

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

Xiaochen Zheng^{1*}, Zhiwen Jiang^{2*}, David Tokar^{1†}, Yexiang Cheng^{1†},
Alvaro Serra¹, Melanie Guerard³, Klas Hatje², Tatyana Doktorova¹

¹Predictive Modelling, Roche, Basel, Switzerland

²Computational Sciences Center of Excellence, Roche, Basel, Switzerland

³Translational Safety, Roche, Basel, Switzerland

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 labour-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 specialised subagents, each responsible for a single report section. Each subagent retrieves evidence from curated biomedical sources through standardised tool interfaces and produces individually citable, evidence-grounded sections, with behaviour 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.

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 characterisation 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 (Appendix A.1). Performed at the target nomination stage and conducted rigorously, TSA can de-risk programmes before resources are committed to preclinical and clinical studies (Plenge et al., 2013).

Despite its importance, the generation of TSA reports remains largely manual. Scientists must synthesise 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;

* Equal contribution. † Equal contribution as second authors.
Corresponding authors: xiaochen.zheng@roche.com,
zhiwen.jiang@roche.com

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 evidence domain (Figure 1); (ii) a **three-layer instruction architecture** separating coordination logic, domain skills, and runtime user intent; (iii) a **programmatically enforced 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 TSASSISTANT: System Design

2.1 Hierarchical Instruction Architecture

Agent behaviour 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) defines agent roles, coordination logic, and output structure, shared immutably across all targets. **Layer 2** (domain skill modules) encodes section-specific retrieval strategies, evidence weighting, and writing guidelines as independently versioned, composable modules. **Layer 3** (runtime user instructions) injects 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 retrieval guidance independently of coordination logic, and users to customise individual assessments without modifying either reusable layer.

2.2 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 and sequence-level computation to higher-level domain logic encoding acceptance criteria and multi-source cross-referencing strategies. 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 3). A complete catalog of the public data sources backing these tools is provided in Appendix B.

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. Two persistent stores bridge the isolated agents: *tool memory* records every tool-call output with provenance metadata, 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.

2.3 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

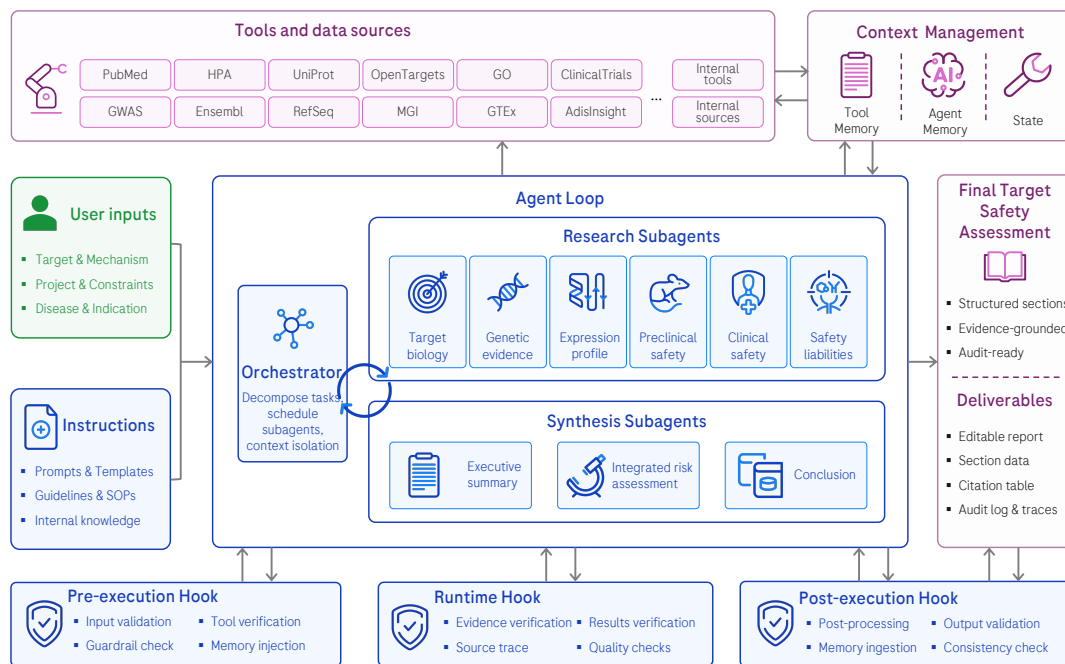


Figure 1: **Hierarchical agent architecture of TSASSISTANT.** An *Orchestrator* decomposes the assessment into *Research Subagents* and *Synthesis Subagents*, each targeting a single TSA domain. Execution hooks and three persistent context stores (tool memory, agent memory, execution state) provide programmatic enforcement; all subagents interface with curated biomedical data sources through standardised MCP tool interfaces. Mechanism detail in Section 2.2.

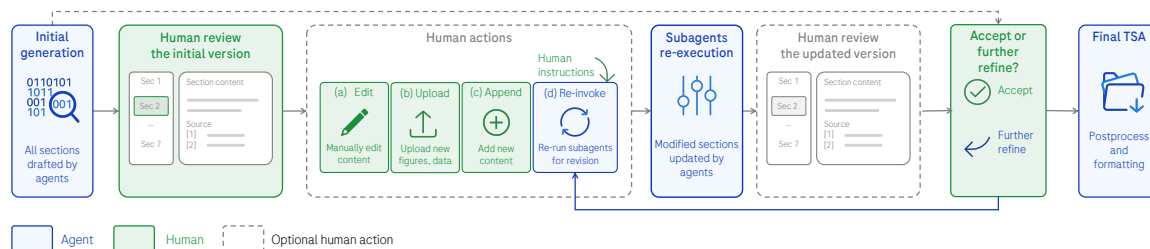


Figure 2: **Interactive refinement loop in TSASSISTANT.** After initial section generation, the user reviews each section and may edit, append, upload sources, or re-invoke the responsible subagent for targeted revision; conversational memory preserves context across iterations and user feedback progressively adapts retrieval strategies and prompt templates. Action types detailed in Section 2.3.

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

3 Evaluation Framework

We evaluate TSASSISTANT along two complementary axes: *automatic intrinsic* evaluations (Section 3.1) that probe report-level reliability without

a human reference, and *evaluations with human reference and oversight* (Section 3.2) that capture human expertise as a frozen reference and as an interactive reviewer.

3.1 Automatic Intrinsic Evaluation

3.1.1 Claim-Level Self-Consistency

The integrated risk assessment produced by TSASSISTANT consolidates findings into a structured table of fifteen standardised toxicity endpoints (Table 2 in Appendix E) aligned to MedDRA System Organ Classes (Brown et al., 1999). We formalise 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, beyond what holistic text similarity captures when the same observation is phrased differently or attributed to 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 $c = (\text{observation}, \text{species})$, where *species* encodes the biological context in which the observation was made (see Appendix E for claim terminology). 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 Appendix D).

Experimental setup. We evaluate across 35 drug targets, covering indications across Cardiovascular and Metabolism, Immunology, Infectious diseases and Ophthalmology, Neuroscience and Rare diseases, and Oncology and Haematology¹, 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 (compute footprint and judge-model rationale: Appendix E). 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.

¹ <https://www.roche.com/innovation/structure/pred>

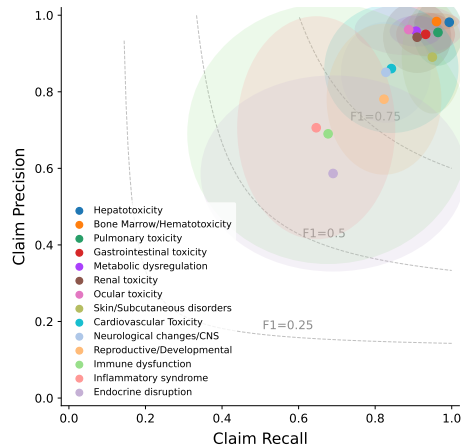


Figure 3: Claim-Level Self-Consistency by Safety Endpoint. Symmetric Precision and Recall (Section 3.1.1) for each of the 15 endpoints, aggregated over 45 run pairs per target across 35 targets. The shaded ellipsoidal regions around each endpoint are axis-aligned and span ± 1 sample standard deviation 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.

Results. Figure 3 reports claim-level self-consistency across the 15 safety endpoints; 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, reflecting the deterministic nature of structured sources and database retrieval, 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.

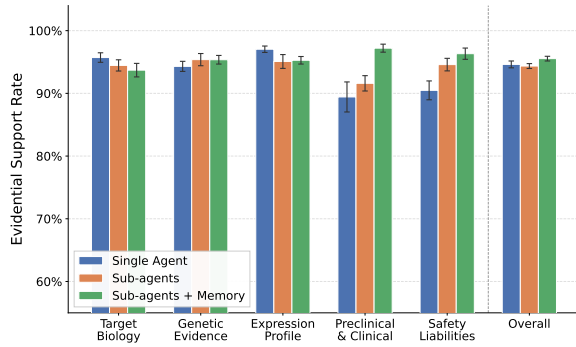


Figure 4: Evidential support rate by section; error bars $\pm$ SE.

3.1.2 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 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.

Setup and results. Figure 4 compares the ESR across three architectural conditions (single agent, subagents, and subagents with memory), averaged over 35 drug targets. 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 sections and the accumulated context grows, its ESR drops to $\sim 90\%$ 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 largely eliminates the quality degradation observed in later sections, confirming that both task decomposition and cross-section evidence propagation contribute to

reducing hallucination in TSA report generation.

As a complementary atomic-claim-level audit on a single target, 20 fully ungrounded key evidential units (KEUs) were observed to propagate to 30 downstream claims (Appendix I), illustrating how localized grounding failures can bypass the sentence-level ESR metric and compound through inter-agent memory. Taken together, the ESR and KEU analyses are complementary: ESR measures grounding under the assumption that memory content is trustworthy, while the KEU audit quantifies the extent to which that assumption is violated.

3.2 Evaluation with Human Expertise

3.2.1 Benchmark: Sequence Homology

Target homology is one of the six evidence pillars of TSA (Appendix A.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 programmes, 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 Appendix C).

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). 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 upstream error: the agent anchored to an incorrect, substantially longer human isoform, causing the correct candidates to be discarded by the deterministic $\pm 20\%$ 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. Second, the human analysts drew on specialised databases not yet accessible to the agent, yielding isoforms the agent could not reach. 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. The first two are tractable engineering gaps; the third is an open research problem flagged for human review in deployment.

3.2.2 HITL Refinement Regression Test

Whereas Section 3.2.1 treats the human contribution as a frozen reference, the refinement loop itself (Section 2.3) 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 intervention 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* = $(B - H) / (B + N + H)$ 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.

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
<i>By question type</i>					
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%
<i>By report section</i>					
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, 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. *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 reorganising content already in scope. *Accuracy* reaches +85.0% but aggregates a single intervention. *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.

4 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 3.1.1), achieves high sentence-level grounding under subagent decomposition with cross-section memory (Section 3.1.2), tracks closely with a human reference on the ortholog benchmark (Section 3.2.1), and yields net-positive expert-driven refinement (Section 3.2.2). 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. 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 pharmaceutical scientists, and aims to contribute to a broader vision of AI-augmented pharmaceutical safety science where agentic AI and human expertise are complementary.

Limitations

A natural evaluation strategy would be to compare TSASSISTANT outputs against expert-authored reports on the same targets in a holistic, head-to-head manner. We deliberately did not pursue this approach because TSA reports lack a unique ground truth: different domain experts writing independently on the same target may legitimately be different while all producing scientifically valid assessments. This makes holistic comparison metrics ill-defined, as a claim absent from an expert report may reflect a deliberate scoping decision rather than an omission.

Our evaluation strategy instead decomposes report quality into independently measurable properties: reproducibility, evidential grounding, task-level accuracy, and controllability under expert oversight. This decomposition enables targeted measurement but introduces two blind spots that holistic comparison would not have.

First, self-consistency (Section 3.1.1) measures whether independent runs agree, not whether they agree on correct claims. A system that deterministically retrieves the same incorrect information from the same databases will score high on Claim-F1. The metric therefore captures reproducibility but remains silent on factual accuracy.

Second, the ESR metric (Section 3.1.2) treats memory provenance as a grounding signal by design: sentences supported solely by upstream sub-agent memory are classified as grounded regardless of whether the upstream content itself was verified. ESR therefore provides an upper bound on true grounding. Quantifying the extent to which ungrounded content propagates through inter-agent memory remains an open question for future work.

References

Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Saleema Amershi, Maya Cakmak, W. Bradley Knox,

and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

Anthropic. 2024. Model context protocol. <https://modelcontextprotocol.io>.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

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

Juan M Banda, Lee Evans, Rami S Vanguri, Nicholas P Tatonetti, Patrick B Ryan, and Nigam H Shah. 2016. A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific data*, 3(1):1–11.

Ewan Birney, T. Daniel Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, James Cuff, Val Curwen, Tim Cutts, and 1 others. 2004. An overview of Ensembl. *Genome Research*, 14(5):925–928.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624:570–578.

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

Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, and 1 others. 2019. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. 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)*, pages 7066–7085.

David Cook, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterthwaite, and Menelas N Pangalos. 2014. Lessons learned from the fate of astrazeneca’s drug pipeline: a five-dimensional framework. *Nature reviews Drug discovery*, 13(6):419–431.

Fiona Cunningham, James E. Allen, Jamie Allen, Jorge Alvarez-Jarreta, M. Ridwan Amode, Irina M. Armean, Olanrewaju Austine-Orimoloye, Andrey G.

- Azov, If Barnes, Ruth Bennett, and 1 others. 2022. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995.
- Jason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, and 1 others. 2024. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*.
- Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and 1 others. 2025. Democratizing ai scientists using tooluniverse. *arXiv preprint arXiv:2509.23426*.
- Alireza Ghafarollahi and Markus J. Buehler. 2025. *Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning*. *Advanced Materials*, 37(22):2413523.
- Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, and 1 others. 2022. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D419–D426.
- Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.
- GTEX Consortium. 2020. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330.
- Richard K Harrison. 2016. Phase II and phase III failures: 2013–2015. *Nature reviews Drug discovery*, 15(12):817–818.
- Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. 2014. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. 2011. ICH harmonised tripartite guideline S6(R1): Preclinical safety evaluation of biotechnology-derived pharmaceuticals. Technical report, ICH. <https://www.ich.org/page/safety-guidelines>.
- Yuyang Jiang, Chacha Chen, Shengyuan Wang, Feng Li, Zecong Tang, Benjamin M. Mervak, Lydia Chelala, Christopher M Straus, Reve Chahine, Samuel G. Armato Iii, and Chenhao Tan. 2025. CLEAR: A clinically grounded tabular framework for radiology report evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.
- John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, and 1 others. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.
- Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. 2023. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D483–D489.
- Minoru Kanehisa and Susumu Goto. 2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.
- Max Karlsson, Cheng Zhang, Loren Méar, Wen Zhong, Andreas Digre, Borbala Katona, Evelina Sjöstedt, Lynn Butler, Jacob Odeberg, Philip Dusart, and 1 others. 2021. A single-cell type transcriptomics map of human tissues. *Science advances*, 7(31):eabh2169.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.
- Philippe Laban, Tobias Schnabel, and Jennifer Neville. 2026. Llms corrupt your documents when you delegate. *arXiv preprint arXiv:2604.15597*.
- Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, and 1 others. 2018. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and Sebastian Riedel. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474. Curran Associates, Inc.

- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. [Generating with confidence: Uncertainty quantification for black-box large language models](#). *Transactions on Machine Learning Research*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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*, pages 9004–9017.
- David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, and 1 others. 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100.
- Paul Morgan, Dean G Brown, Simon Lennard, Mark J Anderton, J Carl Barrett, Ulf Eriksson, Mark Fidock, Bengt Hamren, Anthony Johnson, Ruth E March, and 1 others. 2018. Impact of a five-dimensional framework on r&d productivity at astrazeneca. *Nature reviews Drug discovery*, 17(3):167–181.
- Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.
- Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, and 1 others. 2015. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860.
- David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Asier Gonzalez-Urriarte, Cinzia Malangone, Alfredo Miranda, Luca Fumis, Denise Carvalho-Silva, Michaela Spitzer, and 1 others. 2021. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic acids research*, 49(D1):D1302–D1310.
- David Ochoa, Andrew Hercules, Miguel Key, Deepika Chandran, John Sheridan, Ian Dunham, Ahmet Gunes, Joannella Morales, Ellen M. McDonagh, Tudor I. Oprea, and 1 others. 2023. The next-generation open targets platform: Reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51(D1):D1353–D1359.
- Nuala A. O’Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciuffo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, and 1 others. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745.
- Harry Olson, Graham Betton, Denise Robinson, Karluss Thomas, Alastair Monro, Gerald Kolaja, Patrick Lilly, James Sanders, Glenn Sipes, William Bracken, and 1 others. 2000. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regulatory toxicology and pharmacology*, 32(1):56–67.
- Irene Papatheodorou, Pablo Moreno, Jonathan Manning, Ana M. Perez Fuentes, Nuno George, Silvie Fexova, Nuno A. Fonseca, Anja Füllgrabe, Mark Green, Nancy Huang, and 1 others. 2020. Expression atlas update: From tissues to single cells. *Nucleic Acids Research*, 48(D1):D77–D83.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulators of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST ’23)*, pages 1–22. ACM.
- Janet Piñero, Juan Manuel Ramírez-Anguaita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855.
- Robert M Plenge, Edward M Scolnick, and David Altshuler. 2013. Validating therapeutic targets through human genetics. *Nature reviews Drug discovery*, 12(8):581–594.
- Rita Santos, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I. Oprea, and John P. Overington. 2017. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34.

- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in neural information processing systems (NeurIPS)*, volume 36, pages 8634–8652.
- Duxin Sun, Wei Gao, Hongxiang Hu, and Simon Zhou. 2022. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062.
- Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.
- The Gene Ontology Consortium. 2023. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031.
- Mathias Uhlén, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, and 1 others. 2015. Tissue-based map of the human proteome. *Science*, 347(6220):1260419.
- Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, and 1 others. 2010. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250.
- UniProt Consortium. 2023. Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51(D1):D523–D531.
- Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, and 1 others. 2025. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pages 2250–2261.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.
- Jacob White. 2020. PubMed 2.0. *Medical Reference Services Quarterly*, 39(4):382–387.
- David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, and 1 others. 2018. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Auto-gen: Enabling next-gen LLM applications via multi-agent conversations. In *First conference on language modeling (CoLM)*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems (NeurIPS)*, 36:11809–11822.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations (ICLR)*.
- Harrison G Zhang, Peter Eckmann, Jiacheng Miao, Andrew B Mahon, and James Zou. 2026. The virtual biotech: A multi-agent ai framework for therapeutic discovery and development. *bioRxiv*, pages 2026–02.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems (NeurIPS)*, 36:46595–46623.
- Yucheng Zhou, Lingran Song, and Jianbing Shen. 2025. Mam: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25319–25333.
- Yinghao Zhu, Lei Gu, Zixiang Wang, Haoran Sang, Dehao Sui, Wen Tang, Lan Mi, Yasha Wang, Junyi Gao, Liang Yao, Tianfan Fu, Ewen Harrison, Lequan Yu, and Liantao Ma. 2026a. Auditing medical multi-agent ai reveals risks of false consensus. *Preprint*, arXiv:2510.10185.
- Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2026b. 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*.

A Background

A.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 programmes 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.

A.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., 2026b).

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 2.1).

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 2.2).

Autonomous Scientific Agents. Recent work extends multi-agent architectures to full scientific workflows. SciAgents (Ghafarollahi and Buehler, 2025) grounds multi-agent hypothesis generation in large-scale ontological knowledge graphs, sampling relational paths between distant concepts to propose and critique research directions in bio-inspired materials. 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, reproducibility, 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 2.

A.3 Human-in-the-Loop Systems

HITL systems integrate human judgement 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 judgement. 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 2.3.

B Data Sources

The data sources backing the tool interfaces span the five evidence domains described in Appendix A.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 and 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.gov² supplies clinical study records including safety endpoints and adverse event reports from interventional trials. AdisInsight³ provides competitive intelligence, including known severe adverse events associated with drugs modulating the same target.

² <https://clinicaltrials.gov> ³ <https://adisinsight.springer.com>

C Sequence Homology Procedure

This appendix provides the full procedure summarised in Section 3.2.1.

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 — 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).

D Claim-Level Self-Consistency: Full Metric Definition

This appendix provides the complete metric definition summarized in Section 3.1.1.

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 $c = (\text{observation}, \text{species})$, 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:

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

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:

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

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 (Appendix E). Contradictory claims (e.g., “dose-related neutropenia” vs. “no significant haematologic toxicities”) 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, signalling 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:

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

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

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

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.

E Evaluation Terminology Definitions

Atomic claim terminology. 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.

Compute footprint and cross-LLM-judge limitation. A single pairwise LLM-as-a-judge evaluation requires $35 \times 45 \times 15 \times n$ API calls with $n \in [0, 20]$, 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.

Table 2: Fifteen standardised 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 <i>in vitro</i> systems (iPSC-derived cells, primary cells)
Alternative screening models	Zebrafish embryo

F Evaluation 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_organisms": 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_organisms, 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.

G Claim Comparison Statistics

Table 4 provides a full breakdown of all the claim pairs generated during the self-consistency evaluation described in Section 3.1.1, 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 localised 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 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.

Category	Rule	Same Species		Diff. Group		Same Group	
		67.4% of all records		31.7% of all records		0.9% of all records	
		% 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%

H Target-Specific Refinement Questions

Target A (inhibitor)

HITL_A_Q1 (Clinical Safety):

"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):

"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):

"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):

"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):

"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):

"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):

"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):

"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):

"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):

"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):

"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):

"Revise the risk assessment to explicitly distinguish between complete loss-of-function (genetic 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 (3.2.2). Each question is designed by a safety scientist to probe a specific report section along one of four dimensions: accuracy, completeness, prioritisation, or action. Target names are anonymised; the full question set spans 15 targets with 10 questions each.

I Key Evidential Unit Analysis Methodology

We define a Key Evidential Unit (KEU) as an atomic, self-contained biological or clinical statement, a specific metric or measurement, a conclusion derived from other statements, or a specific reference to a paper or mechanism. To ensure strict traceability, each KEU must be fully resolved (e.g., pronouns replaced with explicit entities) and traceable to a verbatim source quote.

We adapt concepts introduced by (Zhu et al., 2026a), while MedAgentAudit uses a bottom-up approach to evaluate open-ended peer interactions—tracking whether evidential units survive rounds of unstructured debate—our framework applies a top-down approach. We track KEU provenance across subagents to determine exactly where a fact originated and through which edges it propagated.

I.1 Agent Pipeline Architecture

The system divides KEU generation into two distinct operational tiers. **Research Agents:** Six specialized subagents that sequentially execute external tool calls (e.g., PubMed, Ensembl) and pass memory forward. **Synthesis Agents:** Three downstream subagents that consume the generated research reports to formulate higher-level summaries. These agents operate in a read-only capacity and do not have independent external tool access.

I.2 Extraction, Grounding, and Propagation Tracking

The auditing pipeline processes the outputs of these agents in four sequential steps to construct a complete lineage graph:

Step 1: KEU Extraction. The raw text generated by each subagent is processed by an LLM to extract candidate claims along with their verbatim text. Each candidate is evaluated for atomicity: multi-fact candidates are split, and pronouns are resolved. The shortest verbatim quote is then extracted, citations are stripped, and a unique identifier (e.g., KEU-BIO-001) is assigned. A per-section deduplication process ensures only strict semantic duplicates are removed, retaining more precise variants.

Step 2: Memory Retention Mapping. For the research agents, the system evaluates each extracted KEU against the agent’s compressed memory document to determine if the fact was retained or dropped,

extracting a verbatim memory quote as proof of retention.

Step 3: Origin Grounding (Research Agents). An LLM auditor receives the extracted KEUs, memory retention maps, and the full interaction payload (including specific tool calls and inherited context). Because a single KEU can synthesize information from multiple sources, the auditor performs a cascading, three-pass fractional grounding analysis. First, it maps specific phrases or concepts within the KEU to inherited memory from prior agents. Second, it evaluates the interaction logs to map any remaining or overlapping claims to specific external tool calls. Third, it identifies any residual fragments of the KEU that lack an identifiable source and flags them as explicitly ungrounded. A secondary validation pass then cross-references these mapped claims against actual tool call IDs and upstream KEU IDs to identify LLM hallucinations.

Step 4: Synthesis Grounding. Synthesis KEUs are mapped against the full corpus of validated research KEUs. Each synthesis claim is categorized as either successfully traced to a parent research KEU or ungrounded if the claim is not justified by any underlying research finding. A final validation ensures the parent IDs exist and semantically match the downstream synthesis claim.

I.3 State Classification and Error Propagation

Based on the grounding steps, every node in the KEU network is assigned a specific verification state:

- **Fully Grounded:** The KEU is supported entirely by valid external tool calls or traceable memory inheritance, containing no ungrounded components.
- **Partially Grounded:** The KEU relies on at least one valid source but also incorporates ungrounded or unverifiable claims.
- **Fully Ungrounded:** The KEU lacks any valid identifiable source or supportive grounding entries.

To analyse how errors compound within multi-agent architectures, we calculate a *references ungrounded* dependency flag for each node. Any downstream KEU that directly inherits memory from a fully ungrounded parent KEU is explicitly flagged. This allows the system to trace how unsupported hallucinations bypass intra-agent filters and propagate as foundational truths to subsequent research and synthesis tiers.

I.4 Case Study: NRF2

We illustrate the methodology with a case study on a report generated for the NRF2 target, comprising 572 KEUs across the various subagents and 1,944 cross-references between them. Within this network, 474 KEUs were fully grounded, 78 were partially grounded (i.e., a part of the reasoning used to generate the KEU was not grounded), and 20 were fully ungrounded. The 20 fully ungrounded KEUs propagated to 30 subsequent KEUs, illustrating how errors can compound in multi-agent architectures, as downstream agents base their work on the unverified results of previous agents. A comprehensive visualization of this case study is available for download in the Supplementary Materials.

I.5 Prompt Specifications

Step 1: KEU Extraction - Pass 1 (Extraction)

You are an expert scientific data auditor specializing in pharmaceutical safety. Your task is to read a Target Safety Assessment (TSA) report section and extract every "Key Evidential Unit" (KEU).

A Key Evidential Unit is:

- * A non-trivial biological/clinical statement.
- * A specific metric, statistical value, or disease linkage.

Rules:

1. Extract the `exact\_text` verbatim from the provided report so it can be string-matched.
2. Try to break complex paragraphs down into atomic (individual) KEUs.
3. Ignore formatting headers (e.g., "## 2.1") unless they contain a specific claim.
4. Give each KEU a short, 1-3 word title.

5. Keep the `reasoning` field extremely short (under 10 words).

Step 1: KEU Extraction - Pass 2 (Atomicity and Refinement)

You are an expert scientific data refiner specializing in pharmaceutical safety.
You will be given a 'Candidate Key Evidential Unit (KEU)' extracted from a pharmaceutical report.

A Key Evidential Unit (KEU) is defined as:

- * Any non-trivial biological or clinical statement.
- * A specific metric, measurement, or statistical value.
- * A conclusion derived from other statements.
- * A specific reference to a paper, disease, pathway, or mechanism.

Your job is to evaluate if this candidate is TRULY ATOMIC.

1. If the text contains multiple distinct facts or metrics, you MUST split it into multiple, separate atomic KEUs.
2. If the text is already a single atomic fact, return a list containing just ONE KEU based on the original.

CRITICAL RULES FOR REFINEMENT:

- * `standalone\_fact`: Rewrite the fact to be completely self-contained. Resolve any pronouns (change "The enzyme" or "It" to the actual target name).
- * `exact\_source\_quote`: Extract the SHORTEST possible verbatim substring that proves this fact. You MUST completely strip out and remove reference citations (e.g., remove "[3,4]" or "[1]").
- * `reasoning`: Keep under 10 words.

Step 1: KEU Extraction - Pass 3 (Deduplication)

You are an expert scientific data auditor specializing in pharmaceutical safety.
You will be given a list of Key Evidential Units (KEUs) extracted from a SINGLE section of a Target Safety Assessment report.

Your task is to perform STRICT deduplication. Identify the IDs of the KEUs that should be REMOVED because they are LITERALLY REPHRASING another KEU.

Do NOT drop KEUs that are merely related, or where one is a subset of another. Only drop exact semantic duplicates where the core atomic fact is identical.

When choosing between strict duplicates, keep the one that is more comprehensive, precise, or better phrased, and drop the other(s).

Return ONLY the list of KEU IDs to drop and a brief reason (e.g., "Exact rephrasing of KEU-BIO-015").

If there are no strict duplicates, return an empty array.

Step 2: Memory Retention Mapping

You are an expert scientific auditor tracing data lineage.
You will be provided with a 'Subagent Memory' document and a single 'Key Evidential Unit' (KEU).
Your task is to determine if the core factual claim of the KEU is present in the memory document. The wording does not need to be an exact verbatim match to the KEU, but the specific scientific fact, metric, or concept MUST be explicitly retained in the memory.

If it is retained, you MUST provide the `exact\_memory\_quote` verbatim from the memory document that proves it.

If it is not retained, set `is\_retained` to false and leave `exact\_memory\_quote` empty.

Step 3: Origin Grounding Map - Pass 1 (Memory Grounding)

You are an expert scientific auditor verifying memory retention lineages.
Analyze the KEU against the provided RETAINED MEMORY from previous subagents.
Identify if the memory context supports all or any partial sections/phrases of this KEU.

CRITICAL RULES:

1. The `reasoning` field MUST NOT exceed 12 words under any circumstances. Keep it ultra-short.
2. Specify exactly which partial statement or concept of the KEU is supported.

Step 3: Origin Grounding Map - Pass 2 (Tool Call Grounding)

You are an expert scientific auditor verifying tool call groundings.
Analyze the KEU against the provided INTERACTION LOG tool results.
You will also see the groundings already mapped to memory during Pass 1.

Your goal is to evaluate if tool logs fully or partially support any remaining, ungrounded, or overlapping fractions of the KEU text.

CRITICAL RULES:

1. The `reasoning` field MUST NOT exceed 12 words under any circumstances. Keep it ultra-short.
2. Specify exactly which partial phrase, value, or claim is supported by the tool result.

Step 3: Origin Grounding Map - Pass 3 (Ungrounded Reconciliation)

You are an expert scientific auditor reconciling data lineage.
You are given a KEU and a comprehensive compilation of all successful groundings found in prior passes (Memory and Tools).

Your job is to identify if any phrase, metric, claim, or implicit assumption within the KEU remains completely ungrounded.
If the existing groundings completely cover 100% of the KEU's statements, return an empty array.
If any fragment of the KEU is missing from the existing groundings, explicitly flag it as "ungrounded".

CRITICAL RULES:

1. The `reasoning` field MUST NOT exceed 12 words under any circumstances. Keep it ultra-short.
2. State exactly which specific text element or metric is ungrounded.

Step 4: Synthesis Grounding Map - Research Pass

You are an expert scientific auditor verifying cross-section synthesis grounding maps.
Analyze the candidate Synthesis KEU against the complete set of atomic facts from the provided Research section.

Identify if this specific research section fully or partially supports any text fractions, statements, or conclusions in the Synthesis KEU.

CRITICAL RULES:

1. The `reasoning` field MUST NOT exceed 12 words under any circumstances. Keep it ultra-short.
2. Specify exactly which partial concept or metric of the synthesis KEU is supported by the research KEU ID.

Step 4: Synthesis Grounding Map - Final Pass

You are an expert scientific auditor reconciling data lineage for synthesis reports.
You are given a synthesis KEU alongside a compilation of all valid groundings successfully found across all research sections.

Your job is to identify if any phrase, metric, claim, or calculation within this Synthesis KEU remains completely ungrounded or unsupported.
If the existing groundings already account for 100% of the synthesis KEU's text, return an empty array.
If any fragment or data point is missing from the compiled list, explicitly flag it as "ungrounded".

CRITICAL RULES:

1. The `reasoning` field MUST NOT exceed 12 words under any circumstances. Keep it ultra-short.
2. State exactly which specific text element or metric is ungrounded.