Building a Robust Activation Monitor

Davide Baldelli
Polytechnique Montréal, Mila, MATS

An activation monitor is a classifier that takes a model's internal activations as input and detects a property of the model's state or of its output. In its most common form it is a linear probe trained on residual-stream activations [1] to detect properties such as deception [2], harmful intent [3], or hallucination [4]. Monitors are attractive for safety for two reasons. They do not depend on what the model chooses to show, since a model whose visible output is benign may still represent internally what it is about to do. They are also cheap, one dot product per token, so they can run on every token of every response at deployment scale, where an LLM judge reading the same transcripts would be prohibitive. Early work established that truthfulness is linearly decodable [4–6]. Recent probes separate strategic deception almost perfectly on their evaluation sets [2], and frontier labs have begun deploying probes in production [7].

In the last year, several groups independently found that high accuracy on a clean benchmark does not imply robustness. Much of what a probe captures is the surface text reflected in the activation vectors. When the textual evidence of the monitored behavior is removed, probe accuracy drops substantially, and on model organisms that never verbalize the behavior it falls close to chance [8]. Probes fail to transfer across datasets and attack families [9]. On organisms whose beliefs verifiably differ from their statements, every activation-based detector drops sharply [10]. Models can be fine-tuned to evade them, and the evasion transfers zero-shot to monitors the attacker never saw [11, 12]. The signal is distributed over several dimensions rather than a single direction [13]. Some fixes to specific weaknesses have been proposed, for example multi-dimensional probes and style-augmented training data [13], persona-coordinate projections [14], follow-up questions [10], conflict signatures [15], and keyed activation watermarking [16].

I believe that these weaknesses are consequences of the training recipe, and I propose to change the recipe. A monitor today is a linear classifier trained on a few thousand examples from one model, one concept, and one narrow distribution, and it monitors a model trained on trillions of tokens. Language models themselves became robust through scale and diversity of training data, and monitors were never trained this way. The one recipe-level intervention tried so far supports this reading, since augmenting the style of the probe's training data alone recovered near-perfect detection at every model scale [13].

The object of the project is a single activation monitor trained across concepts, across models, and on context-augmented data. The monitor is conditioned on the concept it is asked to detect, so one model covers the whole concept list and can be queried on concepts it never saw. To train it on the activations of different models, each model gets a projection layer that maps its activation space into the monitor's shared space, and the monitor is trained on top of the projections.

Recent literature supports each part of this design, and no published work combines them. Behavioral directions align across model families once projected into a shared coordinate space [17]. Affine maps transfer probes and steering vectors between models [18, 19], and steering-based safety interventions transfer across Llama, Qwen, and Gemma [20]. SAE feature spaces are similar across LLMs up to simple transformations [21], refusal interventions transfer across architectures through a shared concept basis [22], and in vision models a single sparse concept space has been trained over

several networks through per-model projections [23]. To my knowledge, no published work trains one monitor on many models' activations, and the closest paper describes cross-family alignment as an open problem [18].

We will start by collecting a dataset for each concept and augmenting all of them with longer and more realistic contexts, embedding the labeled examples into real-world chats. We will then explore the design axes: which layer to take activations from, whether one or several, the architecture and the training parameters, how to condition the monitor on the concept text, and how to build the projection from each model's activation space to the monitor's space. Finally, we will evaluate on held-out concepts and held-out models, where a new model is onboarded by training only its projection while the monitor stays frozen.

The ideal deliverable is a single activation monitor that is model agnostic, concept agnostic, and robust to context.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

[2] Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes, 2025. URL https://arxiv.org/abs/2502.03407.

[3] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

[4] Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827.

[6] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

[7] János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for gemini, 2026. URL https://arxiv.org/abs/2601.11516.

[8] Gerard Boxo, Aman Neelappa, and Shivam Raval. Linear probes rely on textual evidence: Results from leakage mitigation studies in language models, 2026. URL https://arxiv.org/abs/2509.21344.

[9] Cheng Wang, Zeming Wei, Qin Liu, and Muhao Chen. False sense of security: Why probing-based malicious input detection fails to generalize, 2025. URL https://arxiv.org/abs/2509.03888.

[10] Alan Cooney, David Africa, and Geoffrey Irving. "did you lie?" evaluating lie detectors across

model scale and belief-verified model organisms, 2026. URL https://arxiv.org/abs/2606.12618.

[11] Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass llm latent-space defenses, 2025. URL https://arxiv.org/abs/2412.09565.

[12] Max McGuinness, Alex Serrano, Luke Bailey, and Scott Emmons. Neural chameleons: Language models can learn to hide their thoughts from unseen activation monitors, 2025. URL https://arxiv.org/abs/2512.11949.

[13] Sachin Kumar. Pressure-testing deception probes in llms: Scaling, robustness, and the geometry of deceptive representations, 2026. URL https://arxiv.org/abs/2605.27958.

[14] Prasad Mahadik and Adrians Skapars. Do linear probes generalize better in persona coordinates?, 2026. URL https://arxiv.org/abs/2605.09391.

[15] Petr Nyoma. Rift: A conflict signature for deception in language models, 2026. URL https://arxiv.org/abs/2606.17229.

[16] Toluwani Aremu, Daniil Ognev, Samuele Poppi, and Nils Lukas. Robust safety monitoring of language models via activation watermarking, 2026. URL https://arxiv.org/abs/2603.23171.

[17] Su-Hyeon Kim and Yo-Sub Han. Cross-family universality of behavioral axes via anchor-projected representations, 2026. URL https://arxiv.org/abs/2605.09875.

[18] Femi Bello, Anubrata Das, Fanzhi Zeng, Fangcong Yin, and Liu Leqi. Linear representation transferability hypothesis: Leveraging small models to steer large models, 2025. URL https://arxiv.org/abs/2506.00653.

[19] Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching, 2025. URL https://arxiv.org/abs/2506.06609.

[20] Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. Activation space interventions can be transferred between large language models, 2025. URL https://arxiv.org/abs/2503.04429.

[21] Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders, 2025. URL https://arxiv.org/abs/2410.06981.

[22] Tony Cristofano. Universal refusal circuits across llms: Cross-model transfer via trajectory replay and concept-basis reconstruction, 2026. URL https://arxiv.org/abs/2601.16034.

[23] Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment, 2026. URL https://arxiv.org/abs/2502.03714.