VariantBench: A Framework for Evaluating LLMs on Justifications for Genetic Variant Interpretation

Humair Basharat, Simon Plotkin, Michael Pink, Isabella Alfaro
Charlotte Le*, Kevin Zhu†
Algoverse AI Research
humairbasharat@gmail.com, simon.m.plotkin@vanderbilt.edu,
isabella.alfaro77@qmail.cuny.edu, kevin@algoverse.us

Abstract

Accurate classification in high-stakes domains requires not only correct predictions but transparent, traceable reasoning. We instantiate this need in clinical genomics and present VariantBench, a reproducible benchmark and scoring harness that evaluates both the final American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) labels and criterion-level reasoning fidelity for missense single-nucleotide variants (SNVs). Each case pairs a variant with deterministic, machine-readable evidence aligned to five commonly used criteria (PM2, PP3, PS1, BS1, BA1), enabling consistent evaluation of large language models (LLMs). Unlike prior work that reports only final labels, our framework scores the correctness and faithfulness of per-criterion justifications against numeric evidence. On a balanced 100-variant freeze, Gemini 2.5 Flash and GPT-4o outperform Claude 3 Opus on label accuracy and criterion detection, and both improve materially when the decisive PS1 cue is provided explicitly. Error analyses show models master population-frequency cues yet underuse high-impact rules unless evidence is unambiguous. VariantBench delivers a substrate to track such improvements and compare prompting, calibration, and aggregation strategies in genomics and other rule-governed, safety-critical settings.

1 Introduction

Accurate classification in high-stakes domains requires not only correct predictions but also transparent, traceable reasoning. Errors in fields such as healthcare and finance can lead to serious consequences, from patient harm to erosion of public trust. In the study of clinical genomics, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) create guidelines that require experts to review structured evidence and determine the pathogenicity of missense single-nucleotide variants (SNVs), criterion by criterion (Richards et al., 2015). Here, a missense SNV is a one-base substitution that changes a codon, which then replaces one amino acid in the encoded protein (Cheng et al., 2023). According to the ACMG guidelines, there are two types of criteria: those used to classify pathogenic or likely pathogenic variants, and those used to classify benign or likely benign variants (Richards et al., 2015). The five commonly used criteria we address are pathogenic, weighted as strong (PS1), moderate (PM2), supporting (PP3), and benign, weighted as stand-alone (BA1) or strong (BS1). While LLMs have shown they can predict the final pathogenicity label, they rarely provide traceable, criterion-level reasoning.

Recent work highlights both progress and limitations. Proteome-wide pathogenicity resources such as AlphaMissense provide valuable priors but do not map outputs to ACMG criteria (Cheng et al., 2023). LLM benchmarks targeting variant interpretation have emphasized final labels without assessing reasoning quality (e.g., Li et al., 2024). AutoPM3 explored LLM evaluation for the PM3 segregation rule, but focuses on a single criterion (Li et al. 2025). Beyond genomics, few benchmarks in high-stakes domains combine expert-labeled criteria, curated machine-readable evidence, and reproducible scoring frameworks.

We introduce VariantBench, a benchmark and evaluation harness designed to measure both decision accuracy and criterion-level reasoning fidelity. While our testbed focuses on genomic variant interpretation, the framework applies to any domain where decisions must be justified against structured, expert-defined rules. Each case pairs a randomly sampled missense variant from the Genome Ag-

*Equal contribution
†Corresponding author

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 314–321

gregation Database (gnomAD; via dbNSFP 5.2a, GRCh38) with automatically derived evidence aligned to five commonly used ACMG/AMP criteria (PM2, PP3, PS1, BS1, BA1). (Liu et al., 2020). At a high level, PM2 captures rarity or absence from population databases, PP3 supporting evidence from deleterious in-silico predictions, PS1 strong evidence when the amino-acid change matches a known pathogenic variant, and BS1/BA1 benign evidence when population allele frequencies are higher than expected for a rare monogenic disorder (with BA1 functioning as a standalone benign rule). We require models to output both a classification decision and a structured criterion-level justification in JSON format containing a 5-tier classification label (Pathogenic, Likely Pathogenic, VUS, Likely Benign, Benign), boolean flags for PM2/PP3/PS1/BS1/BA1, and a brief rationale. We evaluate LLMs against deterministic rule-based ground truth, using exact-match accuracy, micro and macro F_1 for criterion detection, and a faithfulness metric verifying correct evidence citation, across two settings: Track A, where no PS1 evidence is provided to test knowledge-only behavior, and Track B, where a PS1 yes/no hint is provided to test rule-application consistency. Baselines include heuristic, logistic, and ablated LLM variants. The source code is available at https://github.com/VariantBench. Results show that VariantBench not only diagnoses where and why reasoning fails in genomic medicine, but also offers a reproducible framework adaptable to other high-stakes, rule-governed decision-making tasks.

In this work, we introduce the following contributions:

- A replicable benchmark and scoring harness for ACMG/AMP-aligned reasoning over missense SNVs.
- A measurement substrate for tracking improvements and comparing prompting, calibration, and aggregation strategies.
- Comparative analyses of models across two tracks that support structured prompting and explicit evidence supplementation.

2 Methodology

We designed VariantBench to evaluate whether LLMs can reproduce ACMG/AMP reasoning when given the same structured, numeric evidence used by clinical curators. Rather than retrieving textual snippets, which proved too sparse and unreliable, we adopted a deterministic evidence generation pipeline that programmatically derives the inputs for five ACMG criteria (PM2, PP3, PS1, BS1, and BA1), directly from curated databases and fixed thresholds.

2.1 Variant Sampling and Filtering

We drew candidate variants from dbNSFP 5.2a (GRCh38) as a proxy for gnomAD coverage, querying single-nucleotide substitutions with one-base REF/ALT and available gnomAD allele frequency (AF) values. Each variant includes a reference (REF) and alternate (ALT) allele, denoting the original and substituted nucleotides at a specific genomic position, respectively. We specifically chose dbNSFP 5.2a over earlier versions due to its comprehensive integration of gnomAD v3.1.2 data, which includes 75,000 genomes and provides more robust population frequency estimates across diverse ancestries. We then enforced a strict missense filter at the HGVS protein level using a regex form (e.g., p.Gly137Arg), excluding stopgain, frameshift, indel, and splice annotations. The regex pattern specifically matches p[A-Z][a-z]2+[A-Z][a-z]2 to ensure consistent HGVS formatting and prevent edge cases like synonymous variants (p.=) or complex multi-amino acid changes from entering the dataset. HGVS formatting provides a standardized way to describe sequence changes at the DNA, RNA, or protein level, ensuring that genetic variants are reported unambiguously across databases and studies. We manually logged and excluded any entries that failed this pattern to prevent parser drift. As a result, we produced a broad, gene-agnostic pool spanning a wide AF range and diverse in-silico scores.

2.2 Deterministic Evidence Computation

For each variant, we compute five rule flags with fixed logic implemented in `helpers.py`:

PM2 (Moderate evidence of pathogenicity): `True` if $\mathrm{AF}_{\mathrm{popmax}} < 10^{-4}$ *or* $\mathrm{AF}_{\mathrm{popmax}}$ is missing, modeling *absent/ultra-rare*. We treat missing AF as satisfying PM2 following ACMG guidelines that consider absence from population databases as supporting evidence (Richards et al., 2015) though we flag these cases separately for sensitivity analysis.

BS1 (Strong evidence of benignity): `True` if

$10^{-4} \leq \mathrm{AF}_{\mathrm{popmax}} < 0.05$. This threshold aligns with the 2015 ACMG guidelines' definition of "greater than expected for disorder" while avoiding overlap with the BA1 threshold.

BA1 (Stand-alone evidence of benignity): `True` if $\mathrm{AF}_{\mathrm{popmax}} \geq 0.05$. This 5% threshold represents the standard ACMG cutoff for "too common to cause disease" and automatically results in a Benign classification regardless of other evidence.

PP3 (Supporting evidence of pathogenicity): PP3 is triggered by concordant in silico evidence that a missense substitution is likely to be functionally damaging. We set PP3 to `True` if at least 3 of 7 in silico tools predict the variant to be *damaging/deleterious* (SIFT, PolyPhen2_HDIV, MutationTaster, MutationAssessor, PROVEAN, MetaSVM, MetaLR) *or* if REVEL > 0.5. Missing values do not contribute to the count. The REVEL override (REVEL > 0.5) follows recent ACMG/AMP recommendations that recognize REVEL as a higher-performing ensemble meta-predictor for missense variants. Individual tools are mapped to binary calls using canonical thresholds: SIFT < 0.05, PolyPhen2_HDIV > 0.909, MutationTaster $\in \{\mathtt{D}, \mathtt{A}\}$, MutationAssessor > 1.9, PROVEAN < -2.5, MetaSVM > 0, and MetaLR > 0.5.

PS1 (Strong evidence of pathogenicity): `True` if any canonical protein change from VEP/snpEff (annotation tools that predict how genetic variants affect genes and proteins, such as whether a change results in a missense or stop-gain mutation) exactly matches an amino-acid change in ClinVar, a publicly accessible database maintained by the U.S. National Center for Biotechnology Information (NCBI) that archives and aggregates the clinical significance of human genetic variants, and that is annotated "Pathogenic", "Likely_pathogenic", or "Pathogenic/Likely_pathogenic". We map three-letter amino acid codes to one-letter codes and keep only missense (no stop-gains/frameshifts). Our PS1 lookup table is built from ClinVar's March 2025 release, filtering for variants with $\geq$ 2-star review status to ensure clinical validity. We normalize protein changes by stripping transcript identifiers and resolving alternative amino acid nomenclature (e.g., selenocysteine) to prevent false negatives.

2.3 Gold Benchmark Freeze

From a large random sample of 100 variants meeting our filtering criteria, we produced a label per variant with a deterministic `combine()` function. The `combine()` function implements standard ACMG combining rules (Richards et al., 2015).

- BA1 alone $\rightarrow$ Benign
- BS1 without contradicting evidence $\rightarrow$ Likely Benign
- PM2 + PP3 + PS1 $\rightarrow$ Likely Pathogenic
- Strong pathogenic evidence without benign evidence $\rightarrow$ Pathogenic
- Conflicting or insufficient evidence $\rightarrow$ Variant of Uncertain Significance (VUS).

We then froze a 100-example benchmark by stratified sampling 20 variants per label (Pathogenic, Likely Pathogenic, VUS, Likely Benign, Benign), yielding balanced coverage across tiers. This balanced design prevents models from exploiting class imbalance and ensures equal weighting of performance across all clinical decision points. In clinical genetics, these five categories support differential actions: Pathogenic and Likely Pathogenic variants can prompt surveillance, cascade testing of relatives, or changes in treatment, whereas Likely Benign and Benign variants are generally not used to alter care. Variants of Uncertain Significance (VUS) are typically non-actionable but can generate follow-up work and patient anxiety. Benchmarks that probe how LLMs reason about these labels therefore speak directly to the safety and audit ability of AI-assisted genomic interpretation, even when used in a research-only context. Although 100 variants is modest by modern benchmarking standards, this size is sufficient to distinguish the models we study and to support detailed error analysis. At temperature 0, headline accuracies in Figure 1 range from ≈ 0.21 (Claude) to ≈ 0.47–0.52 (Gemini), and Matthews correlation coefficients (MCC) from ≈ 0.02 to ≈ 0.40–0.42. Under a simple binomial approximation with $n = 100$, the standard error of an accuracy estimate is at most

$$\sqrt{p(1-p)/n} \leq \sqrt{0.25/100} \approx 0.05,$$

yielding 95% confidence intervals of roughly ± 0.10. The observed accuracy gaps of $\approx$

0.15–0.30 and MCC gaps of $\approx$ 0.18–0.40 between Gemini, GPT–4o, and Claude therefore exceed this sampling margin, indicating that VariantBench–100 is large enough to meaningfully separate model behaviours, even though it is not sufficient for precise estimates of clinical-grade performance. We then saved two files under `results/FrozenBenchmark/`: the full gold table (`variantbench_100_gold.csv`, including flags and label) and the public input table (`variantbench_100_inputs.csv`) that hides gold flags but retains fields needed to build prompts. Both files include cryptographic checksums (`SHA-256`) to ensure reproducibility and detect any data corruption.

2.4 Prompt Construction

We developed two evaluation tracks to isolate the contribution of external knowledge versus structured reasoning:

2.4.1 Track A (No PS1 cue)

The model receives HGVS, $\mathrm{AF}_{\mathrm{popmax}}$, and a compact in-silico summary (CADD, SIFT, PolyPhen2_HDIV, MetaLR, FATHMM-XF, AlphaMissense when present). In-silico scores are presented as raw values rather than pre-interpreted categories to test whether models can apply an appropriate threshold. The prompt explicitly instructs the model to evaluate only PM2, PP3, PS1, BS1, and BA1, and to return a single JSON object with lowercase booleans and a one-line rationale. The JSON schema is strictly enforced:

```
{
  "pm2": true/false,
  "pp3": true/false,
  "ps1": true/false,
  "bs1": true/false,
  "ba1": true/false,
 "label": "Pathogenic"|"Likely_pathogenic"
  |"VUS"|"Likely_benign"|"Benign",
  "rationales": { ... }
}
```

No PS1 evidence is provided; the model must rely on its pretrained knowledge to decide PS1.

2.4.2 Track B (PS1 evidence provided)

Similar to Track A, but we add a single line `PS1 evidence (ClinVar {clinvar_release}): {ps1_yes_no} # "yes" or "no"`, where yes/no is computed deterministically by our PS1 helper. This ablation test evaluates whether models can integrate provided evidence or rely on potentially outdated training data. The ClinVar release date is explicitly stated to signal data currency. The prompt fixes PS1 semantics ("set PS1=true iff the evidence line is 'yes' "). This track isolates whether the model applies PS1 correctly when the evidence is explicit.

We write prompts per track to `results/prompts/`, and one JSONL with a variant ID per variant and a human-readable preview. We then fed the prompts to the following models in zero-shot: GPT-4o, Claude 3 Opus, and Gemini 2.5 Flash.

Additional prompt engineering considerations:

- We prepend a brief ACMG primer (50 words) explaining that variants should be classified based on population frequency and computational predictions, without defining specific thresholds, to activate relevant knowledge without biasing toward particular cutoffs.
- All numeric values are formatted consistently (scientific notation for AF, two decimal places for scores) to prevent parsing ambiguities.
- We include a "`chain-of-thought`" instruction asking models to "briefly explain your reasoning before providing the JSON" to improve accuracy through intermediate reasoning steps.
- Temperature is set to 0 for all primary experiments to ensure deterministic outputs, with a `temperature=0.7` ablation to assess robustness.

Quality control measures:

- Each prompt–response pair is validated for JSON parseability before scoring.
- We implement retry logic (maximum three attempts) for API failures or malformed outputs.
- All model outputs are archived with timestamps and model version identifiers for reproducibility.
- We conduct spot checks on 10% of responses to verify that rationales reference the correct evidence types (e.g., PM2 rationales mention allele frequency).

3 Results and Discussion

Figure 1 compares Gemini, GPT-4o, and Claude across five headline metrics at temperature 0. Gemini emerged as the strongest model for final ACMG label prediction, reaching ~ 0.50–0.52 accuracy and ~ 0.40–0.42 MCC. Roughly a 40% improvement over GPT-4o and more than double Claude, whose MCC hovered near zero. This indicates that Gemini not only classifies more variants correctly but also achieves a better balance across true/false positives and negatives.

At the criterion level, micro-F1 scores were uniformly higher than overall accuracy, showing that all models were more consistent in detecting individual ACMG rules than in combining them into final labels. Gemini and GPT-4o achieved strong micro-F1 (0.78–0.88), while Claude lagged at ~ 0.65. Macro-F1 further highlighted model differences: Gemini remained stable across tracks ($\sim$ 0.61–0.78), GPT-4o improved substantially once PS1 evidence was supplied ($0.41 \rightarrow 0.61$), and Claude plateaued, suggesting limited adaptability.

Faithfulness exposed the sharpest divide. Gemini and GPT-4o exceeded 95%, meaning their explanations consistently cited the numeric cues aligned with invoked criteria. Claude, by contrast, plateaued at $\sim 42\%$, reflecting a tendency to provide generic or hallucinated rationales rather than evidence-grounded reasoning. This gap underscores that even when Claude flagged the criteria correctly, it often failed to justify them in a clinically auditable way.

As illustrated in Figure 2, population frequency rules are handled well by Gemini and GPT-4o and less reliably by Claude. For PM2, Gemini and GPT-4o are stable around 0.92–0.93 F1 in both tracks, whereas Claude trails at ~ 0.77. For PP3, GPT-4o leads (0.93–0.95) over Gemini (0.87–0.89), with Claude at ~ 0.56. Decisive rules reveal the most apparent separation. Without PS1 evidence (Track A), all models are ~ 0 on PS1; with a single explicit PS1 cue (Track B), Gemini and GPT-4o jump to ≈ 1.00 while Claude remains low (~ 0.08). BA1 is near-ceiling for Gemini and GPT-4o (0.97–0.98) but negligible for Claude (~ 0.02). BS1 remains challenging across models. Gemini and GPT-4o reach only 0.28–0.31, and Claude is ~ 0.02. This reflects the rule's narrow frequency threshold and the scarcity of BS1-positive examples. Overall, Gemini and GPT-4o reliably apply frequency evidence and, when provided explicit cues, execute decisive ACMG rules. Claude's competence appears confined mainly to simpler, frequency-based criteria.

3.1 Confusion Matrix Analysis

Overview: Across models, most mistakes collapse to VUS when evidence is incomplete or conflicting. Providing an explicit PS1 cue (Track B) reduces this collapse for GPT-4o and Gemini but not for Claude.

GPT-4o: Figure 3 shows GPT-4o is accurate on Benign and VUS (≈80–90% correct across tracks). On Track B, the model undercalls pathogenicity: ≈80% of true Pathogenic shift to Likely Pathogenic, and ≈72.5% of true Likely Pathogenic shift to VUS. This mirrors its per-flag pattern (strong PM2/PP3, weaker PS1/BS1), yielding conservative decisions when high-impact evidence is absent or ambiguous.

Gemini: Additionally, figure 3 shows Gemini is very strong on Benign and VUS (≥95% correct across tracks). With the PS1 cue (Track B), Gemini recovers more Pathogenic cases (≈40% accurate, roughly 2× GPT-4o). Its weakness is the intermediate tiers: Likely Pathogenic accuracy ≈25%, and Likely Benign ≈12.5% (vs. GPT-4o ≈60% for LB), reflecting difficulty with mid-frequency benign signals (BS1) relative to GPT-4o.

Claude: Marked VUS bias across tracks. In Track B, ≈70% of Likely Benign, ≈95% of Likely Pathogenic, and ≈87.5% of Pathogenic are predicted as VUS, explaining low label accuracy and MCC despite mid-range flag F1. This indicates limited integration of high-impact rules and weak use of explicit PS1 cues.

Effect of Temperature: Figure 4 illustrates aggregate temperature sweeps.

- **Accuracy & MCC:** Gemini benefits most from higher temperature in both tracks (accuracy $+ \sim 0.06$ in Track A, $+ \sim 0.13$ in Track B; MCC $+ \sim 0.06$ and $+ \sim 0.16$). GPT-4o is relatively temperature-stable. Claude changes little.
- **Macro-F1:** In Track A, GPT-4o and Gemini see slight increases up to $\tau = 0.3$ (Gemini: $0.605 \rightarrow 0.625$, GPT-4o peaks near $\tau = 0.3$).

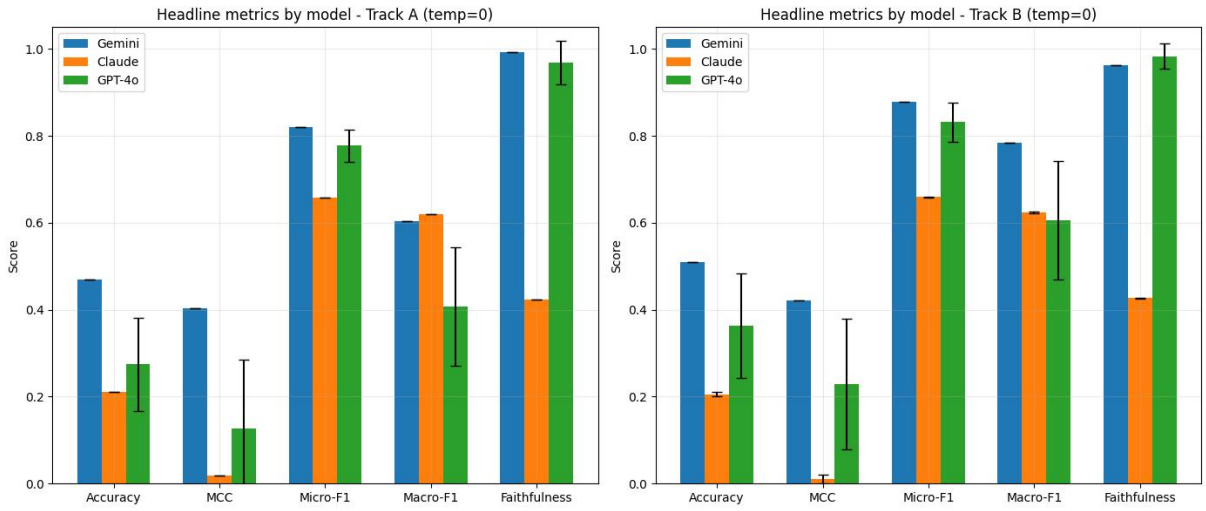

Figure 1: Headline metrics by model on Track A (left) and Track B (right) at temperature 0. Bars show mean scores and error bars denote variability across runs.

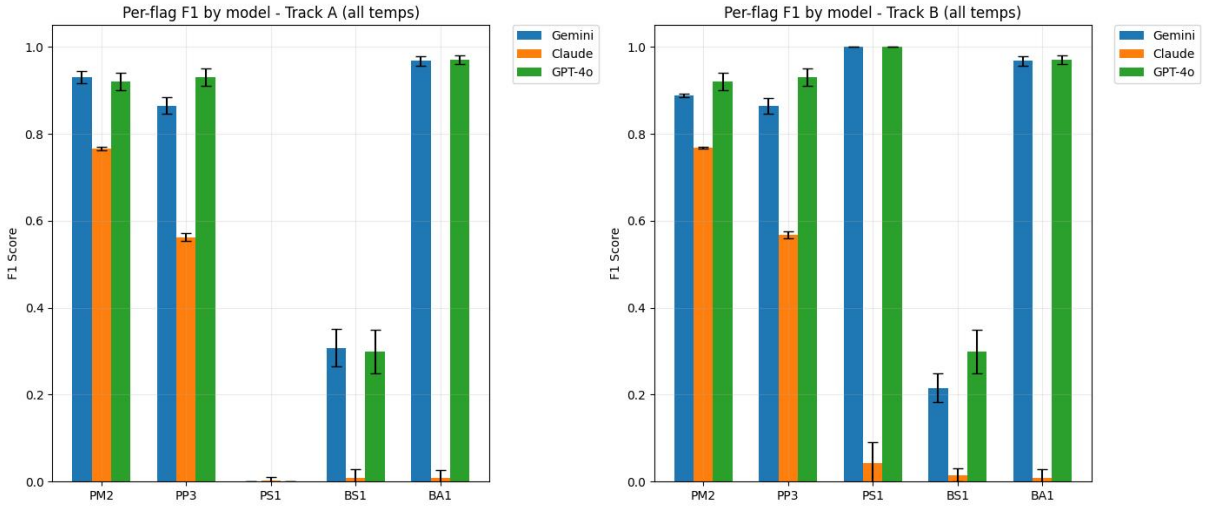

Figure 2: Per-criterion performance by model on Track A (left) and Track B (right) at temperature 0. Bars show mean scores, and error bars denote variability across runs.

OpenAI's macro-F1 in Track B is already high ($\sim$ 0.84–0.85) and flat.

- **Interpretation:** Mild stochasticity helps Gemini explore alternatives that improve final labels without hurting criterion detection. GPT-4o is already near its optimum at low temperature.

4 Conclusion

We introduced *VariantBench*, a reproducible benchmark and scoring harness for ACMG/AMP-aligned reasoning over missense SNVs. In contrast to prior work that scores only the final label, VariantBench evaluates criterion-level correctness (PM2/PP3/PS1/BS1/BA1) and faithfulness to numeric cues using a deterministic pipeline derived from public databases. On VariantBench-100, Gemini 2.5 Flash and GPT-4o outperform Claude on both final labels and rule detection. Across models, population-frequency evidence (PM2/PP3) is learned reliably, while high-impact rules (PS1/BA1/BS1) are brittle unless the signal is made explicit in the prompt. These findings suggest that structured prompting + explicit evidence injection can convert pretrained knowledge into auditable, rule-consistent reasoning, and that VariantBench provides the measurement substrate for tracking such gains and comparing prompting, calibration, and aggregation strategies.

Limitations:

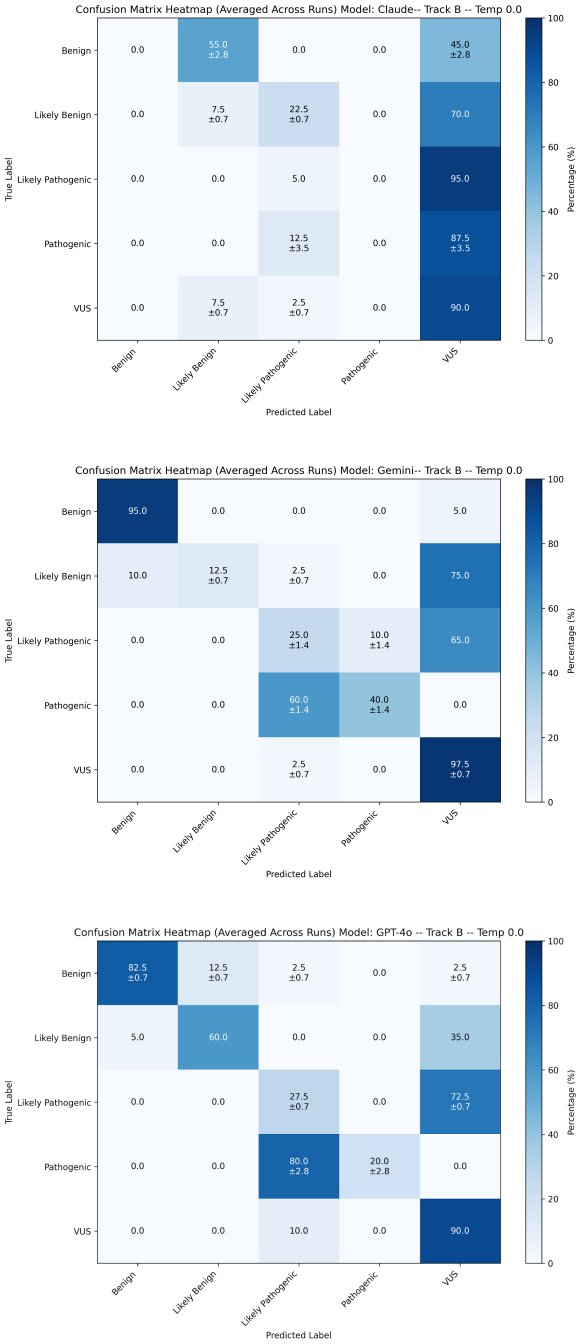

Figure 3: Confusion matrices by model on Track B (temperature 0). Top: Claude. Middle: Gemini. Bottom: GPT-4o. Percentages are averaged across runs.

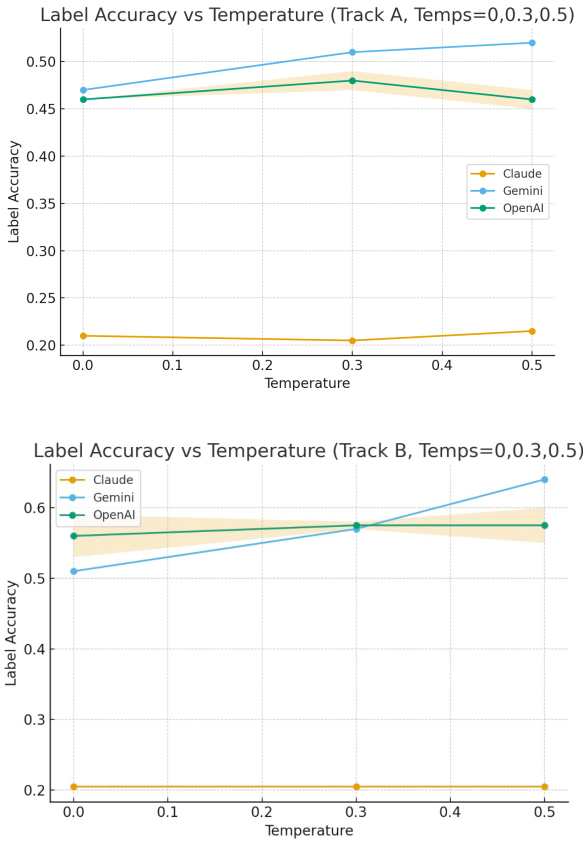

Figure 4: Effect of temperature on label accuracy across models. Top: Track A shows modest accuracy gains for Gemini and GPT-4o up to $\tau = 0.3$. Bottom: Track B highlights Gemini's stronger improvement at higher temperatures. Claude remains flat in both tracks. Error bands show run variability.

- **Rule scope.** VariantBench-100 evaluates reasoning over only five ACMG/AMP criteria (PM2, PP3, PS1, BS1, BA1). Full clinical curation uses additional rules and more complex combinations, so our results should be interpreted as evidence about relative model behaviors under a constrained subset, not as comprehensive estimates of real-world diagnostic performance.

- **Dataset size and balance.** VariantBench-100 is small and label-balanced by design (20 variants per tier) to enable clear comparisons and exhaustive error analysis. This controlled setting prevents exploitation of class imbalance but does not reflect the skewed distributions and edge cases encountered in practice.

- **Faithfulness metric.** Our "cue-citation" score is a surface-level proxy: it checks whether rationales explicitly mention the numeric evidence that should support each criterion. This can undercount valid paraphrases that omit explicit values and overcount boilerplate text that repeats numbers without truly using them in the decision. We therefore view cue-citation as a conservative, first order approximation to reasoning faithfulness.

- **Prompt/decoding sensitivity.** All results are conditional on a particular prompt family, JSON schema, and a single snapshot of three

closed-weight models. Different prompts, decoding parameters, or model versions may change the absolute scores and some qualitative patterns. VariantBench is best viewed as a reusable harness for comparing models and prompting strategies, rather than as a fixed leaderboard.

- **Not a clinical device.** Outputs are non-diagnostic and intended solely for benchmarking research.

Future work will extend to full ACMG/AMP coverage, scale data with stratified sampling, replace string matching with structured evidence auditing (e.g., numeric attribution and counterfactuals), and assess uncertainty calibration.

References

Jun Cheng and 1 others. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, 381(6664):eadg7492.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT Press.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.

Nilah M. Ioannidis, Joseph H. Rothstein, Vikas Pejaver, and 1 others. 2016. Revel: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, 99(4):877–885.

Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, and 1 others. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443.

Melissa J. Landrum and 1 others. 2025. Clinvar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Research*, 53(D1):D1313–D1323.

Shumin Li, Yiding Wang, Chi-Man Liu, Yuanhua Huang, Tak-Wah Lam, and Ruibang Luo. 2025. Autopm3: enhancing variant interpretation via llm-driven pm3 evidence extraction from scientific literature. *Bioinformatics*, 41(7):btaf382.

X. Li, Y. Wang, and 1 others. 2024. Clinvarbert: Benchmarking large language models on clinvar variant classification. *Bioinformatics*. Preprint/early access; update when published.

Xiaoming Liu. 2025. dbnsfp project website and v5.x release notes. `https://www.dbnsfp.org/`. Accessed Sep 15, 2025; see blog "Highlights in dbNSFP v5.1a" for 5.x changes.

Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong, and Yicheng Tu. 2020. dbnsfp v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. *Genome Medicine*, 12(1):103.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424.