BIORELAY: MULTI-AGENT ORCHESTRATION WITH EVIDENCE CONTINUITY FOR BIOMEDICAL KNOWLEDGE DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical deep-research agents retrieve evidence from literature and databases, reason over it, and synthesize conclusions into static reports. Yet whether the evidence needed to support those conclusions survives the intermediate stages from retrieval to final output remains largely unmeasured. We formalize this property as evidence continuity and introduce ProteInsight, a benchmark of 2,796 tasks over 699 disease–protein cases with typed evidence assembled from biomedical databases and literature. We find that evidence omission, rather than unsupported generation, is the dominant failure mode: GPT-4o produces grounded answers in 98.7% of cases but with omission accounting for 90.7% of failures and worsening as tasks require more heterogeneous evidence. To preserve evidence across stages, we develop BioRELAY, a multi-agent system that maintains a shared typed evidence state across search, reasoning, and synthesis, with deterministic gates enforcing evidence linkage and state consistency, supporting narrative report and interactive interface. BioRELAY performs strongly on standardized biomedical QA, challenging protein-function reasoning, and end-to-end evidence synthesis. Ablations further show that evidence continuity is distinct from schema validity and model capability, establishing evidence continuity as a measurable failure for multi-agent system and providing a mechanism to localize where evidence is lost.

1 INTRODUCTION

Biomedical researchers increasingly use AI-generated analyses and reports to support research interpretation and decision-making (Wang et al., 2025; Zhang et al., 2024). Existing web-search LLMs and deep research agents can retrieve evidence, browse the literature, and generate citation-grounded answers (Jin et al., 2025; Li et al., 2025; Liu et al., 2025b; Shao et al., 2026), and biomedical agents such as Biomni further execute complex biomedical tasks (Huang et al., 2025). However, their outputs are typically delivered as static textual reports. In disease biology, researchers need to interrogate how a conclusion was produced, which evidence supports it, where competing interpretations remain plausible, and how claims link back to the source (Gao et al., 2024; Gierend et al., 2024). In a static report, evidence established upstream may be silently omitted or disconnected from its provenance as it passes through successive agent stages. Interactive interfaces can help users navigate complex evidence through structured overviews and visual representations (Leviathan et al., 2026; Dai et al., 2026; Martin-Boyle et al., 2026), but they do not by themselves ensure that the displayed claims remain accurate and supported. Biomedical decision-making therefore requires an evidence representation that supports both high-level overview and detailed inspection without obscuring functions or overwhelming users (Wong et al., 2025; Leviathan et al., 2026).

We argue that scientific evidence should persist across the workflow as a **shared evidence state**, rather than being repeatedly reconstructed at each reasoning stage or for each output representation. Inspired by the Biolink Model (Unni et al., 2022), this state records normalized biomedical entities, typed and qualified associations, accepted supporting evidence, synthesized claims, uncertainty, and provenance links. Both narrative reports and interactive views can therefore be derived from the same shared evidence state.

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

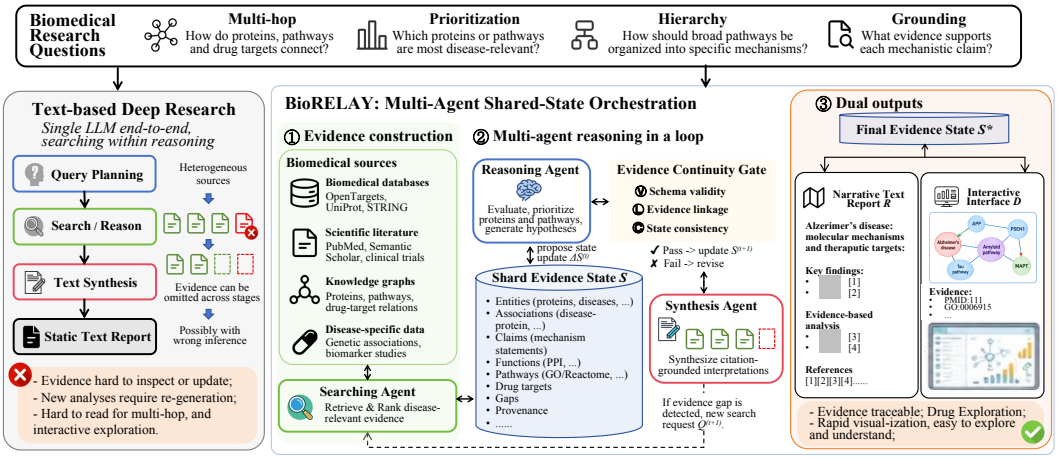


Figure 1: Overview of BioRELAY. The Search and Reasoning Agents iteratively construct and refine a versioned shared evidence state under orchestration and budget constraints. The Synthesis Agent writes the narrative report as the state is built; once the state is finalized, the report is assembled and the interactive evidence interface is projected from it, both deterministically.

We propose **BioRELAY**, an automated multi-agent framework built around this shared evidence state to synthesize heterogeneous disease-associated protein evidence into multiple evidence-grounded representations. Within its Search–Reasoning loop, an independent Search Agent retrieves and ranks candidate literature and database evidence without constructing biomedical claims, while a Reasoning Agent evaluates the retrieved evidence and is the only component authorized to update the schema-constrained shared evidence state. The narrative report is written and gate-checked section by section as the state is built; once the state is finalized, the report is assembled and an interactive evidence interface is projected from it, both deterministically. A deterministic Evidence-Continuity Gate checks schema validity, evidence linkage, and state consistency at every stage, routing evidence gaps back to retrieval and projection failures to the corresponding projection stage. The interface is thus guaranteed to resolve to the state, whereas for the report the gate checks citations and claims.

Interpreting disease-associated proteins often requires integrating evidence from protein–disease associations, pathway annotations, publications, phenotypes, and target–drug relationships. These sources differ in structure and provenance and are often linked through multi-hop relations, making it difficult to preserve the supporting evidence as information moves across stages. We call this property **evidence continuity**, where decision-relevant evidence should remain valid, traceable, and available from retrieval through reasoning and synthesis to downstream representations. We therefore construct **ProteInsight**, a challenging benchmark for disease-associated protein interpretation, to measure how often correct biomedical conclusions remain fully supported by evidence end to end. Different from established genomics and biomedical-agent benchmarks, such as GeneTuring (Shang et al., 2025) and the Biomni (Huang et al., 2025), which score whether a final answer is correct, ProteInsight additionally pairs every open-ended, multi-hop case with a required-evidence set assembled without language models, so that it can test whether the evidence needed for a correct conclusion survives every stage of a multi-agent pipeline. We evaluate BioRELAY at four levels: overall biomedical question-answering, iterative evidence construction, shared-state governance, and cross-stage evidence continuity, with ablation experiments. To summarize, our contributions are threefold:

- We formulate **evidence continuity** as a measurable property of multi-stage biomedical reasoning: whether the evidence required to support a scientific conclusion remains grounded and traceable as it passes through retrieval, reasoning, and downstream generation.
- We propose **BioRELAY**, a harness-centered multi-agent framework in which an independent Search Agent and a Reasoning Agent build a shared evidence state yielding a narrative report and an interactive evidence interface, while an Evidence-Continuity Gate checks every stage.
- We introduce **ProteInsight**, a disease-centered benchmark for protein interpretation in which each case pairs a reference conclusion with a required-evidence set constructed without language model, enabling evaluation of multi-hop reasoning, evidence grounding, and evidence continuity.

2 Related Work

2.1 Search and Deep Research Agents

Agentic LLM systems make search, tool use, and reasoning part of the generation process (Yao et al., 2023; Wu et al., 2024). Recent search agents further push this boundary to multi-step web search and long-form research writing. Search-R1 and ASearcher train models to interleave reasoning with search (Jin et al., 2025; Gao et al., 2025), while WebThinker (Li et al., 2025), WebExplorer (Liu et al., 2025b), Tongyi DeepResearch (Tongyi DeepResearch Team, 2025) focus on text reports over retrieved evidence. In parallel, deep research agents have expanded into scientific and biomedical domains (Shao et al., 2026; Huang et al., 2025). However, reliability is typically assessed through retrieval quality or final-output correctness, providing limited visibility into what happens to evidence between these endpoints. Our experiments show that, in this setting, evidence loss can be a more prominent failure mode than unsupported generation: valid evidence may be acquired but subsequently omitted or detached from the claims it supports. BioRELAY targets this intermediate failure mode by maintaining a shared typed evidence state and auditing evidence continuity across agent transitions and downstream representations.

2.2 Biomedical Agents and Structured Evidence Systems

Biomedical research builds on structured knowledge resources and knowledge graphs such as Open Targets (Koscielny et al., 2017), Monarch (Mungall et al., 2017), Biolink (Unni et al., 2022). However, they traditionally depend on pre-built pipelines or manual curation. To improve flexibility, recent work has automated the construction of biomedical relations from literature (Arazkhani et al., 2026; Zheng et al., 2026). In parallel, domain-specific agents, such as Biomni (Huang et al., 2025), BioRAGent (Bi et al., 2025), and BioMedSearch (Liu et al., 2025a), coordinate biological tools and literature search for complex queries, while BioResearcher (Luo et al., 2025) and BioMedAgent (Bu et al., 2026) extend agentic workflows toward end-to-end research and tool-aware biomedical analysis. Multi-agent systems have also been explored for biomedical hypothesis generation (Qi et al., 2024; Xu et al., 2025) and medical reasoning through specialized clinical roles (Ma et al., 2026; Miao et al., 2025; Nourmohammadi et al., 2026). BioRELAY addresses a complementary question: whether evidence acquired by a multi-stage workflow remains linked, traceable, and available as it passes between agents and into downstream representations. To this end, it maintains a shared typed evidence state that is reused to generate both narrative reports and interactive interfaces, while explicitly auditing cross-stage evidence continuity.

3 Method

We present **BioRELAY**, a multi-agent framework for evidence-centered interpretation of disease-associated proteins. It operationalizes evidence continuity through shared-state reasoning and shared-state projection.

3.1 Task Formulation

Existing biomedical agent evaluation primarily asks whether a final interpretation is scientifically correct or supported by retrieved evidence. However, a multi-stage agent can produce no new unsupported claim while still losing required evidence during retrieval, reasoning, synthesis, or representation. We therefore treat evidence continuity as a distinct requirement, that evidence that has been acquired and accepted should remain resolvable throughout subsequent transformations.

Prior network medicine research motivates interpreting disease-associated proteins through disease modules and network neighborhoods, with higher interactome coverage being preferable (Menche et al., 2015). We formulate **disease-associated protein interpretation** as constructing a supported evidence chain that progressively links protein–disease associations, pathway and network contexts, and disease-specific literature evidence.

Given a disease d and a set of disease-associated proteins $\mathcal{P} = \{p_i\}_{i=1}^{n}$ with observed association signals $\mathbf{z} = \{z_i\}_{i=1}^{n}$, where z_i is p_i's association with d (a UK Biobank hazard ratio, 95% CI, P-value), the task is to contextualize these signals, synthesize the relevant information, and construct

a supported interpretation. The required evidence is heterogeneous but interdependent (Løvsletten et al., 2024), and the final interpretation can require hierarchical exploration (Santos et al., 2022) and multi-hop connections (Menche et al., 2015). Accordingly, we define a **shared evidence state** $\mathcal{S}$ based on the Biolink Model (Unni et al., 2022). The state contains normalized biomedical entities, typed associations, supporting evidence, synthesized claims, protein function records, interaction records, and provenance links. The narrative text report R is written as the state is built and assembled once it is finalized as $\mathcal{S}^\star$; the interactive evidence interface D is a deterministic projection of $\mathcal{S}^\star$.

Transforming biomedical evidence into multiple projections through several agent stages may detach claims from their supporting sources or drop key evidence (Simhi et al., 2024). We therefore define **evidence continuity** as the requirement that decision-relevant evidence remain valid, traceable, and available as it moves through intermediate states to the final representation. Let O_k denote the output produced at stage k, and let $\mathcal{X}_k^{(t)}$ denote its validation context at iteration t, comprising the stage input contract and the applicable shared evidence state. Evidence continuity requires

$$\Gamma\left(O_k, \mathcal{X}_k^{(t)}\right) = 1, \qquad k \in \mathcal{K}, \tag{1}$$

where $\mathcal{K}$ comprises the Search, Reasoning, Synthesis, and Visualization stages, and Γ is implemented by the Evidence-Continuity Gate. For D, the validation context is the finalized state $\mathcal{S}^\star$; for each section of R, it is the state held when the section is written. The gate reads identifiers, so Eq. 1 holds by construction for D but only for the citations and claim anchors of R.

3.2 Multi-Agent Orchestration

3.2.1 Shared-State Orchestration

Building on the disease-centered evidence chain depicted in Figure 1, BioRELAY organizes evidence retrieval and reasoning around three request types: candidate pathways, interaction modules, and target/drug evidence. These requests guide iterative literature retrieval and shared-state refinement while important evidence gaps or unresolved contradictions remain. An Orchestrator controls the agent execution order, maintains the current state version, and routes requests and validation failures without producing biomedical claims.

Agents interact through typed intermediate artifacts with distinct permissions. During evidence construction, the Reasoning Agent is the only agent authorized to propose biomedical updates to $\mathcal{S}$. The Search Agent receives structured evidence requests $\mathcal{Q}^{(t)}$ and returns ranked evidence packets $\mathcal{B}^{(t)}$. The Reasoning Agent evaluates these packets in the context of $\mathcal{S}^{(t)}$ and produces a schema-constrained update $\Delta\mathcal{S}^{(t)}$, which may add or revise normalized entities, typed associations, accepted evidence, qualified claims, and provenance links; the state then advances to $\mathcal{S}^{(t+1)}$. The Synthesis Agent writes R section by section as $\mathcal{S}$ is built, and the Visualization Agent projects the finalized $\mathcal{S}^\star$ into D (Section 3.2.3). The Evidence-Continuity Gate validates the output of every stage, including R and D, and returns validation feedback without modifying the source state.

3.2.2 Iterative Search and Reasoning

BioRELAY separates evidence retrieval from biomedical reasoning. The Reasoning Agent plans evidence acquisition and constructs the semantic content of the shared state, whereas the Search Agent independently retrieves and ranks candidate publications.

Evidence planning. To guide the downstream literature review, the Reasoning Agent first formulates targeted hypotheses about underlying biological mechanisms from an initial state $\mathcal{S}^{(0)}$. Disease-associated proteins are mapped to enriched biological terms via g:Profiler (Kolberg et al., 2023), and near-duplicate terms are merged using BioBERT (Lee et al., 2020) embedding similarity, yielding a ranked set of candidate pathways $\mathcal{T}^{(0)}$. Pathways are prioritized by both enrichment support from the input protein set and disease-specific literature evidence. The Reasoning Agent converts $\mathcal{T}^{(0)}$ into the initial requests $\mathcal{Q}^{(0)}$, constraining subsequent search to explicit disease–pathway and protein–pathway hypotheses rather than unconstrained literature queries. Ranking and deduplication details are provided in Appendix D.3.

Search and evidence acquisition. At iteration t, for each structured request $q \in \mathcal{Q}^{(t)}$, the Search Agent queries academic databases (e.g., PubMed and Semantic Scholar) to obtain a candidate set of publications $\mathcal{A}_q$. To prioritize high-quality literature, each publication $a \in \mathcal{A}_q$ is scored by

$$f(q,a) = \min\Big(\Phi\big(s_{\mathrm{kw}}(q,a),\, s_{\mathrm{sem}}(q,a),\, s_{\mathrm{cit}}(a)\big) \cdot w_J(a),\, f_{\max}\Big), \tag{2}$$

where $s_{\mathrm{kw}}(q,a)$ is the lexical disease–pathway keyword overlap, $s_{\mathrm{sem}}(q,a)$ is the BioBERT semantic similarity, $s_{\mathrm{cit}}(a)$ is a log-scaled citation count, and Φ is a weighted linear combination (weights $0.45/0.35/0.20$). The raw score is modulated by a venue weight $w_J(a) \in [0.95, 2.0]$ and capped at $f_{\max} = 2.2$. Citation count and venue do not measure evidential quality; unlike the citation term, the venue factor changes which pathways are selected. Publications scoring below an empirically set threshold $\tau = 0.25$ are discarded; the remainder are sorted by $f(q,a)$ and truncated to a per-packet budget b (we use $b = 20$), forming an ordered evidence packet $B_q^{(t)} = (a_1, \ldots, a_{m_q})$ with $m_q \le b$ and $f(q,a_1) \ge \cdots \ge f(q,a_{m_q}) \ge \tau$. The retrieval output of iteration t is the set of these packets:

$$\mathcal{B}^{(t)} = \big\{B_q^{(t)} \mid q \in \mathcal{Q}^{(t)}\big\}. \tag{3}$$

Each retained item records the candidate publication, its retrieval metadata, and its relevance score. Crucially, $f(q,a)$ establishes retrieval priority but does not verify whether an article supports a biomedical hypothesis. The Search Agent therefore forwards $\mathcal{B}^{(t)}$ to the Reasoning Agent and does not write these publications into the shared evidence state $\mathcal{S}$.

Reasoning and evidence state update. At iteration t, the Reasoning Agent evaluates the evidence packets $\mathcal{B}^{(t)}$ against the current state $\mathcal{S}^{(t)}$ and proposes an update $\Delta\mathcal{S}^{(t)}$:

$$\mathcal{S}^{(t+1)} = \mathcal{S}^{(t)} \oplus \Delta\mathcal{S}^{(t)}, \tag{4}$$

where $\oplus$ denotes a gate-checked, revision-tracked merge. Each accepted update preserves not only the biomedical assertion but also its typed relation, evidential support, provenance identifier, and qualification status. Thus, state refinement is additive or explicitly revision-tracked rather than an unconstrained textual rewrite.

If unresolved gaps require further investigation, the Reasoning Agent encodes them as targeted requests $\mathcal{Q}^{(t+1)}$ for the next Search–Reasoning iteration. The loop terminates when no further critical request is produced or when the iteration budget $T_{\max}$ is exhausted ($T_{\max} = 2$). The finalized state $\mathcal{S}^\star$ retains accepted evidence, explicit qualifications, conflicting findings, and any unresolved gaps.

3.2.3 Deriving the Report and the Interface

Both representations draw only on the shared state. The Synthesis Agent composes each pathway section of R once its reasoning artifacts are admitted to $\mathcal{S}$, adds follow-up sections as gaps are resolved, and writes the introduction last. It may not introduce new facts and uses only the accepted and qualified claims: it consolidates redundant mechanistic accounts, distinguishes primary findings from supporting or exploratory ones, and organizes the retained claims into coherent disease–pathway–protein mechanism narratives. After the loop, R is assembled from its gate-checked sections without model calls and D is projected deterministically from $\mathcal{S}^\star$.

The two representations emphasize distinct dimensions of the same evidence base. The narrative report delivers a sequential mechanistic synthesis, whereas the visual projection exposes topological relations, preserves explicit claim–evidence traceability, and enables hierarchical inspection at the protein, pathway, publication, and claim levels. Evidence continuity (Eq. 1) is required along the entire chain of transformations

$$\mathcal{S}^{(0)} \to \mathcal{S}^{(1)} \to \cdots \to \mathcal{S}^\star \to \{R, D\}. \tag{5}$$

3.2.4 Evidence-Continuity Gate

To enforce the evidence-continuity requirement in Eq. 1, the Orchestrator applies a deterministic, stage-aware gate at every agent stage. For stage k, the gate evaluates the candidate output O_k against its validation context $\mathcal{X}_k^{(t)}$:

$$G_k\left(O_k, \mathcal{X}_k^{(t)}\right) = \mathbb{I}\left[V_k \wedge L_k \wedge C_k\right], \tag{6}$$

where V_k, L_k, and C_k implement three complementary checks:

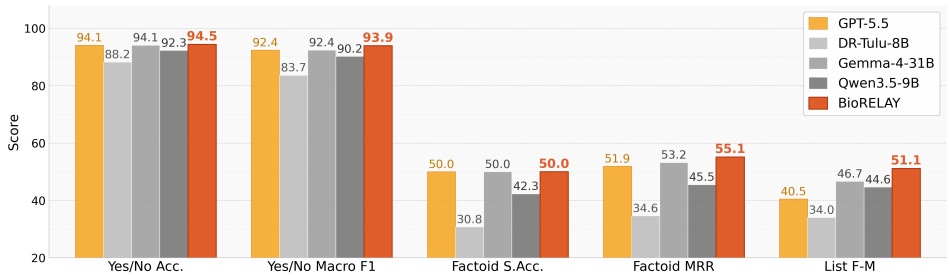


Figure 2: Results on the BioASQ Phase B exact-answer task. All five metrics are higher-is-better; BioRELAY achieves the best or tied-best score on all five.

- **Schema Validity** (V_k) checks required fields, unique identifiers, field types, allowed values, numeric bounds, and valid graph endpoints.
- **Evidence Linkage** (L_k) checks that every entity, claim, support link, and publication identifier in O_k resolves against the shared state, and that required state items are not omitted.
- **State Consistency** (C_k) checks that the qualification status, uncertainty, and relation types of the input state are preserved in the output artifact.

Together, L_k operationalizes evidence support, provenance preservation, and critical-evidence coverage, while C_k preserves qualification and uncertainty fields under projection. An artifact satisfies the evidence-continuity requirement only when $G_k = 1$. A failed artifact is rejected and returned to its producing agent for revision. Evidence insufficiency identified during Reasoning initiates another Search–Reasoning iteration, whereas a continuity failure in R or D triggers only local revision by the Synthesis or Visualization Agent, with the state held fixed. BioRELAY uses GPT-5.4 for all agents.

4 Experiments and Results

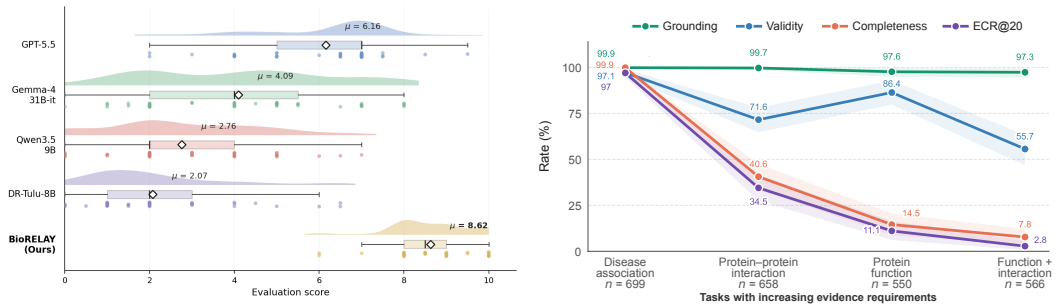
4.1 Experimental Setup

Baselines. We compare BioRELAY with five baselines chosen to cover complementary system settings. GPT-5.5 + Web Search provides a strong frontier search-enabled baseline, while Qwen3.5-9B + Search tests an open-weight model of similar scale to DR-Tulu-8B under the same search objective. Gemini Deep Research represents a commercial deep-research system, and DR-Tulu-8B an open model trained for long-form, citation-grounded research (Shao et al., 2026). Finally, BioRELAY (w/o Search) removes the Search Agent to test the contribution of explicit evidence retrieval. Search-enabled systems receive the same evidence-seeking objective and comparable retrieval budgets where applicable. QA benchmarks also include Gemma-4-31B-it. Settings and baselines are detailed in Appendix E, prompt templates in Appendix J, and rubrics in Appendix B.4.

4.2 Biomedical Evidence Reasoning

Standard biomedical QA. To test whether multi-agent orchestration and structured evidence handling preserve standard biomedical question-answering ability, we first evaluate BioRELAY on BioASQ Phase B (Tsatsaronis et al., 2015), an established benchmark covering yes/no, factoid, and list questions. BioRELAY achieves the best or tied-best performance across all five metrics (Figure 2), with the largest gains on questions requiring the ranking or extraction of multiple biomedical entities. Because the systems also differ in backbone and retrieval, this shows that orchestration preserves standard QA ability, not which component produces the gains.

ProteInsight: multi-hop, evidence-annotated protein interpretation. BioASQ Phase B tests concise answers from supplied documents, whereas disease-centered protein interpretation also requires wide search (Wong et al., 2025) and multi-hop evidence integration. We therefore construct ProteInsight, a benchmark of 699 disease–protein cases spanning 10 diseases. The benchmark consists of open-ended, multi-hop questions that require systems to retrieve and integrate functional annotations, protein–protein interactions, pathway context, disease mechanisms, and evidence for protein ranking. It has four task types of increasing difficulty: disease association, protein–protein interaction, protein



(a) Final-answer score distributions (0–10) on the ProteInsight protein-function questions.

(b) Evidence continuity (ECR@20) degrades on tasks requiring heterogeneous evidence integration.

Figure 3: Protein-function reasoning and evidence continuity on ProteInsight.

function, and function + interaction. Each task pairs a gold conclusion $y^\star$ with a required-evidence set $E^\star$, constructed without language models from UniProt, GO, Reactome, STRING, and PubMed. This design lets us evaluate answer correctness and, separately, how much of the evidence required to support that answer survives multi-stage reasoning (see details in Appendix C).

BioRELAY improves biomedical reasoning. We evaluate five systems on the ProteInsight protein-function questions and use GPT-5.5 as a judge to score each of the 48 answers from 0 to 10 (Figure 3a). BioRELAY obtains the highest mean score (8.62), with scores concentrated in the high range, suggesting more stable protein-centered reasoning and evidence ranking.

4.3 Evidence Continuity

Metrics. A correct, well-cited answer can still omit required evidence, so for each task u we score three properties. (1) **Validity** $\mathsf{Val}(u)$ holds if an LLM judge (GPT-5.5) without access to $E^\star$ finds the predicted conclusion consistent with $y^\star$. (2) **Grounding** $\mathsf{Gnd}(u)$ holds if the answer cites at least one record and every citation resolves to a record in its budget-b packet. (3) **Completeness** $\mathsf{Cmp}(u)$ holds if the answer is grounded and cites every required record in $E^\star$. Both check identifiers, not whether a record entails the claim citing it, so ECR is necessary but not sufficient for evidence support (App. C.3). Domain experts checked the validity labels, the required-evidence matching and claim-level entailment. The **Evidence Continuity Rate** is the share of tasks that satisfy all three:

$$\mathrm{ECR@}b = \frac{1}{|\mathcal{U}_b|} \sum_{u \in \mathcal{U}_b} \mathbf{1}\{\mathsf{Val}(u) \wedge \mathsf{Gnd}(u) \wedge \mathsf{Cmp}(u)\}\,, \tag{7}$$

where $\mathcal{U}_b$ contains the tasks that have required evidence and can be completed within budget b.

Evidence continuity degrades with evidence integration. Figure 3b scores GPT-4o on the $|\mathcal{U}_{20}| = 2{,}473$ ProteInsight tasks, answering from a fixed $b = 20$ packet of the case's own records that holds all of $E^\star$, a favorable condition by design; this is a diagnostic of one model, not a comparison of systems (App. C.3). Although 98.7% of answers were grounded, only 44.0% were complete and ECR@20 was 39.7%, and this gap widens systematically with task complexity. Correctness did not imply continuity: of the 1,940 answers judged valid, at least 927 (47.8%) cited only resolvable records yet omitted a required one, and only 107 of the 1,491 failures were complete (App. C.2). Among valid interaction answers, where only 2 of 658 tasks list interchangeable records, at least 51.4% omitted one; in function tasks Cmp also requires records the gold marks as interchangeable (App. C.3). Even with every required record in hand, preserving heterogeneous evidence, not resolving citations, is thus a primary bottleneck.

4.4 End-to-End Biomedical Discovery Report Generation

Given a disease and its associated proteins, each system returns a narrative report R and an interactive evidence interface D. Following guidance for evidence synthesis and generative scientific evaluation (Page et al., 2021; Clark et al., 2025; Flemyng et al., 2025), we evaluate five diseases spanning distinct biomedical decision regimes: Alzheimer's disease (AD), depression (MDD), atrial fibrillation

Table 1: Evidence resolution and coverage of the narrative report (R) and interactive evidence interface (D). Unsupported rate (U) is the proportion of rendered entities, protein–protein relations, or citations that cannot be traced to the acquired evidence; coverage (Cov) is the share of acquired proteins, pathways, or publications that reaches the representation. n_{cit} is the mean number of distinct rendered PMIDs. Values are means over five diseases; 95% bootstrap intervals are reported in Appendix Table 13.

		V alidity (%)				Coverage (%)		
System	**Rep.**	$U_{\text{ent}} \downarrow$	$U_{\text{rel}} \downarrow$	$U_{\text{cit}} \downarrow$	n_{cit}	$\text{Cov}_{\text{prot}} \uparrow$	$\text{Cov}_{\text{path}} \uparrow$	$\text{Cov}_{\text{pap}} \uparrow$
GPT-5.5 (raw)	R	35.29	75.50	0.91	37.0	32.45	7.77	3.55
	D	42.00	70.00	0.96	28.2	22.42	5.98	2.58
GPT-5.5 (typed)	R	0.29	0.36	1.05	74.4	51.17	89.55	5.48
	D	9.08	1.33	0.75	28.6	32.83	46.57	2.28
BioRELAY	R	8.60	19.27	**0.00**	292.4	31.76	**100.00**	**21.95**
	D	**0.00**	**0.00**	0.00	94.4	**37.32**	37.96	9.49

and flutter (AF), chronic kidney disease (CKD), and rheumatoid arthritis (RA). Disease selection and per-disease results are given in Appendix B, and evaluation protocols in Appendix B.4.

We automatically score all six systems twice on coverage, biomedical validity, evidence traceability, and prioritization and research depth. To isolate reasoning and synthesis from retrieval quality, all "+ Search" variants use the literature retrieved by BioRELAY. Experts then rate the top four systems on these dimensions plus readability, and BioRELAY receives high overall expert rating (Figure 4). Biomedical validity separates the leading systems least, suggesting that it depends largely on the base model, whereas traceability and ranking depend more on workflow structure. Removing the Search Agent lowers all four automatic scores, indicating that explicit retrieval grounds reasoning and improves protein prioritization[1].

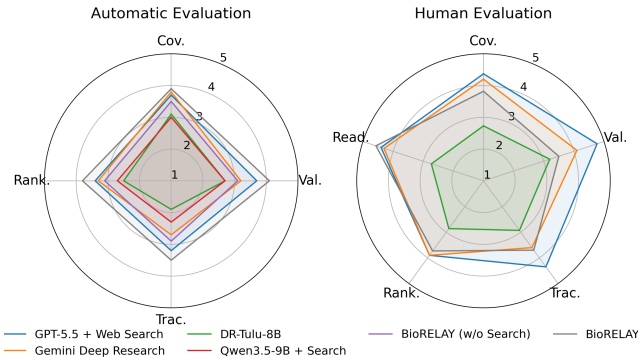


Figure 4: Mean automatic and expert scores of end-to-end reports over five diseases. Cov.: coverage; Val.: biomedical validity; Trac.: evidence traceability; Rank.: prioritization and research depth; Read.: readability and usability (expert only). Higher is better.

Experts also noted that BioRELAY alone consistently surfaced disease-association evidence and that its interface helped them check pathways, proteins, and citations. They were also willing to use clear, well-organized Markdown reports, suggesting that the interface's value lies in evidence traceability and navigation rather than visual presentation. Some coverage is lost as information passes through retrieval, reasoning, synthesis, and visualization. This loss did not undermine report usefulness here, but it suggests that retrieving more evidence is useful only when that evidence can be ranked, grounded, and inspected. Section 4.5 examines where this loss occurs.

4.5 Where Does Acquired Evidence Disappear?

To trace evidence loss after retrieval, we construct a reference set A_d for each of the five disease runs from the evidence available before synthesis. Rather than serving as an external gold standard, A_d records what the pipeline had already acquired. We then evaluate each downstream representation along two dimensions (Table 1). The unresolved rates U_{ent}, U_{rel}, and U_{cit} measure the fractions of rendered entities, protein–protein relations, and citations that cannot be resolved to acquired evidence. The coverage measures Cov_{prot}, Cov_{path}, and Cov_{pap} capture how much of the acquired evidence, such as proteins, pathways, and publications, respectively, remains observable in the representation.

[1]An anonymized example output is available at this webpage.

We compare BioRELAY's narrative report R and interactive evidence interface D with those produced by GPT-5.5 from either raw or typed evidence. For both baselines, the interface D is generated from the report R, whereas BioRELAY derives both representations from its shared evidence state. $\mathrm{Cov}_{\mathrm{path}}$ of BioRELAY's R is 100% because its report includes an inventory table of all acquired pathways. We treat diseases as the unit of analysis and report 95% bootstrap percentile intervals over 10,000 resamples, with paired resampling across systems for ratios.

Typing removes unsupported content but not omission. Typing nearly eliminates unsupported content in the GPT-5.5 baseline. As shown in Table 1, unresolved relations fall from 75.5% to 0.4% in R and from 70.0% to 1.3% in D. Yet the typed report retains only 5.5% of acquired publications. We claim that once evidence is typed, unsupported entities, relations, and citations become rare, yet most acquired evidence never reaches the final representation. From the same evidence, BioRELAY raises this coverage to 22.0% while keeping every citation resolvable. Its remaining unresolved content in R stems primarily from free-text pathway names and uncited mechanistic statements, which the gate does not read, and in D both are zero by construction, because D is a deterministic projection of $\mathcal{S}^\star$ whereas the baselines' D is generated from R by a model.

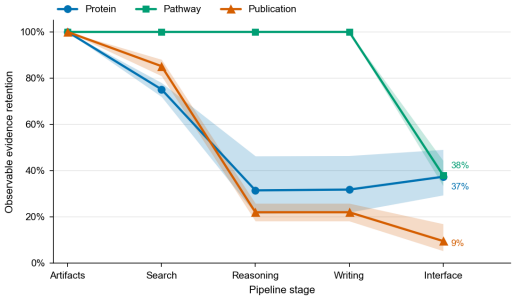


Figure 5: Per-stage retention of acquired proteins and publications relative to A_d (mean over five diseases; shading indicates 95% bootstrap intervals). Search and Reasoning are sequential stages, whereas the report R and interface D are independently derived from the finalized shared state $\mathcal{S}^\star$ rather than from each other.

Localizing evidence loss across stages. Per-stage retention (Figure 5) locates the main contraction at Reasoning, where protein retention drops from 75.1% after Search to 31.4% and publication retention from 85.2% to 22.0%, with little further change during Synthesis. Gate records separate this contraction from deliberate filtering by making deterministic omissions attributable. Across the five disease runs, Search retains 1,069 of 2,182 retrieved records in 78 packets, with all 1,113 omissions explained by deduplication, missing PubMed identifiers, or the per-packet budget ($b = 20$). Together with 2,956 records omitted under the Visualization display budget, this accounts for all 4,069 declared omissions (Section 4.6). Undeclared loss is therefore confined to model-mediated stages, where the model decides which supplied evidence to carry forward.

4.6 What Keeps Evidence Continuous?

We ablate the gate by its check classes (none, V, VC, VL, VLC). Schema checks alone cannot detect cross-stage evidence loss. Of 154 Reasoning calls, 28 silently ignored all supplied evidence while still producing schema-valid artifacts, and only the evidence-linkage check (L) rejected them. State consistency (C) adds no further rejections once linkage is enforced. Finally, holding $\mathcal{S}^\star$ fixed and swapping the Synthesis Agent's backbone, GPT-4o incurred 15 gate-detected publication drops versus none for GPT-5.4, while state-anchored outcomes remained nearly unchanged. At a fixed backbone, the gate therefore serves to detect and attribute evidence loss, while the amount of loss depends on the underlying model. No report or interface artifact was retained after a gate failure; all 28 retained failures are Reasoning artifacts (App. G.2).

5 Conclusion

We introduced **evidence continuity** as a measurable property of multi-stage biomedical agents and **BioRELAY** as a system for maintaining and auditing it. Across BioASQ, ProteInsight and end-to-end disease runs, the dominant failure shifted from hallucination to omission: models typically used resolvable evidence but failed to retain all required records, especially as evidence heterogeneity increased. BioRELAY's shared typed evidence state and deterministic gates substantially improved evidence retention while preserving fully resolvable citations and interface elements.

AI USE STATEMENT

Large language models are components of the proposed method and evaluation; their roles, model versions, and evaluation settings are documented in the corresponding method sections and appendices. During the preparation and revision of this manuscript, we used generative AI tools to assist with language editing and polishing, and with literature discovery to identify potentially relevant related work. AI tools were also used to assist with drafting and refining analysis code and manuscript text. For example, we use prompts such as "Help me correct the typos and grammatical errors in the above text, and streamline the logic to make it clear and easy to understand." All AI-assisted outputs, analyses, citations, and manuscript revisions were reviewed and verified by the authors.

ETHICS AND REPRODUCIBILITY STATEMENT

Ethics. Six domain experts evaluated anonymized system outputs, with additional expert review of sampled ProteInsight labels, evidence matches, and claim-level entailment (Appendix B.4; Section 4.3). All experiments use summary-level protein–disease association statistics derived from UK Biobank and public biomedical databases; no individual-level participant data are used. BioRELAY is intended as a research aid, not a source of clinical recommendations, and its outputs should be verified against the cited evidence before downstream use.

Reproducibility. The paper and appendices document the evidence state and gate invariants, experimental settings, benchmark construction, evaluation protocols, and statistical procedures, with all prompt templates provided verbatim in Appendix J. We will release the code, ProteInsight, and analysis scripts. Given saved outputs, grounding, completeness, interface projection, and downstream analyses are deterministic. Model-backed stages rely on proprietary APIs and may vary across runs; Appendix G.2 characterizes this variation through repeated runs.

REFERENCES

Niloofar Arazkhani, Maciej Kotecki, Brent Cochran, and Natasa Miskov-Zivanov. KALIMBA: Knowledge-assisted literature mining for biological interaction analysis. In Dina Demner-Fushman, Sophia Ananiadou, Kirk Roberts, and Junichi Tsujii (eds.), *BioNLP 2026*, pp. 880–889, San Diego, California, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-434-7. doi: 10.18653/v1/2026.bionlp-1.71. URL https://aclanthology.org/2026.bionlp-1.71/.

Manlian Bi, Zhijie Bao, Dongna Xie, Xiaohan Xie, Changxiao Yang, Tao Wang, Yongtian Wang, and Jiajie Peng. Bioragent: Natural language biomedical querying with retrieval-augmented multiagent systems. *Briefings in Bioinformatics*, 26(5):bbaf539, 2025. doi: 10.1093/bib/bbaf539.

Dechao Bu, Jingbo Sun, Kun Li, Zihao He, Wei Huang, Jinlin Hu, Shanshan Zhang, Shuangshuang Lei, Peipei Huo, Zhihao Wang, et al. Empowering ai data scientists using a multi-agent llm framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nature Biomedical Engineering*, pp. 1–16, 2026.

Justin Clark, Belinda Barton, Loai Albarqouni, Oyungerel Byambasuren, Tanisha Jowsey, Justin Keogh, Tian Liang, Christian Moro, Hayley O'Neill, and Mark Jones. Generative artificial intelligence use in evidence synthesis: A systematic review. *Research synthesis methods*, 2025.

Dasen Dai, Biao Wu, Meng Fang, and Wenhao Wang. Papervoyager: Building interactive web with visual language models. *arXiv preprint arXiv:2603.22999*, 2026.

Ella Flemyng, Anna Noel-Storr, Biljana Macura, Gerald Gartlehner, James Thomas, Joerg J Meerpohl, Zoe Jordan, Jan Minx, Angelika Eisele-Metzger, Candyce Hamel, et al. Position statement on artificial intelligence (ai) use in evidence synthesis across cochrane, the campbell collaboration, jbi, and the collaboration for environmental evidence 2025. *Campbell systematic reviews*, 21(4): cl2–70074, 2025.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl, 2025. URL https://arxiv.org/abs/2508.07976.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, October 2024. doi: 10.1016/j.cell.2024.09.022. URL https://doi.org/10.1016/j.cell.2024.09.022.

Kerstin Gierend, Frank Krüger, Sascha Genehr, Francisca Hartmann, Fabian Siegel, Dagmar Waltemath, Thomas Ganslandt, and Atinkut Alamirrew Zeleke. Provenance information for biomedical data and workflows: Scoping review. *Journal of Medical Internet Research*, 26:e51297, August 2024. doi: 10.2196/51297. URL https://www.jmir.org/2024/1/e51297/.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf H. Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N Carter, Xin Zhou, Matthew T. Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael P. Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:279170602.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023. doi: 10.1093/nar/gkad347.

Gautier Koscielny, Peter An, Denise Carvalho-Silva, Jennifer A. Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, Nikiforos Karamanis, Michael Maguire, Eliseo Papa, Andrea Pierleoni, Miguel Pignatelli, Theo Platt, Francis Rowland, Priyanka Wankar, A. Patrícia Bento, Tony Burdett, Antonio Fabregat, Simon Forbes, Anna Gaulton, Cristina Yenyxe Gonzalez, Henning Hermjakob, Anne Hersey, Steven Jupe, Şenay Kafkas, Maria Keays, Catherine Leroy, Francisco J. Lopez, María Paula Magariños, James Malone, Johanna McEntyre, Antonio Muñoz-Pomer Fuentes, Claire O'Donovan, Irene Papatheodorou, Helen Parkinson, Barbara Palka, Justin Paschall, Robert Petryszak, Naruemon Pratanwanich, Sirarat Sarntivijal, Gary Saunders, Konstantinos Sidiropoulos, Thomas Smith, Zbyslaw Sondka, Oliver Stegle, Y. Amy Tang, Edward Turner, Brendan Vaughan, Olga Vrousgou, Xosé Watkins, Maria J. Martin, Philippe Sanseau, Jessica Vamathevan, Ewan Birney, Jeffrey Barrett, and Ian Dunham. Open targets: A platform for therapeutic target identification and validation. *Nucleic Acids Research*, 45(D1):D985–D994, 2017. doi: 10.1093/nar/gkw1055.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

Yaniv Leviathan, Dani Valevski, Matan Kalman, Danny Lumen, Eyal Segalis, Eyal Molad, Shlomi Pasternak, Vishnu Natchu, Valerie Nygaard, Srinivasan Venkatachary, James Manyika, and Yossi Matias. Generative ui: Llms are effective ui generators. *arXiv preprint arXiv:2604.09577*, 2026.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2504.21776. URL https://arxiv.org/abs/2504.21776.

Congying Liu, Xingyuan Wei, Peipei Liu, Yiqing Shen, Yanxu Mao, and Tiehan Cui. Biomedsearch: A multi-source biomedical retrieval framework based on llms, 2025a. URL https://arxiv.org/abs/2510.13926.

Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents, 2025b. URL https://arxiv.org/abs/2509.06501.

Yi Luo, Linghang Shi, Yihao Li, Aobo Zhuang, Yeyun Gong, Ling Liu, and Chen Lin. From intention to implementation: automating biomedical research via llms. *Science China Information Sciences*, 68(7):170105, 2025.

Per Olav Løvsletten, Birk Stokke Hunskaar, Anja Fog Heen, Geertruida Bekkering, Louise Vanden Poel, Dena Zeraatkar, Mieke Vermandere, Bert Aertgeerts, Nicolas Delvaux, Frankie Achille, Jason W. Busse, Thomas Agoritsas, and Per Olav Vandvik. Physicians found an interactive tool displaying structured evidence summaries for multiple comparisons understandable and useful: a qualitative user testing study. *Journal of clinical epidemiology*, 172:111399, 2024.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 845–853, 2026.

Anna Martin-Boyle, Cara Leckey, Martha Brown, and Harmanpreet Kaur. Papertrail: A claim-evidence interface for grounding provenance in llm-based scholarly q&a. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–25, 2026.

Jörg Menche, Amitabh Sharma, M. Kitsak, S. Ghiassian, M. Vidal, J. Loscalzo, and A. Barabási. Uncovering disease-disease relationships through the incomplete human interactome. *Science*, 347 (6224):1257601, 2015. doi: 10.1126/science.1257601.

Yongming Miao, Jiaxin Wen, Yuemei Luo, and Jin Li. Medarc: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. *International Journal of Medical Informatics*, pp. 106136, 2025.

Christopher J. Mungall, Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, Erin Foster, J.P. Gourdine, Julius O.B. Jacobsen, Dan Keith, Bryan Laraway, Suzanna E. Lewis, Jeremy NguyenXuan, Kent Shefchek, Nicole Vasilevsky, Zhou Yuan, Nicole Washington, Harry Hochheiser, Tudor Groza, Damian Smedley, Peter N. Robinson, and Melissa A. Haendel. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722, 01 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1128. URL https://doi.org/10.1093/nar/gkw1128.

Naeimeh Nourmohammadi, Md Meem Hossain, The Anh Han, Safina Showkat Ara, and Zia Ush Shamszaman. A multi-agent framework for medical ai: Leveraging fine-tuned gpt, llama, and deepseek r1 for evidence-based and bias-aware clinical query processing. *arXiv preprint arXiv:2602.14158*, 2026.

Matthew James Page, Joanne McKenzie, Patrick Bossuyt, Isabelle Boutron, Tammy Hoffmann, cindy mulrow, Larissa Shamseer, Jennifer Tetzlaff, Elie Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy Grimshaw, Asbjørn Hróbjartsson, Manoj Mathew Lalu, Tianjing Li, Elizabeth Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley Stewart, James Thomas, Andrea Tricco, Vivian Andrea Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. *arXiv preprint arXiv:2407.08940*, 2024.

Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, et al. A knowledge graph to interpret clinical proteomics data. *Nature biotechnology*, 40(5):692–702, 2022.

Xinyi Shang, Xu Liao, Zhicheng Ji, and Wenpin Hou. Benchmarking large language models for genomic knowledge with geneturing. *Briefings in bioinformatics*, 26(5):bbaf492, 2025.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. Dr tulu: Reinforcement learning with evolving rubrics for deep research. In *International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2511.19399. URL https://arxiv.org/abs/2511.19399.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Distinguishing ignorance from error in llm hallucinations. *arXiv preprint arXiv:2410.22071*, 2024.

Tongyi DeepResearch Team. Tongyi deepresearch technical report, 2025. URL https://arxiv.org/abs/2510.24701.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015. doi: 10.1186/s12859-015-0564-6.

Deepak R. Unni, Sierra A. T. Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J. Harry Caufield, Paul A. Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, Gustavo Glusman, Jennifer J. Hadlock, Nomi L. Harris, Arpita Joshi, Tim Putman, Guangrong Qin, Stephen A. Ramsey, Kent A. Shefchek, Harold Solbrig, Karthik Soman, Anne E. Thessen, Melissa A. Haendel, Chris Bizon, Christopher J. Mungall, and The Biomedical Data Translator Consortium. Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, 15(8):1848–1855, 2022. doi: https://doi.org/10.1111/cts.13302. URL https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/cts.13302.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509, August 2025. doi: 10.1038/s41746-025-01840-7. URL https://doi.org/10.1038/s41746-025-01840-7.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and Ke Wang. Widesearch: Benchmarking agentic broad info-seeking, 2025. URL https://arxiv.org/abs/2508.07999. Benchmark for large-scale agentic information collection.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Qidi Xu, Claudio Soto, Mohammad Shahnawaz, Xiaozhong Liu, Xiaoqian Jiang, and Yejin Kim. Multi agent large language models for biomedical hypothesis generation in drug combination discovery. *Iscience*, 28(12), 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Gongbo Zhang, Qiao Jin, Denis Jered McInerney, Yong Chen, Fei Wang, Curtis L. Cole, Qian Yang, Yanshan Wang, Bradley A. Malin, Mor Peleg, Byron C. Wallace, Zhiyong Lu, Chunhua Weng, and Yifan Peng. Leveraging generative ai for clinical evidence synthesis needs to ensure trustworthiness. *Journal of Biomedical Informatics*, 153:104640, May 2024. doi: 10.1016/j.jbi.2024.104640. URL https://doi.org/10.1016/j.jbi.2024.104640.

Yikai Zheng, Wanquan Liu, Bi Zeng, Yichun Feng, Xiawei Du, Lu Zhou, and Yixue Li. Automating biomedical knowledge graph construction for context-aware scientific inference. *bioRxiv*, 2026. URL https://api.semanticscholar.org/CorpusID:284815117.

A LIMITATIONS

Our evaluation deliberately isolates evidence continuity from several adjacent questions. The end-to-end study covers five diseases from a single plasma-proteomics cohort and is therefore intended as a controlled comparison rather than a broad estimate of cross-cohort performance. ECR likewise assumes that all required evidence has already been retrieved, separating evidence preservation from retrieval quality; its references are constructed and evaluated with models from one family, and the operational definition of required evidence introduces unavoidable uncertainty in completeness estimates. The strength of the continuity guarantee also differs across outputs: the interactive interface is deterministically projected from and audited against $\mathcal{S}^\star$, whereas the narrative report is generated during state construction and checked for citation and claim anchoring rather than exact equivalence to the final state (Table 1). In addition, artifacts that remain noncompliant after one revision are logged rather than repeatedly regenerated, allowing failures to remain observable for analysis. Finally, the current gates and ECR assess whether required evidence remains available and traceable, not whether each record semantically entails the claim that uses it. Extending the framework to broader cohorts, retrieval settings, model families, and explicit entailment verification is therefore an important direction for future work.

B EXPERIMENTAL SETUP

B.1 DISEASE SELECTION

The end-to-end evaluation uses five diseases chosen to span distinct biomedical decision regimes (Table 2): neurodegenerative, psychiatric, cardiac rhythm, chronic renal/metabolic, and autoimmune/inflammatory. Each case is a complete pipeline run followed by expert rating, so the set is small by design, and disease is the unit of inference in every end-to-end analysis (Appendix B.3).

Table 2: Disease cases used in the end-to-end evaluation and what each case tests.

Disease	**Type**	**Key mechanistic features**	**What the case tests**
Alzheimer's disease (AD)	Neurodegenerative	Protein aggregation, neuroinflammation, synaptic degeneration, aging	Integration of complex neurological mechanisms and long-term progression evidence.
Depression (MDD)	Psychiatric	Neurotransmitters, inflammation, endocrine regulation, gene–environment interactions	Interpretation under heterogeneous mechanisms and uncertain disease biology.
Atrial fibrillation and flutter (AF)	Cardiac rhythm	Electrophysiological and structural remodeling, inflammation, coagulation risk	Integration of clinical risk, biomarker, and proteomic evidence.
Chronic kidney disease (CKD)	Chronic renal metabolic	Inflammation, fibrosis, oxidative stress, tubular injury	Interpretation of chronic progression, organ damage, and multi-pathway protein evidence.
Rheumatoid arthritis (RA)	Autoimmune/ inflammatory	Immune-cell activation, cytokines, synovial inflammation, tissue destruction	Interpretation of immune pathways and target-relevant protein evidence.

B.2 END-TO-END REPORT GENERATION.

Each disease run starts from UK Biobank plasma-proteomic associations filtered at a Bonferroni-corrected threshold of $P < 1 \times 10^{-9}$; only summary-level association data are used. Disease names are fixed from the cohort phenotype definitions and used consistently across retrieval, prompting, and state construction to avoid retrieval instability observed with model-generated query anchors during development. Unless otherwise stated, all artifact-level, stage-wise, and ablation analyses use a single frozen generation of the five disease runs from 2026-09-18, with all three gate invariants enabled (VLC), the default `keep` failure policy, and drug evidence provided to the Synthesis Agent; results are not pooled across generations. Per-disease input and finalized-state statistics are reported in Table 3, with cost and runtime reported in Appendix I. The expert evaluation in Table 5 is the sole exception: it evaluates reports from an earlier April generation that used model-generated disease names.

Table 3: Per-disease inputs and finalized evidence states in the reported generation. Publications are counted as distinct PubMed identifiers and drugs as distinct compounds.

Disease	Proteins	Pathways	Packets	Publications	Claims	Drugs
AD	10	6	14	122	23	8
AF	503	8	16	123	69	499
CKD	1,077	8	16	239	91	787
MDD	215	8	16	180	54	200
RA	315	8	16	199	61	389
Total	2,120	38	78	863	298	1,883

Table 4: Rubric for report evaluation. All dimensions are scored from 1 (poor) to 5 (excellent).

Dimension	What is assessed	Score 1	Score 5
Coverage (comprehensiveness)	Coverage and integration of proteins, pathways, functions, mechanisms, interactions, therapeutic evidence and citations.	Fragmentary or list-like	Complete, integrated interpretation
Biomedical validity	Correctness, disease relevance, mechanistic plausibility and appropriate qualification of biomedical claims.	Major errors or unsupported claims	Accurate and appropriately qualified
Evidence grounding and traceability	Whether conclusions can be traced to visible statistics, database records, literature and other supporting evidence.	Claims largely untraceable	Evidence visibly linked to conclusions
Prioritization and research depth	Whether important proteins, pathways and mechanisms are prioritized using interpretable, evidence-weighted reasoning.	Arbitrary or shallow ranking	Evidence-weighted, actionable prioritization
Readability and usability	Narrative coherence and, where applicable, the ability to inspect and navigate supporting evidence through the interface.	Difficult to follow or use	Clear, coherent and source-traceable

B.3 Statistical procedures

End-to-end and artifact-level analyses treat the disease as the unit of inference ($n = 5$). Intervals are 95% percentile intervals from $B = 10{,}000$ bootstrap resamples of the five diseases, and comparisons between systems resample diseases jointly (paired), and ratios are computed within each resample. The continuity analysis on ProteInsight (Section 4.3) instead resamples the 699 cases, each with all of its tasks. We report intervals rather than hypothesis tests. Declared omissions (attrition causes and display budgets) are always reported separately from undeclared loss.

B.4 Report Evaluation Protocol

Reports are evaluated on five dimensions using a shared five-point rubric (Table 4). Automatic evaluation uses the first four dimensions, and expert evaluation additionally assesses readability and interface usability.

Human evaluation. Six domain experts evaluated the five disease cases in Table 2. For each disease, evaluators received anonymized outputs from GPT-5.5 + Web Search, Gemini Deep Research, DR-Tulu-8B and BioRELAY in randomized order. They evaluated the generated report and, when available, its interface using the five dimensions in Table 4, assigning a score from 1 to 5 with a brief justification. All systems were evaluated on the same cases with the same rubric and presentation protocol. Each output was rated independently by the six experts, and its score is the mean of their ratings. The expert guideline named the third and fourth dimensions evidence synthesis depth and ranking quality; the human Trac. and Rank. columns report those ratings. Inter-rater agreement, measured with a one-way random-effects model, is ICC(1,6) = 0.76 for the six-expert mean score, and Krippendorff's α = 0.66.

Table 5: Automatic and human evaluation of end-to-end reports per disease (1–5, higher is better). Cov. coverage, Val. biomedical validity, Trac. evidence grounding and traceability, Rank. prioritization and research depth, Read. readability and interface usability.

		Automatic				Human				
Disease	System	Cov.	Val.	Trac.	Rank.	Cov.	Val.	Trac.	Rank.	Read.
AD	GPT-5.5 + Web Search	4.0	4.0	3.0	3.5	4.00	4.75	4.50	4.25	4.50
	Gemini Deep Research	4.0	3.0	3.0	3.5	4.50	3.50	4.00	4.00	4.50
	DR-Tulu-8B	3.0	2.5	2.0	2.5	3.00	4.00	3.67	3.33	3.67
	Qwen3.5-9B + Search	3.0	2.5	2.0	2.5	–	–	–	–	–
	BioRELAY (w/o Search)	3.5	3.0	3.0	3.0	–	–	–	–	–
	BioRELAY	4.0	4.0	3.5	4.0	3.75	4.50	4.50	4.50	4.50
CKD	GPT-5.5 + Web Search	3.5	4.0	3.5	3.5	4.33	5.00	4.33	3.67	4.00
	Gemini Deep Research	4.0	4.0	2.5	3.5	5.00	4.50	3.50	4.00	4.50
	DR-Tulu-8B	3.0	2.5	1.5	2.5	2.33	3.00	2.00	2.33	2.33
	Qwen3.5-9B + Search	3.0	3.0	2.5	3.0	–	–	–	–	–
	BioRELAY (w/o Search)	3.5	3.0	3.0	3.0	–	–	–	–	–
	BioRELAY	3.5	4.0	3.5	3.5	4.67	4.75	3.67	4.00	4.50
MDD	GPT-5.5 + Web Search	3.5	4.0	3.5	4.0	4.33	4.67	4.33	3.67	4.33
	Gemini Deep Research	4.0	3.0	2.5	3.5	4.50	4.50	3.50	4.50	4.50
	DR-Tulu-8B	2.5	2.5	1.5	2.0	3.00	2.33	3.33	2.33	2.00
	Qwen3.5-9B + Search	3.0	3.0	2.5	3.0	–	–	–	–	–
	BioRELAY (w/o Search)	3.5	3.0	2.5	3.0	–	–	–	–	–
	BioRELAY	4.0	4.0	3.5	4.0	3.67	4.00	4.50	4.25	4.50
RA	GPT-5.5 + Web Search	4.0	4.0	3.0	3.5	4.67	4.67	4.33	3.67	4.67
	Gemini Deep Research	3.5	3.0	2.5	3.0	4.50	4.50	3.50	3.50	4.50
	DR-Tulu-8B	3.0	2.0	1.5	2.5	2.33	3.33	2.33	3.00	2.33
	Qwen3.5-9B + Search	3.0	2.5	2.0	2.5	–	–	–	–	–
	BioRELAY (w/o Search)	3.5	3.0	3.0	3.0	–	–	–	–	–
	BioRELAY	4.0	4.5	3.5	4.0	3.33	4.75	4.33	3.67	4.83
AF	GPT-5.5 + Web Search	3.5	2.5	3.0	2.5	4.50	4.75	4.25	4.25	4.50
	Gemini Deep Research	3.5	3.0	3.0	3.0	2.50	3.50	3.50	3.50	3.50
	DR-Tulu-8B	4.0	4.0	3.0	3.0	3.00	3.33	3.33	3.33	3.33
	Qwen3.5-9B + Search	3.0	2.5	2.5	2.5	–	–	–	–	–
	BioRELAY (w/o Search)	3.5	3.5	3.0	3.5	–	–	–	–	–
	BioRELAY	4.0	4.0	3.5	3.5	3.67	4.00	4.25	4.50	4.50

Automatic evaluation. Automatic evaluation uses the same 1–5 rubric for the first four dimensions, excluding readability and interface usability. The judge is GPT-5.5 (reasoning effort medium) and receives the anonymized report together with the corresponding input protein table and interface. Each output is evaluated in two independent passes, and their mean is reported.

B.5 Per-disease report evaluation

Table 5 gives the per-disease scores summarized in Figure 4. Scores follow the protocol in Appendix B.4; only the four systems included in the human evaluation received expert ratings.

C The ProteInsight Benchmark

ProteInsight evaluates whether the evidence required for a reference conclusion remains identifiable through multi-stage biomedical reasoning. It contains 699 disease–protein cases and 2,796 tasks spanning four levels of evidence complexity.

C.1 Benchmark construction

Cases and evidence pool. We start from 699 disease–protein cases spanning 10 diseases: chronic kidney disease (254), type 2 diabetes (231), depression (91), atrial fibrillation (66), hypertension

Table 7: Evidence requirements across the four ProteInsight task types. $|E^\star|$ is the mean number of required record identifiers per task; Types is the mean number of required evidence branches.

Task	UKB	Func.	PPI	Path.	Lit.	$\lvert E^\star \rvert$	Types
Disease association	699	–	–	–	–	1.00	1.00
Protein–protein interaction	–	5	661	1	3	6.68	0.96
Protein function	–	680	273	12	560	6.44	2.18
Function + interaction	–	676	661	9	544	10.37	2.70

(19), rheumatoid arthritis (13), Alzheimer's disease (9), heart failure (8), Parkinson's disease (6), and stroke (2). For each case, we reconstruct evidence from UniProt function and disease annotations, GO terms, Reactome pathways, STRING interactions, PubMed records, and UK Biobank association statistics. STRING interactions are restricted to combined score ≥ 700 and at most 50 partners per query protein. The resulting pool contains 32,293 resolvable records (Table 6).

Table 6: The ProteInsight evidence pool.

Evidence type	Records	Cases
STRING interaction	15,378	661
UniProt/GO function	9,102	688
PubMed literature	5,047	635
Reactome pathway	2,064	531
UK Biobank association	702	699
Total	32,293	–

Tasks and required evidence. Each disease–protein case is instantiated under four question types: *disease association*, *protein–protein interaction*, *protein function*, and *function + interaction*, yielding 2,796 tasks. The tasks require progressively more heterogeneous evidence: disease association uses a single association branch, whereas the later tasks combine function, interaction, pathway, and literature records.

For each task u, we retain its reference conclusion $y_u^\star$ and construct a required-evidence set $E_u^\star$ from the reconstructed evidence pool. A record enters $E_u^\star$ only if it resolves to its underlying source and supports a statement required by $y_u^\star$. Table 7 summarizes the resulting evidence requirements. The mean number of required evidence types increases from 1.00 for disease association to 2.70 for function + interaction, while the mean number of required identifiers increases from 1.00 to 10.37.

Across the benchmark, the corresponding evidence graphs contain 699 required disease–protein edges and 11,039 required protein–evidence paths. Literature identifiers are attached as provenance for supported statements rather than represented as biological graph nodes.

Budgeted evidence packets. For the evaluation in Figure 3b, each task is answered from a packet containing the top b records of each evidence type in its case-specific pool. Interactions are ranked by STRING score, function records by annotation tier, and publications by the number of supported statements. We score a task only when $E_u^\star$ is non-empty and every required record occurs in the packet. At $b = 20$, 65 tasks have no resolvable required evidence and 258 have at least one required record outside the packet, leaving 2,473 tasks in $\mathcal{U}_{20}$. Packets in $\mathcal{U}_{20}$ contain 32.1 records on average, of which 21.0% belong to $E^\star$. Because every required record is available to the model, $\mathrm{ECR@20}$ isolates evidence preservation after retrieval rather than retrieval recall. It should therefore be interpreted as a conditional measure of evidence continuity, not as an end-to-end system metric.

C.2 Evidence continuity results

Table 8 reports validity, grounding, completeness, and $\mathrm{ECR@20}$ by task type. Grounding remains high across all four task types (97.3–99.9%), while completeness falls from 99.9% for disease association to 7.8% for function + interaction. Across all 2,473 tasks, 98.7% of answers are grounded but only 44.0% are complete. Among the 1,491 answers that fail at least one criterion, 1,353 (90.7%) are grounded but incomplete. Evidence omission therefore accounts for most failures as the required evidence becomes more heterogeneous.

C.3 Scope and limitations of ECR

ECR assumes that all required evidence has already been retrieved and therefore measures whether supplied evidence survives subsequent reasoning rather than retrieval quality; system-level com-

Table 8: Evidence continuity of GPT-4o at $b = 20$ on the 2,473 scorable tasks (%). The bracket gives the 95% case-level bootstrap interval for the overall ECR@20 (Appendix B.3).

Task type	Tasks	Val	Gnd	Cmp	ECR@20
Disease association	699	97.1	99.9	99.9	97.0
Protein–protein interaction	658	71.6	99.7	40.6	34.5
Protein function	550	86.4	97.6	14.5	11.1
Function + interaction	566	55.7	97.3	7.8	2.8
All	2,473	78.4	98.7	44.0	39.7

Table 9: Exposure of Cmp to interchangeable required records on the 2,473 scorable tasks at $b = 20$. $|E^\star|$: mean required records per task (over $\mathcal{U}_{20}$, so it differs from Table 7); Min.: mean records a minimal citation set needs under the declared rules; Exposed: tasks with at least one *any-of* branch listing two or more records. Strict: Table 8. Rules: the range within which each rate lies when completeness follows the declared rules (%).

					Cmp		ECR@20	
Task type	Tasks	$\lvert E^\star\rvert$	Min.	Exposed	Strict	Rules	Strict	Rules
Disease association	699	1.00	1.00	0 (0.0%)	99.9	99.9	97.0	97.0
Protein–protein interaction	658	7.06	7.05	2 (0.3%)	40.6	[40.6, 40.9]	34.5	[34.5, 34.8]
Protein function	550	5.23	2.38	412 (74.9%)	14.5	[14.5, 89.5]	11.1	[11.1, 86.0]
Function + interaction	566	9.01	6.42	379 (67.0%)	7.8	[7.8, 74.7]	2.8	[2.8, 55.7]
All	2,473	5.39	4.16	793 (32.1%)	44.0	[44.0, 76.1]	39.7	[39.7, 68.5]

parisons instead require matched retrieval APIs, query budgets, and evidence availability, and are evaluated separately in Sections 4.4 and 4.5. Its required-evidence set $E^\star$ is an operational definition of the evidence needed by the reference conclusion: disease-association references are constructed deterministically from cohort statistics, while the remaining reference conclusions are generated by GPT-5 from UniProt and STRING records, after which required records are matched and resolved against the underlying databases without a language model. Thus, every record in $E^\star$ is externally resolvable, although the reference signal is not fully model-independent. Both Gnd and Cmp operate at the identifier level: they test whether cited records resolve to the supplied evidence and whether the identifiers in $E^\star$ remain represented in the answer, rather than whether each cited record semantically entails the claim in which it appears. This also makes Cmp deliberately strict, since interchangeable support—such as a UniProt annotation and its corresponding GO term, or multiple PMIDs supporting the same claim—is counted separately. We therefore interpret $E^\star$ as a non-exhaustive operational definition of required evidence and Cmp as strict identifier-level evidence preservation rather than semantic evidence completeness. Expert validation of reference matching and answer judgments is reported in Section 4.3.

D System Design

D.1 Shared Evidence State

The shared evidence state $\mathcal{S}$ is the typed object over which evidence continuity is defined. Reasoning artifacts update the state, gates validate model-written outputs against it, the report is written against it as it is built, and the interface is projected from its finalized form $\mathcal{S}^\star$. ProteInsight gold graphs use the same schema and identifier constructors, enabling record-level alignment between benchmark evidence and system states.

All entities use normalized Biolink categories and deterministic CURIE constructors, such that a record retains the same identifier across stages. The Reasoning Agent is the only model component that proposes biomedical state updates. Updates add or revise records by identifier and are versioned with provenance rather than rewriting the state as free text. A publication enters `evidence` only when it is cited by a reasoning artifact and resolves to the evidence packet supplied to that call; unresolved retrieved records remain in `candidates`. These rules make evidence linkage and loss traceable across

Table 10: Typed collections in the shared evidence state.

Collection	Content
`entities`	Normalized biomedical entities identified by CURIE, with Biolink categories (disease, gene, pathway and other terms, drug). Gene attributes carry the input association statistics.
`requests`	Structured evidence requests q (pathway, interaction, or target–drug) that produced the records above.
`evidence`	Publications the Reasoning Agent accepted as support, each with a stance (`supports`, `contradicts`, or `qualifies`).
`candidates`	Publications scored for a request but not accepted as evidence, with $f(q, a)$, its components, and a rejection cause of `threshold` or `budget`.
`associations`	Typed, directed relations: gene–term membership with enrichment statistics, protein associations, drug–target edges, and drug–indication edges.
`claims`	Synthesized statements, each linked to its supporting evidence, with a `certainty` of high, moderate, or low.
`protein_functions`	Per-protein function statements, each with its organism, condition, and provenance.
`interactions`	Protein associations typed `functional_association` or `physical_interaction` from their evidence channels, with a score.
`drug_targets`	A drug's action on a protein: action type, clinical phase, and the indications it is studied or approved for.
`gaps`	Unresolved evidence gaps, which may drive the next retrieval round.

state transitions. Drug evidence is added deterministically from Open Targets rather than generated by a model. Drug–target relations retain action type and clinical phase; drug–indication relations are typed as `biolink:treats` at phase ≥ 4 and otherwise as `biolink:related_to`, with phase retained as a qualifier. For synthesis, drug relations are projected onto the proteins belonging to each analyzed pathway.

D.2 Evidence-Continuity Gate

The Evidence-Continuity Gate is a deterministic, model-free validator applied to artifacts produced across the BioRELAY trajectory. For an artifact O_k at stage k, it evaluates three invariant classes against the evidence and state available to that stage: schema validity (V), evidence linkage (L), and state consistency (C). Evidence continuity is the property enforced by these invariants rather than a separate invariant.

The implementation defines 56 finding codes across the three invariant classes and the gate itself. Each finding has a severity; errors cause rejection, whereas warnings are logged without rejecting the artifact. Table 11 describes the codes on which the analyses in this paper rely.

Validation is stage-specific: each artifact is checked against the evidence and state available when it is produced, so the gate tests continuity with actual upstream evidence rather than an external reference set. When an artifact fails, the producing agent may revise it once using the gate findings; under the default `keep` policy, residual failures are retained and logged, while the alternative `drop` policy discards them. Reported gate pass rates are computed from first attempts only. The gate establishes structural validity, evidence linkage, and consistency with supplied state, but does not verify biomedical entailment or completeness with respect to evidence never supplied to the system; external completeness is evaluated separately using ProteInsight.

D.3 Search Agent and Retrieval

The Search Agent is deterministic and uses no language model. Given a frozen retrieval corpus, enrichment, retrieval, ranking, deduplication, and packet construction are deterministic. Initial requests are derived from pathway enrichment, whereas follow-up requests are proposed by the Reasoning Agent and processed by the same retrieval procedure.

Enrichment and retrieval. Input proteins are mapped to enriched biological terms with g:Profiler. Near-duplicate terms are merged by gene-set overlap and semantic similarity, and disease-relevant

Table 11: Gate finding codes referenced in this paper. The prefix gives the invariant class (Section 3.2.4).

Code	Finding
`L.packet_unused`	A Reasoning artifact uses none of the evidence in its input packet (Appendix G.1).
`L.citation_not_in_packet`	A cited publication does not resolve to the packet supplied to that call.
`L.record_citation_not_in_packet`	The same check for typed records (e.g., GO terms); active only when packet items carry record identifiers.
`L.claim_omitted`	A Synthesis output omits a claim that the state held when the call was made; counted per call, not per document (Appendix F).
`L.publication_dropped`	A Synthesis output drops a publication present in the call's recorded input (Appendix G.3).
`L.budgeted_omission`	A state relation that the interface does not draw and that a declared display budget accounts for; recorded as a declared omission, not as a failure.
`C.unsupported_literature_claim`	A literature claim in a call whose context declares that no literature was supplied (the two overview calls; Appendix E.2).
`C.unsupported_relation_claim`	(Warning.) A physical or regulatory verb stated between two protein symbols with no citation in the sentence or its paragraph; runs only where publications were supplied.
`C.unsupported_record_in_prose`	A typed record named in prose that was not supplied to the call; active under the same condition as `L.record_citation_not_in_packet`.

pathways and biological processes are prioritized for downstream analysis. At most 12 pathways are retained per disease.

For each evidence request, the Search Agent retrieves literature from PubMed and Semantic Scholar using the disease and pathway or follow-up query. Retrieved records are ranked by the fixed relevance function in Eq. 2, which combines lexical relevance, BioBERT semantic similarity, citation information, and a fixed venue prior. Low-scoring and duplicate records are removed, and each request contributes at most 20 publications to the evidence packet. The same ranking and filtering procedure is used for follow-up requests.

Retrieval provenance and attrition. Search-stage removal is explicitly recorded as duplication, missing identifier, relevance filtering, or packet-budget truncation. These declared removals are distinguished from downstream evidence loss in Appendix F. External retrieval responses are cached by request and frozen within each generation, so repeated processing of the same request uses the same retrieved evidence. Retrieval quality itself is not evaluated here; the end-to-end experiments therefore measure evidence continuity under this fixed retrieval policy rather than robustness to alternative search or ranking strategies.

E Implementation Details

This appendix summarizes the implementation details needed to interpret and reproduce the reported experiments. Prompt templates for all model-backed calls are provided in Appendix J; retrieval is described in Appendix D.3.

E.1 Model and pipeline configuration

The Reasoning and Synthesis Agents use GPT-5.4. The Search Agent, Orchestrator, evidence-continuity gate, and interface projector are deterministic and call no language model. Model calls use the provider defaults for temperature and reasoning effort. Reasoning outputs are parsed into typed state updates, while Synthesis outputs are report sections linked to claims in the shared evidence state. Every model-backed artifact is checked by the gate before it is incorporated downstream.

The reported runs use at most two Search–Reasoning iterations, with up to four follow-up requests per iteration. Evidence packets contain at most 20 publications. The gate enforces validity, evidence linkage, and state consistency (VLC), permits one revision after a failed check, and retains the artifact

with its failure record if the revision still fails. These settings are fixed across the five end-to-end disease runs.

For the synthesis-model substitution experiment (Appendix G.3), the recorded Synthesis inputs and final evidence state $\mathcal{S}^\star$ are held fixed and only GPT-5.4 is replaced by GPT-4o.

E.2 Model-backed operations

Only reasoning over retrieved evidence and synthesis of the final report are model-backed. During reasoning, the model analyzes pathway evidence, protein–protein interaction clusters, protein functions, and unresolved evidence gaps. Each output cites identifiers from its supplied evidence and is converted into a typed update $\Delta\mathcal{S}^{(t)}$ to the shared evidence state (Appendix D.1).

During synthesis, the model converts the accumulated state into report sections. Generated passages carry temporary claim anchors linking them to claims in $\mathcal{S}$; these anchors are checked by the gate and removed before rendering. The interface is not generated by a language model but projected deterministically from the finalized state.

E.3 External resources

Agents do not invoke external resources directly. Retrieval and database access are handled by deterministic components, while model-backed agents express information needs through typed requests that the Orchestrator routes to the Search Agent. The pipeline uses UK Biobank summary statistics for the input associations, g:Profiler for enrichment, PubMed and Semantic Scholar for literature retrieval, STRING for protein interactions, Open Targets for drug relations, and a local protein-function knowledge base for functional annotations. BioBERT is used for semantic matching during evidence processing. Retrieval settings are described in Appendix D.3.

E.4 What the comparisons isolate

The experiments differ in which components they hold fixed, and therefore support different conclusions. Comparisons with external systems are end-to-end and should be interpreted at the system level; the contributions of individual components are isolated only by controlled comparisons within BioRELAY.

F Artifact Faithfulness and Stage-wise Evidence Loss

For each disease d, we construct an acquired-evidence reference set A_d from the evidence available to downstream generation after retrieval. It contains the grounded input proteins, selected pathways, and retained publications. Records removed during Search for a declared reason, such as duplication, missing identifiers, or the packet budget, are excluded. Thus, A_d is not an external gold standard; it measures whether evidence already available to the pipeline remains represented downstream.

For representation $X \in \{R, D\}$, let $\mathcal{E}_X$, $\mathcal{P}_X$, and $\mathcal{M}_X$ denote its rendered entities, protein–protein relations, and distinct PMIDs. We measure the unresolved rate as the fraction of rendered content that cannot be resolved to acquired evidence:

$$U_{\text{ent}}(X) = \frac{|\{e \in \mathcal{E}_X : e \notin A_d\}|}{|\mathcal{E}_X|}, \tag{8}$$

$$U_{\text{rel}}(X) = \frac{|\{r \in \mathcal{P}_X : r \text{ is unsupported by } A_d\}|}{|\mathcal{P}_X|}, \tag{9}$$

$$U_{\text{cit}}(X) = \frac{|\mathcal{M}_X \setminus A_d^{\text{pap}}|}{|\mathcal{M}_X|}. \tag{10}$$

Coverage measures the complementary direction—how much acquired evidence remains observable in the representation:

$$\text{Cov}_k(X) = \frac{|A_d^k \cap X|}{|A_d^k|}, \qquad k \in \{\text{prot}, \text{path}, \text{pap}\}. \tag{11}$$

Table 12: Fate of every acquired publication in the reported generation (% of acquired, per disease). Not selected: retrieved only to rank an enriched term that pathway prioritization did not select. Function ref.: UniProt reference of a function record, not supplied as literature. Uncited: supplied to the Reasoning Agent but not cited, the only undeclared class.

Disease	Acquired	Rendered in R	Not selected	Function ref.	Uncited
AD	501	24.4	53.3	0.8	21.6
AF	1,210	19.3	67.3	7.0	6.4
CKD	2,207	24.4	63.3	8.5	3.7
MDD	1,650	15.9	76.5	1.8	5.8
RA	1,714	17.8	75.2	2.3	4.7
Mean	1,456	20.4	67.1	4.1	8.4

The two measures therefore separate unsupported rendered content from omission of evidence that was already available.

Extraction. Identifiers, protein symbols, PMIDs, and table contents are parsed deterministically. Entities and relations expressed only in prose are extracted with GPT-5.5 from the complete report and deduplicated across overlapping text windows. Interfaces (D) are parsed deterministically.

The unresolved entities in BioRELAY's report are primarily unmatched pathway names (110 of 124), with no unsupported proteins detected across the five reports; lexical variation may therefore contribute to its measured U_{ent}. Most unresolved protein–protein relations likewise occur in prose without a citation. We conservatively retain these cases as unresolved. In contrast, the deterministic projection to D introduces no unresolved entities, relations, or citations.

Stage-wise retention and attributable loss. For Figure 5, retention of evidence type k after stage t is

$$\mathrm{Ret}_k(t,d) = \frac{|A_d^k \cap O_{t,d}^k|}{|A_d^k|}, \qquad k \in \{\mathrm{prot}, \mathrm{pap}\}, \tag{12}$$

where $O_{t,d}^k$ is the type-k evidence observable after stage t. Pathways are omitted because every selected pathway receives a report section and an interface node by construction. Search and Reasoning form the sequential evidence-processing path; R and D are parallel representations of the finalized shared state rather than successive stages.

Retention alone does not distinguish intentional filtering from evidence lost without an explicit reason. We therefore pair it with the recorded transition outcomes. Across the five runs, all deterministic reductions at Search and interface projection have declared causes, including deduplication, missing identifiers, packet limits, and display budgets. Undeclared loss appears instead at model-mediated transitions: 28 of 154 Reasoning artifacts use none of the evidence supplied in their packet despite remaining schema-valid. At Synthesis, per-call omissions are evaluated only after report assembly; all 298 claims in $\mathcal{S}^\star$ appear in the final reports. These quantities are interpreted at their native units rather than combined into a single loss rate.

Fate of acquired publications. Publication coverage uses all acquired publications as its denominator, including records that were never intended to appear in the final report. Table 12 therefore separates publications rendered in R from those used only during pathway prioritization, attached to protein function records, or supplied to Reasoning but left uncited.

On average, 67.1% of acquired publications are filtered during pathway selection and 4.1% are function-record references. Among publications actually supplied to the Reasoning Agent, the remaining undeclared class is the 8.4% that is not cited. Every publication admitted to $\mathcal{S}^\star$ as support for a claim reaches the final report, as do all 298 claims. This analysis concerns retention of acquired evidence and is distinct from ECR, which evaluates completeness against the task-specific required set $E^\star$.

Table 13: Unresolved rates and evidence coverage with bootstrap uncertainty (intervals for Table 1). Coverage (Cov) is the share of acquired evidence that reaches each representation (higher is better); unresolved rates (U) are shares of rendered items that do not resolve to acquired evidence (lower is better). Each cell is the mean over five diseases, with the 2.5th and 97.5th percentiles of 10,000 bootstrap resamples of the diseases beneath it (each cell resampled independently). n_{cit} is the mean number of distinct rendered PMIDs, the denominator of U_{cit}, and is not a percentage.

		Unresolved (%)				Coverage (%)		
System	**Rep.**	$U_{\mathrm{ent}} \downarrow$	$U_{\mathrm{rel}} \downarrow$	$U_{\mathrm{cit}} \downarrow$	n_{cit}	$\mathrm{Cov}_{\mathrm{prot}} \uparrow$	$\mathrm{Cov}_{\mathrm{path}} \uparrow$	$\mathrm{Cov}_{\mathrm{pap}} \uparrow$
GPT-5.5	R	35.29 [27.69, 45.69]	75.50 [52.50, 98.00]	0.91 [0.00, 5.17]	37.0 [33.8, 40.0]	32.45 [14.69, 57.02]	7.77 [3.33, 12.77]	3.55 [2.06, 5.85]
(raw)	D	42.00 [27.63, 57.74]	70.00 [30.00, 100.00]	0.96 [0.00, 4.52]	28.2 [24.0, 32.8]	22.42 [8.94, 37.65]	5.98 [1.67, 10.30]	2.58 [1.65, 3.99]
GPT-5.5	R	**0.29** **[0.00, 0.67]**	**0.36** **[0.00, 1.07]**	1.05 [0.00, 3.16]	74.4 [31.4, 117.4]	**51.17** **[34.89, 68.46]**	89.55 [83.31, 95.78]	5.48 [2.38, 9.22]
(typed)	D	9.08 [2.04, 16.13]	1.33 [0.00, 4.00]	0.75 [0.00, 2.10]	28.6 [14.4, 43.4]	32.83 [15.67, 57.67]	**46.57** **[21.33, 71.33]**	2.28 [1.34, 3.58]
BioRELAY	R	8.60 [6.23, 10.40]	19.27 [9.17, 27.31]	0.00 [0.00, 0.00]	292.4 [180.8, 428.6]	31.76 [21.91, 46.39]	100.00 [a] [100.00, 100.00]	**21.95** **[18.13, 25.77]**
	D	**0.00** **[0.00, 0.00]**	**0.00** **[0.00, 0.00]**	0.00 [0.00, 0.00]	94.4 [83.0, 104.0]	37.32 [29.21, 49.03]	37.96 [33.41, 44.29]	9.49 [5.21, 16.89]

[a] BioRELAY's report contains inventory tables that list every pathway in A_d, and a pathway counts as covered when its identifier or name appears anywhere in the representation, so this cell reflects the inventory listing.

G More Controlled Ablations

We provide the controlled analyses underlying the gate results in Section 4.6. They test which invariant exposes cross-stage evidence loss, how the gate failure policy affects retained artifacts, and whether residual synthesis loss changes when the evidence state is held fixed but the model is replaced.

G.1 Invariant ablation

We ablate the gate using cumulative subsets of its three invariant classes: schema validity (V), evidence linkage (L), and state consistency (C). The five arms are `none`, V, VC, VL, and VLC. Rejection is computed counterfactually from each artifact's first-attempt gate findings, before any revision, so the comparison isolates detection capacity.

Schema validity detects one malformed artifact, whereas adding evidence linkage raises the number detected to 29. The 28 additional failures are `L.packet_unused`: schema-valid Reasoning outputs that use none of the evidence supplied in their input packet. Evidence linkage therefore detects continuity failures that are invisible to schema validation alone. State consistency identifies no additional failure in these runs once linkage is enforced; this indicates only that no additional violation of that class occurred in this experiment.

Table 14: Gate invariant ablation over 323 first-attempt artifacts. V/L/C: validity/linkage/consistency.

Arm	**Rejected**	**Rate (%)**
none	0	0.00
V	1	0.31
VC	1	0.31
VL	29	8.98
VLC	29	8.98

G.2 Failure policy and gate scope

Table 15 follows each gated artifact to its final state. Under the default `keep` policy, an artifact that still fails after one revision is retained with its failure record; under `drop`, it would be discarded. In the reported generation, the 28 unresolved failures all occur during Reasoning and are retained under `keep`. The single Synthesis failure is repaired, and all Search and Visualization artifacts pass.

The gate guarantees continuity only for the structures it checks. For the report R, it validates identifiers and claim anchors rather than arbitrary relations expressed in free text; after assembly, all 298 claims in $\mathcal{S}^{\star}$ are represented. The interface D is more strongly constrained because it is

Table 15: Final state of every gated artifact in the reported generation. Revised: failed on the first attempt and received one revision. Retained failing: still failing after revision and kept under the default `keep` policy; a `drop` policy would discard exactly these. Visualization: one interface per disease.

Stage	Artifacts	Passed first	Revised	Repaired	Retained failing
Search	78	78	0	0	0
Reasoning	154	126	28	0	28
Synthesis	91	90	1	1	0
Visualization	5	5	0	0	0
Total	328	299	29	1	28

Table 16: Alzheimer's disease under the reported generation and two reruns on the same frozen corpus. Reasoning: artifacts passing on the first attempt. Retained failing: kept after a failed revision. R and D columns as in Table 1 (%); U of D is zero for entities, relations and citations in all three runs.

					R			D
Run	Reasoning	Retained failing	Claims	Evidence	U_{ent}	U_{rel}	Cov_{pap}	Cov_{pap}
Reported (`keep`)	25/26	1	23	134	4.17	0.00	27.11	24.00
Rerun, `keep`	24/26	1	23	131	22.03	0.00	26.04	21.89
Rerun, `drop`	25/26	0	21	116	26.79	42.86	25.56	21.52

projected deterministically from $\mathcal{S}^\star$, and every rendered element is independently checked against that state. Thus, stricter handling of failed artifacts and stricter validation of report prose are separate interventions.

`keep` and `drop` reruns. To examine the direct effect of the failure policy, we reran Alzheimer's disease on the same frozen retrieval corpus under `keep` and `drop` (Table 16). The same protein-function artifact failed in all three runs. Under `drop`, removing it reduces the final state by two protein-level claims and their associated evidence, while the remaining pipeline completes normally. Because the policy operates on whole artifacts, it can also remove supported content contained in the same failed artifact.

The report-level metrics vary substantially even between the two `keep` runs. With one disease and one rerun per policy, these results therefore establish the direct structural effect of dropping a failed artifact, but do not estimate its effect on report quality or coverage.

G.3 Synthesis-model substitution

To separate model capability from the evidence architecture, we hold the final typed evidence state $\mathcal{S}^\star$ and all upstream artifacts fixed and replay only the Synthesis calls with GPT-4o in place of GPT-5.4. The recorded inputs and report-assembly roles are unchanged, so any difference arises downstream of the same evidence state.

Holding $\mathcal{S}^\star$ fixed leaves the assembled output largely stable: publication coverage changes by -0.61 percentage points, all 298 claims reach the report under both models, and neither produces unresolved citations. The gate nevertheless detects 15 publication drops with GPT-4o and none with GPT-5.4. Thus, the shared state stabilizes state-anchored final coverage, while model capability still affects residual evidence loss during individual synthesis transitions.

We do not compare aggregate `L.claim_omitted` counts across the two runs because their gate contexts have different denominators: the replay is checked against the fixed final state $\mathcal{S}^\star$, whereas calls in the original trajectory are checked against the state available at their execution time.

Table 17: Synthesis-model substitution with the typed evidence state $\mathcal{S}^\star$ and upstream artifacts held fixed. Publication coverage and claim transfer describe the assembled report; gate-detected publication drops are measured at individual synthesis transitions against the evidence supplied to each call.

	GPT-5.4	**GPT-4o**
Publication coverage (%)	45.38	44.76
Coverage difference (pp)	–	−0.61
Claims transferred from $\mathcal{S}^\star$	298 / 298	298 / 298
Unresolved citations	0	0
Gate-detected publication drops	0	15
Diseases with publication drops	0 / 5	2 / 5

H CASE STUDY: ALZHEIMER'S DISEASE

We use Alzheimer's disease (AD) to illustrate the interface as an evidence-centered workspace rather than a single report. AD is a useful case because its protein signals must be read across synaptic dysfunction, lipid metabolism, glial activation, axonal injury and neurodegeneration instead of one dominant pathway.

Given the disease name, cohort metadata and 10 significant plasma proteins, BioRELAY builds a chain from protein statistics to pathways, publications, reasoning artifacts and the interface in Figure 6. Enrichment proposes candidate pathways, the Search Agent retrieves disease-specific literature for each, the Reasoning Agent integrates pathway, protein, interaction and drug evidence into $\mathcal{S}$, and the interface renders the finalized state for inspection.

The resulting workspace keeps the interpretation tied to the observed protein signals. APOE appears as a cross-pathway protein linking lipid receptor biology, vesicle organization, synaptic processes and axon-related hypotheses. GFAP and NEFL support glial and axonal-injury interpretations, while SNAP25 and SYT1 form a presynaptic vesicle and chemical-synapse module. Proteins with weaker pathway or literature support remain visible but are marked exploratory.

The case also shows why enrichment and literature support are displayed side by side. A pathway can be statistically enriched yet weakly supported by AD-specific literature, while another has moderate enrichment but clearer disease relevance. Showing both lets a reader follow the path from proteins to pathways, citations and mechanisms and judge each step, rather than accept a single narrative ranking.

I COMPUTATIONAL COST

Across the five disease runs, generation required 279 GPT-5.4 calls and cost $12.40 (about $2.48 per disease; 1.35M input and 0.60M output tokens). Reasoning accounted for 182 calls and Synthesis for the remaining 92; Search used no model calls, and retrieval relied on public APIs without additional model cost. Most computation is therefore spent once in constructing the shared evidence state, after

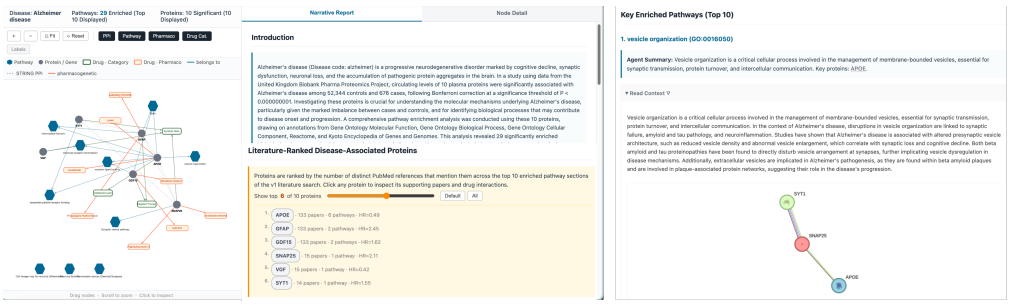


Figure 6: Case study on Alzheimer's disease.

which additional representations are inexpensive: the interface is deterministically projected from the fixed $\mathcal{S}^\star$ without model calls in approximately 0.1 s per disease, while replaying only Synthesis against recorded inputs cost $0.17 across all five diseases with GPT-4o (Appendix G.3). This "build once, project many" design also avoids repeatedly passing evidence through generative transformations: in the report-first typed baseline, generating D from R reduces mean protein coverage from 51.17% to 32.83% and publication coverage from 5.48% to 2.28% (Table 13). Thus, the upfront cost of constructing $\mathcal{S}^\star$ is amortized across downstream representations while deterministic projection adds negligible model cost without introducing an additional generative evidence-loss step.

J PROMPT TEMPLATES

This appendix reproduces, verbatim, the prompt templates of every model-backed call in BioRELAY as used in the reported generation. Each call sends one system and one user message (Appendix E.1). Placeholders in braces are filled at run time; the disease name inserted into every template is the fixed name derived from the cohort phenotype definition (Appendix B.2), never a model-generated one, and an evidence packet is rendered as numbered entries giving each publication's PubMed identifier, journal, date, relevance score, title and abstract.

Listings 1–3 are shared components: the evidence-assessment requirements inserted into the four Reasoning prompts, the correction block appended when the gate rejects an output (Appendix D.2), and the system prompt of the Synthesis Agent. Listings 4–7 are the Reasoning calls and Listings 8–12 the Synthesis-stage calls, in pipeline order. The two overview templates describe only the input they are given and ask for no literature support, and the writing examples they contain satisfy the same gate checks as their outputs.

Listing 1: Shared by the four Reasoning calls: evidence assessment requirements, inserted where marked.

```
# Evidence Assessment Requirements (IMPORTANT):
- For EVERY PubMed identifier you cite, add one entry to ‘evidence_assessment‘ recording whether that publication
      ‘supports‘, ‘contradicts‘, or ‘qualifies‘ the statement it is attached to.
- Use ‘contradicts‘ when the publication reports a finding inconsistent with the statement, and ‘qualifies‘ when
     it supports the statement only under stated conditions, in a specific population, or with an explicit
     caveat.
- Do NOT discard a contradictory or qualifying publication. Report it with its stance and reflect that tension in
      the narrative.
- Set ‘certainty‘ to ‘high‘ only when the cited evidence is consistent and direct. If ANY cited publication
     contradicts the statement, ‘certainty‘ must be ‘moderate‘ or ‘low‘.
- List every unresolved evidence gap in ‘gaps‘: a specific question the supplied publications do not answer. Use
     an empty list when there is none. Do not invent a gap to fill the field.
```

Listing 2: Appended to any prompt whose output the gate rejected; `{revision_note}` lists the gate findings.

```
{prompt}

# REQUIRED CORRECTIONS
{revision_note}
```

Listing 3: System prompt of the three Synthesis Agent calls (network overview, revision, introduction).

```
You are a biomedical research writer tasked with generating well-structured, concise scientific reports in
     Markdown format.

# Guidelines:
- Maintain a formal and scientific tone throughout the writing.
- Integrate numerical data and statistical evidence wherever possible.
- Use data-driven reasoning to support all claims and observations.
- Use as much number and data statistics as possible!
- Always use the data to support the writing.
- Don’t use bullet points. Be coherent and narrative.
- KEEP all embedded figures, tables, and links in valid Markdown format.
- Markdown tables (GitHub-flavored): put one table row per line; include a header separator line with dashes (|
     --- | --- |); leave a blank line before and after each table so renderers recognize it.
- Figures use image syntax ONLY for local image files: ![descriptive alt text](imgs/file.png). Ordinary
     hyperlinks use [link text](https://...). NEVER use image syntax (![...]) for any web URL — always use plain
      link syntax ([text](url)) for STRING, PubMed, or any other website.
- DO NOT use any acronym! Don’t use any abbreviation!
```

Listing 4: `Pathway_Analysis` (Reasoning). System prompt, then user prompt.

```
[SYSTEM]
You are a biomedical research assistant who helps summarize biological findings. Please response in the academic
      writing style. This will be used as part of academic report.

[USER]
Here is the description and the relevant publications for a biological pathway, please analyze the pathway in the
       context of {disease_name} disease:

# Pathway Description:
{pathway_info}

# Relevant Publications:
{publication_info}

# Task:
Please analyze the provided publications, and find out how this pathway is related to {disease_name}.
Please provide a detailed introduction of the pathway, and a narrative coherent and comprehensive paragraph that
      explains how this pathway is related to {disease_name}.

# Citation Requirements (IMPORTANT):
- Place each citation INLINE, in parentheses, immediately AFTER the specific sentence or claim it supports, using
       the exact form ‘(PubMed:XXXXXXXX)‘.
- When a single claim is supported by more than one publication, list them together as ‘(PubMed:XXXXXXXX, PubMed:
      YYYYYYYY)‘.
- Cite the SPECIFIC publication(s) whose Title/Abstract actually support that claim. DO NOT attach the full
      publication list to a single sentence, and DO NOT dump all citations at the end of the paragraph.
- Every literature-supported claim in both ‘description‘ and ‘explanation‘ must carry its own inline citation so
      a reader can trace each statement to its evidence.
- Only cite PubMed IDs that appear in the Relevant Publications above. Do not invent identifiers.
DO NOT use any acronym! Don’t use any abbreviation!

[Evidence Assessment Requirements: Listing 1]

# Response Format:
- A json code block with the following format:
‘‘‘json
{
  "description": (String)"Detailed introduction of the pathway, with inline (PubMed:XXXXXXXX) citations placed
      right after each supported claim",
  "explanation": (String)"A narrative coherent and comprehensive paragraph that explains how this pathway is
      related to {disease_name}, with inline (PubMed:XXXXXXXX) citations placed right after each supported claim",

  "citations": (List) "The complete set of PubMed IDs used inline above (for reference indexing), in the format
      of ['PubMed:XXXXX', 'PubMed:XXXXX', ...]",
  "evidence_assessment": (List) "One entry per cited PubMed ID, in the format of [{'pmid': 'PubMed:XXXXX', '
      stance': 'supports|contradicts|qualifies', 'note': 'one clause on what it establishes or limits'}, ...]",
  "certainty": (String) "high, moderate, or low",
  "gaps": (List) "Unresolved questions the supplied publications do not answer, as plain strings; [] when there
      is none"
}
‘‘‘
```

Listing 5: `Cluster_Function_Analysis` (Reasoning), one call per PPI cluster.

```
[SYSTEM]
You are a biomedical research assistant who helps summarize biological findings.  Please response in the academic
       writing style. This will be used as part of academic report.

[USER]
The input is a list of proteins and their functions:

{formatted_proteins_functions}

# PPI network evidence for this cluster:
{ppi_network_summary}

# Retained PPI edge evidence:
{ppi_edge_evidence}

# Task:
The input list of proteins is a connected cluster in the supplied high-confidence protein association network.
Please summarize the function of the cluster, and how it may contribute to {disease_name}. Use only the supplied
      protein functions, PPI edge scores, connected components, hub degrees, and citations.
Do not claim that an interaction is experimentally direct physical binding unless the supplied evidence type
      explicitly supports that. If PPI evidence is sparse, state that the network evidence is limited under the
      current retained-edge criteria.

# Citation Requirements (IMPORTANT):
```

```
- Place each citation INLINE, in parentheses, immediately AFTER the specific sentence or claim it supports, using
      the exact form ‘(PubMed:XXXXXXXX)‘.
- When a single claim is supported by more than one publication, list them together as ‘(PubMed:XXXXXXXX, PubMed:
     YYYYYYYY)‘.
- Cite the SPECIFIC publication(s) that support each claim; do NOT group all citations at the end of the
     paragraph.
- Only cite PubMed IDs supplied in the input above. Do not invent identifiers.

[Evidence Assessment Requirements: Listing 1]

# Response Format:
- A json code block with the following format:
‘‘‘json
{
  "cluster_function": (String) "The function of the whole cluster",
  "explanation": (String) "A narrative coherent and comprehensive paragraph that explains why this function is
     important, and how it is related to {disease_name}, with inline (PubMed:XXXXXXXX) citations placed right
     after each supported claim",
  "citations": (List) "The complete set of PubMed IDs used inline above (for reference indexing), in the format
     of ['PubMed:XXXXX', 'PubMed:XXXXX', ...]",
  "evidence_assessment": (List) "One entry per cited PubMed ID, in the format of [{'pmid': 'PubMed:XXXXX', '
     stance': 'supports|contradicts|qualifies', 'note': 'one clause'}, ...]",
  "certainty": (String) "high, moderate, or low",
  "gaps": (List) "Unresolved questions the supplied publications do not answer; [] when there is none"
}
‘‘‘
```

Listing 6: `Protein_Function_Analysis` (Reasoning).

```
[SYSTEM]
You are a biomedical research assistant who helps summarize biological findings. Please response in the academic
     writing style. This will be used as part of academic report.

[USER]
 The input is a list of proteins and their functions:
{formatted_proteins_functions}

# Task:
Please identify and summarize the most interesting or relevant protein functions that may have a meaningful
     connection with {disease_name}.
Highlight the proteins that stand out, explain why they are important, and relate them to disease mechanisms
     where possible.
There are citation information for each protein function, please use them to support your analysis.

# Citation Requirements (IMPORTANT):
- Place each citation INLINE, in parentheses, immediately AFTER the specific sentence or claim it supports, using
      the exact form ‘(PubMed:XXXXXXXX)‘.
- When a single claim is supported by more than one publication, list them together as ‘(PubMed:XXXXXXXX, PubMed:
     YYYYYYYY)‘.
- Cite the SPECIFIC publication(s) that support each claim; do NOT group all citations at the end of the
     paragraph.
- Only cite PubMed IDs supplied in the input above. Do not invent identifiers.
- When a function record is marked as described in a species other than human, say so in the explanation rather
     than presenting it as the human function.

[Evidence Assessment Requirements: Listing 1]

# Response Format:
- A json code block with the following format, one entry per protein:
‘‘‘json
{
  "protein_name": {
    "function": (String)"The function of the protein",
    "explanation": (String)"A narrative coherent and comprehensive paragraph that explains why this function is
      important, and how it is related to {disease_name}, with inline (PubMed:XXXXXXXX) citations placed right
      after each supported claim",
    "citations": (List)"The complete set of PubMed IDs used inline above (for reference indexing), in the format
      of ['PubMed:XXXXX', 'PubMed:XXXXX', ...]",
    "evidence_assessment": (List) "One entry per cited PubMed ID, in the format of [{'pmid': 'PubMed:XXXXX', '
      stance': 'supports|contradicts|qualifies', 'note': 'one clause'}, ...]",
    "certainty": (String) "high, moderate, or low",
    "gaps": (List) "Unresolved questions the supplied publications do not answer; [] when there is none"
  }
}
‘‘‘
```

Listing 7: `Gap_Resolution` (Reasoning), one call per follow-up request.

```
[SYSTEM]
```

```
You are a biomedical research assistant. You are given one unresolved question raised during an earlier round of
      analysis, together with publications retrieved specifically to address it. Judge only what the supplied
      publications establish. Please respond in the academic writing style; this will be used as part of an
      academic report.

[USER]
An earlier round of analysis of {disease_name} raised an unresolved question, and publications were retrieved
      specifically to address it.

# Unresolved question:
{gap_description}

# Context in which the question was raised:
{gap_context}

# Retrieved Publications:
{publication_info}

# Task:
Determine what the retrieved publications establish about this question, in the context of {disease_name}.
- If they answer it, state the answer and set ‘resolved‘ to true.
- If they address it only partially, state what is now known, set ‘resolved‘ to false, and record what remains
      open in ‘gaps‘.
- If they do not address it, say so plainly and set ‘resolved‘ to false. DO NOT manufacture an answer, and do not
       restate general background as if it resolved the question.
Write one self-contained paragraph suitable for a follow-up section of a report.

# Citation Requirements (IMPORTANT):
- Place each citation INLINE, in parentheses, immediately AFTER the specific sentence or claim it supports, using
       the exact form ‘(PubMed:XXXXXXXX)‘.
- When a single claim is supported by more than one publication, list them together as ‘(PubMed:XXXXXXXX, PubMed:
      YYYYYYYY)‘.
- Only cite PubMed IDs that appear in the Retrieved Publications above. Do not invent identifiers.
DO NOT use any acronym! Don’t use any abbreviation!

[Evidence Assessment Requirements: Listing 1]

# Response Format:
- A json code block with the following format:
‘‘‘json
{
  "resolution": (String)"A self-contained paragraph reporting what the retrieved publications establish about the
       question, with inline (PubMed:XXXXXXXX) citations placed right after each supported claim",
  "resolved": (Boolean) true when the question is answered by the supplied publications, false otherwise,
  "citations": (List) "The complete set of PubMed IDs used inline above, in the format of ['PubMed:XXXXX', '
      PubMed:XXXXX', ...]",
  "evidence_assessment": (List) "One entry per cited PubMed ID, in the format of [{'pmid': 'PubMed:XXXXX', '
      stance': 'supports|contradicts|qualifies', 'note': 'one clause'}, ...]",
  "certainty": (String) "high, moderate, or low",
  "gaps": (List) "What remains unanswered; [] when the question is fully resolved"
}
‘‘‘
```

Listing 8: `Pathway_Section_Overview` (gated at Synthesis).

```
[SYSTEM]
You are a biomedical research assistant who summarizes enrichment results. You describe only what the table you
      are given contains; the literature is discussed, with citations, in later sections of the report.

[USER]
Table 1 lists the top-ranked pathways and biological processes enriched among the proteins associated with {
      disease_name}. Each row gives the term name, its identifier, its enrichment p-value, its ranking score, the
       number of associated proteins, the term's description, and its evidence type:

{formatted_pathways}

# Task:
Write one concise overview paragraph of Table 1 for a scientific audience.

# What the paragraph may rest on:
Only the rows above. The score is a ranking cost that combines the term's enrichment significance with the
      relevance of the literature retrieved for it, and a lower score ranks higher, so the order need not follow
      the p-values. No publication was supplied to this step: the retrieved literature is discussed, with
      citations, in the per-pathway sections that follow.

# Response Instructions:
- Describe the terms with their identifiers, p-values and numbers of associated proteins, and group terms whose
      names or descriptions share a biological process.
- Say only what each term's name and description state it covers. Many pathway descriptions just repeat the term
      name; for those, say no more than the name says.
```

```
- Do not add mechanisms, disease effects or clinical consequences that no description states.
- Do not rest any statement on the literature, prior studies or established knowledge (for example "consistent
      with the literature", "is known to", "has been implicated", "well-established"). Nothing supplied here can
      support such a statement.
- Do not cite publications or give PubMed identifiers.
- Do not name individual proteins or genes unless the name appears in a term's name or description.
- An enrichment suggests involvement; it does not establish a role. Say that the enrichments point to processes
      that may be involved in {disease_name}.
- Only genuine pathways (Reactome, KEGG, WikiPathways) and Gene Ontology biological processes are ranked here.
      Phenotype, tissue or cell expression, disease-map and regulatory terms are reported separately as
      supporting evidence, so do NOT describe those as pathways.
- DO NOT use any acronym! Don't use any abbreviation!

# Writing Example:
----
Table 1 ranks six enriched terms by combining each term's enrichment significance with the relevance of the
      literature retrieved for it, so the order does not follow the p-values alone: nuclear signaling by ERBB4 (
      REAC:R-HSA-1251985, p = 1.1 × 10−3, 2 associated proteins) ranks first, while the acetylcholine
      neurotransmitter release cycle (REAC:R-HSA-264642, p = 6.5 × 10−4, 2 associated proteins) has the lowest p
      -value and ranks third. Two of the ranked terms concern synaptic transmission, the acetylcholine
      neurotransmitter release cycle and the synaptic vesicle pathway (WP:WP2267, p = 8.4 × 10−3, 2 associated
      proteins), and one, blood based neuroinflammation markers (WP:WP5548, p = 2.3 × 10−3, 2 associated
      proteins), concerns markers of inflammation. The two Gene Ontology biological processes are regulation of
      developmental growth (GO:0048638, p = 1.1 × 10−2, 3 associated proteins), which covers processes that
      modulate the frequency, rate or extent of developmental growth, and intermediate filament organization (GO
      :0045109, p = 1.8 × 10−2, 2 associated proteins), which covers the spatial arrangement of intermediate
      filaments. Each term involves two or three of the associated proteins, so these enrichments point to
      processes that may be involved in Alzheimer disease rather than establish their role.
----
```

Listing 9: `Function_Enrichment_Overview` (Synthesis); system prompt in Listing 3.

```
Please write the overview of the function enrichment analysis following the response structure.

# Data:
Disease: {disease_name}
PathwayName: {pathway_name}
ProteinCount: {protein_count}
PPIFilename: {ppi_filename}
PPIWebURL: {ppi_web_url}
PPINetworkSummary:
{ppi_summary}

# What the overview may rest on:
Only the data above. PPINetworkSummary is computed from the STRING associations retained in the local knowledge
      base. No publication was supplied to this step: what the individual proteins do is covered, with citations,
       in the sections that follow.

# Response Instructions:
- Don't use bullet points. Be coherent and narrative.
- Use the numbers in PPINetworkSummary (protein count, retained edge count, connected proteins, mean score,
      highest-degree proteins) exactly as given.
- Name only proteins listed in PPINetworkSummary. Do not invent partners, hubs, mechanisms, or connectivity.
- A retained edge is a scored STRING association. Do not say that proteins bind, form a complex, or physically
      interact unless the edge evidence listed in PPINetworkSummary says so.
- Do not describe what the proteins do, what the network implies for the disease, or what the literature reports.
- Do not add citations or PubMed identifiers.
- If PPIFilename is provided, embed the PPI figure with Markdown image syntax using exactly the relative path
      given for PPIFilename (same filename and imgs/ prefix). If PPIFilename is empty or None, do not embed an
      image.
- Add the STRING resource as a normal Markdown link [descriptive text](URL), not as an image.
- If there is no STRING network link available, don't mention it.
- If the retained PPI edge count is zero, say only that network evidence is limited under the current retained-
      edge criteria, and do not imply that STRING has no associations for the protein.
- DO NOT use any acronym! Don't use any abbreviation!

# Response Structure:
- Introduce how many proteins are analyzed in the function enrichment analysis.
- One or two sentences summarizing retained PPI edge count, connected proteins, mean score, and highest-degree
      proteins when those values are provided.
- One sentence to mention that we include the PPI network visualization and the link to the STRING webpage when
      available.

# Writting Example:
---
There were 20 proteins associated with Alzheimer disease in the synaptic vesicle exocytosis pathway. Among them,
      14 retained associations connect 9 proteins with a mean score of 0.82, and the highest-degree proteins are
      SNAP25 (degree 5) and SYT1 (degree 4). The figure below shows the association network of these proteins,
      and the interactive network is available on STRING.
![Protein-protein interaction network]({ppi_filename})
```

```
[Open interactive network on STRING]({ppi_web_url})
---
```

Listing 10: `RevisionCoherence` (Synthesis, seed and append); system prompt in Listing 3.

```
You are a scientific writer who helps revise and format scientific reports.

# Task:
Please revise the report draft by:
1. Revise the writing to be more coherent and logical.
2. Adding smooth transition sentences between paragraphs to improve flow and coherence
3. Maintaining all embedded links, tables, and other markdown formatting
4. Keep the original format of the report draft.
5. If there is subsections (e.g. ### A, ### B, ### C), make sure to keep the subsections in the same order, and
     add smooth transition sentences between subsections.

# Report Draft:
{report_draft}

# Response Instructions:
- Keep all original content and insights intact
- Only revise the writing to be more coherent and logical, don't change the content.
- For novel pathways, avoid using conclusive or assertive language—reasoning should be presented cautiously and
     based solely on the information provided, maintain a tentative and low-key tone throughout the explanation.
- Preserve all Markdown formatting: tables must remain valid GitHub-flavored Markdown tables — each row on its
     own line, pipe characters aligned, header separator row preserved, blank line before and after every table.
      Never merge table rows into a single line or strip newlines inside tables.
- Local image files (e.g. imgs/xxx.png) must stay as image syntax: ![alt text](imgs/xxx.png).
- Web URLs (https://...) must ALWAYS use plain hyperlink syntax [text](https://...). NEVER use image syntax
     ![...](https://...) for any web URL, including STRING, PubMed, or any other site.
- Preserve links, figure references, and code blocks.
- Keep literature citations immediately after the paragraph they support; do not move paragraph-level citations
     into a distant section-level citation block.
- The draft contains HTML comments of the form <!--claim {...}--> ... <!--/claim-->. Reproduce every one of them
     CHARACTER FOR CHARACTER, in the same order, each still wrapping the same passage. They are invisible in the
      rendered report and carry the identifier of the finding the passage states. Do not add, remove, reorder,
     reword, merge or split them, and do not move a passage out of the pair that encloses it. You may rewrite
     the prose inside a pair and add transition sentences outside the pairs.
- Insert natural transition sentences between paragraphs
- Return the complete formatted report in markdown
```

Listing 11: Drug context appended to an append-mode `RevisionCoherence` prompt (`--drug-context required`).

```
# Drug-target relations held in the evidence state
The evidence state records the clinical drug candidates below as acting on this pathway's member proteins. Each
     line gives the drug, its action on the target protein, its clinical stage, and the diseases it is approved
     for or studied in. These relations come from Open Targets, not from the literature in the draft, and carry
     no PMIDs.

{drug_context}

Add one passage to this section, after its existing content, that states these relations. For each drug, name the
      drug, the protein it acts on, the direction of that action, and its clinical stage, and name the diseases
     it is approved for or studied in. Relate them to this pathway where the draft supports doing so.
- Describe an approved drug as treating, or approved for, its indications. Describe any drug short of approval as
      studied in them. Never describe a candidate as a treatment.
- Do not add drugs, targets, actions, stages or indications that are not listed.
- Do not attach PMIDs to these statements, and do not move existing citations onto them.
- This passage is the only exception to the instruction not to change the content. Everything already in the
     draft must still be kept, and every claim anchor must be reproduced exactly as instructed above. Place the
     new passage outside every claim anchor pair.
```

Listing 12: `Introduction` (Synthesis); system prompt in Listing 3.

```
You are a professional bioinformatics researcher with expertise in scientific writing.

# Task:
- Please write the *Introduction* section of a bioinformatics report based on the provided context and report
     draft.
- Follow the Response Structure, which is the outline of the introduction.
- Refer the Writing Example to understand the writing style and the structure of the introduction.
- Contain only the introduction text. Do not include any additional explanation or headings.
- Incorporate quantitative evidence and statistics wherever possible to support key claims.
- DO NOT use any acronym! Don't use any abbreviation!

# Writing Example:
```

```
---
Heart failure (Disease code: I9_HEARTFAIL) is a debilitating clinical syndrome characterized by the heart's
      diminished capacity to maintain adequate blood circulation to meet the body's metabolic needs. In the
      analysis of the UK Biobank Pharma Proteomics Project data, circulating levels of 678 plasma proteins were
      found to be significantly associated with heart failure (50,331 controls and 2263 cases), after applying
      Bonferroni correction (P < 1 × 10−5).
Understanding the molecular underpinnings of heart failure reflected from these identified proteins is essential
      for developing targeted interventions and improving patient outcomes. We performed an extensive functional
      enrichment analysis on the 678 proteins to delineate the molecular landscape of heart failure, screening
      genuine biological pathways from Reactome, the Kyoto Encyclopedia of Genes and Genomes, and WikiPathways,
      together with Gene Ontology biological processes. After correcting for multiple testing, we identified
      1,963 significantly enriched terms, and we rank the genuine pathways and biological processes in Table 1;
      complementary phenotype, tissue and cell expression, disease-map, and regulatory associations are
      summarized separately as supporting evidence in Table 2. These pathways provide a window into the complex
      interplay of various biological processes, including signal transduction, immune response, and vascular
      development, which are critical to heart failure progression. In the following sections, we delve into
      specific pathways that have emerged as significant contributors to the disease's molecular framework.
---

# Context:
- Investigated disease: {disease_name}
- Bonferroni correction: P < {level}
- Number of controls: {N_control}
- Number of cases: {N_case}
- Total number of pathways analyzed: {total_pathways_count}
- Total number of proteins analyzed: {total_protein_count}

# Report Draft:
{report_draft}

# Response Structure:
- Begin with a one-sentence clinical and biological definition of the disease, followed by its disease code in
      parentheses.
- State the number of plasma proteins significantly associated with the disease from the UK Biobank Pharma
      Proteomics Project data.
  - Include the number of controls and cases used in the analysis, and the Bonferroni-corrected significance
      threshold (e.g., P < X × 10^−7).
- Motivate the need to understand the biological mechanisms underlying the disease using the identified proteins.
- Mention that a comprehensive functional enrichment analysis was performed on {total_protein_count} proteins,
      using resources including Reactome, KEGG and WikiPathways pathways, and Gene Ontology biological processes
      (Table 1, ranked), with phenotype, tissue/cell expression, disease-map and regulatory associations reported
       separately as supporting evidence (Table 2).
- Report the total number of significantly enriched terms (e.g., "We identified {total_pathways_count}
      significantly enriched terms; the genuine pathways and biological processes are ranked in Table 1").
      Reserve the word "pathway" for genuine pathways and biological processes; do not call phenotype, expression
       or regulatory terms pathways.
- Highlight the types of biological processes revealed (e.g., signaling, immune regulation), as seen in the
      example.
- Conclude with a transition sentence pointing to the following sections, which will elaborate on key pathways
      and interactions.
```