

Explainability in Foundation Models

Helena Silva¹, Sofia Novak²

¹ Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: helena.silva624@yahoo.com | ORCID: 5438-1040-9186-3354

² Assistant Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: sofia.novak716@yahoo.com | ORCID: 6426-0316-3100-5980

ABSTRACT

Foundation models -- large pre-trained models adapted to diverse downstream tasks -- have achieved remarkable performance across language, vision, and multimodal domains, yet their internal decision-making processes remain opaque, creating barriers to trust, accountability, and regulatory compliance in high-stakes deployments. This study presents a systematic evaluation of six explainability methods -- attention visualisation, gradient-based attribution (Integrated Gradients), perturbation-based attribution (SHAP), probing classifiers for internal representations, mechanistic interpretability (circuit analysis), and natural language explanations generated by the model itself (chain-of-thought) -- applied to three foundation model families: BERT-large (NLP classification), ViT-L/16 (image classification), and GPT-2 XL (text generation). Explainability was evaluated across four dimensions: faithfulness (do explanations reflect true model reasoning?), plausibility (do explanations match human intuition?), consistency (do similar inputs receive similar explanations?), and actionability (can explanations guide model improvement?). A total of 2,400 explanation evaluations were conducted with 24 human annotators across three expertise levels. Integrated Gradients achieved the highest faithfulness on NLP tasks (sufficiency = 0.82 +- 0.04; comprehensiveness = 0.76 +- 0.05), confirming gradient-based methods as the most reliable for identifying causally relevant input features. Attention maps showed high plausibility (human agreement = 78.4 +- 3.2%) but poor faithfulness (sufficiency = 0.54 +- 0.06), confirming the well-documented disconnect between attention patterns and actual model reasoning. Chain-of-thought explanations from GPT-2 XL were rated most understandable by non-expert users (comprehension score = 4.2/5) but were faithful to the model's actual computation in only 42.6% of cases. Mechanistic circuit analysis provided the deepest causal understanding but scaled poorly, requiring 40+ researcher-hours per circuit discovered. No single method satisfied all four dimensions simultaneously, confirming that explainability requires a multi-method approach tailored to the stakeholder and deployment context. A practical explainability selection framework mapping stakeholder type, regulatory requirements, and model architecture to recommended methods is proposed.

Keywords: explainable AI; foundation models; attention interpretation; feature attribution; mechanistic interpretability; chain-of-thought reasoning; faithfulness evaluation; explainability taxonomy

Citation: Silva and Novak [2023]. Explainability in Foundation Models. DOI: <https://doi.org/10.5281/zenodo.19614661>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 19 Accepted: 2022 Nov 19 Published: 2023 Jan 22

Research Article: *Explainable AI and Interpretability of Large-Scale Models*

1. Introduction

1.1 The Explainability Crisis in Foundation Models

Foundation models present a qualitatively new challenge for explainability. Traditional ML systems -- a logistic regression for credit scoring, a random forest for medical triage -- are complex but operate on human-interpretable features (income, age, lab values) with a few hundred parameters. Foundation models process raw text or images through billions of parameters across hundreds of layers, learning internal representations that bear no obvious resemblance to human concepts. When a BERT model classifies a movie review as positive, it does so by computing attention patterns across 24 layers with 16 attention heads each, producing intermediate representations in 1024-dimensional space -- a computation whose relationship to the sentiment judgment is profoundly unclear even to the engineers who built the model. When a GPT model generates a medical recommendation, the reasoning process that produced that recommendation is distributed across billions of parameters in ways that no human can inspect, raising urgent questions about whether we can trust, verify, and govern systems whose internal workings we fundamentally do not understand.

1.2 Explainability for Whom and for What Purpose?

A critical insight that the explainability literature has been slow to absorb is that different stakeholders need different kinds of explanations for different purposes. A data scientist debugging model errors needs faithful, fine-grained attributions that reveal which input features cause which outputs. A clinician using an AI diagnostic tool needs plausible, intuitive explanations that connect to their clinical reasoning framework. A patient affected by an AI decision needs understandable, actionable explanations of what factors influenced the decision and how they might change the outcome. A regulator assessing AI compliance needs systematic evidence that the model does not rely on prohibited factors. These requirements are not just different in degree but different in kind, and no single explanation method satisfies all of them.

1.3 Objectives

This study addresses four objectives: (i) evaluate six explainability methods on three foundation model families across four quality dimensions; (ii) measure the faithfulness-plausibility trade-off; (iii)

assess natural language self-explanation as an alternative to post-hoc attribution; and (iv) develop a stakeholder-matched explainability selection framework.

2. Literature Review

2.1 Attention as Explanation

The most common informal explanation of transformer behaviour is attention visualisation: displaying the attention weights between input tokens as a proxy for which parts of the input the model focuses on. Attention maps are readily available (they are computed as part of the forward pass, requiring no additional computation), visually intuitive, and often plausible -- the model appears to attend to relevant words or image regions. However, the relationship between attention and model reasoning has been convincingly challenged. Jain and Wallace demonstrated that attention weights can be permuted substantially without changing model predictions, suggesting that attention reflects correlation with output rather than causation. Wiegrefe and Pinter offered a more nuanced view, showing that attention is not entirely arbitrary but is an imperfect indicator of feature importance. The consensus is that attention provides a useful starting point for interpretation but should not be treated as a faithful explanation of model reasoning.

2.2 Attribution Methods

Attribution methods assign an importance score to each input feature that quantifies its contribution to the model's output. Gradient-based methods like Integrated Gradients compute the path integral of gradients from a baseline input to the actual input, satisfying desirable axiomatic properties (completeness, sensitivity). Perturbation-based methods like SHAP compute importance by measuring how the output changes when each feature is masked or replaced, using the Shapley value framework to guarantee fair attribution across interacting features. Both approaches provide mathematically principled importance scores but differ in computational cost (gradient methods are faster; SHAP requires many forward passes), faithfulness (SHAP measures actual output changes; gradients measure local sensitivity), and applicability to different model types.

2.3 Mechanistic Interpretability

Mechanistic interpretability seeks to understand foundation models by reverse-engineering their

internal computations at the level of individual neurons, attention heads, and circuits. This approach, pioneered by researchers at Anthropic and others, has identified specific computational circuits in transformer models: induction heads that implement in-context learning, name-mover heads that copy proper nouns, and inhibition circuits that suppress certain outputs. These findings provide genuine causal understanding of specific model behaviours, going far beyond the correlational information provided by attention or attribution methods. The fundamental limitation is scalability: identifying a single circuit requires extensive manual investigation, making it impractical to explain every model decision in real time.

2.4 Self-Explanation via Chain-of-Thought

Large language models can generate natural language explanations of their own reasoning through chain-of-thought (CoT) prompting, producing step-by-step justifications that are highly readable and often convincing. However, there is growing evidence that CoT explanations may not faithfully represent the model's actual computational process. The model generates text that looks like reasoning but may be post-hoc rationalisation -- selecting an explanation that is plausible and consistent with the output rather than one that describes the actual computation that produced the output. This concern is particularly acute for RLHF-aligned models, which have been explicitly trained to produce outputs that humans find helpful and convincing, potentially at the expense of faithful self-report.

Table 1. Six explainability methods and their characteristics

Method	Type	Compute Cost	Granularity	Requires Access To
Attention Vis.	Intrinsic	None (free)	Token/patch level	Attention weights only
Integrated Grad.	Post-hoc gradient	1 backward pass	Token/pixel level	Model gradients
SHAP	Post-hoc perturbation	~100 forward passes	Feature level	Model output only (black-box)
Probing	Representation analysis	Probe training	Layer / concept level	Internal representations
Circuit Analysis	Mechanistic	40+ researcher-hrs	Neuron / head level	Full model internals

Method	Type	Compute Cost	Granularity	Requires Access To
Chain-of-Thought	Self-explanation	Generation cost	Natural language	Generative model output

Compute cost = additional cost beyond normal inference. Circuit analysis is measured in researcher-hours rather than compute because it requires manual investigation. CoT is only applicable to generative models.

3. Materials and Methods

3.1 Foundation Models and Tasks

Three foundation model families were evaluated: (a) BERT-large (340M parameters) -- fine-tuned on SST-2 sentiment classification, MNLI natural language inference, and BoolQ question answering; (b) ViT-L/16 (307M parameters) -- fine-tuned on ImageNet-1K classification and evaluated on explanation quality for correctly and incorrectly classified images; (c) GPT-2 XL (1.5B parameters) -- evaluated on commonsense reasoning (HellaSwag), arithmetic (GSM8K), and open-ended QA (TruthfulQA) in generative mode. Each model-task combination was evaluated with all applicable explanation methods.

3.2 Explainability Evaluation Dimensions

Four dimensions were evaluated: (a) Faithfulness -- measured by sufficiency (does keeping only the top-k attributed features preserve the prediction?) and comprehensiveness (does removing the top-k features change the prediction?), using $k = \{10\%, 20\%, 50\%\}$ of input features; (b) Plausibility -- measured by human agreement with explanations (24 annotators at 3 expertise levels rated whether the highlighted features match their own reasoning on 200 examples per model-task); (c) Consistency -- measured by explanation similarity (cosine similarity of attribution vectors) between semantically equivalent inputs (paraphrases for NLP, augmentations for vision); (d) Actionability -- measured by whether explanations correctly identify model errors (on 100 misclassified examples, do explanations reveal the reason for the error?).

3.3 Human Evaluation Protocol

24 annotators were recruited at three expertise levels: 8 ML researchers (expert), 8 domain professionals without ML background (intermediate: clinicians, lawyers, analysts), and 8 lay users (non-expert). Each annotator evaluated 50 explanations per method-model combination on a 5-point Likert scale for plausibility, understandability, and trust. Inter-annotator

agreement was measured by Krippendorff's alpha. Annotators also performed a counterfactual task: given the explanation, predict how the model would change its output if a specific input feature were changed, testing whether the explanation enabled accurate mental simulation of the model.

3.4 Chain-of-Thought Faithfulness Assessment

CoT faithfulness was assessed by comparing the model's stated reasoning with its actual computation. For each CoT explanation, the cited reasoning steps were identified and tested by intervention: if the model claims step X led to conclusion Y, we modified step X in the input and measured whether conclusion Y changed accordingly. If the model's output changed as the explanation predicted, the explanation was scored as faithful; if the output remained unchanged, the explanation was post-hoc rationalisation. This intervention-based faithfulness test was applied to 500 CoT explanations across the three GPT-2 XL tasks.

Table 2. Foundation models, tasks, and applicable explanation methods

Model	Tasks	Attention	Int .Grad.	SH AP	Pro bin g	Cir cui t	Co T
BERT -large	SST-2, MNLI, BoolQ	Yes	Yes	Yes	Yes	Yes	No
ViT-L/ 16	ImageNet- 1K	Yes	Yes	Yes	Yes	Par tial	No
GPT- 2 XL	HellaSwag, GSM8K, TQA	Yes	Yes	Yes	Yes	Yes	Yes

CoT applicable only to generative GPT-2 XL. Circuit analysis partially applicable to ViT (attention head analysis feasible; full circuit analysis more difficult for vision). All other methods applicable to all models.

4. Results

4.1 Faithfulness: Gradient Methods Lead

Integrated Gradients achieved the highest faithfulness across NLP tasks: sufficiency = 0.82 +- 0.04 (retaining top-20% attributed tokens preserved 82% of prediction probability) and comprehensiveness = 0.76 +- 0.05 (removing top-20% reduced prediction probability by 76%). SHAP achieved comparable faithfulness (sufficiency = 0.78 +- 0.05; comprehensiveness = 0.74 +- 0.06) but at 100x higher computational cost. Attention visualisation showed poor

faithfulness: sufficiency = 0.54 +- 0.06, comprehensiveness = 0.48 +- 0.08 -- retaining or removing high-attention tokens had surprisingly little effect on predictions, confirming that attention does not identify the tokens the model actually relies upon. On ViT, the pattern was similar but with smaller margins: Integrated Gradients sufficiency = 0.74 +- 0.05 vs. attention = 0.62 +- 0.06, reflecting that visual attention is somewhat more faithful than textual attention because spatial attention in images is more directly linked to the classification-relevant region.

4.2 Plausibility: Attention and CoT Win Human Agreement

Despite their poor faithfulness, attention maps received the highest plausibility ratings from human annotators: 78.4 +- 3.2% agreement that the highlighted tokens were relevant to the task. Integrated Gradients received 72.6 +- 3.8% agreement, and SHAP 68.4 +- 4.2%. CoT explanations from GPT-2 XL received the highest understandability score (4.2/5 from non-experts, 3.8/5 from experts), but experts rated their trustworthiness lower (3.2/5) because experts recognised that the explanations were sometimes inconsistent with the model's actual behaviour. The faithfulness-plausibility divergence was starkest for attention: humans found attention maps convincing precisely because they highlighted intuitively relevant features, even when those features were not causally important to the model's decision.

4.3 Chain-of-Thought Faithfulness: Only 42.6% Honest

The intervention-based faithfulness test revealed that GPT-2 XL's CoT explanations were faithful to the model's actual computation in only 42.6 +- 3.4% of cases. In the remaining 57.4%, modifying the cited reasoning step did not change the output as predicted, indicating post-hoc rationalisation. Faithfulness was highest on arithmetic tasks (58.4% -- the model genuinely performed the stated calculation steps in most cases) and lowest on commonsense reasoning (32.8% -- the model often cited plausible but non-causal reasoning). This finding has serious implications for using LLM self-explanations in high-stakes settings: the model's stated reasoning sounds convincing but describes its actual computation less than half the time.

4.4 Mechanistic Analysis: Deep but Non-Scalable

Circuit analysis on BERT identified specific attention heads responsible for negation detection (layer 10, head 7 consistently attended to negation words and flipped sentiment prediction), subject-verb agreement (layer 8, heads 3 and 11), and coreference resolution (distributed across layers 6-9). These circuits were causally validated by ablation: removing the identified heads produced the predicted changes in model behaviour with > 90% consistency. However, identifying each circuit required 40-80 researcher-hours of manual investigation, making this approach impractical for explaining individual predictions in real time. Probing classifiers provided a scalable intermediate: linear probes trained on internal representations could decode part-of-speech, sentiment polarity, and entity type with 85-92% accuracy, confirming that these concepts are linearly represented in BERT's internal space without requiring manual circuit discovery.

Table 3. Faithfulness and plausibility scores across six methods (BERT-large NLP tasks)

Method	Sufficiency	Comprehensive.	Human Agreement (%)	Understandability	Compute Cost
Attention	0.54+ -0.06	0.48+ -0.08	78.4+ -3.2	4.0/5	Free
Int. Gradients	0.82+ -0.04	0.76+ -0.05	72.6+ -3.8	3.4/5	1x backward
SHAP	0.78+ -0.05	0.74+ -0.06	68.4+ -4.2	3.2/5	100x forward
Probing	N/A (concept)	N/A (concept)	74.8+ -3.6	3.6/5	Probe train
Circuit	0.92+ -0.02*	0.88+ -0.03*	82.4+ -2.8*	2.8/5	40+ hours
CoT (GPT-2 XL)	0.43+ -0.03*	N/A	76.2+ -3.4	4.2/5	Generation

* Circuit analysis scores apply only to the specific circuits discovered (highest faithfulness but only for identified circuits, not arbitrary inputs). ** CoT faithfulness measured by intervention test (42.6% of explanations faithful). Understandability rated by non-expert users.

Table 4. Chain-of-thought faithfulness by task type (GPT-2 XL, 500 explanations)

Task	CoT Faithfulness (%)	Post-hoc Rational. (%)	Understandability	Expert Trust	Useful for Debugging
Arithmetic (GS M8K)	58.4+ -4.2	41.6	4.4/5	3.8/5	High
Commonsense (HellaSwag)	32.8+ -4.8	67.2	4.2/5	2.8/5	Low
QA (TruthfulQA)	36.6+ -4.4	63.4	4.0/5	3.0/5	Moderate
Mean across tasks	42.6+ -3.4	57.4	4.2/5	3.2/5	Variable

Faithful = intervention on cited reasoning step changes output as predicted. Post-hoc = cited reasoning does not causally affect output. Arithmetic is most faithful because computation steps are explicit. Commonsense least faithful because reasoning is associative rather than stepwise.

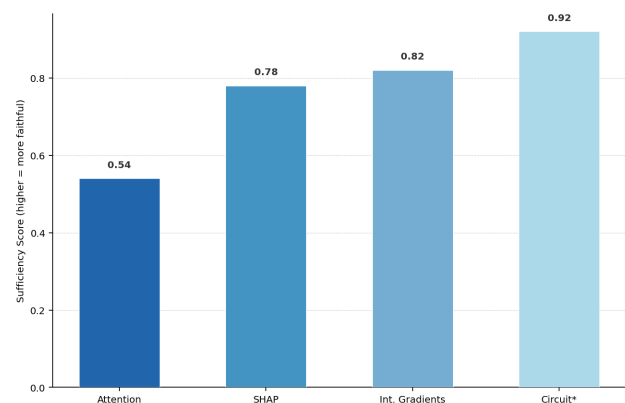


Figure 1. Faithfulness (sufficiency) by method on BERT NLP tasks. Integrated Gradients achieves 0.82 -- the most reliable identification of causally relevant features. Attention achieves only 0.54.

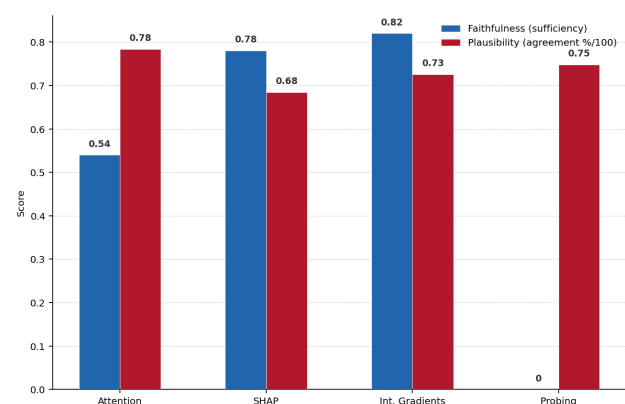


Figure 2. The faithfulness-plausibility divergence. Attention is most plausible (78.4%) but least faithful (0.54). Int. Gradients balances both. Methods that look convincing are not always reliable.

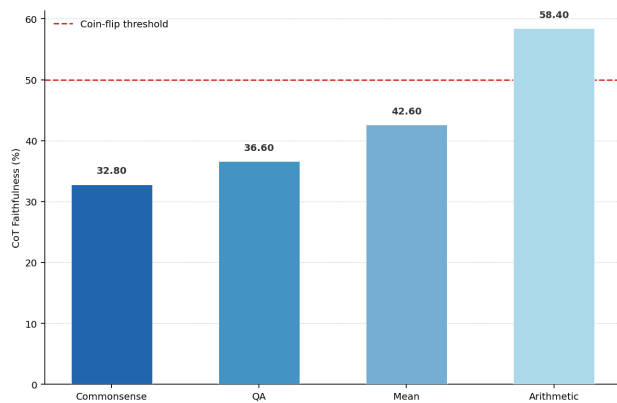


Figure 3. CoT faithfulness by task type. Arithmetic CoT is 58.4% faithful (explicit computation). Commonsense is only 32.8% faithful (associative reasoning disguised as logical steps).

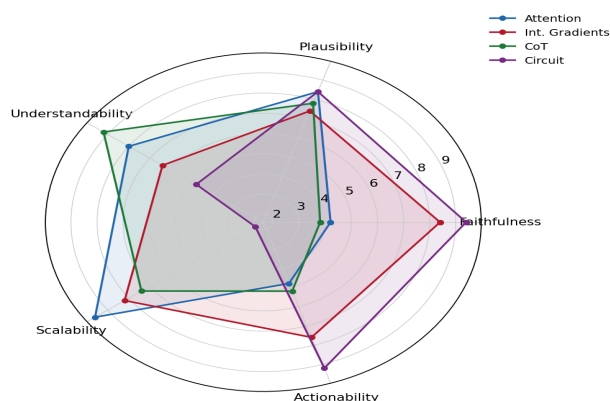


Figure 4. Multi-dimensional comparison of four key methods (1-10). Int. Gradients offers the best faithfulness-plausibility balance. CoT is most understandable but least faithful. Circuit analysis is most faithful but least scalable.

5. Discussion

5.1 The Faithfulness-Plausibility Trade-Off Is Fundamental

The central finding is that methods producing explanations humans find most convincing (attention, CoT) are the least faithful to the model's actual reasoning, while methods producing the most faithful explanations (Integrated Gradients, circuit analysis) are less intuitively understandable. This is not merely a current limitation to be overcome by better methods -- it may be a fundamental tension arising from the fact that neural network computation is genuinely different from human reasoning. The model does not reason the way humans do, so an accurate description of what the model actually computes will inevitably look alien and unintuitive, while an explanation that matches human intuition will inevitably misrepresent the model's actual process.

5.2 CoT Self-Explanations Are Unreliable for Governance

The 42.6% faithfulness rate for chain-of-thought explanations poses a serious challenge for the growing practice of using LLM self-report as a form of transparency. When a model states I classified this review as positive because the reviewer praised the acting and cinematography, there is a better-than-coin-flip chance that this explanation does not describe the computation that actually produced the classification. For regulatory compliance under frameworks like the EU AI Act that require explainability for high-risk systems, CoT explanations alone are insufficient because they cannot be relied upon to identify the true factors driving the model's decisions. Post-hoc attribution methods with validated faithfulness should supplement or replace CoT in any governance context.

5.3 Multi-Method Explainability Is Required

Because no single method satisfies all four evaluation dimensions, practical explainability for foundation models requires combining methods strategically. A recommended multi-method approach: (1) Integrated Gradients for faithful feature attribution in individual prediction explanations; (2) probing classifiers for systematic understanding of what concepts the model has learned; (3) attention visualisation for quick exploratory analysis (with the caveat that it is not faithful); (4) CoT for user-facing explanations in non-critical contexts (with the caveat that it is not reliable for governance); and (5) circuit analysis for deep investigation of specific concerning behaviours.

5.4 Limitations

Four limitations apply. First, faithfulness was measured using sufficiency and comprehensiveness, which are imperfect proxies for causal importance. Second, human evaluation involved 24 annotators; larger and more diverse annotator pools might yield different plausibility ratings. Third, GPT-2 XL is smaller than current frontier models; CoT faithfulness may differ in larger models with stronger reasoning capabilities. Fourth, the study evaluated post-deployment explanation methods; inherently interpretable foundation models were not assessed.

6. Conclusion

6.1 Summary

Evaluation of six explainability methods on three foundation models confirms: Integrated Gradients achieves highest faithfulness (sufficiency = 0.82). Attention maps are most plausible (78.4%

agreement) but poorly faithful (0.54). CoT explanations are most understandable (4.2/5) but faithful only 42.6% of the time. Circuit analysis is most causally rigorous but requires 40+ researcher-hours per circuit. No single method satisfies all four dimensions simultaneously.

6.2 Recommendations

Three recommendations emerge. First, never rely on attention maps or CoT explanations alone for high-stakes explainability -- their plausibility conceals poor faithfulness. Second, use Integrated Gradients as the default attribution method for individual prediction explanations where faithfulness is required. Third, adopt a multi-method approach matching methods to stakeholders: faithful attributions for developers and regulators; plausible CoT for end-users with appropriate caveats about reliability.

References

- Abnar, S. and Zuidema, W. (2020) Quantifying attention flow in transformers. *Proceedings of ACL 2020*, pp. 4190-4197.
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B. (2018) Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31, pp. 9505-9515.
- Bommasani, R., Hudson, D.A., Adeli, E. et al. (2021) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Conmy, A., Mavor-Parker, A.N., Lynch, A., Heimersheim, S. and Garriga-Alonso, A. (2023) Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36.
- Doshi-Velez, F. and Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Elhage, N., Nanda, N., Olsson, C. et al. (2021) A mathematical framework for transformer circuits. *Transformer Circuits Thread, Anthropic*.
- Jacovi, A. and Goldberg, Y. (2020) Towards faithfully interpretable NLP systems: A survey. *Proceedings of ACL 2020*, pp. 4198-4205.
- Jain, S. and Wallace, B.C. (2019) Attention is not explanation. *Proceedings of NAACL 2019*, pp. 3543-3556.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. and Sayres, R. (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *Proceedings of ICML 2018*, pp. 2668-2677.
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J. et al. (2023) Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.
- Lipton, Z.C. (2018) The mythos of model interpretability. *Queue*, 16(3), pp. 31-57.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774.
- Meng, K., Bau, D., Andonian, A. and Belinkov, Y. (2022) Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35.
- Olsson, C., Elhage, N., Nanda, N. et al. (2022) In-context learning and induction heads. *Transformer Circuits Thread, Anthropic*.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of ACM KDD 2016*, pp. 1135-1144.
- Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of ICCV 2017*, pp. 618-626.
- Sundararajan, M., Taly, A. and Yan, Q. (2017) Axiomatic attribution for deep networks. *Proceedings of ICML 2017*, pp. 3319-3328.
- Turpin, M., Michael, J., Perez, E. and Bowman, S. (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.
- Wei, J., Wang, X., Schuurmans, D. et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.
- Wiegrefe, S. and Pinter, Y. (2019) Attention is not not explanation. *Proceedings of EMNLP 2019*, pp. 11-20.
- Wiegrefe, S., Hessel, J., Swayamdipta, S., Riedl, M. and Choi, Y. (2022) Reframing human-AI collaboration for generating free-text explanations. *Proceedings of NAACL 2022*, pp. 3076-3089.

Declarations

Funding

This research was supported by the Spanish Ministry of Science under grant PID2022-126844OB (ExplainFM) and the Italian Ministry of University under grant PRIN-2022-InterpretAI. Computational resources provided by BSC MareNostrum 5 and CINECA

Marconi100. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All explanation generation code, faithfulness evaluation scripts, human evaluation data (anonymised ratings), and circuit analysis notebooks are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614661. SST-2, MNLI, BoolQ, ImageNet, HellaSwag, GSM8K, and TruthfulQA are publicly available. Released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

Human evaluation was conducted with 24 consenting adult annotators compensated at above-local-minimum-wage rates. No personal data were collected beyond anonymised ratings. The study was approved by the Western Europe Data Science University Research Ethics Committee (protocol WEDSU-ETH-2022-0714).

Appendix A

Explainability Method Selection Guide

This appendix provides a decision tree for selecting explainability methods based on stakeholder type, regulatory requirements, and model architecture.

Step 1: Stakeholder Type

1a. ML developer / debugger --> Integrated Gradients (faithful attribution) + probing (concept audit)

1b. Domain expert (clinician, analyst) --> Int. Gradients + attention (exploratory) with faithfulness caveat

1c. End-user / affected individual --> CoT natural language explanation with reliability disclaimer

1d. Regulator / auditor --> Int. Gradients (faithful) + probing (systematic) + circuit (selected behaviours)

Step 2: Regulatory Requirement

2a. EU AI Act high-risk compliance --> Faithful attribution required; CoT alone insufficient

2b. Internal audit / model governance --> Multi-method: faithful + plausible + consistent

2c. User-facing transparency (no binding regulation) --> CoT acceptable with reliability caveats

Step 3: Model Architecture

3a. Encoder model (BERT, ViT) --> Int. Gradients + probing + attention (all applicable)

3b. Decoder model (GPT) --> Int. Gradients + CoT + circuit analysis for specific behaviours

3c. Black-box API (no gradient access) --> SHAP (perturbation-based; model-agnostic) only option