EmCellLLM: Human Peri-Implantation Embryonic Cell Annotation Based on Large Language Models

Xiaorui Guo[1] **Zhiwei Liu**[2] **Qianqian Xie**[3,4] **Sophia Ananiadou**[2,5]
[1]The University of Edinburgh [2]The University of Manchester
[3]School of Artificial Intelligence, Wuhan University
[4]Center for Language and Information Research, Wuhan University [5]ELLIS Manchester
X.Guo-46@sms.ed.ac.uk, {zhiwei.liu,sophia.ananiadou}@manchester.ac.uk
xieq@whu.edu.cn

Abstract

The advent of single-cell RNA sequencing has enabled unprecedented resolution of cell fate decisions and regulatory mechanisms during peri-implantation human embryogenesis, in which accurate cell type annotation is a fundamental prerequisite and the first step for subsequent fate and mechanism inference. Large language models (LLMs) have demonstrated outstanding performance in various fields. However, current studies mostly rely on traditional methods and have not explored the application of LLMs in the field of human embryonic cell annotation. The main reason is the lack of instruction tuning datasets and evaluation benchmarks. In this paper, we proposed EmCellLLM, the first open sourced LLMs that are specialized for human embryonic cell type prediction task based on fine-tuning Qwen3-8B with EmCell4Instruction, the first embryonic cell type prediction instruction dataset. To support LLM instruction tuning, we also build EmCellBench, the first benchmark for evaluating human embryonic cell type prediction ability of LLMs. We compare our models with a variety of LLMs on EmCellBench, where our model outperforms all other open-sourced LLMs as well as DeepSeek.

1 Introduction

Human peri-implantation embryos (approximately days 5–14) represent a critical developmental window in which the embryo transitions from an undifferentiated state to distinct lineages while interacting with the endometrium (Zhai et al., 2022; David et al., 2023). Abnormal development during this period is strongly associated with implantation failure, early pregnancy loss, placental dysfunction, and developmental origins of disease (Norwitz et al., 2001).

Performing single-cell sequencing on the peri-implantation stage and accurately annotating the cells allows the resolution of lineage trajectories, dynamic transitional states, and key regulatory networks at single-cell resolution, thereby reconstructing a comprehensive map of early human development (Ostrer et al., 2006; Latham, 2023; Anger et al., 2025). The challenges of single-cell annotation in peri-implantation human embryos stem from three main factors. (1) Human embryo samples are scarce due to ethical and practical limitations, and each developmental stage contains only a small number of cells, resulting in sparse and noisy datasets. (Petropoulos et al., 2016; Gul and Zhang, 2025). (2) This stage is highly dynamic, with rapid cell fate diversification and continuous, making clear cell-type boundaries and marker-based annotation difficult (Molè et al., 2021). (3) Reference atlases are usually constructed by integrating datasets from multiple sources, but strong batch effects arising from different experimental protocols and sequencing platforms can obscure biological signals and lead to inconsistent or inaccurate annotation (Weatherbee et al., 2023). Accurate annotation is therefore essential for understanding early human development and its applications in reproductive medicine and stem cell modeling.

Traditional embryonic cell annotation relies on known marker genes and manual expert judgment to identify cell types and manually assign cluster labels which is time-consuming, hard to scale, and limited by incomplete marker knowledge and subjective interpretation (Luecken and Theis, 2019; Cheng et al., 2023). Automated cell annotation methods reduce manual curation by leveraging marker-based approaches (e.g., CellMarker 2.0), reference mapping and label transfer methods (e.g., Seurat, SingleR, scmap, TemporalVAE), and supervised or deep learning classifiers (e.g., CellTypist, scTab, scANVI/scVI) (Hu et al., 2023; Stuart et al., 2019; Aran et al., 2019; Kiselev et al., 2018; Liu et al., 2025b; Domínguez Conde et al., 2022; Fischer et al., 2024; Lopez et al., 2018; Xu et al., 2021a). These methods assign cell identities by

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 382–391

mapping query data to reference atlases or learning expression–label relationships. While effective for well-characterized cell types, their performance depends on reference coverage and annotation consistency and is sensitive to batch effects and cross-dataset variation (Cheng et al., 2023; Zhao et al., 2025). In human embryogenesis, where cell states are continuous and reference atlases integrated from multiple sources and incomplete, these methods often fail to resolve rare populations, leading to misannotation or low-confidence predictions.

LLM-based automated cell annotation methods have advanced rapidly in recent years, with the core idea of transforming single-cell transcriptomic data into representations that can be processed by LLMs (Hou and Ji, 2024). For example, the C2S approach converts single-cell gene expression profiles into text sequences by ranking genes in descending order of expression levels, enabling large language models to directly "read" gene expression patterns. While preserving the complexity of single-cell data, this formulation generates biologically meaningful cell representations that can be used for downstream cell type prediction (Levine et al., 2024).

To address the above issues, we construct the first instruction-tuning dataset for embryonic cell type prediction, EmCell4Instruction, to support LLM fine-tuning, which includes a single-cell transcriptome atlas integrating eight human embryo datasets spanning developmental stages from day 3 to day 18.5. Based on this dataset, we propose EmCellLLM, the first open-source LLM for embryonic cell type prediction, which developed by fine-tuning Qwen3-8B on EmCell4Instruction. To systematically evaluate model performance, we further introduce EmCellBench, a benchmark for embryonic cell type annotation. Experimental results on EmCellBench demonstrate that EmCellLLM achieves state-of-the-art performance among open-source LLMs, and we will release both the dataset and the model to the public.

Our main contributions are as follows:

(1) We construct EmCell4Instruction, the first embryonic cell type prediction instruction-tuning dataset.

(2) We develop EmCellLLM, the first open-sourced LLMs that are specialized for the embryonic cell type prediction task.

(3) We build EmCellBench, the first benchmark to evaluate the embryonic cell type prediction ability of LLMs. The results on EmCellBench demonstrate that our model overtakes other open-sourced LLMs.

2 Related work

2.1 Cell Type Annotation

There are few studies semi-automatic or automatic annotate human embryonic cells. Most of them are based on traditional label transfer methods or deep learning methods. Xu et al. (2021b) use Seurat-based integration combined with marker gene curation and reference-guided label transfer for semi-automatic annotation. Zhao et al. (2025) use Seurat-based integration and label transfer to project query datasets onto a reference human embryo atlas for cell type annotation. TemporalVAE annotates cell types by mapping new cells into an atlas-informed latent space and transferring labels from nearest neighbors (Liu et al., 2025b). (Proks et al., 2025) use deep learning models to integrate single-cell RNA-seq data from mouse and human pre-implantation embryos for cross-species cell type annotation and developmental state prediction. However, there is currently no open-sourced LLM specifically designed for the annotation of human embryonic cells.

2.2 Open Sourced Large Language Models

Significant research efforts have focused on developing open-source LLMs as alternatives to closed-source models such as GPT series, with the aim of enabling more accessible research on improving and applying LLMs. Well-known series of open-sourced, general-purpose language models include LLama series (Dubey et al., 2024), DeepSeek series (Liu et al., 2025a), Qwen series (Yang et al., 2025). There are also many open-sourced LLMs for specific domains, including FMDLlama (Liu et al., 2025d) for financial misinformation detection, human cell C2S LLMs (Levine et al., 2024) for somatic cell analysis, ExTES-LLaMA (Zheng et al., 2023) for emotional support chatbots, MeLLaMA for Medical Applications (Xie et al., 2024), and ConspEmoLLM (Liu et al., 2024, 2025c) for conspiracy detection. In this work, we extend the inventory of domain-specific LLMs, by developing the first open-sourced LLM for embryonic cell type annotation.

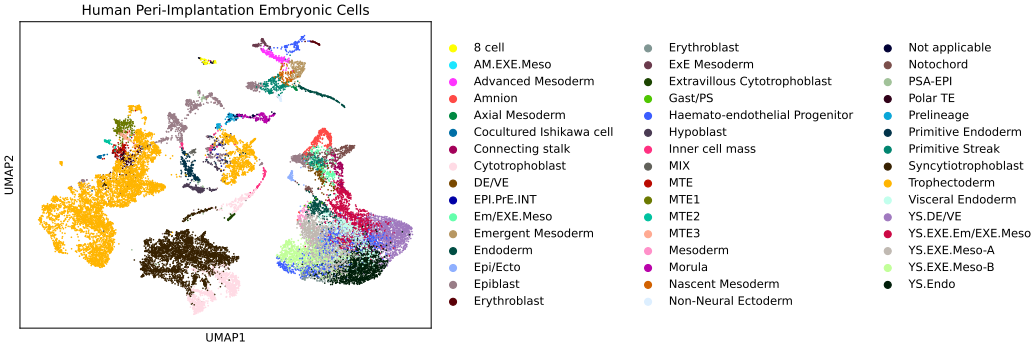

Figure 1: UMAP of Human Peri-Implantation Embryonic Cells

3 Methods

3.1 Task formalization

We approach cell type detection as a generative task, applying a generative model as a foundation. This generative model is an autoregressive language model $P_\phi(y|x)$, parameterized using pre-trained weights ϕ. Each task (t) is represented as a set of context-target pairs: $D_t = (q_i^t, r_i^t)_{i=1,2,...N_t}$, where the context q is a token sequence containing the task description, cell sentence, and query, and r is a further token sequence containing the answer to the query. The model is optimized based on the training dataset, with the aim of maximizing the objective of conditional language modeling to improve prediction and generation performance.

3.2 Construction of instruction tuning dataset

Data	Train	Test
Embryonic cells	19652	4913

Table 1: Dataset statistics. Detailed statistics of each cell type can be found at Table 3.

3.2.1 Raw data

We build our instruction tuning dataset based on a single-cell human embryonic cells transcriptome atlas which integrates 8 human embryo datasets and spanning from embryonic development day 3 to day 18.5 (Liu et al., 2025b). Among them, data from Xiao et al. (2024), Cui et al. (2025) and Tyser et al. (2021) all come from one in vivo human embryo, data from Petropoulos et al. (2016) and Molè et al. (2021) come from 88 and 16 ex vivo embryos, and data from Xiang et al. (2020), Zhou et al. (2019) and Liu et al. (2022) come from 42, 65 and 25 in vitro embryos, individually. Figure 1 presents the UMAP of the datasets.

3.2.2 Construction of the EmCell instruction tuning dataset (EmCell4Instruction) and EmCell benchmark (EmCellBench)

We use the raw dataset without batch correction as the basis for building the instruction dataset. We split the original data into train and test sets. The dataset statistics are presented in Table 1. We construct instruction-tuning data based on the templates in Table 4. One of the 15 instruction templates is randomly selected for each item to enhance diversity. *[num_genes]* represents the top K highly expressed genes used for cell type classification. In this paper, we set K to 200. *[cell_sentence]* is obtained by converting cell expression profiles into natural language sentences in C2S (Levine et al., 2024), which transforms single-cell gene expression data into rank-ordered gene sequences by expression level, enabling LLMs to effectively process transcriptomic information while leveraging standard transformer libraries. LLMs directly output the predicted cell type as their response. Figure 2 presents examples used to fine-tune the LLM.

After constructing the instruction data. We collect the train data as instruction-tuning data (EmCell4Instruction) and test data as the embryonic cells type prediction benchmark (EmCellBench), which are used to fine-tune the LLMs and evaluate the ability of LLMs in the embryonic cells type prediction domain, respectively.

3.3 EmCellLLM

We built EmCellLLM by fine-tuning Qwen3-8b (Yang et al., 2025) using the EmCell4Instruction dataset. The models are trained based on the AdamW optimizer (Loshchilov and Hutter, 2017) for five epochs, using DeepSpeed (Rasley et al., 2020) to reduce memory usage. We set the batch size to 128. The initial learning rate is set to 1e-5 with a warm-up ratio of 0.1. All models are trained

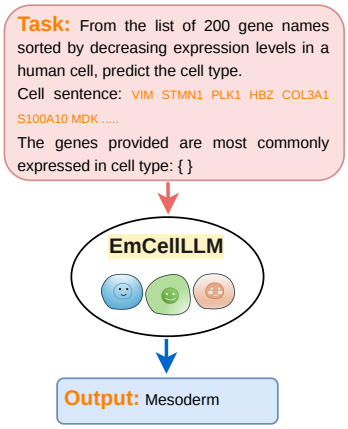

Figure 2: An overview of instruction tuning of EmCellLLM

Model	ACC	PRE	REC	F1
Qwen3-8b-R	0.135	0.023	0.029	0.008
Qwen3-14b-R	0.161	0.053	0.056	0.028
Qwen3-32b-R	0.148	0.029	0.035	0.014
Qwen3-8b	0.138	0.018	0.030	0.007
Qwen3-14b	0.129	0.027	0.045	0.012
Qwen3-32b	0.120	0.034	0.046	0.019
Qwen2.5-72b	0.078	0.004	0.005	0.002
Llama3.1-8b	0.062	0.015	0.025	0.009
Llama3.3-70b	0.092	0.039	0.063	0.026
DeepSeek	0.168	0.068	0.052	0.032
C2S-Pythia-410M	0.884	0.770	0.752	0.745
EmCellLLM	**0.905**	**0.858**	**0.811**	**0.827**

Table 2: Results on EmCellBench.

on four NVIDIA L40s GPUs, each with 48GB of memory. We compare with the domain-specific C2S-Pythia-410M (Levine et al., 2024) cell type prediction model, which has been trained on multiple datasets from CellxGene and the Human Cell Atlas for cell type prediction. We further continue training based on EmCell4Instruction as the baseline. Figure 2 provides an overview of instruction tuning of EmCellLLM for the embryonic cell type prediction task.

4 Experiments

4.1 Baseline models

Our extensive evaluation of open-source and proprietary LLMs included the following reasoning-focused models: Qwen3 reasoning variants (8B-R, 14B-R, and 32B-R) (Yang et al., 2025). We also assessed a diverse set of no-reasoning LLMs, DeepSeek-V3.2-Chat (DeepSeek-C) (Liu et al., 2025a), the Qwen3 no-reasoning models (8B, 14B, and 32B), Qwen2.5-72B-Instruct (Qwen72B) (Qwen et al., 2025), Llama-3.1-8B-Instruct, and Llama-3.3-70B-Instruct (Dubey et al., 2024). Additionally, we compare with the domain-specific C2S-Pythia-410M cell type prediction model, which has been trained on multiple datasets from CellxGene and the Human Cell Atlas for cell type prediction. We further continue training based on EmCell4Instruction as the baseline.

4.2 Evaluation methods

We uses metrics such as Accuracy, Precision, Recall, Macro-F1 for embryonic cell type detection task.

4.3 Results

Table 2 presents the results on EmCellBench. From the table, we can see EmCellLLM achieve SOTA results among all other open-sourced LLMs, including DeepSeek. C2S-Pythia-410M follows closely, demonstrating the effectiveness of the instruction-tuning approach. Other open-source general-purpose LLMs perform poorly, likely because they lack domain-specific understanding and require further training to acquire such knowledge. Reasoning models of the same size perform slightly better than their non reasoning counterparts such as the Qwen3 series, indicating that reasoning provides an advantage when handling unseen and complex knowledge tasks. Larger models generally achieve better performance, and among the untrained models, DeepSeek performs the best, suggesting that models with more parameters tend to possess richer knowledge and stronger overall capabilities.

5 Conclusion

In this paper, we propose EmCellLLM, the first LLM for embryonic cell annotation. We also construct an embryonic cell prediction instruction dataset (EmCell4Instruction) and an evaluation benchmark (EmCellBench). We conduct a comprehensive analysis of the performance of EmCellLLM, as well as a variety of LLMs on the EmCellBench benchmark. The results indicate that EmCellBench performs exceptionally well in the embryonic cell annotation task, achieving SOTA compared to the other open-sourced LLMs.

In the future, we aim to augment the EmCell4Instruction and EmCellBench datasets by collecting a more comprehensive single-cell sequencing atlas of human embryonic development, which can help further improve the EmCellLLM and evaluate the embryonic cell annotation ability of LLMs more comprehensively.

Limitations

The potential limitations of our work may be summarized as follows:

Due to restricted computational resources, we only carried out instruction-tuning/evaluation of the task using 8b/14b/32b/70b/72b LLMs. As such, we have not considered the impact of using larger models on the Embryonic cell type prediction task.

Due to the scarcity and limited temporal coverage of in vivo human embryo data, we integrate ex vivo and in vitro datasets to increase diversity and scale, but these remain less ideal than in vivo data due to potential domain shifts and batch effects.

While our dataset covers embryonic stages from day 3 to day 18.5, spanning the peri-implantation period, extending it to include the full developmental trajectory from day 0 to birth would enable the construction of a more comprehensive human embryonic atlas and improve the robustness and generalizability of annotation models.

Acknowledgments

This work was supported by the British Heart Foundation Manchester Research Excellence Award (RE/24/130017). This work was supported by the CCF-Tencent Rhino-Bird Open Research Fund (CCF-Tencent RAGR20250115), Wuhan Natural Science Foundation Exploratory Program (Morning Light Program) Project (2026040301020029) and the Intelligent Computing Center of the National Cybersecurity Talent and Innovation Base, Wuhan.

References

Martin Anger, Katerina Komrskova, Ahmed Z Balboula, Paolo Rinaudo, and Jason G Knott. 2025. Fertilization and early embryogenesis: from research to clinical practice. *Frontiers in Cell and Developmental Biology*, 12:1547205.

Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172.

Changde Cheng, Wenan Chen, Hongjian Jin, and Xiang Chen. 2023. A review of single-cell rna-seq annotation, integration, and cell–cell communication. *Cells*, 12(15):1970.

Lina Cui, Sirui Lin, Xiaolong Yang, Xinwei Xie, Xiaoyan Wang, Nannan He, Jingyu Yang, Xin Zhang, Xiaojian Lu, Xiaodi Yan, et al. 2025. Spatial transcriptomic characterization of a carnegie stage 7 human embryo. *Nature cell biology*, 27(2):360–369.

Laurent David, Alexandre Bruneau, Thomas Fréour, Nicolas Rivron, and Hilde Van de Velde. 2023. An update on human pre-and peri-implantation development: a blueprint for blastoids. *Current opinion in genetics & development*, 83:102125.

César Domínguez Conde, Chao Xu, Louie B Jarvis, Daniel B Rainbow, Sara B Wells, Tamir Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. 2022. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Felix Fischer, David S Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J Theis. 2024. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611.

Samina Gul and Chengting Zhang. 2025. Single cell rna sequencing and its impact on understanding human embryo development. *International Journal of Molecular Sciences*, 26(16):7741.

Wenpin Hou and Zhicheng Ji. 2024. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature methods*, 21(8):1462–1465.

Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. 2023. 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.

Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. 2018. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362.

Keith E Latham. 2023. Preimplantation embryo gene expression: 56 years of discovery, and counting. *Molecular reproduction and development*, 90(4):169–200.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, et al. 2024. Cell2sentence: teaching large language models the language of biology. *BioRxiv*, pages 2023–09.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Dandan Liu, Yidong Chen, Yixin Ren, Peng Yuan, Nan Wang, Qiang Liu, Cen Yang, Zhiqiang Yan, Ming Yang, Jing Wang, et al. 2022. Primary specification of blastocyst trophectoderm by scrna-seq: New insights into embryo implantation. *Science Advances*, 8(31):eabj3725.

Yijun Liu, Fangxin Cai, Melania Barile, Yi Chang, Dandan Cao, and Yuanhua Huang. 2025b. Temporalvae: atlas-assisted temporal mapping of time-series single-cell transcriptomes during embryogenesis. *Nature Cell Biology*, pages 1–11.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024. Conspemollm: Conspiracy theory detection using an emotion-based large language model. *arXiv preprint arXiv:2403.06765*.

Zhiwei Liu, Paul Thompson, Jiaqi Rong, and Sophia Ananiadou. 2025c. Conspemollm-v2: A robust and stable model to detect sentiment-transformed conspiracy theories. *arXiv preprint arXiv:2505.14917*.

Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2025d. Fmdllama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1153–1157.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. 2018. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Malte D Luecken and Fabian J Theis. 2019. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):MSB188746.

Matteo A Molè, Tim HH Coorens, Marta N Shahbazi, Antonia Weberling, Bailey AT Weatherbee, Carlos W Gantner, Carmen Sancho-Serra, Lucy Richardson, Abbie Drinkwater, Najma Syed, et al. 2021. A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nature communications*, 12(1):3679.

Errol R Norwitz, Danny J Schust, and Susan J Fisher. 2001. Implantation and the survival of early pregnancy. *New England Journal of Medicine*, 345(19):1400–1408.

Harry Ostrer, DI Wilson, and NA Hanley. 2006. Human embryo and early fetus research. *Clinical genetics*, 70(2):98–107.

Sophie Petropoulos, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. 2016. Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026.

Martin Proks, Nazmus Salehin, and Joshua M Brickman. 2025. Deep learning-based models for preimplantation mouse and human embryos based on single-cell rna sequencing. *Nature methods*, 22(1):207–216.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902.

Richard CV Tyser, Elmir Mahammadov, Shota Nakanoh, Ludovic Vallier, Antonio Scialdone, and Shankar Srinivas. 2021. Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, 600(7888):285–289.

Bailey AT Weatherbee, Carlos W Gantner, Lisa K Iwamoto-Stohl, Riza M Daza, Nobuhiko Hamazaki, Jay Shendure, and Magdalena Zernicka-Goetz. 2023. Pluripotent stem cell-derived model of the post-implantation human embryo. *Nature*, 622(7983):584–593.

Lifeng Xiang, Yu Yin, Yun Zheng, Yanping Ma, Yonggang Li, Zhigang Zhao, Junqiang Guo, Zongyong Ai, Yuyu Niu, Kui Duan, et al. 2020. A developmental landscape of 3d-cultured human pre-gastrulation embryos. *Nature*, 577(7791):537–542.

Zhenyu Xiao, Lina Cui, Yang Yuan, Nannan He, Xinwei Xie, Sirui Lin, Xiaolong Yang, Xin Zhang, Peifu Shi, Zhifeng Wei, et al. 2024. 3d reconstruction of a gastrulating human embryo. *Cell*, 187(11):2855–2874.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. 2021a. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):MSB20209620.

Yichi Xu, Tengjiao Zhang, Qin Zhou, Mengzhu Hu, Yao Qi, Yifang Xue, Lihui Wang, Yuxiao Nie, Zhirong Bao, and Weiyang Shi. 2021b. A single-cell transcriptome atlas of human early embryogenesis. *BioRxiv*, pages 2021–11.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jinglei Zhai, Zhenyu Xiao, Yiming Wang, and Hongmei Wang. 2022. Human embryonic development: from peri-implantation to gastrulation. *Trends in cell biology*, 32(1):18–29.

Cheng Zhao, Alvaro Plaza Reyes, John Paul Schell, Jere Weltner, Nicolás M Ortega, Yi Zheng, Åsa K Björklund, Laura Baqué-Vidal, Joonas Sokka, Ras Trokovic, et al. 2025. A comprehensive human embryo reference tool using single-cell rna-sequencing data. *Nature methods*, 22(1):193–206.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Fan Zhou, Rui Wang, Peng Yuan, Yixin Ren, Yunuo Mao, Rong Li, Ying Lian, Junsheng Li, Lu Wen, Liying Yan, et al. 2019. Reconstituting the transcriptome and dna methylome landscapes of human implantation. *Nature*, 572(7771):660–664.

A Data Details

Table 3 presents the statistics of each cell type. Table 4 shows the instruction templates for EmCell4Instruction and EmCellBench.

B Error Analysis

Figure 3 presents the confusion matrix of EmCell-LLM on EmCellBench. The confusion matrix reveals the misclassification among several cell types. For example, Gast/PS cells are predominantly predicted as Endoderm. This pattern is likely because of the extremely small training sample size (only 4 cells) prevents the model from learning robust class-specific features. Additionally, EPI.PrE.INT cells are dispersed across multiple predicted classes, including Epiblast, Primitive Endoderm, and Trophectoderm. EPI.PrE.INT cells is labeled as a mix cell population in training data which including Epiblast, Primitive Endoderm, and intermediate cells. Therefore, it's it is reasonable to predict Epiblast as Primitive Endoderm. However, as there are only 12 cells used to train model, model did not receive sufficient information for training causing a part of cells were mislabeled as Trophectoderm.

Cell Type	Full Name	Train	Test
8 cell	8-cell stage embryo	50	14
AM.EXE.Meso	Amnion-associated Extraembryonic Mesoderm	1	35
Advanced Mesoderm	Advanced Mesoderm	129	0
Amnion	Amnion	239	60
Axial Mesoderm	Axial Mesoderm	18	5
Cocultured Ishikawa cell	Cocultured Ishikawa cells	31	12
Connecting stalk	Connecting stalk	160	34
Cytotrophoblast	Cytotrophoblast	868	216
DE/VE	Definitive Endoderm / Visceral Endoderm	106	21
EPI.PrE.INT	Epiblast–Primitive Endoderm Intermediate	12	4
Em/EXE.Meso	Embryonic / Extraembryonic Mesoderm	137	45
Emergent Mesoderm	Emergent Mesoderm	139	46
Endoderm	Endoderm	290	77
Epi/Ecto	Epiblast / Ectoderm	54	17
Epiblast	Epiblast	956	228
Erythroblast	Erythroblast	29	3
Erythroblast	Erythroblast	226	64
ExE Mesoderm	Extraembryonic Mesoderm	70	13
Extravillous Cytotrophoblast	Extravillous Cytotrophoblast	37	3
Gast/PS	Gastrulation / Primitive Streak	4	2
Haemato-endothelial Progenitor	Haemato-endothelial progenitor	995	265
Hypoblast	Hypoblast	171	31
Inner cell mass	Inner Cell Mass	81	15
MIX	Mixed population	31	8
MTE	Mural Trophectoderm	111	23
MTE1	Mural Trophectoderm subtype 1	229	46
MTE2	Mural Trophectoderm subtype 2	35	4
MTE3	Mural Trophectoderm subtype 3	59	12
Mesoderm	Mesoderm	52	12
Morula	Morula	132	29
Nascent Mesoderm	Nascent Mesoderm	74	24
Non-Neural Ectoderm	Non-neural ectoderm	21	8
Not applicable	Not applicable	75	21
Notochord	Notochord	213	68
PSA-EPI	Post-implantation Stage Epiblast	37	7
Polar TE	Polar Trophectoderm	164	51
Prelineage	Pre-lineage stage	86	20
Primitive Endoderm	Primitive Endoderm	181	65
Primitive Streak	Primitive Streak	293	60
Syncytiotrophoblast	Syncytiotrophoblast	2600	655
Trophectoderm	Trophectoderm	5085	1248
Visceral Endoderm	Visceral Endoderm	378	76
YS.DE/VE	Yolk Sac Definitive Endoderm / Visceral Endoderm	1156	242
YS.EXE.Em/EXE.Meso	Yolk Sac Embryonic / Extraembryonic Mesoderm	995	242
YS.EXE.Meso-A	Yolk Sac Extraembryonic Mesoderm subtype A	1151	303
YS.EXE.Meso-B	Yolk Sac Extraembryonic Mesoderm subtype B	615	169
YS.Endo	Yolk Sac Endoderm	1076	310
Total		19652	4913

Table 3: Detailed statistics of each cell type.

Instruction Template
(1) The following is a list of {num_genes} gene names ordered by descending expression level in a human cell. Your task is to give the cell type which this cell belongs to based on its gene expression. Cell sentence: {cell_sentence}. The cell type corresponding to these genes is:
(2) Below is a list of {num_genes} gene names in order of descending expression level from a human cell. Based on this, predict what the cell type of this cell is. Cell sentence: {cell_sentence}. These genes are most likely associated with cell type:
(3) Given the list of {num_genes} gene names ordered by descending expression level from a human cell, identify the cell type. Cell sentence: {cell_sentence}. The probable cell type for these genes is:
(4) Analyze the following list of {num_genes} genes sorted by decreasing expression levels in a human cell and determine its cell type. Cell sentence: {cell_sentence}. Based on these genes, the corresponding cell type is:
(5) From the list of {num_genes} gene names ordered by decreasing expression level in a human cell, infer the cell type of the cell. Cell sentence: {cell_sentence}. These genes suggest the cell type is most likely:
(6) The {num_genes} gene names below are listed by descending expression level in a human cell. Predict the cell type based on this information. Cell sentence: {cell_sentence}. These genes are indicative of cell type:
(7) Below is a list of {num_genes} gene names sorted by descending expression in a human cell. Determine the cell type of this cell from its expressed genes. Cell sentence: {cell_sentence}. The associated cell type for these genes appears to be:
(8) The following list contains {num_genes} gene names ordered by descending expression level in a human cell. Deduce the cell type based on this. Cell sentence: {cell_sentence}. These genes typically correspond to cell type:
(9) Here is a list of {num_genes} genes in order of descending expression level from a human cell. Identify the cell type of this cell based on this information. Cell sentence: {cell_sentence}. The expected cell type based on these genes is:
(10) Examine the list of {num_genes} gene names sorted by decreasing expression in a human cell, and from this, identify its cell type. Cell sentence: {cell_sentence}. These genes are commonly found in cell type:
(11) The {num_genes} gene names below are arranged by descending expression level in a human cell. Determine the cell type of this cell. Cell sentence: {cell_sentence}. The cell type that these genes are most commonly linked with is:
(12) Here is a list of {num_genes} genes ordered by expression level in a human cell. Based on this information, identify the cell type. Cell sentence: {cell_sentence}. Based on the expression levels, the cell type would likely be:
(13) From the list of {num_genes} gene names sorted by decreasing expression levels in a human cell, predict the cell type. Cell sentence: {cell_sentence}. The genes provided are most commonly expressed in cell type:
(14) Below is a list of {num_genes} genes ordered by descending expression in a human cell. Use this information to determine the cell type. Cell sentence: {cell_sentence}. These genes suggest the cell type is most likely:
(15) The {num_genes} gene names below are listed by descending expression levels in a human cell. Based on this, predict the cell type. Cell sentence: {cell_sentence}. The cell type corresponding to these genes is:

Table 4: Instructions used for EmCell4Instruction and EmCellBench.

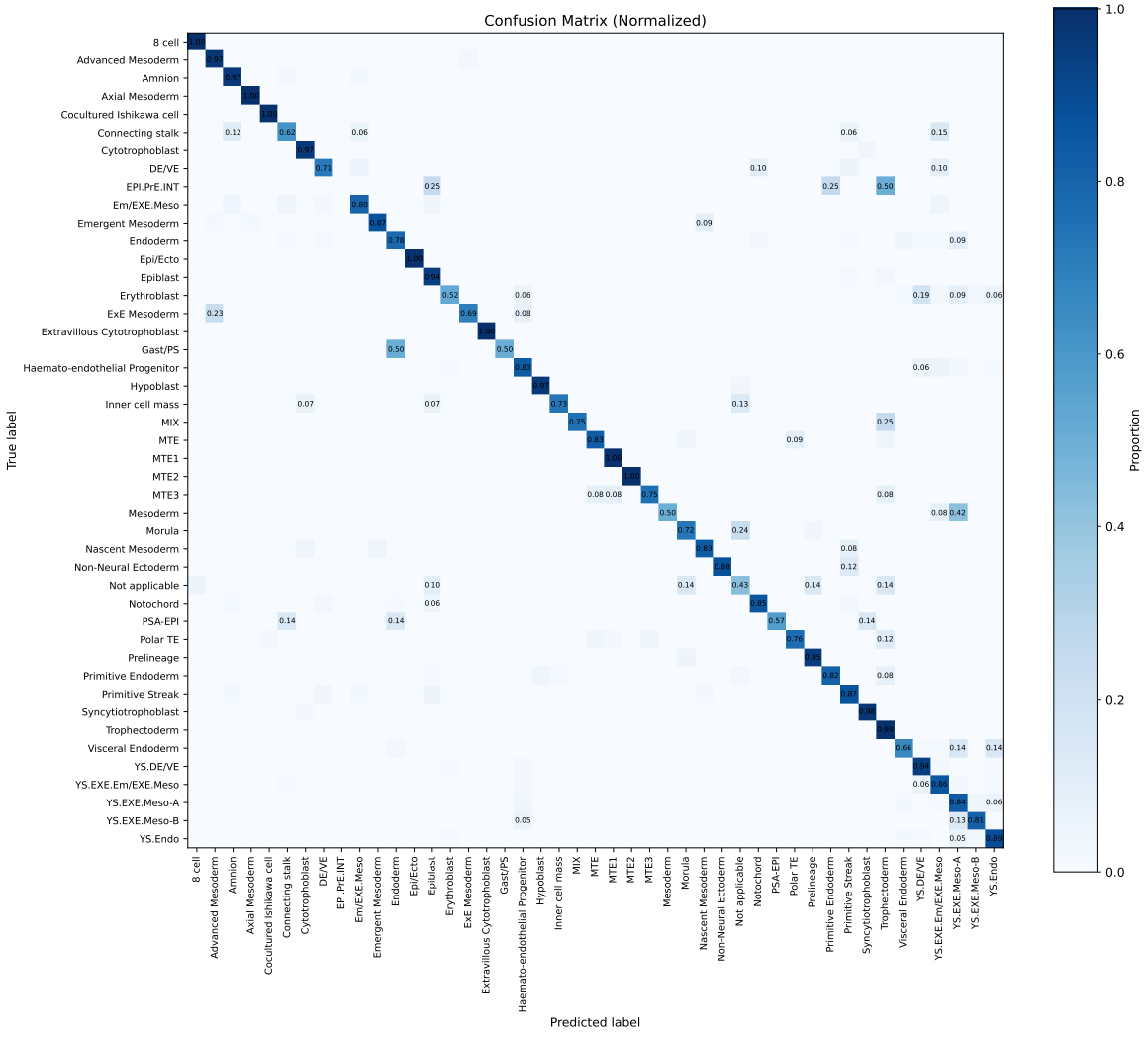

Figure 3: Confusion matrix of EmCellLLM on EmCellBench.