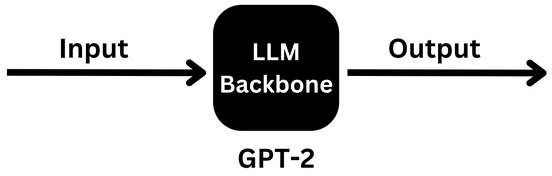
Improving Accessibility of Concept Bottleneck Layers for Scalable, Accurate, Interpretable Models

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Gabriel Cha (gcha), Steven Luong (sxluong) Mentor: Lily Weng (lweng@ucsd.edu) github.com/gabrielchasukjin/cbm-gui-frontend

Background

Large Language Models (**LLMs**) are powerful but often lack transparency. Concept Bottleneck Layers (**CBLs**) improve interpretability by linking predictions to human-understandable concepts.

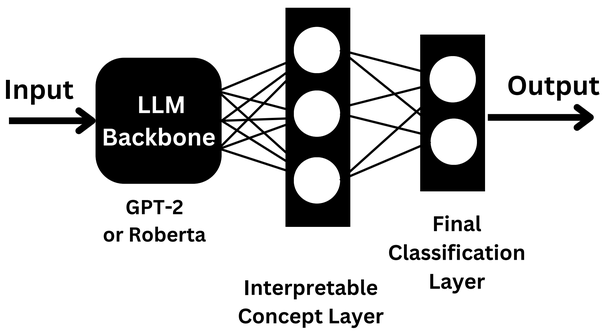


Goals & Objective

The goal of this project is to create an interface that integrates CBLs into LLMs:

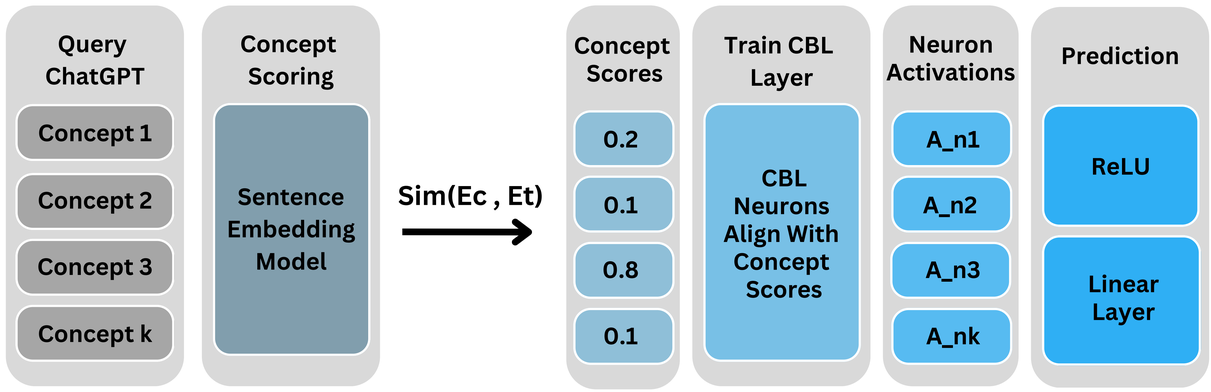
- **Fine-tuning** CBLs within pre-trained LLMs.
- **Visualizing** highly activated concepts during model inference.
- **Pruning** biased or low-impact concepts to enhance fairness.
- **Comparing** model iterations to ensure transparency and reproducibility.

Our project makes AI workflows intuitive and accessible for all users with a simple, one-click solution:



Methodology

The Concept Bottleneck Layer (CBL) generation process maps neuron activations in a pretrained LLM to human-understandable concepts by scoring and adjusting concept relevance.



Graphical Interface

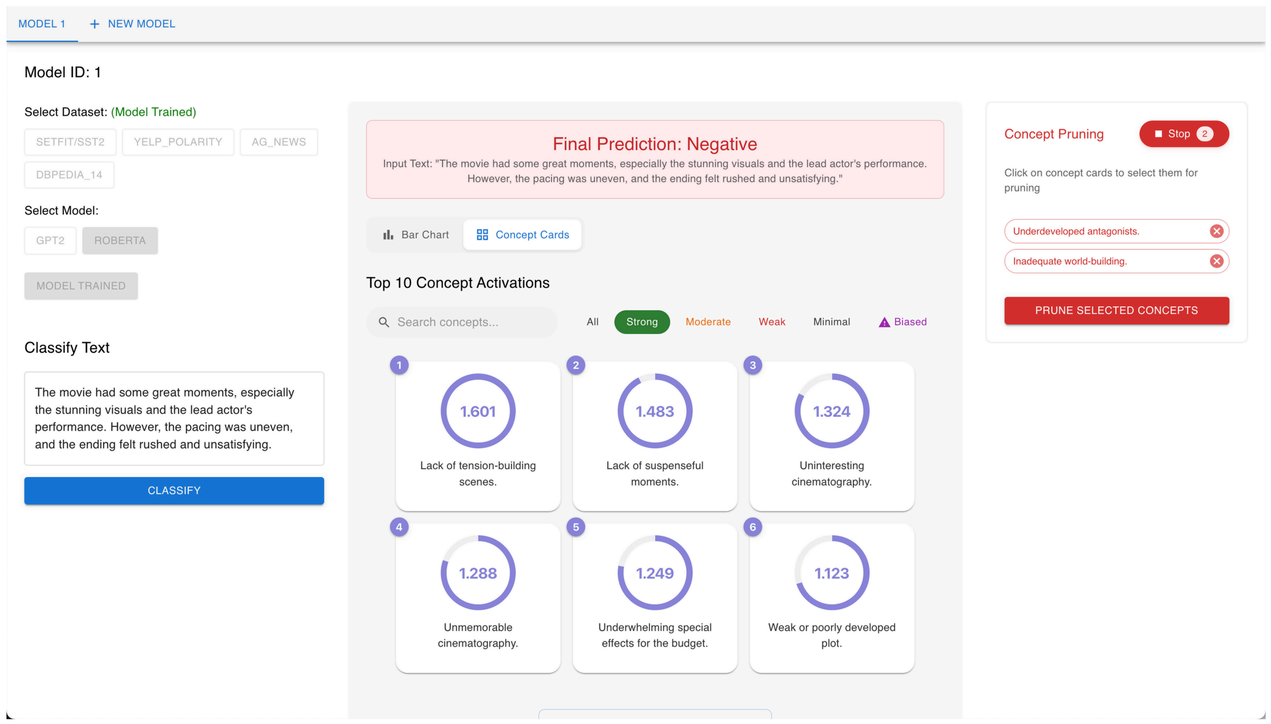


Figure 1: The model classifies the input and visualized the top activated concepts. Users can interactively select and prune concepts to refine the model's performance and fairness.

Tech Stack

Built on Django-React framework:

- **Frontend** (React + JavaScript): Provides an interactive GUI for users to integrate, visualize, and refine Concept Bottleneck Layers (CBLs).
- **Backend** (Django + Python): Orchestrates model training workflows, handles API requests, and manages all core logic.
- **Database** (SQLite3): Stores trained models, concept mappings, and user history for reproducibility.

Conclusion & Discussion

Our CBL-GUI improves LLM interpretability by connecting neuron activations to human-understandable concepts, enhancing transparency and fairness. Built on Lily Weng et al.'s CBL framework, it simplifies concept visualization for users. In the future, we aim to extend CBL integration to image classification models, bringing the same level of interpretability to vision tasks.

Website

References

[1] Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. "Concept Bottleneck Large Language Models". ICLR, 2025