

Multimodal LLM Engineer Take-Home Challenge

# Important note about submission

Please submit your completed challenge by the deadline by sending an email to [yasin.senbabaoglu@czbiohub.org](mailto:yasin.senbabaoglu@czbiohub.org) with the link to your shared document(s) or folder. **Do not submit your solution as an attachment, the CZ Biohub IT system will silently block your submission and we will not know about it.**

# Part 1: Model Proposal

# Challenge Overview

Recent advances in conversational AI have made strides in specialized fields like biology and medicine. Notable examples include:

* [**ChatNT** (Richard et al., 2024)](https://www.biorxiv.org/content/10.1101/2024.04.30.591835v2)
* [**Llava-Med** (Li et al., 2023)](https://arxiv.org/abs/2306.00890)

Additionally, models such as **L2G** (Chung et al., 2024) demonstrate the effective use of pre-trained large language models (LLMs) for genomic predictions by integrating multi-modal data.

Dr. Golgi, a cell biologist, seeks to understand the downstream effects of various gene knock-outs and their impact on cell state. Two key approaches to measure cell state include:

* **Imaging:** Dr. Golgi has acquired image data from 1000 gene knock-outs using an Optical Pooled Screen experiment, where each cell is tagged with its specific gene knock-out.
* **Transcriptomic Profiling:** A corresponding single-cell Perturb-seq experiment has been conducted on the same set of 1000 knock-outs, with approximately 500 cells captured per gene knock-out.

Given that both datasets originate from the same cell line and genetic perturbations, there is a unique opportunity to train a model that maps image features to transcriptomic states. This capability would allow the prediction of a cell's transcriptomic profile from its image.

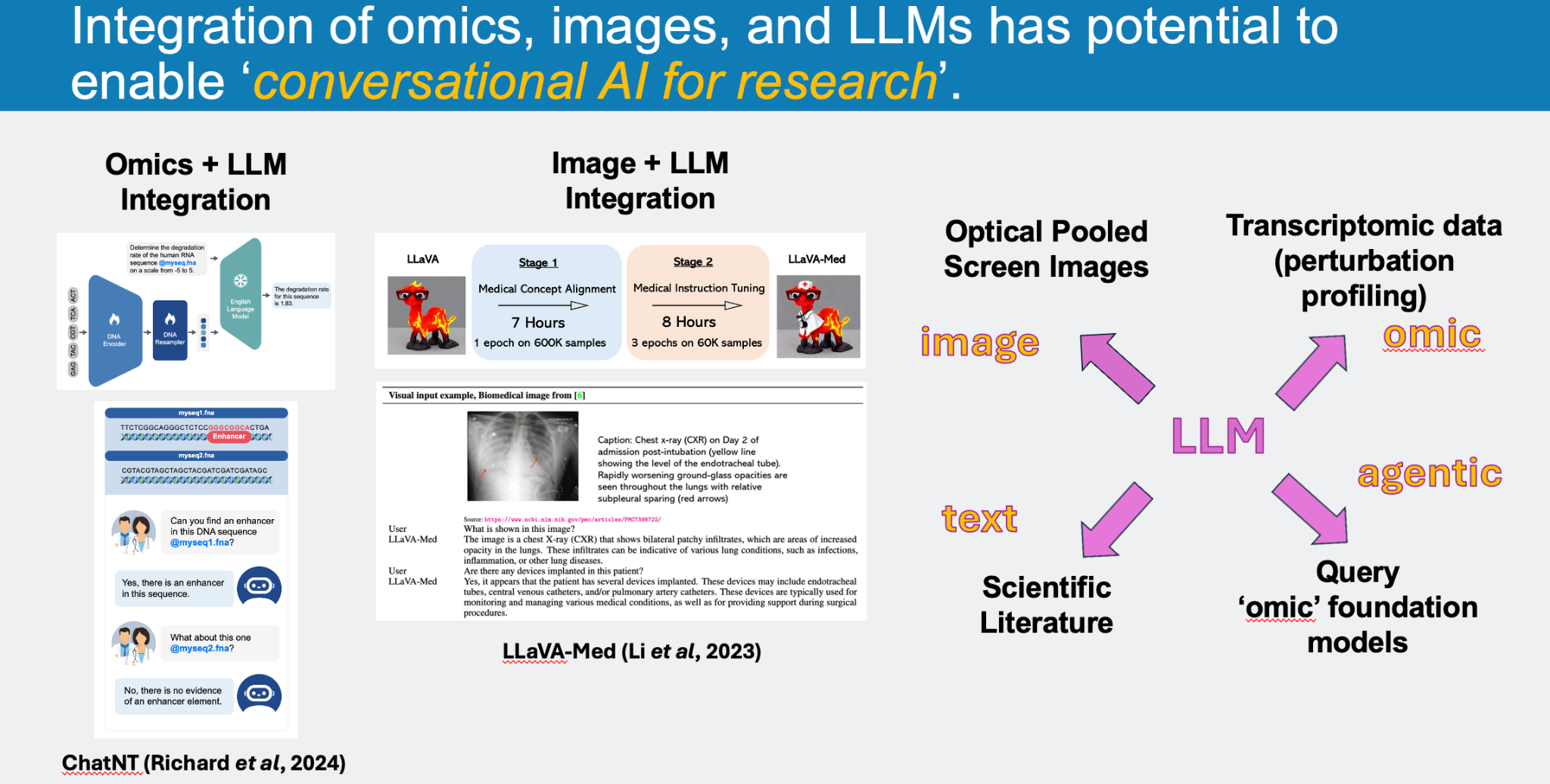
### **Objective**

Propose the architecture and training strategy for a conversational tool that can seamlessly integrate and interpret both image and transcriptomic data (see figure below). The tool should:

* **Input Flexibility:** Accept text prompts, images, or transcriptomic data.
* **Contextual Intelligence:** Leverage its training on scientific literature to provide contextually rich responses.
* **Integration with Existing Models:** Function as part of a broader system, potentially incorporating insights from transcriptomic foundation models (e.g., Geneformer) to assess the expected effects of gene knock-outs across various cell types.

Your model / system should enable Dr. Golgi to query the data, literature, or an existing foundation model to get comprehensive information on how a single gene knockout should affect cell state. No coding necessary, but be detailed in describing:

1. Additional data you need
2. Model architecture
3. Training strategy
4. How agents will complement your model



Note: Feel free to think flexibly about the design. The figure shows LLMs at the center of image & omic data, scientific literature, and existing foundation models. However, this is not a model specification, you can choose a different design. The evaluation criteria will be your creativity, level of detail, and implementation feasibility of your system.

# Part 2: Coding Challenge

# Challenge Overview

**Problem Statement:**

Develop a multimodal learning system integrating textual embeddings (from cell type labels) and single cell transcriptome embeddings to predict donor identifiers.

We would like to answer the question whether single cell gene expression data (aligned with information from cell type labels) have power to predict donor identifiers. Since the sex variable could be a confounder for predicting donors, your model should use an adversarial loss to control for the sex variable. Run the code on a GPU. To develop the code, feel free to leverage online resources and LLMs but cite your resources appropriately. As current LLMs are powerful enough to generate correct or almost correct code for a given task, demonstrate your level of understanding by **explaining all of your design & parameter choices in a very detailed way**. Also pay attention to modularizing training components to enable efficient parallel processing.

**Dataset:**

Employ the **Geneformer** foundation model’s single cell embeddings for transcriptomic data (512-dimensional embedding vectors). The raw omics data should not be used directly in the model. Both the embeddings and associated metadata (e.g. cell type, donor id, sex) can be obtained from the CZ CELLxGENE Discover Census platform. Here are two helpful links

* [Access CELLxGENE collaboration embeddings](https://chanzuckerberg.github.io/cellxgene-census/notebooks/api_demo/census_access_maintained_embeddings.html)
* [Access CLLxGENE-hosted embeddings](https://chanzuckerberg.github.io/cellxgene-census/notebooks/api_demo/census_embedding.html)

For this challenge, we are interested in human RNA measurements from the central nervous system (tissue\_general set to 'central nervous system’). Obtain the cell\_type, donor\_id and sex information as well as ‘geneformer’ embeddings for each cell.

* **Deliverable: Report the number of cells in the data and generate a UMAP plot of cells using Geneformer embeddings (See the Deliverables section)**

The second modality (i.e. textual embeddings) will come from cell type labels. Obtain tokens and embeddings for cell type labels utilizing a **biology-focused pretrained language model** in Huggingface. For a given cell, textual embeddings should be obtained from the first ([CLS]) token embedding.

In the single cell data, use the ‘sex’ variable as the confounding variable (categorical), and ‘donor\_id’ as the target variable for prediction (categorical).

**Model development**

1. **Multimodal Integration:**

Integrate textual data and Geneformer-processed omics embeddings using advanced techniques like cross-attention and Perceiver Resampler modules. All cross-attention modules should use multi-head attention. **Justify your choice of key, query, and value in cross-attention modules.**

1. **Multimodal Pretraining:**

Pretrain the Perceiver Resampler modules (text and omics encoders) using a contrastive learning task to extract meaningful representations.

1. **Adversarial Training:**

Include an adversarial loss to mitigate the influence of the confounding variable and improve robustness.

**Hyperparameter Optimization:**

Use Bayesian optimization (e.g. via Optuna's TPESampler or another tool) to optimize the following hyperparameters:

1. Learning rate
2. Dropout probability
3. Number of latent vectors in the Perceiver Resampler
4. Latent dimensions in the Perceiver Resampler (the dimension/size of each latent vector)
5. Weight for adversarial loss

**Regularization:**

Incorporate dropout and other regularization techniques to improve model generalization.

**Learning Rate Scheduling:**

Implement a cosine annealing learning rate scheduler with an optional warmup phase.

**Evaluation Metric:**

Accuracy of donor id predictions

**Bonus: Explainability:**

Provide methods to interpret the model's behavior, such as visualizing cross-attention weights as heatmaps.

# Deliverables

1. **Figures:**
   * Information on number of cells in the data, and a UMAP plot of cells using Geneformer embeddings
   * Main task loss and adversarial loss curves across epochs (for the model with best parameters)
   * Prediction accuracy vs epoch (for the model with best parameters)
   * **Bonus:** (2-dimensional) Heatmap for cross-attention weights from a sample slice of the attention weight tensor.
2. A **Python script or Jupyter Notebook** containing the code for your solution. Please ensure that your code executes without errors (include instructions for running the code if necessary). As current LLMs will greatly help you generate the code for this challenge, be **highly detailed** **about explaining all of your design & parameter choices (include these explanations as comments in your code)**. As a rule of thumb, if you think there is more than one ‘good’ way to implement something, provide an explanation as to why you chose your particular approach.
3. **Results Summary:** A clear and concise summary of your results, including the performance of your model based on the evaluation metric. Was donor ID prediction successful?
   * **Bonus question:** Did the pretraining of the Perceiver Resampler help or hurt model performance? Explain with evidence.
   * **Bonus question:** Did the information from cell type labels (alignment with transcriptomic embeddings) help or hurt model performance? In other words, was the transcriptomic data sufficient to achieve the same level of accuracy? Explain with evidence.

# Evaluation Criteria

Your submission will be evaluated based on the following criteria:

* **Correctness:** Does the code meet the requirements and execute without errors?
* **Code Quality:** Is the code clean, and well-documented?
* **Justification of design/parameter choices:** Are the design choices and any hard-coded parameter values sufficiently justified?
* **Efficiency:** Is the model modular and computationally efficient? Would it enable parallel processing?
* **Explainability:** Are the provided explainability methods clear and insightful?
* **Optimization:** Are hyperparameters well-tuned, and is the use of Bayesian optimization effective?

# Submission

Please submit your completed challenge by the deadline by sending an email to [yasin.senbabaoglu@czbiohub.org](mailto:yasin.senbabaoglu@czbiohub.org) with the link to your shared document(s) or folder. **Do not submit your solution as an attachment, the CZ Biohub IT system will silently block your submission and we will not know about it.**

# Additional Notes

* You are encouraged to use any relevant libraries or tools.
* Feel free to ask clarifying questions if you need any further information.
* We value creativity and innovation in your approach.

We look forward to reviewing your submission. Good luck!