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

Anonymous Authors[1]

Abstract

Target Safety Assessment (TSA) requires systematic integration of heterogeneous evidence, including genetic, transcriptomic, target homology, pharmacological, and clinical data, to evaluate potential safety liabilities of therapeutic targets. This process is inherently iterative and expert-driven, posing challenges in scalability and reproducibility. We present TSAssistant, a multi-agent framework designed to support TSA report drafting through a modular, section-based, and human-in-the-loop paradigm. The framework decomposes report generation into a coordinated pipeline of specialised subagents, each targeting a single TSA section. Specialised subagents retrieve structured and unstructured data as well as literature evidence from curated biomedical sources through standardised tool interfaces, producing individually citable, evidence-grounded sections. Agent behaviour is governed by a hierarchical instruction architecture comprising system prompts, domain-specific skill modules, and runtime user instructions. A key feature is an interactive refinement loop in which users may manually edit sections, append new information, upload additional sources, or re-invoke agents to revise specific sections, with the system maintaining conversational memory across iterations. TSAssistant is designed to reduce the mechanical burden of evidence synthesis and report drafting, supporting a hybrid model in which agentic AI augments evidence synthesis while toxicologists retain final decision authority.

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

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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. 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 programmes before resources are committed to preclinical and clinical studies (Plenge et al., 2013).

Despite its importance, the generation of TSA remains largely manual, with content composition not standardized and the level of detail and depth of analysis often varying between individuals. Additionally, 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; Mosqueira-Rey et al., 2023).

We introduce TSASSISTANT, a human-in-the-loop multi-agent framework for TSA. Our principal contributions are:

- A **section-based multi-agent pipeline** in which specialized subagents independently ground and cite each TSA evidence domain (Figure 1);
- A **three-layer instruction architecture** separating coordination logic, domain skills, and runtime user intent;
- 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;
- 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 programms 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 pipeline (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, 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 3.

2.3. Human-in-the-Loop Systems

HITL systems integrate human judgement into automated pipelines 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 structually 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 adopts a human-in-the-loop design in which user interaction is integrated at the level of modular report sections. Each section can be reviewed, edited, or regenerated independently, allowing experts to iteratively refine outputs while preserving traceability and provenance. This section-level interaction model enables early error detection and targeted correction, while maintaining the efficiency benefits of automated evidence retrieval and alignment. The concrete implementation of this design is 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 theraputic area and compound modality, TSASSISTANT produces a structured TSA report comprising individually citable, evidence-grounded sections.

A central design challenge is that LLM agent behaviour is governed by prompts, which function as *soft constraints*: they bias model outputs statistically but cannot guarantee correctness, structural compliance, reproducibility or evidence fidelity.

In a monolithic single-agent approach, this limitation is compounded by *error amplification*: if an incorrect or hallucinated fact enters the shared context, the model treats it as established evidence and propagates it into all subsequent reasoning, causing the error to compound across the entire report. Naively decomposing the task across multiple agents does not resolve this; recent empirical work shows that multi-agent systems without cross-verification mechanisms amplify errors by up to 17.2× (Kim et al., 2025), motivating the need for hard-constraint mechanisms beyond agent decomposition alone.

TSASSISTANT addresses both challenges through a complementary architecture combining soft constraints, hard constraints, and human oversight (Figure 1). At the soft-constraint level, a *hierarchical instruction architecture* (Section 3.1) organises agent behaviour through layered prompts that decouple coordination logic from domain expertise and user intent. At the hard-constraint level, each subagent executes in *context isolation* to structurally prevent erroneous retrievals from propagating across sections. A *programmatic enforcement layer* (Section 3.2) of execution hooks and persistent tool and agent memory stores further reduce, though cannot eliminate entirely, the risk of errors entering downstream context. As the final validation layer, an *interactive refinement loop* (Section 3.3, Figure 2) embeds expert oversight at the section level, enabling humans to intercept residual errors before the final TSA is generated.

3.1. 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). The three layers separate concerns that evolve at different rates and are maintained by different stakeholders: stable logic and constrains, section-specific domain knowledge curated by experts, and per-assessment user constraints.

Layer 1: System Prompts. The base layer defines agent roles, coordination logic, output structure, and quality constraints. System prompts are target-agnostic and shared across all assessments, providing a stable coordination framework that is reused without modification.

Layer 2: Domain Skill Modules. The second layer encodes section-specific expertise as reusable, composable modules. Each skill module specifies retrieval strategies, relevant databases, evidence weighting heuristics, and writing guidelines for a single TSA domain. Encapsulating domain expertise in independently versioned modules allows toxicologists to update one section's guidance (e.g., adding a new interpretation heuristic for knockout phenotypes) without affecting coordination logic or other domains.

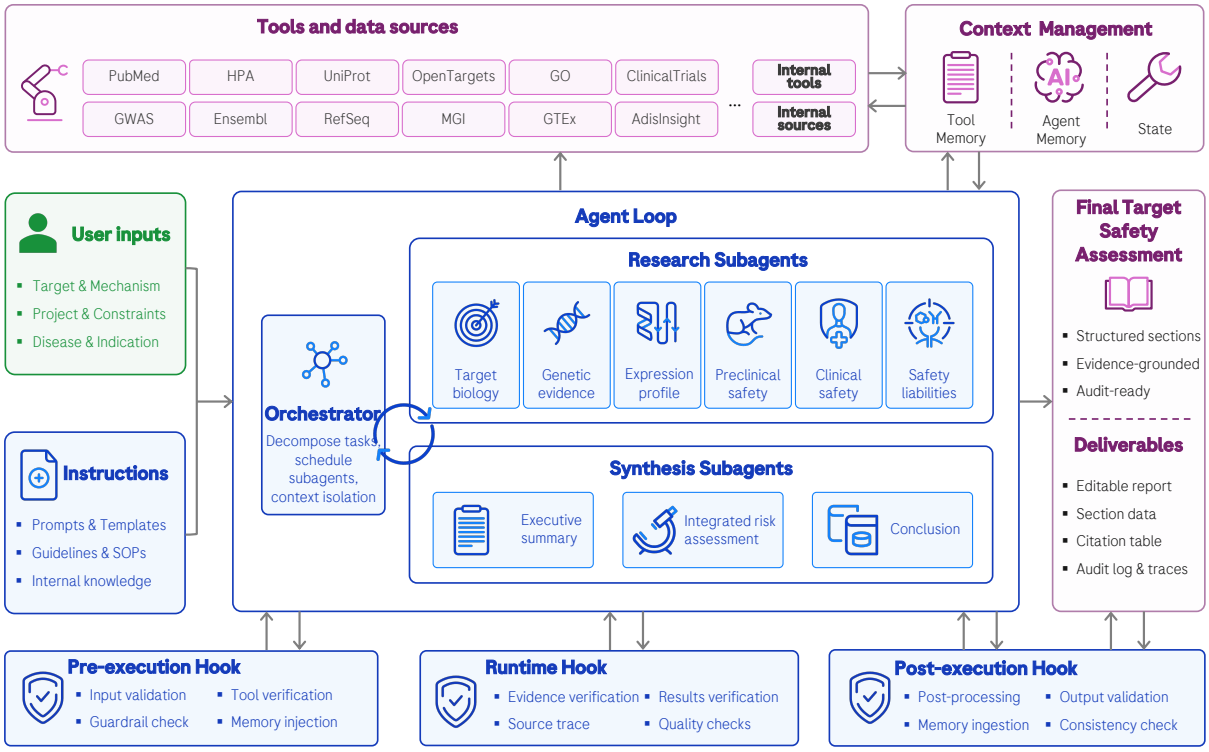

Figure 1. **Hierarchical agent architecture of TSASSISTANT.** An *Orchestrator* decomposes the assessment into *Research Subagents* and *Synthesis Subagents*, each targeting a single TSA domain. Pre-execution hooks handle security checks, memory injection, path validation, and sequential control; post-execution hooks perform citation validation, memory compression, state tracking, and output verification. Runtime hooks provide continuous monitoring (evidence verification, cross-section validation, quality checks). Context management maintains tool memory, agent memory, and execution state across the pipeline. All subagents interface with curated biomedical data sources through standardised MCP tool interfaces.

Layer 3: Runtime User Instructions. The top layer injects project-specific constraints (*e.g.*, target indication, species context, proprietary data references) at the highest priority, overriding lower-layer defaults where necessary. This ensures that each assessment can be customised without modifying the reusable layers underneath.

At runtime, the three layers compose hierarchically, with higher layers overriding lower ones where project context demands. Because prompts are soft constraints that bias but do not guarantee model behaviour, the instruction architecture alone is insufficient for safety-critical applications; the programmatic enforcement mechanisms in the next section provide the complementary hard constraints.

3.2. Agent Pipeline

TSASSISTANT decomposes report generation into a pipeline of specialised subagents: six *research subagents*, each responsible for exactly one TSA evidence domain, followed by three *synthesis subagents* that integrate findings across sections. Section-based decomposition enables independent grounding and citation per section, targeted revision without full-report regeneration, and section-level human review. However, decomposition alone does not ensure reliability. The pipeline enforces correctness through three complementary hard-constraint mechanisms: structured tool interfaces that integrate data processing and evidence indexing before results reach the LLM, programmatic execution hooks that validate outputs at each stage, and persistent context management that maintains a shared tool memory and agent memory across the pipeline.

Tool Interfaces. Each subagent retrieves evidence through MCP-standardized tool interfaces (Anthropic, 2024) connected to curated public biomedical databases and proprietary internal data sources, organized by TSA evidence domain to mirror the five-section report structure and reproduce the workflow of human experts (Section 2.1).

Crucially, the designed tools do not expose a raw API to the LLM. Each MCP server encodes a multi-step processing pipeline at different levels. Target-agnostic data-engineering level processing includes filtering, de-duplication, cross-referencing and aggregation before returning structured, analysis-ready outputs. Where relevant, tools additionally embed domain-specific analytical logic, such as variant effect or expression based thresholding and computational

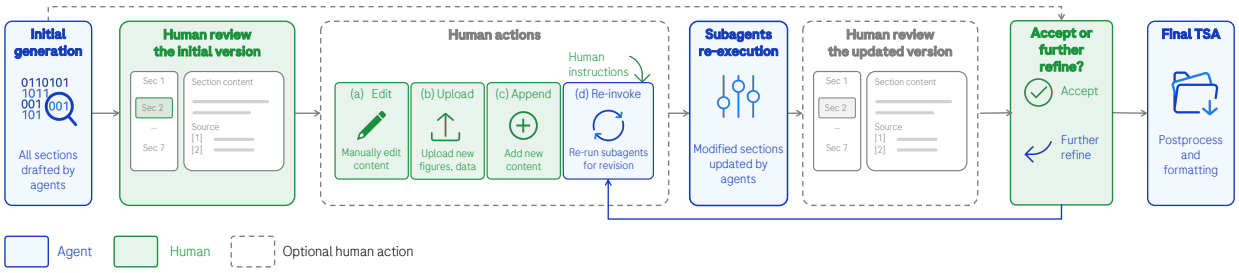

Figure 2. **Interactive refinement loop in TSASSISTANT.** After initial section generation, the user reviews each section and may: **(a)** manually edit content, **(b)** append new information, **(c)** upload additional sources or graphics, or **(d)** re-invoke the subagent for targeted revision. Conversational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

processes, that would otherwise require separate human expert intervention. Every returned record is automatically indexed in the tool's paired memory store, so downstream components operate on structured, provenance-tagged evidence rather than raw query results. This design mirrors the analytical workflow a domain expert would execute manually, and enforces consistent methodology and reference across targets. The framework additionally integrates internal proprietary tools, of which the details are withheld for confidentiality.

The data sources backing these tools span the five evidence domains, 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[1] supplies clinical study records including safety endpoints and adverse event reports from interventional trials. AdisInsight[2] provides competitive intelligence, including known severe adverse events associated with drugs modulating the same target.

Execution Hooks. The pipeline wraps each subagent's execution lifecycle with programmatic hooks (Figure 1). *Pre-execution hooks* perform security checks, inject cross-section memory from upstream dependencies, validate output paths, and enforce sequential control to prevent concurrent subagent conflicts. *Post-execution hooks* validate inline citations against retrieved evidence via natural language inference, compress the section output into dense factual representations for downstream subagents, track execution state for resumability, and verify output integrity.

[1] https://clinicaltrials.gov
[2] https://adisinsight.springer.com

Runtime hooks provide continuous monitoring including evidence verification, cross-section consistency validation, and quality checks throughout execution.

Context Management. Each subagent executes in an isolated context with no visibility into other subagents' inputs, outputs, or retrieved evidence. This *context isolation* prevents context pollution, keeps each agent focused on its assigned domain, and avoids context window overflow. To bridge isolated agents in a controlled manner, the pipeline maintains three categories of persistent context (Figure 1).

Tool memory is a persistent, structured store that records every output returned by tool calls throughout the pipeline. Each record is tagged with provenance metadata (invoking subagent, tool, query, pipeline stage), assigned a globally unique identifier, and supports full CRUD operations. The design principle is to externalise knowledge from the LLM's ephemeral context into a durable, queryable store: downstream subagents and post-execution hooks can look up, cross-reference, or invalidate any previously retrieved evidence without relying on the LLM to remember it. This decouples evidence persistence from the model's context window. During interactive refinement, this allows the user to verify a agent generated claim against the original evidence text without re-executing the search.

Agent memory stores compressed factual representations of completed research sections, preserving all quantitative data, tables, and risk classifications while removing prose, and selectively injects them into downstream subagents based on a predefined dependency graph. Compression is necessary because injecting full upstream sections would exceed context limits; selective injection avoids flooding subagents with irrelevant cross-section data.

Execution state is persisted across the pipeline, enabling session resumability: if execution is interrupted, the system resumes from the last completed section rather than restarting, preserving all accumulated tool memory 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 evidence synthesis 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 target safety decisions and the practical limitations of current LLMs in resolving ambiguous or conflicting 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 self-consistency across independent system runs. External evaluation is, however, inherently difficult to scale: evidence selection and interpretation involve expert judgement, and two independently authored reports on the same target may legitimately differ in scope and emphasis while both being scientifically valid. We therefore evaluate TSASSISTANT through a *self-consistency* paradigm, measuring claim-level concordance across independent runs on the same target at the safety endpoint level. This provides a rigorous, scalable reliability estimate that directly probes the properties most consequential for safety decision-making: factual stability across runs and citation consistency across biological contexts.

4.1. Safety Endpoint Concordance

The integrated risk assessment produced by TSASSISTANT consolidates findings from all upstream evidence domains into a structured table of toxicity endpoints. Fifteen standardised endpoints [3] are defined in alignment with MedDRA System Organ Classes (Brown et al., 1999). We formalise pairwise consistency measurement as a *structured concordance problem*: given two independent TSASSISTANT runs on the same target, how well do they agree on the specific evidentiary claims supporting each safety endpoint? We

[3] hepatotoxicity, renal toxicity, bone marrow toxicity/hematotoxicity, neurological/CNS/neurodevelopment toxicity, cardiovascular toxicity, reproductive/developmental toxicity, ocular toxicity, inflammatory syndrome, immune dysfunction, metabolic dysregulation, skin and subcutaneous tissue disorders, gastrointestinal toxicity, pulmonary toxicity, endocrine disruption and pregnancy/puerperium/perinatal conditions

design a symmetric metric that captures claim-level concordance without designating either run as ground truth.

Claim-Level Concordance. Holistic text similarity fails to capture whether two reports cite the same specific observations from the same biological contexts. Following FActScore (Min et al., 2023), which introduced the paradigm of decomposing generated text into independently verifiable atomic facts, and SAFE (Wei et al., 2024), which extended this to long-form F1 aggregation, we decompose each endpoint's findings into atomic claims.[4]

Unlike prior work where each atomic fact is a plain-text statement, toxicological evidence is inherently structured: the same phenotype observed in different species or from different evidence types carries fundamentally different implications for human risk.

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:

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

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 definition reflects a hierarchy of biological translatability informed by regulatory toxicology practice (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2011) [5].

[4] 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.

[5] The species group is defined as follow: *rodent* (mouse, rat, hamster, guinea pig), *non-rodent* (NHP including cynomolgus macaque and marmoset; dog; minipig; rabbit), *human* (encompassing human genetic evidence, clinical adverse event data, and human-derived *in vitro* systems such as iPSC-derived cells and primary cells), and *alternative screening models* (zebrafish embryo). Non-human *in vitro* systems (*e.g.*, hERG assays, rodent hepatocytes, genotoxicity panels) are classified as experimental modalities.

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 ($\phi_s = 0.75$). Inter-species differences within a group indicate that same-group 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 informative but not always predictive of human outcomes. The reasons are numerous and include inter-species differences in gene expression, target abundance, receptor/channel functionality, downstream signaling pathways, metabolism, tissue distribution, and compensatory physiological mechanisms. Contradictory claims (*e.g.*, "dose-related neutropenia" *vs.* "no significant haematologic toxicities") are explicitly scored as zero by the LLM judge.

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.

4.2. Experimental Setup and Results

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 Gemini 2.5 Flash; similarity assessment uses Claude Opus 4.6. Claim pairs that fail the species context filter ($\phi_s = 0$) are excluded from the matching rather than scored as zero, as direct translata-

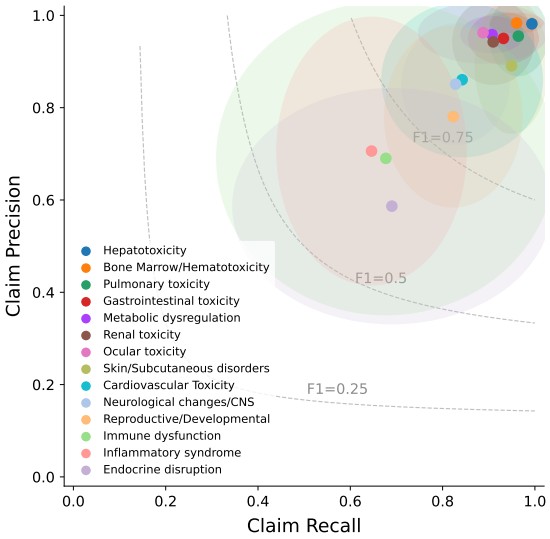

Figure 3. **Claim-Level Self-Consistency by Safety Endpoint.** Symmetric Precision and Recall (Section 4.1) for each of the 15 endpoints, aggregated over 45 run pairs per target across 35 targets. The shaded ellipsoidal regions around each endpoint are axis-aligned and span ± 1 sample standard deviation 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.

bility 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 analogous to Krippendorff's α (Krippendorff, 2011).

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 1 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. Additional retrieval guidance, refined skill modules, or more targeted human review in the refinement loop (Section 3.3) are most warranted in subsequent iterations.

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. Anticipated challenges in deploying systems of this kind include sensitivity to ambiguous or underspecified target queries, variability in evidence prioritisation 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 represents a critical and resource-intensive step in drug discovery, requiring significant time and expert effort across the pharmaceutical industry to synthesise heterogeneous evidence and guide downstream experimental and clinical decisions. While recent advances in LLM-based systems have aimed to support such workflows, TSASSISTANT is, to our knowledge, among the first frameworks specifically tailored to the needs of toxicologists and pharmaceutical scientists. TSASSISTANT aims to reduce the mechanical burden of evidence synthesis and report drafting, leaving safety judgement to expert reviewers while providing a structured and accessible interface for leveraging recent advances in agentic AI within a regulated, high-stakes context.

In summary, we build TSASSISTANT to produce traceable, evidence-grounded TSA reports while reducing the manual drafting and reviewing efforts required of human experts. The interactive refinement loop is built to preserve expert oversight at every stage, supporting progressive personalisation alongside scientific accountability. We hope TSASSISTANT contributes to a broader vision of AI-augmented pharmaceutical safety science where agentic AI and human expertise are complementary.

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.

Krippendorff, K. Computing krippendorff's alpha-reliability. 2011.

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.

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.

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. Evaluation Prompt Specifications

Evaluation Agent Instructions

```
# Self-consistency evaluation


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
 exampple, "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".


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, no_match.


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

Figure 4. Full instruction prompt used for the evaluation agent.

Toxicity Finding Instructions

```
# Granular Toxicity Endpoint


HUMAN_EXTRACTION_PROMPT = """You are a pharmacology/toxicology expert. Read the
following Target Safety Assessment report and extract findings for each of the 15
predefined toxicity endpoints.

Target: {target}

Report:
\"\"\"
{report_text}
\"\"\"

For each of the following 15 toxicity endpoints, extract the relevant findings from the
report. If the report does not mention a particular endpoint, mark it as not reported.

Endpoints to check:
1. Hepatotoxicity
2. Renal toxicity
3. Bone Marrow Toxicity/Hematotoxicity
4. Neurological changes/CNS/Neurodevelopment
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. Pregnancy, puerperium and perinatal conditions
13. Gastrointestinal toxicity
14. Pulmonary toxicity
15. Endocrine disruption

Return a JSON object where keys are these exact endpoint identifiers and values are
objects with "reported" (boolean) and "finding" (string or null):

{{
  "hepatotoxicity": {{"reported": true, "finding": "summary of what the report says
  about hepatotoxicity"}},
```

```
  "renal_toxicity": {{"reported": false, "finding": null}},
  ...
}}

Use these exact keys: hepatotoxicity, renal_toxicity,
bone_marrow_toxicity_hematotoxicity, neurological_changes_cns_neurodevelopment,
cardiovascular_toxicity, reproductive_developmental_toxicity, ocular_toxicity,
inflammatory_syndrome, immune_dysfunction, metabolic_dysregulation,
skin_and_subcutaneous_tissue_disorders, pregnancy_puerperium_and_perinatal_conditions,
gastrointestinal_toxicity, pulmonary_toxicity, endocrine_disruption

Rules:
- Only extract what the report explicitly states. Do NOT infer or add information.
- The "finding" should be a factual summary of what the report says, preserving key
details (species, mechanism, specific observations).
- If the endpoint is discussed but no safety concern is identified, still mark
reported=true and summarize what was said.
- Return ONLY valid JSON."""


# Claim extraction

EXTRACTION_PROMPT = """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 homozygos 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 5. Full instruction prompt used for the toxicity endpoint.

B. Claim Comparison Statistics

Table 1 provides a full breakdown of all the claim pairs generated during the self-consistency evaluation described in Section 4.1, stratified by species relation and scoring rule. Each pair is scored using the two-factor scheme of Figure 4: 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 1. 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 4. 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 67.4% of all records		**Diff. Group** 31.7% of all records		**Same Group** 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%
	No Match (Judge)	3.37%	2.27%	4.33%	1.37%	10.23%	0.09%