Analyze the Neurons, not the Embeddings: Understanding When and Where LLM Representations Align with Humans

Masha Fedzechkina [*] **Eleonora Gualdoni** [*] **Sinead Williamson** [*] **Katherine Metcalf**
Skyler Seto **Barry-John Theobald**
Apple
{mfedzechkina, e_gualdoni, sa_williamson, kmetcalf, bjtheobald, sseto}@apple.com

Abstract

Modern large language models (LLMs) achieve impressive performance on some tasks, while exhibiting distinctly non-human-like behaviors on others. This raises the question of how well the LLM's learned representations align with human representations. In this work, we introduce a novel approach to study representation alignment: we adopt an activation steering method to identify neurons responsible for specific concepts (e.g., "cat") and then analyze the corresponding activation patterns. We find that LLM representations captured this way closely align with human representations inferred from behavioral data, matching inter-human alignment levels. Our approach significantly outperforms the alignment captured by word/sentence embeddings, which have been the focus of prior work on human-LLM alignment. Additionally, our approach enables a more granular view of how LLMs represent concepts — we show that LLMs organize concepts in a way that mirrors human concept organization.

1 Introduction

Large language models (LLMs) exhibit impressive performance on a variety of tasks from text summarization [2, 10] to zero-shot common-sense reasoning [17, 22], and are increasingly deployed as a human proxy [11, 12, 6, 18]. At the same time, there is a growing body of evidence suggesting that LLMs exhibit patterns of behavior distinctly different from humans, such as hallucinating information [5, 13] or memorizing complex patterns to solve reasoning tasks [26]. Such behaviors raise the question of how closely the conceptual representations learned by these models align with human conceptual representations, as safe and trustworthy deployment of LLMs may require such alignment.

Here, we propose a novel way to study human–LLM alignment in concept representation. We borrow a method from activation steering [23, 24, 20], to identify which neurons are most responsible for processing and understanding a particular concept, so-called *expert neurons*. We show that the neurons discovered with this method provide information about how models represent concepts and capture the dimensions meaningful to humans, providing a reliable method to test model alignment. Additionally, we analyze how human-model alignment evolves with model training and depends on model capacity: it emerges early in training, with model size playing only a small role.

[*] Core author.

2 Methods

2.1 Finding expert neurons

We adopt the *finding experts* approach introduced by Suau et al. [23] for activation steering, to study representational alignment. In this approach, a concept c is defined through a set of example sentences $N = N_c^+ + N_c^-$, where N_c^+ is a set of sentences that contain c (henceforth *positive set*) and N_c^- is a set of sentences that do not contain c (*negative set*). Next, we obtain the activations $z_m^c = \left\{z_{m,i}^c\right\}_{i=1}^N$ for every neuron m in the model in response to the inputs from both sets of sentences. z_m^c is then treated as a prediction score for the presence of c. The performance of each neuron as a classifier for the concept (i.e., its *expertise*) is measured as the area under the precision-recall curve (AP) on this task – an expert is a neuron with AP score of above 0.5. We calculate the AP score for all units in the MLP and attention layers.

2.2 Data and Models

To understand the alignment between human and model representations, we examine how patterns in expert neurons relate to perceived concept similarity in humans obtained from the MEN dataset [4], which contains word pairs annotated with human-assigned similarity judgments. For each word in a pair, we generate a set of sentences containing that word. We use three models of different performance levels: GPT-4 [16], Mistral-7b-Instruct-v0.2 [9], and an internal 80b-chat model. The negative sets are sampled from the datasets for the remaining non-target concepts (e.g., if we are considering 1000 concepts, one of which is "cat", the negative set is sampled from 999 concepts excluding "cat").

We use GPT-2 [19] to select hyper-parameters (e.g., the size of the positive and negative datasets) and the Pythia family [3] – 70m (smallest), 1b, and 12b (largest) for the main experiment to understand the impact of model size on representational alignment. For each Pythia model, we work with checkpoints 1, 512, 1k, 4k, 36k, 72k, and 143k, to track how representational alignment develops throughout training.

3 Experts capture stable representations across different dataset characteristics

While the success of expert-based methods at steering model activations is well-documented – activating the experts for a particular concept steers the model to generate text consistent with a particular concept [23, 24], our interest is in studying model representations through the patterns in experts. Given the novel application of the method, we first explore the impact of dataset size, the model used to generate the dataset, and the exact sentences used to represent a concept on the stability of the discovered expert sets. We sample 50 word pairs from the training split of the MEN dataset. For each word in a pair, we generate a positive set containing 7000 sentences from the three models above. We sweep over multiple positive and negative set sizes (see Fig. 1) and for each positive and negative set combination, we repeat expert extraction eight times (folds) with the sets randomly sampled from the full pool of sentences. We examine how sensitive the discovered experts are to the specific slice of the positive and negative sets (the eight folds). We look at Jaccard similarity between expert sets across folds, using a range of AP thresholds τ, which can be thought of as the quality of an expert neuron — the larger the value of τ, the more expert a neuron is for a given concept.

The expert neurons discovered across different data configurations and folds (indicated by the error bars) are stable, as indicated by a high (~ 0.8) overlap proportion, and show little sensitivity to our manipulations. Interestingly, the LLM (line color) used to generate the probing dataset matters little — while stronger models generate more diverse datasets (mean type/token ratio of 0.34, 0.21 and 0.18 for GPT-4, internal 80b-chat, and Mistral-7b-Instruct-v0.2 respectively), resulting in a somewhat higher expert overlap, the gain is too small to warrant their increased cost. Expert overlap increases with every increase in the size of the positive set, but the increases are small beyond 300 sentences, and performance for 400 sentences is virtually indistinguishable from 500 sentences. Interestingly, a larger negative set results in lower expert overlap at higher τ values and an increased variability across folds. Based on these findings, we conduct all subsequent analyses with a positive set of

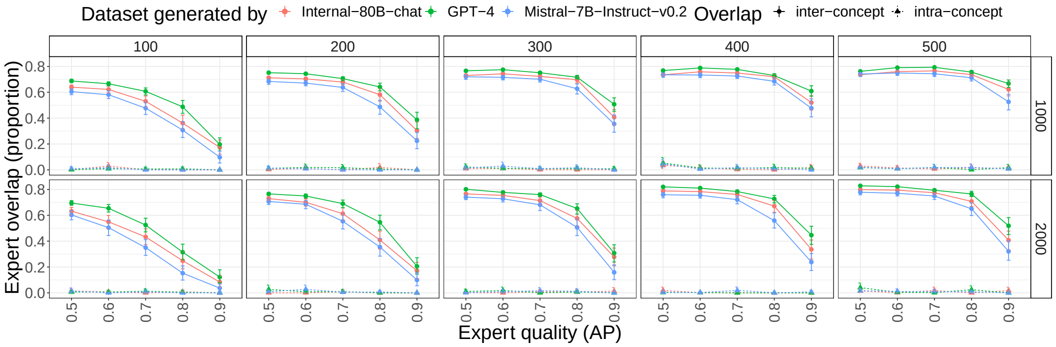

Figure 1: Expert discovery is relatively stable across various dataset characteristics. Points represent condition means; error bars represent bootstrapped 95% confidence intervals. Columns and rows represent the size (number of unique sentences) of the positive and negative sets respectively. Inter-concept is within-concept expert overlap; intra-concept is expert overlap averaged across randomly sampled pairs of concepts.

400 sentences and a negative set of 1000 sentences, all generated with Mistral-7b-Instruct-v0.2 and consider the expertise threshold τ=0.5.

4 Expert-captured representations are more aligned with humans than the embeddings

We measure the alignment between LLM and human representations as the correlation between the human versus the LLM's similarity score for each pair of concepts in the test split of the MEN data (1000 pairs). The experts-based similarity score is the Jaccard similarity between expert sets. We compare this representation to two types of embeddings: single-word embeddings from the embedding layer in line with prior work on LLM-human concept alignment [7] and the average of the sentence embeddings from the positive set for a given word from the final hidden layer. Here, we compute cosine similarity between the embeddings for each word pair in the MEN test split.

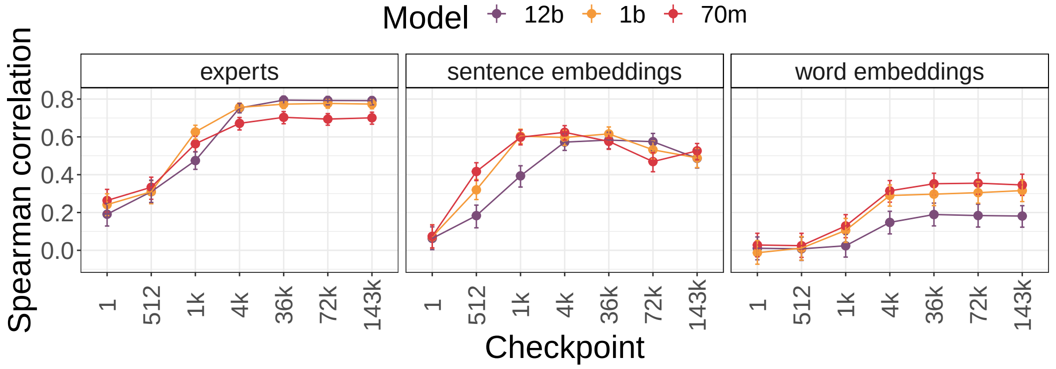

Figure 2: Expert-based representations of similarity are more closely aligned with human ones than the embeddings. Points are Spearman correlations between LLM similarity and human similarity in the MEN dataset (significant after Checkpoint 1, p<0.05); error bars are bootstrapped 95% confidence intervals. The subplots are similarity type.

Expert-based representations are closely aligned with humans (Fig. 2). At the final checkpoint, the Spearman correlations between expert overlap and MEN similarity are 0.70, 0.77, 0.79 for 70m, 1b, and 12b respectively. For reference, agreement between humans has a correlation of 0.84. Interestingly, model size has a small impact on this alignment in line with findings in vision from [15]: the 1b and 12b models are virtually indistinguishable, with the 70m model being slightly less aligned. The correlations with human similarity are significantly lower for both single word and sentence embeddings compared to the experts (p-values<0.05).

5 Expert-captured representations mirror human conceptual structure

We now ask whether the experts capture a broader human-interpretable representation of concepts that goes beyond pairwise (dis)similarity. Specifically, we ask if the concepts are clustered in the expert space in a way that aligns with human-interpretable knowledge structures. Humans organize concepts into domains [8, 14, 21]. For example, "dog", "cat" and "horse" are all *animals* and "bike", "bus", and "car" are all *vehicles*. This raises the question of whether LLMs organize concepts in a similar way. For the case study, we manually generate lists of ten domains with four concepts per domain (e.g., the domain *animal* containing concepts "cat", "dog", "cheetah", and "horse").

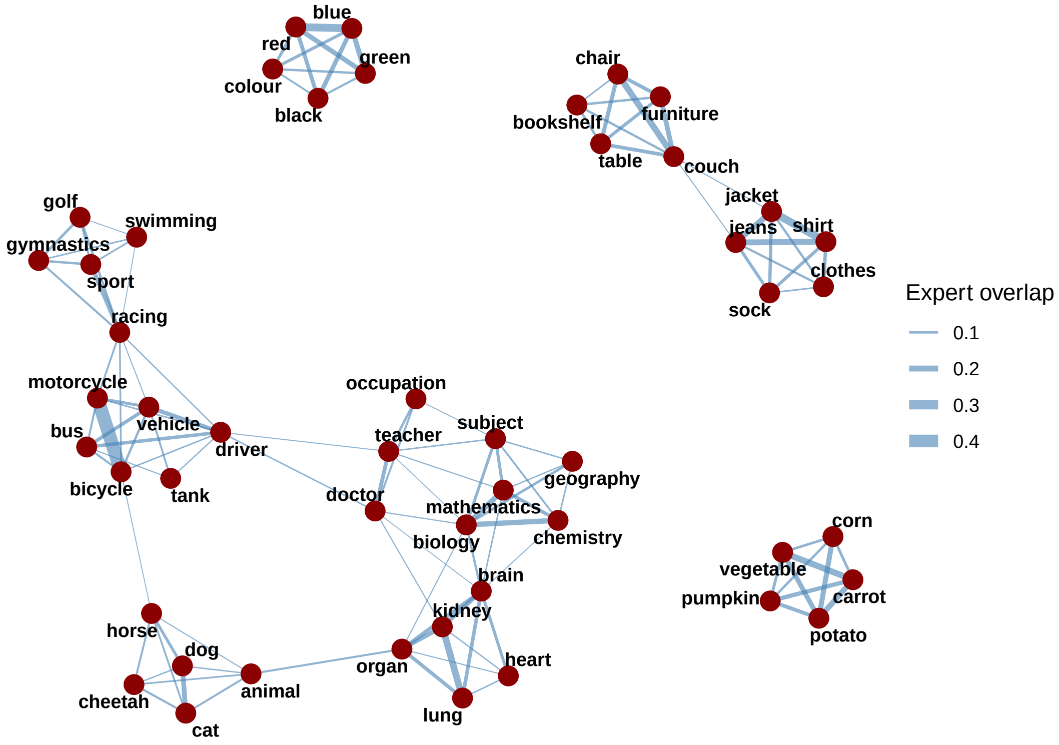

Figure 3: Concept domain reconstruction based on expert overlap in Pythia 12b. Each node represents a concept; edge thickness corresponds to the degree of reciprocal expert overlap between concepts.

Fig. 3 provides a visualization of the concept structure in the expert space, revealing a clear domain organization: concepts belonging to the same domain are strongly associated (e.g., all color terms are connected to each other, but not to other domains), while cross-domain associations are notably sparser. Expert sets, however, also uncover meaningful between-domain connections. For instance, while "driver" is an *occupation*, its expert set is also strongly associated with "bus" or "vehicle". Similarly, "racing" connects the *sports* domain with the *vehicles* domain.

6 Conclusion

We present a novel approach to study alignment between human and model representations based on the patterns in expert neurons. Representations captured by these neurons align with human representations significantly more than word/sentence embeddings, and approach human alignment levels, adding to the growing body of evidence that different learning systems –and, in our case, humans and LLMs– can converge on similar representations [1, 25]. Consistent with prior work in vision, [15], we find that model size has little influence on alignment.

References

[1] Khai Loong Aw et al. "Instruction-tuned LLMs with World Knowledge are More Aligned to the Human Brain". In: *UniReps: the First Workshop on Unifying Representations in Neural Models*. 2023. URL: https://openreview.net/forum?id=qqdHkqHmfA.

[2] Lochan Basyal and Mihir Sanghvi. *Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models*. 2023. arXiv: 2310.10449 [cs.CL]. URL: https://arxiv.org/abs/2310.10449.

[3] Stella Biderman et al. "Pythia: A suite for analyzing large language models across training and scaling". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 2397–2430. URL: https://arxiv.org/pdf/2304.01373.

[4] Elia Bruni, Nam Khanh Tran, and Marco Baroni. "Multimodal Distributional Semantics". In: *J. Artif. Intell. Res.* 49 (2014), pp. 1–47. URL: https://api.semanticscholar.org/CorpusID:2618475.

[5] Sébastien Bubeck et al. *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. 2023. arXiv: 2303.12712 [cs.CL]. URL: https://arxiv.org/abs/2303.12712.

[6] Ganqu Cui et al. "ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback". In: *Forty-first International Conference on Machine Learning*. 2024. URL: https://arxiv.org/pdf/2310.01377.

[7] Jan Digutsch and Michal Kosinski. "Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans". In: *Scientific Reports* 13.1 (2023), p. 5035. URL: https://www.nature.com/articles/s41598-023-32248-6.

[8] Caroline Graf et al. "Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions". In: *Cognitive Science* (2016). URL: https://api.semanticscholar.org/CorpusID:9066747.

[9] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[10] Hanlei Jin et al. *A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods*. 2024. arXiv: 2403.02901 [cs.AI]. URL: https://arxiv.org/abs/2403.02901.

[11] Hoang Anh Just et al. "Data-Centric Human Preference Optimization with Rationales". In: *arXiv preprint arXiv:2407.14477* (2024).

[12] Martin Klissarov et al. "Motif: Intrinsic motivation from artificial intelligence feedback". In: *arXiv preprint arXiv:2310.00166* (2023).

[13] Stephanie Lin, Jacob Hilton, and Owain Evans. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. 2022. arXiv: 2109.07958 [cs.CL]. URL: https://arxiv.org/abs/2109.07958.

[14] Gregory Murphy. *The Big Book of Concepts*. MIT Press, 2004.

[15] Lukas Muttenthaler et al. *Human alignment of neural network representations*. 2023. arXiv: 2211.01201 [cs.CV]. URL: https://arxiv.org/abs/2211.01201.

[16] OpenAI, Josh Achiam, Steven Adler, et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[17] Hyuntae Park et al. *Zero-shot Commonsense Reasoning over Machine Imagination*. 2024. arXiv: 2410.09329 [cs.AI]. URL: https://arxiv.org/abs/2410.09329.

[18] Andi Peng et al. "Learning with Language-Guided State Abstractions". In: *ICLR* (2024). arXiv: 2402.18759 [cs.RO]. URL: https://arxiv.org/abs/2402.18759.

[19] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[20] Pau Rodriguez et al. "Controlling Language and Diffusion Models by Transporting Activations". In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: https://openreview.net/forum?id=l2zFn6TIQi.

[21] Eleanor Rosch. "Principles of Categorization". In: *Cognition and Categorization*. Ed. by Eleanor Rosch and B. B. Lloyd. Hillsdale, NJ: Erlbaum, 1978, pp. 27–48.

[22] Vered Shwartz et al. "Unsupervised Commonsense Question Answering with Self-Talk". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 4615–4629. DOI: 10.18653/v1/2020.emnlp-main.373. URL: https://aclanthology.org/2020.emnlp-main.373/.

[23] Xavier Suau, Luca Zappella, and Nicholas Apostoloff. *Self-conditioning pre-trained language models*. 2023. arXiv: 2110.02802 [cs.CL]. URL: https://arxiv.org/abs/2110.02802.

[24] Xavier Suau et al. *Whispering Experts: Neural Interventions for Toxicity Mitigation in Language Models*. 2024. arXiv: 2407.12824 [cs.CL]. URL: https://arxiv.org/abs/2407.12824.

[25] Tahereh Toosi. “Representational constraints underlying similarity between task-optimized neural systems”. In: *UniReps: the First Workshop on Unifying Representations in Neural Models*. 2023. URL: https://openreview.net/forum?id=tOW3IWHw8G.

[26] Tomer Ullman. *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks*. 2023. arXiv: 2302.08399 [cs.AI]. URL: https://arxiv.org/abs/2302.08399.