Understanding representational geometry of safety-relevant features in language models

Motivation Large language models (LLM) internally encode concepts of space, time [Gurnee and Tegmark, 2023] factual associations [Meng et al., 2022], and features relevant to AI safety. Representation engineering aims to read and further manipulate these representations to augment attributes critical for safety, like morality, fairness, truthfulness, and honesty [Zou et al., 2023]. For instance, decoding a model's honesty involves creating two sets of prompts: one instructing honesty and the other instructing dishonesty. The internal activations for the prompts are recorded, from which an honesty axis is extracted that best separates these two sets. This axis can, therefore, function as a lie detector or an honesty controller (e.g., via a steering vector [Rimsky et al., 2023]). However, these methods often fail to generalize to out-of-distribution samples [Levinstein and Herrmann, 2023], casting doubt on the reliability and robustness of such lie detectors and posing potential risks for real-world applications. By examining the representational geometry of these LLMs, my research aims to deepen our understanding of their internal representations of safety-relevant features and corresponding context-sensitive behaviors. This endeavor will pave the way for better lie detectors, more robust behavior controllers, and more trustworthy and transparent advanced AI systems.

Questions How are safety-relevant features like honesty geometrically represented in LLMs? Do they acquire an abstract format that enables reliable generalizations?

Methodology Here I aim to investigate the representational geometry of hidden layer representations – more specifically, the abstract format of honesty that underpins the generalization capabilities of lie detectors. Operationally, the representation of a variable x is defined as abstract, if a decoder trained to predict x under specific conditions (e.g., some values of context variable y) can correctly generalize its prediction to novel, unseen conditions (e.g., previously unexperienced values of context variable y). This capability, termed cross-condition generalization performance (CCGP) [Bernardi et al., 2020], distinguishes itself by emphasizing the abstraction and disentanglement of neural representations across different contexts, as opposed to traditional generalization metrics where the decoder is trained on samples from all conditions. The abstractness of representations is critical to the robustness of detectors; those detectors based on non-abstract variables may perform well in trained conditions but easily fail in novel, unseen scenarios [Johnston and Fusi, 2023].

Preliminary analysis I have examined the honesty representations in Mistral-7B-Instruct-v0.1. I designed four sets of prompts that instructed the model to behave honestly or dishonestly (akin to [Zou et al., 2023]) in a medical environment (where honesty is crucial for patient health) or a business environment (characterized by a significant incentive to exaggerate). I trained honesty detectors on residual stream activations in one environment and tested their performance in the same environment or the other environment. Ideally, if the honesty variable is entirely abstract, then the other-environment generalization (CCGP) should exhibit a similar performance comparable to the same-environment generalization (traditional metric), suggesting a high reliability of the detectors. However, I found that the other-environment generalization consistently underperforms relative to the same-environment generalization across all layers (Fig. 1, left), hinting at an imperfect abstraction of honesty. To corroborate this intuition, I applied multidimensional scaling to the averaged patterns of the last-layer representations in the four conditions (two traits $\times$ two environments). The four conditions occupy the vertices of a tetrahedron (Fig. 1, middle), indicating that the honesty axis in one environment is only partially aligned with that in the other. In the ideal scenario, the axes in both environments are perfectly parallel (Fig. 1, right), thus the detectors based on one environment can be directly generalized to unseen, novel environments.

Milestone 1: Completion of preliminary analysis Objective: To identify key factors (e.g., contexts in my preliminary analysis) that interfere with the geometrical representations of honesty and other safety-relevant features. This involves a comprehensive literature review and experimental investigations using Mistral-7B. Expected Output: The identification of factors that can best reflect the out-of-distribution challenge faced by associated lie detectors and a detailed analysis of the geometric structure of their internal representations.

Milestone 2: Investigation of model sizes and training stages Objective: To examine how the abstract formats of honesty and other safety-related features evolve across different model sizes and throughout various stages of training, specifically targeting the Pythia series [Biderman et al., 2023]. Expected Output: While the abstract formats of these features are expected to emerge gradually in larger models during training, challenges persist in achieving ideal representations.

Milestone 3: Development of training techniques Objective: To design training techniques aiming at enhancing the abstractness of safety-related variables (with a focus on models trained on TinyStories dataset [Eldan and Li, 2023]): This will involve (1) synthesizing artificial data to serve as multitask learning objectives for safety-related concepts [Johnston and Fusi, 2023]; (2) developing auxiliary loss functions that increase the relative distances of hidden layer representations across different values of a targeted variable (i.e., honesty). Expected Output: The development and validation of training techniques that lead to an ideal abstraction of safety-relevant features within models' representations.

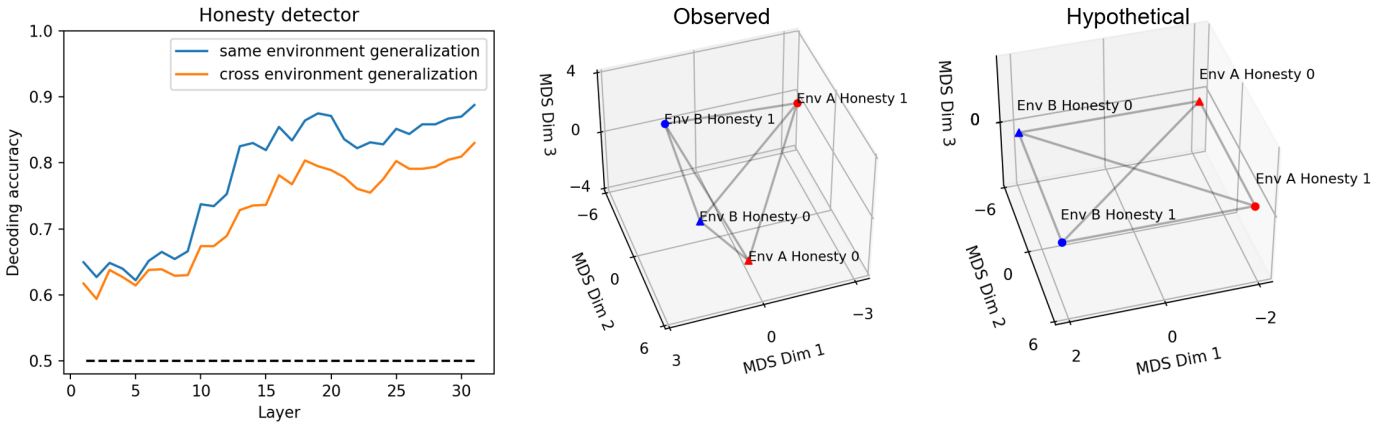

Fig 1: (Left) The honesty detector is trained on the hidden-layer representations in Mistral-7B-Instruct-v0.1 in one environment (medical or business) and tested in the same or the other environment. (Middle) Multidimensional scaling plot of the averaged last-layer representations in four conditions (two environments $\times$ two honesty traits). (Right) Multidimensional scaling plot of an idealized hidden layer representation in which the geometric alignment of honesty facilitates perfect generalization across environments.

References

[Bernardi et al., 2020] Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.

[Biderman et al., 2023] Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

[Eldan and Li, 2023] Eldan, R. and Li, Y. (2023). Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

[Gurnee and Tegmark, 2023] Gurnee, W. and Tegmark, M. (2023). Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

[Johnston and Fusi, 2023] Johnston, W. J. and Fusi, S. (2023). Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nature Communications*, 14(1):1040.

[Levinstein and Herrmann, 2023] Levinstein, B. and Herrmann, D. A. (2023). Still no lie detector for language models: Probing empirical and conceptual roadblocks. *arXiv preprint arXiv:2307.00175*.

[Meng et al., 2022] Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

[Rimsky et al., 2023] Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. (2023). Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

[Zou et al., 2023] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.