# CellMaster: A Collaborative LLM Agent for Cell Type Annotation in Single-Cell Analysis

**Anonymous authors**
Paper under double-blind review

## Abstract

Cell-type annotation remains difficult for rare, transitional, and context-dependent populations because fixed references have incomplete tissue and state coverage. We present CellMaster, an expression-grounded LLM agent that iteratively generates hypotheses, proposes marker panels, inspects cluster-level expression and detection patterns, and assigns labels with auditable rationales. Given a preprocessed scRNA-seq object and short study context, CellMaster-Core annotates without querying an external annotated atlas, fixed marker database, or specialist tool at inference time. Rather than forcing a label, it preserves unresolved clusters for another iteration, targeted sub-clustering, or selective escalation to a specialist or human expert. We evaluate this design with a nine-dataset breadth benchmark and a deeper comparison on three biologically distinct datasets against contemporary agents, zero-shot scGPT, and reference-based specialists. Core outperforms Biomni on each depth dataset, with up to a $17.0\%$ relative increase in Cell Ontology score, while optional CellTypist closure provides smaller additional gains. Under fixed closure, removing marker memory causes the largest degradation among the tested components. The collaborative interface exposes the same expression evidence used by the agent and records interventions, making uncertainty and escalation auditable rather than hiding them behind a single final label.

## 1 Introduction

Cell-type annotation is a bottleneck in single-cell RNA sequencing (scRNA-seq): rare populations, transitional states, and context-dependent subtypes often cannot be identified from a fixed marker list. Reference-transfer and supervised tools can be highly effective when target tissue and label vocabulary are well represented in references, but behavior degrades when coverage, nomenclature, or preprocessing shifts. They also provide limited support for revising an annotation from explicit, cluster-specific expression evidence when the reference itself is uncertain Luecken & Theis (2019); tab (2020); Domínguez Conde et al. (2022). Manual annotation remains effective but does not scale to atlas-sized datasets.

LLM-based annotators can reason over per-cluster differentially expressed (DE) genes and textual context Achiam et al. (2023); Hou & Ji (2024). Recent systems such as scExtract Wu & Tang (2025), ReCellTy Han et al. (2025), CellTypeAgent (Chen et al., 2025), and scPilot (Gao et al., 2026) extend this direction with retrieval or agentic workflows. The unresolved question is how an annotator should expose and act on uncertainty: whether to retain a label, revisit it with new evidence, consult a specialist, or defer to a human. A one-shot label and multi-agent decomposition alone do not specify this decision process.

We frame the missing capability as *selective escalation*. A useful annotator should produce expression-grounded labels and rationales, stabilize confident clusters, and treat uncertain clusters as explicit states that can be iterated, sub-clustered, routed to a specialist, or escalated to a human. The same evidence trace should support autonomous and collaborative operation. This separates two questions that are often conflated: whether an agent can propose a biologically plausible label, and whether a system can decide when that proposal is sufficiently supported to retain, when it should seek another source of evidence, and when it should abstain.

We introduce CellMaster, an iterative annotation agent with Hypothesis, Proposal, and Annotation stages. Given a preprocessed scRNA-seq object, a short study context, and cluster-level DE evi-



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

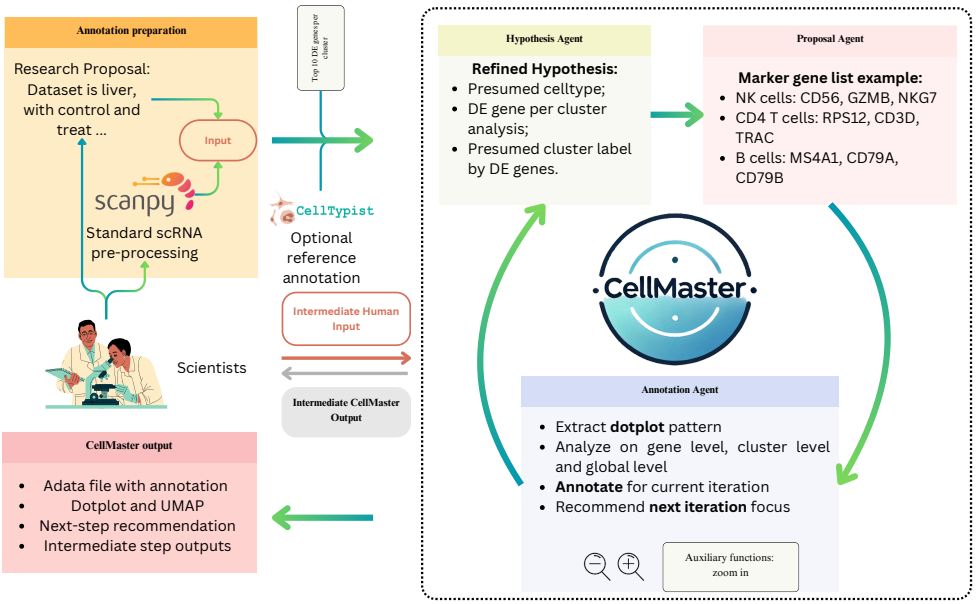



Figure 1: Workflow of CellMaster. Scientists provide a Scanpy object and a one-paragraph hypothesis; three specialised agents (Hypothesis, Proposal, Annotation) iteratively propose, test, and revise per-cluster labels. Intermediate outputs are surfaced in a collaborative UI for optional human-in-the-loop refinement.

dence, the system proposes markers, inspects expression and detection fractions in dot plots, and returns labels, confidence, and rationales. Marker memory suppresses failed proposals, confidence pinning (stabilization) preserves resolved clusters, and abstention preserves uncertainty. The interface records edits and provenance. We distinguish *CellMaster-Core* (reference-free reasoning), *CellMaster-Ecosystem* (confidence-gated specialist assistance), and *CellMaster-HITL* (expert collaboration).

Our evaluation separates breadth from depth. The breadth benchmark spans nine datasets and eight tissues using the original baseline suite and reports the automatic CellMaster+AF setting. The depth benchmark uses PBMC3k, Liver, and Retina, where matched runs are available for contemporary LLM agents, zero-shot scGPT, reference-based specialists, and CellMaster-Core without AF. Core obtains Cell Ontology scores of $0.708$, $0.536$, and $0.667$ on these datasets, respectively; applying CellTypist to residual abstentions increases the corresponding scores to $0.725$, $0.550$, and $0.705$.

Our contributions are:

- An expression-grounded iterative annotation agent with marker memory, confidence pinning, abstention, and auditable rationales.
- A selective-escalation architecture that unifies reference-free reasoning, specialist-tool reconciliation, targeted sub-clustering, and provenance-tracked HITL correction.
- A two-scale evaluation comprising nine-dataset breadth and a contemporary PBMC3k/Liver/Retina comparison with component ablations, a post-cutoff evaluation, local backbones, and expert collaboration.

## 2 Related Work

**Single-cell analysis pipelines.** scRNA-seq analysis typically proceeds through preprocessing, quality control, dimensionality reduction, unsupervised clustering, cell-type annotation, and downstream analyses such as lineage / trajectory inference, cell–cell interaction modelling, and dataset integration Luecken & Theis (2019). Each step has spawned dedicated tools, with hundreds of algorithms used for scRNA-seq.

**Cell-type annotation.** Annotation strategies include marker-table methods such as CellMarker 2.0 (Hu et al., 2023), reference-transfer methods such as Seurat label transfer (Stuart et al., 2019), and supervised classifiers such as CellTypist (Domínguez Conde et al., 2022) and scTab (Fischer et al., 2024). These methods can be highly effective when reference coverage, nomenclature, and preprocessing match the target, but performance can degrade under tissue, state, or batch shift and the resulting decisions are not always accompanied by an auditable expression-grounded rationale. Manual expert annotation remains the reference practice but scales poorly to atlas-sized data.

**LLMs and agentic systems for biology.** Recent work positions LLMs as biomedical reasoners, building on the transformer architecture Vaswani et al. (2017) and generative pretraining Radford et al. (2018). Beyond passive prediction, biomedical AI agents proactively acquire information, invoke external tools, and refine hypotheses against evidence Gao et al. (2024); Boiko et al. (2023); M. Bran et al. (2024). For cell-type annotation specifically, GPT-4 can recover labels from per-cluster DE genes Achiam et al. (2023); Hou & Ji (2024); downstream systems such as scExtract Wu & Tang (2025) and ReCellTy Han et al. (2025) pair LLM-driven retrieval with curated validation. CellTypeAgent (Chen et al., 2025) uses a multi-agent workflow for cell annotation, scPilot (Gao et al., 2026) studies LLM reasoning for automated single-cell analysis, and MAT-Cell (Yang et al., 2026) explores tree-structured multi-agent reasoning. CellMaster is closest in spirit to these agentic annotators, but emphasizes an iterative expression-grounding loop, memory and stabilization across iterations, and a shared uncertainty state that can be escalated to specialists or humans.

**Reference-based and ecosystem annotation.** SCimilarity (Heimberg et al., 2025) learns a pan-tissue single-cell representation and performs annotation through similarity to annotated references, while PopV (Ergen et al., 2024) combines multiple reference-transfer and classification methods through ontology-aware consensus. TranscriptFormer (Pearce et al., 2025) studies a broader cross-species generative single-cell representation regime. These methods require external reference resources and are therefore evaluated separately from our reference-free CellMaster-Core. We use them as optional specialists in CellMaster-Ecosystem, where the agent reconciles their proposals with marker and dot-plot evidence rather than treating correlated tools as independent votes.

# 3 METHODOLOGY

## 3.1 PROBLEM FORMULATION AND SYSTEM OVERVIEW

Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$ be a preprocessed scRNA-seq dataset with $N$ cells and $M$ genes, partitioned into clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$ by Leiden clustering (Traag et al., 2019). Let $\mathcal{Y}$ denote the label space induced by the Cell Ontology (CL) (Meehan et al., 2011). We seek a per-cluster annotation $\alpha_r : \mathcal{C} \to \mathcal{Y} \cup \{\perp\}$, where $\perp$ denotes abstention and $r$ indexes an iteration. CellMaster performs $R$ iterations (default $R = 3$); each iteration consists of hypothesis generation, marker selection, expression analysis, and result evaluation coordinated by Hypothesis, Proposal, and Annotation agents (Figure 1).

We distinguish three operating modes. *CellMaster-Core* uses the iterative agent and expression evidence and leaves unresolved clusters as $\perp$. *CellMaster-Ecosystem* optionally invokes external specialist annotators for selected uncertain clusters. *CellMaster-HITL* allows an expert to edit the hypothesis, marker proposals, or labels through the collaborative interface. Here, *reference-free* means that Core does not query an external annotated atlas, fixed marker database, or specialist annotation tool at inference time; the pretrained LLM's general biological prior is not treated as an external annotation reference.

## 3.2 ITERATIVE ANNOTATION PIPELINE

At iteration $r$, the system maintains a study hypothesis $h_r$, candidate marker sets $\mathbf{P}_r = (P_r^{(1)}, \ldots, P_r^{(K)})$, current annotations $\alpha_r$, and expression evidence computed from the preprocessed object. For each cluster $C_k$ and gene $g$, let $\tilde{x}_{i,g}$ be log-normalised expression for cell $i$. We compute

$$E_{k,g} = \frac{1}{|C_k|} \sum_{i \in C_k} \tilde{x}_{i,g}, \qquad F_{k,g} = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{1}[\tilde{x}_{i,g} > 0], \tag{1}$$

where $E_{k,g}$ is mean expression and $F_{k,g}$ is the detection fraction.

These quantities persist across iterations rather than serving a one-shot prompt. The next hypothesis receives current labels and unresolved cases; marker memory carries informative genes forward; and pinning protects accepted labels from unnecessary revision. Later iterations therefore concentrate on clusters whose evidence remains weak or contradictory while preserving why earlier decisions were accepted.

**(i) Hypothesis generation.** The Hypothesis Agent receives the top-$N$ differentially expressed genes per cluster $G_k = \mathrm{top}_N(\mathrm{DE}(C_k))$, ranked by $\log_2$ fold-change, together with the previous hypothesis $h_{r-1}$ and annotations $\alpha_{r-1}$. It produces a refined $h_r$ containing candidate cell types and contextual cues such as developmental stage, perturbation context, and lineage relationships.

**(ii) Marker selection.** Conditioned on $h_r$ and marker memory $\mathcal{M}_r = (\mathcal{M}_r^{\mathrm{succ}}, \mathcal{M}_r^{\mathrm{fail}})$, the Proposal Agent emits cluster-specific marker sets $\mathbf{P}_r$. Previously successful markers can be reused, while failed proposals are suppressed in later iterations.

**(iii) Expression analysis.** For each proposed $(k, g)$, the Annotation Agent receives the evidence in Eq. 1, including dot-plot colour for expression and point size for detection fraction. It uses this evidence to confirm, revise, or reject a candidate label rather than matching marker names alone. The agent also receives gene-level specificity, cluster-level marker signatures, cluster-pair differences, and contamination flags.

**(iv) Result evaluation.** The Annotation Agent returns a proposed label $\hat{\tau}_k$, a confidence value $q_{r,k}$, and a stabilized set $S_r$ under the implemented confidence policy. Labels for clusters in $S_r$ are pinned so that later iterations cannot overwrite them; other clusters remain eligible for revision. CellMaster-Core leaves unresolved clusters as $\perp$. The CellMaster+AF setting is introduced and evaluated separately in Section 4; it is not part of the reference-free Core definition. Per-stage prompts are reproduced in Appendix L.

### 3.3 Adaptive heuristics

**Marker memory.** Marker memory records whether a proposed gene produced sufficiently strong expression and prevalence evidence in any cluster. Successful markers can be reused in later proposals, while failed markers are suppressed to reduce repeated uninformative queries. The evidence thresholds are implementation parameters; we do not assume an additional relative-maximum score beyond the expression and detection measurements in Eq. 1.

**Confidence pinning (stabilization).** The Annotation Agent returns $q_{r,k}$ for each proposed label. The implemented policy uses these confidence values and the iteration state to select $S_r$; labels in $S_r$ are pinned, while low-confidence or contradictory labels remain revisable or abstained. This policy makes uncertainty explicit without treating a model-generated confidence value as a calibrated probability.

**Cluster-pair check.** For two clusters $i$ and $j$, we count the number of genes with materially different expression or detection:

$$D_{ij} = \left|\{g \in G : \ |\log_2 E_{i,g} - \log_2 E_{j,g}| > \tau_E \text{ or } |F_{i,g} - F_{j,g}| > \tau_F\}\right|. \tag{2}$$

Pairs with $D_{ij} \leq K_{\min}$ are flagged for targeted sub-clustering at an adjustable Leiden resolution. The default implementation uses $\tau_E = 1$, $\tau_F = 0.1$, and $K_{\min} = 5$.

**Contamination rules.** Clusters with a Scrublet doublet score above a predefined dataset-specific threshold are flagged (Wolock et al., 2019). Clusters with mitochondrial-gene fraction above $10\%$ are also flagged before evaluation. The Annotation Agent can emit $\perp$ and recommend quality-control review for flagged clusters.

### 3.4 CellMaster-Ecosystem

CellMaster-Ecosystem treats CellMaster-Core as a reasoning and evidence-reconciliation layer over optional specialist tools. For an uncertain or abstained cluster, a router may request predictions from SCimilarity, PopV, CellTypist, or another registered annotator. An external candidate is retained only when its confidence exceeds the tool-specific threshold and it is not contradicted by marker

expression or the dot plot. Disagreement triggers a broader parent label, another annotation iteration, abstention, targeted sub-clustering, or HITL review. Outputs from correlated constituent classifiers are not counted as independent votes. Because specialist tools use external references, standalone and integrated results are reported separately from the reference-free CellMaster-Core benchmark. The targeted integration study in Section 4 does not imply that every dataset uses every specialist.

### 3.5 CellMaster-HITL

The interface runs locally and keeps the raw expression matrix on the user's machine. When a remote LLM backend is selected, only the text and evidence summaries required by the agent are transmitted through the configured API; fully local backbones avoid this external call. At each iteration, an expert can edit $h_r$, accept, override, or pin labels in $\alpha_r$, modify markers in $\mathbf{P}_r$, or request a zoom-in on selected clusters at adjustable Leiden resolution. Every edit is logged with provenance and re-enters the next iteration. Thus specialist and human escalation operate on the same expression evidence and uncertainty state rather than on separate annotation pipelines. Component-level details and the complete interface walkthrough are in Appendix H and Appendix Figure 6.

## 4 Experiments

### 4.1 Experimental setup

We evaluate CellMaster at two scales. The *breadth* benchmark covers nine public datasets spanning eight tissues and human/mouse data, using the original annotation baselines. The *depth* benchmark uses PBMC3k, Liver, and Retina, three representative datasets spanning immune, developmental, and neural contexts, for contemporary agents, zero-shot foundation-model inference, reference-based specialists, and component ablations. Dataset statistics are provided in Appendix Table 6.

Unless otherwise noted, *CellMaster-Core* denotes the reference-free system, *CellMaster+AF* denotes the historical CellTypist AutoFill setting applied to residual abstentions, *CellMaster-Ecosystem* uses confidence-gated specialist tools, and *CellMaster-HITL* adds expert interaction. The nine-dataset benchmark reports CellMaster+AF; no-AF Core results are available for the three depth-benchmark datasets.

**Preprocessing and inputs.** We obtained processed Seurat/AnnData objects where available and otherwise reprocessed raw count matrices following the original publications. Counts were normalised with `scanpy.pp.normalize_total` and log-transformed with `scanpy.pp.log1p` (Wolf et al., 2018); clusters and reference annotations followed the source metadata. Cluster-level methods receive identical Leiden partitions and per-cluster DE lists. Cell-level methods receive the count matrix and their predictions are majority-voted to the same clusters.

**Methods.** The original baseline suite contains GPTCelltype (Hou & Ji, 2024), CellTypist (Domínguez Conde et al., 2022), CellMarker 2.0 (Hu et al., 2023), scTab (Fischer et al., 2024), and Biomni (Huang et al., 2026). The depth benchmark adds CellTypeAgent (Chen et al., 2025), scPilot (Gao et al., 2026), zero-shot scGPT (Cui et al., 2024), SCimilarity (Heimberg et al., 2025), and PopV (Ergen et al., 2024). SCimilarity and PopV use external annotated references and are not reference-free baselines. Fine-tuned foundation-model settings are not compared directly with zero-shot CellMaster.

**Metric.** We use deterministic Cell Ontology (CL) agreement: cleaned predictions and references are mapped to CL identifiers; an exact match scores $1.0$, a direct parent/child match scores $0.5$, and all other labels, including abstentions, score $0.0$. We report cluster-macro-averaged score and coverage, the fraction of clusters receiving a non-abstained label. Mapping details and plausible ontology disagreements are in Appendix H.

**Implementation.** Experiments use Python 3.11 with Scanpy 1.10, SciPy 1.13, NumPy 1.26, and statsmodels 0.14. Differential expression uses two-sided Wilcoxon tests with Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) and $|\log_2 \mathrm{FC}| > 1$. Exact model versions, prompts, dates, and local-model settings are in Appendix I.

Table 1: **Full nine-dataset breadth benchmark (CL score).** CellMaster+AF applies CellTypist AutoFill to residual abstentions and is reported as mean±s.d. over five runs; baselines are point estimates. $\Delta$ is relative to the strongest available baseline on each dataset, identified in parentheses. "–" marks unavailable results.

| | **Reference-based** | | | **LLM baselines** | | **CellMaster** | |
|---|---|---|---|---|---|---|---|
| Dataset | CellTypist | scTab | CellMarker 2.0 | GPTCelltype | Biomni | CellMaster+AF | $\Delta$ vs. best |
| PBMC3k | 0.500 | 0.375 | 0.250 | 0.613 | 0.646 | 0.725±0.050 | +0.079 (Biomni) |
| Liver | 0.429 | – | 0.304 | 0.404 | 0.464 | 0.550±0.040 | +0.086 (Biomni) |
| Retina | 0.263 | 0.632 | 0.632 | 0.300 | 0.570 | 0.705±0.086 | +0.073 (scTab/CellMarker) |
| BCL | 0.333 | 0.500 | 0.333 | 0.611 | – | 0.656±0.042 | +0.045 (GPTCelltype) |
| Myeloid | 0.409 | 0.500 | 0.182 | 0.464 | – | 0.573±0.068 | +0.073 (scTab) |
| Large PBMC | 0.409 | 0.455 | 0.227 | 0.345 | – | 0.464±0.059 | +0.009 (scTab) |
| Brain | – | – | 0.333 | 0.650 | – | 0.650±0.063 | +0.000 (GPTCelltype) |
| Large Intestine | – | – | 0.281 | 0.431 | – | 0.506±0.072 | +0.075 (GPTCelltype) |
| Limb Muscle | – | – | 0.167 | 0.456 | – | 0.589±0.044 | +0.133 (GPTCelltype) |

Table 2: **Depth benchmark and backbone sensitivity** (CL score; higher is better). **(a)** Original baselines, contemporary agents, zero-shot scGPT, reference-based specialists, and CellMaster on PBMC3k, Liver, and Retina. Core queries no external annotation reference and leaves abstained clusters unresolved; +AF applies CellTypist to those residual abstentions. **(b)** Backbone sensitivity in the +AF setting; local/open backbones provide a privacy-preserving alternative with a capacity–performance tradeoff. Values are mean±s.d. where repeated runs are available; "–" is unavailable. No cross-dataset mean is reported.

| Method | Setting | PBMC3k | Liver | Retina |
|---|---|---|---|---|
| **(a) Contemporary comparison and AutoFill analysis** | | | | |
| *Original baselines* | | | | |
| CellTypist | pretrained classifier | 0.500 | 0.429 | 0.263 |
| scTab | pretrained classifier | 0.375 | – | 0.632 |
| CellMarker 2.0 | curated markers | 0.250 | 0.304 | 0.632 |
| GPTCelltype | one-shot LLM | 0.613 | 0.404 | 0.300 |
| Biomni | biology agent | 0.646 | 0.464 | 0.570 |
| *Contemporary agents and models* | | | | |
| CellTypeAgent | cell-annotation agent | 0.625±0.063 | 0.494±0.045 | 0.614±0.030 |
| scPilot | single-cell agent | 0.646±0.017 | 0.512±0.002 | 0.675±0.011 |
| scGPT zero-shot | no fine-tuning | 0.604±0.095 | 0.363±0.072 | 0.491±0.101 |
| *Reference-based specialists* | | | | |
| SCimilarity | external reference | 0.750 | 0.393 | 0.684 |
| PopV | external reference | 0.813 | 0.464 | 0.658 |
| *CellMaster and AutoFill effect* | | | | |
| **CellMaster-Core** | **reference-free** | **0.708±0.072** | **0.536±0.034** | **0.667±0.061** |
| **CellMaster+AF** | **CellTypist AutoFill** | **0.725±0.050** | **0.550±0.040** | **0.705±0.086** |
| Core − Biomni | per-dataset difference | +0.062 | +0.072 | +0.097 |
| **(b) Backbone sensitivity (CellMaster+AF)** | | | | |
| GPT-4o | cloud | 0.725±0.050 | 0.550±0.040 | 0.705±0.086 |
| Qwen3-4B | local/open | 0.542±0.036 | 0.399±0.027 | 0.561±0.040 |
| gpt-oss-120b | local/open | 0.667±0.072 | 0.542±0.021 | 0.684±0.053 |

### 4.2 Breadth and contemporary comparison

Table 1 establishes breadth across the nine-dataset benchmark. CellMaster+AF is competitive across eight tissues, with the largest gains on Limb Muscle (+0.133), Liver (+0.086), and PBMC3k (+0.079). The consolidated table reports the complete original method-by-dataset matrix and identifies the strongest available comparator for each dataset.

Table 3: **Component ablation on PBMC3k, Liver, and Retina.** The CellMaster+AF setting is fixed across rows. Removing marker memory produces the largest degradation on every dataset, followed by multi-agent decomposition and confidence pinning.

| Variant | PBMC3k | Liver | Retina |
|---|---|---|---|
| Full CellMaster+AF | 0.725±0.050 | 0.550±0.040 | 0.705±0.086 |
| Without multi-agent decomposition | 0.625±0.125 | 0.441±0.055 | 0.597±0.081 |
| Without marker memory | 0.583±0.072 | 0.393±0.071 | 0.544±0.081 |
| Without confidence pinning | 0.667±0.072 | 0.488±0.055 | 0.667±0.030 |

Table 4: **Unseen Liver.** CL score on a 2026 healthy-human-liver dataset with 11 retained clusters.

| Method | CL score |
|---|---|
| CellMaster | 0.606±0.053 |
| scPilot | 0.546±0.091 |
| Biomni | 0.515±0.053 |

Table 2 provides the denser method comparison on PBMC3k, Liver, and Retina. CellMaster-Core scores $0.708$, $0.536$, and $0.667$, respectively, without external annotation references, exceeding Biomni by $0.062$, $0.072$, and $0.097$; AF further increases the Core scores by $0.017$, $0.014$, and $0.038$. The reference-based specialists remain complementary: PopV is strongest on PBMC3k ($0.813$), SCimilarity is strongest on Retina ($0.684$), and CellMaster-Core is strongest on Liver among the reported methods ($0.536$). These results motivate CellMaster-Ecosystem as a reconciliation layer rather than a replacement for every specialist tool.

### 4.3 What makes the iterative agent useful?

Table 3 reports component ablations with the same CellMaster+AF setting. On PBMC3k, Liver, and Retina, removing marker memory reduces the score by $0.142$, $0.157$, and $0.161$, respectively; removing multi-agent decomposition reduces it by $0.100$, $0.109$, and $0.108$, and removing confidence pinning by $0.058$, $0.062$, and $0.038$. AF is held fixed across these rows; its per-dataset contribution is shown in Table 2. Dot-plot inspection is common to these variants and supplies the expression evidence used to confirm, revise, or reject a candidate label; this table does not isolate it as a separate causal ablation.

The method improves through the early iterations but can drift after high-confidence clusters have stabilised. We therefore use three unattended iterations as the default and route ambiguous pairs to sub-clustering or escalation; longer-run dynamics are shown in Appendix C. Table 2 also shows that local execution is feasible: gpt-oss-120b approaches GPT-4o on Liver ($0.542$ vs. $0.550$) and Retina ($0.684$ vs. $0.705$), with a larger gap on PBMC3k ($0.667$ vs. $0.725$); Qwen3-4B is lower on all three datasets. On *Unseen Liver*, an additional healthy-human-liver dataset whose journal version appeared in 2026 (Rahman et al., 2026), CellMaster obtains $0.606 \pm 0.053$ versus $0.546 \pm 0.091$ for scPilot and $0.515 \pm 0.053$ for Biomni (Table 4). This post-cutoff evaluation reduces reliance on canonical benchmarks but does not eliminate memorisation concerns.

Figure 2 stratifies performance by cluster cell count and shows that the gains are concentrated in difficult regimes, including small clusters, subtype-level labels, and neural or T-cell families. The full breakdown remains useful as a complement to the benchmark because it tests whether the method's advantage is tied only to a particular tissue or dataset size.

### 4.4 Selective escalation and collaboration

CellMaster treats uncertainty as an explicit state. In CellMaster-Ecosystem, SCimilarity, PopV, or CellTypist are invoked only for uncertain or abstained clusters; candidate labels are accepted only when specialist confidence and expression evidence agree. On the 28-cluster Liver dataset, confidence-gated SCimilarity/PopV routing increases the CL score from $0.536 \pm 0.034$ for Core to $0.589$ with full coverage, compared with $0.550 \pm 0.040$ for CellMaster+AF (Appendix Table 8).

Table 5: **Iteration dynamics on Liver** (CL score, mean±s.d. over five runs). HITL uses one sentence of expert guidance per iteration.

| Iteration | Automatic | HITL |
|---|---|---|
| 1 | 0.207±0.058 | 0.257±0.039 |
| 2 | 0.393±0.036 | 0.429±0.036 |
| 3 | 0.464±0.025 | 0.525±0.029 |
| 3 + AF | 0.561±0.027 | **0.647±0.024** |

In CellMaster-HITL, two independent biology experts annotated all nine datasets without seeing ground-truth labels. They used the normal interface, dot plots, metadata, and editable controls; full

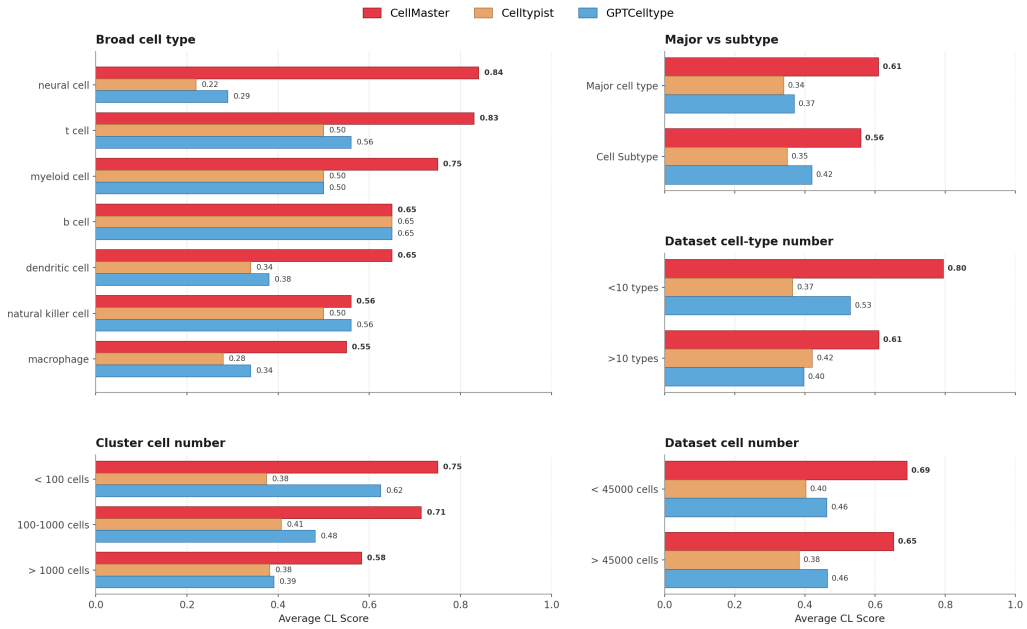



Figure 2: Fine-grained breakdown of CL-score performance, comparing CellMaster (red), CellTypist (orange), and GPTCelltype (blue). *Top left*: broad cell-type families. *Top right*: major cell types vs. subtypes. *Bottom left*: per-cluster cell-count strata. *Right column*: dataset-level cell-type-count strata and total-cells strata.

passes were not repeated by the same expert, and final annotations differed by no more than one cluster per dataset. On Liver, a single sentence of guidance per iteration improves every reported stage and raises the final iteration-plus-AF score from $0.561 \pm 0.027$ to $0.647 \pm 0.024$ (Table 5). This measures the value of expert collaboration, not superiority over automatic baselines; the full protocol is in Appendix Table 9.

### 4.5 Case study and targeted refinement

The developmental mouse Liver case study illustrates the escalation path. CellMaster first assigns broad labels, then identifies an ambiguous Neutrophil region for zoom-in and sub-clustering. Marker evidence for *Lcn2*, *Ltf*, *Camp*, and *Mmp9* supports immature, intermediate, and mature neutrophil states before the subset is merged back into the global annotation (Liang et al., 2022) (Figure 3). The figure records the user's acceptance of an agent recommendation and is not a fully automatic Core result. The same workflow resolves pro-B, large/small pre-B, and naive B states in the appendix. The case study is qualitative: it documents targeted refinement rather than independent biological validation. It does not imply that CellMaster replaces specialist references or expert review.

## 5 Discussion

**Evidence and scope.** The experiments support CellMaster primarily as an expression-grounded iterative annotation framework rather than a universal replacement for reference-based annotators. The nine-dataset breadth benchmark extends the automatic CellMaster+AF evaluation across tissues and species, while the PBMC3k/Liver/Retina depth benchmark isolates CellMaster-Core from AF. Fixed-AF ablations support the roles of marker memory, multi-agent decomposition, and confidence pinning in the automatic pipeline; dot-plot inspection is shared evidence rather than an independently ablated component.

**Selective escalation.** Specialist results reinforce complementarity: PopV and SCimilarity can be stronger when their references cover the target population, while Core is stronger in other contexts.

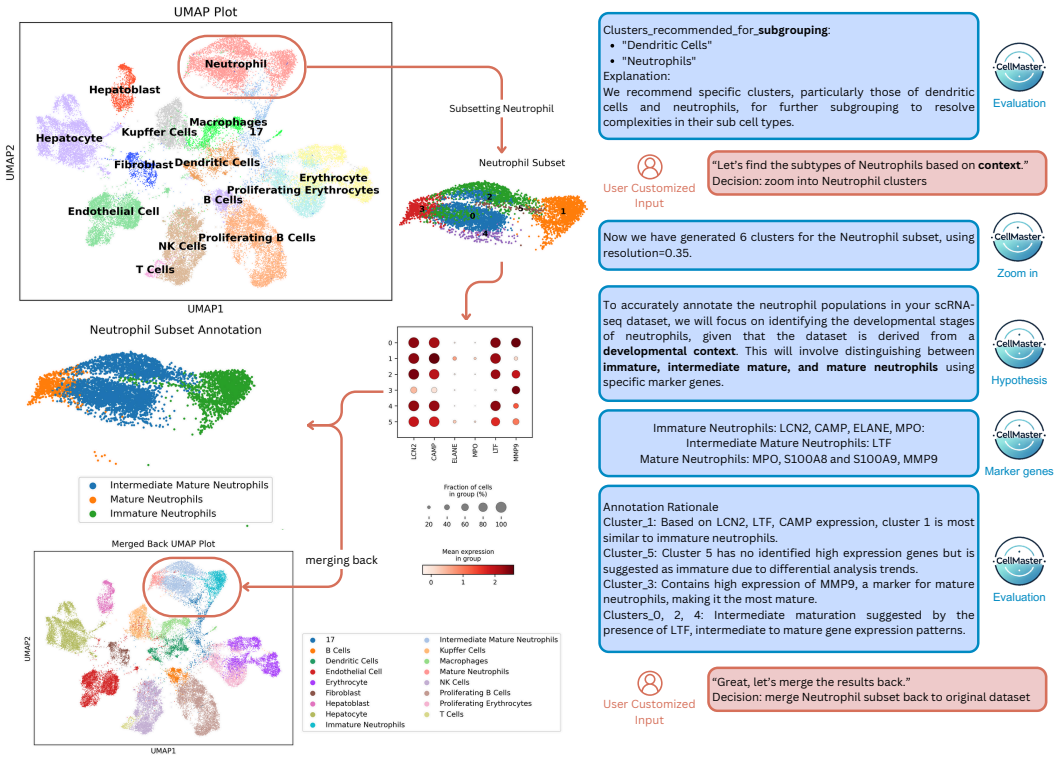



Figure 3: Case study on the developmental mouse Liver dataset. After two automatic iterations, CellMaster produces the major-cell-type UMAP (top left). Following an evaluation-stage recommendation, the user zooms into the Neutrophil cluster; CellMaster generates six sub-clusters and a dotplot over candidate developmental markers (centre), then assigns immature / intermediate / mature neutrophil labels (lower-left panel) supported by a developmental-context rationale. Merging back yields the final annotated UMAP. The right column shows the corresponding agent reasoning, marker selection, and rationale at each step.

CellMaster-Ecosystem is therefore a reconciliation layer for unresolved clusters, not a mechanism for turning correlated specialist outputs into votes. HITL extends the same principle to expert review: the two-expert study shows consistent steering through the shared evidence trail, but the lack of repeated annotations per expert prevents a population-level estimate of user variance.

**Operational limitations.** CellMaster remains sensitive to initial clustering, LLM stochasticity, marker and ontology coverage, and over-iteration after high-confidence clusters stabilize. External specialists inherit the coverage and biases of their references, and remote LLM backends introduce provider, cost, retention, and governance dependencies. The post-cutoff *Unseen Liver* evaluation reduces but does not eliminate memorisation concerns, while local execution avoids remote model calls only subject to local storage, logging, serving, and UI configuration. In practice, conservative iteration limits, explicit abstention, targeted escalation, and review of rare or contested populations are preferable to forcing complete fine-grained annotation.

## 6 Conclusion

We presented CellMaster, an expression-grounded iterative agent for cell-type annotation. Across nine datasets it provides cross-tissue evaluation; on PBMC3k, Liver, and Retina, CellMaster-Core obtains Cell Ontology scores of 0.708, 0.536, and 0.667 without external annotation references, increasing to 0.725, 0.550, and 0.705 with CellTypist AutoFill. Marker memory, multi-agent decomposition, and confidence pinning provide consistent gains under fixed AF. Unresolved clusters remain abstained or are selectively escalated to specialist tools or human experts through an auditable evidence trail. This design supports practical collaboration while keeping reference-based specialists and HITL intervention distinct from the automatic reference-free benchmark. Limitations and broader impact are discussed in Appendices M and N.

### AI USE STATEMENT

CellMaster uses generative AI models as core methodological components for hypothesis generation, marker proposal, expression-evidence interpretation, and cell-type annotation; the model identities, prompts, operating modes, and human-interaction protocol are described in Sections 3–4 and Appendix L. During manuscript preparation, generative AI tools were also used for language editing, organization, and source-level consistency checks. They were not used to fabricate datasets, ground-truth labels, numerical results, or citations; these were sourced or computed as described in the paper. The authors reviewed AI-assisted text and verified numerical results, citations, tables, and claims against the source files and reported experiments. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with AI assistance.

### ETHICS STATEMENT

This work uses publicly sourced single-cell datasets and does not recruit human participants, collect new identifiable data, or make clinical decisions; institutional review was therefore not applicable to the reported experiments. Automated annotations can propagate biological, reference, nomenclature, and model biases, and should not be interpreted as clinical diagnoses. We mitigate these risks by exposing expression evidence, preserving abstention, separating reference-based tool calls from the reference-free core, reporting limitations, and recommending expert review for rare or contested populations. The interface keeps the raw expression matrix local; remote backends receive only the text and evidence summaries required by the configured call, subject to the user's data-governance and provider-retention policies. Source datasets, models, and software remain subject to their original licenses and terms of use.

### REPRODUCIBILITY STATEMENT

The problem formulation and annotation pipeline are specified in Section 3; dataset composition, preprocessing, baselines, scoring, and repeated-run protocol are specified in Section 4. Appendix H provides the deterministic name-normalization and Cell Ontology scoring procedure, Appendix L provides the agent prompts and workflow details, and Appendix I records model identifiers, software versions, run dates, and local-backbone settings. The reported score is computed without an LLM judge. An anonymized code and processed-data package can accompany the submission; the public repository URL is omitted during double-blind review.

## REFERENCES


A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817):590–595, 2020.

10x Genomics. 10x genomics 3k pbmcs dataset. `https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k`, 2017. Accessed: 2025-05-15.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Jiawen Chen, Jianghao Zhang, Huaxiu Yao, and Yun Li. Celltypeagent: Trustworthy cell type annotation with large language models. *arXiv preprint arXiv:2505.08844*, 2025. doi: 10.48550/arXiv.2505.08844.

Hong Cui, Chao Wang, H. Maan, et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

C. Ergen, G. Xing, C. Xu, et al. Consensus prediction of cell type labels in single-cell data with popv. *Nature Genetics*, 56:2731–2738, 2024. doi: 10.1038/s41588-024-01993-3.

Fabian Fischer, David S. Fischer, Roman Mukhin, et al. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15:6611, 2024. doi: 10.1038/s41467-024-51059-5.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

Yiming Gao, Zhen Wang, Jefferson Chen, Mark Antkowiak, Mengzhou Hu, JungHo Kong, Dexter Pratt, Jieyuan Liu, Enze Ma, Zhiting Hu, and Eric P. Xing. scpilot: Large language model reasoning toward automated single-cell analysis and discovery. *arXiv preprint arXiv:2602.11609*, 2026.

Dezheng Han, Yibin Jia, Ruxiao Chen, Wenjie Han, Shuaishuai Guo, and Jianbo Wang. Recellty: Domain-specific knowledge graph retrieval-augmented llms workflow for single-cell annotation. *arXiv*, 2025. doi: 10.48550/arXiv.2505.00017. URL https://doi.org/10.48550/arXiv.2505.00017.

Gregory Heimberg, T. Kuo, D. J. DePianto, et al. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638:1085–1094, 2025. doi: 10.1038/s41586-024-08411-y.

Wenpin Hou and Zhicheng Ji. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, pp. 1–4, 2024.

Chao Hu, Tian Li, Yiming Xu, et al. Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. *Nucleic Acids Research*, 51(D1):D870–D876, 2023. doi: 10.1093/nar/gkac947.

K. Huang, S. Zhang, H. Wang, et al. Autonomous biomedical research with an artificial intelligence agent. *Science*, 393(6813):eadz4351, 2026. doi: 10.1126/science.adz4351.

Yan Liang, Kota Kaneko, Bing Xin, Jin Lee, Xin Sun, Kun Zhang, and Gen-Sheng Feng. Temporal analyses of postnatal liver development and maturation by single-cell transcriptomics. *Developmental Cell*, 57(3): 398–414.e5, 2022. doi: 10.1016/j.devcel.2022.01.004.

Nianping Liu, Chen Jiang, Xinfeng Yao, Minghao Fang, Xiaolong Qiao, Lin Zhu, Zongcheng Yang, Xuyuan Gao, Ying Ji, Chaoshi Niu, Chuandong Cheng, Kun Qu, and Jun Lin. Single-cell landscape of primary central nervous system diffuse large b-cell lymphoma. *Cell Discovery*, 9(1):55, 2023. doi: 10.1038/s41421-023-00583-1.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8. arXiv:2304.05376.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Terrence F. Meehan, Angela M. Masci, Abdullah Abdulla, et al. Logical development of the cell ontology. *BMC Bioinformatics*, 12:6, 2011. doi: 10.1186/1471-2105-12-6.

Madhvi Menon, Shahin Mohammadi, Jose Davila-Velderrain, Brittany A. Goods, Tanina D. Cadwell, Yu Xing, Anat Stemmer-Rachamimov, Alex K. Shalek, John Christopher Love, Manolis Kellis, and Brian P. Hafler. Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nature Communications*, 10:4902, 2019. doi: 10.1038/s41467-019-12780-8.

OpenAI. Gpt-4o system card, 2024.

OpenAI. gpt-oss-120b and gpt-oss-20b model card, 2025. arXiv:2508.10925.

J. D. Pearce, S. E. Simmonds, G. Mahmoudabadi, et al. A cross-species generative cell atlas across 1.5 billion years of evolution: The transcriptformer single-cell model. *bioRxiv*, 2025. doi: 10.1101/2025.04.25.650731.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

R. U. Rahman, E. T. Epstein, S. Murphy, et al. Single-cell transcriptomics reveals the impact of sex and age in the healthy human liver. *JHEP Reports*, 8(5):101773, 2026. doi: 10.1016/j.jhepr.2026.101773.

Tim Stuart, Andrew Butler, Paul Hoffman, et al. Comprehensive integration of single-cell data. *Cell*, 177(7): 1888–1902.e21, 2019. doi: 10.1016/j.cell.2019.05.031.

Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562: 367–372, 2018. doi: 10.1038/s41586-018-0590-4.

Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019. doi: 10.1038/s41598-019-41695-z.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 6000–6010, 2017.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: 10.1186/s13059-017-1382-0.

Shahar L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8(4):281–291.e9, 2019. doi: 10.1016/j.cels.2018.11.005.

Yuxuan Wu and Fuchou Tang. scextract: leveraging large language models for fully automated single-cell rna-seq data annotation and prior-informed multi-dataset integration. *Genome Biology*, 26:174, 2025. doi: 10.1186/s13059-025-03639-x. URL `https://doi.org/10.1186/s13059-025-03639-x`.

Chuan Xu, Martin Prete, Simone Webb, Laura Jardine, Benjamin J. Stewart, Regina Hoo, Peng He, Kerstin B. Meyer, and Sarah A. Teichmann. Automatic cell-type harmonization and integration across human cell atlas datasets. *Cell*, 186(26):5876–5891.e20, 2023. doi: 10.1016/j.cell.2023.11.012.

An Yang, An Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yehui Yang, Zelin Zang, Xienan Zheng, Yuzhe Jia, Changxi Chi, Jingbo Zhou, Chang Yu, et al. Mat-cell: A multi-agent tree-structured reasoning framework for batch-level single-cell annotation. *arXiv preprint arXiv:2604.06269*, 2026.

Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049.

# A Data and Code Availability

The complete source code for CellMaster itself, dataset preprocessing, evaluation metrics, and benchmark drivers is inside the CellMaster Anonymous zip file. All of the raw and processed single-cell RNA-seq datasets used in CellMaster benchmarking are publicly sourced collections. For each dataset, we release the processed objects (H5AD format for scRNA datasets, and csv format for DEG data), along with cell-type labels. The link to the dataset (anonymous google drive) is also provided in the CellMaster Anonymous folder.

The *Unseen Liver* evaluation is kept as a separate post-cutoff healthy-human-liver dataset with its preprocessing record, 11-cluster partition, labels, and method outputs. Cached LLM responses, intermediate rationales, final annotations, and deterministic Cell Ontology scoring files should be released with the benchmark drivers.

# B Additional benchmark and evaluation details

**Dataset coverage.** Table 6 records the nine breadth-benchmark datasets, spanning eight tissues, two species, and approximately 2.6K–66K cells. These metadata clarify the biological and computational range of the breadth evaluation; the processed objects, labels, and release contents are described in Appendix A.

Table 6: **Datasets used in benchmarking.** Nine published scRNA-seq datasets covering eight tissues across human and mouse, ranging from ∼2.6K to ∼66K cells. $N_{\text{cells}}$: cells used; $N_{\text{genes}}$: genes retained after preprocessing.

| Dataset | Tissue | Species | $N_{\text{cells}}$ | $N_{\text{genes}}$ |
|---|---|---|---|---|
| PBMC3k (10x Genomics, 2017) | PBMC | human | 2,638 | 13,714 |
| Liver (Liang et al., 2022) | liver | mouse | 41,000 | 2,000 |
| Retina (Menon et al., 2019) | retina | human | 20,091 | 19,719 |
| BCL (Liu et al., 2023) | lymphoma | human | 49,910 | 25,780 |
| Myeloid (Xu et al., 2023) | bone marrow | human | 51,552 | 36,601 |
| Large PBMC (Zheng et al., 2017) | PBMC | human | 65,943 | 17,732 |
| Brain (Tabula Muris Consortium, 2018) | brain | mouse | 7,856 | 23,341 |
| Large Intestine (Tabula Muris Consortium, 2018) | intestine | mouse | 3,938 | 23,341 |
| Limb Muscle (Tabula Muris Consortium, 2018) | muscle | mouse | 3,900 | 23,341 |

Table 7: Paired comparison of CellMaster+AF and Biomni on the three depth-benchmark datasets. The small number of datasets makes these tests underpowered; values are descriptive rather than evidence of broad statistical significance.

| Dataset | CellMaster+AF − Biomni | Wilcoxon $p$ | 95% CI |
|---|---|---|---|
| PBMC3k | +0.079 | 0.1250 | [0.005, 0.157] |
| Liver | +0.086 | 0.0625 | [0.023, 0.142] |
| Retina | +0.135 | 0.0625 | [0.052, 0.218] |

**Descriptive robustness check.** Table 7 reports per-dataset comparisons with Biomni. The five repeated CellMaster+AF runs yield positive confidence intervals on all three depth datasets, but the dataset count is too small for a broad significance claim; the tests are therefore descriptive checks rather than a pooled cross-dataset result.

**Selective specialist integration.** On the 28-cluster Liver dataset, either SCimilarity or PopV assistance resolves three of the four Core abstentions, while routing both specialists reaches full coverage and a CL score of $0.589$ (Table 8). These results concern confidence-gated escalation and do not redefine the reference-free Core setting.

**HITL protocol.** Table 9 documents the blinded two-expert protocol and the evidence visible during annotation. The Liver iteration trajectory is reported in main-text Table 5; it remains a collaboration study rather than an automatic-baseline comparison.

**Capability scope.** Table 10 summarises the intended system-level distinction: CellMaster combines coding-free interaction and user-controlled granularity with the ability to operate without an external marker reference in Core mode. The checkmarks describe supported workflow capabilities, not empirical accuracy rankings.

Table 8: Confidence-gated specialist-tool integration on the 28-cluster Liver dataset. SCimilarity and PopV use external references and are not part of CellMaster-Core.

| Variant | CL score | Coverage | Abstention |
|---|---|---|---|
| CellMaster-Core, no AF | 0.536±0.034 | 24/28 | 4/28 |
| CellMaster+AF | 0.550±0.040 | 28/28 | 0/28 |
| Core + SCimilarity assist | 0.554 | 27/28 | 1/28 |
| Core + PopV assist | 0.554 | 27/28 | 1/28 |
| Core + SCimilarity + PopV router | 0.589 | 28/28 | 0/28 |

Table 9: HITL evaluation protocol. HITL measures expert collaboration on top of the automatic system and is not an automatic-baseline comparison.

| Item | Protocol |
|---|---|
| Experts | Two independent biology experts |
| Datasets | All nine benchmark datasets |
| Ground truth | Hidden during annotation |
| Visible evidence | Outputs, dot plots, metadata, editable controls |
| Intervention | Short natural-language feedback per iteration |
| Repeated full passes | Not performed by the same expert |
| Agreement | At most one differing cluster per dataset |

Table 10: **Capability comparison of cell-type annotation approaches.** ✓ indicates the property is supported; ✗ indicates it is not. *Manual* = expert hand-annotation. *Reference-based* covers supervised classifiers and marker-table methods (e.g., CellTypist, scTab, CellMarker 2.0). *Naive LLM* refers to one-shot LLM annotation without iterative refinement (e.g., GPTCelltype).

| Property | Manual | Reference-based | Naive LLM | CellMaster |
|---|---|---|---|---|
| No specialized knowledge required | ✗ | ✓ | ✓ | ✓ |
| Coding-free operation | ✗ | ✗ | ✓ | ✓ |
| Low time and effort | ✗ | ✓ | ✓ | ✓ |
| No pre-training data restrictions | ✓ | ✗ | ✓ | ✓ |
| No external marker reference required | ✗ | ✗ | ✓ | ✓ |
| Adjustable annotation granularity | ✓ | ✗ | ✗ | ✓ |
| User-defined inputs supported | ✓ | ✗ | ✗ | ✓ |

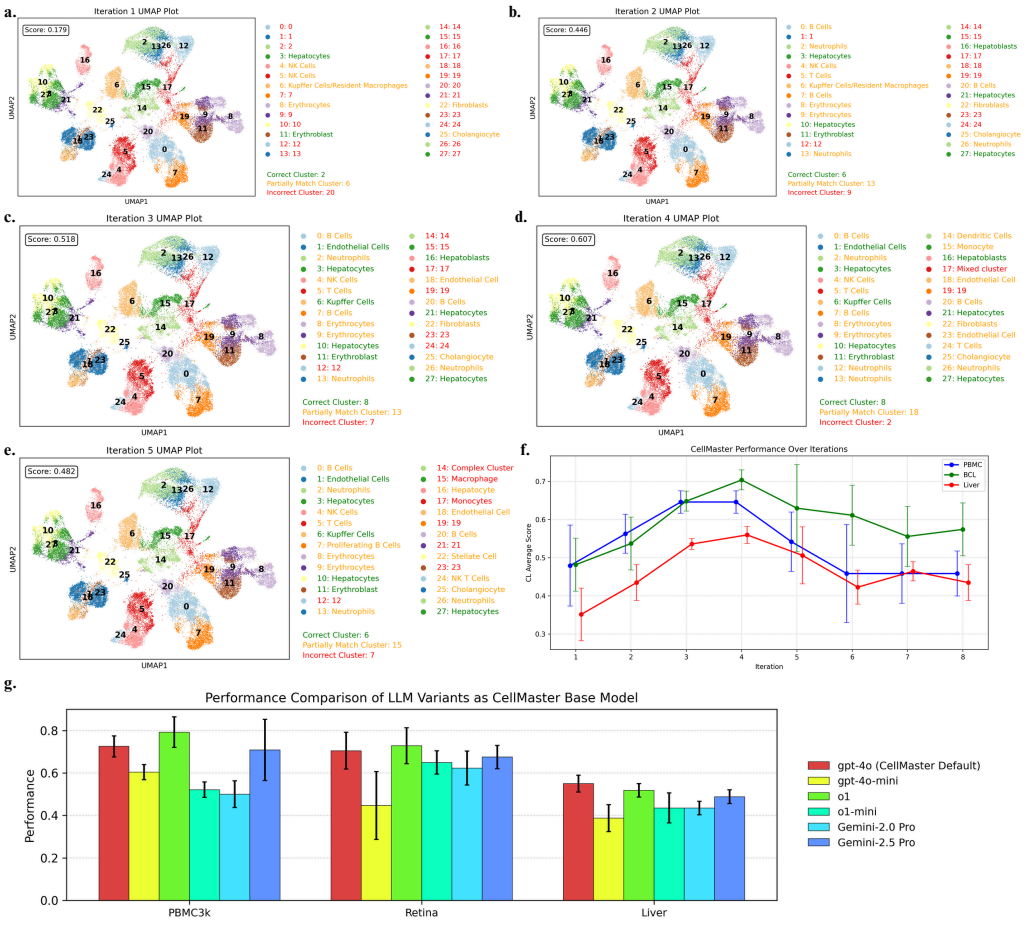



Figure 4: Ablations. *(a–e)* Per-iteration UMAP evolution on Liver (correct / partial / incorrect cluster counts annotated). *(f)* CL agreement per iteration on PBMC, BCL, Liver (3 runs each); accuracy peaks at iterations 3–4 and decays past iteration 4 under unattended cycling. *(g)* Backbone-LLM comparison on PBMC3k, Retina, Liver: GPT-4o is the strongest balanced default; o1 wins on PBMC3k / Retina but drops on Liver.

## C Iterative Annotation Performance Dynamics

**Subfigures a–e: single–dataset evolution.** The first five panels in Figure 4 tracked CellMaster's fully–automated performance over five successive annotation cycles on the Liver dataset. Accuracy rose monotonically from iteration 1 to iteration 4. This confirmed the agent's iterative feedback-driven refinements continually annotate blank clusters and correct early mis-labels, incorporating new marker evidence. A noticeable decline occurred at iteration 5, suggesting that, beyond a certain point, the marginal benefit of further automated updates is outweighed by error propagation (e.g. compounding low-confidence predictions) and by the diminishing pool of yet-unrefined clusters.

**Subfigure f: three–dataset ablation.** To verify that this pattern is not dataset–specific, we repeated the analysis on PBMC, BCL, and Liver data (8 iterations each run, 3 runs each dataset). All three curves displayed the same qualitative trend: peak accuracy was reached between the third and fifth iteration, with the exact optimum shifting slightly according to dataset complexity. Prolonged unattended cycling beyond this point produced a mild but consistent drop, mirroring the single–dataset findings.

**Interpretation and recommended practice.** The transient downturn is best understood as a form of *over-iteration*: once high-confidence clusters have stabilized, the agent is left to re-interrogate a shrinking subset of ambiguous cells. Iterative self-correction without fresh external cues can then amplify spurious signals, eroding global accuracy. In practice we therefore (i) limit unattended runs to 3 iterations by default in automatic version, (ii) present a "Stabilize Result" option in the UI version to lock converged labels, and (iii) encourage a brief manual review before authorizing additional cycles. These safeguards preserve the demonstrable gains of iterative refinement while preventing the late-cycle regressions observed in Figure 4.

## D Detailed case-study walkthrough

A key goal of advanced cell-type annotation tools is not only accurate classification but also an inspectable refinement workflow. To demonstrate this, we simulated a beginner user analysing the developmental mouse Liver dataset (41,000 cells; 2,000 highly variable genes) without prior tissue knowledge.

**Annotation process.** Dataset description and hypothesis were entered as: "41,000 cells across five developmental stages…likely cell types include T, B, hepatocytes, endothelial…". Two automatic iterations with one brief instruction ("Look for complicated clusters and identify their types") yielded the major cell types shown in the upper-left UMAP of Figure 3. In its evaluation output at iteration 2, CellMaster recommended subgrouping the Neutrophil cluster. The user adopted the suggestion and zoomed into Neutrophils; CellMaster generated six sub-clusters at resolution $0.35$.

Focusing on the developmental context of the dataset, CellMaster proposed developmental subtypes (immature, intermediate, mature neutrophils) with marker genes *Lcn2, Ltf, Camp, Mmp9*. Through holistic analysis of the marker genes and their developmental-stage associations, the agent annotated the six sub-clusters with the rationale: *Lcn2, Ltf, Camp* for immature; partial *Ltf* for intermediate; *Mmp9* for mature. The user accepted the recommendation, merged the subset back, and obtained the updated UMAP automatically. This qualitative workflow is consistent with the developmental context described in the original publication (Liang et al., 2022); it is not independent biological validation of the subtype labels.

**Comparison with baselines.** Unlike manual annotation methods or standard marker-based approaches, CellMaster's iterative feedback loops and context-aware suggestions deliver a more nuanced understanding of cell subtypes. GPTCelltype's outputs sometimes included subtypes inconsistent with the developmental context, such as disease-specific neutrophil types; in contrast, CellMaster's context-sensitive annotations remained aligned with project goals, substantially reducing off-target labels. By grounding decisions in markers + context-aware reasoning, CellMaster bridges computational annotation and biological interpretation.

**Cross-lineage comparison.** The parallel B-cell analysis in Figure 5 follows the same select–recluster–inspect–merge sequence, illustrating that targeted refinement is not restricted to the Neutrophil lineage.

## E BIOLOGICAL INSIGHT EXAMPLE OF B-CELL SUBGROUPING

Figure 5 applies the targeted-refinement workflow to the developmental B-cell compartment. After the broad B-cell population is selected, the user requests context-specific subtypes and CellMaster reclusters the subset at Leiden resolution 0.3, yielding six local clusters for marker-based inspection.

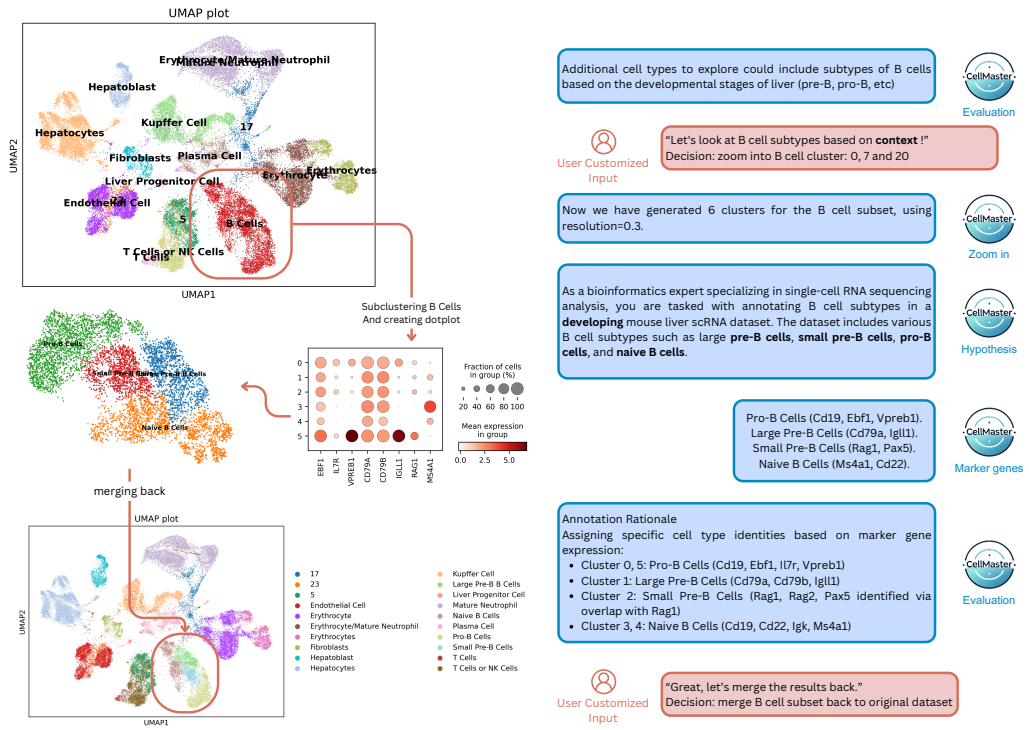



Figure 5: **Targeted B-cell subgrouping in developmental mouse Liver.** CellMaster identifies the broad B-cell compartment, generates six local clusters, proposes developmental markers, assigns pro-B, large/small pre-B, and naive B labels, and merges the refined labels back into the global UMAP.

The proposed panel includes early-development markers such as *Ebf1*, *Il7r*, and *Vpreb1*, recombination genes *Rag1/2*, and mature B-cell markers including *Cd19*, *Cd22*, *Igk*, and *Ms4a1*. Their dot-plot patterns support broader assignments to pro-B, large/small pre-B, and naive B states, after which the subset labels are merged into the full atlas. This example shows that selective zoom-in can be reused across lineages while retaining the user-visible evidence trail; it is a qualitative illustration rather than independent validation of each developmental subtype.

## F UI WALKTHROUGH

Figure 6 shows the four conversational steps experienced by an expert running CellMaster in HITL mode. The Hypothesis Agent presents the refined hypothesis and waits for any user override; the Proposal Agent presents the marker-gene list and accepts user-provided substitutions; the Annotation Agent extracts dot-plot evidence (gene-level, cluster-level, and population-level), produces the per-cluster labels and a UMAP visualisation (McInnes et al., 2018), and proposes the next-iteration focus.

**Interface components.** The React interface centres on an upload and hypothesis panel, linked UMAP and dot-plot panels, and an analysis panel containing the hypothesis, marker proposals, and iteration summary. Users can compare CellTypist or one-shot GPT suggestions, navigate iterations,

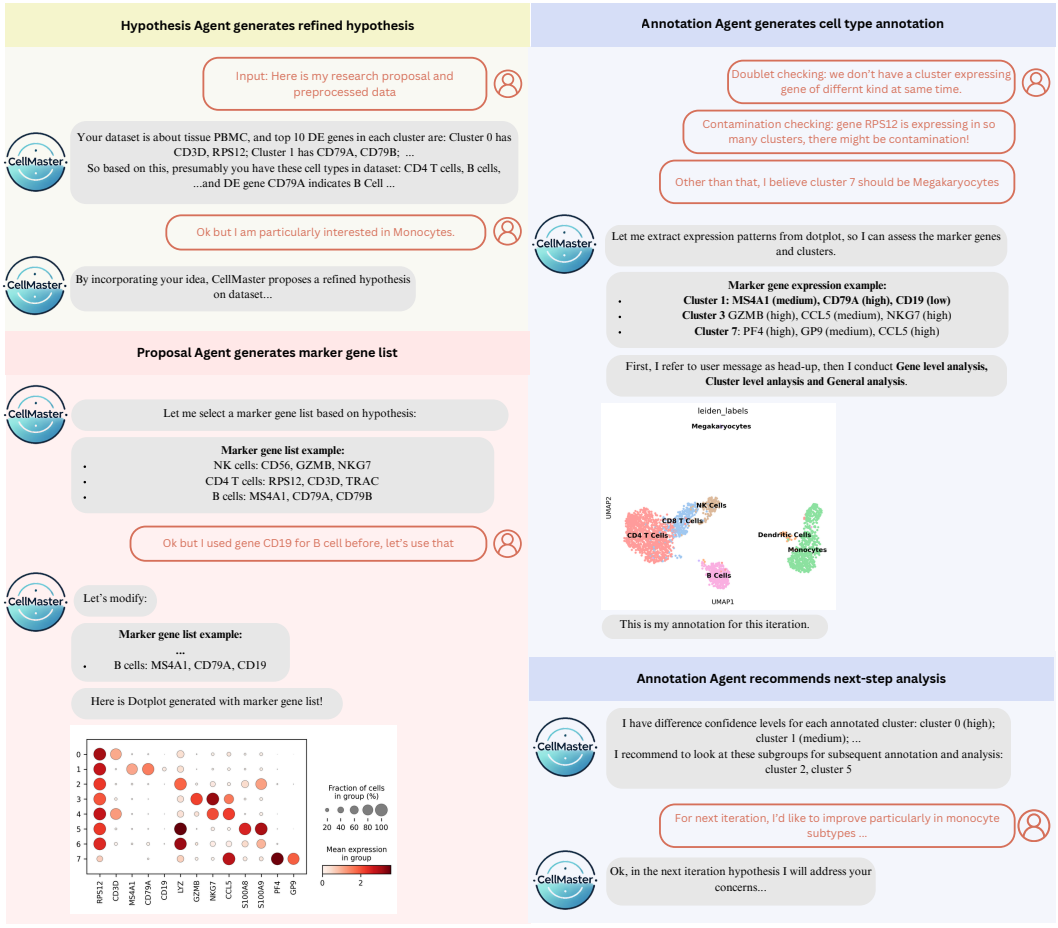



Figure 6: Detailed UI walkthrough across the four conversational steps. The Hypothesis Agent refines the hypothesis from user input; the Proposal Agent selects markers (here the user customises CD19 for B cells); the Annotation Agent extracts dot-plot evidence and produces a UMAP-based annotation; finally, the Annotation Agent recommends next-step subgrouping for low-confidence clusters.

stabilize converged labels, apply AutoFill to residual clusters, and zoom into selected populations with adjustable Leiden resolution before merging annotations back. Context-sensitive input accepts hypothesis edits, experimental notes, contamination flags, or evaluation feedback, while export records the annotated H5AD, analysis history, and visualisations. State is synchronised through React Context, and a **PipelineManager** coordinates the hypothesis, proposal, expression-analysis, and evaluation stages.

## G MANUAL-ANNOTATION WORKFLOW INTEGRATION

Manual annotation of single-cell data typically requires experts to assemble marker-gene lists, examine dot-plots, and assign cell-type identities based on domain knowledge. CellMaster's web interface implements this workflow directly, letting experts formalise their reasoning, provide iterative feedback, and refine computational predictions without departing from familiar practices. The interface runs locally and keeps the raw expression matrix on the user's machine. When a remote LLM backend is selected, only the text and evidence summaries required by the agent are transmitted through the configured API; fully local backbones avoid this external call.




Figure 7: Screenshot of the complete CellMaster web interface.

## CellMaster Annotation Workflow

### Example Input (Liver Development Dataset)

- **Dataset:** 41,000 mouse liver cells, 2,000 genes, spanning five developmental stages.

- **Objective:** Characterize hepatocytes, T cells, B cells, and endothelial cells to study liver development.

**Structured Workflow in CellMaster**

1. **Initial Hypothesis:** Provide project background and expected cell types.
2. **Refined Hypothesis:** Elaborate on annotation goals (e.g., highlight specific lineages).
3. **Marker Gene List:** Add or adjust marker genes to tailor reference panels.
4. **Dot Plot Review:** Inspect expression patterns across clusters and flag uncertainties.
5. **Iteration Feedback:** Evaluate and refine annotations until a satisfactory solution is reached.

## H Detailed methodology

**Adaptive heuristics.** Marker memory records whether a proposed gene produced sufficiently strong expression and prevalence evidence in any cluster. Successful markers can be reused in later proposals, while failed markers are suppressed to reduce repeated uninformative queries. The evidence thresholds are implementation parameters and are applied to the mean expression and detection fraction defined in Eq. 1.

The cluster-pair check counts genes with materially different expression or detection:

$$D_{ij} = |\{g \in G : \ |\log_2 E_{i,g} - \log_2 E_{j,g}| > \tau_E \text{ or } |F_{i,g} - F_{j,g}| > \tau_F\}| \, . \tag{3}$$

Pairs with $D_{ij} \leq K_{\min}$ are candidates for targeted sub-clustering. The default thresholds are $\tau_E = 1$, $\tau_F = 0.1$, and $K_{\min} = 5$.

**Detailed evaluation pipeline.** Raw cell-type names often exhibit inconsistencies (mixed casing, redundant suffixes, ambiguous abbreviations). An automatic cleaning step singularises plurals (e.g., *cells* $\rightarrow$ *cell*), normalises whitespace and punctuation, and retains biologically meaningful symbols (e.g., the slash in `CD4/CD8`). Cleaned names are matched against a versioned dictionary of standard nomenclature and Cell Ontology synonyms; unresolved names are retained for manual adjudication rather than passed to an LLM judge. Standardised names are then mapped to CLIDs and higher-level categories. The CL is parsed in OWL format into a directed acyclic graph whose nodes are CLIDs and whose edges denote parent–child relationships; this DAG enables deterministic lineage-aware scoring. Predictions are additionally checked against parent/child terms and higher-level categories to record biologically plausible disagreements.

## I Reproducibility

The reported cloud-backbone experiments use GPT-4o (`gpt-4o-2024-08-06`), GPT-4o-mini (`gpt-4o-mini-2024-07-18`), o1 (`o1-2024-12-17`), and o1-mini (`o1-mini-2024-09-12`), together with Gemini-2.0 Pro and Gemini-2.5 Pro accessed through OpenRouter. GPT-4o is documented by its system card (OpenAI, 2024). Experiments were conducted between June 2025 and July 27, 2026. Provider decoding defaults were used for the reported runs; nondeterministic methods were repeated three times unless a table specifies five runs. The complete release should include the exact provider/model identifier, decoding parameters, API date, prompt version, preprocessing version, clustering seed, raw responses, intermediate rationales, final annotations, and deterministic CL-scoring outputs.

Local/open experiments use Qwen3-4B (Yang et al., 2025) and gpt-oss-120b (OpenAI, 2025). Their release record must include hardware, quantisation, inference engine, context length, and runtime. The same DE lists, cluster partition, initial hypothesis, iteration count, and scoring script must be used across backbone comparisons.

## J BIOLOGICALLY PLAUSIBLE CL DISAGREEMENTS

The CL score can understate biological validity when the predicted label is a synonym, a parent/child term at a different granularity, or a developmental state absent from the reference annotation. Each such case is retained with the predicted label, reference label, CLIDs, marker evidence, and adjudication outcome. These cases are reported separately from the deterministic score and are not used to change benchmark labels after observing the prediction.

## K SPECIALIST-TOOL INTEGRATION PROTOCOL

CellMaster-Ecosystem invokes SCimilarity, PopV, or CellTypist only for uncertain or abstained clusters. The router records the specialist name, confidence, proposed label, marker evidence, dot-plot evidence, and final disposition. A proposal is accepted only when it passes the fixed confidence threshold and is not contradicted by expression evidence. Otherwise the router requests another iteration, emits a broader label, abstains, sub-clusters, or requests HITL review. Standalone specialist predictions and integrated predictions are stored separately.

## L DETAILED WORKFLOW OF CELLMASTER AGENTS

**CellMaster Prompt for Cell-type Annotation: Hypothesis**

**Hypothesis Generation stage:** CellMaster integrates the top marker genes per cluster, dataset context, and (optionally) information from previous iterations to create a hypothesis about the dataset.

content = f"Top {len(self.top_genes)} differentially expressed genes: {self.top_genes}"
if self.reference_dict:
content += f"You can refer to the possible cell types of these top genes in this dictionary-self.reference_dict"
content += f"Current Hypothesis:self.hypothesis"
if annotation_dict:
content += f"The cell type annotation from previous iterations annotation_dict"
if no_gene_cluster: content += f"Clusters without need to be focused on: {no_gene_cluster}"
if iteration_summary:
content += f"This is summary of previous iteration annotation, with information of next steps to take. iteration_summary"
system_role = "You are a research assistant specializing in cell biology. Based on top differentially expressed genes, previous cell type annotation (if provided), Clusters without need to be focused on (if provided), summary of previous iteration annotation (if provided), and failed genes (if provided), refine the given hypothesis to be more accurate and specific."

**CellMaster Prompt for Cell type annotation - marker gene proposal**

**Marker Selection stage:** CellMaster specifically proposes a marker gene list for the cell types of interest.

prompt = f"' You are a bioinformatics expert specializing in liver cell annotation. Your task is to propose an experiment for cell type annotation based on the following information:
Refined hypothesis: self.hypothesis
Instructions:
0. We have already labeled some of clusters, the information is in annotation_dict
1. a) The most important unsolved clusters are no_gene_cluster.
b) Other cell types in annotation_dict is cell types we labeled with high or low confidence. We first need to decide whether we should annotate them again.

2. For the cell types we have already successfully used, the gene marker list is successful_genes.
For large cell types like hepatocyte or B cell, the remaining clusters might still contain them, but for smaller cell types you don't need to think about them again. For the marker genes we already used but did not detect any expression, the list is failed_genes. So you don't need to try these gene and related cell type again.
3. Do not specify the cluster with cell type here. You can just output cell type and related marker genes.
4. Only provide proposal of unlabeled clusters, consider potential overlaps in marker gene expression between cell types:
a) Name of the cell type b) 3-5 marker genes
Output format: 1. List of cell types with their markers: [Cell Type 1]: Gene A; Gene B; Gene C [Cell Type 2] ...
2. Python list of all marker genes: MARKER_GENES = ['GeneA', 'GeneB', 'GeneC', ...]
Remember: - Be specific and concise in your descriptions. - Ensure all cell types have at least 3 marker genes. - Include the Python list of all marker genes at the end of your response, don't use any backtick. '''
system_role = "You are an AI trained to design scientific experiments based on hypotheses and background information."

**CellMaster Functions - Expression Analysis**

**Expression Analysis stage:** CellMaster quantifies and contrasts gene expression programs across identified clusters to derive both broad markers and nuanced distinctions. First, average expression and detection fraction matrices are computed for each gene–cluster pair. High-confidence markers are nominated by threshold filtering (`identify_marker_genes`), ensuring each candidate meets user-defined minima for mean expression and prevalence. To harmonize gene-wise dynamic ranges, CellMaster supports either Z-score normalization (`zscore_normalize_expression`) or min–max scaling (`min_max_scale_expression`). For cluster–cluster comparisons, log2fold changes are calculated (`compute_logfc`), and genes that satisfy combined criteria on expression, detection fraction, and absolute fold change (`identify_distinguishing_markers`) are flagged as distinguishing markers. Finally, by surveying all cluster pairs with a relaxable gene-difference tolerance (`find_similar_cluster_pairs`), the module reveals pairs of clusters that share overall transcriptional landscapes but diverge by only a small set of signature genes, facilitating downstream analyses such as targeted sub-clustering or lineage trajectory inference based on these key transcriptional shifts.

`identify_marker_genes`
Iterates over each cluster (rows of `dotplot_data`) and each gene (columns) to select those with mean expression $\geq$ `exp_thresh` and fraction expressing $\geq$ `frac_thresh`. Returns a `marker_genes` dictionary mapping clusters to lists of candidate markers.

`zscore_normalize_expression`
Applies a per-gene Z-score (zero mean, unit variance) transformation across clusters:

$$\text{normalized}_{g,c} = \frac{\text{expr}_{g,c} - \mu_g}{\sigma_g}, \quad \text{where } \mu_g, \sigma_g \text{ are the mean and standard deviation of gene } g.$$

`min_max_scale_expression`
Scales each gene's expression into [0,1]:

$$\frac{\text{expr}_{g,c} - \min_c(\text{expr}_{g,c})}{\max_c(\text{expr}_{g,c}) - \min_c(\text{expr}_{g,c})}.$$

`compute_logfc`
Computes $\log_2$ fold-change between two clusters with pseudocount 1:

$$\log_2\big(\text{expr}_{g,\,\text{cluster1}} + 1\big) - \log_2\big(\text{expr}_{g,\,\text{cluster2}} + 1\big).$$

`identify_distinguishing_markers`
First computes log-fold change via `compute_logfc`, then selects genes satisfying: at least one cluster's expression $\geq$ `exp_thresh`, fraction $\geq$ `frac_thresh`, and absolute log-fold change $\geq$ `logfc_thresh`. Returns a list of genes distinguishing the two clusters.

`find_similar_cluster_pairs`
Examines every unordered pair of clusters, applies `identify_distinguishing_markers` with relaxed thresholds (`exp_diff_thresh`, `frac_diff_thresh`, `logfc_thresh`), and retains pairs whose number of differing genes is $\leq$ `max_diff_genes`. Outputs a list of (cluster1, cluster2, diff_genes) tuples.

**CellMaster Prompt for Cell type annotation - evaluation**

**Result Evaluation stage:** CellMaster analyzes the dotplot expression data, and makes comprehensive evaluation about the expression and thus perform prediction.

content = f"' Context: You are working on a single-cell RNA sequencing cell type annotation task. The goal is to identify distinct cell types based on gene expression patterns. You have created a dotplot using a set of marker genes to visualize gene expression across different clusters.
MUST remember: you should list the cell types here. All cell types you refer to, should be in here. possible_cell_types
Data: For each cluster, the top genes are in this dictionary: marker_genes The clusters cannot find top genes are: empty_keys These are the genes that are successfully (highly) expressed in some clusters: success_list These are the genes that are failed to express in any clusters: fail_list There are some clusters that have similar expression, so we did differential expression analysis for these cluster pairs. The pairs and top differential genes are in: similar_clusters_dict
Add-on: Duplet and Contamination in dotplot: duplet_rule Any other important instructions: contamination_rule
Please give greatest attention to the Add-on part, if they are provided. Make sure to use them in your analysis.
Instructions: Please analyze the provided data and answer the following questions:
1. Gene-level analysis: a) Which genes are highly expressed in specific clusters? Provide a detailed description. b) Are there any genes that show differential expression across clusters (high in some, low in others)? c) Are there any genes that are not informative for cell type annotation (low expression across all clusters)? You should answer this question based on negation of 1b.
2. Cluster-level analysis: a) Are there any clusters that lack high expression of any marker gene? If so, list the cluster numbers. There are in total cluster_size clusters, index from 0.
3. Overall assessment: a) Based on the gene expression patterns, are there distinct clusters that potentially represent different cell types? b) Can you assign specific cell type identities to any of the clusters based on the marker gene expression? If so, provide your cell type annotations. You should only assign one cell type to each cluster. No doublet or contamination here. c) To refine the cell type annotation, recommend possible additional cell types. d) To refine the cell type annotation, recommend any particular cluster to perform subgrouping.
4. Confidence assessment: a) What are confidence levels of your annotation? Please also assign a confidence score to the process name you selected. This score should follow the name in parentheses and range from 0.00 to 1.00. A score of 0.00 indicates the lowest confidence, while 1.00 reflects the highest confidence. This score helps gauge how accurately

the annotation is. Your choices of confidence score should be a normal distribution (average = 0.5) You should consider if the annotation is widely seen, using deterministic wording and following background of dataset. For instance, if you label a doublet, or using word "probable", the score should be lower.
b) Based on confidence levels, what are the annotation results that you want to "stabilize", that is not change in next steps? if iteration is 1, you should choose top 1/3 confident clusters. if iteration is 2 and beyond, you should choose top 2/3 or more clusters. You can choose this threshold.
Please provide your answers in a structured JSON format, addressing each question separately using "1a" "2a" etc.
Remember, this is one iteration cell type annotation for a liver scRNA-seq dataset. Your insights will guide further refinement of the analysis. ”’ system_role = "You are an expert bioinformatician specializing in single-cell RNA sequencing data analysis and cell type annotation."

## M LIMITATIONS

**Reproducibility, resources, and scoring.** LLM-based methods have higher run-to-run variation than deterministic baselines, and provider defaults, model updates, unavailable random seeds, API cost, and retention policies can affect replication; repeated runs, model snapshots, cached responses, and intermediate outputs are therefore important. Local/on-premise backbones avoid external LLM transmission but require local storage, logging, serving, and UI configuration, and the reported Qwen3-4B and gpt-oss-120b results show a capacity–accuracy tradeoff. The Cell-Ontology score inherits the coverage and granularity of the CL graph, so valid synonyms, broader labels, finer subtypes, and cell states without CL identifiers may score lower than biologically warranted; coverage and abstention are reported, but no single metric captures all annotation validity. CellMaster-Ecosystem additionally depends on the species coverage, ontology, and quality of SCimilarity, PopV, CellTypist, or other registered specialists, which are external references rather than zero-shot components.

**Data, iteration, and human-review limits.** The method assumes that the initial clustering retains biological signal: similarity checks and targeted sub-clustering cannot recover information removed by severely poor partitions, and unattended cycles can replace strong markers with weaker alternatives after stabilization (for example, omitting *Afp* at Liver iteration 5 mis-labelled cluster 16 from Hepatoblast to Hepatocyte), motivating the conservative three-iteration default and confidence pinning. Public datasets and canonical marker signatures may also be recognizable to general-purpose LLMs despite re-clustering, so the 2026 *Unseen Liver* evaluation reduces but does not eliminate leakage concerns and stronger evidence requires private or newly generated datasets with independent labels. Finally, the two-expert HITL study did not repeat full annotation by the same expert: at-most-one-cluster disagreement supports consistency but does not estimate population-level user variance, so HITL should be interpreted as collaboration evidence rather than a definitive human-subject evaluation.

## N BROADER IMPACT

**Scientific utility.** Faster, evidence-linked annotation can shorten the iteration cycle for downstream single-cell analyses. **Access.** A coding-free workflow may broaden access to annotation, although reliable biological interpretation still requires domain knowledge. **Over-reliance.** Because LLM labels and rationales can both be plausibly wrong, rare or consequential populations should be reviewed by an expert. **Privacy.** Remote backends transmit text and evidence summaries under provider policies, whereas fully local backbones avoid that API transmission but introduce local security and performance tradeoffs. **Role of experts.** CellMaster is intended to support, not replace, annotators by making uncertainty, evidence, and intervention points explicit.