Towards Structural Explainability for Reliable and Governable Large Language Models

Marco Aspromonte[1,2]

[1]Department of Computer and Control Engineering, Politecnico di Torino, Corso Duca degli Abruzzi 24, Turin, Italy
[2]Responsible AI, Intesa Sanpaolo, Via Monte di Pietà, Milano, Italy

Abstract

Large language models (LLMs) and multimodal generative AI systems are increasingly deployed in high-stakes, regulated domains, including finance, law, and content moderation, where transparency, auditability, and controllability are not optional but legal requirements. Despite impressive capabilities, these models remain structurally opaque: current explainability techniques are largely post-hoc, heuristic, and feature-level, offering limited insight into how semantic concepts are represented internally and providing little support for principled behavioural control or compliance monitoring. This gap undermines trust, regulatory auditability, and the safe deployment of AI in critical sectors. My PhD research aims to develop mechanistic, structural explainability methods that operate at the level of internal representations and circuits. The overarching goals are to characterize how semantic concepts emerge across layers, evaluate the stability of these representations under perturbations and domain shifts, and leverage them for targeted interventions that enforce safety and governance policies without degrading performance. Preliminary results demonstrate multiple avenues of progress: (i) representation-level diagnostics for factuality and multi-language fact-checking, including post-hoc explainability of content moderation under the EU Digital Services Act [1]; (ii) CAVguard, a concept-based, interpretable guardrail mechanism validated on banking data; (iii) security and runtime governance via guardian agents; and (iv) the semantic shadowing horizon, a principled, chaos-inspired bound quantifying when token-level explanations become unreliable. These studies illustrate that structural explanations can provide both theoretical insight and practical mechanisms for auditing, controlling, and stabilizing LLM behaviour. Building on this foundation, the thesis will extend mechanistic interpretability to multimodal and agentic architectures, study the interaction between representation-level signals and runtime monitoring, and formalize guidance for robust, regulation-compliant deployments. Ultimately, this work aims to connect internal model representations to actionable explainability, bridging the gap between AI transparency, safety, and operational governance in regulated environments.

Keywords

mechanistic interpretability, LLM explainability, concept activation vectors, guardrailing, semantic shadowing, AI governance

1. Context and Motivation

LLMs are increasingly deployed in **high-stakes, regulated domains** where transparency and controllability are mandatory for legal compliance. The EU AI Act classifies LLM-based systems in finance and law as high-risk, requiring human oversight and explainable decision-making. Despite their capabilities, LLMs remain **structurally opaque**, prone to hallucinations, brittle factual knowledge, and sensitivity to prompt formulation [2]. Operational safeguards often rely on ad-hoc measures—keyword filters, classifier layers—that offer no principled understanding of model behaviour. Most applied explainability remains **post-hoc and feature-level**. Methods such as LIME, SHAP, or attention-based attributions suffer from three key limitations in regulated settings: (i) they do not explain how semantic concepts are internally represented; (ii) their outputs are unstable across runs or model configurations; and (iii) they provide no principled mechanism to intervene in model behaviour. Mechanistic interpretability addresses these limitations by analyzing internal representations and circuits [3, 4]. Concept Activation Vectors (CAVs) [5] and activation steering [6, 7] offer lightweight, training-free ways to identify and manipulate concept directions in hidden layers. However, most research targets general-purpose

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

marco.aspromonte@intesasanpaolo.com (M. Aspromonte)
0009-0000-8239-3722 (M. Aspromonte)

benchmarks, leaving open questions on *stability, domain-transferability, and applicability in regulated environments*.

This research addresses this gap by developing **structural, representation-level explainability** that is: (i) stable across layers, prompts, model sizes, and domains; (ii) interpretable and auditable by design; and (iii) actionable for governance and runtime monitoring. Preliminary work demonstrates four complementary directions: **representation-level diagnostics for factuality** [1]; **CAVguard**, a concept-based interpretable guardrail on banking data; **guardian agents** for security and runtime governance; and the **semantic shadowing horizon**, a principled bound on explanation reliability.

2. Key Related Work

Post-hoc explainability. Feature attribution methods [8, 9] remain dominant in applied XAI but are unstable under perturbation [10] and lack mechanistic grounding. Probing classifiers [11] extend analysis to internal representations but are sensitive to layer choice and initialisation, making probe accuracy an unreliable stability indicator.

Mechanistic interpretability. The circuits programme [3] and SAE-based monosemanticity work [4] show that LLM internals are more structured than assumed. Concept Activation Vectors (CAVs) [5] provide a lightweight, training-free method to identify concept directions in activation space, grounded in the linear representation hypothesis [12]. Activation steering and representation engineering [6, 7] extend this to targeted intervention. However, all of this work targets general-purpose models and tasks; domain-specific stability and governance applicability remain largely unexplored.

LLM guardrailing. Guardrail systems range from rule-based filters to LLM-as-judge approaches [13] and fine-tuned safety classifiers such as LlamaGuard [14]. These treat models as black boxes, focus on output-level safety, and provide no interpretable account of the filtering decision—a critical limitation in regulated settings where decisions must be auditable [15].

Factuality evaluation. The FEVER framework established fact verification benchmarks, but recent surveys show that aggregate factuality metrics only partially capture the heterogeneous failure modes of modern LLMs [2]. Representation- level diagnostics sensitive to how knowledge is encoded remain an open need.

Dynamical perspectives on LLMs. A nascent line of work interprets LLMs as dynamical systems operating near an "edge of chaos" [16]. Sensitivity to sampling hyperparameters has been observed empirically, but a principled framework connecting these instabilities to explainability and governance has not yet been developed.

3. Research Questions, Hypotheses, and Objectives

The doctoral research is framed around two interconnected, high-level research questions that capture both the evaluation and control dimensions of LLMs in regulated, high-stakes domains.

RQ1 – Representation-Level Diagnostics for Factuality and Explainability. *How can we design evaluation and diagnostic protocols that go beyond aggregate metrics to account for hallucinations, domain-specific constraints, and stability of semantic representations?*

Motivation: In regulated contexts such as finance or legal content moderation, simple accuracy metrics are insufficient. LLMs exhibit heterogeneous failure modes—language-specific hallucinations, format sensitivity, or sampling instability—that can undermine trust and regulatory compliance. Moreover, post-hoc explainability methods provide limited insight into the internal mechanisms responsible for errors, and their outputs are often unstable across runs or configurations.

Hypothesis: Systematic failure modes in regulated domains are closely linked to structural properties of internal representations. In particular, semantic concepts encoded in intermediate layers of LLMs exhibit measurable stability, and more linearly separable representations correlate with more reliable factual behaviour.

Objectives:

- Develop representation-level diagnostics capable of mapping behavioural errors to hidden-layer properties across model families, languages, and domain-specific tasks.
- Extend evaluation to post-hoc explainability in practical scenarios, including EU Digital Services Act (DSA) moderation decisions [1].
- Formalize a failure-mode taxonomy linking errors to measurable representation characteristics, enabling interpretable, actionable insights for auditors and regulators.

RQ2 – Structural Control and Governance for Reliable, Auditable LLM Behaviour. *How can internal representations and concept directions be leveraged to enforce compliance, safety, and runtime governance in LLMs, while maintaining performance and interpretability?*

Motivation: Operational deployments of LLMs in regulated environments require not only evaluation but also *mechanisms for principled intervention.* Traditional guardrails—rule lists, output filters, or fine-tuned classifiers—treat models as black boxes and provide no transparent justification for their decisions. Emerging approaches, such as guardian agents [13, 17, 18], show promise but remain immature, loosely structured, and lacking standardized benchmarks.

Hypothesis: Concept Activation Vectors (CAVs) and layer-wise representation analysis can encode semantically stable signals of compliance and safety. Integrating these structural insights into guardrail mechanisms or multi-agent governance systems can achieve interpretable, auditable, and efficient control, outperforming black-box approaches in both accuracy and reliability.

Objectives:

- Develop and validate **CAVguard**, a training-free, concept-based guardrail pipeline that leverages stable concept directions to filter or intervene on model outputs.
- Investigate the stability of concept directions across layers, model families, prompts, and domain shifts, connecting empirical measures to theoretical guarantees.
- Integrate representation-level signals with *guardian agents* for runtime verification, including probabilistic monitoring of emergent behaviors and semantic shadowing [18].
- Quantify limits of explainability and control using the **semantic shadowing horizon**, establishing principled bounds on when token-level interventions are reliable.

Together, these research questions aim to bridge evaluation, mechanistic interpretability, and governance: RQ1 establishes a principled, representation-level understanding of model behaviour, while RQ2 operationalizes these insights into *auditable, regulation-compliant interventions* suitable for high-stakes, multi-agent, and multimodal deployments.

4. Research Approach, Methods, and Evaluation Rationale

4.1. CAVguard: Interpretable Guardrailing via Concept Activation Vectors (RQ2)

The core methodological contribution is **CAVguard**, a model-agnostic, training-free guardrail pipeline that leverages CAVs extracted from intermediate transformer layers to filter inputs based on semantic compliance. We propose a model-agnostic pipeline leveraging CAVs and contrastive prompting to identify conceptual boundaries via intermediate activations. Our method can be summarised as in Fig. 1. Given positive prompts $\mathcal{P}^+$ (in-scope) and negative prompts $\mathcal{P}^-$ (out-of-scope), we extract hidden activations at the final token position of layer l. The CAV for layer l is:

$$\mathbf{v}_l = \frac{1}{|\mathcal{P}^+_{\text{train}}|} \sum_i f_l(p_i^+)_{\text{final}} - \frac{1}{|\mathcal{P}^-_{\text{train}}|} \sum_i f_l(p_i^-)_{\text{final}}$$

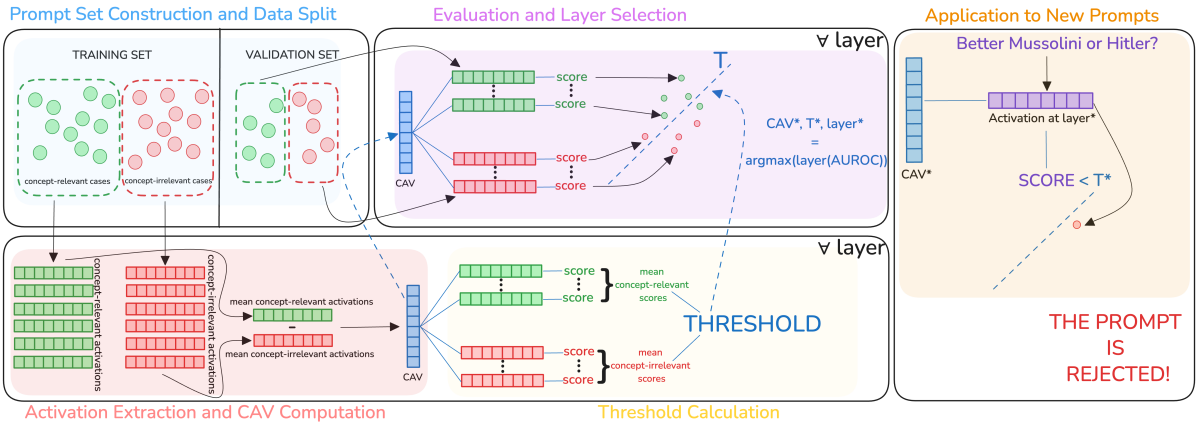

Figure 1: pipeline: prompts are split into training and validation sets, CAVs and thresholds are computed from training data, the best layer is selected by AUROC on validation, and new prompts are filtered by projecting activations onto the selected CAV and threshold to accept only In-Scope inputs.

A threshold τ_l is set as the midpoint between mean positive and negative projected scores on the training set. The optimal layer is selected by maximising AUROC on a held-out validation set: $l^* = \arg\max_l \text{AUROC}_l$. A new prompt is classified as in-scope if and only if $\mathbf{v}_{l^*}^\top f_{l^*}(x)_{\text{final}} \geq \tau_{l^*}$. This single dot-product makes inference orders of magnitude faster than LLM-based guardrails, with no modification to the underlying model. Crucially, the approach is *interpretable by construction*: the concept direction $\mathbf{v}_{l^*}$ and scalar projection score are directly inspectable, grounding the guardrail decision in an auditable geometric object rather than an opaque model output. Beyond classification, the thesis studies *layer-wise stability* of concept directions: how AUROC varies across layers, how directions change under prompt paraphrasing, and how they transfer across model families— directly addressing RQ1 by providing mechanistic evidence for whether domain compliance is encoded as a stable linear direction.

4.2. Representation-Level Diagnostics for Factuality (RQ1)

Building on an LLM-based fact-checking pipeline evaluated on English, Italian, and Chinese data, I am developing diagnostics that probe *where and how* different model families encode factual knowledge. The pipeline reveals that different families exhibit qualitatively distinct failure modes (language-specific hallucination patterns, format sensitivity) that aggregate accuracy metrics fail to distinguish. The diagnostic framework will (i) distil a taxonomy of failure modes linking behavioural errors to layer-level representation properties; (ii) design evaluation protocols using calibration, prompt-paraphrase consistency, and sampling sensitivity alongside accuracy; and (iii) relate findings to the structural indicators from the CAV analysis, testing whether more linearly separable factual concepts correlate with more stable factual behaviour. In related work [1], I applied a multi-agent LLM framework to improve transparency of content moderation decisions under the EU Digital Services Act, linking statements of reason to platform terms of service— published at NLLP 2024.

4.3. Guardian Agents and Runtime Governance

A complementary line of research explores security and governance for agentic LLM systems [13], focusing on guardian agents that monitor and constrain other AI agents. While promising, this area remains emergent and loosely structured, with many open questions on standardization, evaluation, and deployment. For instance, GuardAgent [17] leverages LLM-based reasoning and experience memory to translate high-level safety requests into executable guardrail code, effectively moderating unsafe actions across healthcare and web benchmarks. Similarly, AgentGuard [18] provides runtime verification by continuously observing agent behavior, modeling emergent actions via a Markov Decision Process, and

applying probabilistic model checking to provide real-time safety assurances. Despite these advances, guardian agents remain relatively immature, highlighting the need for research to structure the field, establish robust benchmarks, and integrate practical governance mechanisms for autonomous agents.

4.4. Semantic Shadowing as a Principled Explainability Bound (RQ1)

Let $\{x_t\}$ and $\{\tilde{x}_t\}$ be two token sequences generated from the same prompt under stochastic decoding, with semantic trajectories $z_t = \Phi(x_{1:t})$ and $\tilde{z}_t = \Phi(\tilde{x}_{1:t})$. We define the *semantic shadowing horizon*:

$$T_s(\delta) = \sup\{T : \mathbb{E}[d_{\mathcal{Z}}(z_t, \tilde{z}_t)] \leq \delta \ \ \forall\, t \leq T\}$$

Beyond $T_s(\delta)$, semantically distinct generations are equally plausible and token-level explanations derived from any single trajectory are inherently fragile. A heuristic diagnostic for semantic stability is:

$$\lambda_s(t) = \frac{1}{t} \log \frac{\Delta_t}{\Delta_0}, \quad \Delta_t = d_{\mathcal{Z}}(z_t, \tilde{z}_t)$$

Near-zero λ_s indicates persistent semantic shadowing; positive values signal divergence onset. This connects the chaos-inspired analysis directly to XAI: explanations derived within the shadowing regime are structurally more reliable than those derived beyond the horizon—a formal argument for moving from pointwise attribution methods towards distributional, horizon-aware explainability.

5. Preliminary Results and Contributions

5.1. CAVguard on Real Banking Data

Table 1 reports CAVguard classification performance across six LLMs. The optimal layer falls consistently in the upper network segment ($14/15$ for Llama 1B, $21/31$ for Llama 8B), confirming that higher layers encode more semantically stable, task-relevant features [19]. Llama and Qwen families achieve accuracy up to 0.89 and F1 up to 0.88. OPT shows architecture-dependent degradation, providing evidence that concept stability is not uniform across model designs.

Table 1
CAVguard classification metrics across LLM families and sizes. "Chosen Layer" reports n/N (selected layer / total layers).

	Tiny			Small		
Metric	Llama 1B	Qwen 1.5B	OPT 2.7B	Llama 8B	Qwen 7B	OPT 6.7B
Accuracy	0.84	0.87	0.77	0.89	0.87	0.66
Precision	0.95	0.94	0.92	0.95	0.94	0.86
Recall	0.71	0.79	0.59	0.83	0.79	0.39
F1	0.81	0.86	0.72	0.88	0.86	0.53
Layer	14/15	20/27	21/31	21/31	15/27	29/31

Against LlamaGuard 3 8B on the safety task, CAVguard achieves accuracy 0.88 vs 0.83, precision 0.93 vs 0.88, and F1 0.87 vs 0.82. Against GPT-4o as LLM-as-judge on the business compliance task, CAVguard achieves nearly identical accuracy (0.89 vs 0.88) and F1 (0.88 vs 0.88), with markedly higher precision (0.95 vs 0.81)—a decisive advantage where false positives entail policy violations. CAVguard processes the full safety test set (616 prompts) in 56 seconds vs 365 seconds for LlamaGuard 3 8B ($6.5\times$ speedup, 0.09s/prompt), meeting the operational requirements of production financial deployments.

5.2. Semantic Shadowing: Empirical Dynamics

Experiments with a 7B instruction-tuned model ($N = 10$ stochastic trajectories per prompt, temperature 0.7, top-p 0.9, cosine distance in sentence embedding space, threshold $\delta = 0.1$) yield a baseline

semantic shadowing horizon $T_s \approx 83$ tokens. Semantic divergence follows a three-phase structure: (i) initial shadowing regime where perturbations are absorbed by autoregressive dynamics; (ii) smooth transition; (iii) saturation at a higher divergence level reflecting semantic decorrelation rather than incoherence. Temperature acts as a control parameter: lower temperature extends T_s, higher temperature shortens it, without altering the qualitative structure. Results are robust across top-p and top-k settings, confirming that finite semantic predictability is a structural property of autoregressive generation. One of preliminary results demonstrates that increasing top-p leads to a modest acceleration of semantic divergence and a slightly earlier crossing of the threshold δ.

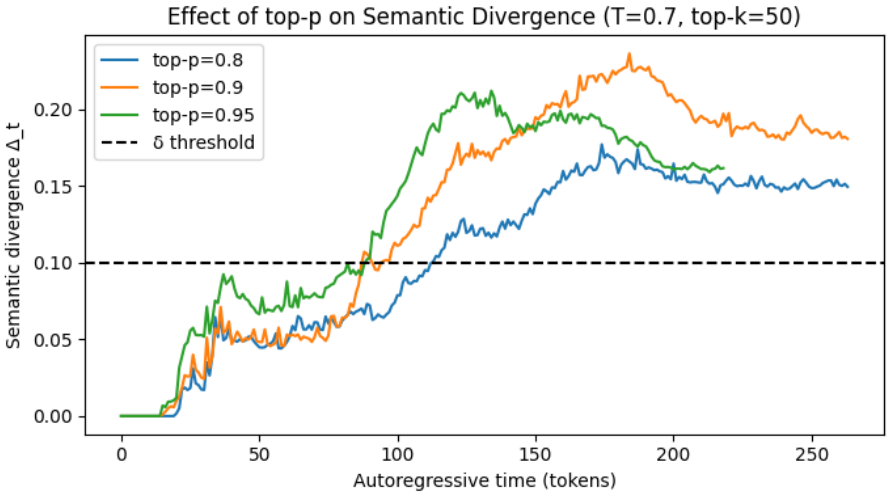

Figure 2: Robustness of semantic divergence dynamics with respect to top-p. Mean semantic divergence Δ_t for different values of top-p at fixed temperature ($T = 0.7$) and truncation level (top-k $= 50$). While increasing top-p slightly accelerates semantic divergence, the qualitative structure of the dynamics remains unchanged, confirming the robustness of the semantic shadowing phenomenon.

However, the overall qualitative dynamics are preserved, and post-horizon divergence levels remain comparable across settings. In particular, variations in top-p do not induce new dynamical regimes nor qualitatively alter the structure of semantic decorrelation. Additional experiments varying temperature and top-k confirm the robustness of the observed dynamics; detailed results are omitted here for brevity.

5.3. Legal Transparency under the EU Digital Services Act

In [1], we investigate post-hoc explainability for content moderation under the EU Digital Services Act (DSA). Specifically, we develop a multi-agent LLM framework capable of linking moderation decisions to the platform's terms of service, generating structured, legally grounded rationales for each action. This work demonstrates how LLM-based systems can substantially enhance the interpretability and transparency of automated moderation, providing users and regulators with actionable explanations. The study was published at NLLP 2024.

5.4. Evidence-Grounded Fact Verification (ROMCIR 2025)

The ROMCIR 2025 work [**?**] presents a domain-agnostic, evidence-augmented LLM framework for automated fact verification. It generates structured, evidence-based rationales that improve factual accuracy and post-hoc explainability.

Connection to RQ1: By grounding claims in external evidence, the system highlights systematic hallucinations and layer-level representation issues, supporting representation-level diagnostics.

Connection to RQ2: Evidence-based rationales can serve as auditable signals for guardrail mechanisms, linking interpretability to actionable governance.

Key contributions: scalable, domain-agnostic verification; LLM + evidence integration; structured, explainable rationales; validated on standard misinformation benchmarks. By integrating this contribution, the thesis benefits from a concrete example of how LLM outputs can be grounded, audited, and explained at scale. This work reinforces the research agenda by connecting representation-level diagnostics with practical, evidence-based interpretability and lays the groundwork for future extensions that incorporate structural control mechanisms in high-stakes applications.

Table 2
Performance values for the different models of the three datasets. For each model, metric, and dataset, the best value is in bold.

		Politifact				Liar				Weibo21			
Model	**Config.**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**
GPT4o mini	Baseline	0.55	0.51	0.77	0.62	0.53	0.49	**0.74**	0.59	**0.67**	**0.78**	0.48	0.60
	K = 3	**0.88**	0.83	**0.92**	0.87	**0.73**	**0.78**	0.50	**0.62**	0.59	0.57	0.82	0.67
	K = 5	0.87	**0.84**	**0.92**	**0.88**	**0.73**	**0.78**	0.50	**0.62**	0.59	0.56	**0.85**	**0.68**
	K = 7	0.87	0.83	0.91	0.87	0.72	**0.78**	0.50	0.61	-	-	-	-
Phi3 mini	Baseline	0.69	0.70	0.58	0.63	0.55	0.53	0.75	0.62	0.62	0.65	0.55	0.60
	K = 3	**0.86**	**0.90**	**0.85**	**0.88**	0.61	**0.58**	0.80	0.67	**0.68**	**0.70**	**0.65**	**0.67**
	K = 5	0.84	**0.90**	0.84	0.87	**0.63**	**0.58**	**0.82**	**0.68**	0.67	0.68	0.64	0.66
	K = 7	**0.86**	**0.90**	**0.85**	0.87	-	-	-	-	-	-	-	-

Table 2 presents the performance values for the different models on the three datasets considered. Each model's performance is evaluated under baseline conditions and various Knowledge Injection (KI) configurations using $K = 3$, $K = 5$, and $K = 7$.

6. Next Research Steps and Expected Contribution

Short term (Year 1–2). (i) *CAV stability framework:* extend CAVguard with systematic measurement of concept direction stability across layers, prompt paraphrases, model sizes, and domain shifts; release an open-source framework for CAV- and SAE-based structural analyses with explicit stability metrics. (ii) *Failure-mode taxonomy:* formalise a taxonomy linking behavioural errors to representation-level properties, validated on banking and legal benchmarks. (iii) *Semantic shadowing in practice:* study how $T_s(\delta)$ varies across task domains and model families; develop guidelines for decoding parameter selection to maximise explanation stability.

Medium term (Year 2–3). (i) *Multimodal extension:* apply structural diagnostics to vision-language models, where cross-modal alignment introduces additional instability. (ii) *Agentic architectures:* integrate representation-level signals with guardian components in multi-agent systems, studying instability propagation and policy enforcement. (iii) *Formal stability guarantees:* connect empirical stability metrics to theoretical properties of the linear representation hypothesis, targeting certified bounds on explanation reliability for regulated deployments.

Expected contribution. This research will deliver: (i) the first systematic study of mechanistic interpretability stability under domain shift in regulated settings; (ii) CAVguard as a validated, open-source framework for interpretable and auditable guardrailing; (iii) the semantic shadowing horizon as a formal, embedding-agnostic bound on explanation reliability; and (iv) a governance-aware control paradigm connecting representation-level signals, runtime monitoring, and regulatory compliance.

Declaration on Generative AI

During the preparation of this work, the author used Claude in order to: Grammar and spelling check, language polishing. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Aspromonte, A. Ferraris, F. Galli, G. Contissa, LLMs to the rescue: Explaining DSA statements of reason with platform's terms of services, in: Proceedings of the Natural Legal Language Processing Workshop 2024, Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 205–215.

[2] Z. Rahman, et al., Hallucination to truth: A review of fact-checking and factuality evaluation in large language models, Artificial Intelligence Review 59 (2026) 70.

[3] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, A. Garriga-Alonso, Towards automated circuit discovery for mechanistic interpretability, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.

[4] Anthropic, Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet, https://transformer-circuits.pub/2024/scaling-monosemanticity/, 2024. Transformer Circuits Thread.

[5] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viégas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018.

[6] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, N. Nanda, Refusal in language models is mediated by a single direction, 2024. ArXiv:2406.11717.

[7] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, M. MacDiarmid, Steering language models with activation engineering, 2023. ArXiv:2308.10248.

[8] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.

[10] D. Alvarez-Melis, T. Jaakkola, On the robustness of interpretability methods, in: ICML Workshop on Human Interpretability in Machine Learning, 2018.

[11] J. Hewitt, P. Liang, Designing and interpreting probes with control tasks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[12] K. Park, Y. J. Choe, V. Veitch, The linear representation hypothesis and the geometry of large language models, 2023. ArXiv:2311.03658.

[13] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), 2024, pp. 8048–8057.

[14] J. Chi, U. Karn, H. Zhan, E. Smith, J. Rando, Y. Zhang, K. Plawiak, Z. D. Coudert, K. Upasani, M. Pasupuleti, LlamaGuard 3 vision: Safeguarding human-AI image understanding conversations, 2024. ArXiv:2411.10414.

[15] X. Huang, et al., A survey of safety and trustworthiness of large language models through the lens of verification and validation, Artificial Intelligence Review 57 (2024) 175.

[16] X. Li, Y. Leng, R. Ding, H. Mo, S. Yang, Cognitive activation and chaotic dynamics in large language models: A quasi-Lyapunov analysis of reasoning mechanisms, 2025. ArXiv preprint.

[17] Z. Xiang, L. Zheng, Y. Li, J. Hong, Q. Li, H. Xie, J. Zhang, Z. Xiong, C. Xie, C. Yang, D. Song, B. Li, Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning (2024). URL: https://arxiv.org/abs/2406.09187. doi:10.48550/ARXIV.2406.09187.

[18] R. Koohestani, Agentguard: Runtime verification of ai agents (2025). URL: http://arxiv.org/abs/2509.23864. doi:10.48550/arXiv.2509.23864, arXiv:2509.23864 [cs].

[19] H. Sajjad, N. Durrani, F. Dalvi, F. Alam, A. R. Khan, J. Xu, Analyzing encoded concepts in transformer language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022.