AI-Driven Formulation Development for Pharmaceutical Applications

Client: Okhee Yoo

Team: Nisha Jha(23945457), Sarah Pinelli(23419054), Shreya Kaushal Patel(24690749), Sungbae Ji(24619726), Philipp Koluzanov(24069852)

AIM

To improve formulation efficiency, reduce costs, and improve treatment adherence in children, we will create an LLM-based recommendation system that uses literature-based human taste data and fine-tuned language models to select the best taste-masking excipient for paediatric antimicrobial formulations.

BACKGROUND AND VALUE PROPOSITION

Paediatric antibacterial treatment faces the combined problem of ensuring medication efficacy while maintaining patient compliance. For children, the unpleasant taste of antibiotics, antifungals, and antibacterials is a major barrier to finishing treatment. To address this, formulations frequently include taste-masking excipients which are inactive substances that improve taste while maintaining efficacy. However, selecting the best excipient remains essentially a trial-and-error procedure supported by literature reviews and professional judgment.

By creating a data-driven recommendation engine that recommends the best taste-masking excipient for a certain approved medicinal molecule, the proposed research aims to revolutionise the excipient selection process. The project intends to produce a tool that can quickly and accurately recommend excipients. This will be achieved by performing literature research to gather human taste-related data into a structured dataset and by investigating the fine-tuning of a large language model (LLM).

The **value proposition** of this work is twofold:

1. **For pharmaceutical formulators** – An intelligent, evidence-based decision support system that boosts formulation efficiency, promotes taste acceptability, and lowers R&D costs.
2. **For paediatric patients and caregivers** – For paediatric patients and caregivers, higher pharmaceutical taste leads to increased treatment adherence and, ultimately, better health outcomes.

DELIVERABLES AND TIMELINE

To meet the project's objectives and client requirements, the following key deliverables were identified:

- A fine-tuned Large Language Model (LLM) that allows the recommendation of the best taste-masking excipient associated with an input antimicrobial medicine molecule. It will consist of the following features:
    - Interactive inference interface (10/09 – 24/09) – user interface to input prompts and receive model outputs instantly.

    - Optional advanced access (2/09 – 25/09) – Availability of model code and fine-tuning scripts for technical users who wish to retrain, adapt, or extend the model.
- Antimicrobial medicines and taste-masking excipients database (11/08 – 2/09) – A structured file compiled from human-trial literature research for model training.
- LLM model documentation (29/09 – 10/10) – Step-by-step documentation how to access and use the model, tailored for non-technical users.
- Multimedia demonstration video (1/10 – 10/10) – Multimedia video showing how to access and use the model following the step-by-step guide.

The project will be completed over a 12-week period, beginning in Week 4 (11/08) after the first client meeting and concluding in Week 12 (10/10). The key phases and their approximate durations are:

1. Project Initiation and Planning (Weeks 4-5): Finalising project requirements and scope with client and creating collaborative tools.
2. Research and Data Preparation (Weeks 4-7): Gathering and processing relevant literature and compiling extracted data into a dataset.
3. LLM Model Fine-Tuning (Weeks 7-12): Training, Validating and testing the LLM model on created dataset.
4. Final Deliverables (Week 12): Handover of final output and model demonstration to client.

For a detailed breakdown of tasks and their dependencies as a Gantt Chart, see Appendix.

COSTS

The project is expected to incur no costs, as it will leverage entirely free and open-source resources. Input data will be sourced primarily from literature research articles and several publicly available datasets, which will be cleaned and normalised in preparation for fine-tuning. Model development will focus on open-source LLMs, such as BioGPT and PubMedBERT, both available for free download. Fine-tuning will be conducted using Google Collab and local machines to efficiently manage resource-intensive tasks without incurring subscription fees. As the project is designed exclusively for the client's research group and not for public deployment, there is no need for a cloud-based web service, thereby eliminating any potential hosting or deployment costs.

METHODS

Our project aims to advance its detailed methodology across five key tasks: 1) exploring how to best leverage available datasets, 2) constructing the training dataset for LLM fine-tuning, 3) selecting a base LLM model and ensuring efficient training, 4) evaluating and enhancing model performance, and 5) determining the deployment and utilisation strategy for the final model.

Task 1: Exploring How to Best Leverage Available Datasets

We will begin by thoroughly exploring publicly available datasets for formulation development, such as the US FDA's DailyMed and the Inactive Ingredients Database (IID). These provide extensive details on ingredients, dosage forms, and safe usage levels. Crucially, as this project aims to improve paediatric antimicrobial formulations by taste-masking, relevant human taste

data is essential. We will obtain this from academic literature and, if necessary, from the client. An Exploratory Data Analysis (EDA) will then assess data quality, distributions, and interconnections to maximise utility and identify gaps for our specific needs.

Task 2: Constructing the Training Dataset for LLM Fine-Tuning

For LLM fine-tuning, we will use Instruction (Prompt) – Response (Completion) pairs to guide the model, expecting a question-answering format where input conditions yield new formulation outputs. For example, a prompt like "Tell me common excipient combinations for a [Dosage Form] drug containing [API Name]" will generate a response detailing product name, API, and excipients with concentrations. We will start by comprehensively collecting DailyMed data (approx. 44GB), which will undergo initial cleaning and preprocessing to standardise formats and remove irrelevant information. The preprocessing stage will include duplicate removal, text normalisation, and mapping ingredient names to standardised vocabularies such as UNII codes to improve data consistency.

Task 3: Selecting a Base LLM Model and Ensuring Efficient Training

Instead of generic LLMs like Llama 2, We will leverage open-source biomedical LLMs such as BioGPT or PubMedBERT, which are pre-trained on biomedical literature. These models offer deeper pharmaceutical language understanding, maximising transfer learning and requiring less data for superior fine-tuning performance. To optimise computational resources, We will use Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA, significantly reducing GPU memory and training time without performance compromise.

Task 4: Evaluating and Enhancing Model Performance

Evaluating our fine-tuned LLM is a crucial, ongoing process. We will use a validation dataset (10-20% of constructed data) for quantitative assessment, focusing on semantic similarity metrics and checking adherence to pharmaceutical constraints. Beyond numbers, qualitative human assessment by domain experts is essential to review generated excipient combinations for plausibility and practical applicability. This expert feedback, combined with systematic error analysis and integrated user feedback, will guide our iterative enhancement strategy to mitigate "hallucinations" and continuously improve the model's reliability and accuracy.

Task 5: Determining the Deployment and Utilisation Strategy for the Final Model

Initially, we will provide the client with a shared Jupyter Notebook environment to enable direct prompt–response experimentation and immediate feedback, allowing for rapid iteration. Once the model's performance stabilises, we will build a standalone terminal-based system tailored for the client's local computing environment. This project focuses solely on the client's research group, so we have not considered a cloud-based web service for a broad user base, instead prioritising maximising model performance. However, if implementing a web-based user interface feedback mechanism is necessary, we will review and proceed with its development for continuous improvement.

APPENDIX

Graph A1

Project Gantt Chart of Team Tasks and Deliverables Timeline