Affinage: Genome-Scale Mechanistic Gene Annotation from the Published Literature

Anonymous Authors[1]

Abstract

Gene-level annotations are a bottleneck for both biologists reasoning about unfamiliar genes and computational pipelines that embed or reason over per-gene descriptions: literature-grounded LLM retrieval is expensive per gene, while curated databases lag the literature. Here, we present Affinage, an LLM pipeline that performs literature retrieval and mechanistic reasoning once per gene — with a biologist-designed reading pass that extracts only direct experimental evidence — and stores the result as a reusable, structured, PubMed ID (PMID)-anchored annotation. The synthesis pass produces a mechanistic narrative, a per-finding mechanistic history with open questions, and a structured mechanism profile; pre-existing database sources are not considered during synthesis. Applied genome-wide to all human protein-coding genes in a two-pass pipeline with deterministic structural-QC retry, Affinage produces a substantive mechanistic narrative for 92% of annotated genes, including 28% of the genome where UniProt's curated function field is empty or under 200 characters. All 19,291 records are available through a public REST API and MCP server at `https://affinage.wi.mit.edu`, designed as a stackable base layer for downstream reasoning systems.

1. Introduction

Large language models are transforming how biologists interact with gene function information. A researcher encountering an unfamiliar gene in a screen or variant list increasingly turns to an LLM for an explanation — but a literature-grounded answer requires a retrieval session that searches PubMed, reads abstracts, and synthesizes across papers, which is expensive per gene and non-reproducible across users. Without retrieval, the model draws on training data that may be stale and cannot provide verifiable citations. At the same time, computational groups building gene-level models — perturbation predictors (Chen & Zou, 2025; 2024), cell-type classifiers (Istrate et al., 2024), and knowledge-graph constructors (Li et al., 2025) — need faithful per-gene representations as input features or context. The default sources are UniProt function descriptions (The UniProt Consortium, 2023) and NCBI gene summaries, but these lag the literature: UniProt is constrained by manual curation cycles, OMIM (Amberger et al., 2019) captures disease links faster than mechanism, and Gene Ontology (Ashburner et al., 2000; Aleksander et al., 2023) cannot express the substrate, structural, or partner-level detail that mechanistic reasoning requires. Groups outside the immediate domain may lack the expertise to design extraction criteria that distinguish direct mechanistic evidence from phenotypic association — yet this distinction determines whether a downstream model reasons over mechanism or over noise.

Here, we present Affinage, which addresses both problems by performing the expensive literature retrieval and mechanistic reasoning *once* per gene, with prompts designed by a biologist to encode what counts as direct experimental evidence, and storing the result as a structured, PMID-anchored annotation that any downstream consumer — human or computational — can query at lookup cost. The system reads only published literature (PubMed (Sayers et al., 2022), Europe PMC (Europe PMC Consortium, 2015), bioRxiv); pre-existing database sources are never shown to the LLM. A deterministic evaluation layer (R1–R9) audits the output with no LLM in the evidentiary chain. The resulting resource is designed as a stackable base layer: downstream systems performing target identification, screen interpretation, or hypothesis generation can query per-gene mechanism at API cost rather than re-deriving it from the literature.

2. Methods

Affinage runs as a two-pass pipeline. Pass 1 annotates every HGNC protein-coding gene; Pass 2 is gated by three deterministic structural rules that detect corpus-level failures and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

re-runs only the flagged subset under an augmented corpus and identity-anchored prompt. Full implementation details, batch mechanics, and cost breakdowns are in Appendices A and A.2.

2.1. Pass 1: genome-wide annotation

For each gene, Pass 1 runs three stages.

Stage 0 (source-paper retrieval) assembles a per-gene literature corpus without any LLM involvement, mirroring the search a biologist would perform when reviewing a gene for the first time. PubMed E-utilities (Sayers et al., 2022) are queried with a title-restricted search across the canonical symbol plus up to five HGNC aliases (Seal et al., 2023) (previous and current symbols), filtered by biological terms to suppress non-biomedical hits. Short symbols ($\leq$4 characters) receive additional disambiguated queries ("{gene} protein" and the UniProt full name), reflecting the practical reality that short gene names collide with common abbreviations. If the title search returns fewer than ten results, the query broadens to title-or-abstract. bioRxiv and medRxiv preprints are added via Europe PMC (Europe PMC Consortium, 2015). All papers are ranked by peer-review status first (published over preprint) and then by NIH iCite citation count (Hutchins et al., 2016), so that the reading pass sees the most-cited, peer-reviewed evidence first.

Stage 1 (reading pass) feeds the ranked source papers to Claude Sonnet 4.6 (Anthropic, 2025). The prompt encodes a biologist's judgment about what constitutes mechanistic evidence: it extracts *only* findings where a direct experiment established something about how the protein works — substrates identified by co-immunoprecipitation or reconstitution, enzymatic activities measured by in-vitro assay or active-site mutagenesis, structures solved with functional validation, pathway positions established by genetic epistasis, post-translational modifications mapped to specific residues, and localization confirmed by imaging or fractionation. Phenotypic associations, expression correlations, and computational predictions without experimental validation are explicitly excluded. The model emits a structured evidence layer: each entry is a dated finding with experimental method, journal, two-axis confidence score (method tier $\times$ evidence preponderance), and supporting PMIDs. The prompt bans references to UniProt, OMIM, or any database. Genes returning zero findings from a non-empty corpus are re-run with a relaxed prompt admitting loss-of-function phenotypes.

Stage 2 (synthesis pass) feeds the evidence layer *alone* — not the raw abstracts — to Claude Opus 4.6. This constraint ensures the synthesis cannot draw on literature the reading pass did not admit, a design choice that mirrors how a careful reviewer would synthesize only from vetted sources. The model produces four outputs: (1) a declarative mechanistic narrative, hard-capped at 3–4 sentences to force synthesis over enumeration; (2) a per-entry mechanistic history recording what each study established; (3) per-entry open questions — the explicit gaps in knowledge at the time of each finding; and (4) a structured mechanism profile mapping the gene onto controlled vocabularies for molecular activity, localization, pathway, named complexes, and named partners. The prompt bans hedging language ("may," "might," "remains to be confirmed"); claims are stated as facts with citations. Prefetched reference data (UniProt, DepMap, OpenCell, HPA, HGNC, AlphaFold) is attached to each record for the viewer but is never shown to the LLM, preserving the no-paraphrase invariant.

2.2. Pass 2: structural QC and augmented re-annotation

Three deterministic rules, computable from the database alone with no LLM-derived signal, trigger Pass 2 re-annotation:

R1 (zero-discovery): Pass 1 returned zero discoveries from a non-empty corpus. **R2 (symmetric alias)**: at least one HGNC alias is itself a current canonical HGNC symbol, so the title search necessarily ingested a different gene's literature. **R3 (corpus-disjointness)**: the title corpus shares less than 5% of the gene's NCBI gene2pubmed PMID set, conditional on the curated set having ≥ 5 PMIDs.

R1$\cup$R2$\cup$R3 selected 6,049 of 19,291 annotated genes (31.4%). Each flagged gene is re-annotated under two changes. First, the corpus is augmented by unioning the title-search papers with NCBI gene2pubmed PMIDs retrieved via `elink` — a curated mapping linked to the Entrez gene ID, collision-free by construction. Second, both LLM stages run a Round-2 prompt: Stage 1 classifies each paper as keep or exclude before extraction, with explicit handling of alias-collision and alt-locus product papers; Stage 2 receives the canonical UniProt full name as a passive identity anchor with explicit refusal instructions if the discovery timeline does not match the canonical protein. The UniProt function paragraph itself is not shown, preserving Pass 1's no-paraphrase invariant.

3. Results

Affinage processed 19,291 of 19,296 HGNC protein-coding genes (99.97%) at a genome-amortized cost of $0.256 per gene ($4,934 total; Appendix A.2). The output database contains 274,637 mechanism findings (median 11 per gene), 178,710 mechanistic history entries with open questions, and 17,065 mechanism profiles (88.5% of annotated genes). Of all findings, 29% cite literature from 2020 or later. Of the annotated genes, 5,485 (28.4%) carry a mechanistic narrative of $\geq$500 characters where UniProt has under 200 (Fig-

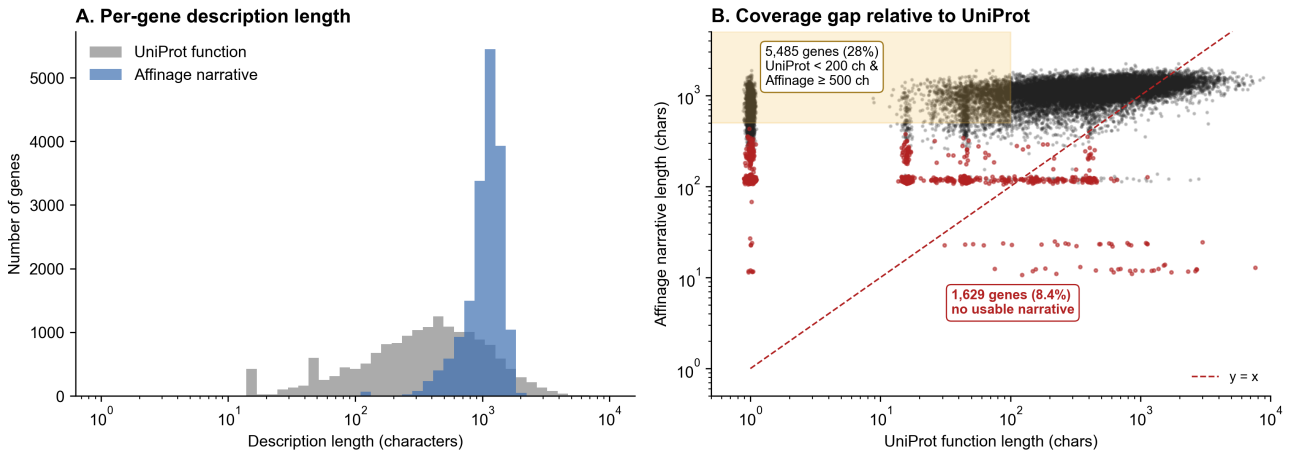

Figure 1. **Affinage extends per-gene description coverage relative to UniProt.** (A) Marginal distributions of UniProt curated function-field length and Affinage mechanistic narrative length, one observation per gene; both axes log-scaled. (B) Per-gene comparison; each point is one of 19,291 annotated genes, dashed line is $y = x$. The shaded tan region (5,485 genes, 28.4% of the genome) marks the coverage gap where UniProt has fewer than 200 characters of curated function and Affinage produces ≥500 characters of mechanistic narrative. The red floor band (1,629 genes, 8.4%) marks genes with no usable mechanistic narrative — 1,599 zero-discovery refusals plus 30 Stage 2 failure stubs.

ure 1). Pass 2 (Section 2.2) re-annotated 6,049 structurally flagged genes: 4,450 (73.6%) recovered a substantive narrative; 1,599 (26.4%) refused, predominantly where on-target literature was genuinely absent. The corpus-disjointness trigger fell from 3,756 to 163 post-Pass 2 (95.7% reduction), verifiable without any LLM judge.

4. Comparison with UniProt

A direct cross-tabulation of the 19,291 annotated genes against UniProt's curated function field (The UniProt Consortium, 2023) (Figure 2) bounds the resource's behavior against the field-standard reference. Joint coverage — UniProt has ≥200 characters of curated function and Affinage returns at least one discovery-anchored finding — accounts for 11,476 genes (59.5%). Pure Affinage wins, where UniProt has no function entry and Affinage produces a discovery-anchored narrative, account for 1,766 genes (9.2%). Affinage upgrades over a sparse UniProt entry (<200 characters) add another 4,420 genes (22.9%). Concordant orphans, where UniProt is sparse-or-empty and Affinage returns no usable narrative, account for 1,492 genes (7.7%). The cell that warrants direct examination is UniProt-rich (≥200 characters) and Affinage no usable narrative: **137 genes (0.71%)**. Of these 137, the majority arise from Pass 2 explicitly refusing on a structurally-flagged corpus rather than from an extraction failure. The single-pass paper would have reported 247 hard misses (1.3%); Pass 2 collapses this cell to 137 not by recovering the misses but by re-evaluating each one against gene2pubmed-curated literature and overwriting Pass 1 narratives that the augmented corpus did not support. Appendix B.1 gives the full breakdown.

5. Case studies

UniProt lists LENG8 as "Leukocyte receptor cluster member 8" with no function entry. All five of its mechanistic papers were published in 2025–2026, making this a gene whose biology postdates any curated database. Affinage resolves LENG8 as a conserved nuclear RNA quality-control factor that enforces retention and degradation of misprocessed mRNAs and noncoding RNAs. It assembles with PCID2 and SEM1 into the REX (TREX-2.1) complex, structurally analogous to the canonical TREX-2 export complex but acting as its dominant-negative antagonist: a conserved trigger loop in LENG8 releases the helicase DDX39B from mRNPs, diverting polyadenylated transcripts away from nuclear export and toward degradation by the PAXT–nuclear exosome pathway. Loss of LENG8 results in cytoplasmic leakage of intron-retained and aberrantly polyadenylated transcripts. The mechanistic history tracks how open questions were resolved across the five papers: the initial structural characterization left substrate specificity beyond GC-content enrichment undefined; subsequent work established the full retention-to-degradation pathway but left open the molecular basis for selective recognition of misprocessed versus correctly processed mRNAs and the structural interface between REX and the PAXT complex. LENG8 illustrates the value of Affinage for genes whose biology is too recent for any curated database to have captured.

Beyond LENG8, the annotation gap extends across diverse biology. KHNYN (16 discoveries, 14 from 2020+, UniProt empty) is resolved as a Mn^{2+}-dependent endoribonuclease and catalytic effector of the ZC3HAV1-mediated antiviral CpG-RNA decay pathway, restricting HIV-1, SARS-

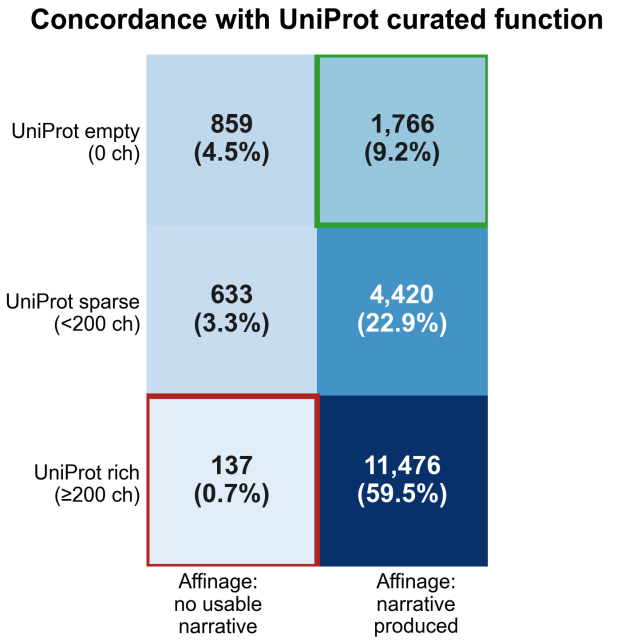

Figure 2. **Comparison with UniProt curated function.** Rows: UniProt function-field length buckets. Columns: Affinage outcomes. Green border: UniProt empty, Affinage produced a narrative (1,766 genes, 9.2%). Red border: UniProt rich but Affinage no usable narrative (137 genes, 0.71%), predominantly Pass 2 explicit refusals.

CoV-2, and influenza. EEPD1 (16 discoveries, UniProt empty) has a dual function: nuclear fork-protection via an Exo1–BLM complex and plasma-membrane PKA signaling; its loss activates cGAS–STING and sensitizes tumors to anti-PD1. CDK19 (21 discoveries, UniProt empty) is a Mediator-associated kinase that acts kinase-independently during IFN-γ responses; *de novo* missense variants cause epileptic encephalopathy. ANKRD22 (17 discoveries, 16 from 2020+, UniProt empty) integrates metabolic reprogramming with innate immune signaling via MAVS and NIK.

6. Limitations

The single largest limitation is that Affinage reads PubMed and Europe PMC abstracts, not full text. Mechanistic detail that lives only in the methods, results, or supplement of a paper is not directly visible to the reading pass, and the synthesis pass cannot produce what the reading pass did not extract. Adding a full-text retrieval path is the most impactful future direction.

Systematic evaluation at genome scale is an open challenge. Because the synthesis pass is instructed to produce 3–4-sentence narratives regardless of evidence depth, it is difficult to distinguish genes where the narrative genuinely reflects an augmented literature corpus from genes where the model is filling the requested format from thin evidence. The deterministic rules R1–R9 catch structural failures (Appendix B.2), but subtler errors — a correct-sounding narrative that omits a key mechanism, or one that over-weights a single paper — are not detectable without domain expertise. We believe the most productive path toward long-term quality assurance is expert feedback from the research community; adding infrastructure for per-gene commentary on the public viewer is a near-term goal.

To surface the structural failures we can catch, we extended the deterministic-rule layer with six output-side rules (R4–R9; Appendix B.2), evaluated on the database alone with no LLM invoked. The detector flags 199 narratives (1.03%) for revision: 17 where the narrative names a different gene, 134 refusals with substantive UniProt evidence, and 48 narratives dominated by a non-coding product of the locus. Flagged genes carry a revision banner on the public viewer.

7. Conclusion

Affinage demonstrates that a biologist-designed LLM pipeline can produce genome-scale, literature-grounded mechanistic annotation that serves both audiences: a biologist receives a ready-made, PMID-anchored account of any gene without running a retrieval session, and a computational group receives a structured, current annotation layer suitable for embedding or downstream reasoning — without needing domain expertise in mechanistic extraction. For 28% of the human genome, Affinage provides a substantive mechanistic narrative where UniProt has fewer than 200 characters. The resource — 19,291 narratives, 274,637 PMID-anchored findings, 178,710 open questions, and 82,684 literature-derived partner edges across 15,698 genes — is live and queryable through a public API and MCP server. As LLM-assisted reasoning becomes routine in biological research, the quality and currency of per-gene annotations becomes infrastructure. Affinage is designed as that infrastructure: a stable, citation-anchored mechanistic layer that downstream agentic systems — for target identification, screen interpretation, or hypothesis generation — can build on without re-deriving mechanism from the literature.

Data and code availability

All gene records, a REST API, and an MCP server are available at `https://affinage.wi.mit.edu`. Source code: `https://github.com/cheeseman-lab/affinage`. Both passes executed May 2026 (Sonnet 4.6, Opus 4.6) and are reproducible from the released code.

References

Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023. doi: 10.1093/genetics/iyad031.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, 47(D1):D1038–D1043, 2019. doi: 10.1093/nar/gky1151.

Anthropic. Claude Sonnet 4.6 and Opus 4.6 model card. Technical report, https://www.anthropic.com/, 2025.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Chen, Y. and Zou, J. GenePert: Leveraging GenePT embeddings for gene perturbation prediction. *bioRxiv*, 2024. doi: 10.1101/2024.10.27.620513.

Chen, Y. T. and Zou, J. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01293-z.

Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Research*, 43(D1):D1042–D1048, 2015. doi: 10.1093/nar/gku1061.

Hutchins, B. I., Yuan, X., Anderson, J. M., and Santangelo, G. M. Relative Citation Ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLOS Biology*, 14(9):e1002541, 2016. doi: 10.1371/journal.pbio.1002541.

Istrate, A.-M., Li, D., and Karaletsos, T. scGenePT: Is language all you need for modeling single-cell perturbations? *bioRxiv*, 2024. doi: 10.1101/2024.10.23.619972.

Li, P.-H., Sun, Y.-Y., Juan, H.-F., Chen, C.-Y., Tsai, H.-K., and Huang, J.-H. A large language model framework for literature-based disease–gene association prediction. *Briefings in Bioinformatics*, 26(1):bbaf070, 2025. doi: 10.1093/bib/bbaf070.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50 (D1):D20–D26, 2022. doi: 10.1093/nar/gkab1112.

Seal, R. L., Braschi, B., Gray, K., Jones, T. E. M., Tweedie, S., Haim-Vilmovsky, L., and Bruford, E. A. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*, 51(D1):D1003–D1009, 2023. doi: 10.1093/nar/gkac1062.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023. doi: 10.1093/nar/gkac1052.

A. Extended methods

A.1. Pipeline mechanics

For genome-scale execution we chunk the HGNC universe (Seal et al., 2023) into batches of $\leq 2{,}000$ genes; Anthropic's batch payload limit at $\sim$256 MB constrains chunk size at the observed mean of $\sim$80 KB/gene of formatted prompt. Chunks run sequentially through prefetch $\rightarrow$ Stage 1 batch $\rightarrow$ sync retry $\rightarrow$ Stage 2 batch $\rightarrow$ SQLite upsert. Prefetched corpora are disk-cached so partial reruns do not re-hit PubMed (Sayers et al., 2022). Failures (NCBI 5xx, Stage 2 parse error) are logged per gene and the run continues. Costs reported here are at Anthropic's batch rate (50% of on-demand; Anthropic 2025).

Five genes failed in the 2026-04-28/29 Pass 1 run: three protocadherin cluster sentinels (`PCDHA@`, `PCDHB@`, `PCDHG@`) violating Anthropic's `custom_id` regex, and two transient NCBI 5xx prefetch errors. Pass 2 uses a smaller chunk size of 1,500 genes because the augmented title-$\cup$-gene2pubmed corpus is denser than title alone; both passes share the same prefetch cache so per-gene external-API calls are not duplicated.

A.2. Operating cost and incremental maintenance

The 2026-04-29 Pass 1 run used Claude Sonnet 4.6 in Stage 1 (442 M input plus 64 M output tokens, \$1,146) and Claude Opus 4.6 in Stage 2 (123 M input plus 48 M output tokens, \$2,741) at Anthropic batch rates (50% of on-demand; Anthropic 2025). The synthesis stage dominates spend ($\sim$70%); per-gene cost scales with corpus richness, with the p95 Pass-1 gene at \$0.41. The 2026-05-06 Pass 2 added \$1,047 across 6,049 flagged genes (\$0.173 per flagged gene), recovering 4,450 narratives at \$0.235 per recovered narrative. The output-side concordance detector (R4–R9, Appendix B.2) is regex-only and adds no API spend; it runs in $\sim$30 seconds against the read-only database. Total LLM spend: \$4,934 (\$0.256/gene amortized).

For ongoing maintenance, the pipeline supports an incremental mode. On a rerun, a per-gene diff reads the prior record's PMID set, re-fetches the current PubMed corpus, and decides per gene whether to re-run: if no new PMIDs have appeared, the prior annotation is copied forward at zero LLM cost; if new PMIDs are present, the gene re-enters Stage 1 and Stage 2 with the full updated corpus.

B. Evaluation framework

B.1. Pass 2: structural QC layer and residual hard misses

Motivating cases. Pass 1 hard misses were dominated by an upstream-of-LLM failure: gene-symbol collision in the PubMed title search. CASP4 carries the alias `TX` and its 100-paper corpus filled with "TX-TL *E. coli* toolbox" papers; CDPF1 carries the alias `p21`/CDKN1A and its Pass 1 narrative described p21's cell-cycle inhibitor role rather than CDPF1's centriole-distal-appendage function. In each case Stage 1 either returned zero discoveries (R1) or dutifully extracted whatever mechanism the contaminated corpus contained (R2/R3 hallucination risk). Stage 1 was disciplined; Stage 0 had handed it the wrong corpus.

Outcomes by rule. Of 6,049 flagged genes, 4,450 recovered a clean narrative and 1,599 refused; outcomes by triggering rule are summarized in Table 1.

Table 1. **Pass 2 outcomes by triggering structural rule.**

Rule(s) fired	Flagged	Recovered	Refused
R3 only	3,608	3,541 (98.1%)	67 (1.9%)
R1 only	1,921	403 (21.0%)	1,518 (79.0%)
R2 only	353	351 (99.4%)	2 (0.6%)
R2,R3	148	142 (95.9%)	6 (4.1%)
R1,R2	19	13 (68.4%)	6 (31.6%)
Total	**6,049**	**4,450 (73.6%)**	**1,599 (26.4%)**

Residual misses. The post-Pass-2 hard-miss cell (UniProt-rich, Affinage no usable narrative) is 137 genes (0.71%). Hand inspection assigns most to: (i) genes whose UniProt entry cites mechanism literature available only in full-text PMC, (ii) genes whose UniProt evidence is flagged *By similarity* without primary mechanism literature, and (iii) genes for which Pass 2 correctly refused on an off-target corpus.

B.2. Output-side concordance rules (R4–R9)

The structural rules R1–R3 (Section 2.2) are evaluated upstream of the LLM and gate Pass 2. We complement them with a six-rule output-side layer that scans the mechanistic narrative and flags concordance failures the two-pass pipeline did not eliminate. Like R1–R3, every rule is regex- and SQL-based; no LLM is invoked, and the flag assignment is reproducible from the database plus the gene2pubmed cache.

R4 (alias-collision domination): the narrative is dominated by a different gene whose symbol overlaps an HGNC alias. **R5 (species-led opener)**: the first sentence is led by a non-human-organism qualifier. **R6 (paralog opener)**: the first noun phrase names a different gene symbol. **R7 (alt-product narrative)**: the opening text is dominated by a non-coding product of the locus. **R8 (rich-UniProt refusal)**: the narrative is a refusal placeholder and UniProt has $\geq$ 200 characters. **R9 (narrative-corpus disjointness)**: the narrative's own citations share $\leq 5\%$ overlap with gene2pubmed.

Table 2. **Output-side concordance flags across 19,291 genes.**

Tier	Rule(s)	Action	Genes
1	R4, R6, R9	rewrite (wrong gene)	17
2	R8	rerun (refusal w/ rich UniProt)	134
3	R7	flag (alt-product narrative)	48
5	R5 alone	cross-species (defensible)	0
Total flagged			**199 (1.03%)**

B.3. Quantified failure modes

Table 3. **Quantified failure modes after Pass 2.** The hard-miss cell (UniProt-rich, Affinage no usable narrative) falls from 247 in Pass 1 alone to 137 after Pass 2.

Failure mode	Genes (%)	Cause
Pass 2 recovered	4,450 (23.1%)	R1$\cup$R2$\cup$R3-flagged genes whose augmented corpus + identity-anchored prompt produced a clean narrative.
Pass 2 explicit refusal	1,599 (8.3%)	Augmented corpus also lacked on-target literature; Pass 2 overwrote with a refusal.
Output-side flags (R4–R9)	199 (1.03%)	17 tier-1, 134 tier-2, 48 tier-3. Reproducible from database alone.
Hard miss vs. UniProt-rich	137 (0.71%)	Majority are Pass 2 explicit refusals rather than extraction failures.
Schema-validation failure	59 (0.31%)	Opus emitted JSON that failed validation; timeline saved, narrative replaced with placeholder.
Symmetric-alias hallucination	11 $\rightarrow$ 2 (0.01%)	82% reduction on a deterministic marker.

C. The resource

C.1. Output layers

Each gene record exposes four layers. The `mechanistic_narrative` is the declarative synthesis described in the body — the closest analog to a UniProt function field, but with per-clause PMIDs and recent biology integrated. The `teleology` block is a per-entry mechanistic history recording what each finding established, together with explicit `gaps` (open questions at the time of each finding). The `mechanism_profile` places each gene on controlled vocabularies: 17,065 of 19,291 genes (88.5%) carry at least one structured term across the molecular-activity, localization, or pathway axes; the median gene has six named partners and at least one localization term.