RISING STARS IN DATA SCIENCE 2024

Stanford Data Science

THE UNIVERSITY OF CHICAGO DATA SCIENCE INSTITUTE

UC San Diego HALICIOĞLU DATA SCIENCE INSTITUTE

ENCORE

The Future of Human-Centric eXplainable AI

is not post-hoc explanations

VINITRA SWAMY

EPFL

Main Takeaways

EVALUATING THE EXPLAINERS

TRUSTING THE EXPLAINERS

Post-hoc explainers have serious problems,
and there is no effective way to validate them

Main Takeaways

MULTIMODN, INTERPRETCC, ILLUMINATE

With interpretable-by-design NNs,
guaranteed interpretability
does not have to come at the cost of performance
or human-understandability

Vinitra Swamy

XAI Research

5th year PhD Student at EPFL

Co-advised by
Tanja Käser (ML4ED) and
Martin Jaggi (MLO)

UC Berkeley → Microsoft AI → EPFL

AI for Education, eXplainable AI

Teaching data science

Overview

1) Why should we care about explainability?

2) Why are existing XAI methods not enough?

3) Three ideas towards an eXplainable AI future:
- MultiModN
- InterpretCC
- iLLuMinaTE

Cost of using neural networks

AI for Education

Problem: Our current SoTA trades transparency for accuracy

Identifying “why” is important for effective, personalized interventions

Solution: Explainable Machine Learning

Why is eXplainable AI important?

1. Building stakeholder trust in models
2. Auditing models when they make mistakes
3. Improving models with their own reasoning

Framework of Explainability Needs for Human-Centric Computing

Consistent (Li et al., 2021)
[multiple generated explanations are the same]

Real-Time (Xu et al., 2017)
[next minute, next lesson, 1 week, after the course]

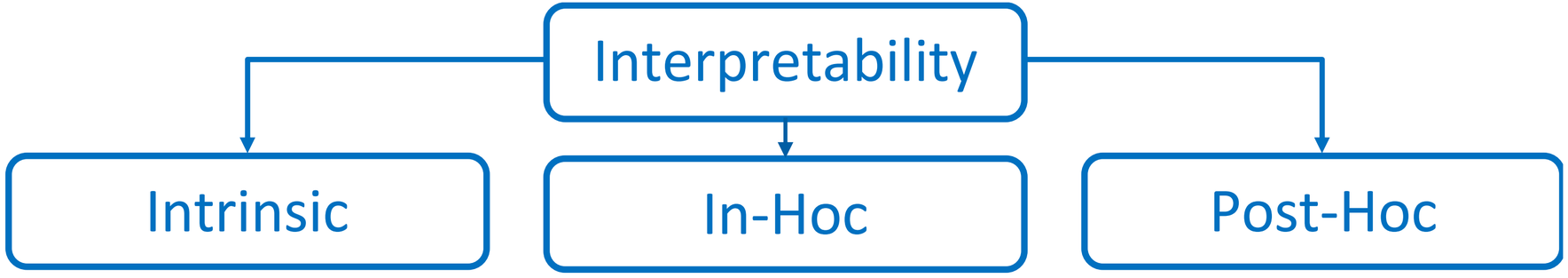
Accurate (Marx et al., 2023; Leichtmann et al., 2023)
[model is confident in the explanation]

Actionable (Joshi et al., 2019)
[able to take a next step based on the insight]

Human-Interpretable (Hudon et al., 2021; Haque et al., 2023)
[easy for a non-data scientist to understand]

Interpretability

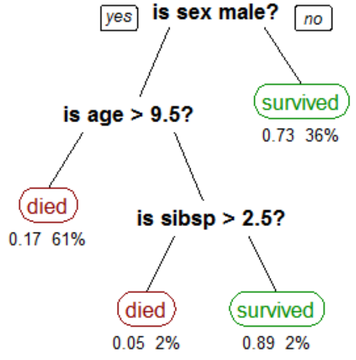
XAI Fundamentals



Decision Trees

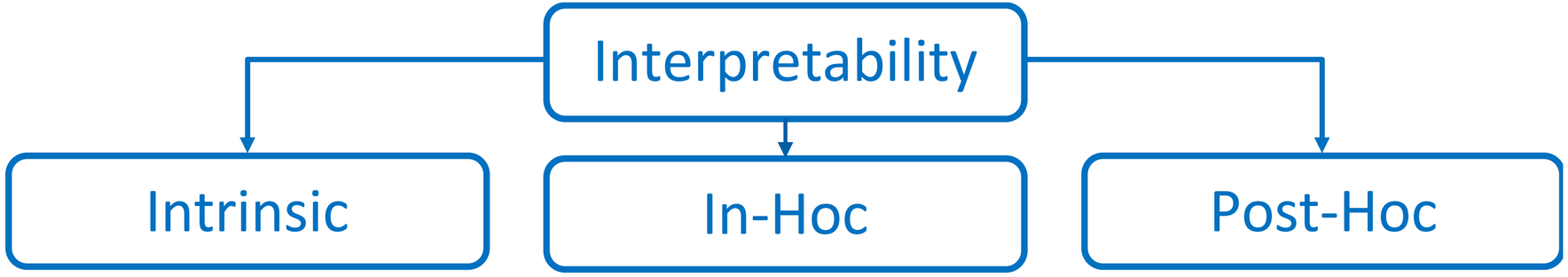
Gradient methods

PDP, LIME, SHAP

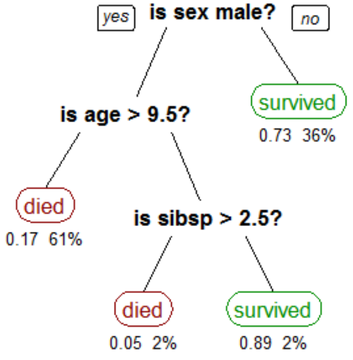


Interpretability

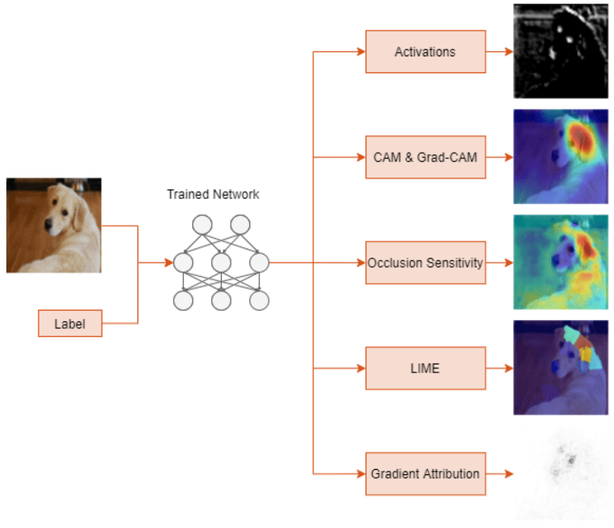
XAI Fundamentals



Decision Trees



Gradient methods

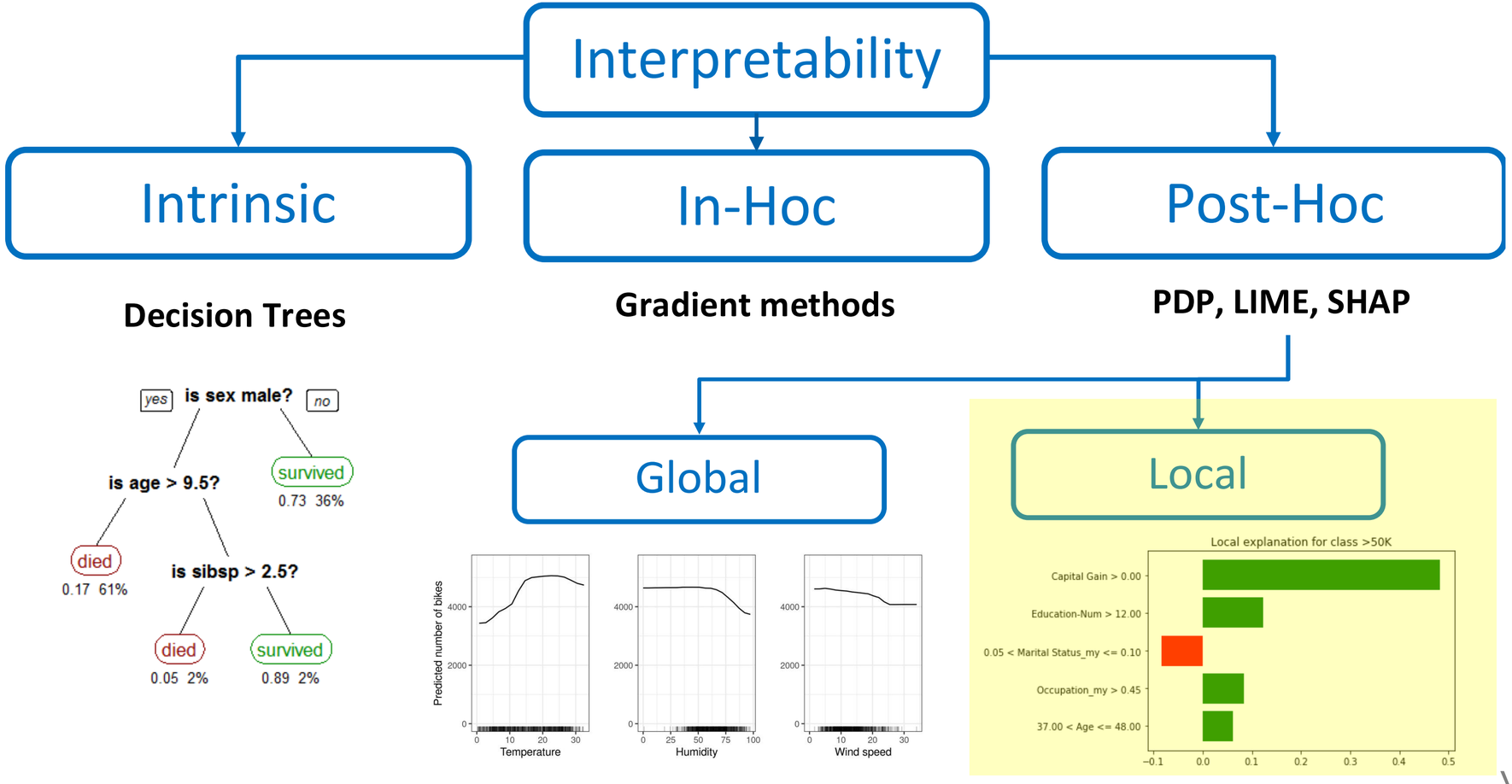


PDP, LIME, SHAP

The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations. Swamy V. et al. (Preprint)

Interpretability

XAI Fundamentals



The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations. Swamy V. et al. (Preprint)

Local XAI

METHODOLOGY

LIME

Local Interpretable Model-Agnostic Explanations

1

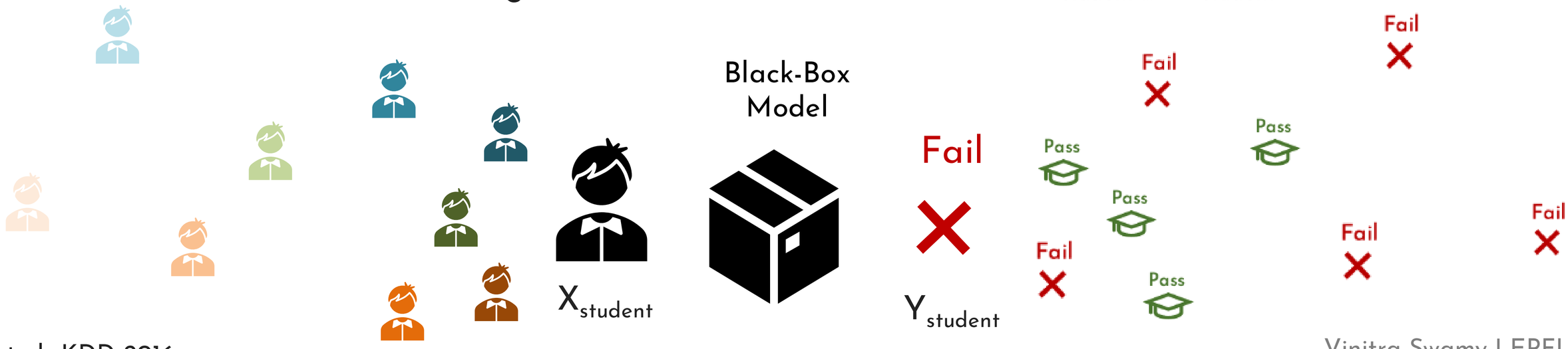
Select a specific point to explain: ($X_{student}$, $Y_{student}$)

2

Perturb features of selected point to get {$X^1_{student}$ ... $X^N_{student}$} neighbors

3

Feed in $X_{student}$ neighbors to the black-box model and get predictions {$Y^1_{student}$ ... $Y^N_{student}$}



Ribeiro et al., KDD 2016.

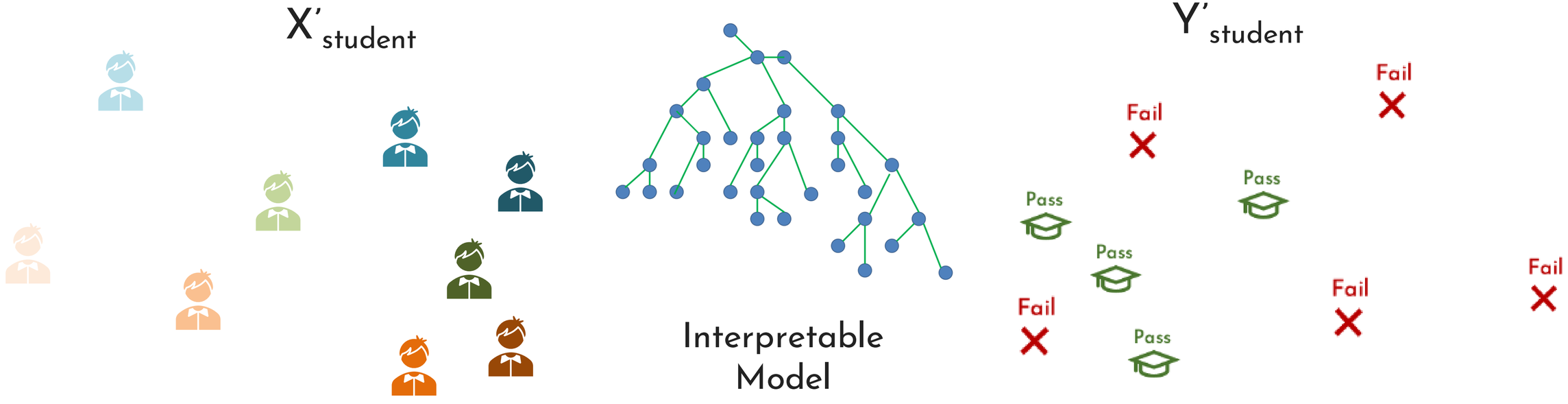
Local XAI

METHODOLOGY

LIME

Local Interpretable Model-Agnostic Explanations

4 Train an interpretable local model using (weighted) $X'_{student}$ and $Y'_{student}$



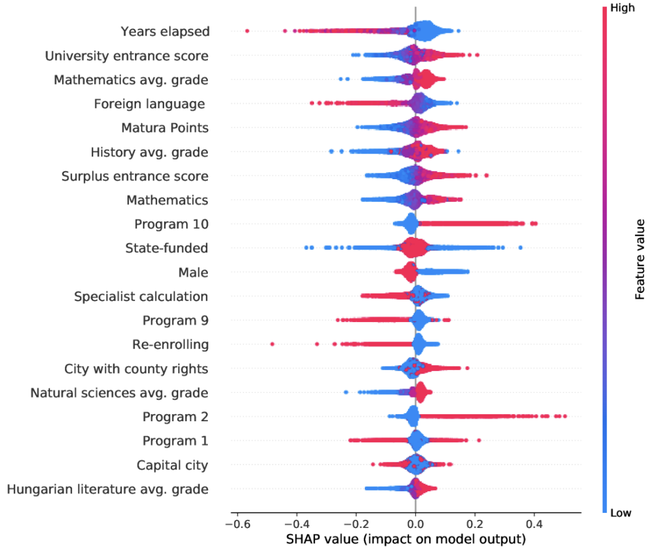
Ribeiro et al., KDD 2016.

Previous Work

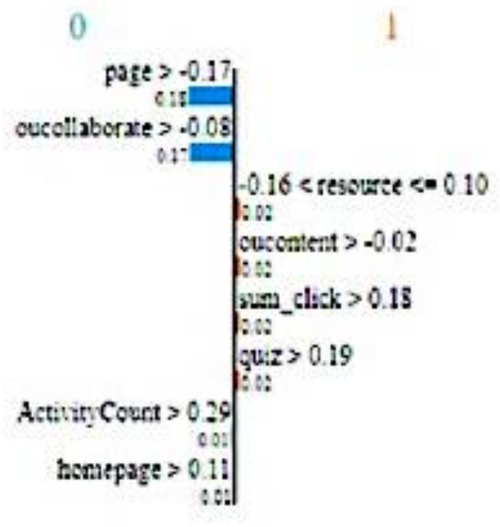
MOTIVATION

Previous work: In (minimal) related literature, only one explainability method is picked per application

SHAP for student dropout[1]



LIME for student advising[2,3]



Baranyi et al. CITE 2020; Scheers and Laet, ECTEL 2021; Pei and Xing, Journal of Edu Computing Research 2021

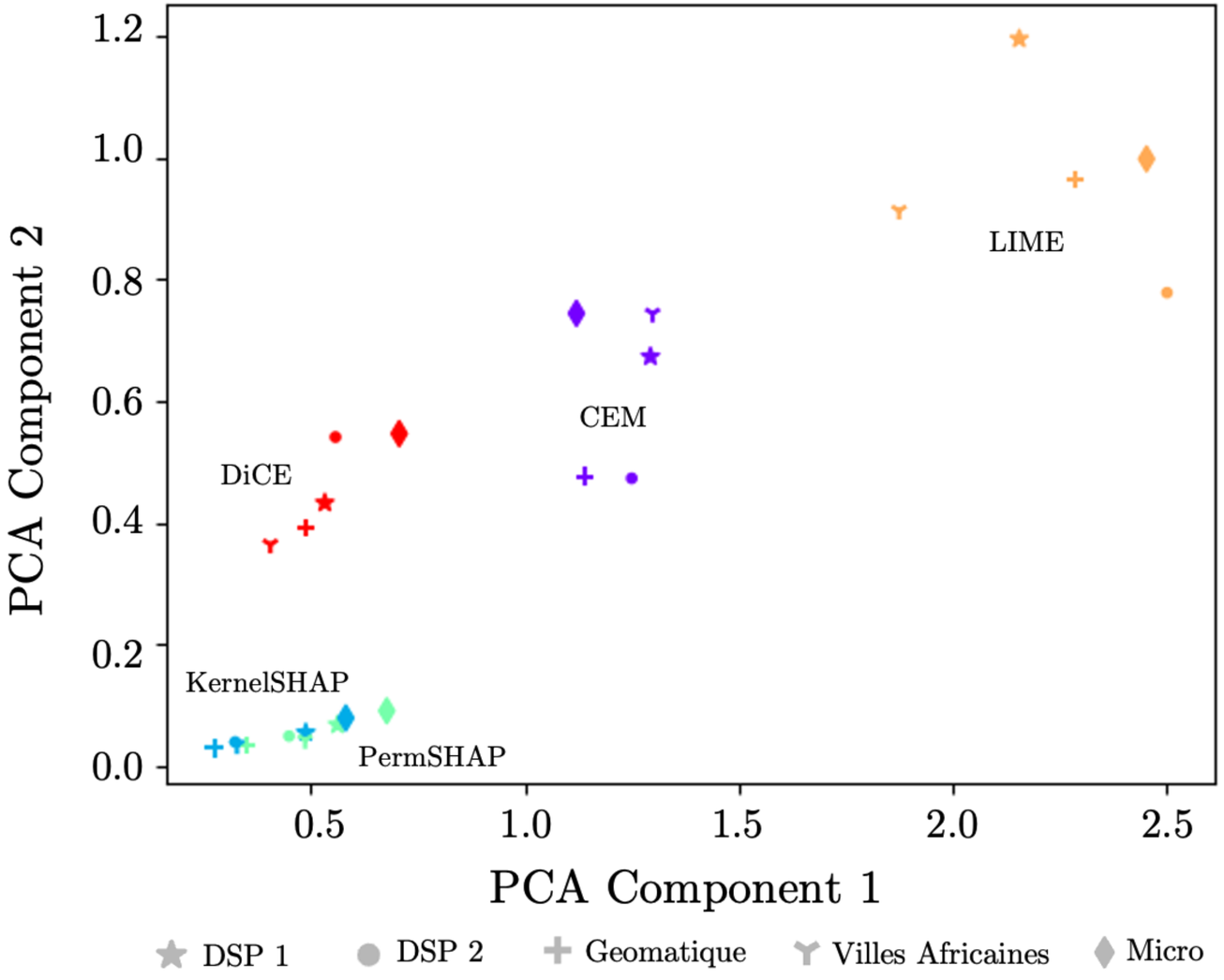
5 Courses

RESULTS

How do explanations (quantitatively) compare across courses?

PCA Analysis

Feature importance clusters by explainability method, not by course



Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. Swamy V., Radmehr B. et al. (EDM 2022)

XAI methods systematically disagree.

How can we build trust in them?

Human expert validation!

Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. Swamy V., Radmehr B. et al. (EDM 2022)

Study Participants

RESULTS

Who can we most trust to validate educational explainers?

Pass

Fail

Over 25 STEM professors

45 minute semi-structured interviews

- diverse in geographic location
- 80.95% identify as male
- age: 36.5 yrs (std: 9 years)
- strong MOOC expertise

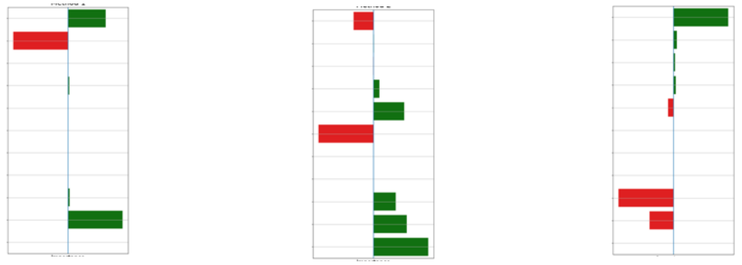
Interview Structure

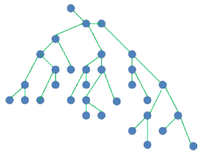
RESULTS

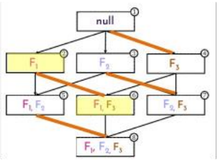
Qualitative Validation

Course pairs that show differences in learning behavior

M1 M2 M3







LIME SHAP

$F_1, F_2, F_3, F_4, \ldots$
$F_{42}, F_{43}, F_{44}, F_{45}$

confounder

25+ expert interviews with educators

Trusting the Explainers: Teacher Validation of Explainable Artificial Intelligence for Course Design. Swamy V., et al. (LAK 2023) – Honorable Mention

HREC 065-2022/27.09.2022

Expert Trust

RESULTS

- diverse prior perceptions of what factors enable student success and failure
- trust explanations that aligned with their beliefs (46.8%) (which led to significant disagreement across experts)
- only 3 chose the same method for both courses in the pair

Main Takeaways

EVALUATING THE EXPLAINERS

TRUSTING THE EXPLAINERS

Explainability methods, systematically,

do not agree

on which features are important for predictions

…. and neither do human experts

Where do we go from here?

Interpretable-by-design neural nets!

1) **MultiModN** – Multimodal Interpretability
2) **InterpretCC** – Mixture of Experts Interpretability
3) **iLLuMinaTE** – role of LLMs in explainability

NeurIPS
2023

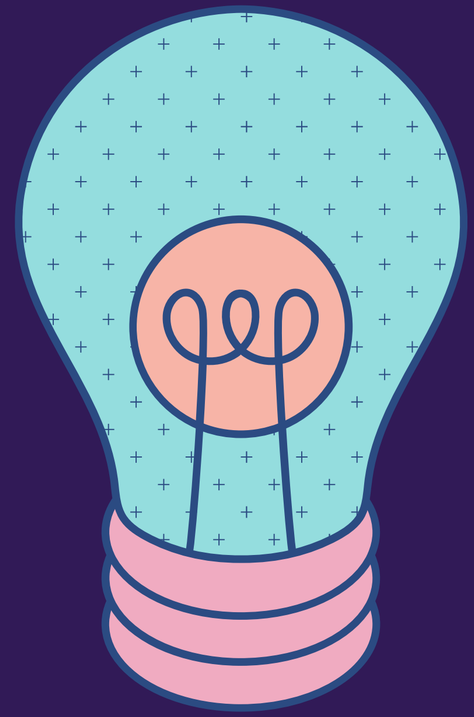
MultiModN

github.com/epfl-iglobalhealth/MultiModN

Multimodal, Multitask, Interpretable Modular Networks

NEURAL INFORMATION PROCESSING SYSTEMS

EPFL Yale




Vinitra Swamy [1*]


Malika Satayeva [1*]


Jibril Frej [1]


Thierry Bossy [1]


Thijs Vogels [1]


Martin Jaggi [1]


Tanja Käser [1*]


Mary-Anne Hartley [1,2*]

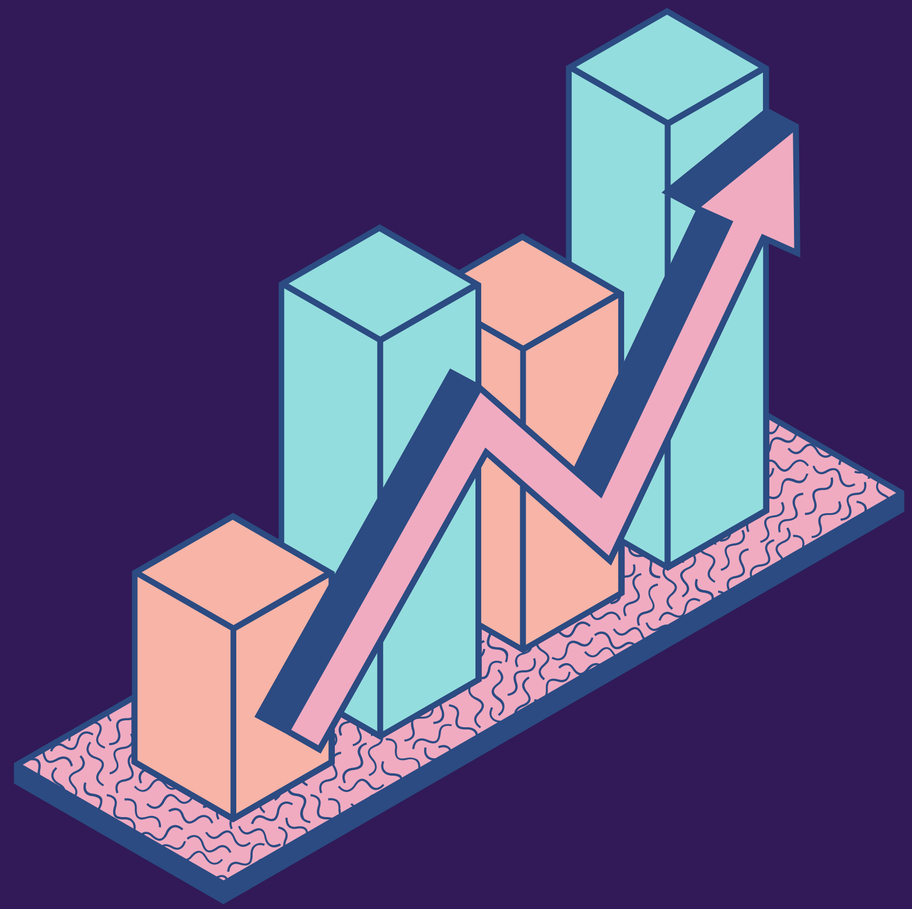
Multimodality

- synergistic predictive potential
- inputs with drastically varying sizes (i.e. images, text, sound)

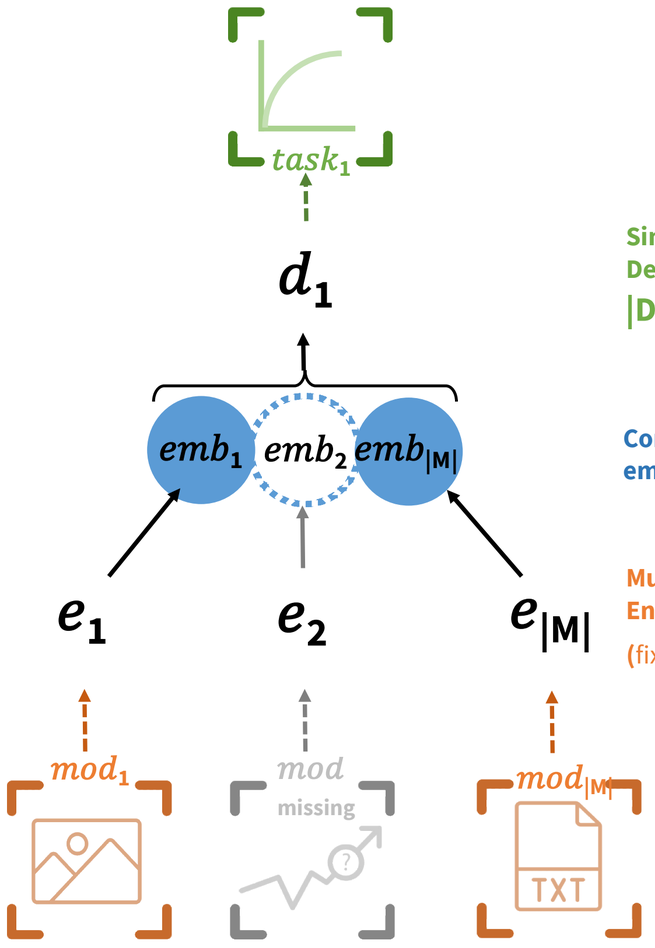
Current Approaches

fuse representations in parallel

- limits interpretability
- dependency on modality availability



Traditional P-Fusion



task

Single-task Decoder
|D| = 1

decoders

Concatenated embedding

embedding

Multimodal Encoders
(fixed number)

encoders

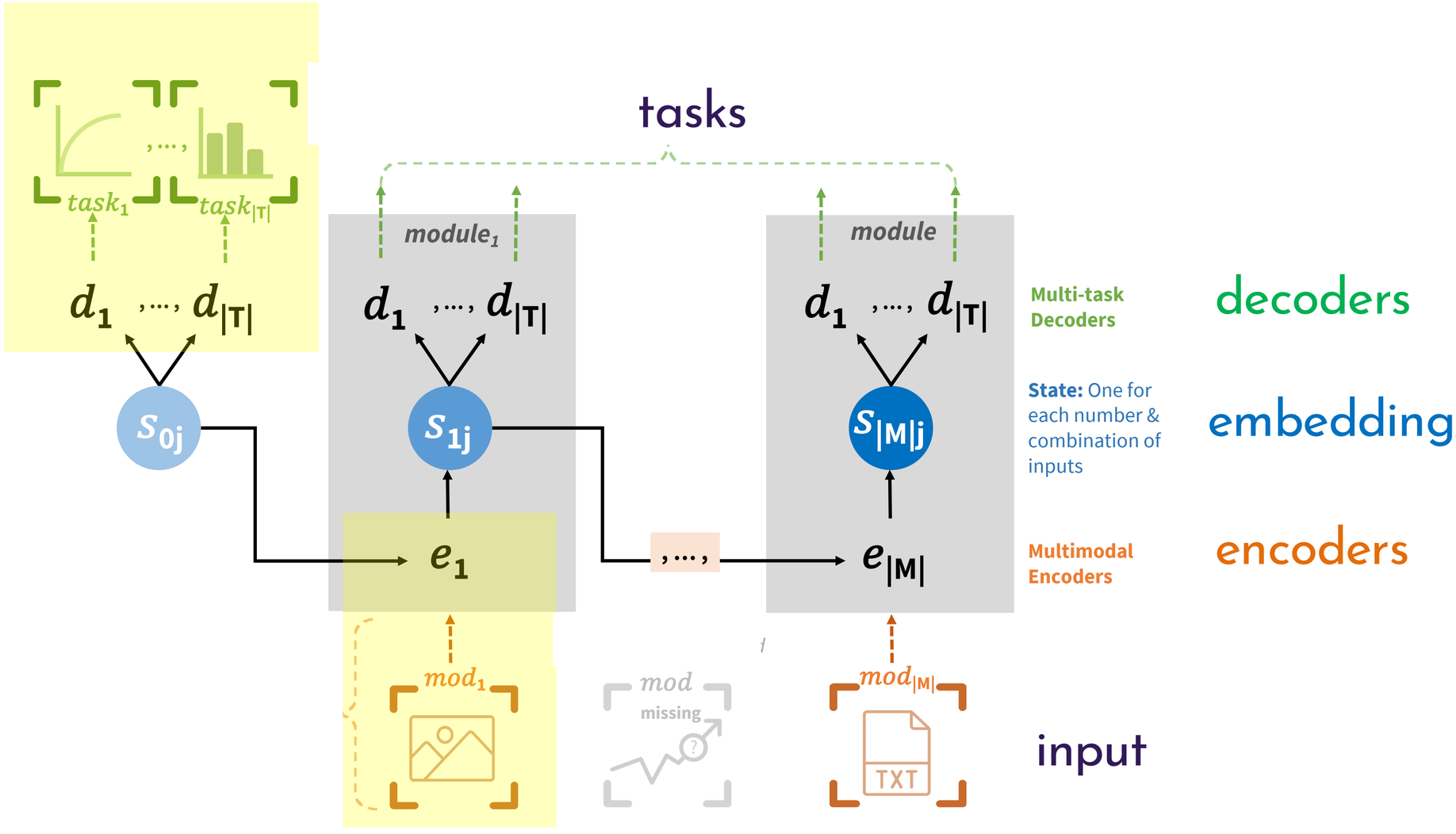
input

fusion in parallel

MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks. Swamy V. et al. (NeurIPS 2023)

Can we address these limitations of current multimodal architectures with sequential fusion?

MultiModN



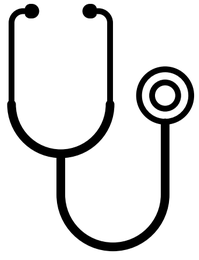
MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks. Swamy V. et al. (NeurIPS 2023)

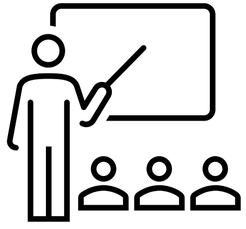
Real World Multimodal Settings

MULTIMODN

Evaluated on 10 real world tasks across multimodal healthcare (MIMIC IV + CXR-JPG), education (EPFL MOOC), and weather (Weather2k) benchmarks

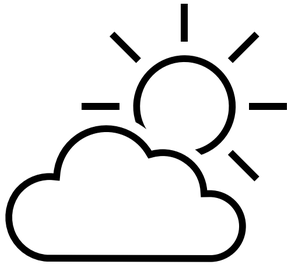
921 patients in tertiary care in a Boston hospital





5,611 students in a MOOC with over 1 million interactions

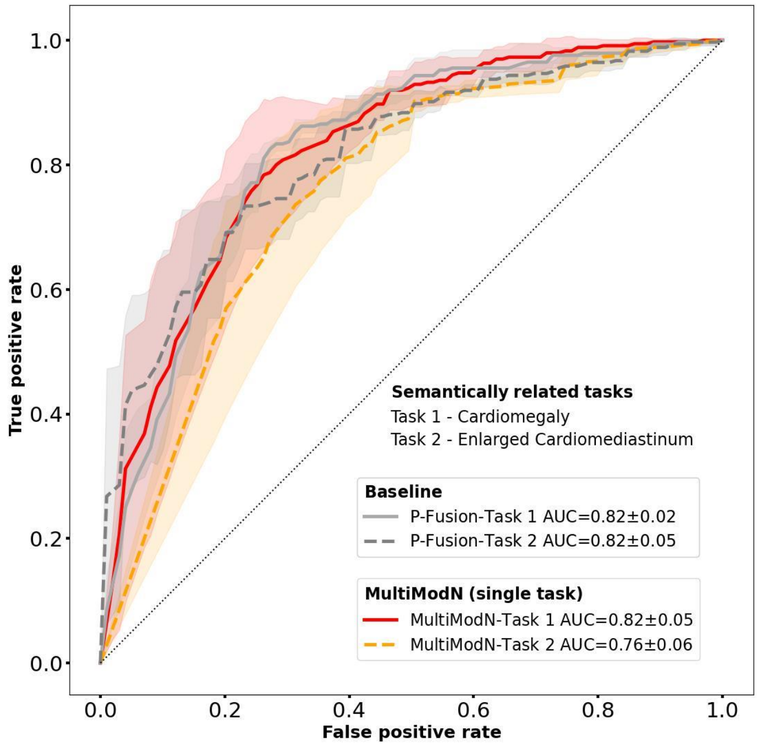
1866 ground weather stations covering 6 million km^2



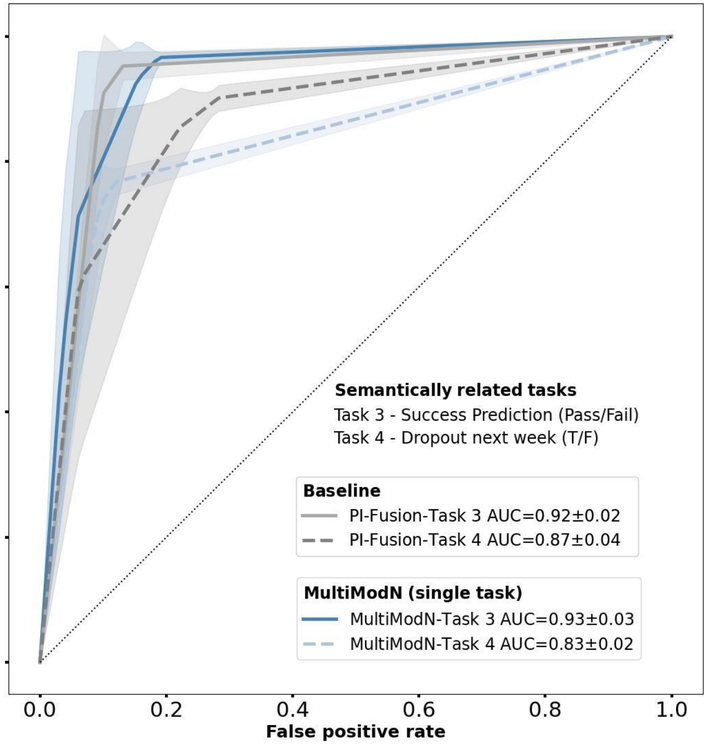
MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks. Swamy V. et al. (NeurIPS 2023)

Performance

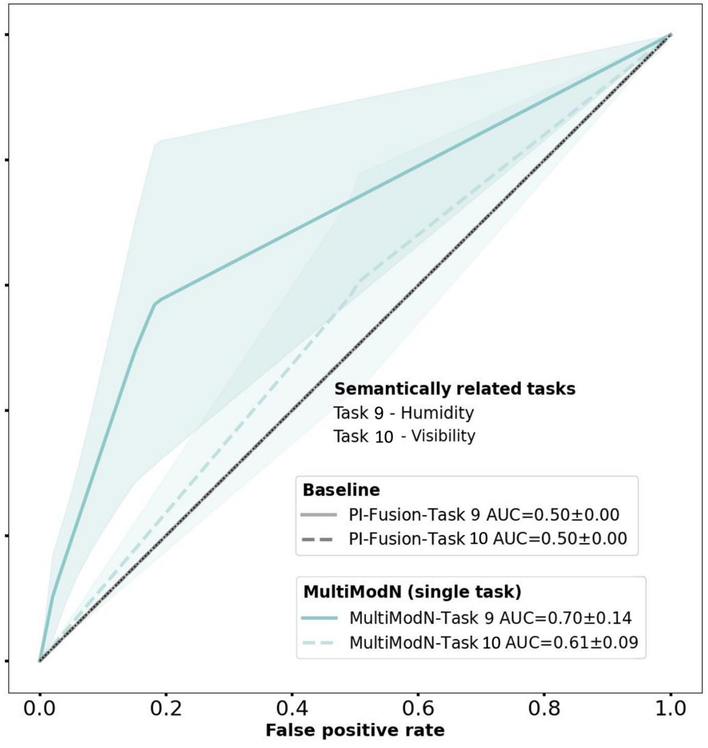
MULTIMODN VS. P-FUSION



MIMIC



EDU



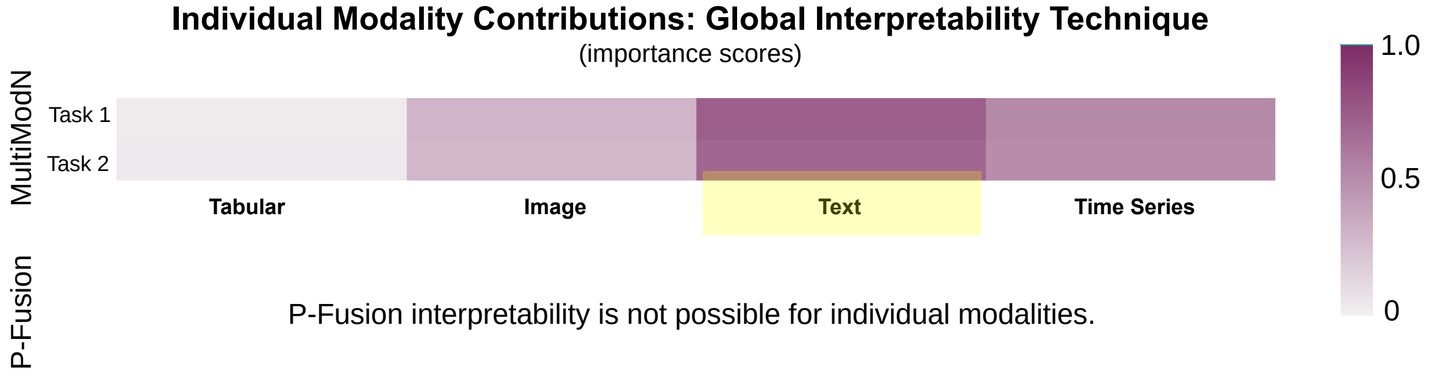
Weather2k

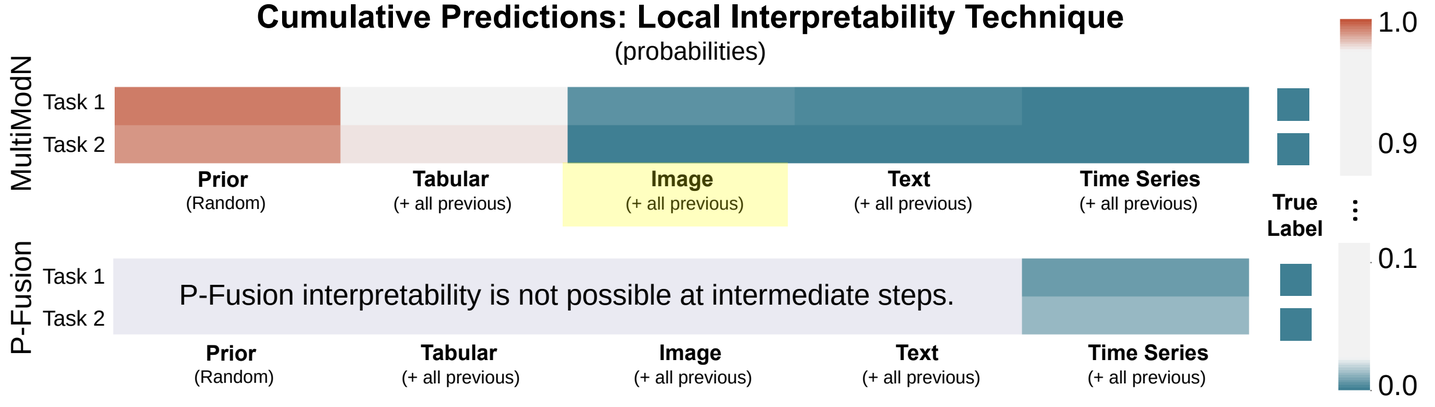
MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks. Swamy V. et al. (NeurIPS 2023)

Interpretable-by-design: MIMIC

MULTIMODN VS. P-FUSION

Which modality is the most important for this task?





Which modality caused the decision to flip?

MultiModN has modality-specific global (IMC) and local (CP) model explainability.

MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks. Swamy V. et al. (NeurIPS 2023)

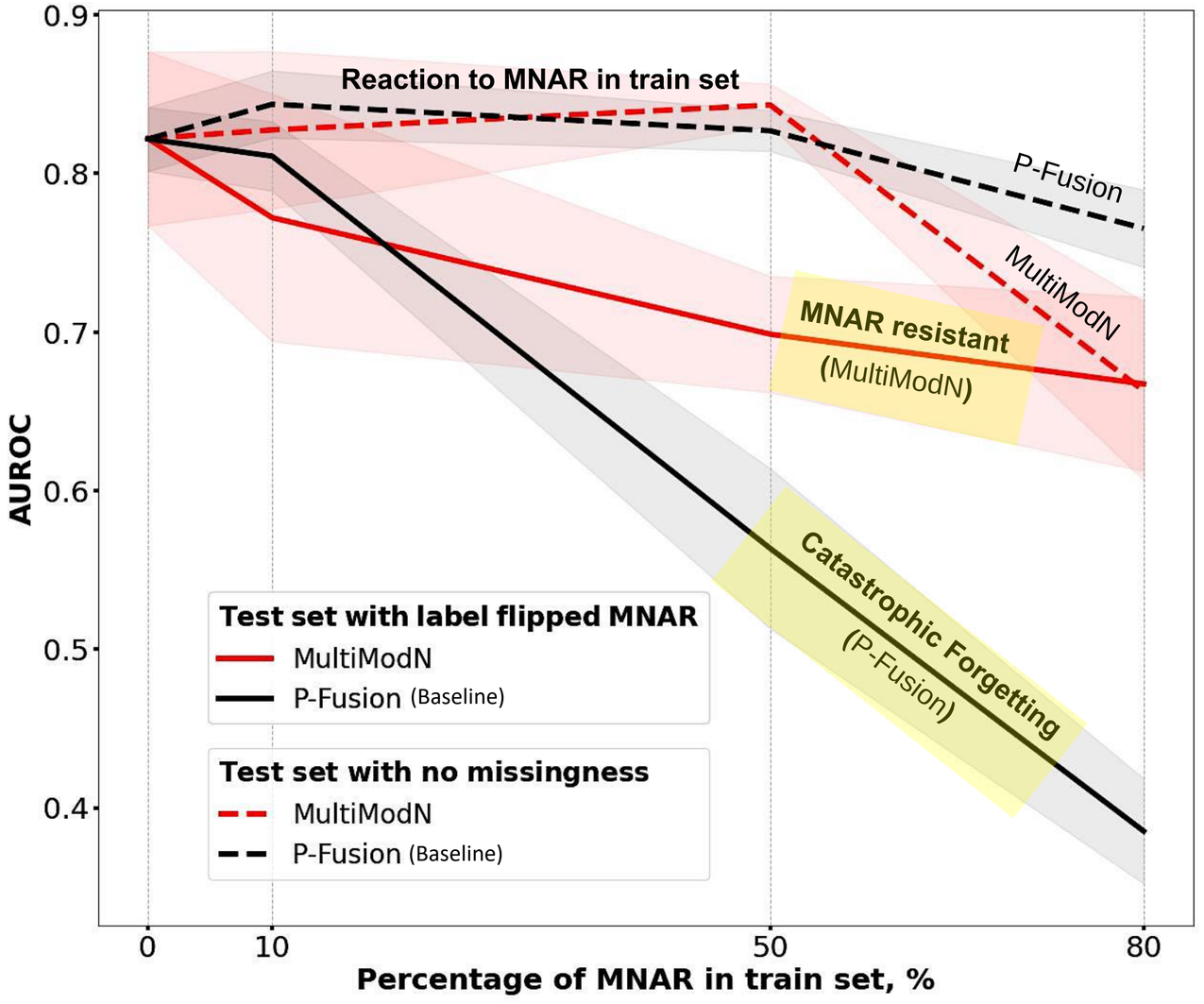
Robust to Missingness

MULTIMODN VS. P-FUSION

MultiModN is robust to bias from missing input modalities.

P-Fusion exhibits catastrophic MNAR failure when missingness patterns change.

Especially relevant in low-resource settings.



✨From the Lab to Real Life✨

MULTIMODN

Prof. Annie Hartley
(MD, PhD)

CHUV

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HÔPITAL UNIVERSITAIRE DE BERNE

clinical studies using MultiModN for pneumonia and tuberculosis diagnosis in **low resource settings**

- trained more than 100 doctors to collect multimodal data (images, ultrasound)
- recruiting 1000s of patients in South Africa, Tanzania, Namibia and Benin

MultiModN-Multimodal, Multi-Task, Interpretable Modular Networks. Swamy V. et al. (NeurIPS 2023)

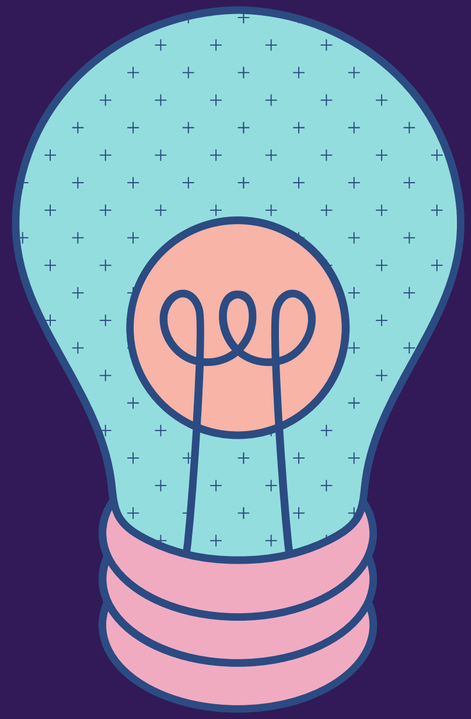
Under Review

InterpretCC

github.com/epfl-ml4ed/InterpretCC

Intrinsic, User-Centric Interpretability through Global MoE

EPFL




Vinitra Swamy


Syrielle Montariol


Julian Blackwell


Jibril Frej

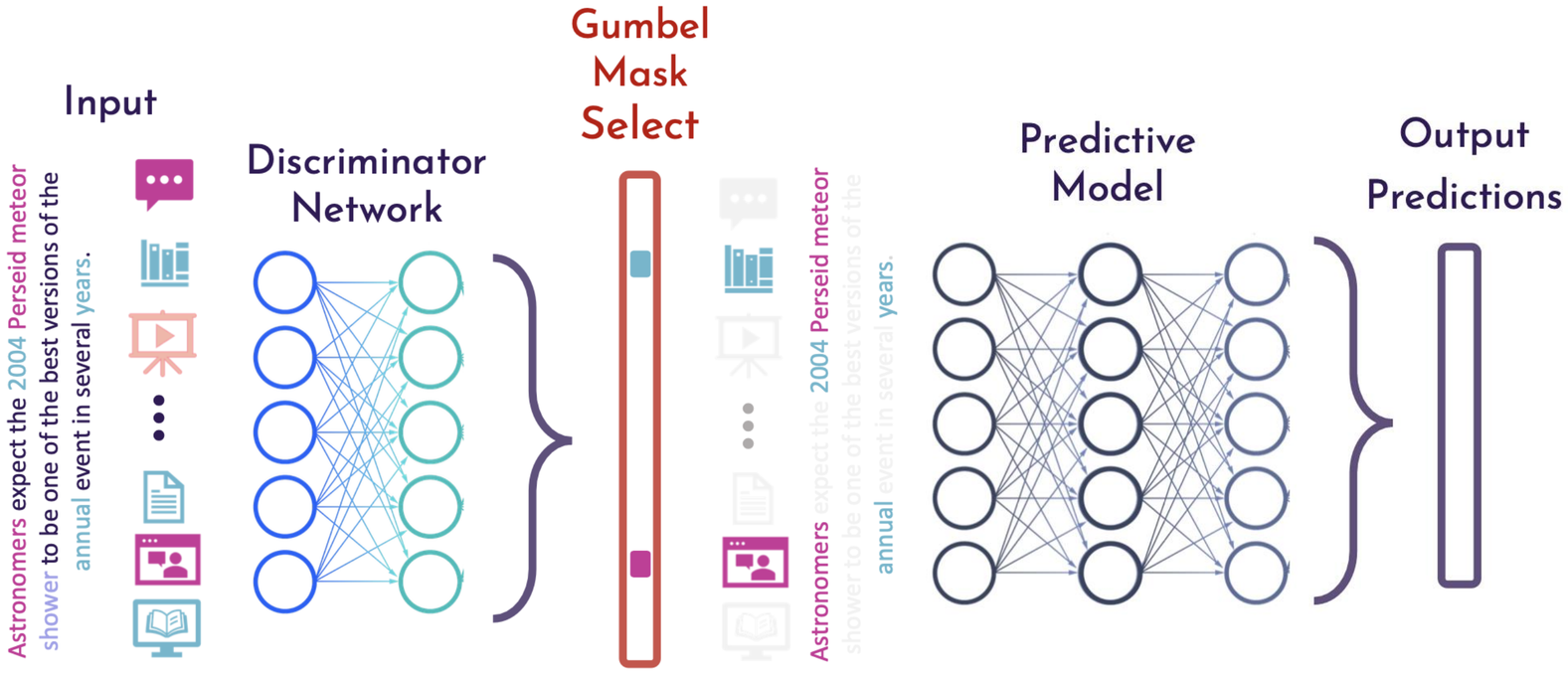

Martin Jaggi


Tanja Käser

InterpretCC: Feature Gating for Interpretability

Swamy, Montariol, Blackwell, Frej, Jaggi, Käser

Adaptive Feature Gating – accuracy vs. interpretability tradeoff

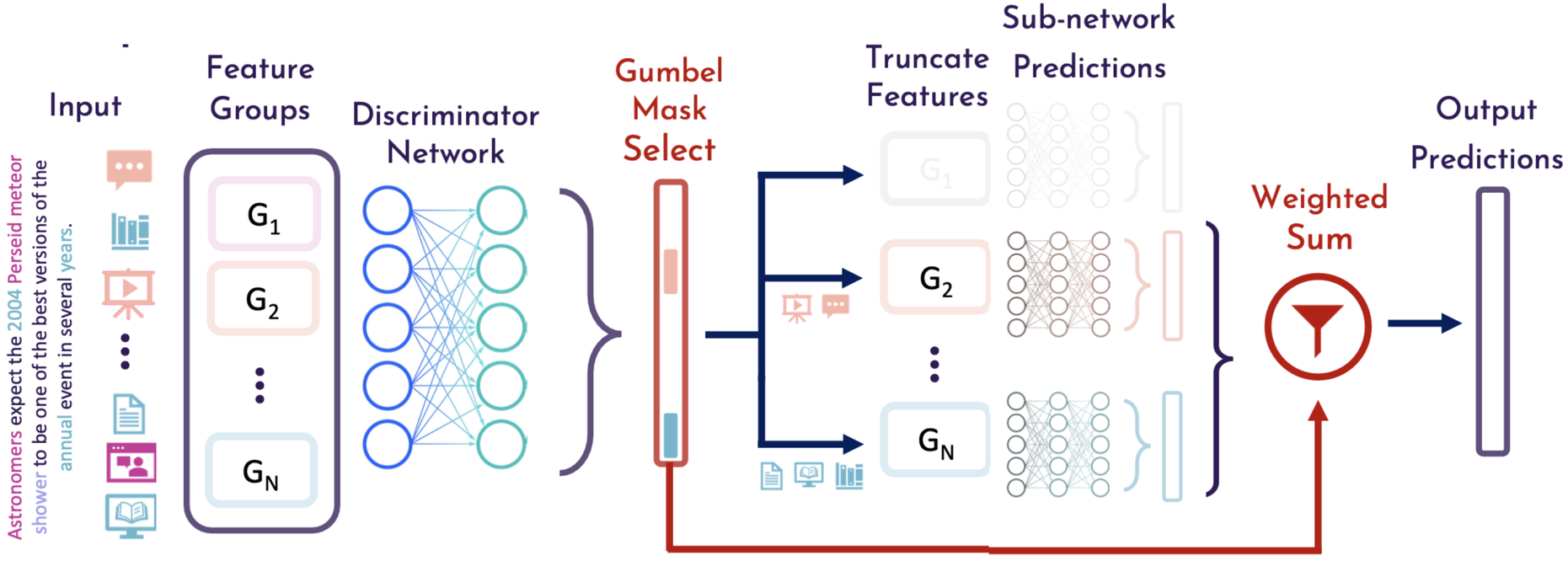


InterpretCC: Intrinsic, User-Centric Interpretability through Global Mixture of Experts. Swamy V. et al. (Under Review at ICLR 2025)

InterpretCC: Mixture-of-Experts for Interpretability

Swamy, Montariol Blackwell, Frej, Jaggi, Käser

Filter the feature space and send relevant parts to relevant experts

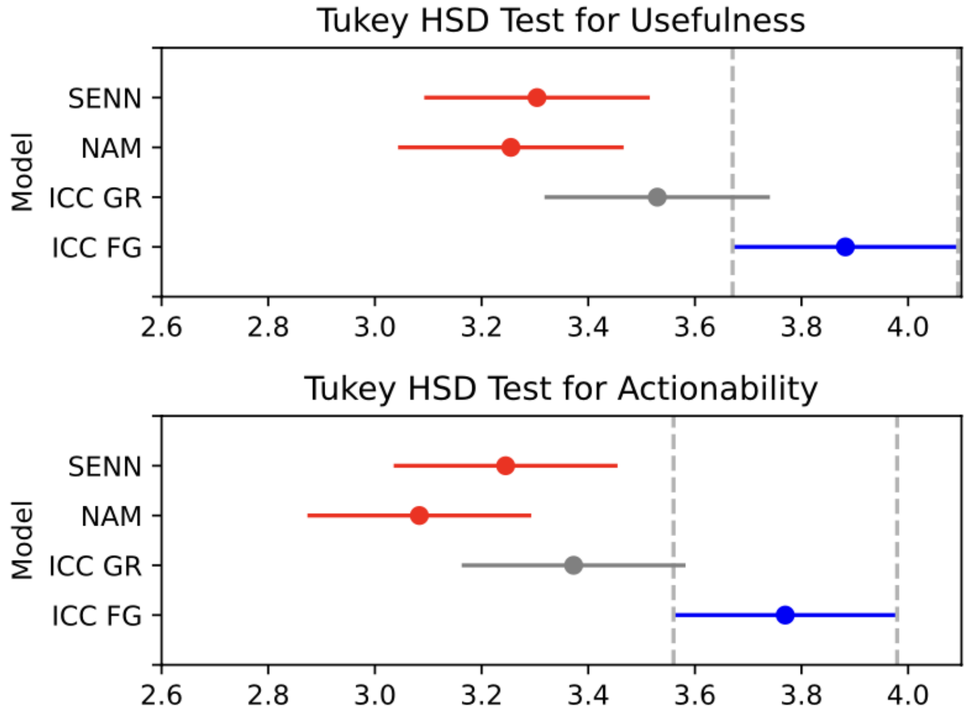


InterpretCC: Results

Swamy, Montariol Blackwell, Frej, Jaggi, Käser

1) Maintains or improves model performance across 8 different benchmark datasets (text, time-series, tabular)

2) Preferred by 56 teachers over other interpretable-by-design approaches

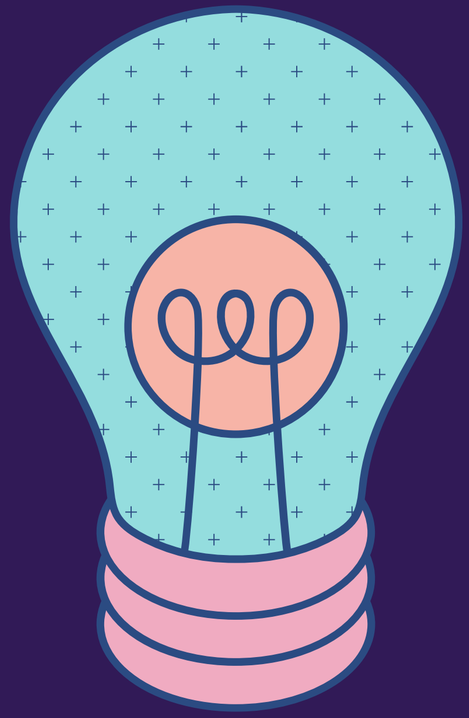


Under Review

iLLuMinaTE

github.com/epfl-ml4ed/illuminate

From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback




*Vinitra Swamy**


*Davide Romano**


Bhargav Desikan


Oana-Maria Camburu


Tanja Käser

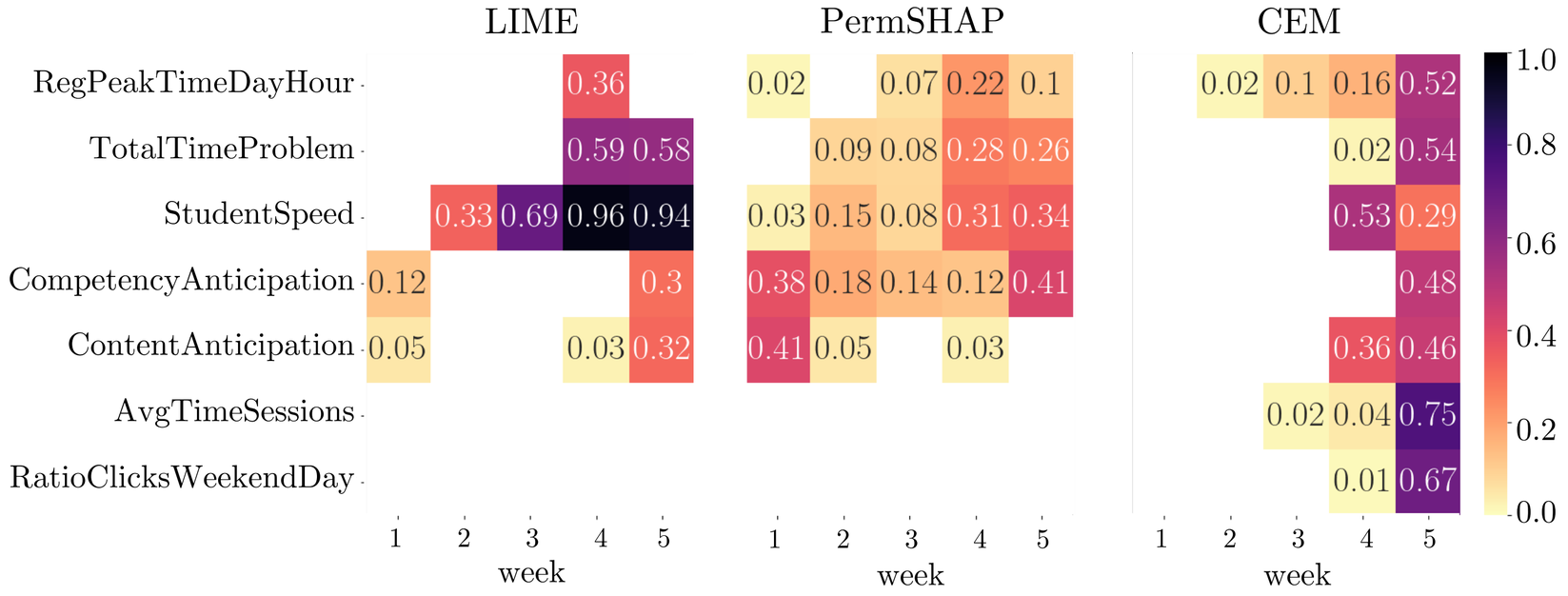
Processed Output of an Explainer

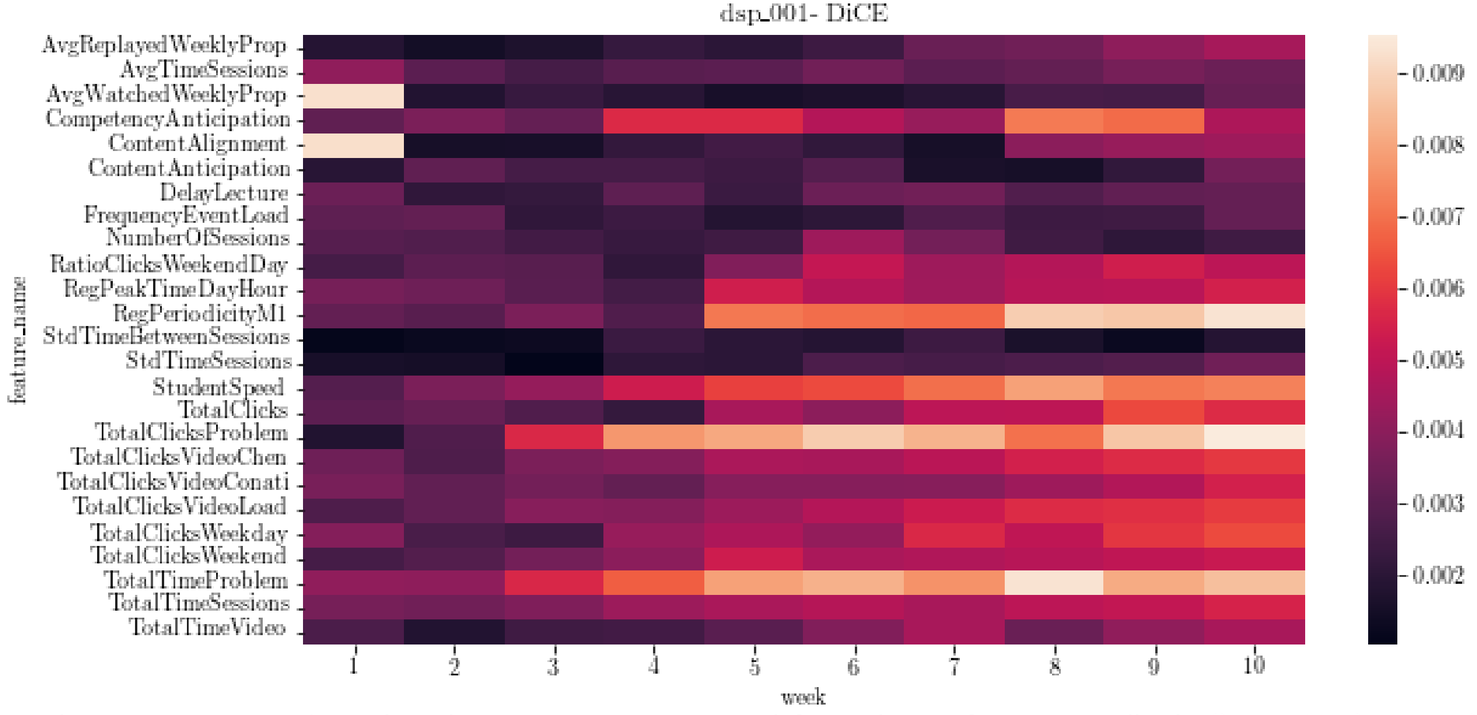
	exp_num	total_clicks_lnWeek1	number_sessions_lnWeek1	time_in__video_sum_lnWeek1	time_in__problem_sum_lnWeek1
66	1575	-4.56E-09	-2.21E-10	0.074035813	1.3699654119458948e-09
67	5252	-5.60E-10	-2.21E-10	2.1067425364992842e-11	-2.06E-10
68	881	1.7555813192071668e-09	-8.83E-10	0.14653082042868498	-2.65E-10
69	2683	-3.42E-09	-2.21E-10	0.24476348871729897	1.264484322804904e-09
70	16963	-8.50E-10	0	0.045331784	-1.61E-09
71	5931	6.354141102171695e-10	1.3576282653637861e-08	0.17988949383993513	-7.48E-10
72	12145	0	0	0.9422763586044312	0
73	11999	0	0	0.03497038	0
74	1571	6.909329608451031e-09	-8.83E-10	0.13568544287717593	1.6364350777231529e-09
75	2220	-5.60E-10	0	0.057424471615632494	0
76	9184	0	0	0.025690399	0
77	9592	-2.28E-09	-2.21E-10	0	-1.84E-10
78	12309	5.789162378644352e-10	-2.21E-10	0	-4.92E-10
79	5394	-2.28E-09	0	0.026910097	0
80	730	-2.28E-09	-2.21E-10	0.029592504314488427	8.244599122332608e-13
81	3318	2.315664951457741e-09	6.7881413268189306e-09	0.014542981	-1.48E-09
82	290	-1.12E-09	0	-3.32E-09	0
83	1959	5.789162378644352e-10	0	1.8329978368480937e-09	4.3817753245245505e-10
84	1424	-1.70E-09	-4.42E-10	-4.82E-09	4.4265635254503444e-10
85	12925	0	0	0.021664523	0
86	11139	0	0	0	0
87	12603	0	0	0	0
88	16987	1.1578324757288705e-09	-2.21E-10	0	-1.46E-10
89	10284	0	0	0	0
90	1316	1.7555813192071668e-09	-8.83E-10	0.015449408282339427	-9.00E-10
91	2601	1.447290594661088e-10	0	0	0
92	2851	-2.84E-09	-4.42E-10	0.037342517841102124	0
93	16707	2.894581189322176e-10	0	0	9.635302451738159e-11
94	11108	0	0	0	0
95	5846	1.1578324757288705e-09	-2.21E-10	-3.86E-11	0
96	16538	1.447290594661088e-10	0	0	0
97	14230	0	0	0	0

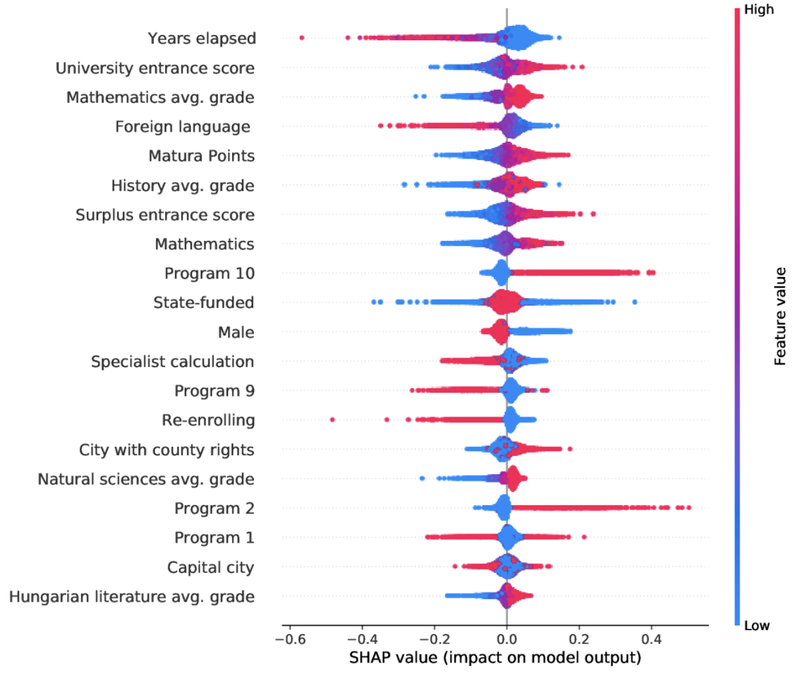
10 weeks x 45 fts
= 450 columns

From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

Heavily Processed Visuals

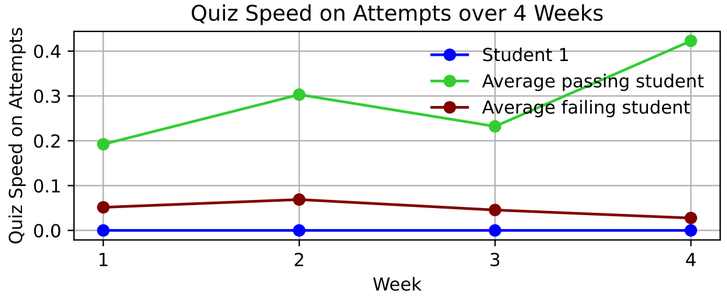


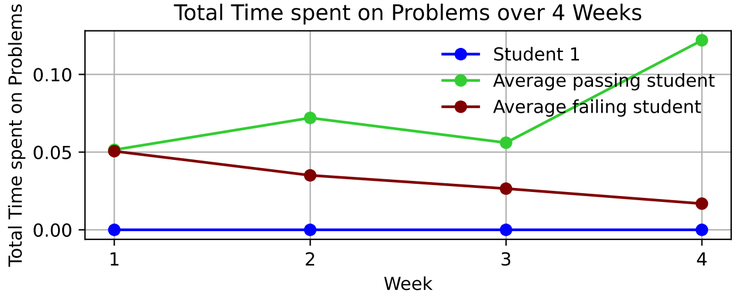


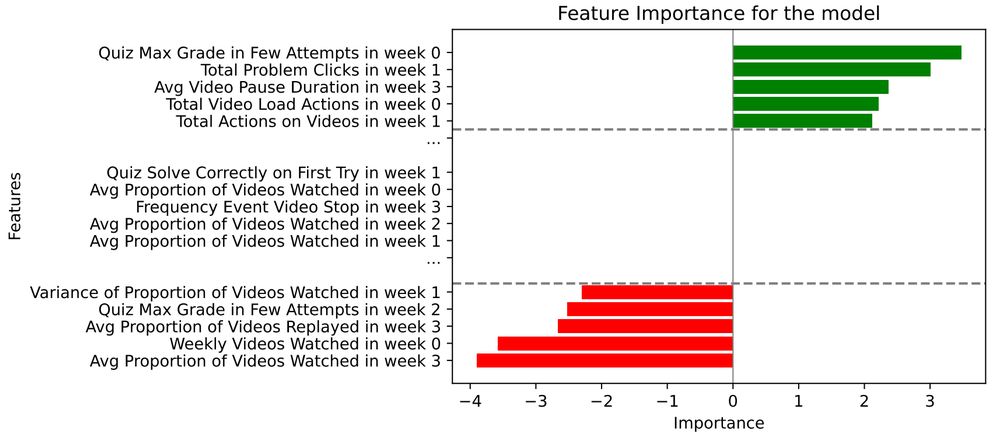


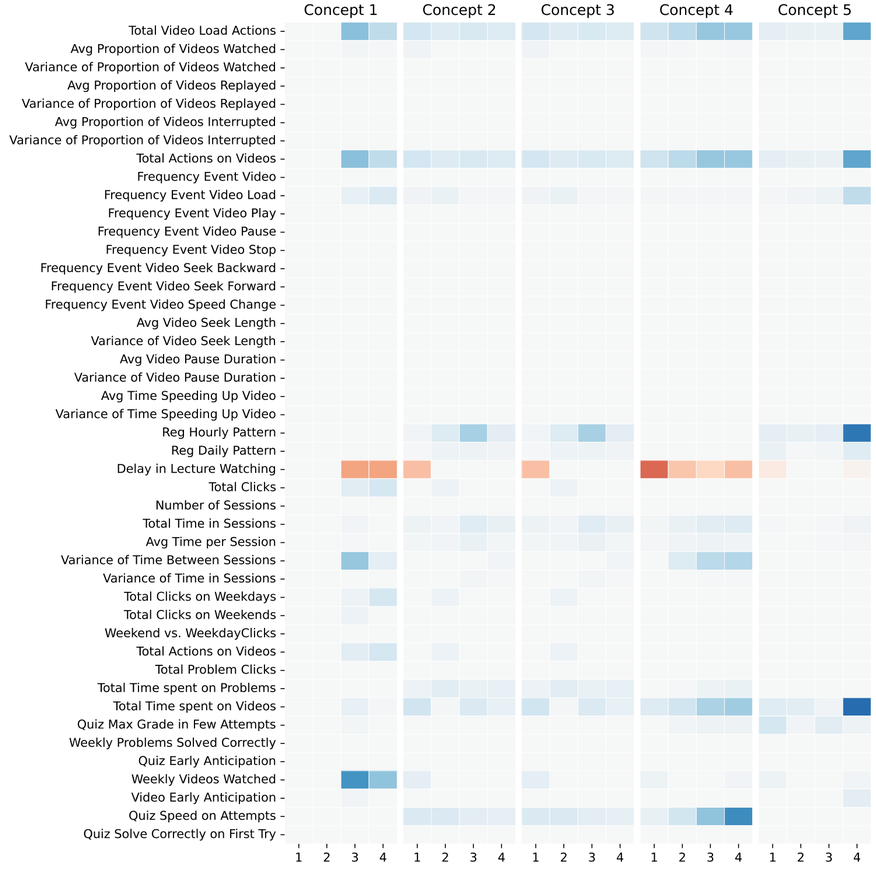
From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

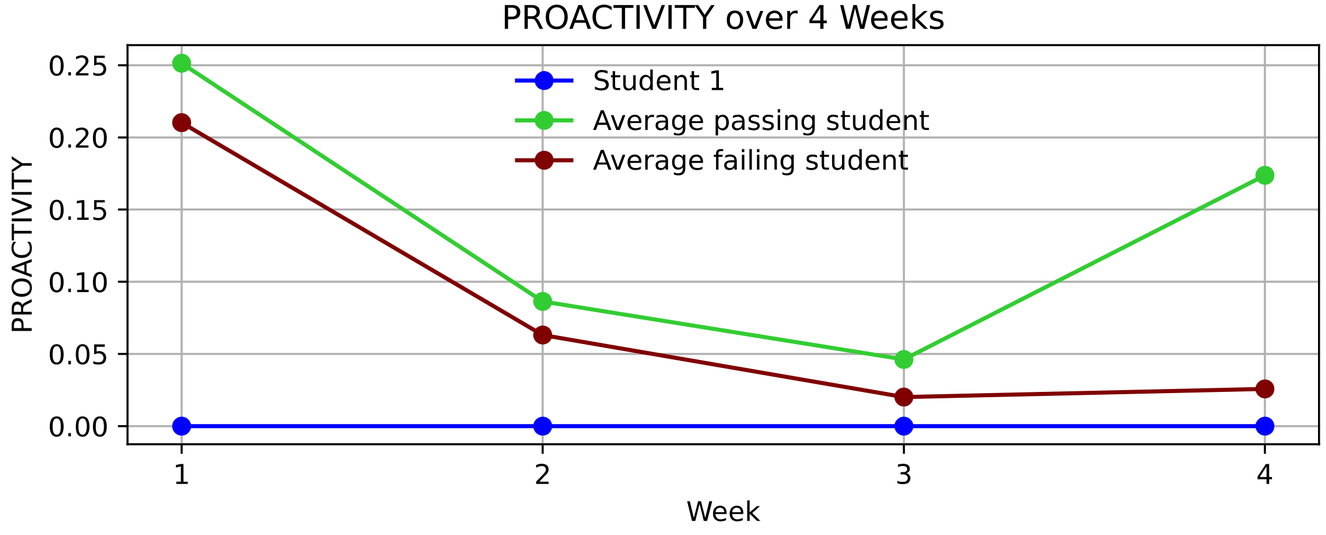
Heavily Processed Visuals





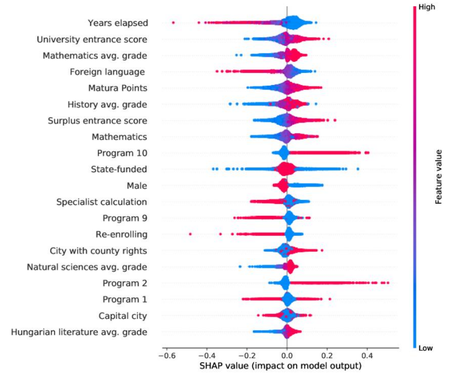


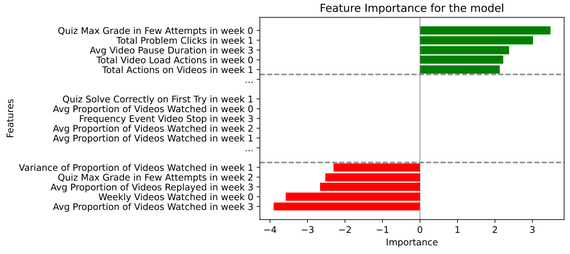




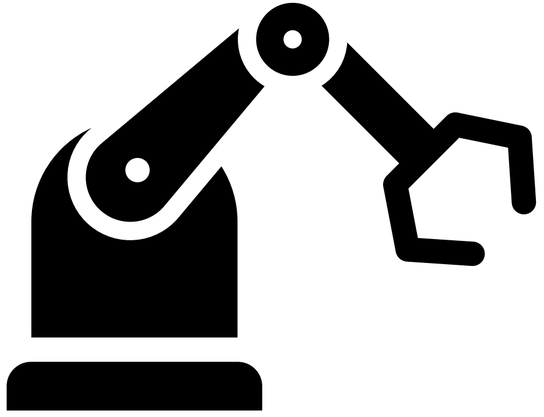
From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

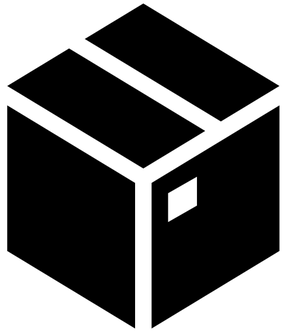
Select what's important





	exp_num	total_clicks_lnWeek1	number_sessions_lnWeek1	time_in_..video_sum_lnWeek1	time_in_..problem_sum_lnWeek1
66	1575	-4.56E-09	-2.21E-10	0.074035813	1.369965411945858e-09
67	5252	-5.60E-10	-2.21E-10	5.106742536490842e-11	-2.06E-10
68	881	1.755581319207l668e-09	-8.83E-10	0.146530820428684698	-2.65E-10
69	2683	-3.42E-09	-2.21E-10	0.244763488717298997	1.264484322804904e-09
70	16963	-8.50E-10	0	0.045331784	-1.61E-09
71	5931	0.3541411021716l95e-10	1.3576282653637861e-08	0.179889493839393513	-7.48E-10
72	12145	0	0	0.942276358604431 2	0
73	11999	0	0	0.03497038	0
74	1571	0.9093296084510 3 1e-09	-8.83E-10	0.135685442877175 93	6.36435077723152 9e-09
75	2220	-5.60E-10	0	0.057424471615632494	0
76	9184	0	0	0.025690399	0
77	9592	-2.28E-09	-2.21E-10	0	-1.84E-10
78	12309	5.789162378644352e-10	-2.21E-10	0	-4.92E-10
79	5394	-2.28E-09	0	0.026910097	0
80	730	-2.28E-09	-2.21E-10	0.029592504314488427	1.244599122332608e-13
81	3318	2.3156649514577 4 1e-09	5.788141326818930 6e-09	0.014542981	-1.48E-09
82	290	-1.12E-09	0	-3.32E-09	0
83	1959	5.789162378644352e-10	0	0.832997836848093 7e-09	1.381775324524550 5e-10
84	1424	-1.70E-09	-4.42E-10	-4.82E-09	5.426563525450344 4e-10
85	12925	0	0	0.021664523	0
86	11139	0	0	0	0
87	12803	0	0	0	0
88	16987	1.157832475728870 5e-09	-2.21E-10	0	-1.46E-10
89	10284	0	0	0	0
90	1316	1.755581319207l668e-09	-8.83E-10	0.015449408282339427	-9.00E-10
91	2601	1.447290594661088e-10	0	0	0
92	2851	-2.84E-09	-4.42E-10	0.037342517841102124	0
93	16707	2.894581189322176e-10	0	0	0.635302451738159e-11
94	11108	0	0	0	0
95	5846	1.157832475728870 5e-09	-2.21E-10	-3.86E-11	0
96	16538	1.447290594661088e-10	0	0	0
97	14230	0	0	0	0



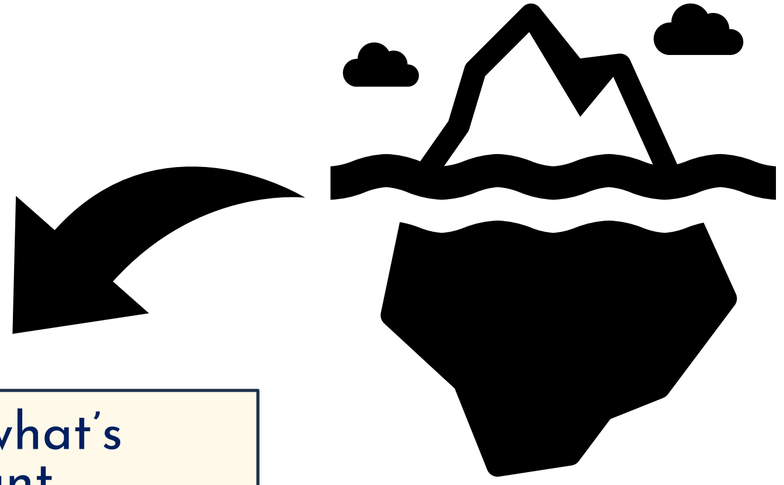


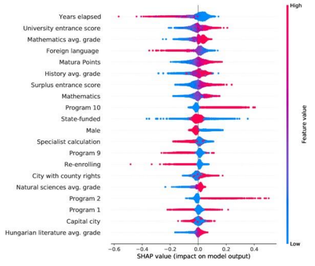
Package it up nicely

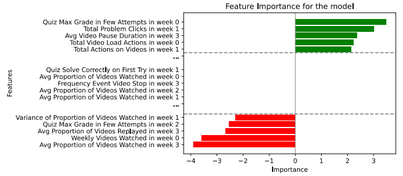
students are happy*!

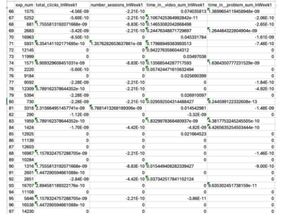
*with their improved learning outcomes

From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

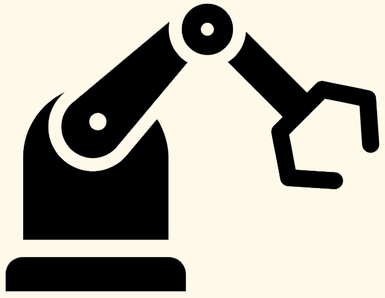








Select what's important



Package it up nicely

students are happy!

From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

Leverage theories of explanation from philosophy and social sciences to define prompting strategies

Hempel

DN model

Statistical Relevance

Relevance Selection

C.S. Peirce

Abductive reasoning

Abnormal Conditions

Malle

Social attribution model

Necessity Robustness

Base Contrastive

P. Lipton

Contrastive explanation

RaR + Contrastive

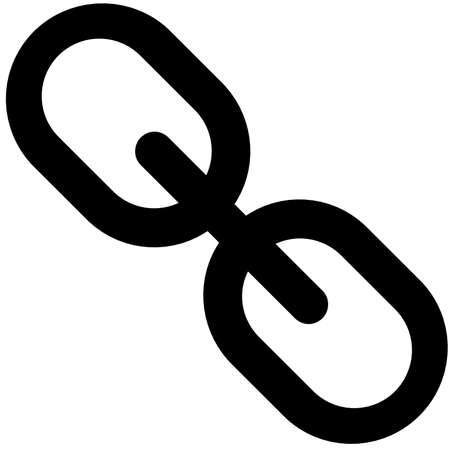
Lombrozo

Cognitive processes of explanation

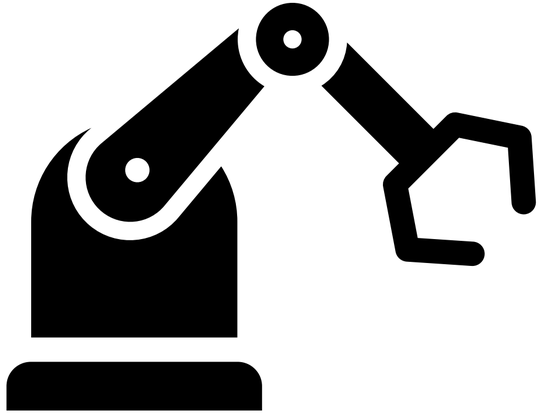
Pearl's Explanation

Chain-of-Thought

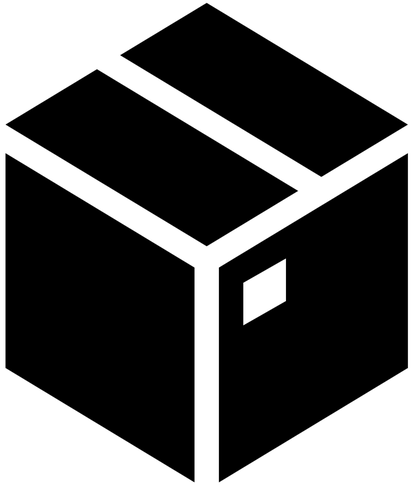
Cognitive Processes of Explanation



Causal Connection



Explanation Selection



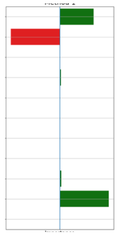
Explanation Presentation

Miller, Tim. "Explanation in Artificial Intelligence: Insights from the Social Sciences."
From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

Explanation Selection Prompting Strategy

Model Prediction

Explanation



+

Feature values

Course context

DSP 1

Theory Instrutions

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).

Language Model

GPT4O / Llama3 70B / Gemma2 9B

Text Explanations

Explanation Presentation

Student Feedback Theory

Where am I going?
How am I doing?
Where to next?

Grice's Maxims

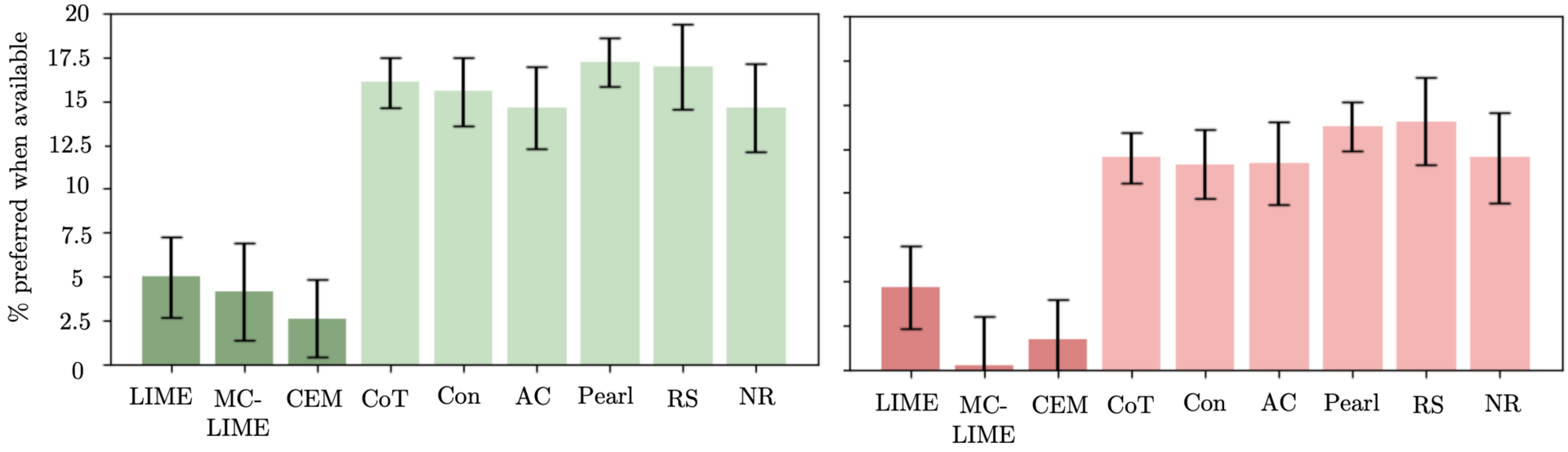
Text Explanations

Large Language Model

GPT4O / Llama3 70B / Gemma2 9B

iLLuMinaTE Explanations

Results



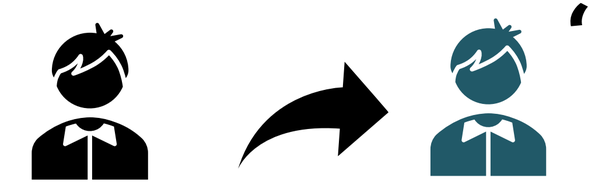
Students prefer iLLuMinaTE explanations significantly over expert processed post-hoc explanations!

text >> visual part of the explanation

From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback . Swamy V., Romano D., et al. (Under Review at AAAI 2025)

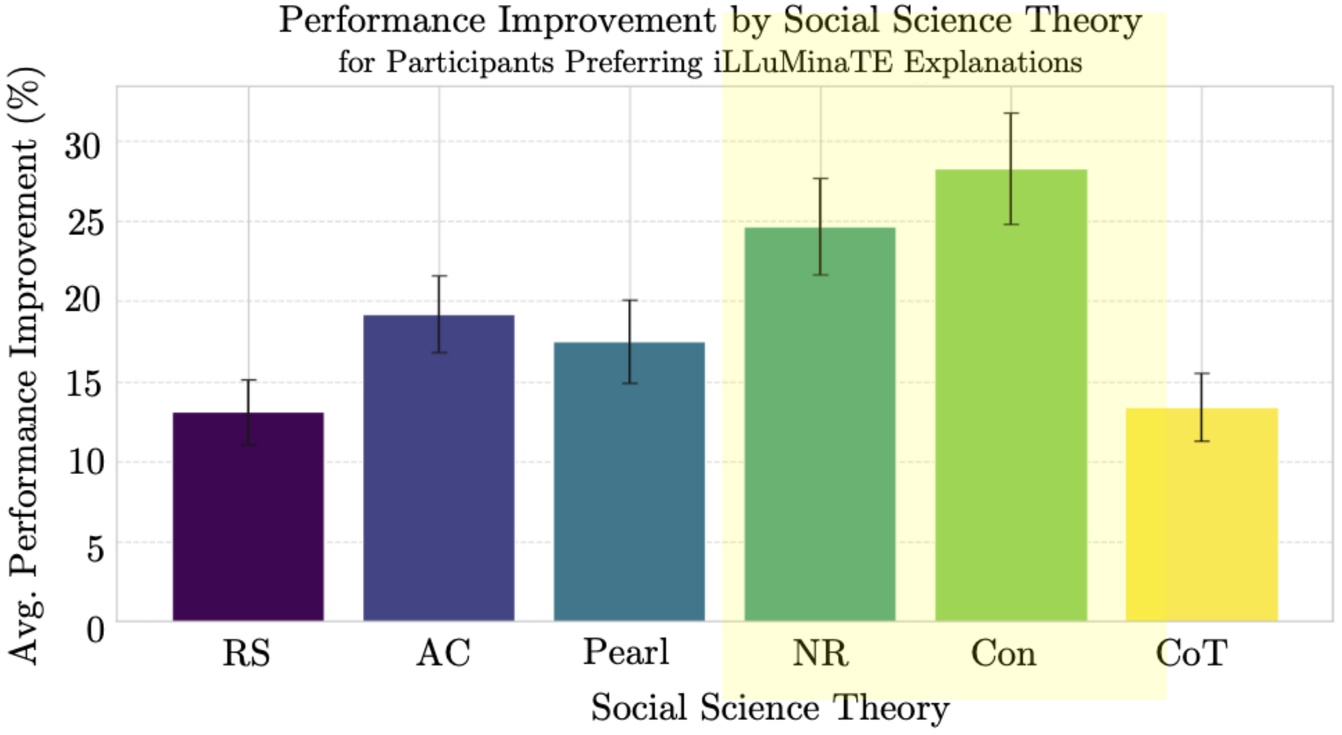
Actionability

students selected an action to take based on their preferred explanation



week 5

week 6 simulated intervention



Contrastive, Necessity Robustness

Main Takeaways

EVALUATING THE EXPLAINERS

TRUSTING THE EXPLAINERS

Post-hoc explainers have serious problems,
and there is no effective way to validate them

Main Takeaways

MULTIMODN, INTERPRETCC, ILLUMINATE

With interpretable-by-design NNs,
guaranteed interpretability
does not have to come at the cost of performance
or human-understandability

Thank you!

Vinitra Swamy

vinitra.swamy@epfl.ch

vinitra.github.io

twitter: @vinitra_s