iLLuMinaTE: An LLM-XAI Framework Leveraging Social Science Explanation Theories Towards Actionable Student Performance Feedback

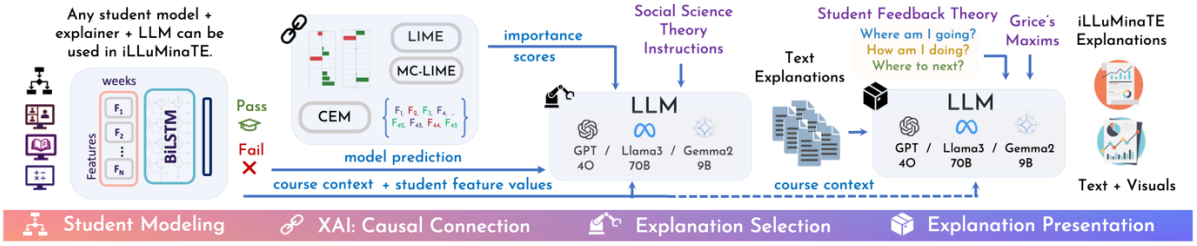

Figure 1: The `iLLuMinaTE` pipeline: from modeling to actionable feedback.

Abstract

Recent advances in eXplainable AI (XAI) for education highlight a critical challenge: ensuring that explanations for state-of-the-art models are understandable for non-technical users such as educators and students. In response, we introduce iLLuMinaTE, a zero-shot, chain-of-prompts LLM–XAI pipeline inspired by Miller (2019)'s cognitive model of explanation. iLLuMinaTE delivers theory-driven, actionable feedback to students in online courses, navigating three stages—causal connection, explanation selection, and explanation presentation—guided by eight social science theories (e.g., Abnormal Conditions, Pearl's Model, Necessity and Robustness, Contrastive Explanation). We evaluate 21,915 explanations generated by three LLMs (GPT-4o, Gemma2-9B, Llama3-70B) across three XAI methods (LIME, CEM, MC-LIME) and three diverse MOOCs. Evaluation covers theory alignment, readability, and a user study with 114 university students including a novel actionability simulation. Students preferred iLLuMinaTE explanations 89.52% of the time. Our work provides a robust, ready-to-use framework for effectively communicating hybrid XAI insights in education, with potential for broader human-centered domains.

Motivation

Traditional XAI methods like LIME or SHAP produce feature-importance vectors that are technically valid but difficult for non-experts to use. In education, students and instructors struggle to connect such outputs to next steps. The challenge is to bridge rigorous XAI methods with feedback that is concise, trustworthy, and actionable.

Methodology

iLLuMinaTE follows a four-stage pipeline:
1) Student Modeling: BiLSTMs predict early success from five weeks of behavioral features (76.8–90.8% balanced accuracy). **2) XAI Causal Connection**: LIME, CEM, and MC-LIME extract local drivers. **3) Theory-Guided Selection**: LLMs apply eight social science theories (e.g., Contrastive, Abnormal Conditions, Pearl).**4) Presentation**: Feedback is structured using Hattie & Timperley's framework and Grice's maxims. **Datasets**: We use data from three MOOCs offered by a European university on the edX platform: Digital Signal Processing (STEM), Villes Africaines (social sciences), and Éléments de Géomatique (applied engineering). Each course attracted an international learner base and combined video lectures, quizzes, and assignments. From raw clickstream logs, we extracted 45 behavioral features capturing regularity, engagement, control, and participation.

Evaluation

We assess: (a) theory alignment of explanation selection, via expert and GPT-4o rubrics; (b) quality, using readability metrics (Flesch–Kincaid, Gunning Fog, SMOG, grammar issues); (c) student preferences, from a 114-participant study; and (d) actionability, by simulating performance gains when students adopt suggested interventions.

Key Results

Alignment: All explainer–theory pairs exceed 0.82 instruction-following; GPT-4o leads, Gemma2-9B and Llama3-70B close behind.
Readability: GPT-4o generates the clearest text; Llama3-70B commits the fewest grammar errors.
114 Student Study: Students preferred iLLuMinaTE explanations in 89.52% of cases, rating them more useful, trustworthy, and actionable.
Actionability: A novel actionability simulation favors +13.5% (LIME), +14.2% (CEM), +20.7% (MC-LIME), with contrastive explanations reaching +28.2%.

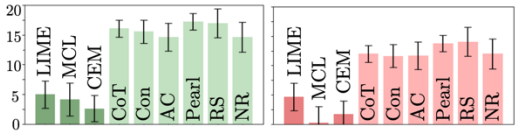

Figure 2: University students overwhelmingly preferred `iLLuMinaTE` explanations over baselines.

Significance & Outlook

iLLuMinaTE shows how LLMs can act as communicators rather than explainers, grounding feedback in social science theory. This modular design generalizes beyond education to healthcare, social services, and recommendation systems. Future work includes interactive explanation dialogues, longitudinal studies, and teacher-in-the-loop integration. Code is available at: https://github.com/epfl-ml4ed/iLLuMinaTE.