LAMP: Extracting Local Decision Surfaces From Large Language Models

[1]Ryan Chen*, [1]Youngmin Ko*, [1]Catherine Cho, [1]Zeyu Zhang,
[2]Sunny Chung, [2]Mauro Giuffré, [3]Dennis L. Shung, [1]Bradly Stadie

[1]Northwestern University, [2]Yale School of Medicine, [3]Mayo Clinic

Northwestern | WEINBERG COLLEGE OF ARTS & SCIENCES

Problem Statement

- LLMs can produce “justifications” on their decisions

Sentiment Classification Example

Review: “The movie was so fun. It was full of great actors and scenes, but the plot was not so creative.”

LLM classifies as positive, reporting:

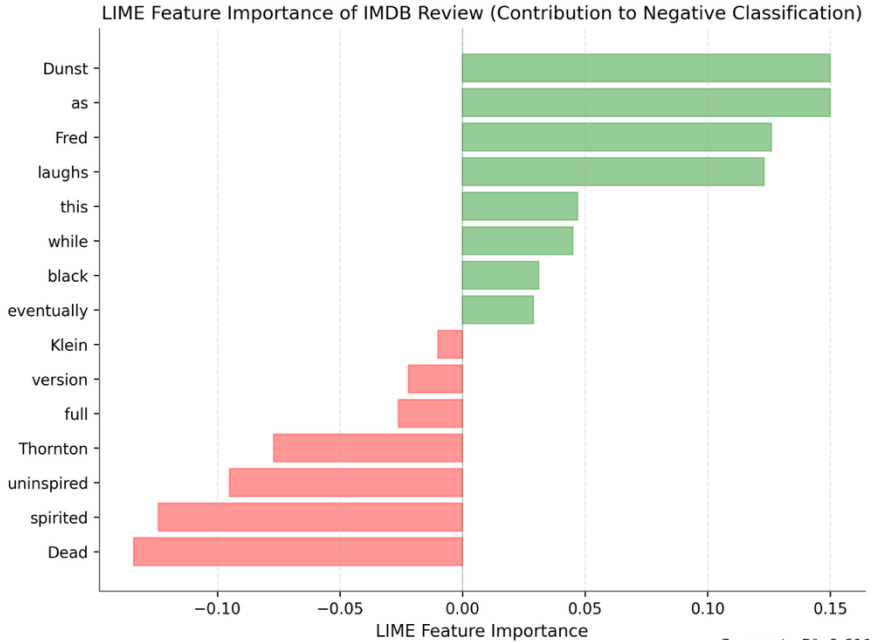
1. Feature 1: “movie was fun”
2. Feature 2: “great actors and scenes”
3. Feature 3: “plot was not creative”

- **Question**: Can these self-reported factors be used to model the LLM’s local decision behavior?

Are there any methods that help?

- Canonical interpretability methods exist: LIME, SHAP, ANCHOR, etc.
- However, most of the methods work in token-level
- But deleting individual words does not always correspond to a meaningful change in the model's stated rationale.

  → LAMP instead probes behavior in a human-readable explanation space.



LAMP: Local Attribution Mapping Probe

Sentiment Classification Example

Review: "The movie was so fun. It was full of great actors and scenes, but the plot was not so creative."

LLM classifies as positive, reporting:

1. Feature 1: "movie was fun" $\rightarrow$ importance: $+0.6$
2. Feature 2: "great actors and scenes" $\rightarrow$ importance: $+0.4$
3. Feature 3: "plot was not creative" $\rightarrow$ importance: -0.3

- Ask the LLM to classify an input and report explanatory factors with importance weights.
- Treat those weights as coordinates in an explanation space.
- Perturb the weights and observe how the model's reported probability changes.
- Fit a local linear surrogate to summarize the model's behavior.

Methods

Step 1: Extraction

- Prompt LLM to classify input (x) and report factors, $\mathrm{f} = (\mathrm{f\_1}, \cdots, \mathrm{f\_n})$, with weights, $w = (w_1, \cdots, w_n)$
- Repeat k times $\rightarrow \mathrm{k} * \mathrm{n}$ features total
- Meta-aggregation: consolidate into fixed factors; re-collect weights

Step 2: Perturbation

- Perturb: $\tilde{\mathbf{w}}^{(j)} = \mathbf{w_0} \odot \left(1 + \boldsymbol{\epsilon}^{(j)}\right)$
- Re-query LLM with same factors but new weights → get $p^{(j)}$

Step 3: Modeling

- Fit OLS: $\beta = \arg\min_{\beta} \|X\beta - y\|_2^2$

Results

- Four Datasets: IMDB, Pseudo-Harmful, HateBenchSet, Gastroenterology Clinical Dataset
- Four LLMs: GPT-4.1-mini, Gemini 2.5 Flash, Claude 3.5 Haiku, Mistral Large
- Evaluation Dimensions:
  - Quality of linear fit
  - Human Evaluation
  - Consistency of the Method in Natural Language

Quality of Linear Fit

Model	IMDB	PH	HateBS
GPT-4.1-mini	0.42 ± 0.03	0.38 ± 0.03	0.17 ± 0.02
Gemini 2.5 Flash	0.26 ± 0.03	0.25 ± 0.03	0.23 ± 0.03
Claude 3.5 Haiku	0.26 ± 0.03	0.21 ± 0.02	0.16 ± 0.01
Mistral Large	0.30 ± 0.03	0.25 ± 0.03	0.20 ± 0.02

- Perturbations in self-reported weight explain substantial variance in predictions
- GPT-4.1-mini consistently strongest
- Additional linearity analyses are included in the paper.

Human Agreement

Dataset	GPT-4.1-mini	Gemini	Claude	Mistral
IMDB	0.840	0.700	0.620	0.680
Pseudo-Harmful	0.780	0.729	0.600	0.660
HateBenchSet	0.600	0.633	0.680	0.640
Gastro (IAA=0.635)	0.697	0.624	0.645	0.461

- Human evaluators judged whether each coefficient's direction matched the expected direction of the factor.
- LAMP captures semantically meaningful patterns in the model's observable behavior.

Consistency of the Method

- **Question:**

  The surrogate might only work when we feed the model artificial weight perturbations.

- **Test:**

  Rewrite the original input to reflect the same factor changes, then test whether the LAMP surrogate predicts the model's new probability.

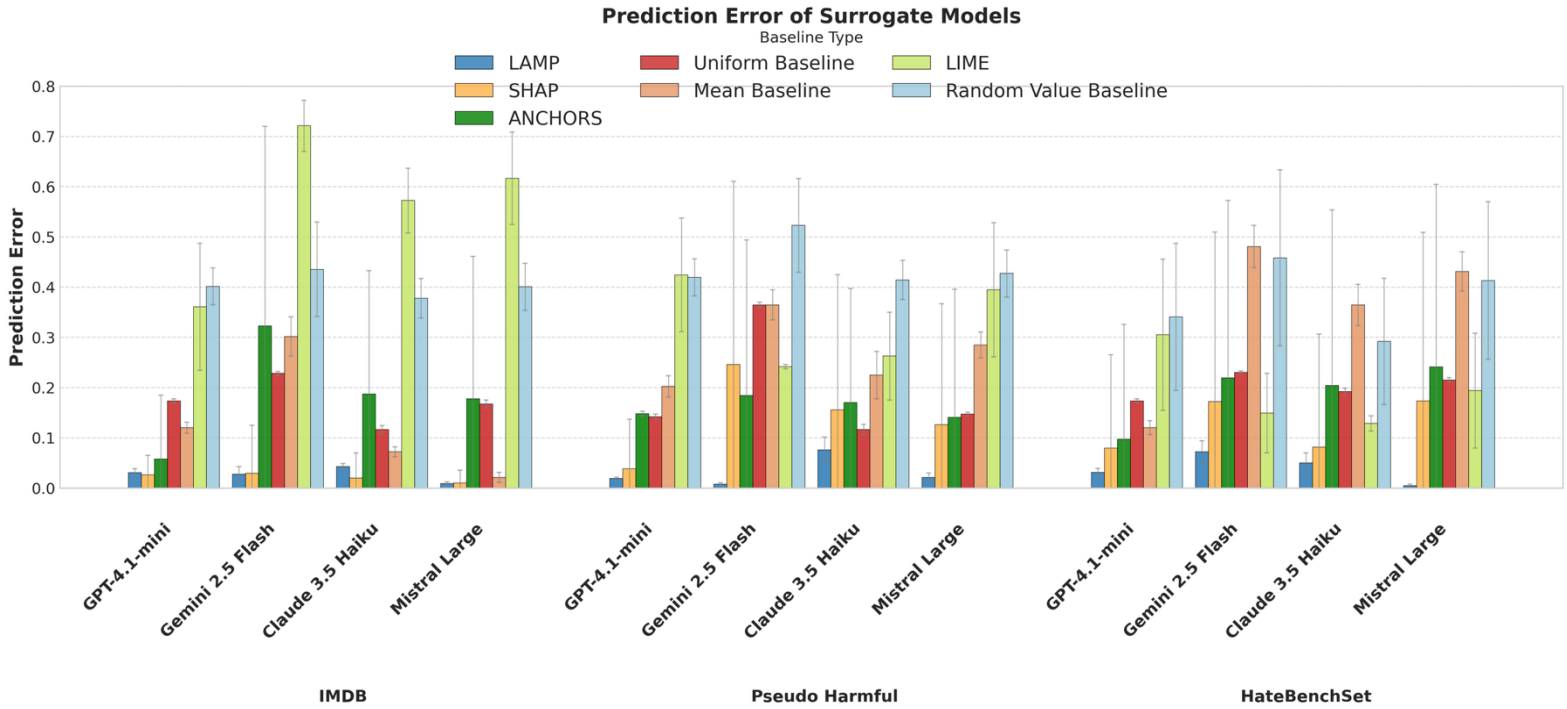
Original x

Dumb is as dumb does, in this thoroughly uninteresting, supposed black comedy [...] of ""The Three Amigos"", only without any laughs. [...] for black comedy to work, it cannot be mean spirited, which ""Play Dead"" is. What ""Play Dead"" really is, is a town full of nut jobs. Fred Dunst does however do a [...].

Modified x_{modified}

Dumb is as dumb does, in this somewhat engaging, supposed black comedy [...] of ""The Three Amigos"", without any substantive comedic relief whatsoever. [...] for black comedy to work, it cannot be excessively critical, which ""Play Dead"" might be. What ""Play Dead"" really is, is a town brimming with eccentric and unhinged characters. Fred Dunst does however do a [...].

Results



LAMP predicts rewritten-input probabilities with lower error than token-level baselines.

Key Takeaways

- **Explanation-space probing:** Uses stated factors as an audit space.
- **Black-box auditing:** Requires no gradients, logits, or activations.
- **Local coherence:** Finds locally structured decision behavior.
- **Expert transparency:** Produces coefficients people can inspect.

Thank you

- Poster Location: 177 (Poster Session 2)