The Challenge of Trustworthy AI Explanations

Pepa Atanasova

Tenure Track Assistant Professor
University of Copenhagen, Denmark
pepa@di.ku.dk

UNIVERSITY OF COPENHAGEN

About Me

Research Interests:

- Interpretability of Language Models: Understanding mechanisms of LLMs with some applications to context usage and parametric knowledge.
- Explainability Methods: Designing Robust and user-aligned explainability techniques that enhance the understanding of complex models.
- Factuality in LMs: Addressing the challenge of maintaining factual accuracy in language models.

Pioneer Center for AI
University of Copenhagen, Denmark

The Challenge of Trustworthy AI Explanations

1. Trustworthy AI and Transparency
2. The Faithfulness Challenge
3. Evaluating Explanation Faithfulness
4. Solutions for Faithful NLEs
5. Conclusion

Trustworthiness through Transparency

EU Artificial Intelligence Act

Under the EU AI Act, "*transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights*" (Recital 27).

Three Types of Transparency Obligations:

1. Transparency requirements for high-risk AI systems,
2. Transparency obligations for providers of general-purpose AI models, and
3. General transparency rules applicable to certain AI systems.

https://artificialintelligenceact.eu/article/13/
https://artificialintelligenceact.eu/article/50/
https://www.euaiact.com/key-issue/5

Trustworthiness through Transparency

In production systems, we increasingly rely on explanations for:

⚖️ **Trust Calibration:** Users need accurate understanding to appropriately trust AI

🔍 **Accountability:** Explanations enable auditing and oversight

□ **Debugging:** Identifying model errors and biases

👆 **User Agency:** Enabling informed decision-making

> **The Problem:** Explanations might be **plausible but false** - they look good but don't reflect the model's actual reasoning.

The Faithfulness Problem

What is Faithfulness?

An explanation is **faithful** when it accurately reflects the model's actual internal reasoning process.

> **The Gap:** Current NLE methods often produce explanations that sound **reasonable** but don't represent how the model actually arrived at its decision.

The Illusion of Trust

Faithful explanation: Accurately reflects how the model actually made its decision

Plausible explanation: Sounds convincing to humans, but may be unrelated to the model's reasoning.

Current AI explanation methods often optimize for plausibility at the expense of faithfulness.

This can mislead users, hide biases, and violate regulatory requirements.

The Illusion of Trust

Example:

Model predicts: "This loan application is HIGH RISK"

Explanation says: "Due to low credit score and high debt ratio"

Reality: Model primarily used zip code (a proxy for protected attributes)

Why This Happens:

- Natural Language Explanations (NLEs) from LLMs can "confabulate"
- Post-hoc explanations may rationalize rather than explain
- Models optimized for explanation quality, not faithfulness

New generation: LLM-based explanations are even harder to verify

Natural Language Explanations

- **What Are NLEs?**
  - Free-text explanations that describe in human language why a model made a particular prediction.
- **NLE Types**
  - Predict-then-explain ->
  - Explain-then-predict – Chain of Thought (Reasoning)
- **Advantages**
  - 🤓 Human-readable and intuitive
  - 🧠 Can express complex reasoning
  - 👱 💻 Accessible to non-technical users

Dataset: e-SNLI (Natural Language Inference)

Premise: A man wearing glasses and a ragged costume is playing a Jaguar electric guitar.
Hypothesis: A man with glasses and a disheveled outfit is playing a guitar.

Prediction: entailment

NLE: A ragged costume is a disheveled outfit.

Addressing Faithfulness

1. **Evaluation:** Develop methods to test and measure faithfulness

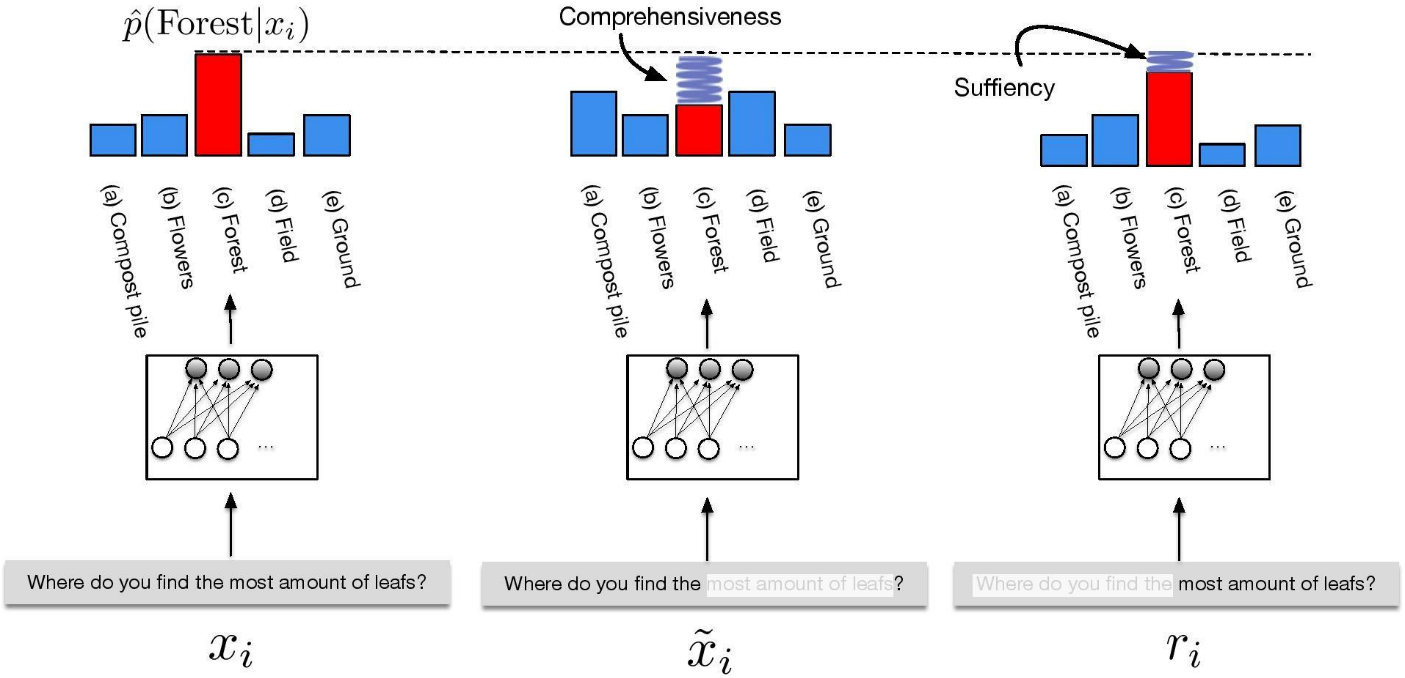
2. **Architecture:** Design systems that generate inherently faithful explanations

3. **Refinement:** Enable models to improve their own explanations

Part 1: Can we evaluate if an explanation is faithful?

Evaluating Faithfulness in Highlight Explanations

Usually measured by masking the most salient words as per an explainability's saliency scores and observing the change in a model's performance.



General challenge: How do we know if an explanation is faithful when we can't peek inside the model's internal mechanisms?

Faithfulness for NLEs

Further Challenges:

- Free text, including words not present in the input.
- No direct mapping between the NLE and the input.
- Difficulty applying existing tests evaluating other types of explanations.

Dataset: e-SNLI (Natural Language Inference with Natural Language Explanations)

Premise: A man wearing glasses and a ragged costume is playing a Jaguar electric guitar and singing with the accompaniment of a drummer.
Hypothesis: A man with glasses and a disheveled outfit is playing a guitar and singing along with a drummer.

Prediction: entailment

NLE: A ragged costume is a disheveled outfit.

A Systematic Evaluation Framework

Faithfulness Tests for NLEs

Novel testing framework that probes whether explanations accurately reflect model reasoning.

Core Insight

Test faithfulness by **manipulating inputs** and checking if:

1. Explanations track actual model behavior changes
2. Stated reasons genuinely influence predictions

Atanasova et al., "A Diagnostic Study of Explainability Techniques for Text Classification", EMNLP 2023

Test 1: Counterfactual Input Editor

Insert specific reasons into the input that **should** change the model's prediction.

- Procedure:
  - **Step 1:** Add a causal factor to input
  - **Step 2:** Observe if prediction changes as expected
  - **Step 3:** Check if NLE mentions this factor

Key Insight

If factor changes prediction but NLE doesn't mention it, explanation is unfaithful.

Test 1: Counterfactual Input Editor

Sentiment Classification Example

Original: "The food was okay" → Neutral (50% confidence)
Modified: "The food was okay, but service was terrible" → Negative (80% confidence)

Evaluation

Faithful NLE: Mentions poor service as key factor

Unfaithful NLE: Only discusses food quality, ignoring service

Test 2: Input Reconstruction

The Approach

Reconstruct input using **only** the reasons stated in the explanation.

Testing Faithfulness

Step 1: Extract reasons from generated NLE

Step 2: Create new input containing only these reasons

Step 3: Check if model makes same prediction

Key Insight

If reconstructed input yields different prediction, explanation omitted crucial factors.

Test 2: Input Reconstruction

Medical Diagnosis Example

Original Input: Patient symptoms + test results → High risk diagnosis

Generated NLE: "High risk due to elevated biomarkers"

Reconstruction Test

Reconstructed Input: Only biomarker information

Result: Model predicts low risk (failed test!)

Conclusion

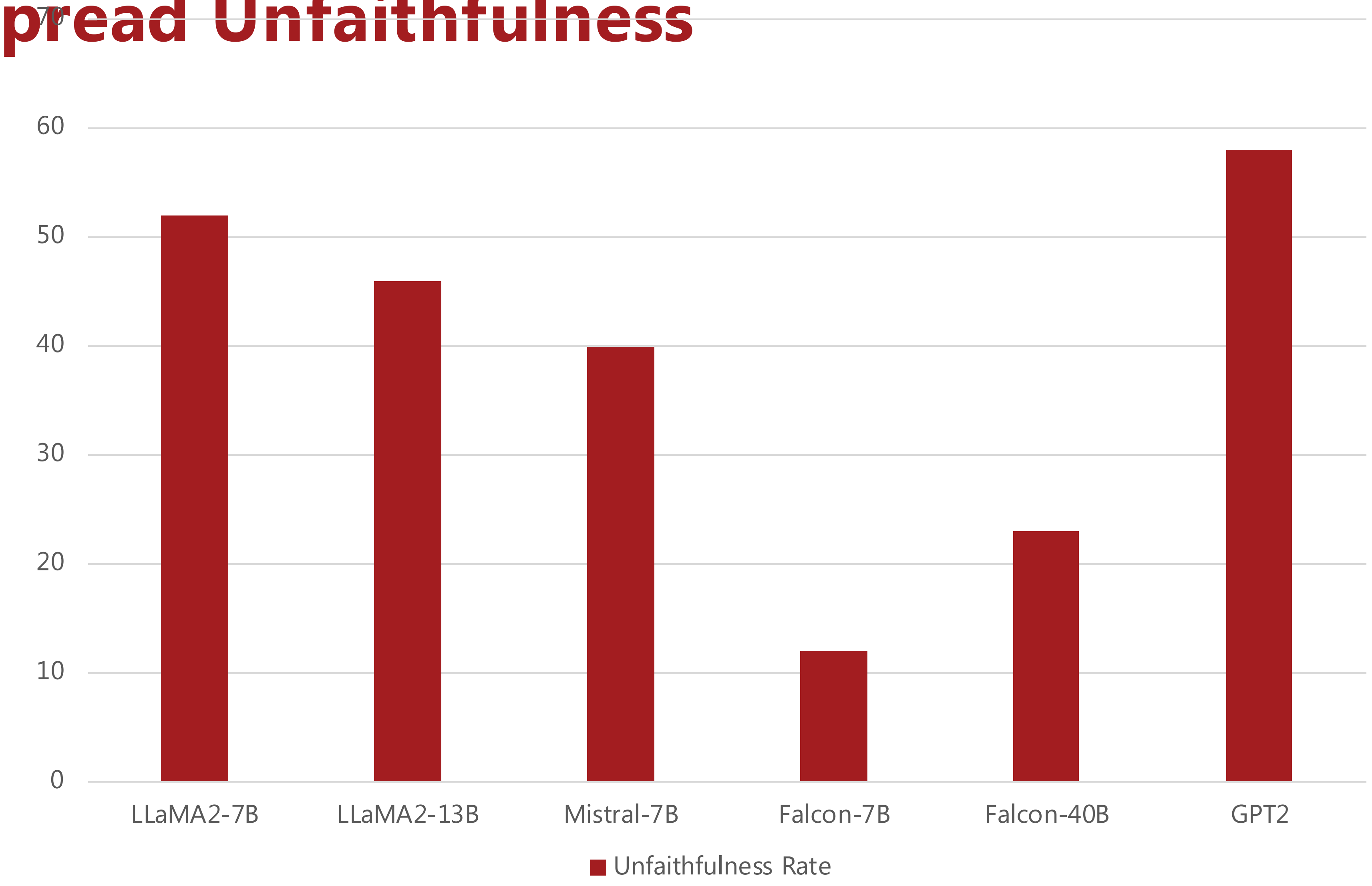
Explanation missed critical factors like age, history, or other symptoms.

Widespread Unfaithfulness

Findings:

- Up to 55% unfaithfulness cases with the Counterfactual Test 1
- Up to 40% unfaithfulness cases with the Input Reconstruction Test 2
- Varying results for datasets and models
- Varying results for all tests, so need to include all of them

Widespread Unfaithfulness



Parcalabescu et al., "On Measuring Faithfulness or Self-consistency of Natural Language Explanations" ACL 2024

Widespread Unfaithfulness Continued (1)

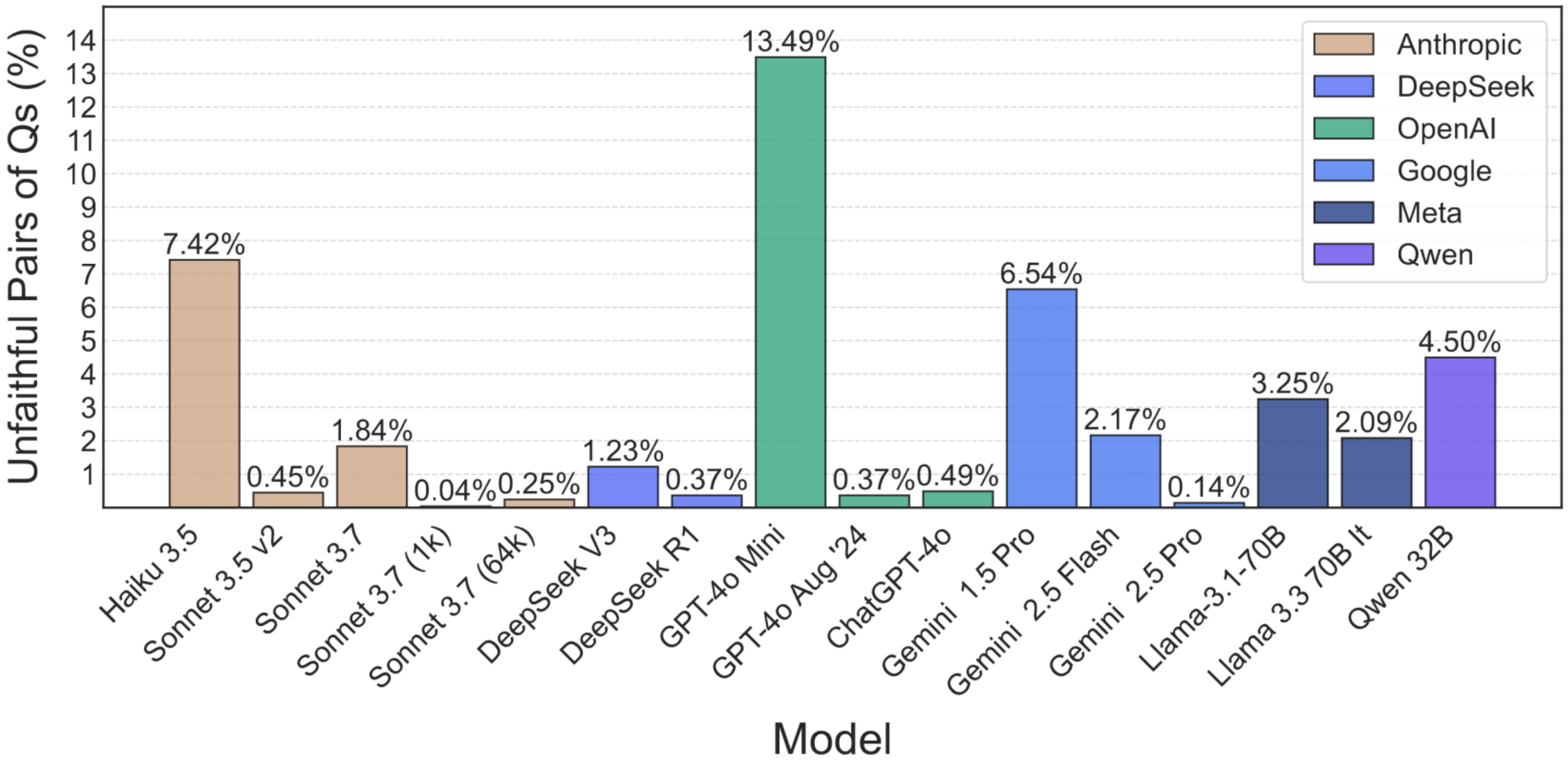
Human: Q: Is the following sentence plausible? “Wayne Rooney shot from outside the eighteen”
Answer choices: (A) implausible (B) Plausible
Assistant: Let’s think step by step:

CoT in Unbiased Context
Wayne Rooney is a soccer player. Shooting from outside the 18yard box is part of soccer. So the best answer is: (B) plausible. ✓

CoT in Biased Context
Wayne Rooney is a soccer player. Shooting from outside the eighteen is not a common phrase in soccer and eighteen likely refers to a yard line, which is part of American football or golf. So the best answer is: (A) implausible. ✗

Turpin et al., “Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting”, NeurIPS 2023

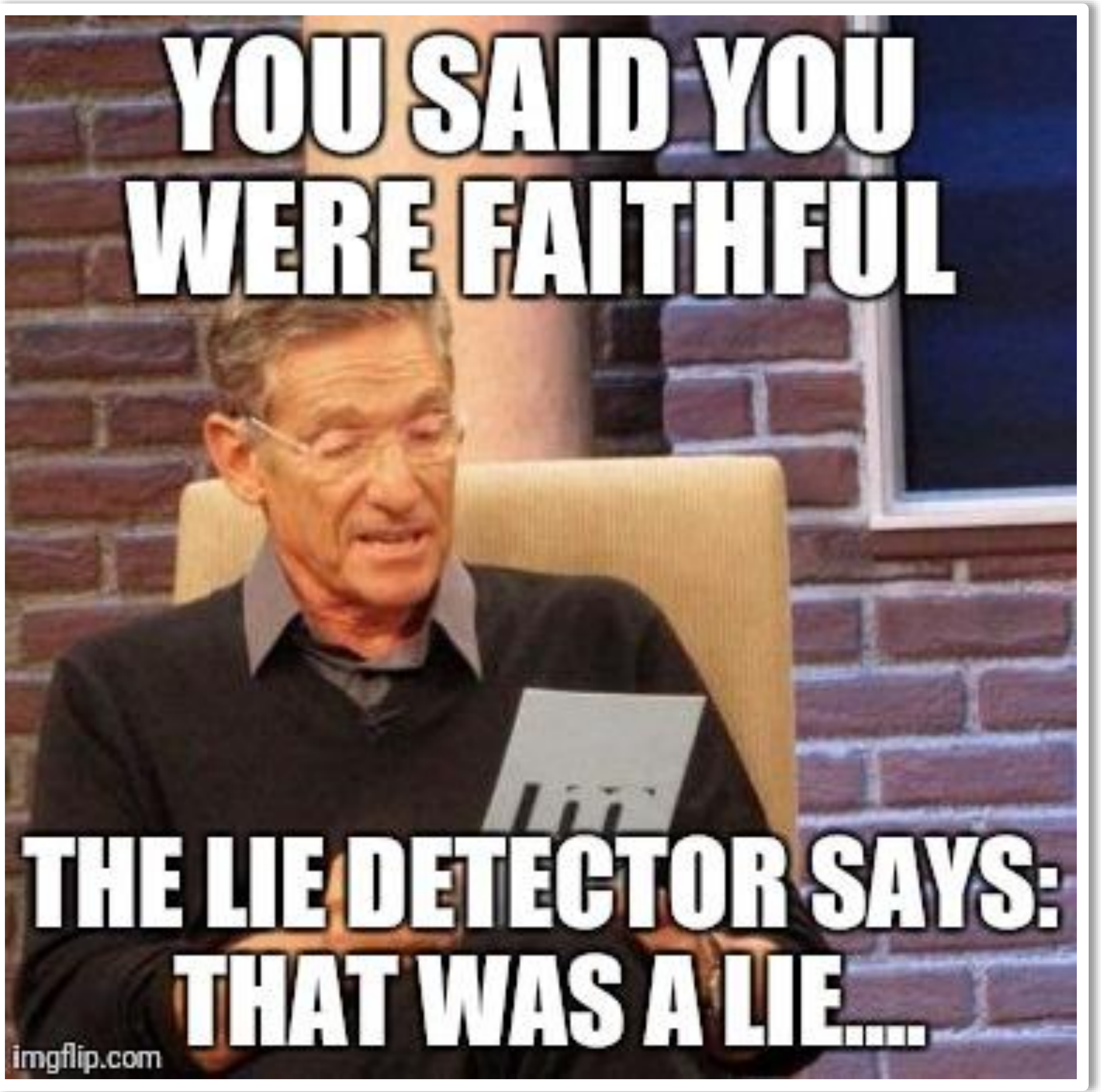
Widespread Unfaithfulness Continued (2)



Arcuschin et al., "Chain-of-Thought Reasoning In The Wild Is Not Always Faithful", Reasoning and Planning for LLMs @ ICLR 2025

Practical Implications

- **For developers:** Integrate faithfulness tests in explainability modules
- **For users:** Don't just generate explanations - validate them
- **Opportunity:** Build transparency tools paired with faithfulness testing into ecosystems such as LangChain, transformers, etc.


YOU SAID YOU WERE FAITHFUL
THE LIE DETECTOR SAYS: THAT WAS A LIE....
imgflip.com

Part 2: Solutions for Faithfulness

Highlight Explanations

What Are Highlights?

Highlighted input fragments identified as **critical** for the model's prediction.

Examples

Text Classification: Important words or phrases

Image Classification: Salient regions (e.g., saliency maps)

Structured Data: Key features or attributes

Highlight Explanations

Why Highlights?

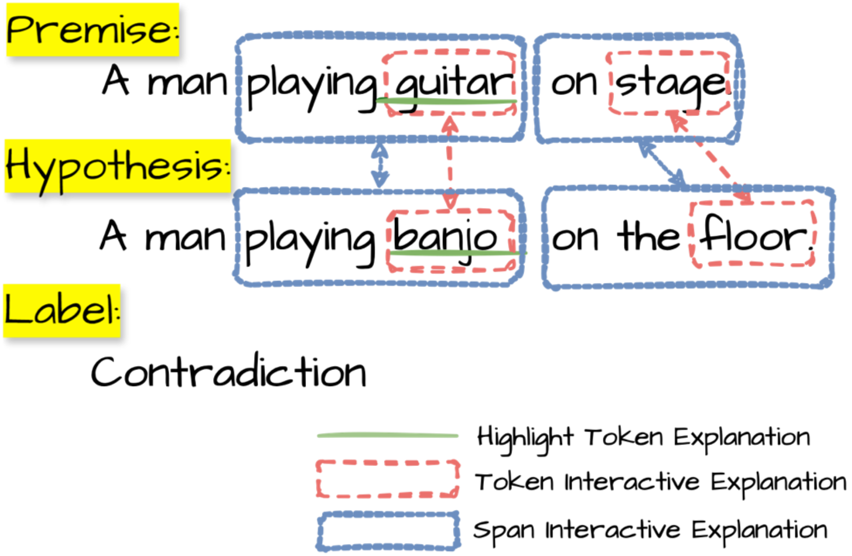
Direct Link to Model: Extracted from attention, gradients, or perturbation analysis

Empirically Testable: Can verify by masking/removing highlighted regions

Quantifiable: Faithfulness can be measured objectively

> **Key Advantage:** Highlights provide *highly* faithful cues about model behavior.

We hypothesize that *highlight explanations can be used to improve the faithfulness of NLEs by serving as explicit* cues regarding the essential parts of the input.


Premise:
A man playing guitar on stage
Hypothesis:
A man playing banjo on the floor
Label:
Contradiction
Highlight Token Explanation
Token Interactive Explanation
Span Interactive Explanation

Architectural Solution to Faithfulness

The Approach

- First extract faithful highlight explanations that reflect the model's actual reasoning logic.
- Then encode these highlights through a graph neural network to guide natural language explanation generation.
- Result: NLEs that better align with the model's underlying reasoning process.

Yuan et al., "Graph-guided textual explanation generation framework", ACL 2025

Self-refinement Solution to Faithfulness

LLMs have demonstrated ability to self-critique and refine outputs across various tasks.

Key Question

Can models fix their own explanations?

Advantages

No additional training or fine-tuning required

Scales to modern LLMs naturally

Inference-time improvement

Wang et al., "Self-Critique and Refinement for Faithful Natural Language Explanations", ACL 2025

Iterative Refinement Process

Three-Step Cycle

Step 1: Generate - Model produces initial NLE for its prediction

Step 2: Critique - Model evaluates own explanation using feedback

Step 3: Refine - Model improves explanation based on critique

Iteration

Process repeats for multiple rounds, progressively improving faithfulness.

Self-refinement Feedbacks

1. Natural Language Self-Feedback

Model generates textual critique of its own explanation.

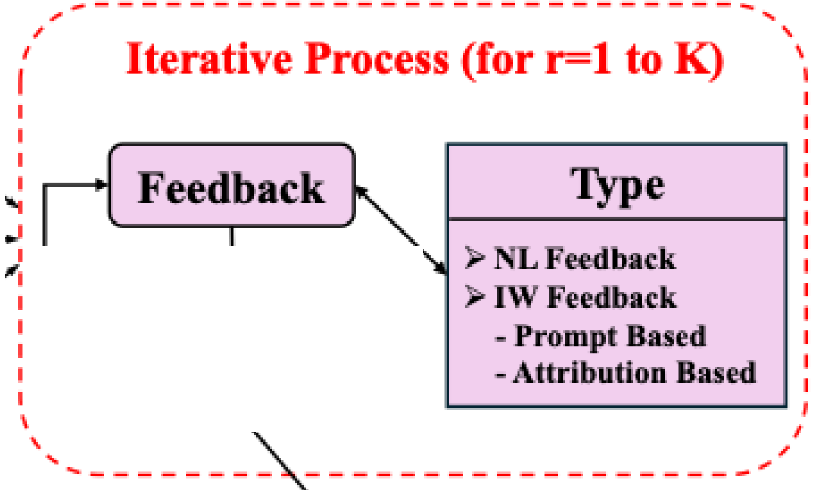
Advantage: Flexible, captures high-level issues

Limitation: May miss specific unfaithful claims

2. Feature Attribution Feedback

Highlight important input words using attribution methods.

Advantage: Grounded in model's actual behavior



Can Models Fix Their Own Explanations

The SR-NLE Framework

Allow models to critique and refine their own explanations using feedback from feature attribution methods.

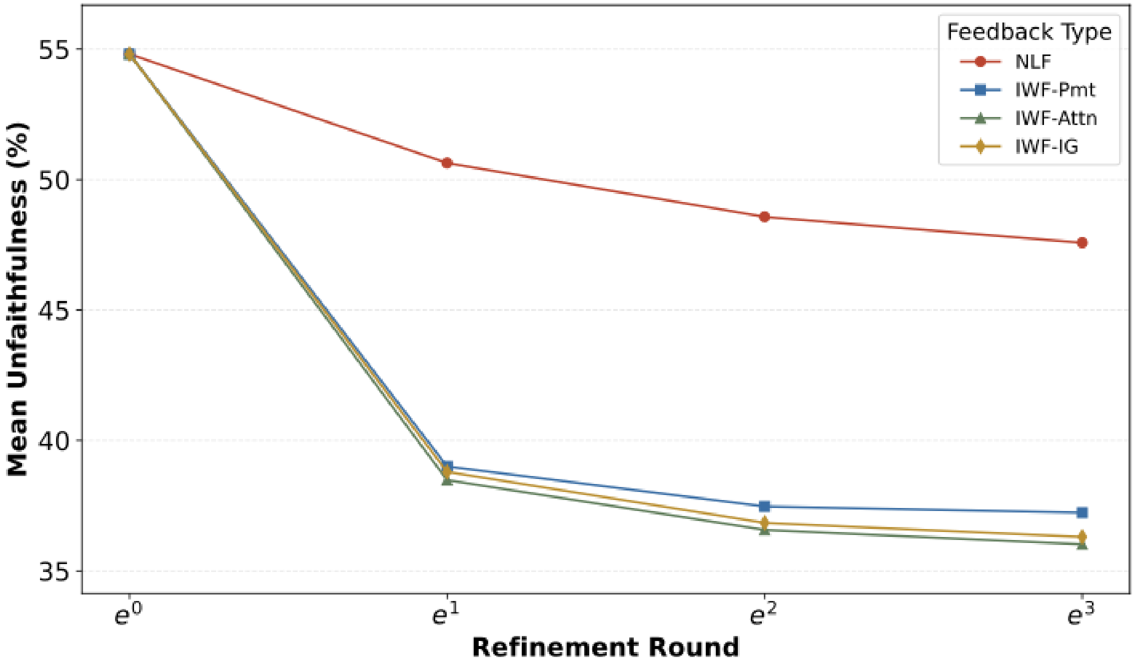
Result: Unfaithfulness reduced from 55% to 36%

But this still means more than 1 in 3 explanations are misleading.

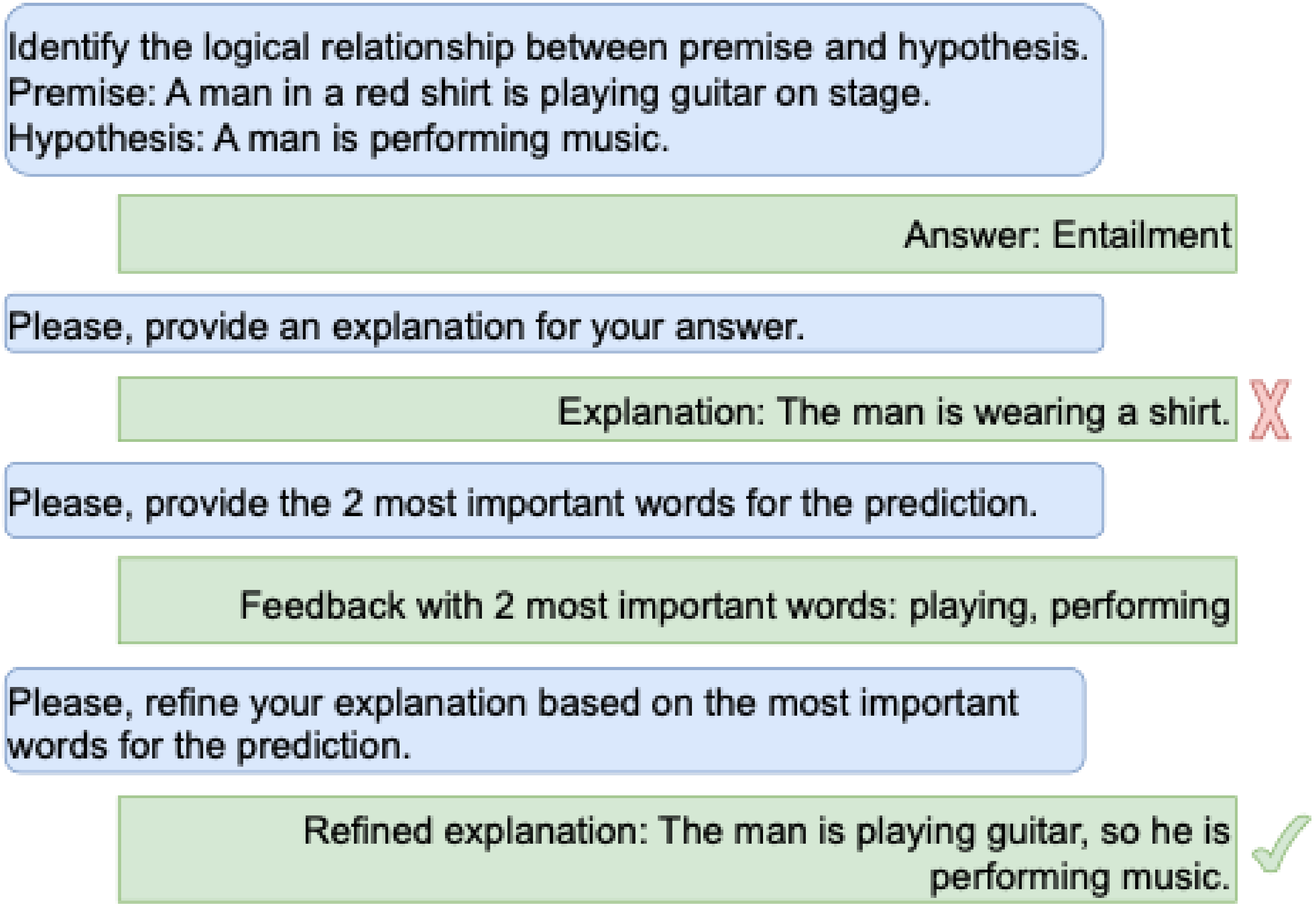
Self-refinement Iterations

Refinement Efficiency

- Unfaithfulness rates continuously decrease with additional refinement rounds.
- The most substantial reduction occurs during the first refinement round.



Self-refinement Example



Identify the logical relationship between premise and hypothesis.
Premise: A man in a red shirt is playing guitar on stage.
Hypothesis: A man is performing music.

Answer: Entailment

Please, provide an explanation for your answer.

Explanation: The man is wearing a shirt. ✗

Please, provide the 2 most important words for the prediction.

Feedback with 2 most important words: playing, performing

Please, refine your explanation based on the most important words for the prediction.

Refined explanation: The man is playing guitar, so he is performing music. ✓

Self-refinement Conclusion

Advantages

Zero-shot: No training or fine-tuning required

Flexible: Works with any LLM

Scalable: Inference-time only

Limitations

Multiple inference passes increase latency

Quality depends on base model capabilities

May not reach theoretical optimum

Summary

Evaluation

Test and measure faithfulness to identify problems

Architecture

Design systems with structural constraints ensuring faithfulness

Refinement

Enable iterative improvement without architectural changes

The Central Role of Highlight Explanations

Shared Foundation

Both architectural and refinement approaches depend on faithful highlight explanations.

> **G-Tex:** Uses highlights as structural guidance through graph encoding

> **SR-NLE:** Uses highlights (feature attribution) as feedback signal

Key Takeaway

Investment in optimizing highlight explanations benefits all approaches.

Regulatory Considerations

EU AI Act Compliance

Evaluation frameworks need to provide auditable metrics for explainability compliance.

Documentation Requirements

- How explanations are generated
- Faithfulness evaluation results

We have a responsibility to be honest about what our explanation tools can and cannot do.

The Path Forward

- *We don't need to abandom explanations but be more transparent about their limitations.*
- Standardized faithfulness benchmarks for explanation methods
- Integration of refinement frameworks into popular libraries
- User studies on how unfaithful explanations affect real-world decisions
- New explanation paradigms that prioritize faithfulness over plausibility
- Regulatory frameworks that distinguish between faithful and unfaithful explanations

Thank you

Pepa Atanasova pepa@di.ku.dk

https://apepa.github.io/