Latent-Aligned Generative Modeling for High-Impact Financial Risk Scenarios (Rare tabular events)

Project Supervisor: He Zhao, Zang Tao

Motivation:

Financial institute increasingly rely on machine learning models to assess *creditworthiness*, *detect anomalies*, and predict *default risk*. However, these models are often trained on real-world datasets that exhibit strong **distributional biases** — for example, **very few high-net-worth clients** default, making such edge cases severely underrepresented.

This scarcity leads to:

- Poor model generalization on critical, high-impact edge scenarios (e.g., VIP clients who unexpectedly default) and difficulty in model training due to high imbalance.
- Insufficient robustness and explainability in downstream decision systems.
- A lack of **counterfactual analysis tools** to answer "what-if" questions in high-stakes financial scenarios, such as
    - "What would a credit profile look like for a high-income client who default?".
    - In response, we would like to have an approach that can generate a transaction table whose embedding fails in the clustering of "default" but having "high-income" features.

Can Embedding Language Models (ELMs) solve this problem?

ELM (e.g., EAGLE from NeurIPS 2024) recently has shown the power to teach LLMs to generate textual output that can possess a desired embedding zone, using an embedding-based utility function (or reward function). This new finding inspires us to propose a framework that

1. **Identifies underrepresented latent clusters** in tabular embeddings of financial records (e.g., defaulting VIP clients),
2. **Generates realistic counterfactual examples** in tabular data format with embedding space-based utility (reward function),
3. **Guides models and analysts** toward better understanding and calibration in the data-sparse regions of the financial landscape.

Related Work:

- **Embedding-Aligned Language Models (NeurIPS 2024)**: Introduced EAGLE, a reinforcement learning agent that edits LLM outputs to match target embedding zones derived from behavioral data. This internship would mainly rely on the EAGLE framework to achieve scarce data synthesis and counterfactual example generation.
- **Tabular Representation Learning**: Techniques like TabNet, FT-Transformer, and TabPFN have shown strong performance in capturing structure in financial data. This internship can use the encoder of TabPFN to project EAGLE generated outputs into embedding space where a designed utility function would provide reward / guidance in embedding space.
- **Counterfactual Explanation** (Wachter et al. 2017, Karimi et al. 2021): Generate alternate data instances to probe model reasoning but often lack control over global embedding properties.
- **Data Augmentation in Tabular Domains**: Still underdeveloped compared to NLP and vision; recent works explore GANs, SMOTE, and few-shot transformers, but often fail to target semantically meaningful gaps in data distribution.
- **Outlier Risk in Finance**: Studies in explainable AI and fairness have emphasized the need for models that remain robust on rare but impactful events, such as strategic default or profile shift in wealthy clients (e.g., “The HNWI Default Paradox”).

Proposed Approach:

Step 1: Latent Space Construction

- Train a **tabular foundation model** (e.g. FT-Transformer or TabLLM encoder) on a financial dataset (e.g., credit bureau or loan portfolio).
- Use the learned embeddings to map all data points into a latent semantic space.
- Use clustering or density estimation (e.g., UMAP + DBSCAN, KDE) to identify **sparse regions** in the latent space — these represent **underrepresented or risky profile zones**.

Step 2: Target Cluster Identification

- Define high-interest clusters, such as:
  - High-income + low default likelihood (normal)
  - High-income + **actual default** (rare)
  - Low-income + non-default (unexpected resilience)
- Tag these clusters using supervised labels or latent behavioral signals for the utility function.

Step 3: Targeted Profile Generation

- Use an embedding-aligned generator (i.e., EAGLE) to:
    - Start from existing data points in dense regions.
    - Iteratively modify tabular features (e.g., utilization ratio, loan mix, age of accounts) to reach the **target embedding zone**.
    - Ensure **semantic and business-rule validity** using learned constraints or domain-informed edit graphs.

Step 4: Applications

- **Model stress testing**: Inject synthetic rare cases into training or evaluation sets to test model robustness.
- **Scenario planning**: Equip risk teams with rich hypothetical profiles to improve early warning systems.
- **Counterfactual explanation**: Provide insights into how slight changes in profile might shift risk prediction.

Potential Dataset

- Taiwan Credit Default (UCI)
- German Credit Data
- Bank Marketing Dataset
- (If extendable to time series data) NAB / SKAB anomaly benchmarks for rare-event modelling for time series data.

Potential Dataset Proposed evaluation metric

- Rare cluster generalization – AUC, Recall, Calibration in rare zone
- Latent diversity – T-SNE / PCA plot, entropy, KL-divergence
- Enhanced foundation model embedding – performance on downstream fine-tuning
- Human expert validation – Human rating on plausibility, usefulness, helpfulness.

Deliverables:

The expected outcome of the project is a submission of the study to a main conference like ICML 2026.

- Before the internship begins: Propose potential solutions to the open challenges and settle down on a workable formulation of the project.
- Month 1: The intern is supposed to setup the candidate opensource LLM models, investigate tabular data benchmarks and foundation model embeddings. As follows, the intern should re-implement a minimal working version of EAGLE framework on any tabular dataset and design an embedding-space utility function (as reward function) that can steer LLMs to generate target outputs.
- Month 2: The intern is supposed to set up the entire pipeline from Step 1 to Step 3, get initial experiment results and set-up the evaluation pipelines.
- Month 3: We expect the intern to continue improving the entire approach and compare against baseline models.
- Month 4: The intern will be devoted to wrapping up the experiments, writing the paper for submission, and cleaning up the code base.