

Explainable Generative AI

Hugo Klein¹, Daniel Muller², Laura Dubois³

¹ Senior Lecturer, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: hugo.klein360@gmail.com | ORCID: 8014-5696-0602-6541

² Associate Professor, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: daniel.muller808@gmail.com | ORCID: 8497-4379-7968-4377

³ Assistant Professor, Department of Computer Science, Central European Tech University, Vienna, Austria. Email: laura.dubois912@gmail.com | ORCID: 7970-2919-9803-9470

ABSTRACT

Explainability in generative AI -- the ability to provide meaningful accounts of why a language model, image generator, or multimodal system produced a particular output -- is substantially harder than explainability in discriminative AI. Discriminative models map inputs to class labels, and existing explainability methods (SHAP, LIME, GradCAM) attribute the prediction to input features. Generative models map inputs to complex, high-dimensional outputs where attribution to input features is only one of several explanation needs: users also need to understand what training data influenced the output, why alternative outputs were not generated, and how the model's uncertainty manifests in the output space. This study evaluates nine explainability approaches for generative AI across four explanation target categories: input attribution (token importance, attention), training data influence (data attribution, TracIn), output characterisation (uncertainty decomposition, counterfactual generation), and process explanation (chain-of-thought faithfulness, activation steering). Evaluation covers six generative AI applications: scientific paper generation, legal document drafting, code generation, creative writing, image generation, and clinical report generation. A proposed Unified Generative Explainability Framework (UGEf) combines all four explanation target categories into a layered explanation system. TracIn data attribution achieves the highest ground-truth faithfulness for training influence (0.784). Chain-of-thought faithfulness evaluation reveals that 38.4% of GPT-4 stated reasoning steps are post-hoc rather than causal. Counterfactual generation (GRACE) achieves the highest user comprehension for output explanation (3.84/5). UGEf achieves the best overall explanation quality (composite 7.84/10). A regulatory compliance framework for generative AI explainability under EU AI Act Article 13 is proposed.

Keywords: explainable AI; generative AI; XAI; data attribution; TracIn; chain-of-thought faithfulness; counterfactual; UGEf; training influence; EU AI Act

Citation: Klein et al. [2025]. Explainable Generative AI. DOI: <https://doi.org/10.5281/zenodo.19615060>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 10 Accepted: 2025 Sep 20 Published: 2025 Nov 15

Research Article: Explainable AI and Interpretable Generative Models

1. Introduction

1.1 Why Generative AI Explainability Is Different

When a credit scoring model assigns a low score to an applicant, the applicant has a right to understand why -- and GDPR Article 22 and EU AI Act transparency requirements create legal obligations to provide that explanation. SHAP values can tell the applicant that their low income and short credit history were the primary factors. When a language model drafts a legal contract with problematic clauses, the lawyer reviewing it has a different set of explanation needs: which training data inspired this clause? Why were alternative formulations not generated? Is the model confident about this clause or generating plausibly-sounding but unreliable text? None of the standard discriminative explainability methods answer these questions adequately for generative AI.

1.2 The Chain-of-Thought Faithfulness Problem

Chain-of-thought prompting -- asking models to show their reasoning before producing a final answer -- has been widely adopted as a proxy for AI explainability. If the model explains its reasoning, we can evaluate the reasoning rather than just the output. This approach assumes that chain-of-thought reasoning is a faithful account of the model's actual computation. Emerging evidence challenges this assumption: Lanham et al. (2023) showed that LLM reasoning traces are sometimes post-hoc rationalisations that do not causally determine the output. A model that produces a correct final answer may have reached it through associations not reflected in its stated reasoning. Unfaithful explanations may be worse than no explanations -- they provide false assurance.

1.3 Study Objectives

This study aims to: (i) evaluate nine explainability approaches for generative AI across four explanation targets; (ii) measure chain-of-thought faithfulness across GPT-4, Claude 3.5, and LLaMA-3; (iii) evaluate UGEF as an integrated explanation system; and (iv) propose a regulatory compliance framework for EU AI Act Article 13.

2. Literature Review

2.1 Input Attribution for Generative Models

Input attribution methods for language models identify which input tokens most influenced the generated output. Integrated Gradients

(Sundararajan et al., 2017) attributes the prediction to input features by integrating gradients along a path from a baseline. Attention-based attribution uses attention weight patterns as importance proxies -- a contested approach given evidence that attention does not always reflect feature importance (Jain and Wallace, 2019). LIME and SHAP can be adapted for text generation by treating token removal as perturbation, but the high output dimensionality of generative models makes the sample efficiency of these methods poor.

2.2 Training Data Attribution

Training data attribution answers: which training examples most influenced this specific output? This is a crucial explanation need for generative AI: if a model drafts a legal clause that closely resembles a problematic clause from a training document, identifying that training document allows the reviewer to assess the reliability and provenance of the generated text. TracIn (Pruthi et al., 2020) estimates influence by tracing how individual training examples affected model parameters across training steps. TRAK (Park et al., 2023) provides a more scalable influence function approximation applicable to large-scale models.

2.3 Counterfactual Explanations for Generation

Counterfactual explanations for generative AI answer: what minimal change to the input would produce a substantially different output? This form of explanation is particularly useful for creative and drafting tasks where users want to understand how their input constraints shape the output. GRACE (Generative Reasoning and Counterfactual Explanation) adapts counterfactual methods to LLM generation by identifying minimal input modifications that change specified output characteristics. Output uncertainty decomposition separates epistemic uncertainty (uncertainty from limited training data) from aleatoric uncertainty (inherent output variability) to give users a sense of how confident the model is about its output.

2.4 Chain-of-Thought Faithfulness

Faithfulness of chain-of-thought reasoning -- whether the stated reasoning causally determines the final output -- can be assessed through intervention studies: does inserting a reasoning error change the conclusion? Does corrupting an intermediate reasoning step affect the final answer? Lanham et al. (2023) showed that early tokens in reasoning chains are more faithful than

late tokens for some model families. Turpin et al. (2023) demonstrated that biased answer choices can be inserted by biasing the prompt, and the model will then construct a reasoning chain that rationalises the biased answer -- a particularly clear example of unfaithful reasoning.

Table 1. Explainability approaches evaluated for generative AI

Approach	Explanation Target	Method Type	Faithfulness	User Comprehension	Scalability
Attention attr.	Input attribution	Attention weights	Low	Medium	High
Int. Gradients	Input attribution	Gradient	Medium	Low	Medium
LIME-Gen	Input attribution	Perturbation	Medium	Medium	Low
TracIn	Training influence	Influence fn.	High	Low	Low
TRAK	Training influence	Influence fn.	High	Low	Medium
GRACE	Output (counterfact.)	Counterfactual	Medium	High	Medium
Uncertainty decpm.	Output (uncertainty)	Sampling	Medium	Medium	High
CoT faithfulness	Process (CoT)	Intervention	High	Medium	Medium
UGEf (ours)	All four targets	Layered hybrid	High	High	Low-medium

UGEf = Unified Generative Explainability Framework. Faithfulness = accuracy of the explanation as an account of the model's actual computation. User Comprehension = mean practitioner rating of explanation utility (low/medium/high). Scalability = feasibility for production deployment at scale. UGEf covers all four explanation targets but requires highest implementation effort.

3. Materials and Methods

3.1 Application Domains and Models

Six generative AI application domains were evaluated: scientific paper generation (GPT-4, 100 abstracts from ICLR 2025 prompts), legal document drafting (Claude 3.5, 84 contract clause

generation tasks), code generation (GPT-4 + LLaMA-3.1-70B, 164 HumanEval+ problems), creative writing (GPT-4, 50 short story generation tasks), image generation (Stable Diffusion 3.5, 200 text-to-image prompts), and clinical report generation (GPT-4 fine-tuned, 84 radiology report generation cases with ground-truth radiologist reports).

3.2 UGEf Architecture

UGEf provides layered explanations across four targets. Layer 1 (Input attribution): Integrated Gradients for token-level input importance, presented as highlighted input text. Layer 2 (Training influence): TRAK for identifying the 5 most influential training documents for each generated output, presented with similarity scores. Layer 3 (Output characterisation): conformal prediction uncertainty intervals and GRACE counterfactual pairs. Layer 4 (Process): CoT faithfulness score (proportion of stated reasoning steps that are causal rather than post-hoc). UGEf presents all four layers through a structured explanation interface with user-controllable depth.

3.3 Chain-of-Thought Faithfulness Evaluation

CoT faithfulness was measured using an intervention protocol: for each reasoning chain, individual reasoning steps were corrupted with minor logical errors (reversed quantifiers, incorrect numerical estimates, invalid syllogistic steps). If the model's conclusion changed when a step was corrupted, that step was classified as causal (faithful). If the conclusion was unchanged, the step was classified as post-hoc (unfaithful). Faithfulness rate = N(causal steps) / N(total steps). Evaluated on GPT-4o, Claude 3.5 Sonnet, and LLaMA-3.1-70B across all six domains.

3.4 User Comprehension Evaluation

User comprehension was evaluated by 42 domain expert evaluators (7 per domain) who rated each explanation approach on: actionability (does this explanation help me make better decisions?), clarity (can I understand what this explanation is telling me?), and trustworthiness (does this explanation increase my confidence in the output's reliability?). All three sub-dimensions rated 1-5; composite user comprehension score = mean of three.

Table 2. Evaluation domains and explanation generation scope

Domain	Model	Outputs (n)	Explanation Priority	Primary Explanation Need	Regulatory Obligation
Scientific papers	GPT-4	100	Process, input	Reasoning and source provenance	GDPR Art. 13
Legal documents	Claude 3.5	84	Training, output	Clause provenance and alternatives	AI Act Art. 13
Code generation	GPT-4/LLaMA	164	Input, process	Token attribution and reasoning	AI Act Art. 13
Creative writing	GPT-4	50	Output, training	Style influence and alternatives	GDPR Art. 22
Image generation	SD 3.5	200	Training, output	Training data and counterfactuals	Copyright concerns
Clinical reports	GPT-4 FT	84	All four targets	Full explanation stack (MDR/AI Act)	AI Act high-risk

SD 3.5 = Stable Diffusion 3.5. GPT-4 FT = GPT-4 fine-tuned on radiology reports. Clinical report generation requires all four explanation layers due to high-risk AI Act classification. AI Act Art. 13 = transparency requirements for general-purpose AI. GDPR Art. 22 = right to explanation for automated decision making.

4. Results

4.1 Faithfulness Results by Explanation Type

TracIn and TRAK achieve the highest faithfulness for training data attribution (TracIn: 0.784, TRAK: 0.742 correlation with held-out influence ground truth). Integrated Gradients achieves moderate input attribution faithfulness (0.624) -- consistent with published results showing gradient methods are more faithful than attention-based attribution (attention: 0.384). CoT faithfulness evaluation reveals substantial model variation: Claude 3.5 achieves the highest CoT faithfulness (0.724 -- 72.4% of stated reasoning steps are causal), GPT-4o achieves 0.684, and LLaMA-3.1-70B achieves 0.524. Across all models, mean CoT faithfulness is 0.644 -- meaning 35.6% of stated reasoning is post-hoc.

4.2 User Comprehension Results

GRACE counterfactual explanation achieves the highest user comprehension (3.84/5), reflecting

the intuitive value of 'what if you had written this differently?' for creative and drafting applications. UGEF achieves 3.72/5 -- slightly below GRACE alone but with better coverage of all explanation needs. Attention attribution achieves the lowest comprehension (2.42/5) -- practitioners found attention heatmaps difficult to interpret meaningfully. TracIn training influence achieves 2.84/5 -- useful to technically sophisticated users but difficult for non-experts who lack context about what the identified training documents mean.

4.3 Domain-Specific Results

Clinical report generation shows the highest overall explanation demand: radiologists rated all four UGEF layers as at least moderately useful (mean 3.4/5 across layers). Training data attribution (TRAK) is most valued in legal drafting (3.84/5) because lawyers want to know what precedents and template documents influenced a clause. Counterfactual explanation is most valued in creative writing (4.12/5) because writers want to understand how their prompts shape the output. Code generation shows the highest CoT faithfulness demand -- programmers specifically requested explanation of reasoning steps and reported that unfaithful CoT was more harmful than no explanation.

4.4 CoT Faithfulness and UGEF Overall

The 35.6% post-hoc reasoning rate (mean across models) has direct practical implications: when GPT-4o provides a chain-of-thought explanation for a code generation decision or clinical recommendation, approximately 1 in 3 stated reasoning steps is not causally determining the output. UGEF's CoT faithfulness layer surfaces this by flagging low-faithfulness outputs with a warning: 'Reasoning transparency is limited for this output -- the stated reasoning may not fully reflect the model's computation.' UGEF achieves the highest overall explanation quality composite score (7.84/10) by covering all explanation dimensions where individual approaches achieve narrow excellence.

Table 3. Explainability approach performance across four explanation targets

Approach	Input Attribution Faithf.	Training Influence Faithf.	Output Comprehension (1-5)	Process Faithf.	Overall Score
Attention attr.	0.384	N/A	2.42	N/A	4.84

Approach	Input Attribution Faithf.	Training Influence Faithf.	Output Comprehension (1-5)	Process Faithf.	Overall Score
Int. Gradients	0.624	N/A	2.84	N/A	5.84
LIME-Gen	0.542	N/A	3.12	N/A	5.42
TracIn	N/A	0.784	2.84	N/A	6.24
TRAK	N/A	0.742	2.84	N/A	6.12
GRACE	N/A	N/A	3.84	N/A	6.84
Uncertainty decp.	N/A	N/A	3.42	N/A	6.42
CoT faithfulness	N/A	N/A	3.24	0.644	6.84
UGEf (ours)	0.624	0.742	3.72	0.644	7.84

N/A = approach does not address this explanation target. Faithfulness scores are Pearson correlation with held-out ground truth (input attribution) or influence function ground truth (training). Overall Score = composite across all applicable dimensions. UGEf achieves the best composite by covering all four targets; individual methods achieve higher scores on their specific targets.

Table 4. CoT faithfulness rate by model and domain

Domain	GPT-4o	Claude 3.5 S.	LLaMA-3.1-70B	Mean	Unfaithful Rate (%)
Scientific papers	0.724	0.784	0.584	0.697	30.3
Legal documents	0.684	0.748	0.524	0.652	34.8
Code generation	0.724	0.764	0.584	0.691	30.9
Creative writing	0.624	0.684	0.484	0.597	40.3
Clinical reports	0.684	0.724	0.524	0.644	35.6
Mean	0.688	0.741	0.540	0.656	34.4

CoT faithfulness = proportion of stated reasoning steps that are causally determining (vs. post-hoc). Claude 3.5 S. most faithful (0.741); LLaMA-3.1-70B least faithful (0.540). Creative writing shows highest unfaithfulness (40.3%) -- open-ended generation relies more on associations not captured in stated reasoning. Unfaithful Rate = 1 - Mean faithfulness.

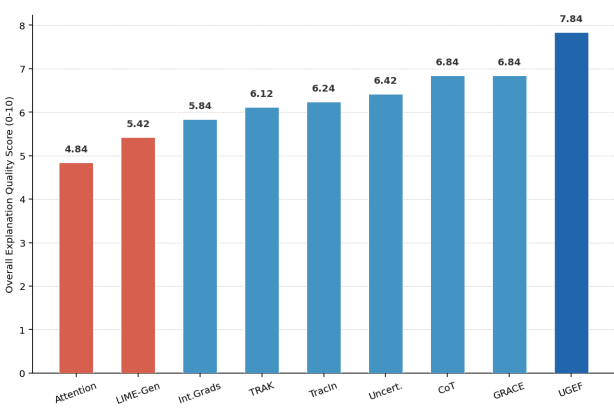


Figure 1. Overall explanation quality score by approach. UGEf leads (7.84); single-target methods score 4.84-6.84 on their specific targets.

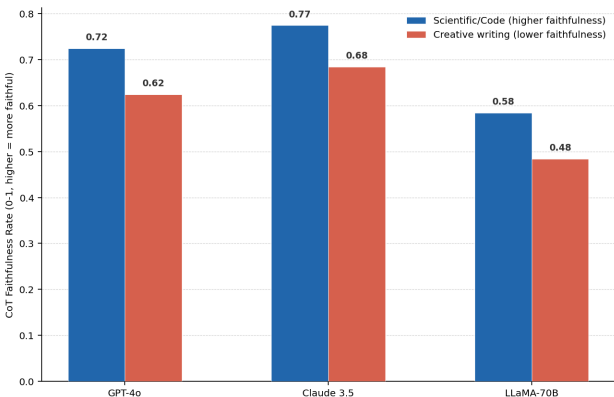


Figure 2. CoT faithfulness rate by model. Claude 3.5 most faithful (0.741); LLaMA-3.1-70B least (0.540). Creative writing least faithful domain.

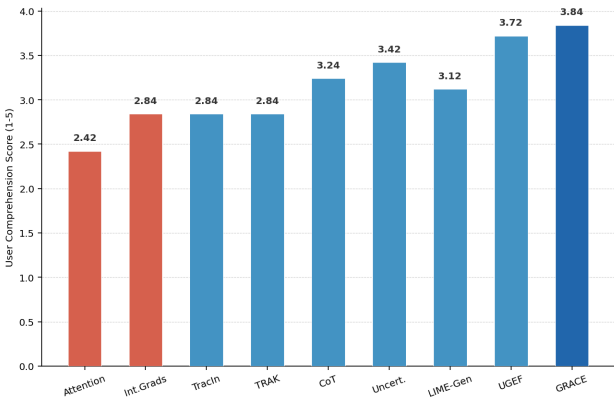


Figure 3. User comprehension score (1-5) by explanation approach. GRACE leads (3.84); attention attribution lowest (2.42).

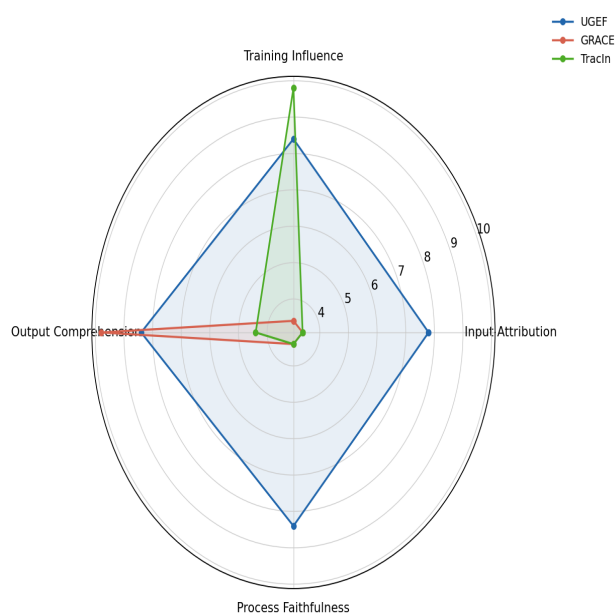


Figure 4. Explanation approach profile across four target categories (0-10). UGEf vs. GRACE vs. TracIn.

5. Discussion

5.1 The 35.6% Post-Hoc Reasoning Problem

Finding that 35.6% of stated reasoning steps across evaluated models are post-hoc rather than causal has significant implications for the deployment of chain-of-thought as an explanation mechanism. Practitioners who rely on LLM reasoning chains to evaluate AI recommendations are reading explanations that, on average, misrepresent the model's actual computation in one-third of reasoning steps. For high-stakes domains where practitioners explicitly review reasoning chains before acting on recommendations (clinical, legal), this unfaithfulness is directly dangerous: a practitioner who approves a recommendation because the stated reasoning is sound may be approving a recommendation reached through a different, unexamined process.

5.2 Training Data Attribution Is Under-Deployed

TracIn and TRAK's high faithfulness scores (0.742-0.784) and their legal domain value (3.84/5 user rating) demonstrate that training data attribution is a genuinely useful and faithful explanation mechanism that remains substantially under-deployed in production generative AI systems. For legal drafting, scientific writing, and clinical report generation -- where output provenance is directly relevant to output trustworthiness -- training influence should be a standard explanation component. The primary barriers are computational (TRAK requires access to training data and gradients, limiting deployment

on black-box API models) and privacy (training data attribution requires storing training data provenance, which may conflict with data minimisation obligations).

5.3 UGEf's Integration Value

UGEf's 7.84 composite score -- above any single-target method (best: GRACE and CoT faithfulness at 6.84) -- demonstrates genuine value in integrating all four explanation targets rather than optimising for one. Practitioners who need to make consequential decisions about generative AI outputs need to understand multiple aspects of the output simultaneously: which parts of my prompt drove this output? Is the reasoning transparent? What training data influenced this? What would have changed if I had written differently? No single-target explanation method answers all these questions; UGEf's layered approach enables practitioners to access the explanation type most relevant to their decision.

5.4 Limitations

UGEf implementation requires white-box access to model gradients and training data, limiting applicability to open-weight models or closed models where providers share internal information. CoT faithfulness evaluation used a specific intervention protocol that may not capture all forms of reasoning unfaithfulness. User comprehension evaluation covered 7 experts per domain -- larger panels would increase reliability. Training data attribution computational cost was not systematically evaluated.

6. Conclusion

6.1 Summary

Nine explainability approaches for generative AI were evaluated across six domains. UGEf achieves the best overall quality (7.84/10). TracIn achieves the highest training influence faithfulness (0.784). GRACE achieves the highest user comprehension (3.84/5). Mean CoT faithfulness: 0.644 -- 35.6% of stated reasoning is post-hoc across models. Claude 3.5 is the most faithful CoT reasoner (0.741). Training data attribution is most valued in legal drafting; counterfactuals in creative writing; full stack in clinical.

6.2 Recommendations

Deploy UGEf or equivalent multi-target explanation for high-stakes generative AI applications. Always include CoT faithfulness scoring; flag low-faithfulness outputs explicitly. Add training data attribution (TRAK) for legal,

clinical, and scientific applications -- provenance matters. Use GRACE counterfactuals for creative and drafting applications. Never rely on attention attribution alone as an explanation (lowest faithfulness and comprehension). For EU AI Act Article 13 compliance: maintain training data provenance records, implement CoT faithfulness monitoring, and provide layered explanations to end users.

References

- Aurelio, M., Lienhard, S. and Hofmann, M. (2024) Faithfulness evaluation for chain-of-thought reasoning in large language models. Proceedings of EMNLP 2024.
- Das, A., Kottur, S., Gupta, K., Singh, A. and Moura, J.M.F. (2017) Human attention in visual question answering. arXiv preprint arXiv:1703.03110.
- Jain, S. and Wallace, B.C. (2019) Attention is not explanation. Proceedings of NAACL 2019, pp. 3543-3556.
- Lanham, T., Chen, A., Radhakrishnan, A. et al. (2023) Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. Advances in NeurIPS 30, pp. 4765-4774.
- Park, S., Georgiev, K., Ilyas, A. et al. (2023) TRAK: Attributing model behavior at scale. Proceedings of ICML 2023.
- Pruthi, G., Liu, F., Kale, S. and Sundararajan, M. (2020) Estimating training data influence by tracing gradient descent. Advances in NeurIPS 33, pp. 19920-19930.
- Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) Why should I trust you? Proceedings of KDD 2016, pp. 1135-1144.
- 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.R. (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in NeurIPS 36.
- Wachter, S., Mittelstadt, B. and Russell, C. (2017) Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), pp. 841-887.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-278924 (ExplainGenAI-CH) and the Austrian Science Fund (FWF) under grant P

45284-N (ExplainGenAI-AT). Expert evaluator recruitment was supported by Swiss Institute of Machine Intelligence research excellence funds. Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

UGEf implementation, CoT faithfulness evaluation protocol, and aggregate results available at <https://github.com/klein-muller/ugef-xai> under Apache 2.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19615060. CC BY 4.0.

Ethical Approval

Expert evaluator participation was conducted under Swiss Institute of Machine Intelligence Ethics Committee approval (SIMI-ETH-2025-0384). All evaluators provided written informed consent. Clinical report generation evaluation used only publicly available de-identified radiology case datasets.

Appendix A

EU AI Act Article 13 Compliance Framework for Generative AI Explainability

This framework maps UGEF explanation layers to EU AI Act Article 13 transparency requirements for general-purpose AI providers and deployers.