NeuroFaith: Evaluating Mechanistic Faithfulness of LLM Free Text Self-Explanation at the Concept Level

Milan Bhan Jean-Noël Vittaut Nicolas Chesneau Sarath Chandar Marie-Jeanne Lesot

{milan.bhan,nicolas.chesneau}@ekimetrics.com {jean-noel.vittaut,marie-jeanne.lesot}@lip6.fr sarath.chandar@mila.quebec

Context and objective

Context: are LLM free text self-explanations faithful?

- LLMs generate free-text **self-NLE** (*predict-then-explain*)
- **Plausibility vs. faithfulness**: self-NLE look credible but may not reflect the true decision process (Jacovi & Goldberg, 2020)
- Existing methods (CI, AA) rely on **behavioral tests**, treat f as a black box ⇒ measure *self-consistency*, not faithfulness

Aim: directly measure self-NLE faithfulness via **concept-level mechanistic analysis** of internal reasoning

Step 1: Concept Extraction

- Extract from self-NLE $e(x)$ a set of concepts $\{c_i\}_{i=1}^{p}$ quoted as important for $f(x)$
- Concepts: interpretable binary high-level features
- Extraction strategies
  - Human annotation: high-quality, domain-expert, but costly
  - Auxiliary LLM (LLM-as-a-Judge): scalable

Step 2: Concept-wise Interpretation

- Attribution score $\mathtt{I}_\Gamma(c)$ per concept, over a circuit Γ
- Circuit $\Gamma = \{(k, \ell)\}$: sparse subgraph of coordinates (token k, layer ℓ) carrying information flow
- Interpreter $\mathtt{I}$
  - **Probing** c from f hidden states: `Patchscopes`, linear probes $\mathtt{I}_\Gamma(c) = \max_{(k,\ell)\in\Gamma} \mathtt{I}(h_k^\ell, c)$
  - **Concept importance**: causal influence on $f(x)$ via `TCAV`, RepE

Step 3: Faithfulness Measurement

- $e(x)$ faithful when extracted concepts are mechanistically important
- Faithfulness: proportion of concepts with positive attribution
  - $F(x, e) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}_{\mathtt{I}_\Gamma(c_i)>0}$
- Interpretation depends on interpreter $\mathtt{I}$
  - **Probing-based**: concept detection rate in hidden states
  - **Importance-based**: causal relevance for $f(x)$

Instantiation 1: The Case of Classification

Task: predict label $\hat{y}$ from task-relevant concepts $\mathcal{C}$; interpreter $\mathtt{I}$ = **concept importance**

Concept Extraction: auxiliary LLM extracts $\{c_i\}_{i=1}^{p} \subset \mathcal{C}$ from $e(x)$; each c = CAV (`diff-mean`, last token)

Importance-based Interpretation:

- Erase c: $h_k^\ell \leftarrow h_k^\ell - \lambda \overrightarrow{c_\ell}$, over Γ (last token, F1 $> 60\%$ layers)
- Probability shift $p_{\hat{y}} = \beta_0 + \beta_1\lambda$ ⇒ importance $\mathtt{I}_\Gamma(c) = \beta_1$
- Faithfulness $F(x, e) \in [0, 1]$

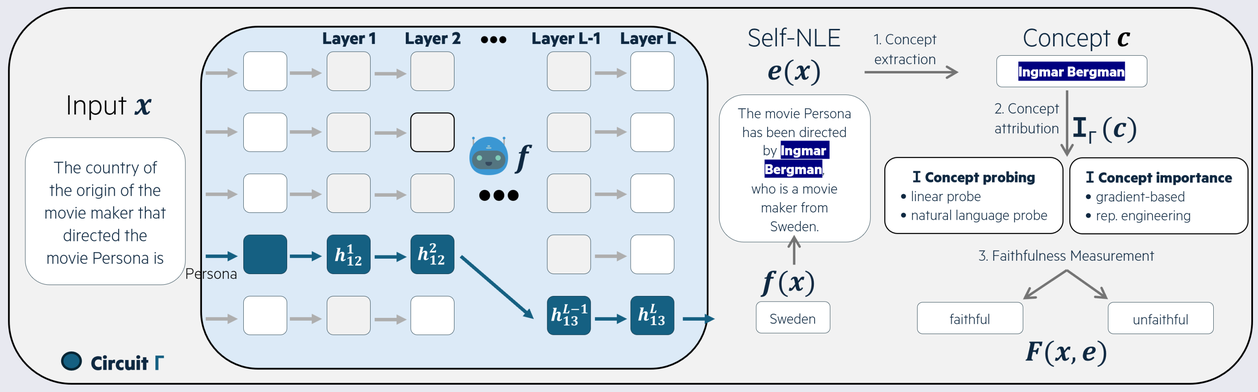
Model	Task Acc.		Self-NLE Faithfulness			
	AGNews	Ledgar	AGNews		Ledgar	
			Accurate	Inaccurate	Accurate	Inaccurate
`gemma-2b`	86.5	40.3	54.6	65.1	58.9	40.5
`gemma-9b`	88.7	59.7	96.0	92.3	56.0	58.3
`gemma-27b`	88.9	58.1	74.0	72.4	53.0	31.6
`mistral-7b`	80.6	82.9	86.6	84.1	94.8	84.4

Table 1: Self-NLE faithfulness from NeuroFaith (%), stratified by prediction correctness.

Our proposal: NeuroFaith

3-step framework

Concept extraction, Concept-wise Interpretation, Faithfulness Measurement



Instantiation 2: The Case of 2-Hop Reasoning

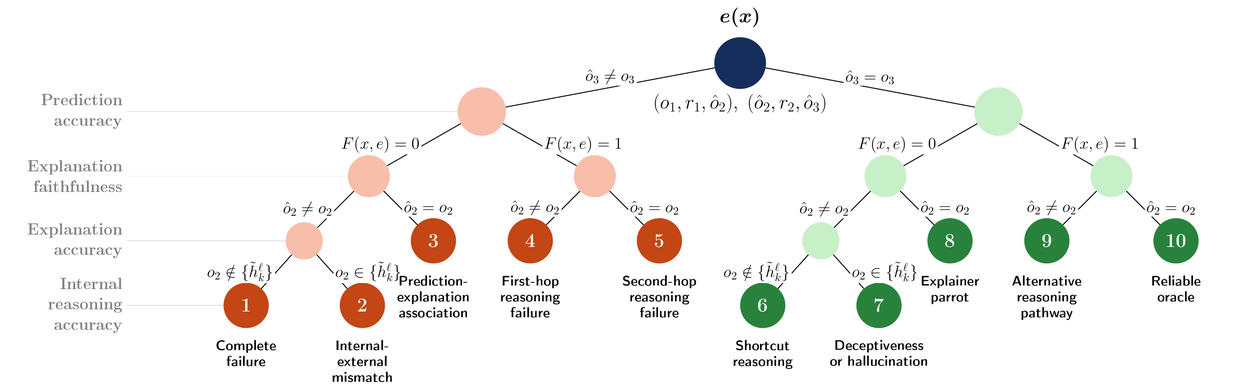
Task: $x = (o_1, r_1, \blacktriangle, r_2, \bullet)$, bridge $\widehat{o_2}$, answer $f(x) = \widehat{o_3}$

- Ex.: *director of "Persona"* → *Bergman* → *Sweden*

Concept Extraction: auxiliary LLM gets bridge $\widehat{o_2} = c$ from $e(x)$

Probing-based Interpretation:

- Detect c in hidden states (`Patchscopes`, unsupervised)
- Γ: bridge resolved early, last token of o_1 (Biran et al., 2024)
- $\mathtt{I}_\Gamma(c) = 1 \iff \exists (k, \ell) \in \Gamma,\ c \in \tilde{h}_k^\ell$

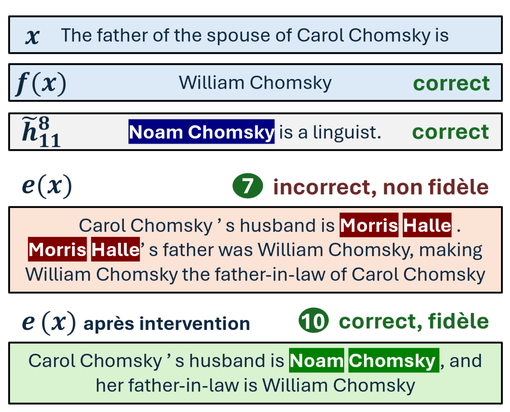


Model	Task Acc.	Incorrect Predictions					Correct Predictions				
		1	2	3	4	5	6	7	8	9	10
`gemma-2b`	7.1	23.8	4.2	14.4	16.0	41.6	27.8	6.1	17.7	8.1	40.3
`gemma-9b`	16.6	26.1	4.5	14.5	14.1	40.9	14.9	4.2	20.0	6.8	54.0
`gemma-27b`	22.3	22.8	4.4	12.5	17.1	43.1	12.8	3.1	15.3	6.4	62.4
`mistral-7b`	21.2	46.5	13.7	7.1	8.4	24.2	20.6	4.0	19.3	5.2	50.9

Table 2: Detailed taxonomy and distribution of categories across models (in %).

Linearly Detecting and Improving Faithfulness

- NeuroFaith faithfulness is a **linear direction** in representation space
- **Detection**: `diff-mean` probe $\overrightarrow{F_\ell}$ detects (un)faithful self-NLE (F1 61–75%); strongest at the bridge object token
- **Improvement**: activation steering $h_k^\ell \leftarrow h_k^\ell + \lambda \overrightarrow{F_\ell}$ converts 8–11% of unfaithful self-NLE into faithful ones
- Largest gains (19–58%) when the bridge is internally computed but absent from $e(x)$ ⇒ steering surfaces what f already knows
- Links faithfulness to **hallucination** and **deceptiveness** directions



Conclusion

NeuroFaith: self-NLE mechanistic faithfulness

- Faithfulness: linear structure
- Fast detection and **enhancement via activation steering**

Mechanistic Interpretability Workshop @ ICML 2026