TSASSISTANT: A Human-in-the-Loop Agentic Framework for Automated Target Safety Assessment

Xiaochen Zheng [*1] **Zhiwen Jiang** [*2] **David Tokar** [1] **Melanie Guerard** [3] **Klas Hatje** [2] **Tatyana Doktorova** [1]

Abstract

Target Safety Assessment (TSA) requires systematic integration of genetic, transcriptomic, target homology, pharmacological, and clinical data to evaluate potential safety liabilities of therapeutic targets. This process is labor-intensive and expert-dependent, posing challenges in scalability and reproducibility. We present TSASSISTANT, a human-in-the-loop multi-agent framework that decomposes TSA report generation into a workflow of specialized subagents: *Research Subagents* that each ground and cite a single TSA domain, and *Synthesis Subagents* that integrate findings across domains. Subagents retrieve and synthesize evidence from curated biomedical sources through standardized tool interfaces and produce individually citable, evidence-grounded sections, with behavior shaped by a hierarchical instruction architecture that separates coordination logic from domain expertise and user intent. To complement these soft constraints, programmatic execution hooks and persistent memory stores enforce hard constraints across the workflow, while an interactive refinement loop allows experts to review and revise individual sections with full conversational context preserved across iterations. Rather than a single holistic comparison, we decompose report quality into reproducibility, evidential grounding, task-level accuracy, and controllability under expert oversight, finding high reproducibility and grounding, substantial agreement with the human reference, and net-positive expert-driven refinement.

[*]Equal contribution [1]Predictive Modelling, Roche, Basel, Switzerland [2]Computational Sciences Center of Excellence, Roche, Basel, Switzerland [3]Translational Safety, Roche, Basel, Switzerland. Correspondence to: Xiaochen Zheng <xiaochen.zheng@roche.com>, Zhiwen Jiang <zhiwen.jiang@roche.com>.

Accepted at the 43rd International Conference on Machine Learning AI for Science workshop, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

1. Introduction

Bringing a new medicine from target identification to regulatory approval typically spans more than a decade, with only a small fraction of candidates that enter clinical trials ultimately reaching patients (Olson et al., 2000; Hay et al., 2014; Sun et al., 2022). Among the root causes of late-stage failure, inadequate early characterization of target-related safety liabilities is consistently identified as a primary contributor (Cook et al., 2014; Harrison, 2016). Target Safety Assessment (TSA) is the systematic process of evaluating whether pharmacological modulation of a biological target is likely to produce adverse on-target or off-target effects in humans. It also supports species selection for subsequent animal testing by assessing the similarity between human gene sequences and those of relevant preclinical species, and integrates insights from the competitive landscape for the same target, including known severe adverse events. Performed at the target nomination stage and conducted rigorously, TSA can de-risk programs before resources are committed to preclinical and clinical studies (Plenge et al., 2013).

Despite its importance, the generation of TSA reports remains largely manual, with content composition not standardized and the level of detail and depth of analysis often varying between individuals. Additionally, scientists must synthesize heterogeneous evidence spanning human genetics (Nelson et al., 2015; Buniello et al., 2019), tissue-expression profiles (Uhlén et al., 2015; GTEx Consortium, 2020), pharmacological interaction data (Wishart et al., 2018; Mendez et al., 2019), and adverse-event repositories (Piñero et al., 2020; Banda et al., 2016). This process is time-consuming, difficult to harmonize and communicate across teams, and prone to knowledge gaps when evidence bases grow and portfolios scale (Morgan et al., 2018).

Recent advances in large language model (LLM) agents offer a promising route to augmenting this workflow. While general-purpose LLMs (*e.g.*, GPT, Gemini) can be applied through prompt-based approaches, such usage alone often lacks consistency and reproducibility. Tool-augmented LLMs (M. Bran et al., 2024; Boiko et al., 2023), multi-agent coordination frameworks (Wu et al., 2024; Hong et al., 2024), and retrieval-augmented generation (RAG) (Lewis

et al., 2020) address these limitations by introducing structured, agentic coordination, and have demonstrated that complex, multi-step scientific reasoning is achievable at scale. Autonomous scientific systems such as the AI Co-Scientist (Gottweis et al., 2025) and the AI Scientist (Lu et al., 2024) further demonstrate the potential of agent-based approaches for knowledge-intensive research tasks. However, pharmaceutical workflows with safety endpoints require more than automation: key requirements include human oversight, transparent evidence traceability, and fine-grained expert control (Amershi et al., 2014; Mosqueira-Rey et al., 2023).

We introduce TSASSISTANT, a human-in-the-loop multi-agent framework for TSA. Our principal contributions are: **(i)** a **section-based multi-agent workflow** in which specialized subagents independently ground and cite each TSA domain (Figure 1); **(ii)** a **three-layer instruction architecture** separating coordination logic, domain skills, and runtime user intent; **(iii)** a **programmatic enforcement layer** comprising execution hooks and persistent memory stores, together with domain-engineered tool interfaces that encapsulate multi-step expert analytical workflows rather than serving as simple API wrappers; and **(iv)** an **interactive refinement loop** with tool and agent memory, allowing section-level targeted revision and expert-guided iteration (Figure 2).

2. Background

2.1. Target Safety Assessment

TSA is conducted at the target nomination stage in drug discovery to evaluate potential adverse effects of modulating a biological target in humans, while informing species selection and integrating competitive landscape insights to de-risk programs before preclinical and clinical development (Cook et al., 2014). A comprehensive assessment typically integrates multiple lines of evidence, grounded in an understanding of target biology, including: **(i)** *Genetic evidence*, human genetic variants linking target perturbation to phenotypic consequences (Plenge et al., 2013; Nelson et al., 2015; Buniello et al., 2019); **(ii)** *Transcriptomic evidence*, tissue- and cell-type-specific expression profiles to identify undesired expression in off-target tissues (Uhlén et al., 2015; GTEx Consortium, 2020); **(iii)** *Target homology*, sequence and structural similarity to proteins that could produce off-target pharmacology (Altschul et al., 1997; Jumper et al., 2021); **(iv)** *Pharmacological evidence*, known effects of approved or investigational drugs modulating the same target (Wishart et al., 2018; Mendez et al., 2019; Santos et al., 2017); and **(v)** *Clinical evidence*, signals of adverse events and disease associations from clinical registries and databases (Piñero et al., 2020; Ochoa et al., 2021). Structured frameworks such as the AstraZeneca 5R guidelines (Cook et al., 2014) have codified these requirements, yet execution remains inconsistently documented, expert-dependent, and difficult to scale across large target portfolios.

2.2. LLM-Based Scientific Multi-Agent Systems

LLMs have evolved from single-pass text generators into systems capable of multi-step reasoning, tool use, and interaction with external knowledge sources (Wei et al., 2022; Yao et al., 2022; Lewis et al., 2020), enabling agentic frameworks in which specialized agents collaborate on multi-step tasks (Chen et al., 2024; Kim et al., 2024; Zhou et al., 2025; Wang et al., 2025; Zhang et al., 2026; Zhu et al., 2026).

Multi-Agent Decomposition and Memory. A single agent's context window must simultaneously manage task planning, domain knowledge, tool interactions, and output formatting, which creates capacity bottlenecks and makes individual concerns difficult to maintain independently (Yao et al., 2022; Shinn et al., 2023; Yao et al., 2023). Multi-agent systems mitigate this by decomposing complex tasks across subagents, enabling modular distribution and specialization (Wu et al., 2024; Hong et al., 2024). In more structured systems, the concerns governing each agent (its role, coordination protocol, and domain knowledge) are factored into separate components rather than fused into a single specification. TSASSISTANT inherits this separation of concerns and extends it to the instruction level through its three-layer architecture (Section 3.1).

A complementary challenge is maintaining coherence across agents without exceeding context limits. Generative Agents (Park et al., 2023) address this through a persistent external memory stream from which relevant observations are retrieved for injection into the current agent context, decoupling memory persistence from the LLM's ephemeral context window. TSASSISTANT instantiates this principle through its tool memory and agent memory adapted to a structured workflow (Section 3.2).

Autonomous Scientific Agents. Recent work extends multi-agent architectures to full scientific workflows. The AI Scientist (Lu et al., 2024) automates the research cycle from idea generation to manuscript preparation, and the AI Co-Scientist (Gottweis et al., 2025) uses tournament-based generation, reflection, and ranking to refine biomedical hypotheses. The Virtual Lab (Swanson et al., 2025) couples computational design with experimental validation, while ToolUniverse (Gao et al., 2025) standardizes tool interfaces for non-expert users.

TSASSISTANT differs in several key respects. It targets structured evidence synthesis over a pre-defined report schema rather than open-ended exploration; it operates in pre-clinical safety assessment, where traceability, repro-

ducibility, and human accountability are primary requirements; and it embeds expert-defined constraints and decision frameworks to replicate the workflow of target safety experts. These requirements motivate the programmatic enforcement layer and section-level human-in-the-loop (HITL) design, as described in Section 3.

2.3. Human-in-the-Loop Systems

HITL systems integrate human judgment into automated workflows and decision processes to improve accuracy, safety, and user trust, especially in high-risk and regulated domains (Amershi et al., 2014; Mosqueira-Rey et al., 2023). Rather than treating models as fully autonomous, HITL systems enable iterative interaction in which users can guide, correct, and validate intermediate outputs.

In pharmaceutical research and development, regulatory guidance explicitly requires expert accountability for safety decisions, making full automation structurally inappropriate regardless of performance (Cook et al., 2014). Existing guidance emphasizes that automated systems must support, rather than replace, human judgment. We apply this constraint to TSA, where erroneous assessments carry downstream consequences for patient safety and research resource allocation. TSASSISTANT instantiates these principles through a section-level HITL design described in Section 3.3.

3. TSASSISTANT: System Design

Given a biological target identifier (*e.g.*, gene ID or symbol, UniProt accession) and optional project context including disease therapeutic area and compound modality, TSASSISTANT produces a structured TSA report comprising individually citable, evidence-grounded sections. The framework decomposes report generation into a workflow of specialized subagents: *Research Subagents*, each responsible for one TSA domain, followed by *Synthesis Subagents* that integrate findings across domains (Figure 1). Each domain draws on one or more of the five lines of evidence (Section 2.1) rather than mapping to them one-to-one, and the Synthesis Subagents integrate across completed domains without issuing new retrievals.

A central design challenge is that LLM agent behavior is governed by prompts, which function as *soft constraints*: they bias model outputs statistically but do not reliably guarantee correctness, reproducibility, or evidence fidelity. In monolithic single-agent approaches, errors that enter the shared context risk compounding across the report, and multi-agent decomposition alone does not resolve this without explicit verification mechanisms (Kim et al., 2025). These observations motivate a complementary architecture in which a *hierarchical instruction architecture* (Section 3.1) organizes agent behavior through layered prompts, a *programmatic enforcement layer* (Section 3.2) of tool interfaces, execution hooks, and context management provides hard constraints, and an *interactive refinement loop* (Section 3.3, Figure 2) embeds expert oversight at the section level.

3.1. Hierarchical Instruction Architecture

Agent behavior in TSASSISTANT is governed by a three-layer instruction hierarchy that applies *separation of concerns* to prompt engineering (Hong et al., 2024; Wu et al., 2024), factoring instructions by change rate and stakeholder.

Layer 1: System Prompts. Define agent roles, coordination logic, and output structure, shared immutably across all targets.

Layer 2: Domain Skill Modules. Encode section-specific research and tool-use strategies, evidence weighting, and writing guidelines as independently versioned, composable modules.

Layer 3: Runtime User Instructions. Inject project-specific constraints at highest priority, overriding lower-layer defaults.

At runtime the layers compose with higher layers taking precedence, allowing domain experts to evolve research guidance independently of coordination logic, and users to customize individual assessments without modifying either reusable layer.

3.2. Programmatic Enforcement Layer

Tool Interfaces. Each subagent operates through MCP-standardized tool interfaces (Anthropic, 2024) that implement complete analytical subroutines rather than serving as data-retrieval endpoints alone. The guiding principle is to decompose expert research workflows into atomic, deterministic operations implemented in auditable code: the tool layer fixes each analytical procedure (data acquisition, computational analysis, rule-based interpretation) in code rather than delegating it to the LLM's stochastic interpretation of raw data. Tool pipelines span multiple levels of abstraction, from low-level operations such as ontology traversal, sequence-level computation, and statistical aggregation, to higher-level domain logic encoding acceptance criteria and multi-source cross-referencing strategies that would otherwise demand expert judgment at each invocation. Every returned record is automatically indexed in the tool's paired memory store with full provenance metadata, enabling systematic comparison against expert reference standards (Section 4). The framework additionally integrates internal proprietary tools, withheld for confidentiality. A complete catalog of the public data sources backing these tools is provided in Section A.

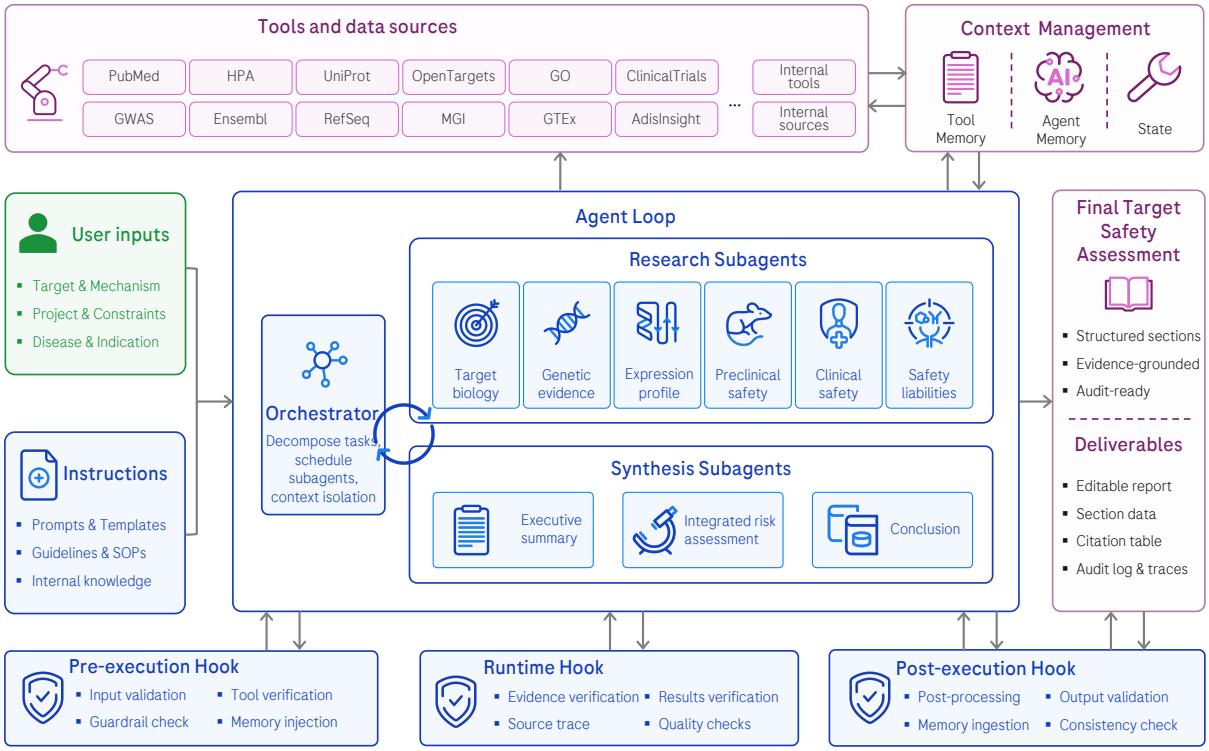

Figure 1. **Hierarchical agent architecture of TSASSISTANT.** An *Orchestrator* decomposes the assessment into *Research Subagents*, each covering one TSA domain, and *Synthesis Subagents* that integrate findings across domains. Execution hooks and three persistent context stores (tool memory, agent memory, execution state) provide programmatic enforcement; all subagents access curated biomedical data sources through standardized MCP tool interfaces. Mechanism detail in Section 3.2.

Execution Hooks. Execution hooks wrap each subagent's lifecycle (Figure 1): before execution, they inject compressed upstream findings and serialize subagent launches; after execution, they validate inline citations against retrieved evidence via natural-language inference and compress the output for downstream consumption; during execution, they cross-reference each newly retrieved record against accumulated evidence to detect contradictions.

Context Management. Each subagent executes in an isolated context to prevent cross-contamination and context window overflow. Three persistent stores bridge the isolated agents.

Tool memory records every tool-call output with provenance metadata (invoking subagent, tool, query, pipeline stage), each assigned a globally unique identifier with full CRUD support, enabling downstream lookup and cross-referencing without re-execution.

Agent memory stores compressed factual summaries of completed sections and selectively injects them into downstream subagents via a predefined dependency graph.

Execution state is persisted for resumability: an interrupted run resumes from the last completed section, preserving accumulated tool and agent memory.

3.3. Human-in-the-Loop Refinement

After initial report generation, TSASSISTANT enters an interactive refinement loop (Figure 2) in which experts and the system iterate on the draft. Users can edit section content directly, append new information, upload additional sources, or re-invoke the responsible subagent to revise a specific section. A conversational memory store carries context across iterations, so users do not need to restate background on each interaction, and captured corrections progressively adapt retrieval constraints and prompt templates to team conventions and project context.

A deliberate design choice is that human validation occurs at the *section* level rather than only at report completion. This enables early error detection and reduces the cost of late corrections, since a flawed upstream section (*e.g.*, an incorrect knockout phenotype) would otherwise propagate into downstream synthesis sections before the expert has an opportunity to intervene.

More broadly, TSASSISTANT adopts a hybrid mode in which the LLM handles research and drafting while domain experts retain final decision authority (Amershi et al., 2014; Mosqueira-Rey et al., 2023). This division reflects both regulatory requirements for expert accountability in safety

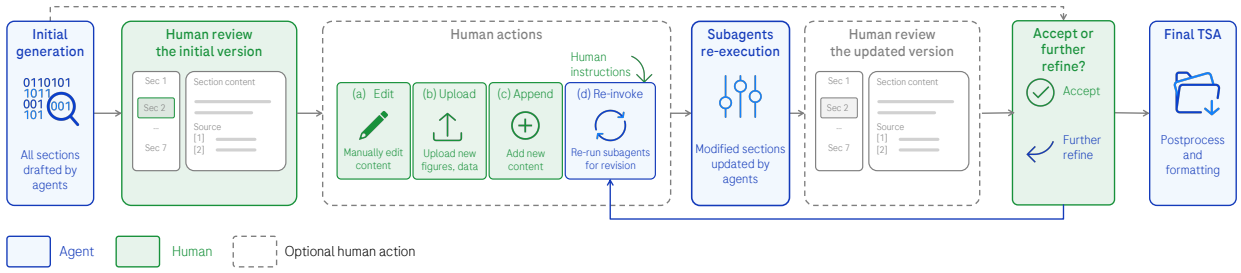

Figure 2. **Interactive refinement loop in TSASSISTANT.** After initial section generation, the user reviews each section and may: **(a)** manually edit content, **(b)** append new information, **(c)** upload additional sources or graphics, or **(d)** re-invoke the subagent for targeted revision. Conversational memory preserves context across iterations; user feedback progressively adapts retrieval strategies and prompt templates.

decisions and the practical limitations of current LLMs in resolving ambiguous evidence (Gabriel et al., 2024).

4. Evaluation Framework

Comprehensive evaluation of TSA report quality can be approached from two directions: externally, by comparing system outputs against expert-authored reference reports across dimensions such as factual consistency, evidence completeness, and structural alignment; and intrinsically, by measuring properties of the generated report itself, including self-consistency across independent runs and traceable grounding of individual claims. External evaluation is, however, inherently difficult to scale: evidence selection and interpretation involve expert judgment, and two independently authored reports on the same target may legitimately differ in scope and emphasis while both being scientifically valid.

All system evaluations are conducted across 35 drug targets spanning diverse therapeutic areas and mechanism classes; the sequence homology benchmark uses a 22-target subset for which expert-authored reference analyses are available. We evaluate TSASSISTANT along two complementary axes. *Automatic intrinsic* evaluations (Section 4.1) probe report-level reliability without any human reference, through claim-level self-consistency across independent runs (Section 4.1.1) and sentence-level evidential grounding (Section 4.1.2). *Evaluations with human reference and oversight* (Section 4.2) capture the two ways human expertise enters the loop: a task-level benchmark (Section 4.2.1) compares the agent's ortholog selection against analyses performed by human computational biologists as a frozen reference, and an expert-in-the-loop refinement test (Section 4.2.2) measures whether interactive expert oversight produces precise, beneficial revisions or introduces collateral damage. Together, our evaluation strategy decomposes report quality into independently measurable properties: reproducibility, evidential grounding, task-level accuracy, and controllability under expert oversight.

4.1. Automatic Intrinsic Evaluation

4.1.1. CLAIM-LEVEL SELF-CONSISTENCY

The integrated risk assessment produced by TSASSISTANT consolidates findings into a structured table of 15 standardized toxicity endpoints (Table 2 in Section C) aligned to MedDRA System Organ Classes (Brown et al., 1999). We formalize self-consistency (Manakul et al., 2023; Kuhn et al., 2023; Lin et al., 2024) measurement as a *structured concordance problem*: given two independent runs on the same target, how well do they agree on the specific evidentiary claims supporting each safety endpoint? Holistic text similarity fails to capture this, as two reports may use different phrasing for the same observation or cite the same phenotype from different biological contexts.

Following FActScore (Min et al., 2023) and SAFE (Wei et al., 2024), we decompose each endpoint's findings into structured atomic claims[1] $c = (\mathit{observation}, \mathit{species})$, where *species* encodes the biological context in which the observation was made. Unlike prior work where each atomic fact is a plain-text statement, toxicological evidence is inherently structured: the same phenotype observed in different species carries different translatability for human risk. Pairs of claims from two independent runs are scored by an LLM-as-judge binary semantic similarity (Zheng et al., 2023; Liu et al., 2023; Jiang et al., 2025), scaled by a species-context weight ϕ_s reflecting biological translatability informed by regulatory toxicology practice (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2011) (same species: 1.0, same species group: 0.75, cross-group: 0.5). Scores are aggregated via symmetric greedy matching (Zhang et al., 2020) into Precision, Recall, and Claim-F1, where neither run is treated as ground truth (full metric definitions in Section B).

[1] In the safety and toxicology context, each atomic claim corresponds to a single empirical observation: either a specific preclinical phenotype (*e.g.*, a knockout mouse phenotype or histopathological finding) or a clinical adverse event (*e.g.*, an SAE or TEAE from a clinical trial). We retain the term "claim" to align with the NLP evaluation literature.

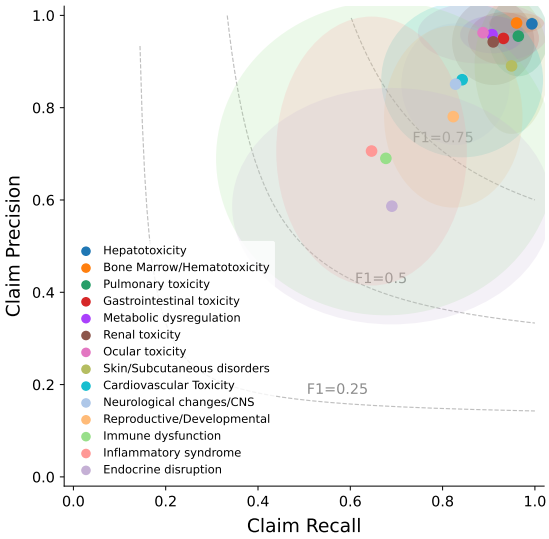

Figure 3. **Claim-Level Self-Consistency by Safety Endpoint.** Symmetric Precision and Recall (Section 4.1.1) for each of the 15 endpoints, aggregated over 45 run pairs per target across 35 targets. The shaded ellipsoidal regions around each endpoint are axis-aligned and span ± 1 sample standard deviation of Precision and Recall computed across all 35×45 (target, run-pair) observations per endpoint. Tighter ellipses indicate that an endpoint's self-consistency is stable across both runs and biological contexts; broader ellipses indicate substantial variation, typically correlating with evidence sparsity or domain-specific phenotype ambiguity. F1 contours are shown at 0.25, 0.50, and 0.75.

Experimental setup. We evaluate across 35 drug targets, each assessed with 10 independent TSASSISTANT-generated reports. For each target, the integrated risk assessment table is parsed to extract the structured endpoint findings, aligned to the 15 predefined endpoints. Claim extraction uses Claude Opus 4.6; similarity assessment uses Claude Opus 4.6 [2]. Claim pairs that fail the species context filter ($\phi_s = 0$) are excluded from the matching rather than scored as zero, as direct translatability between these evidence types cannot be assumed. We score all $\binom{10}{2} = 45$ pairs of TSASSISTANT runs per target using an all-pairs design rather than designating a single run as ground truth (Manakul et al., 2023; Kuhn et al., 2023), yielding a symmetric self-consistency estimate.

Results. Figure 3 reports claim-level self-consistency across the 15 safety endpoints. Because neither run is treated as ground truth, Precision and Recall here quantify how many claims in one run have a species-compatible match in the other and vice versa. Table 4 further breaks down all the atomic claim pairs by species relation and scoring rule.

Most of the endpoints achieve both Precision and Recall above 0.75, with Claim-F1 scores above the 0.75 contour, supporting the conclusion that TSASSISTANT produces reproducible endpoint-level findings across independent runs. Several endpoints, including hepatotoxicity, hematotoxicity, pulmonary toxicity, and gastrointestinal toxicity, approach Claim-F1 values close to 1.0, consistent with the structured and well-characterized nature of evidence sources for these endpoints, indicating that independent runs converge on the same atomic claims with matching species context across diverse targets.

A smaller number of endpoints, particularly endocrine disruption, inflammatory syndrome, immune dysfunction, exhibit lower Claim-F1 driven by reduced Precision and/or Recall. We attribute this to greater mechanistic heterogeneity in the underlying evidence base, sparser per-target data density, and increased ambiguity in phenotype terminology and adverse event classification in these domains, all of which broaden the space of plausible findings any single run may surface and therefore reduce claim-level overlap between independent runs. Additional retrieval guidance, refined skill modules, or more targeted human review in the refinement loop (Section 3.3) are most warranted in subsequent iterations.

4.1.2. SENTENCE-LEVEL GROUNDING

To evaluate the factual reliability of generated TSA reports, we introduce the *evidential support rate* (ESR), which measures the proportion of factual sentences supported by traceable evidence, serving as an inverse indicator of hallucination.

Metric. Each sentence in a report is classified by an LLM judge (Claude Opus 4.6) into one of seven categories: *cited* (has an inline reference), *tool-grounded* (verified by MCP tool call results), *inferable* (derivable from preceding grounded context), *common knowledge*, *procedural* (methodology or guidance), *in-memory* (supported by subagent memory from prior sections), or *unsupported* (an ungrounded factual claim and potential hallucination). The ESR is then computed as the fraction of factual sentences (excluding procedural classifications) that are not unsupported.

Experimental setup. Figure 4 compares the ESR across three architectural conditions (single agent, subagents, and subagents with memory), averaged over 35 drug targets.

Results. The single agent scores highest on early sections (Target Biology, Expression Profile), where its context window is still short and manageable. As it generates more

[2] A single pairwise LLM-as-a-judge evaluation requires $35 \times 45 \times 15 \times \bar{n}$ API calls, where $\bar{n}$ is the average number of claim pairs per endpoint ($\bar{n} \approx 10$, range $[0, 20]$), i.e., roughly 2.3×10^5 calls. We therefore restrict our experiments to a single judge model and leave cross-LLM comparison to future work.

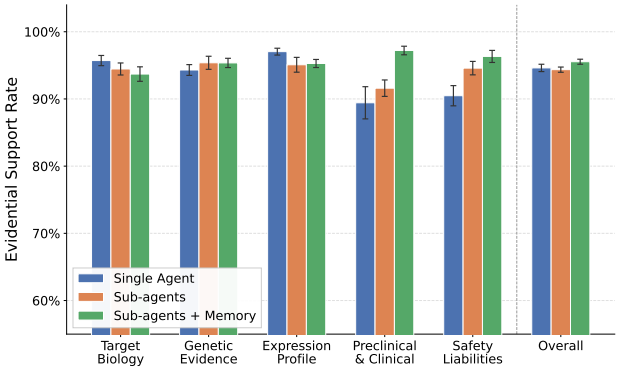

Figure 4. Evidential support rate by section. Error bars indicate standard error.

sections and the accumulated context grows, its ESR drops to ∼90% on Preclinical & Clinical and Safety Liabilities, suggesting that the model struggles to faithfully attend to relevant information buried in a long conversation history. Subagents sidestep this issue by starting each section with a fresh context. With memory enabled, downstream subagents receive prior findings directly, allowing them to cite or build on earlier evidence. This combination yields the highest overall ESR and reduces the late-section degradation observed in the single-agent condition, consistent with both task decomposition and cross-section evidence propagation contributing to grounding in TSA report generation.

4.2. Evaluation with Human Reference and Oversight

4.2.1. Task-Level Benchmark: Sequence Homology

Target homology is one of the five lines of evidence of TSA (Section 2.1), and its computational core is ortholog selection: for each preclinical species, identify the protein sequence most likely to recapitulate the target's biology. We benchmark the agent's ortholog selection against expert-curated ground truth from 22 TSA programs, asking whether it is reliable enough to serve as the homology subagent in deployment.

The task spans 14 ortholog databases with biological and database-driven edge cases (divergent isoforms, long lineage-specific extensions, partial annotations). The agent works within a dedicated memory environment that references sequences by protein ID rather than transcribing full sequences, and applies a deterministic pre-filter that discards candidates differing from the human canonical sequence by more than 20% in length or showing more than a 60% mismatch in pairwise alignment. When no suitable candidate is found, the agent excludes the species and reports every database it searched together with the rejection reason, so the user can override (full procedure in Section F).

For each species both human and agent included in their final results, we compare protein sequences of the orthologs by pairwise alignment rather than by identifier (the same sequence may be retrieved from different databases). Based on the percent sequence identity of this alignment, we classify each comparison into three tiers: an *exact match* (100% identity), where the agent and human selected the same protein sequence; a *partial match* ($\geq 97\%$ identity), where the sequences differ only by minor isoform or annotation variation that is unlikely to alter the homology conclusion; and a *mismatch* ($< 97\%$ identity), where the agent selected a substantively different sequence from the human reference. We define a false negative as a case where the agent rejected a candidate that matched the human's selection.

Results. The agent reviewed 135 sequences across the 22 benchmark cases. Setting aside 11 sequences for which it found no valid protein in any database (9 correct rejections, 2 false negatives), the remaining 124 yielded 86 exact matches (100% identity; 69.4%), 21 partial matches ($\geq 97\%$; 16.9%), and 17 mismatches ($< 97\%$; 13.7%).

Both false negatives trace to the same protein retrieved from two distinct databases; the agent anchored to a substantially longer incorrect human isoform, causing both candidates to be discarded by the deterministic ±20% length pre-filter.

The 17 mismatches arise from three complementary causes. First, in several cases the agent's reasoning correctly identified a candidate as a likely incorrect isoform yet still accepted it, a failure mode addressable through prompt tuning and tool-schema refinement. Second, the human analysts drew on specialized databases not yet accessible to the agent, yielding isoforms the agent could not reach, a limitation mitigable by extending the database catalog. Third, a subset of cases required non-standard expert procedures the agent is not currently equipped to replicate, such as excising a long C-terminal extension absent from the human reference and running a targeted similarity search on the remainder to recover the correct ortholog. The first two are tractable engineering gaps addressable through prompt tuning, tool-schema refinement, and an extended database catalog; the third remains an open research problem flagged for human review in deployment.

4.2.2. HITL Refinement Regression Test

Whereas Section 4.2.1 treats the human contribution as a frozen reference, the refinement loop itself (Section 3.3) places an expert in active control of the system. The question we ask here is: when an expert injects a targeted question to improve a specific section, does the agent produce a precise improvement, or does it introduce collateral damage elsewhere in the document (Laban et al., 2026)?

We diff each report section pre/post a single HITL interven-

tion at the sentence level. A domain expert then labels every changed sentence as **B** (beneficial: correct new information or removal of incorrect content), **N** (neutral: rephrasing without semantic change), or **H** (harmful: factual error, hallucination, or loss of correct information). From these labels we define *net improvement* $= (\mathrm{B} - \mathrm{H}) / (\mathrm{B} + \mathrm{N} + \mathrm{H})$ as the primary metric, alongside benefit rate (B / total), harm rate (H / total), and response rate (questions addressed / total questions).

Twelve target-specific refinement requests are designed for a selected list of targets (Figure 7), spanning the question dimensions of accuracy, completeness, prioritization, and consistency. Each question is delivered as a single intervention on an independent replicate, so that its effect is measured in isolation from other refinements. We adopt a single-reviewer protocol because the B/N/H labels are anchored to verifiable criteria rather than subjective judgment: beneficial requires the added or modified content to be factually correct against publicly available data sources (genetic databases, expression atlases, clinical trial registries), harmful requires the introduction of a factual error or loss of correct information, and neutral is restricted to rephrasing without semantic change. This fact-checking structure, combined with the domain expertise required for verification, makes the labeling task suitable for a single qualified reviewer.

Table 1. HITL refinement outcomes stratified by question type (top) and report section (bottom). n: total changed sentences per stratum.

Stratum	n	**Benefit**	**Neutral**	**Harm**	**Net**
By question type					
Accuracy	20	85.0%	15.0%	0.0%	+85.0%
Completeness	274	79.9%	17.5%	2.6%	+77.4%
Consistency	63	47.6%	38.1%	14.3%	+33.3%
Prioritization	113	66.4%	25.7%	8.0%	+58.4%
By report section					
Clinical Safety	101	82.2%	15.8%	2.0%	+80.2%
Expression Profile	138	64.5%	28.3%	7.2%	+57.2%
Genetic Evidence	135	82.2%	14.8%	3.0%	+79.3%
Risk Assessment	96	60.4%	30.2%	9.4%	+51.0%
Overall	470	72.6%	22.1%	5.3%	+67.2%

Results. Across $n = 470$ changed sentences over 12 interventions, the refinement loop produced $+67.2\%$ net improvement (72.6% B, 22.1% N, 5.3% H), with every intervention triggering at least one change (Table 1).

Completeness ($n = 274$, $+77.4\%$ net, 2.6% H) dominates the changes, indicating that the refinement loop adds new beneficial content far more often than it fabricates. *Prioritization* ($+58.4\%$ net, 8.0% H) shows higher harm when re-ranking demotes content the reviewer judged equally salient. *Consistency* is the only category with non-trivial harm (14.3% H, $+33.3\%$ net): asking the agent to reconcile statements across sections is meaningfully harder than retrieving, locally re-ranking, or reorganizing content already in scope. *Accuracy* was probed by only a single intervention ($n = 20$ changed sentences, $+85.0\%$ net) and is reported as suggestive rather than conclusive.

Clinical Safety ($+80.2\%$) and *Genetic Evidence* ($+79.3\%$) refine cleanly, while *Risk Assessment* ($+51.0\%$ net, 9.4% H) is unsurprisingly the noisiest, since it is the only section whose content is integrative rather than evidence-local.

5. Discussion

TSASSISTANT is designed to explore whether a hierarchical multi-agent architecture with human-in-the-loop design can accelerate TSA report drafting under expert oversight. Across the four evaluations, TSASSISTANT produces reproducible endpoint-level claims (Section 4.1.1), achieves high sentence-level grounding under subagent decomposition with cross-section memory (Section 4.1.2), shows substantial agreement with a human reference on the ortholog benchmark (Section 4.2.1), and yields net-positive expert-driven refinement (Section 4.2.2). Anticipated challenges in deploying systems of this kind include sensitivity to ambiguous or underspecified target queries, variability in evidence prioritization across sections, reconciliation of conflicting evidence across databases, and adaptation to targets with sparse literature coverage. TSASSISTANT incorporates retrieval constraints, mandatory citation enforcement, and embedded human validation checkpoints intended to address these. Future work will further explore automatic evidence conflict resolution, tighter integration with internal proprietary databases, and extension to related early-safety workflows such as genotoxicity and cardiotoxicity assessment.

More broadly, target safety assessment is a resource-intensive step in drug discovery, and TSASSISTANT is, to our knowledge, among the first frameworks specifically tailored to the needs of toxicologists and pharmaceutical scientists. By reducing the mechanical burden of evidence synthesis and report drafting while preserving expert oversight through section-level refinement and progressive personalization, TSASSISTANT aims to contribute to a broader vision of AI-augmented pharmaceutical safety science where agentic AI and human expertise are complementary.

Author Contributions

Doktorova initiated, directed, and supervised the project; Doktorova, Guerard, Hatje, Jiang, and Zheng formulated the problem; Zheng conceived the framework, designed the system architecture and evaluation methodology, and implemented the entire system; Zheng, Jiang, and Tokar de-

signed the evaluation metrics; Zheng and Tokar conducted the experiments; Zheng, Jiang, and Tokar performed the analysis; Doktorova, Guerard, Hatje, and Jiang translated domain expertise into tool design specifications; Doktorova, Zheng, Jiang, and Hatje curated the data sources and provided data-access resources; Jiang, Zheng, and Doktorova wrote the manuscript; all authors provided expert validation, contributed to user requirements and quality control, participated in discussions, and revised the manuscript.

Funding Disclosures

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgements

The authors thank colleagues across Roche for their contributions, including technical input, scientific discussions, and feedback on system design and evaluation. In particular, the authors thank Yexiang Cheng and Alvaro Serra for their contributions to evaluation metric design, experiments, and analysis. Platform and agentic workflow development were carried out within Pharmaceutical Sciences (Predictive Modelling) and the Computational Sciences Center of Excellence (Data and Digital Catalysts, Computational Biology and Medicine). Scientific contributions, including target safety assessment expertise and validation, were provided by Pharmaceutical Sciences (Predictive Modelling, Translational Safety) and the Computational Sciences Center of Excellence (Computational Biology and Medicine). The authors also acknowledge support in data engineering, infrastructure, and tool integration that enabled the development and deployment of the system.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

Anthropic. Model context protocol. https://modelcontextprotocol.io, 2024.

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

Baldarelli, R. M., Smith, C. L., Ringwald, M., Richardson, J. E., and Bult, C. J. Mouse genome informatics: an integrated knowledgebase system for the laboratory mouse. *Genetics*, 227(1):iyae031, 2024.

Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B., and Shah, N. H. A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific data*, 3(1):1–11, 2016.

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., et al. An overview of Ensembl. *Genome Research*, 14(5): 925–928, 2004.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

Brown, E. G., Wood, L., and Wood, S. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20 (2):109–117, 1999.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

Chen, J., Saha, S., and Bansal, M. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, 2024.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G., and Pangalos, M. N. Lessons learned from the fate of astrazeneca's drug pipeline: a five-dimensional framework. *Nature reviews Drug discovery*, 13(6):419–431, 2014.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., et al. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, 2022.

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., et al. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*, 2024.

Gao, S., Zhu, R., Sui, P., Kong, Z., Aldogom, S., Huang, Y., Noori, A., Shamji, R., Parvataneni, K., Tsiligkaridis, T., et al. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*, 2025.

Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews,

L., Gong, C., et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D419–D426, 2022.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369 (6509):1318–1330, 2020.

Harrison, R. K. Phase II and phase III failures: 2013–2015. *Nature reviews Drug discovery*, 15(12):817–818, 2016.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51, 2014.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH harmonised tripartite guideline S6(R1): Preclinical safety evaluation of biotechnology-derived pharmaceuticals. Technical report, ICH, 2011. https://www.ich.org/page/safety-guidelines.

Jiang, Y., Chen, C., Wang, S., Li, F., Tang, Z., Mervak, B. M., Chelala, L., Straus, C. M., Chahine, R., Iii, S. G. A., and Tan, C. CLEAR: A clinically grounded tabular framework for radiology report evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Kanehisa, M. and Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1): 27–30, 2000.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D483–D489, 2023.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., et al. A single–cell type transcriptomics map of human tissues. *Science advances*, 7(31):eabh2169, 2021.

Kim, Y., Park, C., Jeong, H., Chan, Y. S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., and Park, H. W. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

Kim, Y., Gu, K., Park, C., Park, C., Schmidgall, S., Heydari, A. A., Yan, Y., Zhang, Z., Zhuang, Y., Malhotra, M., et al. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Laban, P., Schnabel, T., and Neville, J. Llms corrupt your documents when you delegate. *arXiv preprint arXiv:2604.15597*, 2026.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., and Riedel, S. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6: 525–535, 2024.

Manakul, P., Liusie, A., and Gales, M. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 47 (D1):D930–D940, 2019.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Morgan, P., Brown, D. G., Lennard, S., Anderton, M. J., Barrett, J. C., Eriksson, U., Fidock, M., Hamren, B., Johnson, A., March, R. E., et al. Impact of a five-dimensional framework on r&d productivity at astrazeneca. *Nature reviews Drug discovery*, 17(3):167–181, 2018.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P. C., Li, M. J., Wang, J., et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic acids research*, 49 (D1):D1302–D1310, 2021.

Ochoa, D., Hercules, A., Key, M., Chandran, D., Sheridan, J., Dunham, I., Gunes, A., Morales, J., McDonagh, E. M., Oprea, T. I., et al. The next-generation open targets platform: Reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1):D1353–D1359, 2023.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016.

Olson, H., Betton, G., Robinson, D., Thomas, K., Monro, A., Kolaja, G., Lilly, P., Sanders, J., Sipes, G., Bracken, W., et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory toxicology and pharmacology*, 32(1):56–67, 2000.

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., et al. Expression atlas update: From tissues to single cells. *Nucleic Acids Research*, 48 (D1):D77–D83, 2020.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pp. 1–22. ACM, 2023.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L. I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

Plenge, R. M., Scolnick, E. M., and Altshuler, D. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery*, 12(8):581–594, 2013.

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T. I., and Overington, J. P. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, 2017.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Advances in neural information processing systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023.

Sun, D., Gao, W., Hu, H., and Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, 2022.

Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E., and Zou, J. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 2023.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. Tissue-based map of the human proteome. *Science*, 347(6220):1260419, 2015.

UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1):D523–D531, 2023.

Wang, Z., Zhu, Y., Zhao, H., Zheng, X., Sui, D., Wang, T., Tang, W., Wang, Y., Harrison, E., Pan, C., et al. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pp. 2250–2261, 2025.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

White, J. PubMed 2.0. *Medical Reference Services Quarterly*, 39(4):382–387, 2020.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1): D1074–D1082, 2018.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First conference on language modeling (CoLM)*, 2024.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations (ICLR)*, 2022.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems (NeurIPS)*, 36: 11809–11822, 2023.

Zhang, H. G., Eckmann, P., Miao, J., Mahon, A. B., and Zou, J. The virtual biotech: A multi-agent ai framework for therapeutic discovery and development. *bioRxiv*, pp. 2026–02, 2026.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems (NeurIPS)*, 36:46595–46623, 2023.

Zhou, Y., Song, L., and Shen, J. Mam: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25319–25333, 2025.

Zhu, Y., He, Z., Hu, H., Zheng, X., Zhang, X., Wang, Z., Gao, J., Ma, L., and Yu, L. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.

A. Data Sources

The data sources backing the tool interfaces span the five lines of evidence described in Section 2.1. PubMed (White, 2020) provides primary literature evidence across all sections. Other sources are domain specific, spanning genetics, genomics, transcriptomics, target homology, pharmacological and clinical evidence. We use Ensembl (Birney et al., 2004; Cunningham et al., 2022) as primary source to get genomic coordinates, transcript and protein annotations, cross-species orthologues, protein domain information, germline and somatic variants with phenotype associations, and regulatory feature annotations. NCBI RefSeq (O'Leary et al., 2016) complements this with curated reference sequences for genes and transcripts essential for standardized genomic annotation. UniProt (UniProt Consortium, 2023) provides curated protein sequences, functional annotations, and tissue expression data essential for evaluating both on-target and off-target safety liabilities.

Functional characterization is further supported by Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium, 2023), which supplies standardized molecular functions, biological processes and cellular components, enabling systematic characterization of target biology and potential mechanism-based toxicities. Pathway-level context is provided by KEGG (Kanehisa & Goto, 2000; Kanehisa et al., 2023) and Reactome (Gillespie et al., 2022), which map targets to metabolic, signaling, and disease pathways critical for understanding downstream physiological consequences of target modulation.

Human genetic and clinical evidence is integrated from multiple sources. The NHGRI-EBI GWAS Catalog (Buniello et al., 2019) aggregates genome-wide association signals linking variants to phenotypes. ClinVar (Landrum et al., 2018) supplies curated variant–disease interpretations including pathogenicity classifications. The Human Protein Atlas (Uhlen et al., 2010; Karlsson et al., 2021) provides tissue- and cell-type-level protein and RNA expression profiles. GTEx (GTEx Consortium, 2020) supplies tissue-specific expression and eQTL data across human tissues. The EMBL-EBI Expression Atlas (Papatheodorou et al., 2020) offers curated RNA-seq expression profiles across diseases, cell types, and developmental stages.

To explore translatability, we use Open Targets (Ochoa et al., 2021; 2023), which integrates genetic, genomic, and clinical evidence, to validate disease–target associations and extract reported adverse events from clinical studies and approved drugs. The Mouse Genome Informatics database (MGI) (Baldarelli et al., 2024) provides phenotype ontology annotations, knockout phenotype data and human-mouse disease connections to facilitate exploration of candidate genes and investigation of phenotypic similarity between mouse models and human patients. ClinicalTrials.gov[3] supplies clinical study records including safety endpoints and adverse event reports from interventional trials. AdisInsight[4] provides competitive intelligence, including known severe adverse events associated with drugs modulating the same target.

B. Claim-Level Self-Consistency: Full Metric Definition

This appendix provides the complete metric definition summarized in Section 4.1.1.

Step 1: Claim extraction. Each safety endpoint's finding text is decomposed into atomic claims by an LLM with toxicology-informed system prompt. Each claim is a structured pair $c = (\textit{observation}, \textit{species})$, where *species* encodes the biological context in which the observation was made.

Step 2: Claim pair scoring. For two claims c_1 and c_2 from the two runs being compared, we define:

$$\text{score}(c_1, c_2) = \text{sim}(c_1.\textit{obs},\ c_2.\textit{obs}) \cdot \phi_s(c_1, c_2) \tag{1}$$

where $\text{sim}(\cdot,\cdot) \in \{0, 1\}$ is an LLM-as-judge binary similarity score (Zheng et al., 2023; Liu et al., 2023; Jiang et al., 2025), which is scaled by ϕ_s, a species context matching function, encoding domain-specific constraints on biological translatability:

[3] https://clinicaltrials.gov
[4] https://adisinsight.springer.com

$$\phi_s(c_1, c_2) = \begin{cases} 1.0 & \text{same species,} \\ 0.75 & \text{same species group,} \\ 0.5 & \text{different species group,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The four-level weighting reflects regulatory toxicology practice on inter-species translatability (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2011); species-group definitions are in Table 3 (Section C). Contradictory claims (*e.g.*, "dose-related neutropenia" *vs.* "no significant hematotoxicity") are explicitly scored as zero by the LLM judge. Same-species matches receive full credit ($\phi_s = 1.0$): hepatic fibrosis observed in two independent rat studies is the same evidence type. Same-group matches receive partial credit ($\phi_s = 0.75$), as inter-species differences within a group indicate that findings are convergent but not equivalent. Cross-group matches receive reduced credit ($\phi_s = 0.5$): the same phenotype observed in rodent preclinical studies and in human clinical trials represents qualitatively different evidence, as animal findings are not always predictive of human outcomes due to inter-species differences in gene expression, target abundance, receptor functionality, signaling pathways, metabolism, tissue distribution, and compensatory physiological mechanisms.

Step 3: Symmetric optimal matching. We adopt the greedy optimal matching formulation from BERTScore (Zhang et al., 2020), which computes precision and recall by independently finding the best match for each element, and combine it with SAFE's F1 aggregation:

$$\text{Recall} = \frac{1}{|C_1|} \sum_{c \in C_1} \max_{c' \in C_2} \text{score}(c, c'), \tag{3}$$

$$\text{Precision} = \frac{1}{|C_2|} \sum_{c' \in C_2} \max_{c \in C_1} \text{score}(c', c), \tag{4}$$

$$\text{Claim-F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

The symmetric formulation distinguishes our approach from FActScore and SAFE, which operate unidirectionally (generated $\rightarrow$ reference). In our setting, neither run is assumed to be definitively complete: one run may identify findings the other omits. The greedy matching permits many-to-one alignments, accommodating cases where one run expresses a single comprehensive finding that the other decomposes into multiple granular claims.

C. Claim-Level Self-Consistency: Endpoint and Species Definitions

Table 2. Fifteen standardized safety endpoints used in the integrated risk assessment, aligned to MedDRA System Organ Classes (Brown et al., 1999).

#	Safety Endpoint
1	Hepatotoxicity
2	Renal toxicity
3	Bone marrow toxicity / hematotoxicity
4	Neurological / CNS / neurodevelopmental toxicity
5	Cardiovascular toxicity
6	Reproductive / developmental toxicity
7	Ocular toxicity
8	Inflammatory syndrome
9	Immune dysfunction
10	Metabolic dysregulation
11	Skin and subcutaneous tissue disorders
12	Gastrointestinal toxicity
13	Pulmonary toxicity
14	Endocrine disruption
15	Pregnancy / puerperium / perinatal conditions

Table 3. Species group definitions for the species context matching function ϕ_s. Non-human *in vitro* systems (*e.g.*, hERG assays, rodent hepatocytes, genotoxicity panels) are classified as experimental modalities and excluded from species-based matching.

Species Group	Members
Rodent	Mouse, rat, hamster, guinea pig
Non-rodent	NHP (cynomolgus macaque, marmoset), dog, minipig, rabbit
Human	Human genetic evidence, clinical adverse event data, human-derived *in vitro* systems (iPSC-derived cells, primary cells)
Alternative screening models	Zebrafish embryo

D. Claim-Level Self-Consistency: Prompt Specifications

Self-Consistency Evaluation Instructions

```
You are a toxicology expert. Determine whether these two observations describe the same
safety finding.

Observation A: "{text_a}"
Observation B: "{text_b}"

Score EXACTLY one of:
- 1 = Match: The observations describe the same finding if ANY of these rules apply:
 - "synonym": Synonym or paraphrase. Different wording for the same concept, for
 example, "liver fibrosis" = "hepatic fibrosis"
 - "cause_consequence": Cause and consequence within the same organ/system. One is the
 direct cause or result of the other, both involving the same organ system, for
 example, "reduced K+ currents" = "prolonged QT interval" (both cardiac
 electrophysiology); "VE-cadherin disruption" = "vascular hemorrhage" (both vascular)
 - "mechanism_outcome": Mechanism and clinical outcome within the same system. A
 molecular or cellular mechanism paired with its clinical manifestation in the same
 system, for example, "demyelination" = "nerve conduction slowing" (both peripheral
 nervous system)
```

```
 - "biomarker_disease": Biomarker evidence and the condition it indicates. A measurable
 marker and the disease or toxicity it reflects, such as "elevated ALT" =
 "hepatotoxicity"; "proteinuria" = "podocyte injury".
 - "includes": One includes the other. One observation is a specific instance, subtype,
 or manifestation of the other. This happens when one is vague and the other is
 specific, for example, if one term is a broad category that encompasses the other.
 This includes "fulminant hepatic failure" = "hepatotoxicity", "ventricular
 tachycardia" = "cardiac arrhythmia", "skin reactions" = "acneiform rash" (general term
 includes the specific).
 - "same_phenotype": Same underlying phenotype observed across different experimental
 setups or with different measurements. Both findings are phenotypic consequences of
 modulating the same target within the same organ or system and belong to the
 toxicological profile of the same target perturbation. For example, "hepatic steatosis
 in constitutive knockout" = "hepatic steatosis with antibody treatment" (same
 phenotype confirmed across genetic and pharmacological models).
 - "same_syndrome": Both findings are recognized features of the same syndrome with the
 same clinical grouping. For example, "delayed puberty" = "premature follicle
 depletion" (both are recognized features of premature ovarian insufficiency).
 - "same_pathway": Different molecules in the same biological pathway. Both
 observations describe changes in molecules that belong to the same biological
 signaling pathway. For example, "elevated IL-12p40" = "elevated IL-1alpha" (both
 pro-inflammatory cytokines), "reduced DOPAC levels" = "reduced striatal dopamine"
 (both dopaminergic pathway).
 - "same_pathology": Same pathology finding with different quantitative description.
 Same underlying pathology differing only in time course, frequency, severity or grade
 is a MATCH. This includes "Grade 1 thrombocytopenia" = "grade 3-4 thrombocytopenia";
 "Acute liver injury" = "chronic hepatotoxicity".



- 0 = No match: The observations are different findings if EITHER applies:
 - "different_organs": The findings affect different organs or body systems, even if
 caused by the same gene, drug, or pathway.    For example, "B cell depletion" does not
 match "pericarditis" (immune vs cardiac); "reduced fertility" does not match "reduced
 hypothalamic neurons" (reproductive vs CNS); "headache" does not match "elevated
 IFN-alpha" (neurological symptom vs systemic cytokine).
 - "contradictory": The observations describe opposite directions of the same parameter
 or mutually exclusive states. This includes the cases where "OPC proliferation" does
 not match "OPC depletion", "protection from neurodegeneration" does not match
 "neurodegeneration"; "no hepatotoxicity observed" does not match "hepatotoxicity".
 - "distinct": The observations describe different findings or aspects, even when
 referring to the same organ. They do not contradict each other, but they also do not
 correspond to the same observation.


Species context: Species weighting is handled separately and not considered when
analyzing the finding itself.
First classify which rule applies, then assign the score.
First go through ALL categories in the MATCH cases, and move to the NO MATCH category
ONLY when NONE of the MATCH rules apply.

Categories: synonym, cause_consequence, mechanism_outcome, biomarker_disease, includes,
same_phenotype, same_syndrome, same_pathway, same_pathology, different_organs,
contradictory, distinct.


Return ONLY a JSON object: {{"rule": "<category>", "score": <int>}}
```

Figure 5. Full instruction prompt used for the self-consistency evaluation.

Claim Extraction Instructions

```
You are a pharmacology/toxicology expert. Given a toxicity finding from a Target Safety
Assessment report, decompose it into atomic claims.

Each claim must be a self-contained observation linked to its species.

Toxicity Endpoint: {endpoint_name}
Target: {target}
Finding text:
\"\"\"{finding}\"\"\"

Decompose into a JSON list of claims:

{{
  "claims": [
    {{
      "observation": "specific phenotype or finding as a short phrase",
      "species": "one of: human, mouse, rat, hamster, guinea_pig, dog, minipig, rabbit,
      non_human_primate, zebrafish"
    }}
  ]
}}

Rules:
- Each claim = ONE distinct observation tied to ONE species
- Only extract POSITIVE findings (observed phenotypes, toxicities, or adverse effects).
Do NOT extract negative statements such as "no adverse findings", "no toxicity
observed", "not reported", "no dose-limiting findings identified", or any statement
indicating absence of effect.
- Extract only homozygous phenotypes in animals observed after knock out (0% function
of the target of interest) (e.g., "increased adipose tissue mass upon Inhbe
overexpression", "hepatotoxicity in KO mice")
- Do NOT extract gene/protein expression patterns (e.g., "target X is expressed in
tissue Y", "high expression in brain"), tissue distribution data, or mechanistic
descriptions that are not observed adverse phenotypes. However, DO extract phenotypic
consequences of overexpression/knockout experiments (e.g., "increased adipose tissue
mass upon Inhbe overexpression", "hepatotoxicity in KO mice") -- these are valid
toxicology findings even though the experimental model involves gene manipulation.
- Do NOT extract findings from non-human in vitro systems (e.g., hERG assays, rodent
hepatocytes, genotoxicity panels, cell line experiments). These are experimental
modalities, not in vivo toxicology findings. Only extract in vivo observations or human
clinical/genetic data.
- If a sentence mentions multiple phenotypes, split into separate claims
- "observation": be specific (e.g., "aortic aneurysm formation" not "cardiovascular
effects")
- "species": the species in which this was observed, classified per ICH S6 regulatory
toxicology categories:
  - rodent: mouse, rat, hamster, guinea_pig
  - non-rodent: non_human_primate (cynomolgus macaque, marmoset), dog, minipig, rabbit
  - human: encompasses human genetic evidence, clinical adverse event data, and
  human-derived in vitro systems (e.g., iPSC-derived cells, primary human hepatocytes)
  - alternative screening models: zebrafish
  - Do NOT use "other". If a finding does not fit the above species categories, skip it
  entirely.

Return ONLY valid JSON, no markdown fencing.


SPECIES_ALIASES = {
    "mice": "mouse", "murine": "mouse", "mus musculus": "mouse",
    "rats": "rat", "rattus": "rat",
    "hamsters": "hamster", "syrian hamster": "hamster",
    "guinea pig": "guinea_pig", "guinea pigs": "guinea_pig", "cavia": "guinea_pig",
    "dogs": "dog", "beagle": "dog", "canine": "dog",
```

```
    "minipigs": "minipig", "mini-pig": "minipig", "pig": "minipig", "swine": "minipig",
    "porcine": "minipig",
    "rabbits": "rabbit", "lapine": "rabbit",
    "nhp": "non_human_primate", "monkey": "non_human_primate", "monkeys":
    "non_human_primate",
    "primate": "non_human_primate", "primates": "non_human_primate",
    "cynomolgus": "non_human_primate", "macaque": "non_human_primate", "rhesus":
    "non_human_primate",
    "marmoset": "non_human_primate",
    "zebrafish": "zebrafish", "danio rerio": "zebrafish", "danio": "zebrafish",
    "humans": "human", "patients": "human", "clinical": "human",
}

VALID_SPECIES = {"human", "mouse", "rat", "hamster", "guinea_pig", "dog", "minipig",
                 "rabbit", "non_human_primate", "zebrafish", "other"}
```

Figure 6. Full instruction prompt used for extracting the atomic claims from the safety endpoint.

E. Claim-Level Self-Consistency: Claim Comparison Statistics

Table 4 provides a full breakdown of all the claim pairs generated during the self-consistency evaluation described in Section 4.1.1, stratified by species relation and scoring rule. Each pair is scored using the two-factor scheme of Figure 5: a rule-based scorer first classifies the semantic relationship between the two claims (Match rules contribute a positive score; No Match rules yield zero), and the resulting score is then scaled by the species weight $\phi_s \in \{0.5, 0.75, 1.0\}$ reflecting the biological translatability of the pair.

The majority of pairs fall in the *Same Species* category (67.4%), consistent with the expectation that independent runs on the same target should predominantly surface evidence from the same biological contexts.

Within this category, the dominant No Match rule is *Different Organs* (25.48% of Same Species pairs), which arises when two claims describe the same species but distinct anatomical systems and thus cannot be considered concordant. Cross-species pairs (*Diff. Group*, 31.7%) show a higher concentration of *Mechanism Outcome* matches (21.23%), where one claim describes a molecular or cellular mechanism and the other its clinical manifestation within the same organ system. This pattern is elevated in cross-species comparisons because preclinical runs tend to capture mechanistic observations (*e.g.*, hepatocyte apoptosis) while independent runs drawing on clinical literature surface the downstream adverse manifestation (*e.g.*, drug-induced liver injury), both localized to the same system but attributed to different species groups.

Table 4. Distribution of claim-pair scoring rules stratified by species relation, derived from 35 targets and 15 safety and toxicological safety endpoints. For each target, 10 independent LLM runs were executed under identical prompts, producing $\binom{10}{2} = 45$ run pairs per target. Within each run pair, every claim extracted by run a is cross-compared against every claim extracted by run b for the same endpoint. Each pair is assigned a rule by a rule-based scorer in Figure 5. The *species relation* field captures whether the two claims in a pair reference the same biological species (*Same Species*, 67.4%), species from different groups (*Diff. Group*, 31.7%), or species from the same group but distinct entries (*Same Group*, 0.9%). *% within*: fraction of claim pairs within that species relation category assigned the given rule. *% of total*: fraction across claim pairs.

		Same Species		**Diff. Group**		**Same Group**	
		67.4% of all records		31.7% of all records		0.9% of all records	
Category	**Rule**	% within	% of total	% within	% of total	% within	% of total
Match	biomarker_disease	1.91%	1.29%	2.04%	0.65%	2.53%	0.02%
	cause_consequence	15.73%	10.60%	9.16%	2.90%	3.08%	0.03%
	mechanism_outcome	10.45%	7.05%	21.23%	6.73%	4.95%	0.04%
	same_pathology	7.90%	5.33%	1.50%	0.47%	2.64%	0.02%
	same_pathway	4.85%	3.27%	4.03%	1.28%	0.55%	0.00%
	same_phenotype	8.34%	5.62%	13.37%	4.24%	12.54%	0.11%
	same_syndrome	7.58%	5.11%	4.84%	1.53%	0.22%	0.00%
	synonym	4.47%	3.01%	0.15%	0.05%	3.41%	0.03%
	includes	8.96%	6.04%	8.44%	2.67%	7.37%	0.07%
No Match	contradictory	0.97%	0.65%	1.25%	0.40%	2.86%	0.03%
	different_organs	25.48%	17.18%	29.66%	9.40%	49.61%	0.44%
	distinct	3.37%	2.27%	4.33%	1.37%	10.23%	0.09%

F. Sequence Homology Procedure

This appendix provides the full procedure summarized in Section 4.2.1.
The agent carries out the core of the ortholog identification task in four main steps:

(1) Query the ortholog database for each preclinical species.

(2) Review the ortholog candidates returned and retain those passing the deterministic pre-filter (length deviation $\leq 20\%$, pairwise mismatch $\leq 60\%$).

(3) For any species with no valid candidate from step 2, run a targeted sequence similarity search against the human canonical sequence.

(4) Review the search results and select the highest-confidence candidate that passes the pre-filter.

When no suitable candidate can be found in either pass, the agent excludes the species from the analysis and reports every database it searched together with the rejection reason for each candidate. This allows the user to exercise their own judgment by accepting the exclusion, manually selecting a borderline candidate, or invoking a non-standard procedure (*e.g.*, excising a divergent C-terminal extension and re-running the similarity search on the remainder).

G. HITL: Target-Specific Refinement Questions

Target A (inhibitor)

```
HITL_A_Q1 (Clinical Safety, Completeness):
"Revise this section to include any clinical-stage compounds targeting this target or
the oxysterol pathway, and add available safety data for these compounds."

HITL_A_Q2 (Risk Assessment, Completeness):
"This target is involved in both immune cell migration and lipid metabolism. Update the
integrated assessment to ensure both axes of risk are explicitly reflected and
discussed."

HITL_A_Q3 (Genetic Evidence, Completeness):
"Knockout mice show altered germinal center formation. Ensure this phenotype is
included in the section and add a discussion of its relevance to immunosuppression
risk."

HITL_A_Q4 (Genetic Evidence, Accuracy):
"Update this section to include any reported loss-of-function variants from population
databases (e.g., gnomAD), with correctly described associated phenotypes."
```

Target B (activator)

```
HITL_B_Q1 (Expression Profile, Completeness):
"This target is predominantly neuronal. Update the expression analysis to include any
non-neuronal tissues with detectable expression and discuss potential off-target safety
concerns."

HITL_B_Q2 (Genetic Evidence, Completeness):
"Update this section to include TDP-43 related variants that affect target expression,
and discuss whether these inform the safety profile of target activation."

HITL_B_Q3 (Clinical Safety, Completeness):
"Update this section to cover any clinical programs targeting this protein or the
TDP-43 axis. If none exist, state this explicitly and discuss the implications for
safety assessment."

HITL_B_Q4 (Risk Assessment, Prioritization):
"Given that this is an activator with neuronal-restricted expression, revise the risk
assessment to clearly state the primary safety concern as the top-ranked risk."
```

Target C (inhibitor)

```
HITL_C_Q1 (Genetic Evidence, Completeness):
"Target deficiency causes a lysosomal storage disease. Update this section to include
key phenotypes of knockout models and add a discussion of their relevance to
dose-dependent pharmacological inhibition."

HITL_C_Q2 (Expression Profile, Prioritization):
"This target is a lysosomal enzyme with broad tissue expression. Revise the expression
analysis to prioritize which organs are most vulnerable to substrate accumulation upon
inhibition."

HITL_C_Q3 (Clinical Safety, Completeness):
"Update this section to include any clinical compounds targeting this enzyme or the
ceramide pathway, along with their reported adverse event profiles."

HITL_C_Q4 (Risk Assessment, Consistency):
"Revise the risk assessment to explicitly distinguish between complete loss-of-function
(Y disease) and partial pharmacological inhibition, and add a discussion of the
therapeutic window."
```

Figure 7. Representative target-specific refinement questions used in the HITL regression test (Section 4.2.2). Each question is designed by a safety scientist to probe a specific report section along one of four dimensions: accuracy, completeness, prioritization, or consistency. Target names are anonymized. The figure shows all twelve interventions used in the regression test, spanning three representative targets with four questions each.