kruskal–Wallis), and multiple regression p=0.009 (leukopenia) and p=0.017 (neutropenia).

Step 3: Phenotype Category ✓

PRIOR CONTEXT (FROM STEP 1):

```
{
  "UGT1A1": ["UGT1A1*1, UGT1A1*6, UGT1A1*28"]
}
```

INSTRUCTION:

```
For each predictor identified in Step 1, classify the clinical outcome
that the genetic association targets in this paper.

Options: "efficacy", "toxicity", "dosage", "metabolism/PK", "PD", "other"

Rules:
  efficacy      = treatment response, remission, survival
  toxicity      = adverse drug reactions, side effects
  dosage        = dose requirements, dose adjustments
  metabolism/PK = drug concentration, clearance, AUC, Cmax
  PD            = clinical pharmacodynamic endpoints in patients
  other         = none of the above
```

Gold: `[toxicity]`
Model: `[toxicity]` *Reasoning:* All three outcomes—neutropenia, leukopenia, and diarrhea—are adverse drug reactions associated with UGT1A1 genotype during nal-IRI/FL treatment. No efficacy, dosage, or PK endpoints were primary targets of genetic association.

Step 4: Evidence Selection ✓

PRIOR CONTEXT (FROM STEPS 1–2):

```
Predictor-comparison pairs:
{
  "UGT1A1*1, UGT1A1*6, UGT1A1*28": [
    "*28 + *6 vs *1/*1",
    "*28/*28 + *6/*28 + *6/*6 vs *1/*1"
  ]
}
```

INSTRUCTION:

```
Among the predictor-comparison pairs listed above, which has the
smallest p-value reported for the association between the predictor
and the clinical or pharmacological outcome?

Use only the p-value for the direct predictor-outcome association
test. If multiple p-values are reported for the same pair, use the
smallest among them. If multiple pairs are tied, list all.
```

Gold: `UGT1A1*1, UGT1A1*6, UGT1A1*28 (*28/*28 + *6/*28 + *6/*6 vs *1/*1)`
Model: `UGT1A1*1, UGT1A1*6, UGT1A1*28 (*28/*28 + *6/*28 + *6/*6 vs`

`*1/*1`) *Reasoning:* This comparison has the smallest reported p-value (p=0.009) from the multiple regression analysis associating the homozygous/compound heterozygous genotype with decreased WBC nadir count (Table III). This is smaller than the other reported p-values for this predictor (p=0.017, 0.033, 0.043), and no p-values are reported for the heterozygous vs. wild-type comparison as a standalone analysis.

B Chain Error Examples

We catalogue representative failure modes for each chained question step.

Step 1. On chain 001743, both models returned only `[*B]` for a UGT1A1 question whose gold answer was `[*A, *B, *F]`, illustrating the anchoring-on-prominent-allele failure mode.

Step 2. The three papers on which both models fail are chains 000230, 001317, and 001773. Chain 001317 exhibits simultaneous over- and under-calling of significance within the same paper: one predictor is incorrectly labelled not significant while another is over-stated as significant.

Step 3. All five of Claude Code's zero-reward Step 3 instances produce empty-list outputs (`[]`) across different phenotype classes (`toxicity`, `efficacy`), suggesting a systematic serialisation issue with set-valued outputs.

Step 4. Claude Code errors are within-gene (e.g., selecting rs9923231 `CT+TT vs CC` when the target is rs1045642 `AA vs GG`). Codex errors are cross-gene: on chains 001733 and 001773, it returns rsIDs from entirely different pathways.

C Multiple-Choice QA — Drug

Below we provide representative examples of the Drug multiple-choice task, including both standard questions and "none of the above" variants.

C.1 Standard Example

```
Genotype TT is associated with decreased response to _____ in people
    with
Diabetes Mellitus, Type 2.
```

Option	Drug
a	saxagliptin
b	linagliptin
c	vildagliptin
d	**sitagliptin**

Ground truth: `d`

C.2 None-of-the-Above Example

```
Genotype TT is associated with decreased response to _____ in people
    with
Diabetes Mellitus, Type 2.
```

Option	Drug
a	vildagliptin
b	linagliptin
c	**None of the options**
d	saxagliptin

Ground truth: `c`

D MULTIPLE-CHOICE QA — VARIANT

D.1 STANDARD EXAMPLE

```
_____ is associated with decreased response to sitagliptin in people
    with
Diabetes Mellitus, Type 2.
```

Option	Variant
a	rs2290910
b	rs7754840
c	rs2285676
d	**rs2909451**

Ground truth: `d`

D.2 NONE-OF-THE-ABOVE EXAMPLE

```
_____ is associated with decreased response to sitagliptin in people
    with
Diabetes Mellitus, Type 2.
```

Option	Variant
a	rs2285676
b	rs7754840
c	**None of the options**
d	rs2290910

Ground truth: `c`

E MULTIPLE-CHOICE QA — PHENOTYPE

E.1 EXAMPLE

```
Genotypes AA + AT are associated with decreased risk of _____ due to
    nicotine
as compared to genotype TT.
```

Option	Phenotype
a	Disease: Hypotensive disorder
b	Disease: Hematologic Disorder
c	**Other: Tobacco Use Disorder**
d	Disease: Bipolar Disorder

Ground truth: `c`

F STUDY PARAMETER EXTRACTION

Below we provide representative examples for extracting a p-value for a specific variant–drug association and determining statistical significance. We include counterfactual questions with a modified variant or drug to test grounding.

F.1 Correct Variant–Drug Pair

```
For the association between rs2909451 (gene DPP4) and sitagliptin
    described in
PMC11730665, extract the p-value and determine statistical significance.
```

Ground truth: p-value < `0.001`, significance = `yes`

F.2 Counterfactual (Modified Variant)

```
For the association between rs3765467 (gene DPP4) and sitagliptin
    described in
PMC11730665, extract the p-value and determine statistical significance.
```

Ground truth: p-value = `not found`, significance = `not found`

F.3 Counterfactual (Modified Drug)

```
For the association between rs2909451 (gene DPP4) and azithromycin
    described in
PMC11730665, extract the p-value and determine statistical significance.
```

Ground truth: p-value = `not found`, significance = `not found`

F.4 Example Model Response (GPT-5)

```
{
  "question_id": 1,
  "pmcid": "PMC11730665",
  "question_type": "correct",
  "response_p_value": "<.001",
  "response_significance": "yes"
}
```

G Variant Extraction

Below we provide a representative example for extracting all pharmacogenomic variants discussed in a full-text research article.

G.1 Example Task

```
From the article "Comparative efficacy and safety of sitagliptin or
    gliclazide
combined with metformin in treatment-naive patients with type 2 diabetes"
(PMC11730665), extract all pharmacogenomic variants.
```

Ground truth variants: `rs1799853, rs4664443, rs7754840, rs3765467, rs2285676, rs2909451, rs6923761, rs163184`

Example Model Response (GPT-5)

```
{
  "pmcid": "PMC11730665",
  "response": [
    "rs2909451", "rs4664443", "rs3765467", "rs2285676",
    "rs6923761", "rs163184", "rs7754840", "rs756992",
    "rs1799853", "rs1057910", "rs57803087", "rs4244285",
    "rs4986893", "CYP2C9*2", "CYP2C9*3", "CYP2C19*2", "CYP2C19*3"
  ]
}
```

Note: The model extracted all ground-truth variants but also included additional variants, illustrating the precision challenge in variant extraction.

H CPIC EVIDENCE (AGENT)

An agentic task where the model is given access to a set of research papers and must generate a clinical recommendation following CPIC guideline format. The output is evaluated deterministically (classification match) and via an LLM judge (four dimensions scored on a 1–5 scale, each requiring ≥ 4 to pass).

H.1 EXAMPLE TASK

Drug: desflurane
Gene(s): CACNA1S | RYR1
Patient genotype: RYR1 Malignant Hyperthermia Susceptibility; CACNA1S Uncertain Susceptibility

H.2 EXPECTED OUTPUT

```
{
  "recommendation": "Halogenated volatile anesthetics ... should not be
      used ...",
  "classification": "Strong",
  "implication": "RYR1: ... increased risk ...; CACNA1S: ... uncertain
      susceptibility ..."
}
```

H.3 EVALUATION DIMENSIONS

Dimension	Description
Action Correctness	Does the agent recommend the same clinical action as CPIC?
Recommendation Completeness	Does it capture clinically significant details?
Implication Accuracy	Does it correctly describe the PGx phenotype?
Safety	Is the recommendation safe for the patient?

I CPIC ZERO-CONTEXT (AGENT)

Similar to the CPIC Evidence task, but no papers are provided. The model must produce a CPIC-style recommendation purely from parametric knowledge, testing guideline recall.

I.1 EXAMPLE TASK

Drug: clopidogrel
Gene: CYP2C19
Patient genotype: CYP2C19 Ultrarapid Metabolizer

I.2 EXPECTED OUTPUT

```
{
  "recommendation": "No recommendation",
  "classification": "No Recommendation",
  "implication": "CYP2C19: Increased clopidogrel active metabolite
      formation; ..."
}
```

J PGx Drug QA (Agent)

An agentic task where the model reads papers and answers a set of 10 multiple-choice questions about drug–variant associations.

J.1 Expected Output Format

```
{
  "1": "c",
  "2": "d",
  "3": "b",
  "4": "b",
  "5": "b",
  "6": "a",
  "7": "c",
  "8": "c",
  "9": "c",
  "10": "b"
}
```

K Summary QA (Agent)

An agentic task where the model is given a collection of research papers (with distractors) and must identify all drugs and phenotypes associated with a specific genetic variant, as well as count the number of relevant papers.

K.1 Expected Output Format

```
{
  "drugs": ["clopidogrel", "voriconazole"],
  "phenotypes": ["efficacy", "metabolism"],
  "relevant_paper_count": 5
}
```