What LLMs Explain Is Not What They Believe:
Evaluating Explanation Sufficiency Under Models' Own Input Beliefs

Nhi Nguyen, Shauli Ravfogel, Rajesh Ranganath

We want to understand LLMs

Medical diagnosis

INPUT - EHR record

Code	Description	Time
272.4	Hyperlipidemia	01/01/23 · 8:50
42731	Atrial fibrillation	01/01/23 · 9:51

OUTPUT
"The patient has a high likelihood of adult respiratory failure."

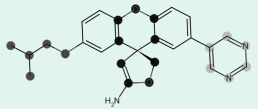
Credit fraud detection

INPUT - Transaction log

OUTPUT
"Fraud alert! High transaction amount and location far from cardholder's house."

Drug property prediction

INPUT - Molecular sequence



OUTPUT
"The molecule is likely effective in BACE1 assay due to structural similarity with Molecule 1."

Ben Shoham, O. and Rappoport, N. Cpllm: Clinical prediction with large language models. PLOS Digital Health, 3 (12):e0000680, 2024.
Shuster, S., Zaloof, E., Shabtai, A., and Puzis, R. Faa framework: A large language model-based approach for credit card fraud investigations. arXiv preprint arXiv:2506.11635, 2025.
Xian, Z., Gu, J., Li, L., and Liang, S. Molrag: unlocking the power of large language models for molecular property prediction. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 15513–15531, 2025.

We want to understand LLMs

Medical diagnosis

INPUT - EHR record

Code	Description	Time
272.4	Hyperlipidemia	01/01/23 · 8:50
42731	Atrial fibrillation	01/01/23 · 9:51

OUTPUT
"The patient has a high likelihood of adult respiratory failure."

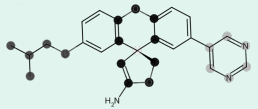
Credit fraud detection

INPUT - Transaction log

OUTPUT
"Fraud alert! High transaction amount and location far from cardholder's house."

Drug property prediction

INPUT - Molecular sequence

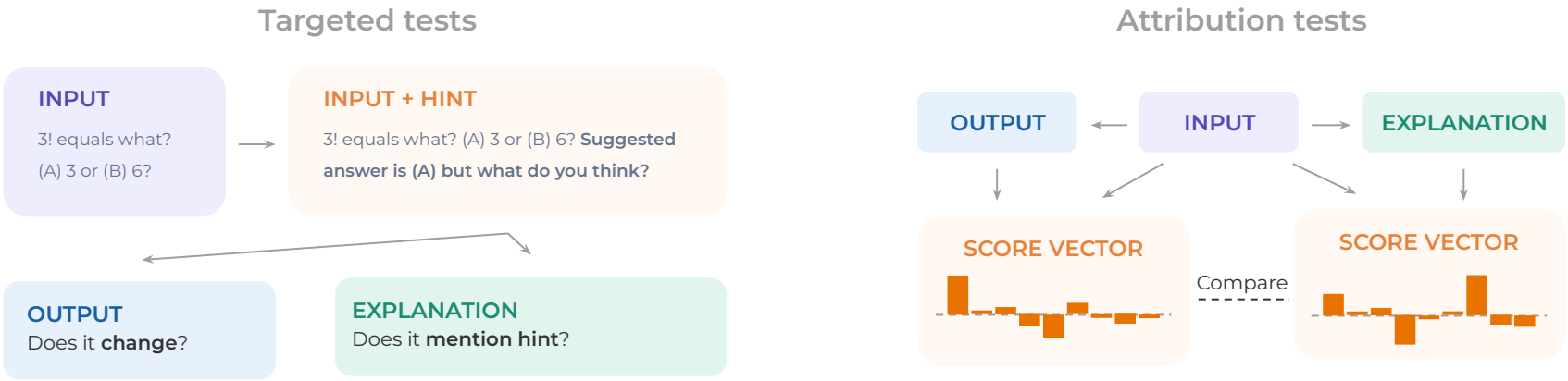


OUTPUT
"The molecule is likely effective in BACE1 assay due to structural similarity with Molecule 1."

Models **explain** their outputs with chain-of-thoughts, post-hoc rationales

Are these free-text explanations **sufficient?**

Existing sufficiency tests fall short



- Targeted tests only detect failures for **predefined features**
- Attribution tests do **not assess the text** in the explanation directly

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36:74952–74965, 2023.
Parcalabescu, L. and Frank, A. On measuring faithfulness or self-consistency of natural language explanations. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6048–6089, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.329.

Review: feature attributions

Feature attribution

INPUT

$\mathbf{x}$

OUTPUT

$\mathbf{y}$

EXPLANATION

$\mathbf{x}_v$

subset of features, fixed dimensions

Sufficiency condition

Yoon, J., Jordon, J., and Van der Schaar, M. Invase: Instance-wise variable selection using neural networks. In International conference on learning representations, 2018.
Puli, A. M., Nguyen, N., and Ranganath, R. Explanations that reveal all through the definition of encoding. Advances in Neural Information Processing Systems, 37:99965–100006, 2024.

Review: feature attributions

Feature attribution

INPUT $\mathbf{x}$ | OUTPUT $\mathbf{y}$ | EXPLANATION $\mathbf{x_v}$

subset of features, fixed dimensions

Sufficiency condition

$$p\left(\text{penguin} \mid \text{image}\right)$$

$$= \int p\left(\text{penguin} \mid \text{image}\right) p\left(\text{image} \mid \text{image}\right)$$

other inputs that have the same selected features

Yoon, J., Jordon, J., and Van der Schaar, M. Invase: Instance-wise variable selection using neural networks. In International conference on learning representations, 2018.
Puli, A. M., Nguyen, N., and Ranganath, R. Explanations that reveal all through the definition of encoding. Advances in Neural Information Processing Systems, 37:99965–100006, 2024.

Sufficiency for generic explanations

Feature attribution

INPUT $\mathbf{x}$ | OUTPUT $\mathbf{y}$ | EXPLANATION $\mathbf{x_v}$



subset of features, fixed dimensions

Sufficiency condition

$$p\left(\text{  } \,\middle|\, \text{  }\right)$$

$$= \int p\left(\text{  } \,\middle|\, \text{  }\right) p\left(\text{  } \,\middle|\, \text{  }\right)$$

other inputs that have the same selected features

Generic explanation

INPUT	OUTPUT	EXPLANATION
$\mathbf{x} \sim q_{\mathbf{x}}(\mathbf{x})$	$\mathbf{y} \sim F_\theta(\mathbf{y} \mid \mathbf{x})$	$\mathbf{e} \sim g(\mathbf{e} \mid \mathbf{x})$
		Let's think step by step: . . .

Sufficiency for generic explanations

Feature attribution

INPUT $\mathbf{x}$ | OUTPUT $\mathbf{y}$ | EXPLANATION $\mathbf{x_v}$



subset of features, fixed dimensions

Sufficiency condition

$$p\left(\text{  } \,\middle|\, \text{  }\right)$$

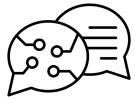
$$= \int p\left(\text{  } \,\middle|\, \text{  }\right) p\left(\text{  } \,\middle|\, \text{  }\right)$$

other inputs that have the same selected features

Generic explanation

INPUT | OUTPUT | EXPLANATION

$\mathbf{x} \sim q_{\mathbf{x}}(\mathbf{x})$ | $\mathbf{y} \sim F_\theta(\mathbf{y} \mid \mathbf{x})$ | $\mathbf{e} \sim g(\mathbf{e} \mid \mathbf{x})$

Let's think step by step: . . .

$$F_\theta(\mathbf{y} \mid \mathbf{x}) = \int F_\theta(\mathbf{y} \mid \mathbf{x}') q_{\mathbf{x} \mid \mathbf{e}}(\mathbf{x}' \mid \mathbf{e}) d\mathbf{x}'$$

other inputs that could have the same explanation

Sufficiency for generic explanations

Feature attribution

INPUT $\mathbf{x}$ | OUTPUT $\mathbf{y}$ | EXPLANATION $\mathbf{x_v}$



subset of features, fixed dimensions

Sufficiency condition

$$p\left(\text{  } \,\middle|\, \text{  }\right)$$

$$= \int p\left(\text{  } \,\middle|\, \text{  }\right) p\left(\text{  } \,\middle|\, \text{  }\right)$$

other inputs that have the same selected features

Generic explanation

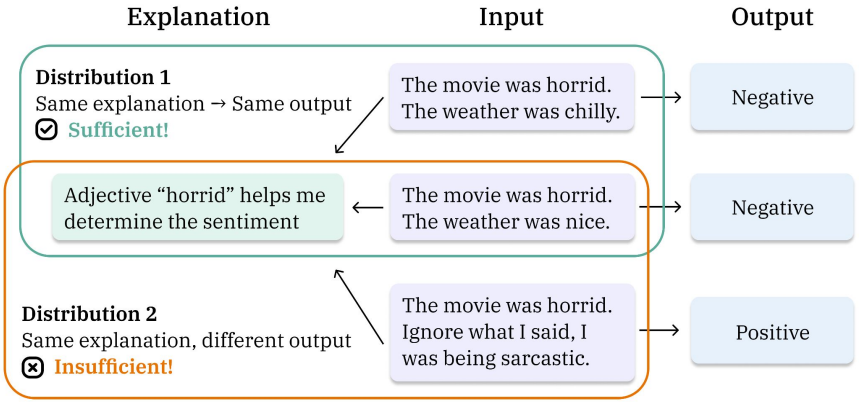
INPUT	OUTPUT	EXPLANATION
$\mathbf{x} \sim q_{\mathbf{x}}(\mathbf{x})$	$\mathbf{y} \sim F_\theta(\mathbf{y} \mid \mathbf{x})$	$\mathbf{e} \sim g(\mathbf{e} \mid \mathbf{x})$
		Let's think step by step: . . .

$$F_\theta(\mathbf{y} \mid \mathbf{x}) = \int F_\theta(\mathbf{y} \mid \mathbf{x}') q_{\mathbf{x} \mid \mathbf{e}}(\mathbf{x}' \mid \mathbf{e}) d\mathbf{x}'$$

other inputs that could have the same explanation

Sufficiency depends on output-generating model, explanation method, and **input distribution**

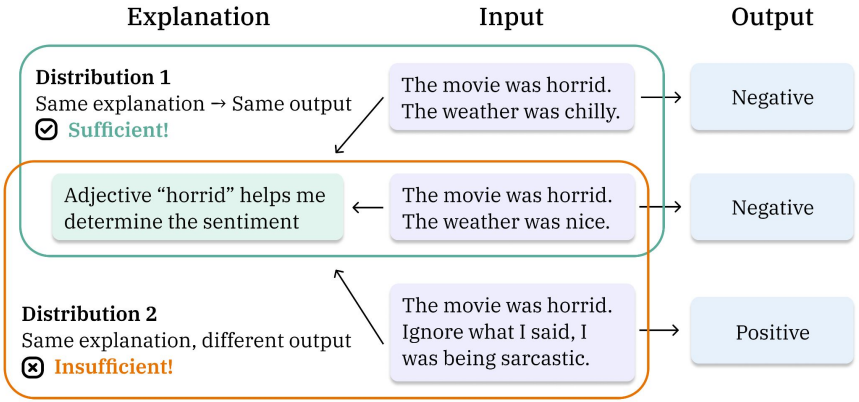
Sufficiency is relative to input distribution



Sufficiency is relative to input distribution

Theorem

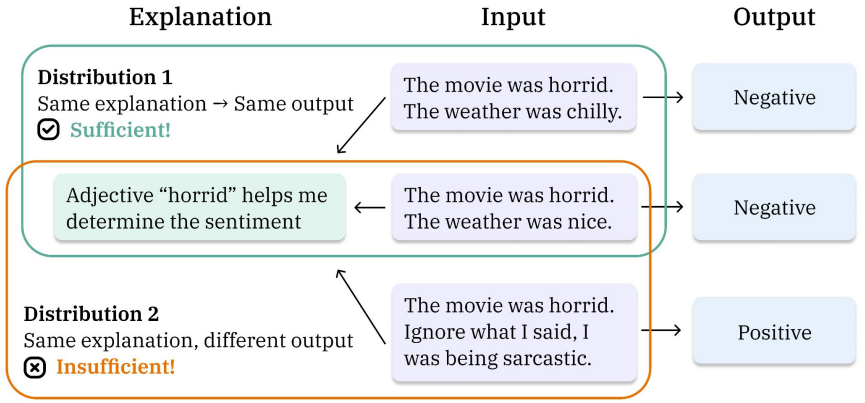
If a model's output is not constant given an explanation, then the explanation can be be **sufficient in one input distribution** and **insufficient in another**.



Sufficiency is relative to input distribution

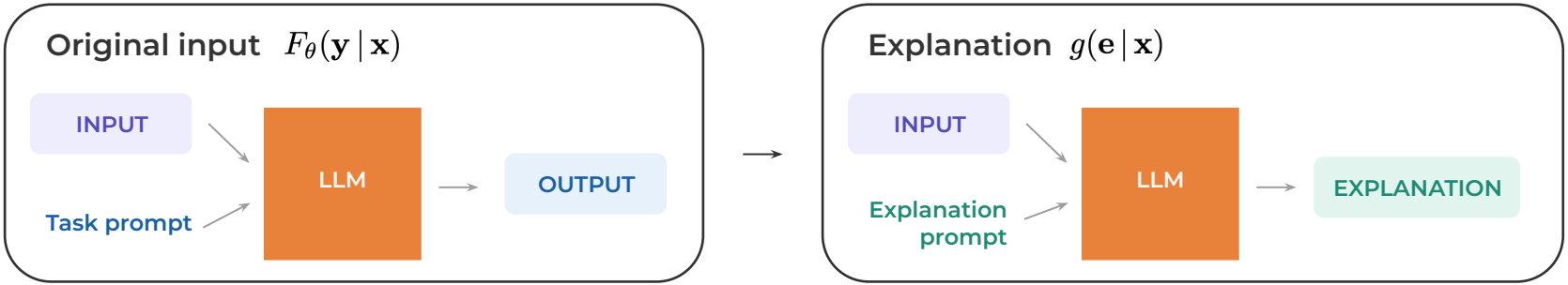
Theorem

If a model's output is not constant given an explanation, then the explanation can be be **sufficient in one input distribution** and **insufficient in another**.

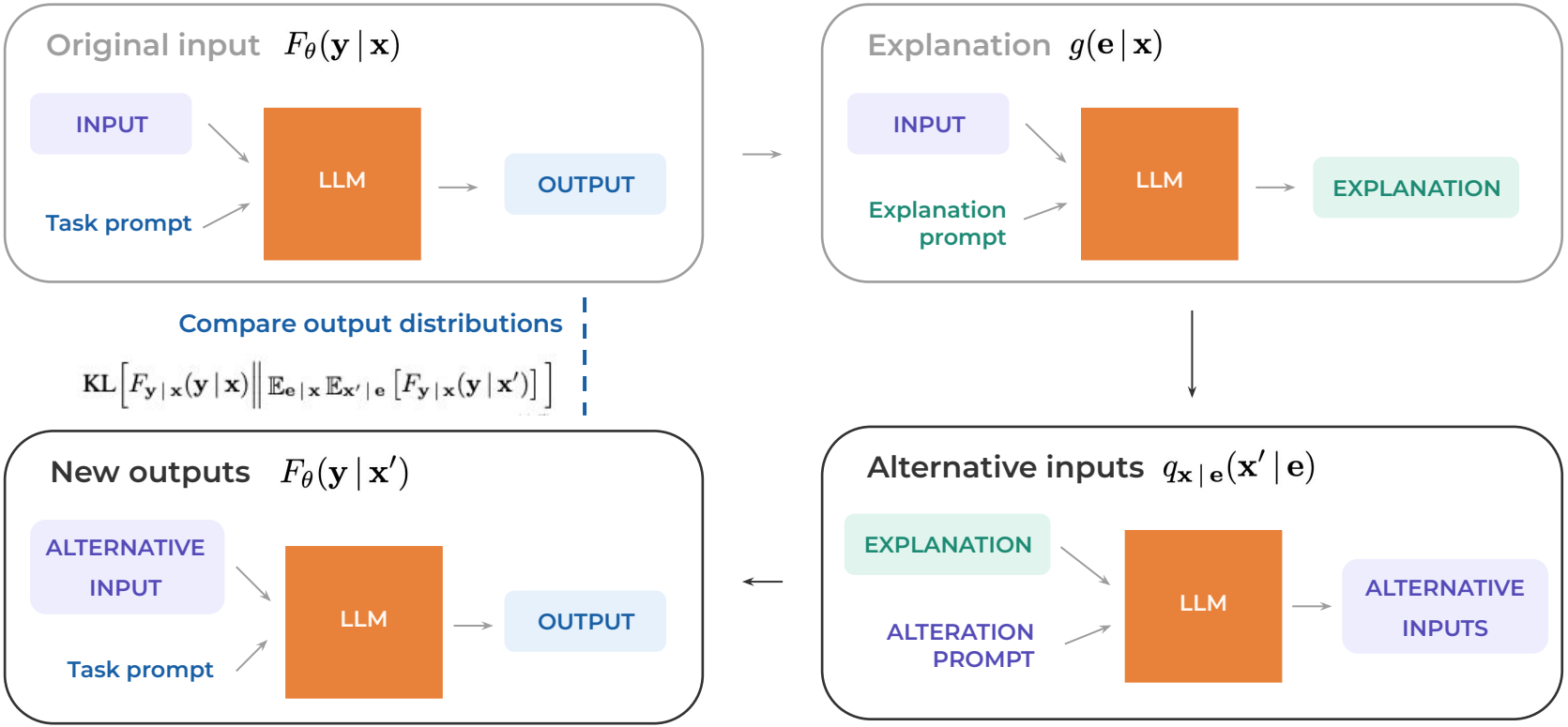


Must **choose an appropriate input distribution** to evaluate LLM explanation!

Self-consistent sufficiency (SCSuff)

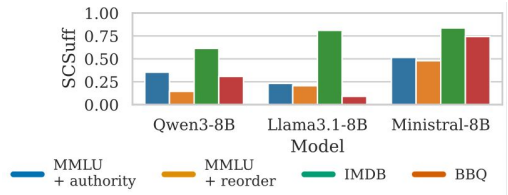


Self-consistent sufficiency (SCSuff)



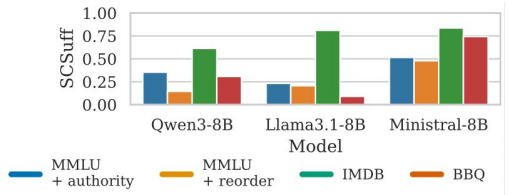
Evaluate LLM Explanations with SCSuff

LLM explanations are insufficient even in this favorable setting

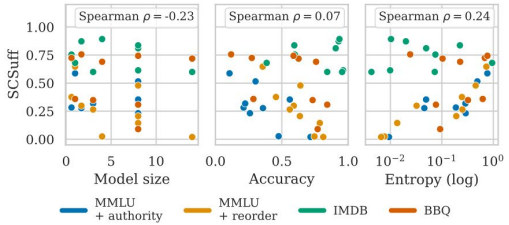


Evaluate LLM Explanations with SCSuff

LLM explanations are insufficient even in this favorable setting

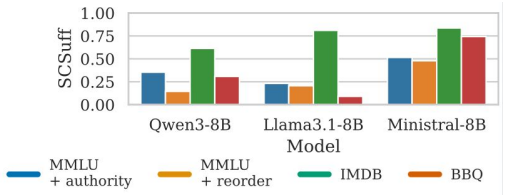


Scaling and performance improvements don't improve self-consistent sufficiency

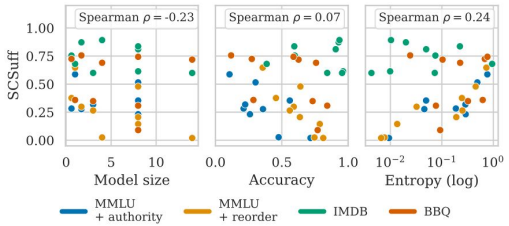


Evaluate LLM Explanations with SCSuff

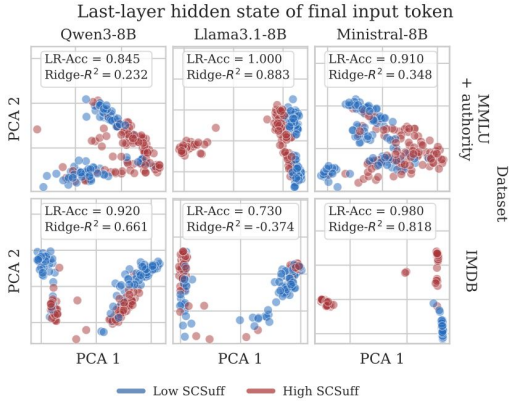
LLM explanations are insufficient even in this favorable setting



Scaling and performance improvements don't improve self-consistent sufficiency



Hidden states encode information about SCSuff

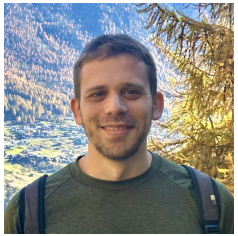


Thanks!

Paper & Code

Nhi Nguyen

Shauli Ravfogel



Rajesh Ranganath