DiagnoLLM: Integrating Bayesian Deconvolution and Large Language Models for Interpretable Disease Diagnosis

Anonymous ACL submission

Abstract

While deep learning models have shown strong performance in clinical disease classification, their black-box nature limits adoption in high-stakes healthcare settings. We present **DiagnoLLM**, a novel framework that combines *Bayesian deconvolution* with *Large Language Model (LLM)*-driven interpretability to bridge this gap. First, DiagnoLLM applies *GP-unmix*, a Gaussian Process-based hierarchical model, to infer cell-type-specific gene expression profiles from bulk RNA-seq and single-cell data. A deep learning model trained on these features achieves high predictive performance in Alzheimer's Disease (AD) classification (**88.0% accuracy**). To enhance transparency, we introduce an LLM-based interpretability plug-in that generates faithful, audience-specific diagnostic reports grounded in model outputs, eQTL signals, and domain knowledge. The resulting reports align with clinical reasoning while maintaining fidelity to underlying predictions. DiagnoLLM demonstrates that LLMs, when used as structured narrative generators rather than classifiers, can play a critical role in building trust in biomedical AI. Code and data are available at: DiagnoLLM.

1 Introduction

Accurate disease diagnosis from transcriptomic data remains a central challenge in biomedical AI, particularly due to the limitations of bulk RNA-seq in capturing cell-type-specific (CTS) expression (Blumenfeld et al., 2024; Natri et al., 2024). Bulk measurements reflect an average over heterogeneous cell populations, masking critical disease signals that are localized to specific cell types. For instance, in Alzheimer's Disease (AD), dysregulation in microglia or astrocytes may be lost when averaged with expression from neurons or other brain cells (Brendel et al., 2022). While single-cell RNA sequencing (scRNA-seq) improves resolution (Tasic et al., 2018; Paik et al., 2020), it remains expensive, technically challenging, and sparsely available in clinical datasets. To address this, deconvolution methods have emerged to estimate CTS profiles from bulk data (Xu et al., 2025). Yet, existing approaches are often sensitive to clustering noise and lack principled uncertainty modeling (Torroja and Sanchez-Cabo, 2019). Furthermore, most prediction pipelines based on gene expression overlook valuable prior knowledge from regulatory genomics.

To overcome the limitations of current disease prediction pipelines, we introduce **DiagnoLLM**, a unified framework for interpretable diagnosis that integrates statistical deconvolution, genetic regulatory reasoning, and natural language explanation. As illustrated in Figure 1, the framework proceeds in two stages, each addressing a fundamental bottleneck in existing biomedical AI workflows. In the first stage, we propose **GP-unmix**, a novel Gaussian Process-based Bayesian model that infers cell-type-specific (CTS) gene expression from bulk RNA-seq and single-cell RNA-seq data (Tasic et al., 2018; Paik et al., 2020). Unlike traditional deconvolution methods (Torroja and Sanchez-Cabo, 2019; Xu et al., 2025), GP-unmix introduces posterior updates that explicitly model biological variability across samples. This allows the model to reliably recover gene-cell-type expression profiles even when reference data and target samples differ by tissue type or species. The result is a high-quality, uncertainty-aware CTS expression matrix suitable for downstream disease modeling (Brendel et al., 2022). In the second stage, we combine these inferred CTS features with **eQTL-derived signals**, genetic variants known to regulate gene expression in a cell-type-specific or disease-associated manner (Nica and Dermitzakis, 2013; Natri et al., 2024). These regulatory features enhance biological specificity and provide causal grounding, enabling a two-layer deep learning model to focus on disease-relevant genes rather than spurious expres-

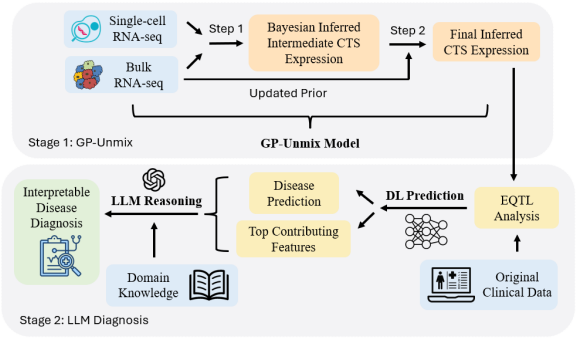

Figure 1: **Overview of the DIAGNOLLM framework.** Stage 1 (**GP-Unmix**) performs Bayesian deconvolution of bulk RNA-seq into CTS expression using single-cell references. Stage 2 combines eQTL-informed DL predictions with LLM-based reasoning to produce human-readable diagnostic reports, linking model outputs with clinical interpretability.

sion fluctuations (Blumenfeld et al., 2024). We find that incorporating eQTL information significantly boosts predictive accuracy in Alzheimer's Disease classification.

While recent studies have explored the use of large language models (LLMs) for numerical reasoning and classification in biomedical domains (Yang et al., 2023; Gao et al., 2024; Hegselmann et al., 2023; Han et al., 2022), we observe that LLMs underperform compared to our deep learning model in direct disease classification tasks. Instead, we deploy the LLM as a structured interpretability plug-in. Given the DL model's predicted probability and feature attributions (e.g., from Integrated Gradients), the LLM generates human-readable, domain-aware reports tailored for physicians and patients. These reports align with the underlying model output and leverage biomedical domain knowledge (Yu et al., 2025; Omar et al., 2024), thereby enhancing trust and transparency in clinical decision-making.

Our key contributions are as follows:

(1) GP-unmix for robust CTS inference: A Bayesian deconvolution method with posterior refinement for accurate, uncertainty-aware CTS expression estimation.

(2) eQTL-informed disease prediction: A deep learning classifier trained on CTS and regulatory features, demonstrating improved disease classification performance.

(3) LLM-based interpretability plug-in: A prompting framework that generates clinically meaningful reports aligned with model outputs and domain knowledge.

By combining mechanistic modeling, biologically informed prediction, and natural language explanation, **DiagnoLLM** offers a comprehensive solution for interpretable AI in clinical genomics.

2 Related Work

LLMs have recently been adapted for numerical and structured data tasks, with methods such as LIFT (Dinh et al., 2022) and TabLLM (Hegselmann et al., 2023) demonstrating few-shot performance on classification and regression by serializing tabular inputs for models like GPT-3 (Brown et al., 2020), GPT-J (Wang and Komatsuzaki, 2021), and T0 (Sanh et al., 2021). Techniques like LUNA (Han et al., 2022) further improve numerical reasoning in transformer models by introducing numerical embeddings into BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). LLMs have also been applied in biomedical (Yang et al., 2023; Gao et al., 2024), financial (Ma et al., 2025; Zhu et al., 2024), and mathematical domains (Schwartz et al., 2024; Lee et al., 2023). In disease analysis, LLMs have supported cell-type-specific (CTS) annotation and biomarker discovery (Omar et al., 2024; Giuffrè et al., 2024; Jagadeesh et al., 2022), with tools like scInterpreter (Li et al., 2024) and Single-Cell Omics Arena (Liu et al., 2024) leveraging pretrained models to annotate scRNA-seq data. Others have proposed explainable LLM frameworks for tracking transcriptional changes (Elsborg and Salvatore, 2023). While prior work uses LLMs for either raw prediction or annotation, our approach uniquely combines a Bayesian CTS deconvolution model (GP-unmix), regulatory priors (eQTLs (Nica and Dermitzakis, 2013)), and LLMs as faithful interpretability modules grounded in deep learning outputs, addressing both prediction accuracy and clinical transparency.

3 GP-unmix: Bayesian Deconvolution for CTS Expression

Estimating gene-level cell-type-specific (CTS) expression from bulk RNA-seq is vital for discovering disease-relevant regulatory signals. Traditional methods such as TCA and bMIND (Xu et al., 2025; Torroja and Sanchez-Cabo, 2019) primarily estimate cell type proportions and fail to model gene-level uncertainty or cross-sample biological variability. In contrast, our goal is to recover full CTS expression matrices across samples, enabling both interpretability and predictive utility.

3.1 Model Overview

We introduce **GP-unmix**, a novel Bayesian framework for cell-type-specific (CTS) gene expression deconvolution from bulk RNA-seq data. Unlike prior methods that focus solely on cell type proportions, GP-unmix recovers full gene-level CTS profiles using single-cell references and operates without ground-truth supervision. Its key innovations include a two-stage posterior refinement process that adapts reference-derived priors to target datasets, and a pretraining strategy that identifies reliably inferable gene–cell-type pairs. Additionally, GP-unmix incorporates a tripartite gene selection mechanism to improve identifiability and downstream biological resolution. This strategy combines (i) literature-curated markers (e.g., *SLC6A12* for microglia), (ii) cross-modal consensus between sc/snRNA-seq and pseudobulk datasets ($p_{\text{FDR}} < 0.01, |\log_2 \text{FC}| > 1$), and (iii) Seurat-based filtering of the top 200 differentially expressed genes ($p < 0.05$, ranked by $\log_2 \text{FC}$). Integrated into both stages of the generative model, it guides the initialization of priors in the reference-informed phase and supports the refinement of reliably inferable gene–cell-type pairs during posterior updating. This selection strategy improves deconvolution accuracy by 37–54% compared to unfiltered gene sets. Collectively, these design choices yield significant gains over existing methods, enabling robust CTS recovery across tissues, species, and applications. The generative model consists of two stages:

Reference-Informed Inference: In the first phase, single-cell/nuclei RNA-seq (sc/snRNA-seq) reference data provide CTS expression priors modeled as multivariate normal distributions: $Z_{ij} \sim \mathcal{N}(\mu_j, \Sigma_j)$, where μ_j represents the mean expression vector and Σ_j the gene-gene covariance matrix for cell type j. Bulk RNA-seq expression profiles X_{ij} are decomposed as: $X_{ij} = w_i' Z_{ij} + \Gamma_j' C_i^{(1)} + w_i' B_j C_i^{(2)} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ where w_i denotes cell-type proportions, $C_i^{(1)}$ adjusts for bulk-level technical confounders (e.g., batch effects), and $C_i^{(2)}$ models CTS biases.

Data-Adaptive Refinement: The second phase introduces dynamic Bayesian updating to address reference-target discrepancies. Priors are refined using posterior estimates from the first phase: $\mu_j^{(2)} \sim \mathcal{N}(\widehat{\mu_j^{(1)}}, 0.5 I_k), \Sigma_j^{(2)} \sim \text{InvWishart}(\widehat{\Sigma_j^{(1)}}, v_0 + v_1)$ where $\widehat{\mu_j^{(1)}}$ and $\widehat{\Sigma_j^{(1)}}$ are posterior means and covariances, respectively. Markov Chain Monte Carlo (MCMC) sampling via the *MCMCglmm* package ensures robust parameter estimation, with convergence assessed using Gelman-Rubin diagnostics ($\hat{R} < 1.05$).

3.2 Results: Benchmarking and Case Study

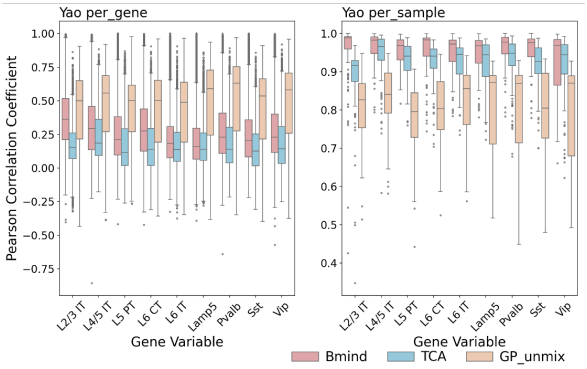

Figure 2: Pearson correlation (per gene) across methods on Yao dataset. GP-unmix outperforms TCA and bMIND in gene-level CTS recovery.

On simulated mouse cortex data, GP-unmix outperforms TCA and bMIND by **up to 66.1%** in gene-level Pearson correlation with ground-truth CTS expression. In the human brain (Yao dataset), median PCC reaches 0.82 for microglia and 0.78 for astrocytes (Figure 3). Cross-species validation shows robust performance (e.g., L5 PT neuron, PCC = 0.78 vs. 0.67 for bMIND). In PBMCs, natural killer cell estimates correlate strongly with flow cytometry ($r = 0.71$), exceeding baselines by 105.1%. Applied to the ROSMAP Alzheimer's dataset, the resulting CTS profiles enable downstream eQTL mapping and differential expression analyses, revealing astrocyte-linked genes enriched for UDP-glucosyltransferase activity—a pathway implicated in neurodegeneration. These results underscore the power of eQTL analysis for uncovering disease-relevant mechanisms and generating biologically grounded hypotheses. Additional results are presented in the Appendix A.

3.3 LLM-Based Classification and Interpretation

We explore whether LLMs can support disease classification when provided with structured features derived from GP-unmix and eQTL-informed representations (Gusev et al., 2016). Unlike traditional models, LLMs can incorporate biomedical knowledge expressed in text, making them suitable for settings where domain priors are critical.

Prompting Strategies. We compare three prompting strategies of increasing structure and domain integration: *Direct Reasoning:* A few-shot prompt of raw examples followed by a test instance and direct class prediction. *Step-by-Step Reasoning:* The model first summarizes feature distributions by class and then reasons through classification. *Step-by-Step + Domain Knowledge:* We embed definitions and biological context for each feature into the prompt (Yu et al., 2025), aligning statistical signals with known disease mechanisms. Examples of these structured prompts are provided in Appendix B and D.

Experimental Setup. We evaluate LLMs in binary Alzheimer's Disease (AD) classification using a clinical dataset with 28 features (biomarkers, vitals, and genetic markers). Two training regimes are considered: low-data (50 samples) and full-data (100 samples). For comparison, we use a two-layer MLP baseline trained on the same features. All models are evaluated on a stratified test set of 60 held-out patients.

Results. Table 1 summarizes the performance across prompting strategies. While the LLM underperforms the MLP in direct reasoning, the structured prompt with domain knowledge (**LLM+Domain**) achieves 90% accuracy in the low-data regime, surpassing the MLP. This demonstrates the value of scaffolded prompting in medical applications where training data is scarce.

3.4 LLM as Interpretability Plug-in

LLMs are effective at generating fluent, human-readable explanations, making them valuable tools for clinical interpretability. While LLMs can sometimes outperform DL models in low-data settings, we retain the DL predictor for its efficiency, stability, and scalability. Instead of serving as classifiers, LLMs act as post-hoc plug-ins to verbalize model outputs in a faithful and accessible manner. To operationalize this interpretability framework, we condition the LLM on structured outputs from the DL model and design prompts that elicit consistent, audience-specific explanations.

Each generation is conditioned on: (a) the DL-predicted probability $p(\mathrm{AD})$, and (b) the top-5 most influential features, identified via Integrated Gradients and annotated with values, attributions, and clinical reference ranges. The LLM outputs: (1) a binary decision aligned with $p(\mathrm{AD})$, (2) a rationale grounded in feature-level evidence, and (3) one or two recommended next steps with justification. To serve both clinicians and patients, we generate two versions of each report: a clinician-facing version using diagnostic terminology, and a layperson-facing version prioritizing plain language and actionable suggestions. This modular plug-in design improves transparency and usability without altering the model's underlying behavior. Prompt templates and output examples are provided in Appendix E.

Method	Train Size = 100		Train Size = 50	
	ACC	F1	ACC	F1
MLP (DL)	**0.88**	**0.86**	0.87	0.88
LLM-Direct	0.48	0.43	0.50	0.49
LLM-Step	0.70	0.62	0.77	0.75
LLM+Domain	0.74	0.69	**0.90**	**0.89**

Table 1: Accuracy and F1 across prompting variants and a deep learning baseline under two training regimes.

3.5 Human Evaluation of Interpretability Outputs

We evaluated 60 LLM-generated reports across five criteria: correctness, rationale completeness, clarity of recommendations, and stylistic appropriateness. All outputs matched DL predictions and true labels (100% alignment). Clinician-facing reports were judged 100% actionable; 88% of layperson-facing reports were similarly rated. Style was appropriate in 88% of cases, with layperson versions considered more readable. Appendix F provides annotated examples. These results confirm that LLMs, when guided by structured and audience-specific prompts, can produce faithful and interpretable explanations to support clinical decision-making.

4 Conclusion

We propose DIAGNOLLM, a unified framework that integrates Bayesian deconvolution, regulatory genomics, and LLMs for interpretable disease diagnosis. Our GP-unmix model enables robust cell-type-specific expression recovery, outperforming prior methods. Combined with eQTL-informed deep learning and LLM-based explanation, the framework delivers both accurate predictions and faithful, audience-specific reports, bridging statistical modeling with clinical transparency. Future extensions will explore broader disease applications and integration of additional omics modalities to enhance generalizability and translational impact.

5 Limitations

While DIAGNOLLM demonstrates promising results in interpretable disease diagnosis, several limitations remain. First, the performance of GP-unmix depends on the availability and quality of single-cell reference datasets, which may not be uniformly available across all tissues or disease contexts. Second, although the LLM-generated reports align well with model outputs, their clinical validity has only been assessed via preliminary human evaluation, and more extensive user studies involving clinicians and patients are necessary. Third, our current evaluation focuses on Alzheimer's Disease; generalizability to other diseases or multi-label diagnostic settings remains to be validated. Finally, while we incorporate eQTL-derived features, other regulatory modalities (e.g., epigenomic or proteomic data) are not yet included, which could limit biological completeness in certain use cases.

6 Ethics Statement

This study aims to enhance the interpretability and transparency of machine learning models in clinical genomics, with the goal of supporting informed decision-making rather than replacing medical expertise. All datasets used in this work are publicly available and de-identified, ensuring no personally identifiable information was accessed or processed. While the proposed framework generates audience-specific diagnostic reports, it is intended as a research tool and not for direct clinical deployment without expert oversight. We acknowledge that language models may inherit biases from training data and emphasize the need for careful validation in diverse patient populations. Future work will involve collaboration with medical professionals to ensure responsible, equitable, and context-sensitive use of these technologies in clinical practice.

References

Jessica Blumenfeld, Oscar Yip, Min Joo Kim, and Yadong Huang. 2024. Cell type-specific roles of apoe4 in alzheimer disease. *Nature Reviews Neuroscience*, 25(2):91–110.

Matthew Brendel, Chang Su, Zilong Bai, Hao Zhang, Olivier Elemento, and Fei Wang. 2022. Application of deep learning on single-cell rna sequencing data analysis: a review. *Genomics, proteomics & bioinformatics*, 20(5):814–835.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. 2022. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784.

Jonas Elsborg and Marco Salvatore. 2023. Using llm models and explainable ml to analyse biomarkers at single cell level for improved understanding of diseases. *bioRxiv*, pages 2023–08.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Matthew Churpek, and Majid Afshar. 2024. When raw data prevails: Are large language model embeddings effective in numerical data representation for medical machine learning applications? *arXiv preprint arXiv:2408.11854*.

Mauro Giuffrè, Simone Kresevic, Nicola Pugliese, Kisung You, and Dennis L Shung. 2024. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. *Liver International*, 44(9):2114–2124.

Alexander Gusev, Alvin Ko, Huwenbo Shi, Gaurav Bhatia, Wei Chung, Brenda W. Penninx, Rick Jansen, Eco J. C. de Geus, Dorret I. Boomsma, Margaret J. Wright, Nicholas G. Martin, Patrick F. Sullivan, Eva Nikkola, Naomi R. Wray, Jordan W. Smoller, Kenneth S. Kendler, Jonathan Flint, Kathryn Roeder, Shaun Purcell, and 3 others. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252.

Hongwei Han, Jialiang Xu, Mengyu Zhou, Yijia Shao, Shi Han, and Dongmei Zhang. 2022. Luna: language understanding with number augmentations on transformers via number plugins and pre-training. *arXiv preprint arXiv:2212.02691*.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Karthik A Jagadeesh, Kushal K Dey, Daniel T Montoro, Rahul Mohan, Steven Gazal, Jesse M Engreitz, Ramnik J Xavier, Alkes L Price, and Aviv Regev. 2022. Identifying disease-critical cell types and cellular processes by integrating single-cell rna-sequencing and human genetics. *Nature genetics*, 54(10):1479–1492.

Nayoung Lee, Kartik Sreenivasan, Jason D Lee, Kangwook Lee, and Dimitris Papailiopoulos. 2023. Teaching arithmetic to small transformers. *arXiv preprint arXiv:2307.03381*.

Cong Li, Meng Xiao, Pengfei Wang, Guihai Feng, Xin Li, and Yuanchun Zhou. 2024. scinterpreter: Training large language models to interpret scrna-seq data for cell type annotation. *arXiv preprint arXiv:2402.12405*.

Junhao Liu, Siwei Xu, Lei Zhang, and Jing Zhang. 2024. Single-cell omics arena: A benchmark study for large language models on cell type annotation using single-cell data. *Preprint*, arXiv:2412.02915.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tianmi Ma, Jiawei Du, Wenxin Huang, Wenjie Wang, Liang Xie, Xian Zhong, and Joey Tianyi Zhou. 2025. Llm knows geometry better than algebra: Numerical understanding of llm-based agents in a trading arena. *arXiv preprint arXiv:2502.17967*.

Heini M Natri, Christina B Del Azodi, Lance Peter, Chase J Taylor, Sagrika Chugh, Robert Kendle, Meii Chung, David K Flaherty, Brittany K Matlock, Carla L Calvi, and 1 others. 2024. Cell-type-specific and disease-associated expression quantitative trait loci in the human lung. *Nature Genetics*, 56(4):595–604.

Alexandra C Nica and Emmanouil T Dermitzakis. 2013. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362.

Mahmud Omar, Dana Brin, Benjamin Glicksberg, and Eyal Klang. 2024. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *American Journal of Infection Control*.

David T Paik, Sangkyun Cho, Lei Tian, Howard Y Chang, and Joseph C Wu. 2020. Single-cell rna sequencing in cardiovascular development, disease and medicine. *Nature Reviews Cardiology*, 17(8):457–473.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Eli Schwartz, Leshem Choshen, Joseph Shtok, Sivan Doveh, Leonid Karlinsky, and Assaf Arbelle. 2024. Numerologic: Number encoding for enhanced llms' numerical reasoning. *arXiv preprint arXiv:2404.00459*.

Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, and 1 others. 2018. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78.

Carlos Torroja and Fatima Sanchez-Cabo. 2019. Digitaldlsorter: deep-learning on scrna-seq to deconvolute gene expression data. *Frontiers in Genetics*, 10:978.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Xintian Xu, Rui Li, Ouyang Mo, Kai Liu, Justin Li, and Pei Hao. 2025. Cell-type deconvolution for bulk rna-seq data using single-cell reference: a comparative analysis and recommendation guideline. *Briefings in Bioinformatics*, 26(1):bbaf031.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. 2023. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857.

Xinyi Yu, Xianghong Hu, Xiaomeng Wan, Zhiyong Zhang, Xiang Wan, Mingxuan Cai, Tianwei Yu, and Jiashun Xiao. 2025. A unified framework for cell-type-specific eqtl prioritization by integrating bulk and scrna-seq data. *The American Journal of Human Genetics*, 112(2):332–352.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 310–318.

A Additional Experimental Result of GP-Unmix

As shown in figure 3, in human peripheral blood mononuclear cells (PBMCs), natural killer cell estimates correlated strongly with flow cytometry ($r = 0.71$), exceeding existing methods by 105.1% in PCC.

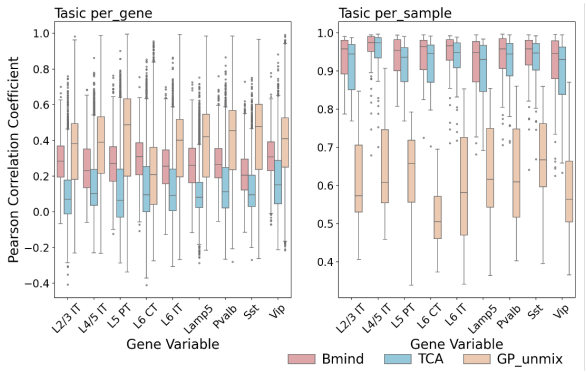

Figure 3: Pearson correlation (per gene) across methods on Tasic dataset. GP-unmix outperforms TCA and bMIND in gene-level CTS recovery.

B Step-by-Step Reasoning

We conduct our experiments using the GPT-4o-mini model to generate diagnostic reports. The model is accessed via the OpenAI API, and we adopt the default decoding settings, including a temperature of 1.0, top-p of 1.0, and a maximum token limit of 512. No additional prompt tuning or system message customization was applied unless otherwise noted, ensuring consistent and reproducible outputs across evaluations.

Step-by-Step Reasoning Task. This task aims to help the LLM distinguish between two cell types—**Oligodendrocytes** (Class A) and **OPCs** (Class B)—based on the distribution of three key statistical features: **BETA** (effect size), **SE** (standard error), and **PVAL** (p-value). Each training sample is a triplet of {BETA, SE, PVAL}, extracted from eQTL summary statistics corresponding to either Class A or B.

Instead of directly predicting the label of a test sample, the LLM is prompted to summarize the *approximate range* of each feature across the two classes, focusing on the central trend (e.g., interquartile ranges) and excluding outliers. For example, the model may infer that BETA values in Class A cluster around 0.02–0.05, while Class B shows more dispersed or weaker effects.

This two-step process—first identifying range boundaries, then comparing test instances—allows the model to reason more transparently. It also facilitates downstream interpretation and aligns well with biological intuition, as different cell types often exhibit characteristic feature distributions.

The prompt and the example output are shown in Figure 4.

C DL Baseline Configuration.

For the binary classification tasks, we implement a two-layer MLP with hidden sizes of 16 and 8. The network consists of: (a) two fully connected layers (input $\rightarrow 16 \rightarrow 8 \rightarrow 1$), (b) ReLU activations between layers, and (c) a sigmoid function at the output layer for binary prediction. We optimize the model using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy (BCE) loss.

All experiments are conducted using a test set of 100 samples, consistent across both classification tasks. For each training regime (low-data or full-data), we use either 50 or 100 training samples, respectively, randomly selected from the full dataset while ensuring class balance. The test set remains fixed across all prompting variants for fair comparison.

D Domain Knowledge Used to Train LLMs

Cell-Type Classification Task. This task further guides the LLM to distinguish between **Oligodendrocytes** (Class A) and **OPCs** (Class B), using the same three statistical features: **BETA**, **SE**, and **PVAL**. As before, each instance is a triplet extracted from eQTL summary statistics.

In addition to reasoning over feature distributions, the LLM is now provided with domain-specific biological priors describing how these features typically manifest in different cell types. For instance, Oligodendrocytes tend to exhibit stronger and more stable regulatory signals—reflected by larger BETA values, lower SE, and more significant PVALs—while OPCs often display weaker or noisier patterns.

By incorporating these biological expectations directly into the prompt, the LLM can better align statistical evidence with known cellular behaviors. This makes the classification process not only more accurate but also more consistent with expert understanding in genomics.

The annotated domain knowledge and prompt

example are shown in Figure 5. The domain knowledge is from (Yu et al., 2025).

E LLM-Based Diagnostic Interpretability

This interpretability task is designed to translate deep learning (DL) model predictions into audience-specific textual reports. The input to the LLM consists of:

- A DL-predicted probability of Alzheimer's disease (e.g., 0.10),
- The top-5 contributing clinical features (e.g., blood pressure, age), annotated with normal ranges and attribution scores.

The LLM is instructed to generate structured outputs tailored to different audiences:

- **Physician Prompt:** Focuses on clinical summary, risk justification, and suggested next steps. It uses medical terminology and diagnostic conventions.
- **Layperson Prompt:** Translates the same prediction into plain language, highlighting behavioral advice and simplifying medical reasoning for non-experts.

Each generation includes a binary judgment, explanation of contributing factors, and actionable suggestions. Examples of both prompt styles and corresponding outputs are shown in Figure 6 and Figure 7.

F Human Evaluation Protocol and Examples

To evaluate the interpretability and factual consistency of LLM-generated diagnostic reports, we conducted a manual annotation study involving one domain-expert rater and one layperson rater. Each report was assessed along four criteria. Both annotators provided informed consent prior to participation and were clearly informed about the study's purpose and how their evaluations would be used in model assessment. The domain-expert was a biological researcher affiliated with a U.S.-based academic institution, and the layperson was a graduate student with no background training. Both participants were volunteers recruited through academic networks and were not financially compensated. Their voluntary participation was based on interest in the research topic, and no sensitive or personally identifiable health information was used during annotation. Each report was assessed along four criteria:

Final Judgment: Whether the binary diagnosis aligns with the predicted probability.

Rationale: Whether the explanation refers to key contributing clinical features.

Next Steps: Whether actionable next steps are included.

Style: Whether the tone matches the intended audience (physician vs. layperson).

No payment or compensation was provided for their participation. Annotators were given clear written instructions on how to assess each report based on the criteria listed above.

Table 2 presents two annotated examples from the human evaluation—one for a physician-facing report and one for a layperson-facing report.

Sample 1: Patient ID 14 – Physician Report		
Component	**Score**	**Annotator Comment**
Final Judgment	Yes	Model probability 1.000 ⇒ likely, report says likely.
Rationale	Yes	Mentions 4 of top features: LDL, TC/HDL Ratio, Age, HDL.
Next Steps	Yes	Found actionable recommendation(s).
Style	Yes	Tone appropriate for physician.
Sample 2: Patient ID 48 – Layperson Report		
Component	**Score**	**Annotator Comment**
Final Judgment	Yes	Model probability 0.992 ⇒ likely, report says likely.
Rationale	Yes	Mentions 2 of top features: Age, BMI.
Next Steps	Yes	Found actionable recommendation(s).
Style	Yes	Tone and language suitable for general audience.

Table 2: Manual evaluation of LLM-generated outputs by audience type.

Step-by-Step Reasoning Prompt and Output Example

`Prompt:` You are a bioinformatics expert. Your task is to analyze and summarize the distribution of BETA, SE, PVAL values from two classes of eQTL samples (Class A and Class B). Please ignore rare outliers and focus only on the range where the majority of values are concentrated.
`Input:`

Class A (Oligodendrocytes):

- BETA: 0.02483, SE: 0.05101, PVAL: 0.51584
- BETA: 0.02870, SE: 0.04923, PVAL: 0.36891
- BETA: 0.02015, SE: 0.05045, PVAL: 0.74187
- ...

Class B (OPCs):

- BETA: -0.05283, SE: 0.08523, PVAL: 0.67051
- BETA: 0.04713, SE: 0.07224, PVAL: 0.80923
- BETA: -0.01930, SE: 0.07116, PVAL: 0.53891
- ...

`Output Format:` The BETA of Class A is mostly within (lower bound)–(upper bound), while the BETA of Class B tends to be within (lower bound)–(upper bound).
The SE of Class A is mostly within (lower bound)–(upper bound), while the SE of Class B tends to be within (lower bound)–(upper bound).
The PVAL of Class A is mostly within (lower bound)–(upper bound), while the PVAL of Class B tends to be within (lower bound)–(upper bound).

`Output:` **Distribution Summary of Class A and Class B:**
The BETA of Class A is mostly within **0.020–0.052**, while the BETA of Class B tends to be within **-0.054–0.048**.
The SE of Class A is mostly within **0.048–0.053**, while the SE of Class B tends to be within **0.065–0.093**.
The PVAL of Class A is mostly within **0.344–0.742**, while the PVAL of Class B tends to be within **0.536–0.810**.

Figure 4: Step-by-step reasoning prompt and LLM output showing feature-wise distribution analysis across cell types.

Example of Domain Knowledge

`Domain Knowledge (for Cell-Type Reasoning):`
BETA (Effect Size):

- Indicates direction and magnitude of gene regulation by the SNP.
- Larger absolute BETA (e.g., > 0.08 or < –0.08) suggests stronger regulation, often seen in mature cells like *Oligodendrocytes*.
- Near-zero BETA values are typical for less differentiated cells (e.g., OPCs), indicating weak or transient regulation.

SE (Standard Error):

- Reflects uncertainty in the BETA estimate.
- Smaller SE (e.g., < 0.06) implies more stable and confident effect size estimates, often in *Oligodendrocytes*.
- Larger SE is frequently observed in *OPCs* due to high variability.

PVAL (p-value):

- Indicates the statistical significance of the SNP-gene association.
- PVAL < 1×10^{-5} denotes strong association; PVAL > 0.05 suggests weak evidence.
- More significant PVALs are often found in *Oligodendrocytes*; *OPCs* tend to yield higher PVALs.

Figure 5: Domain knowledge used in cell-type classification, detailing biological interpretation of BETA (effect size), SE (uncertainty), and PVAL (statistical significance). This knowledge informs the LLM's reasoning about eQTL characteristics of Oligodendrocytes and OPCs.

Physician Prompt and Output Example

`Prompt:`

You are a clinical lab technician. Draft a concise report for physicians.
Predicted probability of Alzheimer's Disease: 0.10
Top contributing features:
Systolic_BP_mmHg = 108.25 (normal: 90-120 mmHg, attribution -0.217)
LDL_mg_dL = 111.93 (normal: <100 mg/dL, attribution -0.168)
Age = 80.37 (normal: N/A, attribution +0.147)
Homocysteine_umol_L = 16.26 (normal: 5-15 µmol/L, attribution +0.146)
Sleep_Duration = 4.96 (normal: 7-9 hours/night, attribution +0.139)
Provide:
(a) Final judgment (Possibility of disease: yes / no) → Ensure your answer matches the predicted probability provided above.
If the probability is below 0.5, say no.
If the probability is above 0.5, say yes.
(b) Plain-language rationale referencing the features above
(c) Typical next-step tests or medications
(d) Justify why those steps follow logically from the attributions and values above
Phys Report:
Clinical Lab Report
Patient ID: [Insert Patient ID]
Date: [Insert Date]
Prepared by: [Your Name], Clinical Lab Technician

`Output:`

Final Judgment: No
Plain-Language Rationale:
The predicted probability of developing Alzheimer's Disease for this patient is 0.10, which is below the threshold of 0.5. This indicates a low likelihood of the disease. The contributing features include:
Systolic Blood Pressure (108.25 mmHg): This value is within the normal range but slightly lower than optimal levels, contributing negatively to the risk.
LDL Cholesterol (111.93 mg/dL): This is above the normal range, which also contributes negatively to the risk.
Age (80.37 years): This is a significant risk factor, contributing positively to the probability of Alzheimer's.
Homocysteine (16.26 µmol/L): Elevated levels are associated with cognitive decline, contributing positively to the risk.
Sleep Duration (4.96 hours/night): This is below the recommended amount of sleep, which negatively impacts cognitive health.
While age and homocysteine levels are concerning, the overall probability remains low.
Typical Next-Step Tests or Medications:
1. Cognitive Assessment: A comprehensive cognitive evaluation to monitor any changes in cognitive function over time.
2. Lifestyle Modification Recommendations: Encourage dietary changes to manage LDL cholesterol and promote better sleep hygiene.
3. Homocysteine Level Monitoring: Regular monitoring of homocysteine levels, with potential supplementation of B vitamins if indicated.
Justification for Next Steps:
Given the low predicted probability of Alzheimer's Disease, immediate pharmacological intervention is not warranted. However, the elevated homocysteine and LDL levels, along with inadequate sleep, suggest that lifestyle interventions and monitoring are prudent. These steps aim to mitigate risk factors that could contribute to cognitive decline in the future, aligning with the patient's current health profile and risk factors identified in the report.

Figure 6: Clinician-facing prompt and output from the interpretability plug-in.

Layperson Prompt and Output Example

Prompt:

Lay Prompt:
You are a family doctor explaining lab results to a patient with no medical background. Use plain everyday language.
Predicted probability of Alzheimer's Disease: 0.10
Top contributing features:
Systolic_BP_mmHg = 108.25 (normal: 90-120 mmHg, attribution -0.217)
LDL_mg_dL = 111.93 (normal: <100 mg/dL, attribution -0.168)
Age = 80.37 (normal: N/A, attribution +0.147)
Homocysteine_umol_L = 16.26 (normal: 5-15 μmol/L, attribution +0.146)
Sleep_Duration = 4.96 (normal: 7-9 hours/night, attribution +0.139)
Explain briefly:
(a) Is the patient likely to have Alzheimer's? (yes / no) → Your answer must align with the predicted probability shown above.
If the probability is below 0.5, say no.
If the probability is above 0.5, say yes.
(b) The main reasons—in simple words
(c) One or two easy next steps the patient can take
(d) Briefly explain why these next steps make sense based on the above findings from the attributions and values

---

Output:

Lay Report:
Sure! Let's go through your lab results together.
(a) Is the patient likely to have Alzheimer's?
No. The predicted probability of having Alzheimer's Disease is 0.10, which is below 0.5. This means it's unlikely you have it.
(b) The main reasons—in simple words:
Your lab results show a few things that could be concerning. You are 80 years old, which is a risk factor for Alzheimer's. Your homocysteine level is a bit high, and you're not getting enough sleep, which can also be linked to memory issues. Your blood pressure and cholesterol levels are okay, but they are not ideal.
(c) One or two easy next steps the patient can take:
1. Improve your sleep: Aim for 7-9 hours of sleep each night.
2. Talk to your doctor about your diet: Consider foods that can help lower homocysteine levels, like leafy greens and beans.
(d) Why these next steps make sense based on the findings:
Getting enough sleep is important for brain health, and since your sleep duration is below the recommended amount, improving it could help your overall cognitive function. Lowering your homocysteine levels through diet can also be beneficial, as high levels can be linked to brain health issues. These steps can help reduce your risk factors and support your brain health moving forward.

Figure 7: Layperson-facing prompt and output from the interpretability plug-in.