A081 · Benchmarking LLM-based cell type annotation for standardized reanalysis of public single-cell RNA-seq data

Presenter: Eva Fast — Pfizer

Authors: Rebecca Weiss, Ruth Marise Elgamal, Mary Piper, Stephen Christensen, Sydney Lavoie, Hendrik Luuk, Eva Fast

Session: Day 1 · Thu Oct 1, 2026 · 2:15–4:15 PM

Reanalyzing public single-cell RNA-seq (scRNA-seq) datasets is a common but time-consuming task. Generative AI could accelerate this process but requires rigorous evaluation. We assessed large language models (LLMs) for cell type annotation, a key bottleneck in scRNA-seq analysis, within a reproducible workflow integrated with LaminDB (https://lamin.ai/), our centralized data lakehouse. We analyzed eight expert-curated public datasets spanning multiple therapeutic areas that had been previously internalized into Pfizer's data ecosystem. The top 100 marker genes per cluster were formatted into prompts and submitted in parallel (8 workers, 3 replicates, 14 models), generating predicted cell types, confidence scores, and reasoning. Predicted and reference labels were standardized to the Cell Ontology using a retrieval-augmented workflow. Ontology terms and synonyms were ranked against each label by cosine similarity, and the 25 closest matches were provided to an LLM that selected the most appropriate term in context. Labels were further annotated with parent cell types to enable hierarchical performance assessment. Replicate agreement was high (Cohen's $\kappa > 0.67$), indicating systematic rather than random error. On average, ontology normalization improved concordance of predictions with expert annotations by four-fold, with further gains at higher hierarchical levels (average match rate: 48%⊠4.9%). Performance varied by both model and cell type: some populations, such as B cells, were identified accurately across models, whereas others exhibited model-specific strengths. No single model consistently outperformed the others. In conclusion, LLM-based annotation coupled with ontology standardization provides scalable, benchmarkable results suitable for large-scale reanalysis, while expert review remains essential for validating biologically meaningful predictions.

Benchmarking LLM-based cell type annotation for standardized reanalysis of public single-cell RNA-seq data

Rebecca Weiss, Ruth Marise Elgamal, Mary Piper, Stephen Christensen, Sydney Lavoie, Hendrik Luuk, Eva Fast*

Pfizer Research and Development, Pfizer Inc., Cambridge, MA, USA.

Reanalyzing public single-cell RNA-seq (scRNA-seq) datasets is a common but time-consuming task. Generative AI could accelerate this process but requires rigorous evaluation. We assessed large language models (LLMs) for cell type annotation, a key bottleneck in scRNA-seq analysis, within a reproducible workflow integrated with LaminDB (https://lamin.ai/), our centralized data lakehouse.

We analyzed eight expert-curated public datasets spanning multiple therapeutic areas that had been previously internalized into Pfizer's data ecosystem. The top 100 marker genes per cluster were formatted into prompts and submitted in parallel (8 workers, 3 replicates, 14 models), generating predicted cell types, confidence scores, and reasoning.

Predicted and reference labels were standardized to the Cell Ontology using a retrieval-augmented workflow. Ontology terms and synonyms were ranked against each label by cosine similarity, and the 25 closest matches were provided to an LLM that selected the most appropriate term in context. Labels were further annotated with parent cell types to enable hierarchical performance assessment.

Replicate agreement was high (Cohen's $\kappa > 0.67$), indicating systematic rather than random error. On average, ontology normalization improved concordance of predictions with expert annotations by four-fold, with further gains at higher hierarchical levels (average match rate: 48%±4.9%). Performance varied by both model and cell type: some populations, such as B cells, were identified accurately across models, whereas others exhibited model-specific strengths. No single model consistently outperformed the others.

In conclusion, LLM-based annotation coupled with ontology standardization provides scalable, benchmarkable results suitable for large-scale reanalysis, while expert review remains essential for validating biologically meaningful predictions.