CASSIA allows for robust, automated cell annotation in single-cell RNA-sequencing data

Elliot Xie[1], Lingxin Cheng[1], Jack Shireman[2], Yujia Cai[1], Jihua Liu[1], Chitrasen Mohanty[1], Mahua Dey[2], Christina Kendziorski[1]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

[2]Department of Neurological Surgery, University of Wisconsin-Madison, Madison, WI, USA

Abstract

Numerous methods are available to assign cell type to subpopulations of cells profiled using single-cell RNA-sequencing. Most require a combination of computational and domain-specific expertise, and they frequently yield inconsistent results that can be challenging to interpret. Large language models have the potential to expand accessibility while reducing manual input and improving accuracy, but existing approaches suffer from hyperconfidence, hallucinations, and lack of interpretation. To address these limitations, we developed CASSIA for automated, accurate, and interpretable cell annotation of single-cell RNA-sequencing data.

Accurate cell type annotation is essential for understanding the cellular composition of tissues, characterizing the functional states of distinct cell types, and identifying dynamic changes in response to disease; and numerous annotation methods are available[1, 2]. Cell annotation typically involves identifying cell clusters and computationally annotating each one by comparing the cluster's average expression profile to lists of cell-type specific marker genes or reference expression profiles of known cell types. Manual annotation often follows to refine those clusters having no or low-confidence annotations, and/or those showing discrepant results across methods[3, 4]. While most investigators consider manual annotation by an expert to be the gold standard, it is time consuming, labor intensive, and requires a combination of computational and domain-specific expertise[3].

The popular large language model (LLM) based approach, GPTCelltype, demonstrates the potential of LLMs to reduce the effort and expertise required for cell type annotation[5]. However, the approach is limited by LLMs' inherent hyperconfidence and hallucinations, and by lack of quality assessments[6, 7] (Supplementary

Section S1). Without annotation-specific quality scores and/or measures of uncertainty, users cannot distinguish high-quality annotations from erroneous ones.

To improve the accuracy and fidelity of cell type annotation in practice, and to appropriately calibrate the confidence with which annotations are reported, we developed a collective agent system for single-cell interpretable annotation (CASSIA). In short, CASSIA is a multi-agent LLM consisting of an onboarding platform and five interconnected LLMs for annotation, validation, formatting, quality scoring, and reporting. Optional LLMs are also available for subclustering, uncertainty quantification, and annotation boosting. For applications requiring highly detailed annotations, CASSIA also provides a retrieval-augmented generation (RAG) agent. Figure 1 provides a detailed overview.

CASSIA was compared with state-of-the-art cell annotation methods on five large scRNA-seq datasets for which gold-standard annotations are available. Figure 2a shows that CASSIA increases fully correct annotations by 12-41% relative to existing approaches; performance averaged across all annotations is shown in Figure 2b. In addition to the five benchmark datasets, we also evaluated CASSIA on more complex cell populations. We first considered scRNA-seq data from two immune cell populations consisting of multiple cell subtypes (e.g., CD4+ T cells, CD8+ T cells, B cells) in varying functional states (e.g., naive, activated, exhausted). As shown in Figure 2c and Supplementary Figure 1, CASSIA accurately classifies immune cell subtypes and their functional states, surpassing the performance of existing methods for immune cell annotation by at least 25% (Figures 2a and 2b). We also evaluated CASSIA using five diverse datasets representing complex tumor microenvironments. As demonstrated in Figure 2d and Supplementary Figure 2, CASSIA outperformed other methods in terms of accurately distinguishing cancer from non-cancer cells within these highly heterogeneous environments; and it was uniquely capable of detecting metastatic signatures across all metastatic datasets. Similar advantages are shown in Supplementary Section S2 when considering results from simulated mixed cell populations, where the cell type being annotated is comprised of two distinct cell types.

Like immune and cancer cell populations, cell annotation in non-model organisms is also challenging, largely due to sparsely annotated single-cell references. We evaluated CASSIA's performance across vertebrate species including cartilaginous fish (shark) and mammals (domestic cat, pangolin, and tiger). As shown in Figure 2e, CASSIA correctly identified 20 of 22 cell subtypes in shark, 57 of 64 in domestic cat, 66 of 79 in

tiger, and 41 of 54 in pangolin, representing a greater than 14-77% improvement in annotation accuracy compared to existing methods (Figure 2e).

While increasing annotation accuracy in both common, complex, and relatively rare cell populations is worthwhile, equally important is the ability to quantify the quality of each annotation. CASSIA's scoring agent functions as an objective evaluator of the annotation process by generating an annotation-specific quality score ranging from 0-100% based on predetermined evaluation criteria applied to the full conversation history. Informative quality scores should be high for annotations that match the gold standard, lower for those that partially match, and lowest for mismatches. Figure 2f shows this to be the case. To evaluate the quality scores further, we considered annotations where CASSIA estimated high quality scores, but the annotations did not match the gold standard. Figure 2g and Supplementary Figure 4 show that for these cases, the gold-standard annotations appear to be incorrect. CASSIA also provides uncertainty quantification, if desired, by calculating a consensus similarity (CS) score to measure similarity across multiple CASSIA runs. Supplementary Section S3 provides examples and guidelines on using the quality and CS scores in practice.

When highly detailed annotations are required, the optional RAG agent can be useful as it leverages established cell marker databases and biological ontologies to achieve refined annotations. Neurons provide an example, as annotating the general class of neurons is relatively straightforward, but a detailed classification requires annotation of major type (excitatory, inhibitory), as well as numerous subtypes and layer location. Using the mouse motor cortex dataset from Azimuth, we find that CASSIA is the only method capable of precise annotations of neuronal subtypes (Figure 2h).

Large language models have the potential to democratize cell annotation, but robust approaches require safeguards to prevent hallucinations and appropriately attenuate hyperconfidence. CASSIA leverages the power of collective LLMs to provide increasingly accurate and interpretable cell type annotations coupled with appropriate measures of quality and uncertainty, thereby improving cell type annotation in scRNA-sequencing experiments.

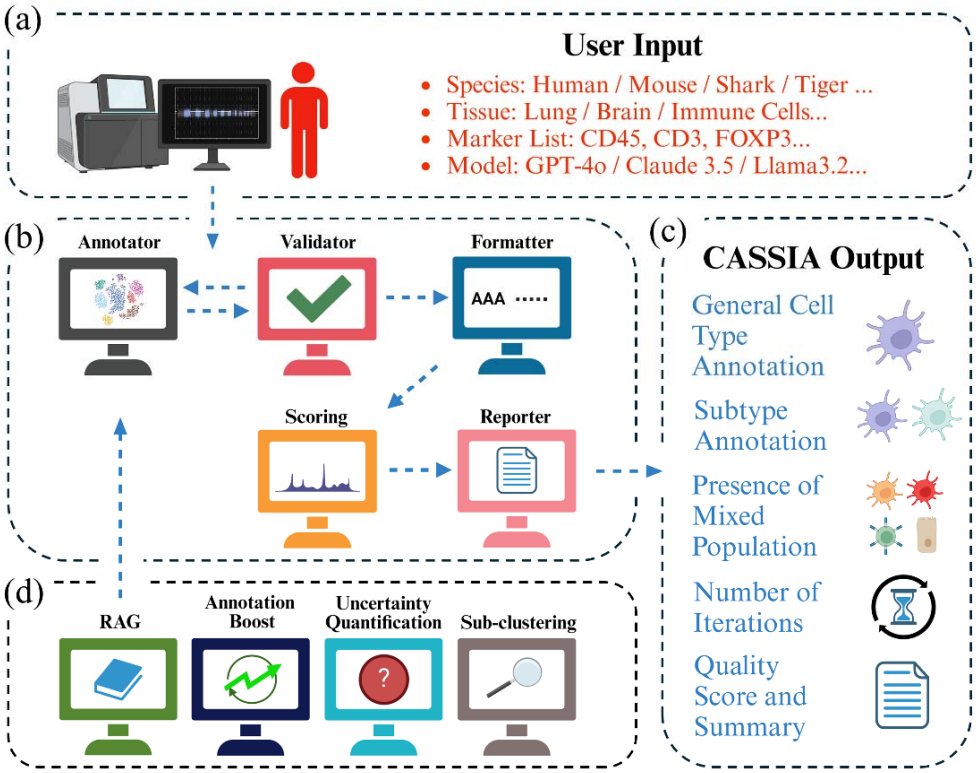

Figure 1: The multi-agent LLM system underlying CASSIA. A user interacts directly with the Onboarding platform by specifying species, tissue type, and a collection of markers associated with cell subtypes within that tissue, if known (a). Any information associated with experimental conditions, interventions, or other sample specific information may be provided. (b) Together, this input is used to create the user prompt given to the Annotator agent. The Annotator agent performs a comprehensive annotation of the single-cell data using a zero-shot chain-of-thought approach that mimics the standard workflow that a computational biologist would typically follow for cell annotation. Results are then passed to the Validator agent to check marker and cell type consistency; results failing validation are passed back to the Annotator and this iterative process continues until results pass validation (or the maximum number of iterations is reached). Results are then moved to the Formatting agent which summarizes each cell annotation; this summary along with the full conversation history is provided to the Scoring agent for quality scoring. The Reporter agent then generates a comprehensive report documenting the complete annotation process, including agent conversations, quality evaluation reasoning, and validation decisions with supporting evidence to facilitate transparent interpretation of results. The output from default CASSIA is shown in (c). Optional agents include those shown in (d).

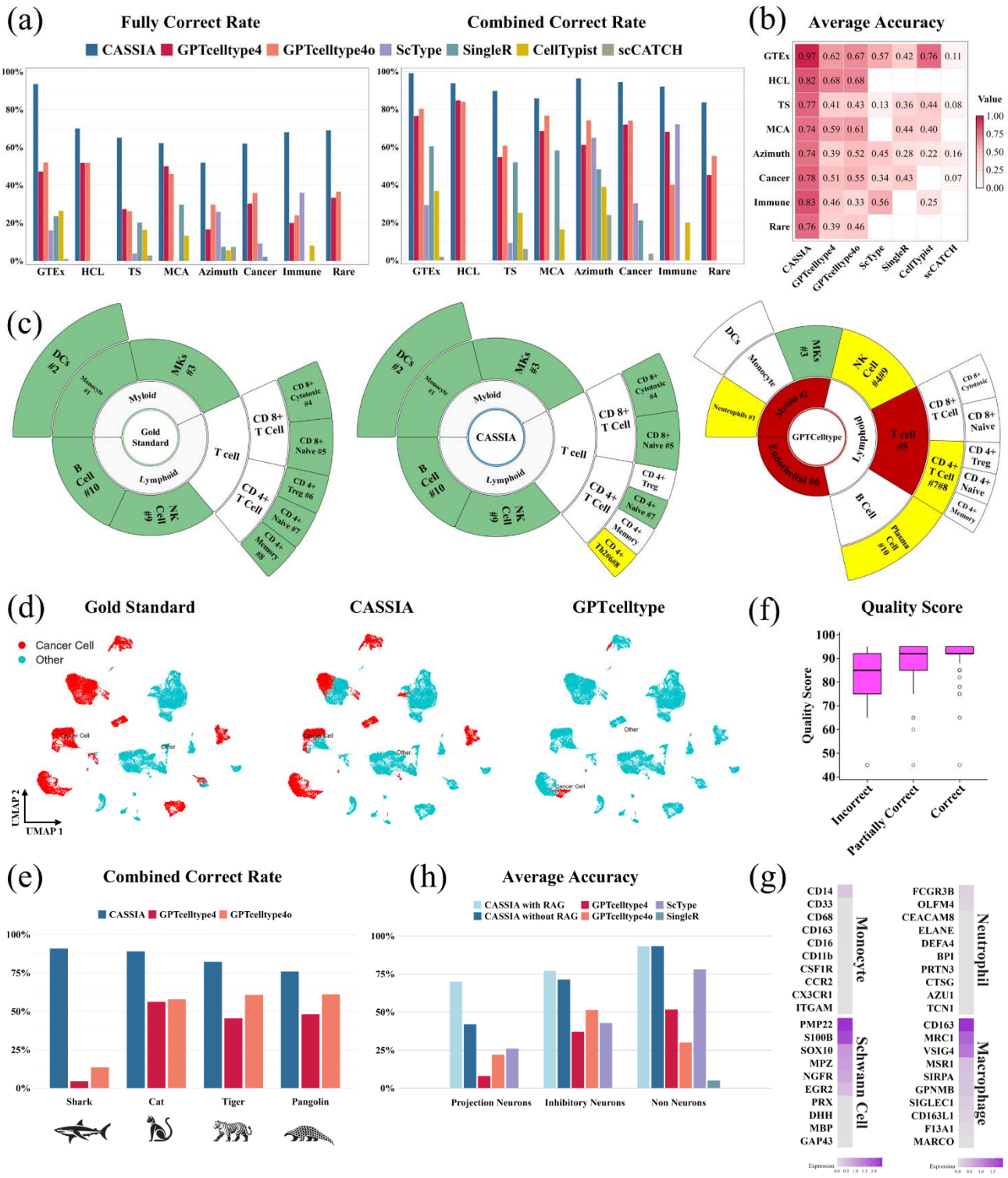

Figure 2: CASSIA outperforms existing approaches on five benchmark datasets, on complex populations of cells from immune and cancer cell populations, and on rare cell types. Shown in panel (a) are the average

percentages of fully correct (left) and combined correct (fully or partially correct, right) annotations for CASSIA and 6 widely used methods, where the average is taken over cell types within each of the 8 datasets shown. Figure 2a shows that CASSIA increases fully correct annotations by 12-41% and combined correct annotations by 9-29% relative to the next-best approach, which in most cases is GPTcelltype4 or 4o. Accuracy averaged across all cell types within each dataset is shown in panel (b). Panel (c) shows that CASSIA accurately classifies immune cell subtypes and their functional states in the PBMC immune cells; the gold-standard annotations are shown left compared with results from CASSIA (middle) and GPTcelltype4 (right). Fully correct annotations are shown in green, partially correct in yellow, and incorrect in red (Supplementary Figure 1 shows similar results in the ProjectTILs immune dataset). Panel (d) shows a UMAP with metastatic brain cancer cells colored according to the gold-standard (left), CASSIA annotation (middle), and GPTCelltype4 annotation (right) as cancer/likely cancer (red) or non-cancer (blue). As demonstrated, CASSIA is able to distinguish cancer from non-cancer cells. Three additional cancer datasets show similar results (Supplementary Figure 2). Panel (e) summarizes performance in relatively rare species including shark, domestic cat, tiger, and pangolin. Panel (f) shows that CASSIA's quality scores are informative as they are high for annotations that match the gold standard, lower for those that partially match, and lowest for mismatches. Panel (g) displays results from the intestine TS dataset, where CASSIA annotated cells from a given cluster as Schwann cells with high quality and CS scores, but the gold-standard annotation is monocytes. Expression averaged across cells in the cluster is shown for monocyte marker genes and Schwann cell marker genes; the low expression of monocyte markers and high expression of Schwann cell markers suggests that the gold standard annotation is incorrect (Supplementary Figure 4 provides additional examples). CASSIA's default model was used for the results presented in panels (a)-(g); optional agents were not used. Panel (h) shows results from CASSIA with the RAG agent implemented to obtain highly detailed annotations (see also Supplementary Section S4).

Data availability: Raw RNA sequencing data generated in this study have been deposited in the Gene Expression Omnibus (GEO) under accession number (pending). All other publicly available datasets analyzed in this study are listed in Supplementary Table 1.

Code availability: The CASSIA package developed for this study is available at https://github.com/ElliotXie/CASSIA. All custom scripts used for data processing, marker analysis, and method benchmarking are available at the same repository.

Author contributions: EX conceived the study and led the method development, implementation, validation, and case study analyses. CK co-led the project. LC contributed to the method development and developed the RAG agent. LC, YC, and JL contributed to case study analyses. MD collected the brain tumor samples and assisted with interpretation of results; JS processed the brain tumor samples for scRNA-seq profiling, contributed to method development, and assisted with interpretation of results from all case study analyses. CK, EX, LC, and JS wrote the manuscript.

Competing interests: The authors declare no competing interests.

Acknowledgments: This work was supported by NIH GM102756 (CK), NIH K08NS092895 (MD), and the IU Value Research grant (MD).

Online Methods

CASSIA is a multi-agent LLM consisting of an onboarding platform and five interconnected LLMs: an annotation agent, a validation agent, a formatting agent, a quality scoring agent, and a reporter agent. Optional LLMs are also available for subclustering, uncertainty quantification, and annotation boosting. For applications requiring highly detailed annotations, CASSIA includes a RAG agent. Each agent is described in detail below.

CASSIA implementation: We provide CASSIA as an R package, a Python package, and a web-based user interface (https://www.cassiacell.com/). Both programming packages support all CASSIA functionalities including optional agents, with the default pipeline executable via a single command. The web interface implements the default pipeline only.

Onboarding: A user provides information to an Onboarding platform by specifying species, tissue type, and a collection of markers associated with cell subtypes within that tissue, if known. We suggest 50 markers be specified (see the Identification of Marker Genes section below). Most analyses will use markers from *FindAllMarkers* in Seurat or *tl.rank_genes_groups* in Scanpy, although markers can be manually added if

desired. Any information associated with experimental conditions, interventions, or other sample specific information may be provided, and will be used to create the user prompt given to the Annotator agent. The specific user prompt reads: "Your task is to annotate a single-cell [species] dataset from [tissue type]. Please identify the cell type based on this ranked marker list: [marker list]. Below is some additional information about the dataset: [additional information]." In some cases (e.g. multiple mixed cell types from an atlas), tissue type is not known. A tissue blind version can also be used (see Supplementary Section S5).

Annotation: The Annotator agent performs a comprehensive annotation of the single-cell data. To minimize hallucinations and prevent oversimplification of complex biological concepts, we implemented multiple prompt engineering techniques: a zero-shot chain-of-thought approach that structures the model's reasoning to follow the same systematic analytical steps a professional computational biologist would take[8, 9], role specification where the Annotator agent acts as a 'professional computational biologist with expertise in single-cell RNA sequencing,' and audience framing as 'an expert in the field.' These role-playing strategies have been shown to improve the likelihood of receiving detailed and technically accurate responses[10]. Finally, we included a hypothetical reward statement: 'you will be rewarded $10,000 if you do a good job' as such reward modeling in LLMs has been shown to prime the model for maximum effort and attention to detail[11].

Taken together, the system prompt to the Annotator Agent reads: "You are a professional computational biologist with expertise in single-cell RNA sequencing (scRNA-seq). A list of highly expressed markers ranked by expression intensity from high to low from a cluster of cells will be provided, and your task is to identify the cell type. You must think step-by-step, providing a comprehensive and specific analysis. The audience is an expert in the field, and you will be rewarded $10000 if you do a good job. Steps to follow:

1. List the Key Functional Markers: Extract and group the key marker genes associated with function or pathway, explaining their roles.

2. List the Key Cell Type Markers: Extract and group the key marker genes associated with various cell types, explaining their roles.

3. Cross-reference Known Databases: Use available scRNA-seq databases and relevant literature to cross-reference these markers.

4. Determine the Most Probable General Cell Type: Based on the expression of these markers, infer the most likely general cell type of the cluster.

5. Identify the Top 3 Most Probable Sub Cell Types: Based on the expression of these markers, infer the top three most probable sub cell types. Rank them from most likely to least likely. Finally, specify the most likely subtype based on the markers.

6. Provide a Concise Summary of Your Analysis.

Always include your step-by-step detailed reasoning. You can say "FINAL ANNOTATION COMPLETED" when you have completed your analysis. If you receive feedback from the validation process, incorporate it into your analysis and provide an updated annotation."

Once FINAL ANNOTATION COMPLETED is reached, annotations are provided to the Validator agent to ensure accuracy and consistency.

Validation: The Validator agent is designed to iteratively cross-check the results provided by the Annotator agent, forming a feedback loop that significantly enhances the robustness of the final annotations. This iterative design is inspired by self-verification techniques, which allow for critical evaluation and continual refinement of annotations[12]. The Validator agent primarily ensures marker and cell type consistency, verifying that the key markers identified by CASSIA are present in the provided marker list and accurately represent the identified cell type. This step is crucial to confirm that the markers used for annotation are both appropriate and reliable. If the validation fails, the Validator agent provides detailed feedback and sends the results back to the Annotator agent for revision. This process iterates up to three times to ensure the highest quality annotation. Regardless of whether validation finally passes or fails, the results are then forwarded to the Formatting agent for further processing. The full system prompt for the Validator agent is provided in Supplementary Section S5.

Formatting: A formatting agent was implemented to process annotation outputs from the previously mentioned two agents. The agent operates in two distinct states: a standard mode that extracts and structures validated cell type annotations including mixed population information, and a diagnostic mode that activates upon validation failure to analyze error sources and generate recommendations for resolution. The full system prompt is provided in Supplementary Section S5.

Quality Scoring: The Scoring agent functions as an objective evaluator that analyzes the complete annotation conversation history to generate a quantitative assessment score (0-100%). It focuses on scientific accuracy of the annotations along with balanced use of multiple markers where high-ranked markers contribute most to the annotation and no single marker drives the annotation. The full system prompt for the Quality Scoring agent is provided in Supplementary Section S5.

To establish thresholds for the quality scores, we implemented a weighted cost optimization approach defined as $\text{Total cost} = (\text{w} \times \text{FP}) + \text{FN}$ where FP represents false positives (correct annotations incorrectly flagged as low quality), FN represents false negatives (incorrect annotations misclassified as high quality), and w is the incorrect weight parameter (default = 2). We assigned greater weight to false positives to minimize the risk of unnecessarily flagging correct annotations for manual review. Supplementary Figure 6 shows that the lower threshold minimizing total cost is ~75% when evaluated on more than 500 cell type annotations from the five benchmark datasets. Given this, annotations with quality scores below 75% are flagged as low-quality. In these cases, the (optional) Annotation Boost agent could be used in an effort to improve the annotation.

Reporter agent: The Reporter agent generates a comprehensive HTML report documenting the complete annotation process, including agent conversations, quality evaluation reasoning, and validation decisions with supporting evidence. The full system prompt is provided in Supplementary Section S5.

Subclustering Agent: CASSIA provides an optional subclustering agent designed to optimize the analysis of specialized cell populations (e.g., CD4+ T cells cluster). Rather than processing individual subclusters separately, this agent analyzes all subclusters within a specific cell population simultaneously. This concurrent analysis strategy improves computational efficiency, enabling better detection of subtle phenotypic differences between closely related cell states. The full system prompt is provided in Supplementary Section S5.

Uncertainty quantification: CASSIA provides an optional feature to assess annotation reliability by quantifying consensus across multiple (defaults to 5) runs. The Consensus Similarity (CS) score described below builds upon recent advances in LLM ensemble methods, which suggest that multiple iterations of a model or cooperation between several different models can achieve superior performance through consensus mechanisms[13, 14]. To ensure consistent comparisons across CASSIA runs, cell type nomenclature is standardized using two complementary approaches. The first approach applies basic text normalization

(converting to lowercase, replacing hyphens with spaces, removing trailing 's'), then maps each normalized label to standardized Cell Ontology terms via the CL API[12]. This mapping matches cell types to their formal ontological definitions, ensuring consistent terminology. After unifying cell type names, we employ a consensus similarity (CS) score to measure the similarity across the *A* CASSIA runs. Recall that the Annotator agent identifies the most probable general cell type and the three most probable sub-cell types, ranked from most likely to least likely. Letting *R* denote the set of all annotation results, each result $r \epsilon R$ is a pair (g, s) where *g* represents the general cell type and *s* represents the most likely sub-cell type. Then

$$\mathrm{CS}_{\mathrm{CL}} = \frac{1}{|R|(w_g + w_s)} \sum_{r \epsilon R} w_g I(g_r = g_c) + w_s I(s_r = s_c) + w_g w_s [I(g_r = s_c) + I(s_r = g_c)]$$

where g_c is the most frequent general cell type in *R*; s_c is the most frequent cell sub-type in *R*; w_g and w_s specify the weights given to matches of the general cell type and cell sub-type, respectively. $I(x)$ denotes the indicator function where $I(x) = 1$ if *x* is true and 0 otherwise; $|R|$ denotes the number of annotations in *R*. We use $\mathrm{CS}_{\mathrm{CL}}$ to emphasize the fact that this CS score was calculated using terms unified by the Cell Ontology CL API. The $\mathrm{CS}_{\mathrm{CL}}$ score takes into account both the general and sub-cell types, giving more weight to matches in the general type while still considering sub-type matches and potential type inversions.

To address potential API mapping inconsistencies, a second approach employs an LLM-based unification agent that harmonizes synonyms and slight variations in cell type nomenclature. A $\mathrm{CS}_{\mathrm{LLM}}$ based score is calculated based on this unification. Finally, to enhance robustness and mitigate potential errors from unification steps, a consensus agent independently estimates the consensus general cell type and subtype. Taken together, three measures of consensus are obtained ($\mathrm{CS}_{\mathrm{CL}}$, $\mathrm{CS}_{\mathrm{LLM}}$, and the consensus agent's estimated consensus score). The maximum score is taken as the final CS score. Annotations with a CS score below 75% are flagged as uncertain and recommended for further validation using the Annotation Boost agent.

Annotation Boost agent: CASSIA provides an optional annotation agent that can be used to improve low-scoring annotations (below 75%). Unlike the default workflow which uses only ranked marker gene lists, the Annotation Boost agent leverages comprehensive statistical metrics from the *FindAllMarkers* file uploaded to the Onboarding platform (e.g. p-values, percentage expression in groups, and log2 fold changes) and examines

expression patterns across all genes. Based on results from the Annotation agent, it generates testable hypotheses and systematically queries gene expression data to assess each hypothesis. This process continues iteratively until either a high-quality annotation is achieved, or the maximum iteration limit (defaults to 5) is reached. Further details and the full system prompt are provided in Supplementary Section S5.

Retrieval-Augmented Generation (RAG) pipeline: When highly detailed annotations are required, an optional multi-agent RAG pipeline can be useful. The RAG pipeline consists of multiple sequential agents that leverage established cell marker databases and biological ontologies to generate comprehensive marker profiles for tissue-specific cell type identification. Specifically, the Marker Database agent queries the CellMarker database[15] to identify validated, canonical cell type markers. The agent will generate a conservative list of markers specific to the tissue of interest. Given the markers, which in many cases contain markers with relatively low sensitivity and specificity, the agent further analyzes the marker list to identify markers of broad cell types as well as those representing refined tissue-specific subtypes. The second agent, the Ontology Database agent, leverages the Owlready ontological database[16] to identify those ontology trees associated with the tissue of interest. Trees are then filtered by the agent to retain only those containing relevant tissue-relevant cell types, eliminating extraneous branches that are not typically present in the tissue of interest. The Hierarchical Feature agent then performs a PCA-like analysis on the ontology trees selected from the second agent to identify key discriminative features. This process identifies hierarchical patterns that effectively categorize cell types into functionally related groups that ultimately inform the cell type annotation. For each identified feature axis, we compile specific markers that can effectively discriminate between cell populations along that axis. For studies involving non-model organisms not directly represented in reference databases, we implemented an optional cross-species comparison agent. This agent analyzes evolutionary conservation between the species of interest and reference species to evaluate marker gene conservation, identify potential inter-species variations, and provide guidance on marker interpretation accounting for species-specific differences. The final component consolidates all generated outputs in the previous four agents into a text format, which is then integrated as additional information into the default CASSIA pipeline. Further details are provided in Supplementary Table 2; a specific example is provided in Supplementary Section S4.

Fully correct vs. partially correct annotations: We established a hierarchical evaluation framework based on the Cell Ontology tree structure to assess annotation accuracy, classifying predictions as fully correct, partially correct, or incorrect based on their taxonomic distance from reference annotations. A prediction one step away from the reference on the ontology tree is considered partially correct (e.g., predicting "T-cell" for a

CD8+ T-cell), while predictions multiple steps removed are classified as incorrect (e.g., predicting "immune cell" for a CD8+ T-cell). For rare or highly specific cell types, we adapted this framework to consider predictions as partially correct when they accurately identify the general cell type category, such as classifying "epithelial cell" as partially correct for Bestrophin 4 (BEST4) + cells. See Supplementary Table 2 for more evaluation examples. A similar, but more detailed, scoring system was used for the RAG agent applied to the mouse motor cortex data (see Supplementary Section S4).

Average accuracy: Cell type annotations were evaluated on a three-tier scale: fully correct (score = 1), partially correct (score = 0.5), and incorrect (score = 0). Dataset accuracy was calculated as the mean score across all annotated cell types.

Identification of marker genes: We evaluated CASSIA's performance across varying marker set sizes (*n* = 10, 30, 50, 70, 100) using the GTEx data, large and small intestine tissues from the Tabula Sapiens (TS), and non-model mammalian samples including domestic cat, tiger, and pangolin. For most datasets, average annotation accuracy plateaued at 50 marker genes (Supplementary Table 3), with minimal improvements observed with larger sets. Based on this analysis, we recommend using 50 marker genes as the optimal balance between computational efficiency and classification accuracy.

Parallel computing and computational time: CASSIA utilizes Python's *concurrent.futures* module with *ThreadPoolExecutor* for efficient parallel processing. Cell type analyses are distributed across a user-defined thread pool (default: 10 workers), enabling concurrent annotation of multiple cell types. This parallel architecture scales efficiently, processing each cell type in about 30 seconds per CPU core with pre-prepared marker files. On a standard 8-core computer, CASSIA completes the annotation workflow for typical single-cell datasets with 20 clusters in under 2 minutes.

Memory management: CASSIA implements selective memory access across its agent network to optimize annotation processing. The annotation and validation agents maintain access to their shared conversation history for iterative refinement, while the scoring agent evaluates the complete dialogue record. For successful annotations, the formatting agent processes only the final validated exchange, whereas for failed cases, it analyzes the full conversation history to generate comprehensive error reports. This targeted memory

management strategy ensures efficient information flow while maintaining contextual isolation between distinct cell type analyses.

Model selection: To evaluate CASSIA's multi-agent framework across different large language models, we conducted comprehensive benchmarking using three distinct models: Claude 3.5 Sonnet (2024-10-22), GPT-4o (2024-08-06), and Llama3.2 90B. We created a benchmark dataset consisting of 100 cell types drawn from five randomly selected tissues within our benchmark datasets to ensure diverse representation. Performance was evaluated against GPTCelltype implementations using both GPT-4 and GPT-4o. Our analysis revealed that CASSIA's architecture enhances annotation accuracy regardless of the underlying language model. Notably, even when using Llama3.2 90B, an open-source model that typically demonstrates lower performance metrics compared to proprietary models, CASSIA achieved superior results compared to GPTCelltype implementations. While Claude 3.5 Sonnet achieved the highest accuracy, we selected GPT-4o as the default model due to its optimal balance between performance and computational efficiency. See Supplementary Figure 5 for detailed benchmarking results.

Model implementation and costs: CASSIA supports flexible API integration through three options: OpenRouter API, OpenAI API, or Anthropic API. While direct API connections to OpenAI and Anthropic are supported, the OpenRouter API option enables users to access most of the commercially available large language models without code modifications, including but not limited to GPT-4, Claude, and Llama variants. Based on our benchmarking results and cost analysis as of October 2024, we established a tiered recommendation system. GPT-4o serves as the default model, costing approximately $0.02 per annotation ($2.50/M input tokens, $10.00/M output tokens). For applications requiring maximum accuracy, Claude 3.5 Sonnet provides enhanced performance at $0.03 per annotation. For large-scale analyses (>10000 clusters) or cost-sensitive applications, Llama3.2 90B offers a cost-effective alternative at $0.003 per annotation through OpenRouter.

Data preprocessing: Dataset-specific markers were used in most cases, if provided. Exceptions included the shark dataset, where we excluded unmapped genes. Details are provided in Supplementary Table 1. For datasets lacking predefined marker genes, we implemented a standardized analysis pipeline. Raw count matrices were normalized using the *NormalizedData* function in Seurat v5, and differentially expressed genes were identified using Seurat's *FindAllMarkers* function with parameters set to a minimum cell expression threshold (min.pct) of 0.1, adjusted p-value cutoff of 0.05, and log2 fold change threshold of 0.25.

Comparative evaluation of existing cell type annotation methods: For comprehensive benchmarking, we compared CASSIA with established annotation methods including GPTCelltype (version 1.0, using GPT4 and GPT4o), ScType[17] (version 1.0), singleR[18] (version 2.4.1), CellTypist[19] (version 1.6.3), and scCATCH[20] (version 3.2.2). Full analysis results are detailed in Supplementary Section S6. The optional agents in CASSIA were not used for benchmarking and so all CASSIA benchmarking results demonstrate performance of CASSIA's default model. As described in the text, the uncertainty quantification agent was used for further analysis of the TS intestinal dataset; the RAG agent was used to analyze the neuronal dataset; and the subclustering agent was applied to the ProjectTILs dataset.

Analysis with each package followed the author's standard workflow. A few methods required us to specify preferences as follows: for analysis with GPTCelltype, we set the marker size parameter to 10, as recommended; tissue specificity was configured according to each dataset's context. Analysis with ScType was restricted to tissues with available marker references; gold-standard annotations were used as reference cluster labels. We used majority voting for cluster annotations within SingleR. Special considerations were required for two key datasets: the Human Cell Landscape (HCL) analysis was restricted to marker-based approaches due to annotation-metadata discrepancies, while the mixed-tissue nature of the Mouse Cell Atlas (MCA) limited our comparison to CASSIA and GPTCelltype.

Immune system analyses: We evaluated CASSIA using two immune datasets. For the PBMC68k dataset, we applied the default CASSIA workflow. For the ProjectTILs dataset, we created a challenging test case for the subclustering agent by integrating specialized human CD4+ and CD8+ T cell reference datasets that capture T cells in various functional states. Since the dataset contains exclusively T cells with subtle state-specific transcriptional differences rather than distinct cell types, it provided an ideal scenario to evaluate CASSIA's subclustering agent for fine-grained annotation, aligning with our hierarchical annotation design.

Cancer dataset analyses: We evaluated CASSIA's performance on cancer datasets, focusing on three aspects: annotation accuracy in cancer-altered transcriptional states, cancer cell detection, and metastatic signal identification. CASSIA successfully identified cancer-specific clusters across all datasets. In three metastasis datasets, CASSIA detected and correctly classified carcinoma cells in non-epithelial anatomical sites (brain and lymphoid tissues). This tissue-discordant detection of carcinoma cells, which originate exclusively from

epithelial tissues, demonstrates CASSIA's ability to identify metastatic signatures through the recognition of cells in developmentally distinct anatomical locations.

Mixed population simulation analysis: To systematically evaluate CASSIA's performance in detecting and annotating mixed cell populations, we developed a controlled benchmarking framework using the GTEx lung dataset, selected for its high-quality cell type annotations and well-defined clusters. We generated synthetic mixed populations through a structured marker-mixing strategy. For each simulation, we selected pairs of distinct cell types and created hybrid marker profiles using two mixing ratios: balanced (50:50) and dominant (80:20). In the balanced scenario, we randomly selected 50% of markers from each cell type and interleaved them in alternating order (A□B□A□B□...An□Bn□). For the dominant scenario, we selected 80% of markers from one cell type and 20% from the other, maintaining the same interleaving pattern. We evaluated these synthetic populations using both CASSIA, GPTcelltype4, and GPTcelltype4o. In the balanced mixing condition, we assessed each method's ability to detect the presence of mixed populations. For the dominant scenario, we evaluated both the detection of mixed populations and the accurate identification of the predominant cell type (defined by 80% marker representation). This framework provides a rigorous assessment of each method's capacity to handle complex cellular compositions while maintaining annotation accuracy.

Brain metastases tissue collection and scRNA-seq profiling: Tissue samples from patients with confirmed brain metastatic lesions were obtained during the published clinical trial duration of NCT03398694 and obtained and processed as detailed in Shireman *et al*. 2024[21]. Briefly, patients with up to 4 brain metastatic lesions with at least one symptomatic lesion were consented and enrolled in the clinical trial which was approved and monitored by the IU Simon Cancer Center institutional review board and data safety monitoring committee. Confirmation of brain metastatic lesion was done based on MRI imaging, IHC analysis under pathologist supervision, and previous known diagnosis of systemic primary cancer. For this study, patients with primary diagnosis of non-small cell lung cancer and an associated brain metastatic lesion needing resection were taken to the OR and a tissue specimen was obtained by the consulting surgeon. The resulting tissue was dissociated and subject to single-cell sequencing analysis using the 10x chromium method according to manufacturer's protocol. For single-cell annotation of tumor tissue, gold standard reference for presence of tumor cells and categorization of brain metastasis or primary CNS tumor was confirmed by patient health record and pathologist determination upon biopsy/resection. For reference genes within the tumor samples, the outputs of *FindAllMarkers* at the cluster level were used as input to CASSIA.

Reference List

1. Cheng, C., Chen, W., Jin, H. & Chen, X. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell-Cell Communication. *Cells* **12** (2023).
2. Pasquini, G., Rojo Arias, J.E., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* **19**, 961-969 (2021).
3. Clarke, Z.A. et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat Protoc* **16**, 2749-2764 (2021).
4. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* **20**, 194 (2019).
5. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* **21**, 1462-1465 (2024).
6. Kwong, J.C.C., Wang, S.C.Y., Nickel, G.C., Cacciamani, G.E. & Kvedar, J.C. The long but necessary road to responsible use of large language models in healthcare research. *NPJ Digit Med* **7**, 177 (2024).
7. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625-630 (2024).
8. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems* **35**, 22199-22213 (2022).
9. Wang, L. et al. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091* (2023).
10. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493-498 (2023).
11. Bsharat, S.M., Myrzakhan, A., Shen, Z. Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. *arXiv* (2023).
12. Weng, Y. et al. Large Language Models are reasoners with Self-Verification. *ArXiv* **abs/2212.09561** (2022).
13. Wang, J., Wang, J., Athiwaratkun, B., Zhang, C. & Zou, J. Mixture-of-Agents Enhances Large Language Model Capabilities. *arXiv preprint arXiv:2406.04692* (2024).
14. Li, J., Zhang, Q., Yu, Y., Fu, Q. & Ye, D. More agents is all you need. *arXiv preprint arXiv:2402.05120* (2024).
15. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* **47**, D721-D728 (2019).
16. Lamy, J.B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artif Intell Med* **80**, 11-28 (2017).
17. Ianevski, A., Giri, A.K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* **13**, 1246 (2022).
18. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163-172 (2019).

19. Dominguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
20. Shao, X. et al. scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* **23**, 100882 (2020).
21. Shireman, J.M. et al. Genomic analysis of human brain metastases treated with stereotactic radiosurgery reveals unique signature based on treatment failure. *iScience* **27**, 109601 (2024).