

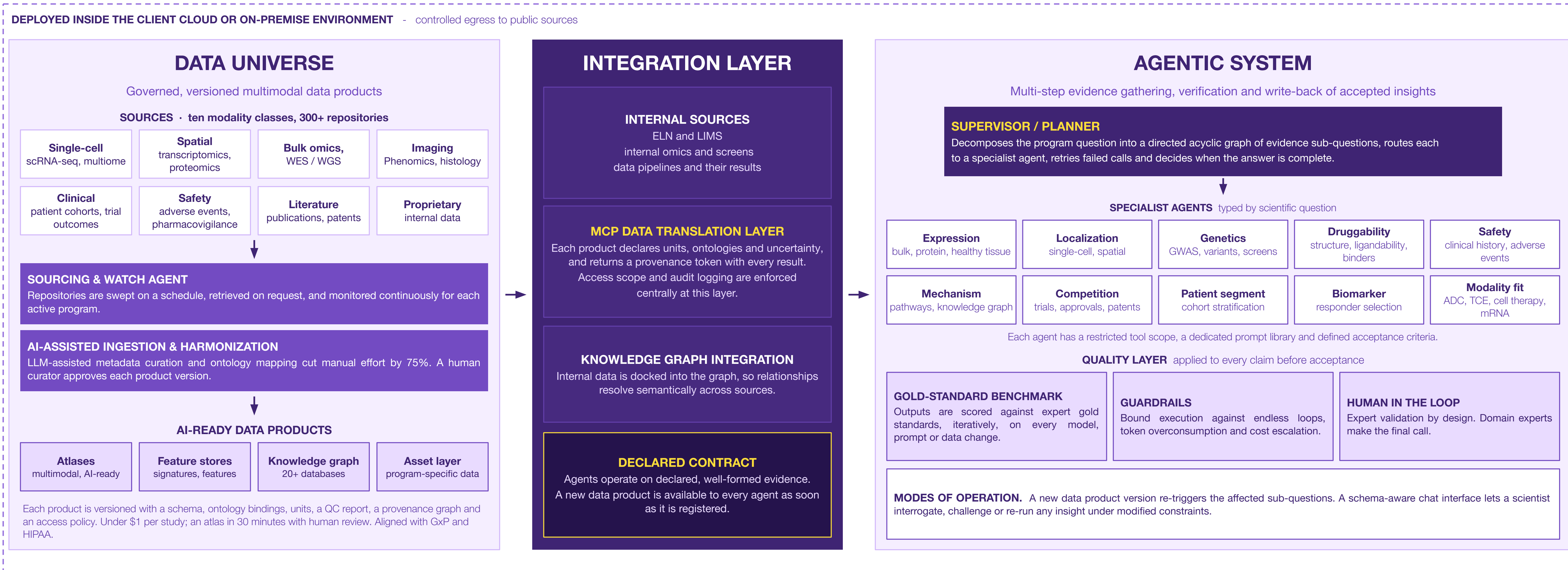


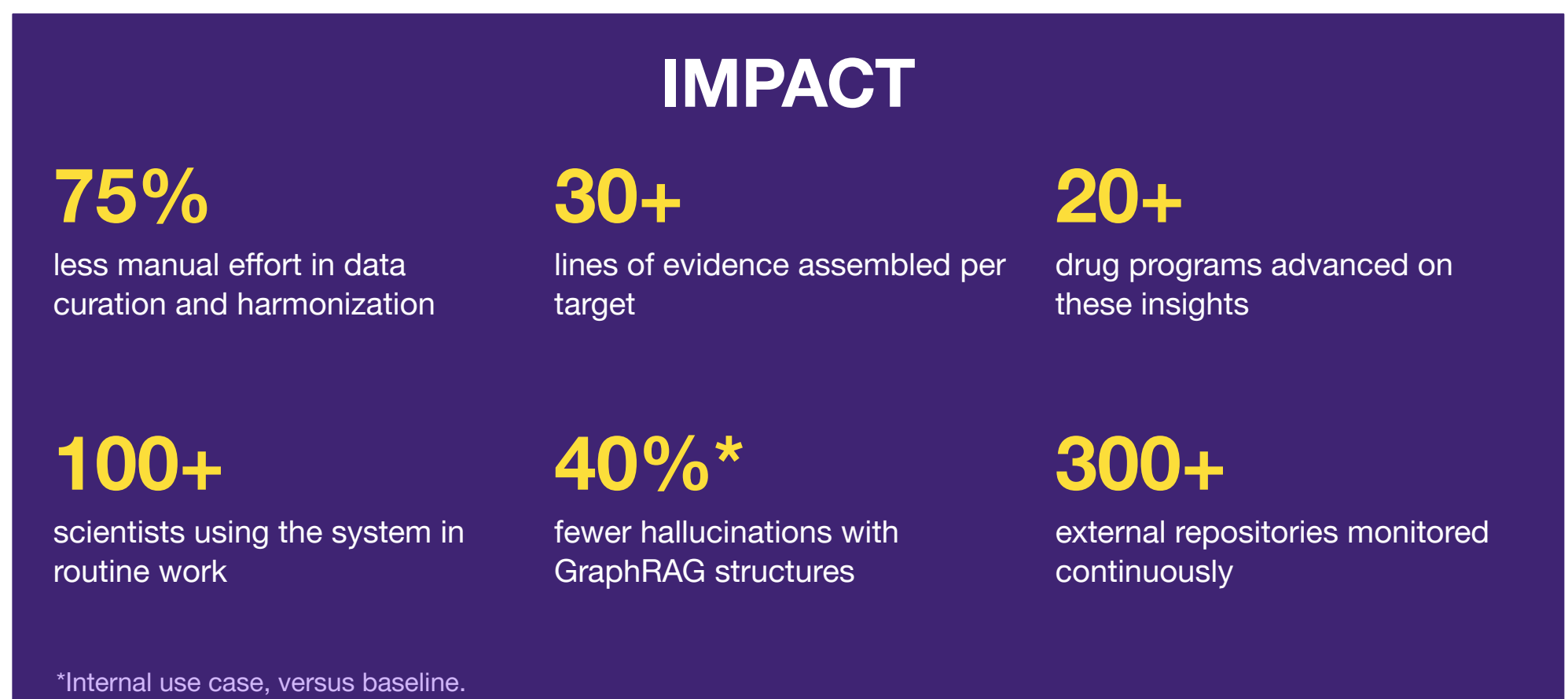
Fig. 1. Scheme of the system, deployed inside the client environment. Multimodal sources are compiled into governed, versioned data products in the Data Universe. All data access runs through an integration layer served over MCP, where a knowledge graph docks internal data into public evidence. In the agentic system, a supervisor decomposes each program question and routes sub-questions to specialist agents, whose outputs pass a quality layer. The result is a portfolio of program insights, each traceable to the records that produced it.

ABSTRACT

Only about 11% of agentic AI systems reach production^[7]. The deciding factor is the system around the model. In drug discovery the gap is sharper, because a drug-program question is never one question. Target selection, modality choice and patient segmentation rest on dozens of dependent sub-questions across incompatible data modalities. Single-agent retrieval flattens that structure, and it answers past the evidence as fluently as within it.

We present a production multi-agent system in which governance sits upstream of reasoning. A Data Universe of versioned data products spans many data modality classes, each with a schema, ontology bindings, a quality control and audit-readiness, kept current by a monitoring agent watching external sources. An integration layer serves each product over Model Context Protocol (MCP), declaring units, ontologies and uncertainty and returning a provenance token with every result. A knowledge graph docks internal data into public evidence. A supervisor agent decomposes each program question into a directed acyclic graph (DAG) of evidence sub-questions, routed to specialists typed by scientific question.

Every output is gated before acceptance: bounded by guardrails, and validated by domain experts within a system scored against gold-standard benchmarks. The output is a panel of program insights, core classes plus modality-specific sets for ADC, bispecific, cell therapy and mRNA vaccine programs.



BACKGROUND

- Human genetic evidence currently supports about two-thirds of FDA-approved drugs, and genetically supported targets succeed more often in the clinic^[1,2]. Evidence for a single target is spread across hundreds of sources, each with its own semantics and identifiers.
- Metadata created by hand in the lab multiplies the work needed to prepare AI-ready data. Harmonization still occupies curators for weeks per dataset.
- Knowledge graphs integrate such sources and support target identification through link prediction and path-based reasoning^[3-6].
- Current practice relies on commissioned static reports, fixed at delivery, or on generic scores with fixed dimensions.
- Where a tool returns an untyped table, the agent must infer units, ontology and confidence, reintroducing at the reasoning layer the fragmentation removed at the data layer.

AI FOR DATA PRODUCT ENGINEERING

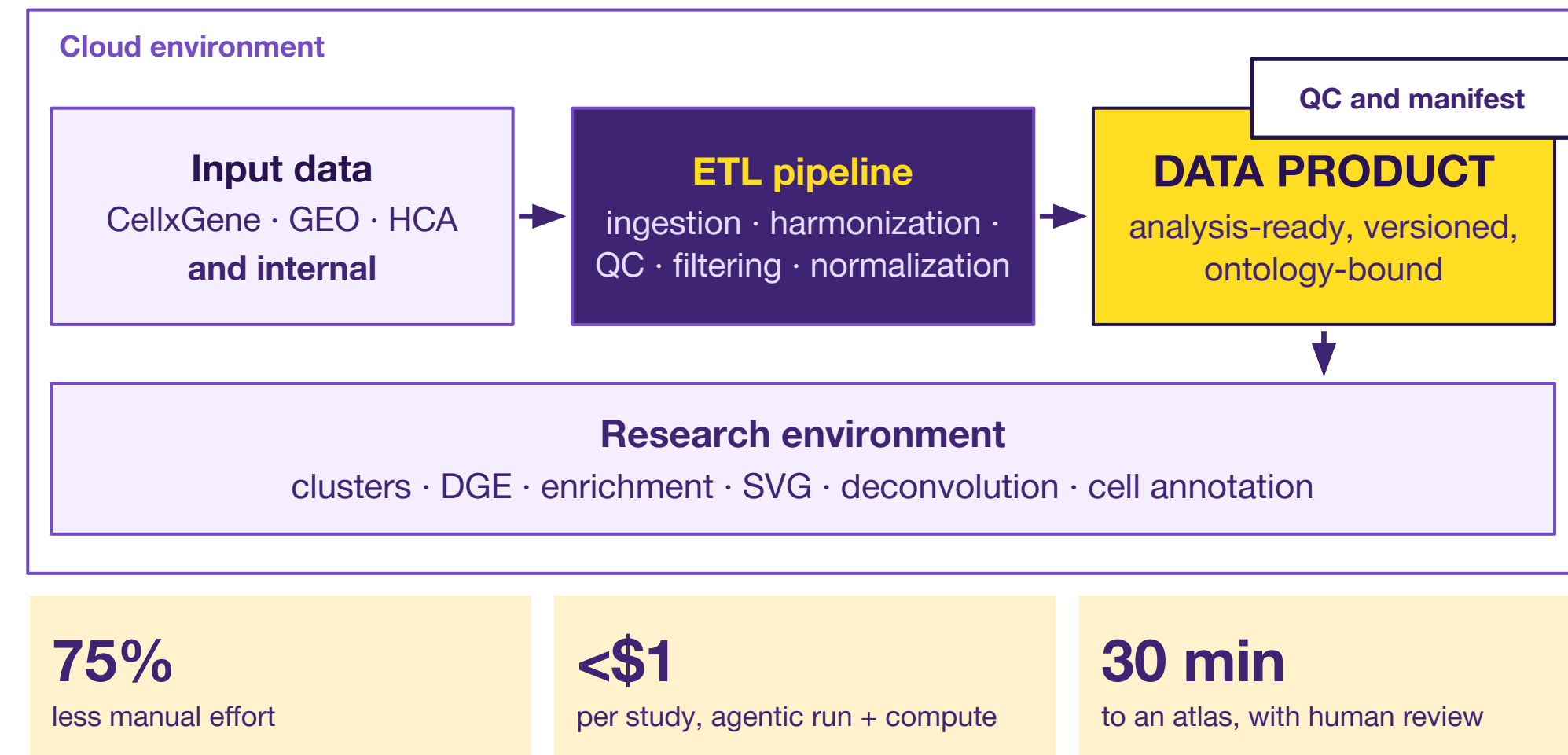


Fig. 2. LLM-assisted ETL and metadata curation inside the client cloud environment. Each data product is published with its QC report.

METHODS

1. Data products

- Each product is a versioned artifact with a schema, ontology bindings (Cell Ontology, UBERON, MONDO, EFO, Oncotree, ChEMBL), a QC report and an access policy.
- Multiple modality classes spanning single-cell, spatial and bulk omics, WES/WGS, imaging, clinical and trial records, adverse events, literature, patents and proprietary internal data. The design is aligned with GxP and HIPAA.

2. Continuous sourcing and ingestion

- An agent monitors data sources on a schedule, enabling ad-hoc retrieval or a standing watch per active program.
- LLM-assisted ETL runs metadata extraction and ontology harmonization, with GPU-accelerated data processing and feature stores (Fig. 2). A human curator approves each new product version.

3. Semantic and integration layer

- A knowledge graph over 20+ databases is built with BioCypher on a Biolink-conformant schema^[4,5], linking proprietary assets to public evidence and contributing 15+ expert-derived scores.
- Each product is published over MCP^[8], standard or custom, with typed operations and a capability manifest; agents reference declared fields only.

4. Agent topology

- A supervisor decomposes the program question into a DAG of evidence sub-questions, routes each to a specialist, retries what fails and decides when the answer is complete (Fig. 3).
- Specialists are typed by scientific question, in a many-to-many relation with data products.

5. Quality control and guardrails

- A gold-standard benchmark scores outputs against program questions answered independently by domain experts, re-run iteratively on every model, prompt or data change. A scientific reviewer agent checks each output against its sources (Fig. 3).
- Guardrails bound loops, token use and cost. Human-in-the-loop checkpoints are built into the design; domain experts make the final call.

6. Insight representation, audit and deployment

- An insight is stored as a claim, an evidence chain (product versions, query hashes, graph paths, references), an effect size with uncertainty, a quality verdict and modality flags; insights are versioned and diffed.
- Every agent action, tool call, retrieved record, prompt and model version is written to an audit trail.
- Deployment runs inside the client environment with role-based access control.

AGENTIC ORCHESTRATION SCHEME

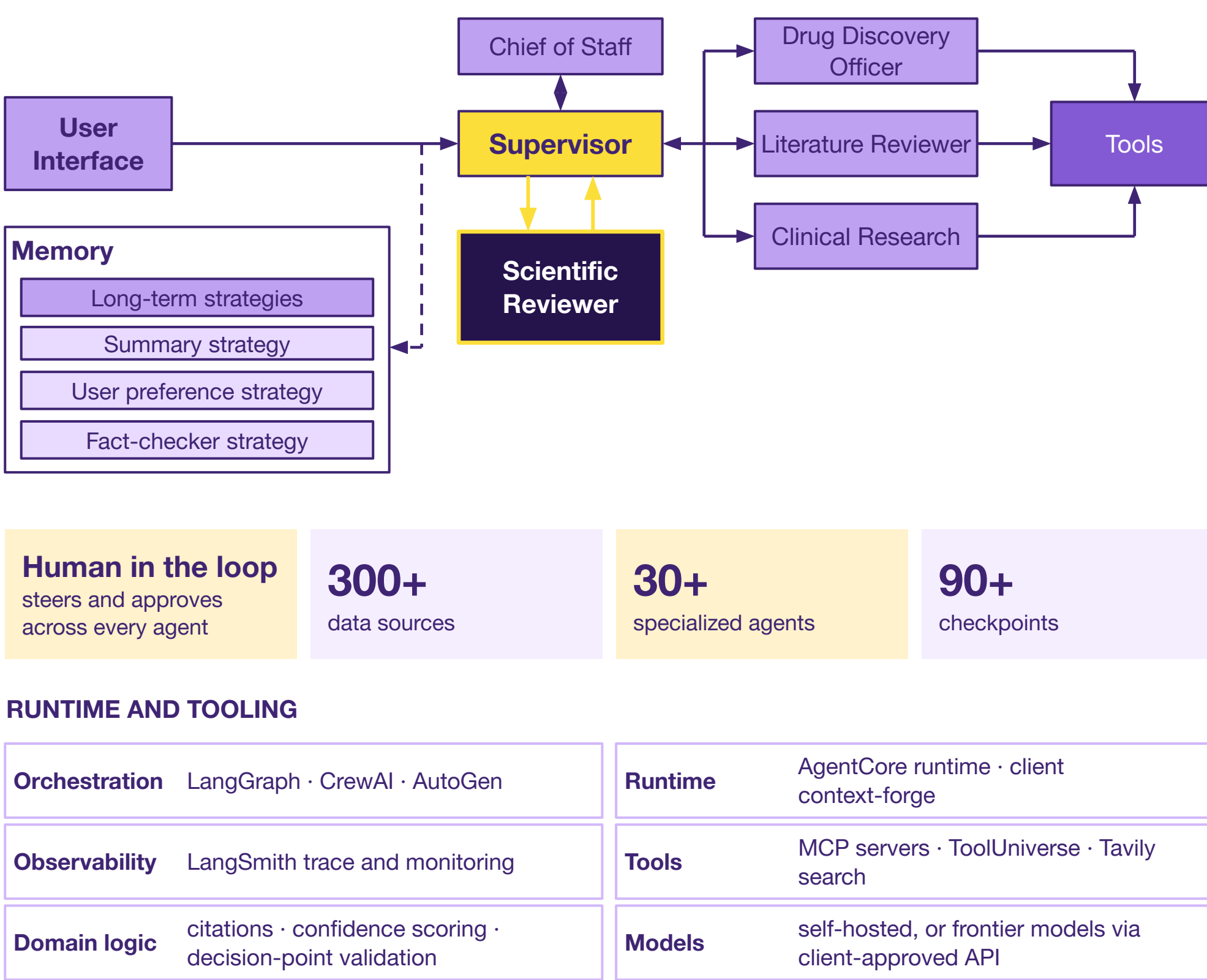


Fig. 3. Agent swarm as deployed. A supervisor plans the approach and routes work to specialist agents with access to shared tools; a scientific reviewer checks every output against its source and against gold standards; memory strategies carry program context across runs.

THE INSIGHT PANEL

Insight class	Biological question	Technical task
Expression	Is the target abundant enough in diseased tissue?	Differential expression at the gene and protein level across indications
	Where else is it present in healthy tissue?	Expression across normal tissues, resolved per tissue
Localization	Which cells carry the target?	Per-cell-type expression and prevalence, diseased versus matched healthy tissue; cell-surface localization annotated
	Is it colocalized in the tissue?	Spatial transcriptomics and proteomics with neighborhood and proximity analysis
	Is it on cell surface?	Protein localization records aggregation and mining
Genetics	Does the target drive the disease?	GWAS, rare-variant and perturbation-screen association
Druggability	Can it be drugged?	Pocket and ligandability analysis against clinical precedent
	Are binders available?	Structural databases, patents, commercial catalogs; epitope and affinity annotation
Safety	What followed prior agents against this target?	Adverse events and clinical history resolved per organ and event class, scored against background reporting rates
Competition	Who else is developing against it?	Trial, approval and patent landscape counted by owner and expiry horizon
Patient segment	Which patients are most likely to respond?	Molecular stratification across annotated clinical cohorts

Table 1. The insight panel. Each class pairs the biological questions a program already has with the technical task run to resolve them. Each class resolves into several independent lines of evidence, giving 30+ per target.

MODALITY MODULES

 technical tasks added per intended drug modality

- ADC** Target internalization and lysosomal trafficking evidence mining. Surface copy number per cell from single-cell and proteomics data. Expression heterogeneity. Exposure mapping in payload-sensitive normal tissues.
- Bispecific & TCE** Single-cell co-expression. Spatial proximity of effector T cells to target-positive cells. Cytokine-release signals from prior agents' adverse-event records.
- Cell therapy** Antigen heterogeneity and loss in relapse cohorts. T-cell infiltration and persistence markers. Immunosuppressive microenvironment scoring from spatial data and deconvolution.
- mRNA vaccine** Tumor-specific antigen expression and clonality. HLA binding and presentation prediction. Immunogenicity scoring. HLA coverage across the target population.

WORKED EXAMPLE

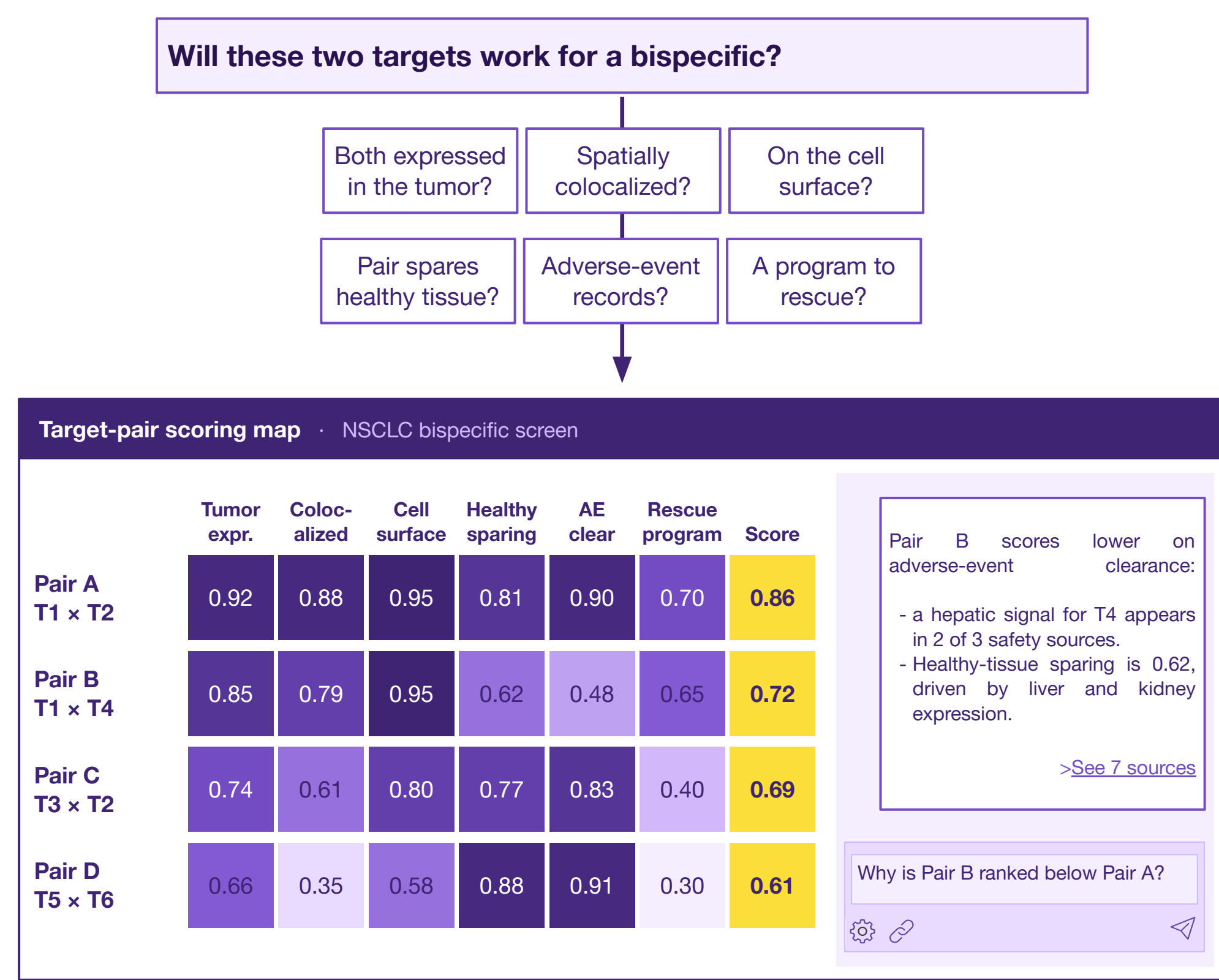


Fig. 4. Illustrative target-pair question for a bispecific program. The supervisor decomposes it into sub-questions (top); each sub-question becomes a scored line of evidence per target pair (bottom left), interrogable through the conversational interface (bottom right, mock data).