VarLitBench and VarLitAgent for Benchmarking and Automating LLM-Assisted Functional Evidence Curation in Genomic Variant Interpretation

Anonymous Authors[1]

Abstract

Linking genomic variants to functional evidence in the literature is a central but labor-intensive step in clinical variant interpretation. We introduce *VarLitBench*, a ClinGen-anchored benchmark for evaluating large language models (LLMs) on variant-specific functional-evidence curation, and *VarLitAgent*, an end-to-end pipeline for human-in-the-loop evidence retrieval, extraction, and reporting. VarLitBench evaluates two tasks. In abstract screening, the model determines whether a paper is likely to report a functional experiment that directly tests one or more genetic variants. In full-paper extraction, the model aligns the target variant to mentions in the paper, extracts experimental readouts, classifies evidence direction, and generates a concise evidence summary. We evaluated `gpt-4o-mini`, `o4-mini`, `claude-haiku-4-5`, and `claude-sonnet-4-5`. All models achieved high recall for abstract screening (0.873–0.904), with `claude-sonnet-4-5` obtaining the best overall F1 score of 0.792. For full-paper PS3 versus BS3 evidence-direction classification, `o4-mini` achieved the highest F1 score, 0.979. We also compared model-generated summaries with expert-written ClinGen curator rationales using an LLM-as-judge protocol. Claude models obtained the highest mean correspondence scores. Evidence strength assignment (e.g., distinctions between pathogenic strong and pathogenic moderate), remained challenging across models. VarLitAgent builds on these findings by taking a genomic variant as input, expanding its identifiers, retrieving candidate literature, screening abstracts, obtaining full texts or PDFs when available, and performing multimodal evidence extraction. The system supports a direct mode for efficient processing and an agentic mode for deeper parsing of figures and tables. Together, VarLitBench and VarLitAgent provide a practical foundation for auditable LLM assistance in functional-evidence curation.

1. Introduction

Clinical genomic variant interpretation requires heterogeneous evidence to be integrated within a standardized pathogenicity framework. The ACMG/AMP guidelines define criteria that incorporate population frequency, segregation, de novo occurrence, computational prediction, and functional studies (Richards et al., 2015). Functional evidence is especially valuable because it can provide mechanistic support linking a molecular perturbation to a disease-relevant effect. Within this framework, PS3 denotes well-established functional studies that support a damaging effect, whereas BS3 denotes well-established functional studies that support no damaging effect (Brnich et al., 2019).

Despite its value, functional evidence remains difficult to use at scale. Variant mentions are inconsistent across the literature and may appear as rsIDs, HGVS expressions across transcripts (Hart et al., 2024), protein-level shorthand, genomic coordinates, or legacy nomenclature. This heterogeneity complicates variant-level retrieval and identity alignment. In addition, the experimental details needed for clinical interpretation are often absent from abstracts and distributed across full text, figures, tables, and supplements. ClinGen recommendations clarify how assay validity, calibration, and concordance with disease mechanism should affect PS3/BS3 application and evidence strength, including supporting, moderate, strong, and very strong levels (Brnich et al., 2019). These signals are rarely reported in a uniform, machine-readable form.

Recent work has begun to operationalize LLMs for evidence-centered clinical genetics workflows. The Evidence Aggregator (EvAgg) demonstrated a generative-AI pipeline for rare disease case analysis that retrieves gene-relevant publications and extracts structured case and variant details for analyst review (Twede et al., 2025). CGBench intro-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country.**AUTHORERR: Missing \icmlcorrespondingauthor.**

Submitted to the ICML Workshop 2026. Do not distribute.

duced a ClinGen-derived benchmark for evaluating whether language models can extract and score evidence from scientific papers under guideline-style instructions, including judge-based evaluation against curator explanations (Queen et al., 2025). AutoPM3 targeted the ACMG/AMP PM3 evidence category by combining retrieval with structured extraction to identify supporting evidence from narrative text and tables (Li et al., 2025). These studies show substantial promise, but they also highlight unresolved challenges. Variant interpretation workflows require scalable literature screening, reliable variant grounding, multimodal understanding of full papers, and human-centered outputs that present traceable evidence for expert verification.

Here, we focus on variant-centered functional evidence mining for PS3/BS3. This setting addresses a practical bottleneck in clinical curation. Curators must determine whether functional evidence in a publication is attributable to the variant under review and interpretable under ACMG/AMP criteria. Because a single paper may discuss multiple variants, assays, and disease mechanisms, functional evidence is clinically useful only when it is linked to the correct variant and reported with sufficient experimental context. Our goal is not to replace expert judgment, but to reduce the search and extraction burden while preserving traceability to the underlying experimental evidence.

We make three contributions. First, we construct VarLitBench, a ClinGen-anchored benchmark for abstract screening and full-paper functional evidence extraction. Second, we evaluate four multimodal LLMs under an explicit variant-matching gate designed to reduce misattribution risk. Third, we implement VarLitAgent, an end-to-end variant-to-report pipeline for human-in-the-loop curator review. Figure 1 summarizes the VarLitAgent workflow.

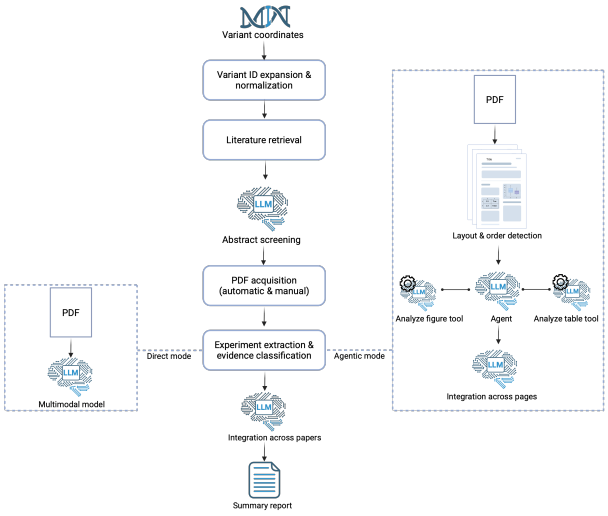


Figure 1. **Overview of the VarLitAgent workflow.** Starting from a user-specified variant, the pipeline performs variant normalization and synonym expansion across common identifiers, retrieves literature and screens abstracts to prioritize likely functional studies, acquires PDFs and performs variant matching, extracts structured experiments and PS3/BS3-aligned evidence interpretations, and generates a report for curator verification and downstream use.

2. Materials and Methods

2.1. ClinGen Curated Variants

We downloaded ClinGen curated variants (ClinGen Evidence Repository) and retained variants annotated with PS3 or BS3 at any strength level, including supporting, moderate, strong, and very strong. For each selected variant, we used the expert-written comments describing the available evidence and the rationale for classifying the variant as pathogenic, benign, or of uncertain significance.

2.2. Structuring Curator Summaries and Retrieving Linked Literature

We transformed curator summaries into structured fields using `gpt-4.1` through the OpenAI API (OpenAI API Documentation). The prompt instructed the model to extract PubMed IDs, PS3/BS3 level assignments, evidence strength, and narrative passages explaining the conclusions drawn from specific publications. Evidence strength was represented as very strong, strong, moderate, or supporting. The full prompt is provided in Appendix A. We manually checked a random subset of 20 variants and confirmed that the extracted summaries were correct in that subset.

Using the extracted PubMed IDs, we retrieved titles and abstracts programmatically with the `metapub` Python library (metapub.org). When open-access full text was available, PDFs were downloaded and cached locally. For benchmarking, we restricted evaluation to programmatically retrievable PDFs to support reproducibility. For the end-to-end pipeline, we also support manual addition of PDFs to the local cache for papers that cannot be retrieved automatically, including publisher-restricted articles.

2.3. Variant Normalization and Synonym Expansion

Variants are referenced in heterogeneous formats across databases and articles, including HGVS strings, rsIDs, protein changes, and genomic coordinates. To support variant-level retrieval and matching, we constructed synonym sets. We used VariantValidator (Freeman et al., 2024) and the Ensembl VEP API (Yates et al., 2015) to normalize each input variant into a consolidated identifier set comprising rsIDs when available, HGVS genomic, coding, and protein expressions (HGVSg, HGVSc, and HGVSp), one-letter and three-letter amino acid notation, genomic coordinates in GRCh38 and GRCh37, and associated gene symbols.

2.4. VarLitBench Dataset and Tasks

VARLITBENCH consists of 463 variants labeled with PS3 or BS3, strength of evidence (i.e., very strong, strong, moderate, or supporting), corresponding PubMedIDs, abstracts, full-paper PDFs, and expert-written summaries. We defined two tasks based on VARLITBENCH: abstract-level screening and full-paper evidence extraction and classification.

2.4.1. ABSTRACT-LEVEL FUNCTIONAL EXPERIMENT SCREENING

The input was the publication title and abstract. The LLM was instructed to return a binary label indicating whether the paper was likely to describe a wet-lab functional experiment directly testing one or more genetic variants. Positive samples were 463 abstracts from VARLITBENCH. Presumed negative samples were 463 abstracts randomly sampled from papers cited in ClinGen curator comments for variants with no PS3/BS3 annotation. Because some cited papers without explicit PS3/BS3 annotation may still contain functional experiments that were not applied as functional evidence, this negative set should be interpreted as a pragmatic screening benchmark rather than a definitive absence-of-experiment reference.

2.4.2. FULL-PAPER EXPERIMENT EXTRACTION AND CLASSIFICATION

The input was the full-paper PDF together with identifiers for the target variant, including the gene symbol, rsID, and HGVS descriptions at the genomic, cDNA, and protein levels, with both one-letter and three-letter amino acid notation when available. The LLM was instructed to return a structured record containing the following fields.

- **Variant matching.** The model identified whether variants mentioned in the paper corresponded to the target variant by building an equivalence set based on rsID, genomic coordinate, cDNA change, or protein change. The `match status` indicated the strength of this match. We used *matched* for exact matches through rsID, genomic, cDNA, or protein notation; *heuristic matching* for plausible matches using non-standard notation, such as the same amino acid substitution written as "R158W mutant" or "R158→W"; *single-variant-study matching* when the paper tested only one specific variant in the gene and no other variants appeared in the functional results; and *unsuccessful* when no plausible variant match was found. The `match type` recorded the identifier used, including rsID, genomic, cDNA, protein, multiple, or heuristic. For each match, the system reported `confidence`, `matched strings`, and brief `notes` explaining the decision.
- **Experiments.** For each relevant experiment, the model extracted the assay type, experimental system, material source, readout with units when available, normal comparator, result direction, effect size or statistics when reported, controls and validation, authors' conclusion, supporting text location, caveats, and confidence that the experiment pertained to the target variant. Result direction was represented as functionally abnormal, functionally normal, intermediate, mixed, or unclear.
- **Overall evidence.** The model assigned an aggregate PS3/BS3 direction, or `not_clear`, together with a strength assignment of very strong, strong, moderate, supporting, or `not_clear`, and a brief rationale.
- **Summary.** The model generated a two- to five-sentence narrative explaining the evidence underlying the final decision.

We used `not_clear` as an explicit abstention outcome when either the variant identity could not be confidently aligned to the target or the direction or strength of functional evidence could not be determined from the paper. We report coverage as the fraction of examples for which the system made a PS3 or BS3 decision.

2.5. VarLitAgent Implementation

We implemented VARLITAGENT as an end-to-end pipeline that takes a variant, represented by chromosome, position, reference allele, and alternate allele, and produces a structured evidence report. The pipeline includes the following steps.

1. **Annotation.** Query the Ensembl VEP REST API on GRCh38 or GRCh37 to obtain the rsID, HGVSc, HGVSp, gene symbol, MANE Select transcript (Morales et al., 2022), and Ensembl transcript ID.
2. **Literature retrieval.** Query the LitVar2 API (Allot et al., 2023) using the rsID to retrieve PubMed IDs of papers mentioning the variant.
3. **Paper details.** Fetch titles and abstracts from PubMed using `metapub`.
4. **Abstract screening.** Filter papers using an LLM to retain those likely to contain variant-level functional experiments. The benchmark includes OpenAI models, such as `gpt-4o-mini` and `o4-mini`, and Anthropic models, such as `claude-haiku-4-5` and `claude-sonnet-4-5`. The pipeline routes calls by provider-specific model name.
5. **PDF acquisition.** For abstracts predicted to report functional evidence, attempt automatic PDF download

through `metapub`. The system also allows users to manually add papers to the cache when programmatic retrieval fails.

6. **Full-text extraction.** Extract evidence from PDFs using either direct mode or agentic mode.
   - **Direct mode.** Submit the full PDF to a multimodal LLM in a single call and request variant-specific functional evidence extraction and classification. This mode is relatively fast and is the default, but it may miss details embedded in visual elements such as figures, plots, and complex tables.
   - **Agentic mode.** Decompose the paper into page-level units and run a lightweight document-understanding pipeline before LLM-based interpretation. The system renders pages to images, applies OCR using PaddleOCR (Cui et al., 2025) or EasyOCR (EasyOCR Repository), infers reading order and section structure using a layout-aware model (LayoutReader; LayoutLMv3-based (Pang, 2024)), routes non-textual regions to specialized vision tools for table and chart parsing, and uses an LLM to aggregate extracted information per page. These components are orchestrated by a LangChain (langchain.com) agent that selects tools for each page or region, repeats extraction when information is incomplete, and produces a normalized intermediate representation of sections, paragraphs, figures, tables, and linked captions. This mode is slower but can provide deeper analysis of figure- and table-heavy papers.

   All benchmark results in the main analysis use direct mode. Agentic mode is evaluated as a prototype on a small balanced subset and is included to illustrate extensibility for visually complex publications.

7. **Evidence integration.** Aggregate extracted experiments across papers to determine overall PS3/BS3 direction, strength, and confidence, guided by ACMG/AMP criteria (Richards et al., 2015) and ClinGen recommendations (Brnich et al., 2019).

8. **Report generation.** Generate an HTML report, optionally exported to PDF, summarizing variant metadata, retrieved papers, extracted experiments, and the integrated assessment. For programmatic use, the system can also return structured JSON containing per-paper extractions and the final decision.

2.6. Models and Evaluation

For direct-mode benchmarking, we evaluated `gpt-4o-mini`, `o4-mini`, `claude-haiku-4-5`, and `claude-sonnet-4-5`. For agentic mode, we report prototype results for `gpt-4o-mini` and `o4-mini` on a random balanced subset of 20 variants, including 10 PS3 and 10 BS3 variants, because this mode has substantially higher cost and runtime.

OpenAI models were accessed through the OpenAI API, and Claude models were accessed through the Anthropic API. Provider routing was determined by model name. We enforced machine-readable outputs using Pydantic schemas (Pydantic repository) and orchestrated LLM and agent workflows with LangChain.

To assess whether an LLM-generated evidence summary faithfully matched the corresponding ClinGen rationale, we used `gpt-4.1` as an independent rater (Zheng et al., 2023). The judge was shown the ClinGen expert-written text restricted to the portion relevant to PS3/BS3 and the LLM-extracted evidence summary. It then scored correspondence between the two on a 1–5 ordinal scale, where 1 indicated a poor match and 5 indicated a near-complete match. The rubric emphasized factual consistency and coverage of key experimental outcomes while penalizing unsupported claims. To reduce stochasticity and mitigate presentation-order bias, we repeated scoring three times per example with A/B order swapping and aggregated scores by majority vote (Wang et al., 2023; Saha et al., 2025).

We report accuracy, precision, recall, F1 score, and specificity for binary tasks; macro-averaged metrics for multi-class tasks; and coverage, defined as the fraction of examples for which the model produced a decision rather than `not_clear`. We also report cost and runtime distributions for direct and agentic modes.

3. Results

3.1. VarLitBench Dataset

Starting from 1,709 ClinGen curated variants carrying PS3 or BS3 evidence at any strength, we restricted the benchmark to variants with programmatically retrievable PDFs, yielding 463 samples. Each sample includes a variant identifier, functional evidence direction (PS3 or BS3), evidence strength (very strong, strong, moderate, or supporting), an abstract, a PDF, and an expert-written curator rationale. For full-paper analysis in direct mode, we used all 463 samples in VarLitBench. For agentic mode, we used a balanced subset of 20 samples. For abstract screening, we augmented VarLitBench with 463 presumed negative abstracts randomly sampled from papers cited in ClinGen curator comments for variants without PS3/BS3 annotations.

3.2. Abstract-Level Functional Experiment Screening

We prompted the LLMs to screen titles and abstracts for whether a paper was likely to report variant-linked functional experiments. Figure 2 shows high recall across all four models (0.873–0.904), consistent with the intended use of abstracts as a first-pass filter in which missed relevant studies are costly. `gpt-4o-mini` achieved the highest recall (0.904), whereas `claude-sonnet-4-5` achieved the highest accuracy (0.769), specificity (0.662), and F1 score (0.792). These results indicate that both model families can surface most candidate functional studies from abstract-level information, although specificity remains limited when abstracts do not provide enough detail to confirm variant-level experimental relevance.

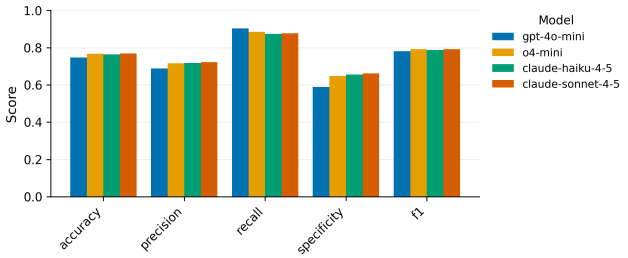


Figure 2. Abstract-level screening performance for identifying variant-linked functional experiments. All benchmarked models achieve high recall, supporting their use as abstract filters to surface candidate functional studies for downstream full-text review.

3.3. Variant Matching

Full-paper extraction is clinically meaningful only when the paper can be linked confidently to the target variant. Figure 3 summarizes variant-matching status by model. When only exact identifier matches were counted as successful, `o4-mini` and `claude-sonnet-4-5` had higher exact-match rates than `claude-haiku-4-5` and `gpt-4o-mini`. When exact, heuristic, and single-variant-study matches were all counted as successful, `claude-haiku-4-5` and `gpt-4o-mini` recovered additional plausible matches. Across models, a model-dependent fraction of variants could not be matched. Given the clinical risk of attributing functional findings to the wrong variant, the system is designed to prefer `not_clear` when alignment is uncertain.

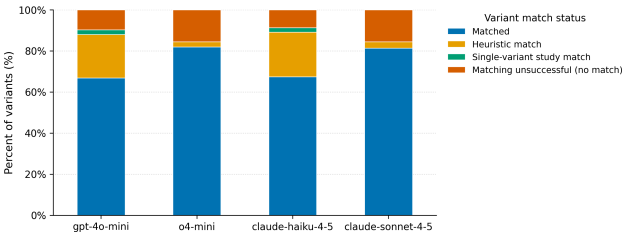


Figure 3. Variant matching outcomes by model. Most successfully resolved pairs are matched through exact identifier detection, with additional matches obtained through heuristic alignment or single-variant-study inference. A model-dependent fraction remains unmatched.

3.4. Full-Paper PS3 Versus BS3 Direction Classification

Figure 4 reports PS3 versus BS3 direction classification on examples for which each model produced a decision after variant matching. `not_clear` outcomes are summarized through coverage. All four models performed strongly on decided cases. `o4-mini` achieved the highest accuracy (0.963) and F1 score (0.979), with high specificity (0.828) and coverage of 0.916. The Claude models were competitive. `claude-sonnet-4-5` reached 0.938 accuracy and 0.965 F1 score at 0.978 coverage, while `claude-haiku-4-5` reached 0.927 accuracy and 0.959 F1 score at 0.972 coverage. Compared with `gpt-4o-mini`, the other three models produced substantially higher specificity, indicating fewer false PS3 calls when BS3 was the ground truth.

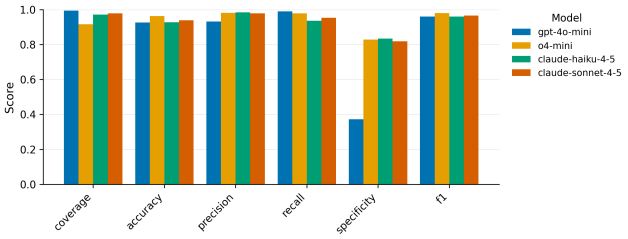


Figure 4. Full-text evidence direction performance (PS3 vs. BS3) on decided examples after variant matching. All models achieved high F1 scores. `o4-mini` and the Claude models improved specificity relative to `gpt-4o-mini`.

3.5. Evidence Strength Classification Remains Challenging

Strength grading remained difficult. Figure 5 shows low accuracy (0.292–0.335) and low macro-F1 (0.147–0.216) for the eight-way classification combining direction (PS3 versus BS3) and strength (supporting, moderate, strong, and very strong). These results suggest that overall functional direction is often recoverable, whereas strength depends on assay validation, calibration, and disease-mechanism concordance signals that may be implicit, fragmented, absent from the main text, or reported only in supplements.

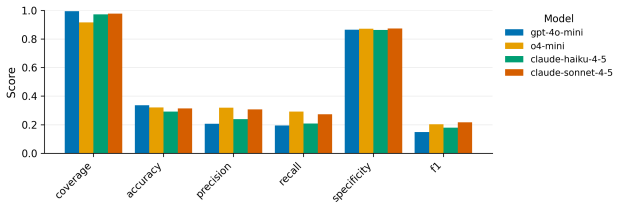


Figure 5. Joint direction and strength classification performance on decided examples after variant matching. The eight classes combine PS3 or BS3 direction with supporting, moderate, strong, or very strong evidence strength.

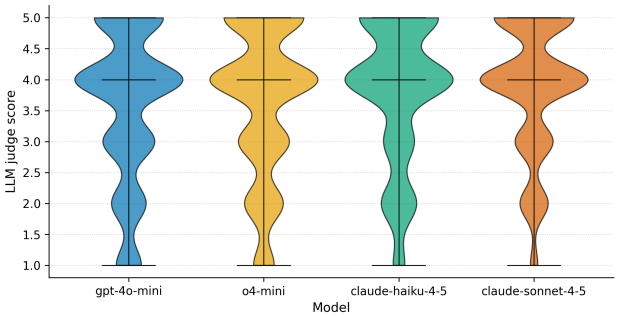
3.6. LLM-as-Judge Evaluation of Evidence Summary Correspondence

Beyond label accuracy, we assessed whether each model's extracted evidence summary matched the expert-written ClinGen PS3/BS3 rationale using an independent LLM rater. The judge assigned a 1–5 correspondence score, with higher values indicating better alignment, and provided a confidence score from 0 to 100 for each assessment. Figure 6 summarizes the evaluated examples. The Claude models achieved the highest mean correspondence scores, with 3.837 for `claude-sonnet-4-5` and 3.766 for `claude-haiku-4-5`, compared with 3.628 for `o4-mini` and 3.563 for `gpt-4o-mini`. Median correspondence was 4.0 for all models, and median judge confidence was 95 for all models except `gpt-4o-mini`, for which it was 92. These results suggest broadly similar summary alignment across models, with modestly stronger mean correspondence for Claude outputs in this evaluation.

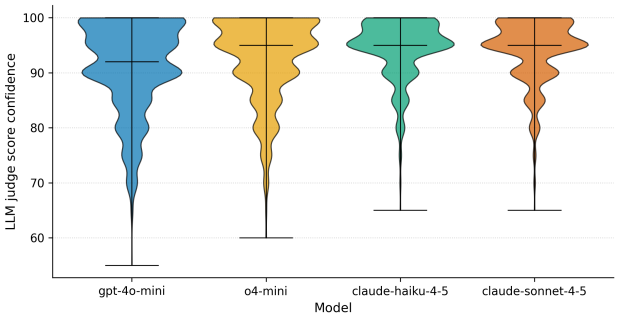
3.7. Comparing Direct and Agentic Full-Paper Extraction

We compared direct and agentic full-paper extraction on a balanced subset of 20 samples, including 10 PS3 and 10 BS3 examples. As shown in Figure 7a, agentic mode improved precision, specificity, and F1 score for both `gpt-4o-mini` and `o4-mini`, suggesting that page-level parsing and figure/table-aware extraction can reduce false direction assignments in difficult cases. Because this evaluation used only 20 samples and two models, these results should be interpreted as preliminary.

The improvement came at higher cost and runtime. Median cost increased from $0.017 to $0.058 per sample for `gpt-4o-mini`, and from $0.045 to $0.174 per sample for `o4-mini` (Figure 7b). Median runtime increased from 40.4 to 241.1 seconds for `gpt-4o-mini`, and from 36.7 to 338.0 seconds for `o4-mini` (Figure 7c). Direct mode is therefore better suited for scalable first-pass extraction, whereas agentic mode may be useful as a selective second pass for complex or ambiguous papers.



(a) Correspondence scores.



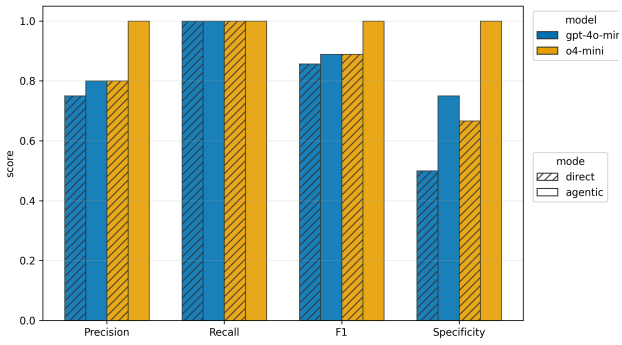
(b) Judge confidence scores.

Figure 6. LLM-as-judge evaluation of evidence-summary correspondence to ClinGen PS3/BS3 rationales. Panel (a) shows correspondence scores and panel (b) shows judge confidence scores.

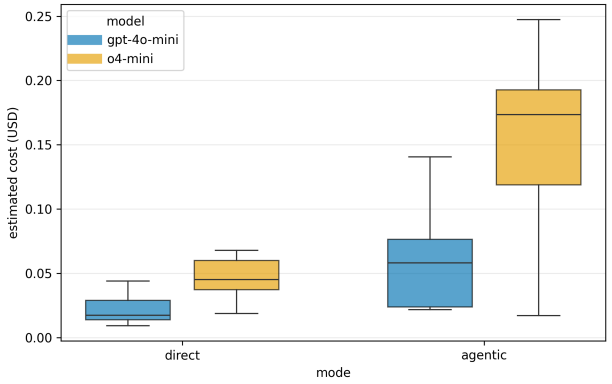
3.8. VarLitAgent: End-to-End Pipeline Output and Report Visualization

Although the benchmark evaluates individual subtasks, real curation requires these steps to be connected into a traceable end-to-end workflow. We therefore implemented a variant-centered system that starts from a genomic variant and produces a curator-oriented evidence package aligned to ACMG/AMP functional criteria.

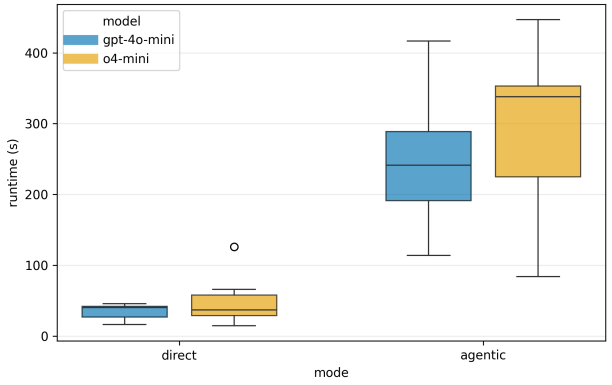
Given chromosome, position, reference allele, and alternate allele, the pipeline normalizes the input into a consolidated identifier set that includes rsID when available, HGVSg, HGVSc, HGVSp, gene symbol, and transcript context. It then retrieves literature through LitVar2, screens candidate abstracts for likely variant-linked functional assays, attempts PDF acquisition, and applies multimodal extraction to retained full papers. For each matched paper, the system extracts structured experiment records covering assay, system, readout, comparator, direction, controls, caveats, and supporting locations. It then integrates evidence across papers into an ACMG/AMP-aligned assessment. The pipeline is conservative under uncertainty. If variant identity cannot be confidently aligned, or if functional direction or strength cannot be supported by the available evidence, it returns a no-decision outcome rather than forcing a PS3 or BS3 call.



(a) Direct versus agentic mode performance.



(b) API cost per sample.



(c) Runtime per sample.

Figure 7. Direct versus agentic full-paper extraction on the balanced 20-sample subset. Agentic mode improves some performance metrics but increases both API cost and runtime.

The primary output is a human-reviewable HTML or PDF report for curator verification. The report consolidates normalized variant metadata and transcript context, retrieved literature with abstract-screening outcomes and download status, per-paper matching notes and extracted experiments with citations to relevant sections or figures when available, and an integrated ACMG/AMP-aligned summary highlighting direction, proposed strength, confidence, and considerations affecting PS3/BS3 use. These considerations include assay validation signals, calibration controls, and concordance with disease mechanism. The pipeline can additionally return structured JSON for downstream analyses.

Supplementary Figure S1 shows representative report sections, including variant normalization, candidate-paper screening, extracted functional experiments, and the integrated curator-facing assessment.

4. Discussion

We assessed whether modern multimodal LLMs from two provider families can assist ACMG/AMP-aligned functional-evidence curation across abstract triage, variant-linked extraction from full papers, and PS3/BS3-oriented interpretation. Three findings emerge. First, abstract screening is already useful as a high-sensitivity filter to surface candidate functional studies for curator review, and this pattern held across both OpenAI and Anthropic models. Second, full-paper evidence interpretation is reliable only when variant identity is aligned confidently, making variant matching a prerequisite rather than a peripheral subtask. Third, conditional on successful matching, PS3 versus BS3 direction is often recoverable from the paper, whereas strength grading remains the least robust component because it depends on validation and calibration signals that are frequently implicit, fragmented across figures and methods, or absent from the main text.

These results support an assistive deployment model. LLMs can add value by organizing papers into structured, traceable experiment records and producing draft curator-facing rationales that are straightforward to verify. At the same time, high-stakes failure modes, especially misattributing results to the wrong variant or making overconfident strength claims, argue for conservative gating and abstention when provenance is weak. In our design, explicit variant matching serves as a safety gate, and `not_clear` acts as a principled no-decision outcome when alignment or interpretability is insufficient.

We operationalize this workflow by connecting variant normalization and synonym expansion to literature retrieval, abstract filtering, full-text multimodal extraction, and report generation aligned to PS3/BS3 application. The curator-facing report emphasizes auditability. Extracted claims

are paired with supporting locations and consolidated into an ACMG/AMP-consistent assessment, enabling rapid verification and targeted follow-up reading. More broadly, the benchmark provides a concrete substrate for measuring progress on end-to-end functional evidence mining under realistic curation constraints.

Several limitations motivate concrete next steps. First, although we benchmarked four models across OpenAI and Anthropic, broader evaluation across additional model families, model sizes, prompting strategies, and inference settings is needed to assess generality and characterize trade-offs such as specificity versus coverage. Second, agentic extraction mode was evaluated only on a small balanced subset and only for two models. The observed improvement is encouraging but preliminary, and larger evaluation is needed to determine when the additional cost and runtime are justified. Third, the main PDF is often insufficient for strength grading and sometimes insufficient for direction classification because validation details and calibration controls are frequently reported in supplementary files, extended methods, or external repositories. Treating supplements as first-class inputs and explicitly tracking whether claims are supported by main-text or supplementary evidence are important extensions. Fourth, the benchmark inherits uncertainty from curated rationales. ClinGen summaries reflect expert interpretation, but judgments can vary across curators, panels, and time as standards evolve. Future work could quantify inter-curator variability where multiple rationales exist and evaluate against multi-annotator references or adjudicated consensus to better separate model error from label ambiguity. Finally, although the pipeline emphasizes auditable extraction, LLMs can still generate unsupported inferences. Strengthening provenance guarantees is therefore a priority. Important extensions include tighter evidence-first constraints, automated consistency checks that flag claims lacking textual or figure support, and robustness testing through systematic perturbations of prompts, formatting, and schemas within a versioned evaluation harness.

In summary, multimodal LLMs offer a practical opportunity to make functional-evidence curation faster and more consistent. They can absorb much of the upfront search and organization burden by converting relevant information into structured, curator-friendly evidence records and draft rationales. By shifting effort from information gathering to focused verification, this approach can support higher-throughput activities such as panel updates, large-scale reinterpretation, and ongoing literature surveillance while keeping final adjudication with expert curators. VarLitBench and VarLitAgent provide an extensible foundation for incorporating additional evidence sources, stronger provenance checks, and reliability safeguards, with the goal of reducing time-to-curation without compromising transparency.

References

Allot, A., Wei, C.-H., Phan, L., Hefferon, T., Landrum, M., Rehm, H. L., and Lu, Z. Tracking genetic variants in the biomedical literature using LitVar 2.0. *Nat. Genet.*, 55(6): 901–903, June 2023.

Brnich, S. E., Abou Tayoun, A. N., Couch, F. J., Cutting, G. R., Greenblatt, M. S., Heinen, C. D., Kanavy, D. M., Luo, X., McNulty, S. M., Starita, L. M., Tavtigian, S. V., Wright, M. W., Harrison, S. M., Biesecker, L. G., Berg, J. S., and Clinical Genome Resource Sequence Variant Interpretation Working Group. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.*, 12(1):3, December 2019.

ClinGen Evidence Repository. URL https://erepo.clinicalgenome.org/evrepo/.

Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., Zhang, Y., Lv, W., Huang, K., Zhang, Y., Zhang, J., Zhang, J., Liu, Y., Yu, D., and Ma, Y. Paddleocr 3.0 technical report, 2025. URL https://arxiv.org/abs/2507.05595.

EasyOCR Repository. URL https://github.com/JaidedAI/EasyOCR.

Freeman, P. J., Wagstaff, J. F., Fokkema, I. F. A. C., Cutting, G. R., Rehm, H. L., Davies, A. C., den Dunnen, J. T., Gretton, L. J., and Dalgleish, R. Standardizing variant naming in literature with VariantValidator to increase diagnostic rates. *Nat. Genet.*, 56(11):2284–2286, November 2024.

Hart, R. K., Fokkema, I. F. A. C., DiStefano, M., Hastings, R., Laros, J. F. J., Taylor, R., Wagner, A. H., and den Dunnen, J. T. HGVS nomenclature 2024: improvements to community engagement, usability, and computability. *Genome Med.*, 16(1):149, December 2024.

langchain.com. URL https://www.langchain.com.

Li, S., Wang, Y., Liu, C.-M., Huang, Y., Lam, T.-W., and Luo, R. AutoPM3: enhancing variant interpretation via LLM-driven PM3 evidence extraction from scientific literature. *Bioinformatics*, 41(7), July 2025.

metapub.org. URL https://metapub.org.

Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J. M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V. S., Kay, M., Kodali, V. K., McGarvey, K. M., McMahon, A., Mudge, J. M., Murphy, D. N.,

Murphy, M. R., Rajput, B., Rangwala, S. H., Riddick, L. D., Thibaud-Nissen, F., Threadgold, G., Vatsan, A. R., Wallin, C., Webb, D., Flicek, P., Birney, E., Pruitt, K. D., Frankish, A., Cunningham, F., and Murphy, T. D. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, 604(7905):310–315, April 2022.

OpenAI API Documentation. URL https://platform.openai.com/docs/overview.

Pang, H. Faster LayoutReader based on LayoutLMv3, February 2024. URL https://github.com/ppaanngggg/layoutreader.

Pydantic repository. URL https://github.com/pydantic/pydantic/tree/main.

Queen, O., Zhang, H. G., and Zou, J. CGBench: Benchmarking language model scientific reasoning for clinical genetics research. October 2025.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., and ACMG Laboratory Quality Assurance Committee. 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. *Genet. Med.*, 17(5):405–424, May 2015.

Saha, S., Li, X., Ghazvininejad, M., Weston, J., and Wang, T. Learning to plan & reason for evaluation with Thinking-LLM-as-a-Judge. 2025.

Twede, H., Pais, L., Bryen, S., O'Heir, E., Smith, G., Paulsen, R., Austin-Tse, C. A., Bloemendal, A., Simons, C., Hall, A. K., Saponas, S., Wander, M., MacArthur, D. G., Rehm, H. L., and Conard, A. M. The evidence aggregator: AI reasoning applied to rare disease diagnostics. March 2025.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. 2023.

Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G. R. S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P. The ensembl REST API: Ensembl data for any language. *Bioinformatics*, 31(1):143–145, January 2015.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena. 2023.

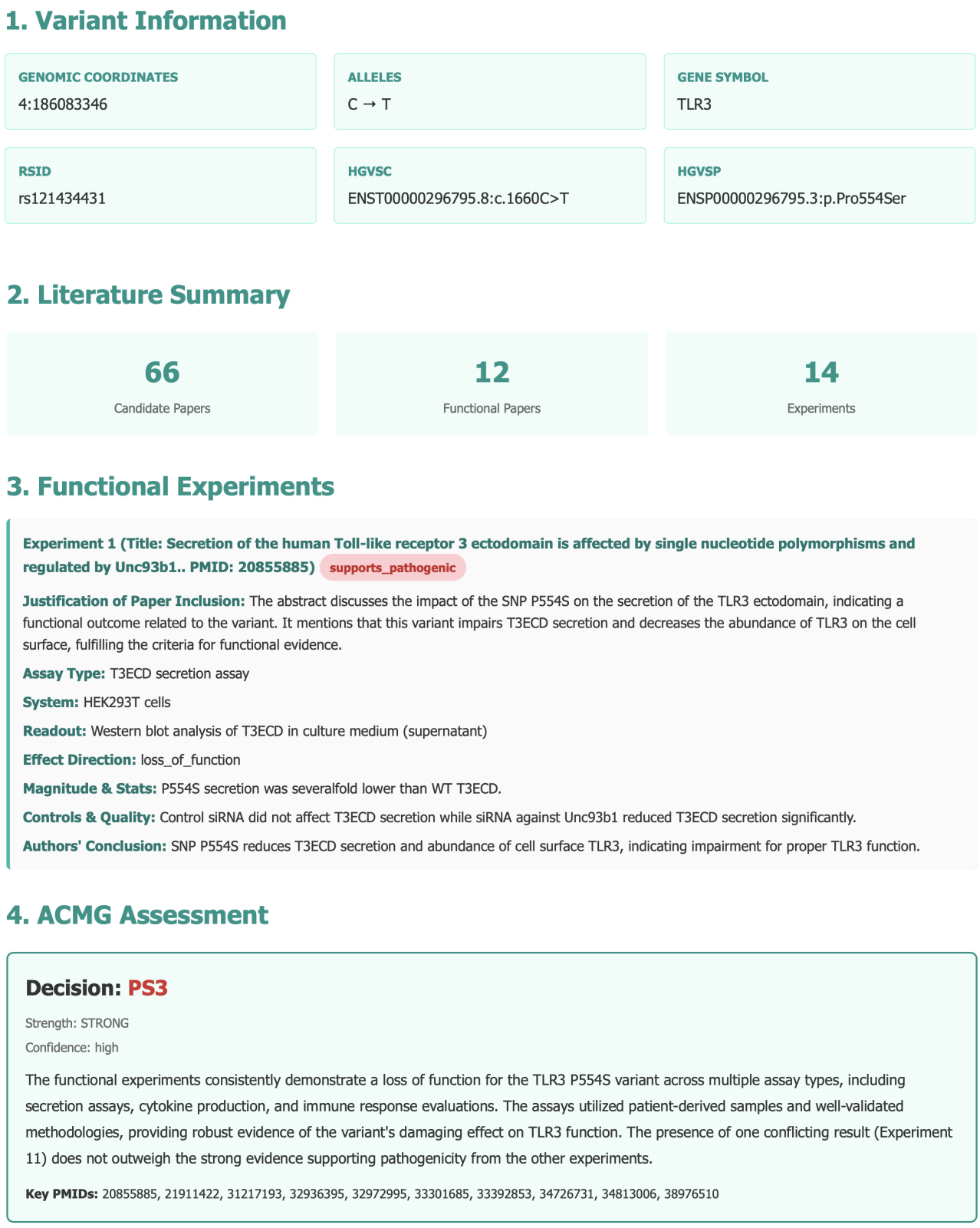


Figure S1. Representative screenshots from the pipeline's evidence report. The report is designed for curator review and compiles variant normalization and transcript context, retrieved literature with screening decisions, per-paper variant matching and extracted functional experiments, and an integrated ACMG/AMP-aligned assessment summarizing PS3/BS3 direction, proposed strength, confidence, and key considerations.

A. Prompt Templates

A.1. Abstract Screening Prompt

```
You are a clinical variant interpretation curator performing an abstract-level screen for
    ACMG/AMP PS3/BS3 relevance.

Goal
Decide ONLY whether the abstract contains ANY experimental (wet-lab) functional evidence
    about the effect of one or more genetic variants/mutations/alleles/mutants on gene
    product function (protein or RNA) or a disease-relevant functional pathway/output.

Output
Return ONLY: functional_experiment = 1 or 0

Bias / sensitivity requirement (important)
This screen is intentionally high-sensitivity. If there is reasonable doubt, classify as 1
     so the paper can be reviewed downstream.
Default to 1 whenever BOTH (i) variant/mutant language and (ii) any wet-lab functional
    assay signal are present, even if details are sparse.

Classify functional_experiment = 1 if the abstract shows BOTH:

A) Variant-or-mutant subject (broad; exact IDs NOT required)
Any of the following counts:
- Specific variant(s) listed (HGVS, rsID, amino-acid change, "c."/"p.", etc.)
- "patient mutations/variants", "disease-causing mutations", "mutant alleles", "allelic
    series"
- "mutant constructs", "site-directed mutants", "missense mutants", "variant panel", "
    mutagenesis"
- Engineered or edited variant models (knock-in, CRISPR-introduced variant, engineered
    mutant protein)
- Patient-derived samples where the abstract links results to the mutation(s) (even
    broadly)

AND

B) Wet-lab functional assay + outcome statement
There is an experimental functional readout and the abstract states an outcome for the
    variant(s), including qualitative direction.
Examples of outcome language: reduced/abolished/impaired, increased/gain, altered,
    disrupted, restored/rescued, mislocalized, unstable, no difference/normal, defective
    splicing, NMD, truncated protein with loss of activity, etc.

Count as functional evidence (any wet-lab) if the abstract includes one or more of:

1) Protein/biochemical function (in vitro or cellular)
- Enzymatic activity/kinetics, catalytic function, substrate turnover
- Binding/interaction/complex formation
- Protein stability/folding/degradation/half-life
- Localization/trafficking/secretion
- Channel transport, receptor/signaling output, post-translational effects tied to
    function

2) Cell-based functional consequences
- Reporter assays, pathway activity, electrophysiology, transport flux
- Rescue/complementation (WT vs mutant; mutant fails to rescue or rescue restores)
- Mechanistic cellular phenotypes tied to function (e.g., DNA repair capacity, metabolic
    function, stress sensitivity) with mutant-vs-WT comparison

3) RNA-level functional assays attributable to a variant
- Splicing assays (patient RNA/cDNA, RT-PCR, minigene) showing aberrant splicing
- mRNA stability / nonsense-mediated decay (NMD) experimentally shown
- Translation/processing efficiency when experimentally measured
```

```
4) Model systems with variant-level manipulation
- Knock-in/engineered variant models with functional or disease-relevant phenotypes and a
    variant-linked readout

5) Patient-derived functional assays (allow, even if confounded)
- Enzyme activity, electrophysiology, pathway output, splicing defects measured in patient
     cells/tissue, when the abstract links findings to the mutation(s)

Strong "bias-to-1" tie-breakers
Return 1 if ANY of the following patterns appear:
- ("mutation/variant/mutant/allele") + a wet-lab assay keyword (activity, assay, measured,
     functional, reporter, localization, stability, splicing, RT-PCR, minigene, NMD,
    electrophysiology, rescue)
- The abstract claims functional impact for mutations ("mutations impair function", "
    variants reduce activity", "mutants show defective splicing"), even without numbers.

Return functional_experiment = 0 ONLY when it is clearly NOT functional variant testing:
- Purely in silico/computational prediction with no wet-lab experiment
- Pure genetic association/segregation/case reports/phenotype-only with no functional
    readout
- Gene/pathway biology experiments (KO/overexpression/mechanism) that do NOT test variants
    /mutant constructs
- Expression/omics profiling alone (RNA-seq, differential expression) without variant-
    linked functional RNA/protein consequences
  (Exception: explicit variant-driven splicing or experimentally shown NMD/mRNA
      instability)

Final rule
If you can point to (A) any variant/mutant subject AND (B) any wet-lab functional readout
    with an outcome claim, output 1. Otherwise output 0.

USER INPUT TEMPLATE

PMID: {pmid}

Abstract:
"""{abstract}"""
```

A.2. Full-Paper Evidence Extraction Prompt

```
You are a clinical variant functional-evidence extractor for ACMG/AMP guidelines PS3/BS3
    criteria.

INPUTS
- TARGET_VARIANT: gene + identifiers (any of rsID, HGVSg, chr_pos_ref_alt,
  HGVSc, HGVSp, aliases).
- PAPER: a full PDF (may include many variants).

GOAL
- Find all plausible variant-level functional experiments that might correspond to the
    TARGET_VARIANT. Read all PDF (text, tables, figure captions, and figure panels/
    embedded labels)
- Be SENSITIVE: when in doubt, extract and clearly mark uncertainty.
- Do NOT hallucinate data.

OUTPUT
- Return ONLY valid JSON that matches the schema exactly.
- Use double quotes for all keys and strings.
- No commentary outside JSON.

-----------------------------------
1. VARIANT MATCHING (SOFT GATE)
-----------------------------------
Build an equivalents set for the TARGET_VARIANT (without inventing mappings):
```

```
- Same rsID
- Same genomic coordinates (exact chr:pos:ref:alt or HGVSg as given)
- Same cDNA change (c.notation; allow formatting variants)
- Same protein change (same ref AA, same position, same alt AA;
  allow 1-letter <-> 3-letter and formatting variants)

Do NOT:
- Change genome build
- Renumber across transcripts unless the paper explicitly gives both
- Guess transcript IDs

Match tiers (you can stop when one is clearly satisfied):

1) STRICT MATCH -> status = "matched"
   - Exact rsID, genomic, cDNA, or protein match from the equivalents set,
     in the correct gene.
   - match_type = "rsid" / "genomic" / "cdna" / "protein" / "multiple"
   - confidence:
     - "high": rsID or genomic
     - "medium": cDNA or protein + clear gene context
     - "low": identifier match but weak context

2) SINGLE VARIANT STUDY -> status = "single_variant_study_matching"
   - Functional experiments in this gene clearly test ONE specific variant only.
   - No other specific variants appear in functional results.
   - confidence:
     - "medium" if gene and clinical context are clear
     - "low" if context is weaker
   - match_type = "single_variant_study"

3) HEURISTIC MATCH -> status = "heuristic_matching"
   Use for plausible, non-strict matches in the same gene. Count applicable clues:

   Clues:
    - Same amino-acid substitution (same ref AA, position, alt AA) but written
    in words or non-standard notation (e.g. "R158W mutant", "R158->W").
    - Explicit numbering / isoform / precursor->mature mapping that links positions
    to the same amino acid change.
    - Different cDNA / protein numbering that the paper directly ties together
    (e.g. "c.472C>T (R158W)").
    - Shorthand label ("mut1", "A", etc.) that is expanded elsewhere to a notation
    matching the TARGET_VARIANT equivalents.
    - Table / figure / text cross-reference that explicitly equates two labels
    as the same variant.
    - Multiplex / saturation screen where the authors systematically test single
    substitutions and the tested set clearly includes the TARGET codon / position
    (e.g. "all single-amino-acid substitutions at residue 158").

    Never call heuristic_matching based only on:
    - Same exon / domain / region, "nearby" codon, or vague proximity.
    - Gene-level statements with no specific variant label.

   - match_type = "heuristic"
   - confidence: "low" if 1 clue; "medium" if >=2 clues

4) NO PLAUSIBLE VARIANT -> status = "variant_matching_unsuccessful"
   - Use ONLY when you find no specific variant in this gene that could
     reasonably be the TARGET_VARIANT.
   - In this case: experiments = [] and overall_evidence.evidence_level =
     "not_clear" and evidence_strength = "not_clear".

IMPORTANT SENSITIVITY RULE:
- If you see any specific variant in the SAME GENE that could plausibly be the
  TARGET_VARIANT, you SHOULD:
  - Assign "matched", "single_variant_study_matching", or "heuristic_matching"
```

```
    with appropriate (often low) confidence.
  - Extract its experiments.
  - Explain uncertainty in variant_match.notes and overall_evidence.basis.
- Only use "variant_matching_unsuccessful" when there is truly no plausible
  candidate.

-------------------------------------
2. EXPERIMENT EXTRACTION
-------------------------------------
Extract experiments ONLY for the variant(s) linked to the TARGET_VARIANT by
your chosen status (matched / single_variant_study_matching / heuristic_matching).

INCLUDE:
- Experiments where the specific variant label (e.g. "R158W", "mut1",
  "c.472C>T") has its own row, bar, lane, or result.
- Variant-level results in tables, figures, or text.

EXCLUDE:
- Purely in silico predictions.
- Case reports or association studies with no functional assay.
- Results where variants are pooled and no individual variant result is given.

For each experiment, record:
- What the assay is (assay)
- The system used (system)
- How the variant material was obtained (variant_material)
- The measured endpoint (readout)
- The explicit comparator (normal_comparator: WT/healthy/threshold)
- The functional direction and any numbers (result.direction and
  result.effect_size_and_stats)
- Controls and validation details (controls_and_validation)
- Authors' explicit conclusion about the variant (authors_conclusion)
- Where it appears (where_in_paper)
- Limitations stated in the paper (caveats)
- Exact variant label in the paper (paper_variant_label)
- How strongly you link that label to the TARGET_VARIANT
  (variant_link_confidence).

If you find NO functional assay on the matched variant:
- experiments = []
- overall_evidence.evidence_level = "not_clear"
- overall_evidence.evidence_strength = "not_clear"
- State this in overall_evidence.basis.

-------------------------------------
3. PS3 / BS3 / not_clear
-------------------------------------
Definitions:
- PS3: Variant shows a functionally abnormal result (for example vs. a normal comparator),
consistent with a damaging effect and disease mechanism.
- BS3: Variant shows functionally normal result (for example vs. a normal comparator).
- not_clear: unclear direction, conflicting or insufficient information.

Strength (very_strong / strong / moderate / supporting / not_clear):
- supporting: comparator present + basic controls described (WT +/- positive/null) but
    limited validation
- moderate: well-established assay with clear controls/replication and/or multiple
    validation controls described
- strong/very_strong: the paper provides rigorous clinical validation/calibration
    supporting high confidence
  (e.g., multiple known benign/pathogenic controls with clear thresholds or explicit
     calibration).

If evidence_level = "not_clear":
- evidence_strength MUST be "not_clear".
```

```
------------------------------------
4. SUMMARY
------------------------------------
In summary:
- Be generous in extraction when the variant is plausibly the TARGET_VARIANT.
- Use status, confidence, variant_link_confidence, and notes to mark how sure
  you are.
- Never invent experiments or numbers.
- Output must be valid JSON according to the schema, with no extra keys.

USER INPUT TEMPLATE

TARGET_VARIANT: {target_variant_string}

Attached: 1 full-text PDF paper.

Follow the system instructions to:
- Match the TARGET_VARIANT to variant labels in the paper,
- Extract all plausible variant-level functional experiments for that variant,
- Summarize PS3/BS3 evidence and strength.

Return ONLY valid JSON that matches the schema exactly.
Do NOT add any text outside the JSON object.
```

A.3. Evidence Integration Prompt

```
You are a clinical variant interpretation curator specializing in ACMG/AMP PS3/BS3
    functional evidence criteria.

Your task is to evaluate functional experiment data extracted from scientific literature
    and determine the appropriate PS3/BS3 classification with evidence strength.

------------------------------------
ACMG/AMP DEFINITIONS (Richards et al., 2015)
------------------------------------

PS3 (Pathogenic Strong): Well-established in vitro or in vivo functional studies
    supportive of a DAMAGING effect on the gene or gene product.

BS3 (Benign Strong): Well-established in vitro or in vivo functional studies show NO
    DAMAGING effect on protein function or splicing.

Key considerations from ACMG/AMP:
- Functional studies can be powerful but not all assays are equally effective
- Consider how closely the assay reflects the biological environment
- Patient-derived samples provide stronger evidence than in vitro expression
- Full biological function assays are stronger than partial function assays
- Validation, reproducibility, and robustness are important factors
- Assays that assess mRNA-level impact (splicing, stability) can be informative

------------------------------------
ClinGen SVI STRENGTH FRAMEWORK (Brnich et al., 2019)
------------------------------------

Evidence strength depends on assay validation and quality:

**very_strong**:
- Rigorous statistical validation with formal odds of pathogenicity (OddsPath) calculation
- Extensive calibration against known pathogenic and benign variants
- Multiple independent, well-validated assays with consistent results
- This level is RARE and requires exceptional evidence

**strong**:
```

```
- Well-established assays with rigorous validation
- Multiple independent studies with consistent results
- Clear controls (wild-type, null/positive) and biological replicates
- Well-documented methodology with good reproducibility
- At least 2 high-quality papers with concordant functional data

**moderate**:
- Validated assay with multiple controls (ideally 11+ variant controls)
- Good experimental design with replicates
- Clear comparator (wild-type or healthy control)
- Single well-validated study OR multiple studies with minor inconsistencies
- Assay measures relevant biological function

**supporting**:
- Basic assay with limited validation or controls
- Single study without extensive replication
- Assay that measures only partial protein function
- Results from patient-derived material without variant-specific controls
- Limited documentation of methodology or controls

------------------------------------
DECISION RULES
------------------------------------

1. **PS3 applies when**:
   - Functional studies consistently show ABNORMAL function
   - Effect is consistent with disease mechanism (e.g., loss-of-function for
     haploinsufficiency)
   - Assay quality supports the claimed strength level

2. **BS3 applies when**:
   - Functional studies consistently show NORMAL function
   - Assay adequately captures relevant protein function
   - No evidence of damaging effect on protein or splicing

3. **not_clear applies when**:
   - Evidence is CONFLICTING (some abnormal, some normal results)
   - Evidence is INSUFFICIENT (too few experiments or poor quality)
   - Evidence is AMBIGUOUS (intermediate effects, unclear interpretation)
   - Assay does not adequately measure relevant function
   - No functional experiments available

------------------------------------
EVALUATION PROCESS
------------------------------------

For each experiment, evaluate:
1. Assay type and biological relevance
2. Quality indicators: controls, replicates, validation
3. Functional outcome: abnormal vs normal vs intermediate
4. Confidence that the tested variant matches the target variant

Then synthesize across all experiments to determine:
- Overall evidence direction (PS3 vs BS3 vs not_clear)
- Appropriate strength level
- Confidence in the assessment

------------------------------------
OUTPUT FORMAT
------------------------------------

Return ONLY valid JSON matching this schema:
{
  "decision": "PS3" | "BS3" | "not_clear",
  "strength": "very_strong" | "strong" | "moderate" | "supporting" | null,
```

```
  "confidence": "high" | "medium" | "low",
  "narrative": "2-4 sentence summary explaining the evidence and rationale for the
      decision",
  "experiment_evaluations": [
    {
      "pmid": "string",
      "assay_type": "string",
      "quality_assessment": "high" | "moderate" | "low",
      "functional_outcome": "abnormal" | "normal" | "intermediate" | "unclear",
      "supports": "PS3" | "BS3" | "neither",
      "notes": "brief note on this experiment"
    }
  ],
  "key_considerations": ["list of key factors that influenced the decision"]
}

IMPORTANT:
- If decision is "not_clear", strength MUST be null
- If no experiments are provided, decision MUST be "not_clear" with strength null
- Be conservative: when in doubt, use lower strength or "not_clear"
- Do not hallucinate or invent experimental details
- Base assessment ONLY on the provided experiment data

USER INPUT TEMPLATE

Evaluate the following functional experiments for variant: {variant_label}

-------------------------------------
EXTRACTED FUNCTIONAL EXPERIMENTS
-------------------------------------

{experiments_text}

-------------------------------------
INSTRUCTIONS
-------------------------------------

Based on the ACMG/AMP guidelines and ClinGen SVI framework described in the system prompt:

1. Evaluate each experiment for quality, relevance, and functional outcome
2. Synthesize the evidence to determine overall PS3/BS3/not_clear classification
3. Assign appropriate strength level based on validation and quality
4. Provide a clear narrative explaining your reasoning

Return your assessment as JSON following the schema in the system prompt.
```

A.4. Evidence Summary Correspondence Judge Prompt

```
SYSTEM MESSAGE

You are a careful, unbiased scientific reviewer.

USER MESSAGE TEMPLATE

You are an impartial scientific judge. Your task is to rigorously determine how well two
    explanations
refer to the same underlying evidence/experiment and make compatible conclusions about it.

Focus on CONTENT, not wording:
- One explanation may be much more terse than the other.
- Extra correct detail is OK and should not be penalized.
- Penalize contradictions, different experiments, or meaningfully different conclusions.

Avoid position bias: do not prefer A or B due to order.
```

```
Score correspondence on a 1-5 Likert scale:
1 = Not the same evidence / contradicts / mostly unrelated
2 = Some overlap but key evidence/conclusions differ or many important misses
3 = Same general evidence but incomplete, vague, or partially mismatched
4 = Same evidence and compatible conclusions; minor omissions OK
5 = Same evidence and conclusions; highly consistent; extra correct detail OK

Return JSON ONLY with:
- score (integer 1-5)
- confidence (0-100)
- rationale (<= 40 words; cite the key reason: same experiment? contradiction? missing key
    result?)

Explanation A:
{A}

Explanation B:
{B}
```