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

Vinitra Swamy*[1], Davide Romano*[1], Bhargav Srinivasa Desikan[2], Oana-Maria Camburu[3], Tanja Käser[1]

[1]EPFL, Switzerland
[2]Institute for Public Policy Research, UK
[3]University College London, UK
vinitra.swamy@epfl.ch, davide.romano@epfl.ch, b.srinivasa-desikan@ippr.org, o.camburu@cs.ucl.ac.uk, tanja.kaeser@epfl.ch

Abstract

Recent advances in eXplainable AI (XAI) for education have highlighted 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 is designed to deliver theory-driven, actionable feedback to students in online courses. iLLuMinaTE navigates three main stages — causal connection, explanation selection, and explanation presentation — with variations drawing from eight social science theories (e.g. Abnormal Conditions, Pearl's Model of Explanation, Necessity and Robustness Selection, Contrastive Explanation). We extensively evaluate 21,915 natural language explanations of iLLuMinaTE extracted from three LLMs (GPT-4o, Gemma2-9B, Llama3-70B), with three different underlying XAI methods (LIME, Counterfactuals, MC-LIME), across students from three diverse online courses. Our evaluation involves analyses of explanation alignment to the social science theory, understandability of the explanation, and a real-world user preference study with 114 university students containing a novel actionability simulation. We find that students prefer iLLuMinaTE explanations over traditional explainers 89.52% of the time. Our work provides a robust, ready-to-use framework for effectively communicating hybrid XAI-driven insights in education, with significant generalization potential for other human-centric fields.

1 Introduction

Over the last decade, AI has seen widespread application in education, encompassing both learner-centric models — such as intelligent tutoring systems (Mousavinasab et al. 2021), knowledge tracing (Piech et al. 2015), and automated feedback systems (Jacobsen and Weber 2023) — and teacher-centric models, including real-time classroom insights (Holstein et al. 2018) and automated question generation (Hang, Tan, and Yu 2024). To adopt these solutions in real-world classrooms, model explainability is essential. Nazaretsky et al. (2022) underscore the importance of transparency for fostering educators' trust in AI-based educational technologies, while Conati, Porayska-Pomsta, and Mavrikis (2018) stress the need for interpretable models in contexts where students see decision outcomes without understanding the underlying reasoning.

Recent literature on eXplainable AI (XAI) in education can be categorized into three main motivations: (1) allowing educational stakeholders to audit model mistakes (Khosravi et al. 2022; Pinto and Paquette 2024), (2) building student and teacher trust in AI (Nazaretsky et al. 2024), and (3) designing personalized interventions for students (Hur et al. 2022; Asadi et al. 2023). The most popular approaches in XAI for education are post-hoc explainers, like LIME (Hasib et al. 2022; Scheers and De Laet 2021) and SHAP (Baranyi, Nagy, and Molontay 2020; Mu, Jetten, and Brunskill 2020). These explainers treat the underlying model as a black-box, enabling explanations after model training.

Despite their rising popularity, XAI methods suffer from a major weakness: a lack of adequate understandability, especially for a non-technical audience. At a course level, STEM professors expressed difficulty in understanding explainer outputs, requesting "more concrete and granular insights" on the scale of individual students (Swamy et al. 2023). At the individual student level, Hur et al. (2022) designed XAI-based interventions, but found they were extensive to integrate and provided limited learning gains compared to expert feedback. With pervasive educational benchmarks like ASSISTments (Heffernan and Heffernan 2014) and MOOCRadar (Yu et al. 2023) having hundreds of features and temporal aspects, it becomes difficult for students and teachers to interpret feature importance.

LLMs can be useful in making XAI more human-interpretable, especially towards building stakeholder trust in AI and designing personalized student interventions. A recent study (Kroeger et al. 2024) suggests that LLMs can act as post-hoc explainers for complex models, finding that LLMs could identify relevant features when given examples of input data and model outputs. Atanasova et al. (2022) integrate explanation generation directly into the LLM's training, optimizing over diagnostic properties like data consistency and confidence. However, in domains where explanations directly influence human decisions, the nature of LLMs

*These authors contributed equally to this work.

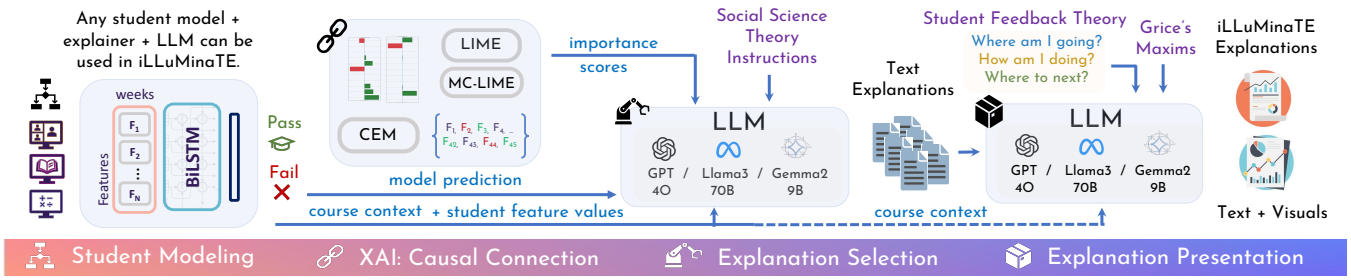

Figure 1: **iLLuMinaTE involves four steps**: 1) modeling course data for student success prediction, 2) using XAI methods to extract feature importance scores, 3) selecting important aspects of the explanation through an LLM aligned with a given social science theory, 4) presenting the explanation with concise and actionable suggestions through an LLM.

as "stochastic parrots" can be accompanied by detrimental side effects (Bender et al. 2021; Sarkar 2024). LLMs, while demonstrating potential in areas such as knowledge tracing or student synthesis, have not yet matured enough to act as student models that can be accompanied by inherent explanations (Neshaei et al. 2024; Nguyen, Tschiatschek, and Singla 2023). We instead propose to use LLMs as *communicators of explanations* (Zytek, Pidò, and Veeramachaneni 2024) to present XAI outputs in aligned text and visual formats that are *actionable* for educational stakeholders.

We therefore present iLLuMinaTE, an in-context, chain-of-prompts, zero-shot LLM pipeline that is inspired by Miller's cognitive processes of explanation (Miller 2019). iLLuMinaTE follows three main stages: (1) causal connection, (2) explanation selection, and (3) explanation presentation. Our experiments range over eight prevalent social science theories of explanation (Hilton 1990; Hilton and Slugoski 1986; Halpern and Pearl 2005; Lombrozo 2010; Woodward and Ross 2021), with three underlying explainers (LIME, Counterfactuals, MC-LIME), data from three online courses, evaluated using three LLMs (GPT-4o, Gemma2 9b, Llama3 70b) and a real-world user study with 114 university students. Notably, we find that students preferred iLLuMinaTE's explanations over baseline hybrid (text and visual) explanations from post-hoc methods 89.52% of the time, and had a particular preference on actionability for abnormal, pearl, and contrastive explanations. With our study, we make the following main contributions:

1. **iLLuMinaTE, a chain-of-prompts framework in the education context** to extract theory-driven natural language explanations (NLE) for student feedback.
2. **An LLM-XAI efficacy analysis** of 216 variations of iLLuMinaTE prompting strategies over explainers, LLMs, social science theories, and student populations.
3. **A real-world evaluation of LLM-XAI preferences** conducted with 114 university students.
4. **An XAI actionability study** simulating student performance gains based on actions they selected from generated explanations.

We provide our modular implementation of iLLuMinaTE publicly with adaptations for LangChain, Groq Cloud, and Replicate[1]. Our work provides a theory-driven methodology to communicate results of XAI to students, with broad generalization potential of explanation theory instruction prompts for other human-centric fields (e.g., healthcare, welfare, product recommendation).

[1]https://github.com/epfl-ml4ed/iLLuMinaTE

2 Methodology

Our iLLuMinaTE pipeline (see Fig. 1) consists of four stages towards communicating explanations in a human-understandable and actionable way through LLMs.

In the *Student Modeling* phase, we extract behavioral features from raw clickstreams of student interactions and use BiLSTMs (Graves and Schmidhuber 2005) to predict student success following prior work (Asadi et al. 2023; Swamy, Marras, and Käser 2022). We then employ post-hoc explainers to obtain feature importance scores, representing the *XAI: Causal Connection* step. With the results from an explainer, the course context, and the student's feature values, we prompt an LLM using *Explanation Selection* instructions specific to social science theories of explanations. We evaluate the obtained explanations using human expert and GPT-4o annotations. We then use *Explanation Presentation* prompts to summarize the often verbose explanation selection reports into concise and actionable feedback for a student, taking into account theory on effective feedback (Hattie and Timperley 2007)'s and maxims for communication (Grice 1975). We then evaluate the final explanations using expert annotations. We assess *Student Preferences* and the actionability of iLLuMinaTE explanations in comparison with the current state-of-the-art in XAI for education in a user study with 114 students.

2.1 Student Modeling

To create the models building the basis for the explanations, we use the same features and model architectures as prior research working with the same datasets (Swamy, Marras, and Käser 2022; Galici et al. 2023; Swamy et al. 2023, 2024).

Data Collection. Our experiments are based on data collected from three MOOCs (Digital Signal Processing (DSP), Villes Africaines (VA), and Elements de Geomatique (Geo)) offered by a European University to a global student audience. The courses were organized into weekly modules including video lectures and quizzes and required students to complete graded assignments to earn course certificates. Students interacted with learning objects (videos, quizzes)

associated with specific course weeks, enabling the creation of course-specific learning indicators. We represent a student's interations as a time series $I_s^c = i_1, \ldots, i_K$, where each interaction i is a tuple (t, a, o), including a *timestamp* t, an *action* a (e.g., video play, pause; quiz submission), and a *learning object* o. The *binary success label* (pass-fail) for student s in course c is denoted as $y_{s,c}$. Data collection and analysis were approved by the university's ethics review board (HREC 058-2020/10.09.2020).

Feature Extraction. We use a broad set of 45 behavioral features h derived from the student interactions, incorporating features from four feature sets shown to be predictive for student performance in MOOCs (Marras, Vignoud, and Käser 2021). **Regularity** features (3) capture consistent study habits (Boroujeni et al. 2016), **Engagement** features (13) measure course involvement (Chen and Cui 2020), **Control** features (22) analyze video usage (Lallé and Conati 2020), and **Participation** features (7) track attendance in scheduled activities (Marras, Vignoud, and Käser 2021).

Modeling. For a course c and student s, our objective is to build a model that predicts $y_{s,c}$ *early*, using the features h_s from the first five weeks. Following prior work (Swamy, Marras, and Käser 2022), we employ a BiLSTM for this task. We provide all reproducibility details in Appendix 5.

2.2 XAI: Causal Connection

We use three popular post-hoc explainers (LIME, CEM, and MC-LIME) to extract local, instance-specific explanations from the student models. We chose these three methods based on their popularity, but any in-hoc or interpretable-by-design model could be used.

LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro, Singh, and Guestrin 2016) provides interpretable explanations for individual predictions by approximating the complex model locally with an interpretable model. This process results in a set of feature weights indicating the positive or negative influence of each feature on the prediction.

CEM (Contrastive Explanation Method) (Dhurandhar et al. 2018) is a counterfactual method that identifies the features that need to be present (pertinent positives) or absent (pertinent negatives) for a model to maintain its prediction for a given instance.

MC-LIME (Minimal Counterfactual LIME) (Cohausz 2022) finds the minimal set of LIME features that, if changed, would alter the outcome. It focuses on features that increase the likelihood of an event (e.g., student dropout). MC-LIME applies changes to individual features, then pairs, and so on, until a change flips the prediction. This approach combines explanation sparseness with the advantages of counterfactuals and LIME.

2.3 Explanation Selection

Our `iLLuMinaTE` pipeline generates explanations aligned to social science theories. Specifically, we have translated eight theories into prompts: *Relevance Selection (RS), Abnormal Conditions (AC), Pearl's Model (Pearl's), Necessity and Robustness (NR), two contrastive explanations (BC, Con), Statistical Relevance (SR),* and *Chain of Thought (CoT)*. Our prompt structure contains (1) descriptions of the model, features, XAI method, and course context, (2) explainer importance scores and relevant student feature values, and (3) social science theory instructions. We present the general prompt template and two theory-specific examples (RS, AC) in this section, all details are in Appendix 6.1.

Explanation selection template You are an AI assistant that analyzes struggling students behavior to help them in their learning trajectories and facilitate learning in the best possible way. You have the following information to help you in your goal:

- A model prediction of student performance at the end of the course, in the form of "pass" or "fail".
- A post-hoc explainable AI approach that identifies which features are important to this student's prediction.
- Data in the form of student's features over 5 weeks that were used by the model. You will see always the most relevant features selected by {*explainer*}.
- Data in the form of student's features over 5 weeks that were used by the model. You will see always the most relevant features selected by {*explainer*}.
- The course {*course_name*} content and structure.
- Detailed instructions on how to reason.

{*model_description*}
{*features_description*}
{*explainer_description*}
{*course_description*}

Take into consideration this data:
{*explainer_importance_scores*}
{*student_feature_values*}

INSTRUCTIONS:
{*theory_instructions*}

QUESTION: Given the information above, follow the instructions precisely and write a small report on what you found. Only use the results from the explainable AI approach and the student's behavior data to justify your conclusions.

Relevance Selection. Based on Hilton's conversational model of explanation (Hilton 1990), relevance-based selection theory aims to "resolve a puzzle in the explainee's mind" by filling gaps in their knowledge. The theory emphasizes that shared knowledge between the explainer and the explainee are presuppositions of the explanations, and the other factors are the causes that should be explained. In short, the explainer should not explain any causes they think the explainee already knows.

Abnormal Conditions. This theory, based upon Hilton and Slugoski's abnormal conditions model (Hilton and Slugoski 1986), suggests that explanations often rely on unusual and temporally proximal events. People do not solely count on statistical likelihood but highlight uncommon factors contributing to an event to explain that event. During a conversation, the explainer is relying on the perceived common prior

Relevance-based selection prompt

[1] Select the causes that are most relevant to the question, context and user

[2] Select the causes that include information that is not already shared with the student

Abnormal conditions model prompt

1. Select potential causes using these criteria:
   - *Abnormality*: Tend to prefer abnormal causes.
   - *Temporality*: Recent events are more relevant for the user and considered more mutable.
   - *Controllability*: focus on the features that the student can control.
2. Select one explanation that follows all of the criteria above (Abnormality, Temporality, Controllability).

knowledge to identify potential causes that are considered abnormal, with greater weight given to temporally proximal events and factors that the explainee can control. This focus on controllable factors helps the explainee understand how to potentially avoid similar situations in the future.

Pearl's Model of Explanation. Halpern and Pearl (2005) present a formal framework for selecting explanations based on epistemic relevance and structural causal models. The model distinguishes between *exogenous variables*, whose values are determined by external factors, and *endogenous variables*, whose values are influenced by relationships with other variables. Within a *context* (a specific assignment of values to variables), the model defines an *actual cause* as a minimal set of events that must occur for an event to happen.

Necessity & Robustness Selection. Two key criteria for selecting strong explanatory causes are necessity and robustness (Lipton 1990). *Necessity* refers to whether a cause is essential for the effect to occur. *Robustness* considers how generally a cause applies (Lombrozo 2010). This idea aligns with the concept of simplicity, where broader explanations with fewer specific requirements are favored.

Contrastive Explanation. This theory suggests explanations are not simply cause and effect statements, but rather comparisons between what happened (the target event) and what could have happened (a counterfactual contrasting event) (Hilton 1990). One possible way to make the identification of the counterfactual event (foil) successful is to ask the module to reformulate the question. This technique is known as Rephrase and Respond (RaR) (Deng et al. 2024).

Statistical Relevance. This method is based on the SR model based on scientific causal reasoning (Woodward and Ross 2021). The SR model explanations can be defined in simple terms as "statistically relevant properties are explanatory and statistically irrelevant properties are not". It follows this structure: *"Based on empirical data, factors A, B and C contribute to the probability of Y by the amount of X"*.

Chain-of-Thought (Baseline). Chain-of-Thought (CoT) prompting (Wei et al. 2023) guides an LLM through sequential reasoning that mimics human thought processes.

Evaluation. To assess whether `iLLuMinaTE` responses align with post-hoc explanations, student feature values, and course context, we developed an annotation rubric based on decomposed questions as a basis for human and LLM annotation. Recent studies (Wang et al. 2023) demonstrated that LLMs can match human annotators, especially when instructions are decomposed into simple criteria phrased as binary ("yes"/"no") questions (Qin et al. 2024). We created four general decomposed questions applicable across all theories and additional theory-specific questions ($1-6$, depending on the theory). These were validated by a computational social scientist who was not involved in the prompt creation process. The general questions used items such as 'Is the generated text correctly using the model's predicted outcome?' or 'Is the generated text analysis based solely on the explainer results provided?'. An example of a theory-specific questions is: 'Is the generated text selecting the causes that are most relevant to the user?' (*RS*). The complete rubric is available in Appendix Table 5.

We instructed three experts to annotate the same block of 42 responses (2 students, 3 explainers, 8 theories) across a set of up to eight decomposed questions for each setting, resulting in an inter-rater agreement of $\kappa = 0.71 \pm 0.13$ (Cohen's Kappa). After that, each annotator proceeded with independently evaluating between 84 to 105 more responses, leading to a total of 315 human-annotated responses and in total over $2,350$ annotations. We then proceeded with instructing GPT-4o to annotate the same 42 responses using the same decomposed questions. We obtained an agreement of $94.68\% \pm 5.20$ (percentage of answers with agreement) between human and GPT-4o annotation. The inter-rater agreement per theory is in Appendix Table 4. Given the high agreement, we annotated all explanations by GPT-4o.

2.4 Explanation Presentation

Inspired by Hilton's conversational model (Hilton 1990), we employ a presentation prompt to refine and condense the output of the explanation selection prompt to make it relevant to the explainee's future actions, current context and prior knowledge. This prompt is formulated using Grice's maxims on communication (1975) and learning science literature on best practices in communicating feedback to students (Shute 2008; Hattie and Timperley 2007). Grice's maxims provide a framework for understanding cooperative conversation, emphasizing providing the right amount of relevant and clear information to achieve a shared goal. Grice further divides this principle into four maxims: *Quality* (truthfulness and evidence-based statements), *Quantity* (providing enough but not excessive information), *Relation* (relevance), and *Manner* (clear and concise communication). Hattie and Timperley (2007) provide a prominent educational feedback framework with three steps: *"Where am I going?"*, *"How am I doing?"*[2], and *"Where to next?"* These steps require the feedback provider to clearly state the learning goal, provide a summary of relevant student performance, and suggest concrete actions for improvement in the

[2]We adapted Hattie and Timperley (2007)'s question *"How am I going?"* to *"How am I doing?"* for ease of understanding.

near future. The full explanation presentation prompt is included in the Appendix 6.2.

Evaluation. Similar to the explanation selection phase, we evaluated the final explanations using a rubric of decomposed 'yes'/'no' questions (the detailed set of questions is illustrated in Appendix Table 5). Given the high agreement between GPT-4o and human annotations in the explanation selection phase, we used GPT-4o as an expert. Additionally, we conducted a readability evaluation using the following metrics: *Flesch-Kincaid Grade Level* (Flesch 1948), *Gunning Fog Index* (Gunning 1952), *SMOG index* (Mc Laughlin 1969), and LanguageTool *Grammar Issues* (Mozgovoy 2011). These metrics evaluate the comprehensibility of the text in terms of sentence length, vocabulary complexity, grammatical correctness, and estimated years of schooling needed for understanding.

2.5 Student Preferences (User Study)

To evaluate students' explanation preferences as well as the actionability of `iLLuMinaTE` explanations, we conducted a user study comparing our explanations **(text and visual)** to post-hoc baselines. We recruited 114 students on `Prolific`, (see Appendix 9.1 for detailed information about the participants' demographics). Students were told that the explanations related to their own performance in three different online courses they were enrolled in.

For each course, we presented participants with eight explanations on their predicted success or failure in that course: four explanations at a time, with three randomly ordered `iLLuMinaTE` variations and one baseline approach. We elected to use the six `iLLuMinaTE` instructions with the highest instruction-following accuracy from the human-expert evaluation for this experiment, and an equal mixture of passing and failing behavior at different model confidence levels. Each explanation was provided as a brief text accompanied by a graph illustrating the features and concepts used by the model. GPT-4o created all `iLLuMinaTE` visuals based on the first two responses (full prompt in Appendix 6.3). LIME visuals were used from the package, and CEM and MC-LIME visuals were expert-created and iterated upon with six pilot participants. Examples of the study format are included in our repository and Appendix 9.2. Participants were asked to choose their preferred explanation for each set of comparisons and explain their choice (open-ended question). They were then asked to compare the explanations based on five criteria (Frej et al. 2024):

1. **Usefulness**: This explanation is useful to understand the prediction based on my learning behavior.
2. **Trustworthiness**: This explanation lets me judge if I should trust the suggestions.
3. **Actionability**: This explanation helps me make a decision on how to improve my learning behavior.
4. **Completeness**: This explanation has sufficient detail to understand why the prediction was made based on my learning behavior.
5. **Conciseness**: Every detail of this explanation is necessary.

Finally, participants were asked to choose one of ten suggested actions for the next week, based on their preferred explanation. These actions were aligned with behavioral features from the model, allowing us to simulate the impact if the student acted according to them. We trained a BiLSTM to predict student success on six weeks of student data, and conducted inference on simulated students increasing the relevant features by 25% percentile. We also asked students which weeks of material they would focus on based on the explanation (between one to three weeks).

3 Results

We evaluated `iLLuMinaTE` explanations by assessing the LLM's instruction-following abilities during explanation selection (Exp 1) and using readability metrics and an automated analysis to measure explanation understandability at the presentation stage (Exp 2). We then analyzed student preferences in the user study (Exp 3) and simulated whether actions derived from the explanations improved student performance (Exp 4).

Experimental Protocol. We optimized the BiLSTM models (one for each course) using a train-validation-test split of 80:10:10 and including a hyperparameter search. We achieve balanced accuracies of 90.8 (DSP), 80.3 (Afr), and 76.8 (Geo) respectively. These results for early student performance prediction at five weeks are in line with prior work (Swamy, Marras, and Käser 2022). Explainers (LIME, CEM) were extracted with the same settings as per related work to ensure a fair comparison (Swamy et al. 2022, 2023).

3.1 Exp 1: `iLLuMinaTE` is aligned with social science theories of explanation

In the first analysis, we evaluated how well the generated explanations aligned with the instructions. We selected 105 representative students per course, distributed across six behavioral dimensions (regularity, effort, consistency, proactivity, control, and assessment) (Mejia-Domenzain et al. 2022). Table 1 shows the average (with standard deviation) scores for each general decomposed question (see Section 2.3 and Appendix 7), for the theory-specific questions, and overall. The general decomposed questions involve whether the response is using the provided data extensively (Q1), whether the analysis in the response is based solely on the results from the explainer (Q2), whether the response is correctly using the model's prediction (Q3), and whether the response is using the course content and structure. The scores represent the ratio of 'Yes' answers averaged across students and courses, with overall scores exceeding 0.82 for all explainers and theories. There were no significant differences between explainers or theories, as indicated by overlapping 95% CIs. There were generally also no differences in scores between questions. Only Q4 ("Using course content and structure") had lower average scores than the other questions, suggesting that generating explanations fully incorporating the course content is a challenge. Notably, inter-rater agreement for Q4 was also lower than the average ($\kappa = 0.52$ vs. $\kappa = 0.71$ overall), indicating lower reliability of annotations for this question.

In a second analysis, we compared different LLMs' abilities to generate explanations according to instructed theo-

Explainer	Theory	Overall	Q1: Using provided data extensively	Q2: Analysis based solely on explainer	Q3: Correctly using model's prediction	Q4: Using course content and structure	Theory-Specific Qs
CEM	*RaR + Contrastive*	**0.985 ± 0.069**	0.99 ± 0.099	0.992 ± 0.087	0.998 ± 0.047	0.967 ± 0.178	0.984 ± 0.085
	Abnormal Conditions	**0.958 ± 0.141**	0.997 ± 0.057	0.997 ± 0.057	0.998 ± 0.047	0.865 ± 0.342	0.952 ± 0.156
	Relevance Selection	**0.949 ± 0.155**	0.998 ± 0.047	0.999 ± 0.033	0.997 ± 0.057	0.898 ± 0.303	0.925 ± 0.195
	Necessity Robustness	**0.949 ± 0.154**	0.996 ± 0.066	0.997 ± 0.057	0.998 ± 0.047	0.676 ± 0.468	0.991 ± 0.047
	Pearl Explanation	**0.948 ± 0.099**	0.979 ± 0.142	0.983 ± 0.131	0.998 ± 0.047	0.869 ± 0.337	0.939 ± 0.071
	Statistical Relevance	**0.884 ± 0.184**	0.988 ± 0.109	0.993 ± 0.081	0.996 ± 0.066	0.453 ± 0.498	0.991 ± 0.001
	(Base) Contrastive	**0.852 ± 0.194**	0.968 ± 0.175	0.647 ± 0.478	1.0 ± 0.0	0.681 ± 0.466	0.879 ± 0.179
	Chain of Thought	**0.872 ± 0.188**	0.979 ± 0.142	0.984 ± 0.127	0.997 ± 0.057	0.585 ± 0.493	0.817 ± 0.001
LIME	*RaR + Contrastive*	**0.975 ± 0.098**	0.974 ± 0.16	0.985 ± 0.123	1.0 ± 0.0	0.946 ± 0.227	0.975 ± 0.111
	Relevance Selection	**0.96 ± 0.112**	0.987 ± 0.114	0.991 ± 0.093	0.998 ± 0.047	0.918 ± 0.274	0.946 ± 0.132
	Abnormal Conditions	**0.951 ± 0.131**	0.99 ± 0.099	0.993 ± 0.081	0.995 ± 0.074	0.893 ± 0.309	0.938 ± 0.156
	Necessity Robustness	**0.929 ± 0.147**	0.982 ± 0.135	0.984 ± 0.127	0.998 ± 0.047	0.659 ± 0.474	0.959 ± 0.103
	Pearl Explanation	**0.896 ± 0.106**	0.904 ± 0.294	0.922 ± 0.269	0.989 ± 0.104	0.836 ± 0.371	0.88 ± 0.114
	Statistical Relevance	**0.859 ± 0.144**	0.919 ± 0.272	0.964 ± 0.186	0.984 ± 0.127	0.476 ± 0.5	0.953 ± 0.001
	(Base) Contrastive	**0.822 ± 0.195**	0.97 ± 0.172	0.45 ± 0.498	0.998 ± 0.047	0.624 ± 0.485	0.884 ± 0.183
	Chain of Thought	**0.82 ± 0.188**	0.945 ± 0.229	0.956 ± 0.204	0.997 ± 0.057	0.526 ± 0.5	0.675 ± 0.001
MC-LIME	*RaR + Contrastive*	**0.985 ± 0.063**	0.987 ± 0.114	0.989 ± 0.104	0.999 ± 0.033	0.968 ± 0.175	0.983 ± 0.073
	Relevance Selection	**0.963 ± 0.132**	0.996 ± 0.066	0.997 ± 0.057	1.0 ± 0.0	0.915 ± 0.279	0.949 ± 0.153
	Abnormal Conditions	**0.96 ± 0.12**	0.991 ± 0.093	0.991 ± 0.093	0.993 ± 0.081	0.879 ± 0.326	0.957 ± 0.135
	Necessity Robustness	**0.932 ± 0.162**	0.99 ± 0.099	0.99 ± 0.099	1.0 ± 0.0	0.609 ± 0.488	0.978 ± 0.083
	Pearl Explanation	**0.919 ± 0.108**	0.942 ± 0.233	0.951 ± 0.216	0.992 ± 0.087	0.806 ± 0.395	0.914 ± 0.104
	Statistical Relevance	**0.876 ± 0.171**	0.97 ± 0.172	0.99 ± 0.099	0.993 ± 0.081	0.446 ± 0.497	0.978 ± 0.001
	(Base) Contrastive	**0.856 ± 0.19**	0.97 ± 0.172	0.615 ± 0.487	0.996 ± 0.066	0.688 ± 0.497	0.895 ± 0.187
	Chain of Thought	**0.858 ± 0.199**	0.973 ± 0.163	0.982 ± 0.135	0.999 ± 0.033	0.565 ± 0.496	0.77 ± 0.001

Table 1: **Alignment of GPT-4o, Gemma2 9b, and Llama3 70B generated explanations with theory**. Average (± std) of "Yes" answers, displayed separately for the first four general decomposed questions as well as averaged over the theory-specific questions and overall. Annotated by experts and GPT-4o. Scores over 95% (less than 65%) are highlighted in green (red).

ries. Table 2 (column *Explanation Selection*) shows the average scores for each model and explainer for the same representative students. Scores reflect the number of 'Yes' answers to the decomposed questions described in Section 2.4, with annotation done automatically using GPT-4o. GPT-4o `iLLuMinaTE` explanations scored the highest, closely followed by Gemma2 9b, and Llama3 70b. However, all of the 95% CI overlap. Again, we found no differences between the different explainers. An experiment on the generalizability of the results to the flipped classroom context (smaller dataset, different domain) is included in Appendix 8.

3.2 Exp 2: `iLLuMinaTE` explanations are understandable

We assessed the presentation of `iLLuMinaTE` explanations using both LLM annotation and readability metrics. Table 2 (column *Explanation Presentation*) shows scores per LLM and explainer, averaged over 300 representative students (see Section 3.1). Scores represent the ratio of "Yes" answers to the set of decomposed questions described in Section 2.4, as annotated by GPT-4o. For this stage, Llama3 70b achieved the highest scores, followed closely by GPT-4o and Gemma2 9b. This finding is in contrast to the explanation selection stage, where GPT-4o reached the highest score, suggesting that depending on the use case (whether selection or presentation is more important), it is possible to use much smaller and open source models for the task at hand.

Figure 2 illustrates the readability scores (Flesch Kincaid,

Model	Explainer	Explanation Selection	Explanation Presentation
Gemma2 9b	*CEM*	0.941 ± 0.202	0.791 ± 0.212
	LIME	0.937 ± 0.217	0.801 ± 0.201
	MC-LIME	0.939 ± 0.208	0.767 ± 0.165
GPT-4o	*CEM*	0.961 ± 0.148	0.992 ± 0.070
	LIME	0.954 ± 0.165	0.992 ± 0.072
	MC-LIME	0.965 ± 0.144	0.992 ± 0.072
Llama3 70b	*CEM*	0.897 ± 0.222	0.998 ± 0.037
	LIME	0.848 ± 0.302	0.997 ± 0.042
	MC-LIME	0.881 ± 0.258	0.998 ± 0.037

Table 2: **Explanation quality by LLM**. Degree of instruction-following for explanations generated by GPT-4o, Gemma2 9b, and Llama3 70b. Average (with standard deviation) of "Yes" answers, annotated by GPT-4o.

SMOG Index, Gunning Fog, Grammar Issues) of the final `iLLuMinaTE` explanations after the explanation presentation prompt. GPT-4o reached the best readability performance of the three LLMs (Fig 2, top), with no overlap in 95% confidence intervals. In contrast, Llama3 70b committed the least grammatical errors. We found no differences in readabilty or grammar between explanations for the different courses Fig 2, middle), demonstrating that our approach is generalizable to diverse educational contexts. Similarly, all explainers achieved similar scores across all four metrics Fig 2, bottom), showing that choice of source explainer does

not have a strong impact on explanation understandability.

3.3 Exp 3: Students prefer `iLLuMinaTE`

For each explanation in the user study, participants were asked to indicate their preferred explanation variation between one base explanation (LIME, CEM, MC-LIME) and three `iLLuMinaTE` explanations. Figure 3 illustrates the percentage of times each method was preferred, separately for either passing or failing student performance. Scores were averaged over all participants and explainers. Students overwhelmingly favored `iLLuMinaTE` explanations over base ones. A Kruskal-Wallis test confirmed this preference, showing a significant difference between the base and `iLLuMinaTE` explanations ($H = 176.38, p < .0001$).

We evaluated students' responses to five Likert-scale questions on the usefulness, trustworthiness, actionability, completeness, and conciseness of explanations. Figure 4 shows the Likert score distribution (1 - 5) for two criteria per method; full results are in Appendix 9.4. Theory-based explanations received consistently high scores with no significant differences between theories. Base explanations were rated substantially lower (e.g., $\text{Usefulness}_{\text{contrastive}} = 4.18$, $\text{Usefulness}_{\text{BASE}} = 3.59$), highlighting the superior usefulness of `iLLuMinaTE` explanations.

3.4 Exp 4: `iLLuMinaTE` explanations can be effective at improving student performance

Participants were asked to choose an action for the next week based on their preferred explanation. Over all explainers and theories, students most frequently selected actions to improve regularity of learning (200 responses) and attempt more problems (147 responses), while the least chosen action was to speed up quiz solving (10 responses). Participants also chose which weeks to focus on in the course, most commonly choosing weeks 6 and 7 (329, 248 responses), which correspond to the weeks directly after the intervention. These were followed by review in weeks 5 (222 responses) and 4 (198 responses). This observation indicates that timing and proximity to the intervention influenced their choices. We also conducted a simulation experiment, applying participants' actions chosen for week 6 to student behavior in that week. For participants preferring `iLLuMinaTE` responses, average performance improved significantly, independent of the underlying explainer: 13.5% for LIME, 14.2% for CEM, and 20.7% for MC-LIME (full results in Appendix 10). MC-LIME's effectiveness may stem from its minimal counterfactual approach, which greedily searches for the smallest set of features that cause the prediction to flip, making it suited for a single- or few-action intervention. Across different theories, both Contrastive and Necessity Robustness explanations result in the most actionable interventions with 28.2% and 24.9% average performance improvement respectively (Appendix Fig. 22).

4 Summary and Outlook

In this work, we addressed the critical need for human-understandable explanations of complex models in education. We introduced `iLLuMinaTE`, a theory-driven framework leveraging LLMs for generating NLEs through a chain

Figure 2: **Readability metrics.** Flesch Kincaid, Gunning Fog, SMOG Index, Grammar Issues across LLM (blue, top), course (green, middle), and explainer (orange, bottom). Lower scores are better.

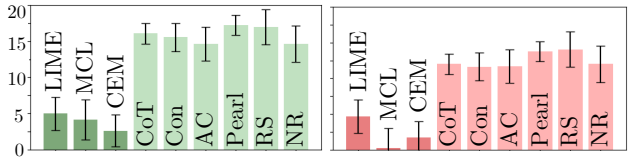

Figure 3: **Student preference of presented explanations for passing (left) and failing (right) student predictions.** Percentage of times a student chose each method when it was available. Higher scores are better.

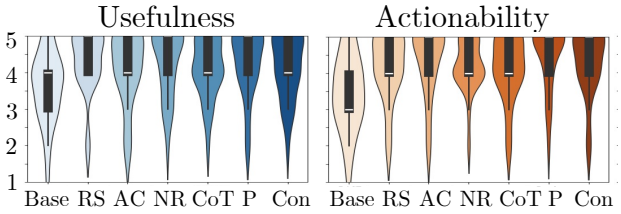

Figure 4: **Comparison of student preferences across two dimensions.** Distribution of Likert scores for *Usefulness* and *Actionability*, averaged over all participants and explainers.

of prompts, consisting of causal connection, explanation selection, and explanation presentation.

We tested our framework on 315 students (105 per course) with all combinations of three post-hoc explainers, three LLMs, and eight prompting strategies, resulting in 21,915 generated NLEs. We evaluated the instruction-following abilities using GPT-4o and human expert annotation, with both decomposed questions and readability metrics. In a user study with 114 university students, we found that students significantly preferred `iLLuMinaTE` explanations over LIME, CEM, and MC-LIME and were able to derive actions that could improve their performance.

Several challenges remain, including explainer variability and the difficulty of evaluating NLEs independently of prior beliefs and knowledge. In future work, we aim to explore interactivity of explanation dialogue with students (Slack et al. 2023) and longitudinal LLM-XAI feedback effects. Our study highlights the shared potential of leveraging LLMs, eXplainable AI, and social science theories together towards scalable, personalized student support.

Acknowledgments

We thank the Swiss State Secretariat for Education, Research and Innovation (SERI) for supporting this project. We also thank Tanya Nazaretsky (EPFL) and Advait Sarkar (MSR) for insightful discussions. Oana-Maria Camburu was supported by a Leverhulme Early Career Fellowship.

References

Asadi, M.; Swamy, V.; Frej, J.; Vignoud, J.; Marras, M.; and Käser, T. 2023. RIPPLE: Concept-Based Interpretation for Raw Time Series Models in Education. *AAAI*.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2022. Diagnostics-guided explanation generation. *AAAI*.

Baranyi, M.; Nagy, M.; and Molontay, R. 2020. Interpretable deep learning for university dropout prediction. *ITE*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAACT*.

Boroujeni, M. S.; Sharma, K.; Kidziński, Ł.; Lucignano, L.; and Dillenbourg, P. 2016. How to quantify student's regularity? In *ECTEL*.

Chen, F.; and Cui, Y. 2020. Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance. *JLA*.

Cohausz, L. 2022. Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science. In *EDM*. Durham, United Kingdom.

Conati, C.; Porayska-Pomsta, K.; and Mavrikis, M. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. *ICML WHI*.

Deng, Y.; Zhang, W.; Chen, Z.; and Gu, Q. 2024. Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves. *arXiv*.

Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. *NeurIPS*.

Flesch, R. 1948. A new readability yardstick. *Journal of App. Psychology*.

Frej, J.; Shah, N.; Knezevic, M.; Nazaretsky, T.; and Käser, T. 2024. Finding Paths for Explainable MOOC Recommendation: A Learner Perspective. In *LAK*.

Galici, R.; Käser, T.; Fenu, G.; and Marras, M. 2023. Do not trust a model because it is confident: Uncovering and characterizing unknown unknowns to student success predictors in online-based learning. In *LAK*.

Graves, A.; and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *IEEE IJCNN*.

Grice, H. P. 1975. Logic and conversation. *Speech acts*.

Gunning, R. 1952. The technique of clear writing.

Halpern, J. Y.; and Pearl, J. 2005. Causes and explanations: A structural-model approach. Part II: Explanations. *BJPS*.

Hang, C. N.; Tan, C. W.; and Yu, P.-D. 2024. MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning. *IEEE Access*.

Hasib, K. M.; Rahman, F.; Hasnat, R.; and Alam, M. G. R. 2022. A Machine Learning and Explainable AIApproach for Predicting Secondary School Student Performance. In *CCC*.

Hattie, J.; and Timperley, H. 2007. The power of feedback. *Review of educational research*.

Heffernan, N. T.; and Heffernan, C. L. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *IJAIED*.

Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin*.

Hilton, D. J.; and Slugoski, B. R. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*.

Holstein, K.; Hong, G.; Tegene, M.; McLaren, B. M.; and Aleven, V. 2018. The classroom as a dashboard: Co-designing wearable cognitive augmentation for K-12 teachers. In *LAK*.

Hur, P.; Lee, H.; Bhat, S.; and Bosch, N. 2022. Using Machine Learning Explainability Methods to Personalize Interventions for Students. *EDM*.

Jacobsen, L. J.; and Weber, K. E. 2023. The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback. *OSF Preprints*.

Khosravi, H.; Shum, S. B.; Chen, G.; Conati, C.; Tsai, Y.-S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; and Gašević, D. 2022. Explainable artificial intelligence in education. *CEAI*.

Kroeger, N.; Ley, D.; Krishna, S.; Agarwal, C.; and Lakkaraju, H. 2024. Are Large Language Models Post Hoc Explainers? *ArXiv*.

Lallé, S.; and Conati, C. 2020. A data-driven student model to provide adaptive support during video watching across MOOCs. In *AIED*.

Lipton, P. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*.

Lombrozo, T. 2010. Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*.

Marras, M.; Vignoud, J. T. T.; and Käser, T. 2021. Can Feature Predictive Power Generalize? Benchmarking Early Predictors of Student Success across Flipped and Online Courses. In *EDM*.

Mc Laughlin, G. H. 1969. SMOG grading-a new readability formula. *Journal of reading*.

Mejia-Domenzain, P.; Marras, M.; Giang, C.; and Käser, T. 2022. Identifying and comparing multi-dimensional student profiles across flipped classrooms. In *AIED*.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Mousavinasab, E.; Zarifsanaiey, N.; R. Niakan Kalhori, S.; Rakhshan, M.; Keikha, L.; and Ghazi Saeedi, M. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *ILE*.

Mozgovoy, M. 2011. Dependency-based rules for grammar checking with LanguageTool. In *FedCSIS*.

Mu, T.; Jetten, A.; and Brunskill, E. 2020. Towards Suggesting Actionable Interventions for Wheel-Spinning Students. *EDM*.

Nazaretsky, T.; Ariely, M.; Cukurova, M.; and Alexandron, G. 2022. Teachers' trust in AI-powered educational technology and a professional development program to improve it. *BJET*.

Nazaretsky, T.; Mejia-Domenzain, P.; Swamy, V.; Frej, J.; and Käser, T. 2024. AI or Human? Evaluating Student Feedback Perceptions in Higher Education. In *ECTEL*.

Neshaei, S. P.; Davis, R. L.; Hazimeh, A.; Lazarevski, B.; Dillenbourg, P.; and Käser, T. 2024. Towards modeling learner performance with large language models. *AIED*.

Nguyen, M. H.; Tschiatschek, S.; and Singla, A. 2023. Large Language Models for In-Context Student Modeling: Synthesizing Student's Behavior in Visual Programming from One-Shot Observation. *EDM*.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep Knowledge Tracing. *NeurIPS*, 28.

Pinto, J. D.; and Paquette, L. 2024. Towards a Unified Framework for Evaluating Explanations. *EDM HEXED*.

Qin, Y.; Song, K.; Hu, Y.; Yao, W.; Cho, S.; Wang, X.; Wu, X.; Liu, F.; Liu, P.; and Yu, D. 2024. InFoBench: Evaluating Instruction Following Ability in Large Language Models. *arXiv*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD*.

Sarkar, A. 2024. Large Language Models Cannot Explain Themselves. *CHI HCXAI*.

Scheers, H.; and De Laet, T. 2021. Interactive and Explainable Advising Dashboard Opens the Black Box of Student Success Prediction. In *ECTEL*.

Shute, V. J. 2008. Focus on formative feedback. *Review of educational research*.

Slack, D.; Krishna, S.; Lakkaraju, H.; and Singh, S. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*.

Swamy, V.; Du, S.; Marras, M.; and Käser, T. 2023. Trusting the Explainers: Teacher Validation of Explainable Artificial Intelligence for Course Design. *LAK*.

Swamy, V.; Marras, M.; and Käser, T. 2022. Meta transfer learning for early success prediction in MOOCs. In *LS*.

Swamy, V.; Radmehr, B.; Krco, N.; Marras, M.; and Käser, T. 2022. Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. *EDM*.

Swamy, V.; Satayeva, M.; Frej, J.; Bossy, T.; Vogels, T.; Jaggi, M.; Käser, T.; and Hartley, M.-A. 2024. Multimodn—multimodal, multi-task, interpretable modular networks. *NeurIPS*.

Wang, J.; Liang, Y.; Meng, F.; Sun, Z.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; and Zhou, J. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv*.

Woodward, J.; and Ross, L. 2021. Scientific Explanation. In *The Stanford Encyclopedia of Philosophy*.

Yu, J.; Lu, M.; Zhong, Q.; Yao, Z.; Tu, S.; Liao, Z.; Li, X.; Li, M.; Hou, L.; Zheng, H.-T.; et al. 2023. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *SIGIR*.

Zytek, A.; Pidò, S.; and Veeramachaneni, K. 2024. LLMs for XAI: Future Directions for Explaining Explanations. *arXiv preprint arXiv:2405.06064*.