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

Anonymous ACL submission

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 a method from research on activation steering 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 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 (Basyal and Sanghvi, 2023; Jin et al., 2024) to zero-shot common-sense reasoning (Park et al., 2024; Shwartz et al., 2020), and are increasingly deployed as a human proxy (Just et al., 2024; Klissarov et al., 2023; Cui et al., 2024; Peng et al., 2024). 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 (Bubeck et al., 2023; Lin et al., 2022) or memorizing complex patterns to solve reasoning tasks (Ullman, 2023). 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 requires such alignment. Unveiling aspects of representation alignment and understanding how to foster it can help us identify and mitigate misaligned LLM behaviors, increasing model trust and safety (OpenAI et al., 2024; Shen et al., 2024).

Prior work has examined the relationship between human-perceived similarity among concepts (i.e., word/image meaning) and various model-based measures of similarity, such as confidence (Shaki et al., 2023) or the embedding distance (Bruni et al., 2012; Digutsch and Kosinski, 2023; Muttenthaler et al., 2023). While there is a large body of work on alignment in traditional NLP models (Auguste et al., 2017; Ettinger and Linzen, 2016; Ruan et al., 2016; Søgaard, 2016) such as GloVE (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013), alignment in LLMs has proved hard to capture. Prompting approaches provide inconsistent results (Shaki et al., 2023); context can have unintuitive influences on alignment (Misra et al., 2020) and single-word embeddings do not fully capture LLM representations. Additionally, approaches relying on embedding similarity or model responses suffer from a major limitation: they do not reveal where in the model the concepts are stored and make it difficult to draw conclusions beyond coarse alignment. For example, the cosine distance between embeddings might indicate that "animal" and "dog" are more similar than "animal" and "daffodil", but it can not tell us if "dog" and "animal" share the same neurons, limiting our ability to understand the model's concept organization.

Here, we propose a novel way to study human – LLM alignment in concept representation. We borrow a method from activation steering (Suau et al., 2023, 2024; Rodriguez et al., 2025), to identify which neurons are most responsible for processing and understanding a particular concept, so-called *expert neurons*. This approach was originally introduced with the goal of directing LLM outputs towards a desired direction (e.g., reducing toxicity).

Specifically, it has been shown that expert neurons play a causal role in the generation of outputs semantically related to the concept they encode. For instance, activating the expert neurons for the concept "dog" steers the model to generate text consistent with it (Suau et al., 2023); conversely, suppressing the experts for toxicity generates less toxic text (Suau et al., 2024).

We investigate a novel application of this approach as a technique to achieve model interpretability. Specifically, 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. In addition, this approach enables us not only to measure alignment between human and model representations, but also to explore additional questions, such as whether LLMs organize concepts in a way that mirrors human conceptual organization (e.g., if "dog", "cat", "cheetah" , and "animal" share a consistent set of neurons). We also track how alignment evolves during training for different model sizes, shedding light on the impact of model capacity on the development of aligned representations — an aspect largely overlooked in previous work on text-based models (Shen et al., 2024; Wei et al., 2022). Ultimately, understanding the factors that lead to mis-alignment can provide valuable insight for designing interventions targeted at guiding model behaviors towards human-like solutions and enhancing their transparency (Fel et al., 2022; Peterson et al., 2018; Toneva, 2022).

In our experiments, we focus on causal LLMs using the Pythia models (70m, 1b and 12b) for which multiple training checkpoints are publicly available (Biderman et al., 2023). Given a diverse set of concepts across multiple domains (see Sec. 3.2), we identify each LLM's corresponding expert neurons. We measure their similarity at the LLM level as the amount of overlap between the expert neurons. We measure alignment in two tasks. First, we look at whether expert overlap is predictive of human-perceived concept similarity. Second, we ask if the alignment goes beyond simple pairwise similarities — specifically, whether LLMs organize concepts in a way that mirrors human conceptual structures (Rosch, 1978). Finally, we study how the model's concept representations develop through training.

Our results show that human-LLM representation alignment matches inter-human alignment, which is not detectable with prior approaches relying on embedding distance. Our contributions are:

1. We show that a method used to identify expert neurons reliably captures concept representations in LLMs and is stable across models and datasets.

2. We show that the representations captured with this approach align closely with human representations matching inter-human alignment, both at the level of concept similarity and in terms of concept organization.

3. We provide an analysis of how human-model alignment evolves with model training and depends on model capacity. Such alignment emerges early in training, with model size playing only a small role.

2 Related work

Representation alignment Studies on the kinds of representations used by humans and machines have been of interest to many fields (e.g., cognitive science, neuroscience, and machine learning; Hebart et al., 2020; Khosla and Wehbe, 2022; Muttenthaler et al., 2023; Tian et al., 2022; Søgaard, 2016). Studies on *representation alignment* (Sucholutsky et al., 2024) look specifically at the extent to which the internal representations of humans and neural networks converge on a similar structure. Across vision and text domains, models show notable alignment with human similarity judgments — typically used as a window into human representational structures. Peterson et al. (2018) report significant alignment between human similarity judgments and representations of object classification networks. In the vision domain, Muttenthaler et al. (2023) find that while the training dataset and objective function impact alignment, model scale and architecture have no significant effect. To our knowledge, there are no investigations of the influences of model size on alignment specifically in LLMs. Additionally, reliably capturing human-LLM alignment has proven hard. Shaki et al. (2023) prompt GPT-3 (Brown et al., 2020) and use response confidence as a measure of LLM representation, which they relate to human behavioral measures. However, they find the approach highly unreliable — small variations in the prompt lead to large changes in alignment. Misra et al. (2020) show that BERT (Devlin et al., 2019) can get distracted by context, assigning lower probability to

related concepts. Digutsch and Kosinski (2023) operationalize LLM concept similarity as the cosine similarity between single word representations extracted from the embeddings matrix. While they find that this measure predicts human-perceived similarity of the concept pair, it is unclear how this method can be scaled to study human-LLM alignment between more complex concepts like toxicity that cannot be represented as single words. Thus, despite the increasing interest in human-LLM alignment, reliable methods to study alignment are still lacking.

Activation steering refers to a class of methods that intervene on a generative model's activations to perform targeted updates for controllable generation (Rodriguez et al., 2025; Li et al., 2024; Rimsky et al., 2024). Suau et al. (2023) propose a method to identify sets of neurons in pre-trained transformer models that are responsible for detecting inputs in a specific style (Suau et al., 2024, e.g., toxic language) or about a specific concept (Suau et al., 2023, e.g., "dog"). Intervening on the expert neuron activations successfully guides text generation into the desired direction. In a similar spirit, Turner et al. (2024) use a contrastive prompt to induce sentiment shift and detoxification, while Kojima et al. (2024) steer multilingual models to produce more target language tokens in open-ended generation. Finally, Rodriguez et al. (2025) introduce a unified approach to steer activations in LLMs and diffusion models based on optimal transport theory. Overall, work on activation steering demonstrates that it is possible to find expert neurons and use them to steer model activations towards a desired direction, thus demonstrating the causal role of these neurons. What we do not know is whether the set of identified expert neurons is stable across inputs (see Sec. 4) and, if so, whether these representations mirror human knowledge structure.

3 Methods

3.1 Finding expert neurons

We adopt the *finding experts* approach introduced by Suau et al. (2023) for activation steering, to study representational alignment. Our motivation is two-fold: a) this approach has been successfully applied to detect neurons responsible for everyday concepts like "dog", which is the focus of this work; and b) it is able to distinguish the different senses of a homophone (e.g., "apple" as a fruit or company), suggesting that this method is able to pick up fine-grained semantic distinctions.

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 (henceforth *negative set*). Next, we obtain the activations $z_m^c = \{z_{m,i}^c\}_{i=1}^N$ 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, since we know the ground truth label. 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. We calculate the AP score for all units. To be agnostic with respect to the sequence length, the output of each layer is max-pooled across the temporal dimension. Formulated this way, the experts approach has several advantages: as discussed above, it is sensitive to context and can distinguish different senses of a homophone; it can also capture concepts that cannot be represented in one word such as toxicity (Suau et al., 2024).

We consider neurons with an AP score above a given threshold, τ, for a concept to be expert neurons for that concept. τ can be thought of as quality of an expert neuron — the larger the value of τ, the more expert a neuron is for a given concept. In our experiments, we consider a range of values for $\tau \in [0.5, 0.9]$ from a low (classification accuracy above chance) to a high level of expertise.

3.2 Data

To understand the alignment between human and model representations, we examine how patterns in expert neurons relate to perceived concept similarity in humans. We obtain human similarity judgments from two datasets: the MEN dataset (Bruni et al., 2014), which contains $3,000$ word pairs annotated with human-assigned similarity judgments crowd-sourced from Amazon Mechanical Turk, and the Semantic Priming Project (hereafter, SPP), a database of behavioral measures for related and unrelated word pairs (Hutchison et al., 2013).

For each concept under consideration, we generate a set of sentences containing that concept. To ensure dataset diversity, half of each positive dataset is generated with a prompt eliciting story descriptions and half of the dataset is generated with a prompt eliciting factual descriptions of the target concept (the prompts, along with sample

generations, are provided in App. A). 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"). For dataset generation, we experiment with three models of different performance levels: GPT-4 (OpenAI et al., 2024), Mistral-7b-Instruct-v0.2 (Jiang et al., 2023), and an internal 80b-chat model.

For the case study in concept organization in LLMs (Sec. 5), we manually generate lists of ten domains with four concepts per domain (e.g., the domain "animal" containing concepts "cat", "dog", "cheetah"', and "horse"; the full set of domains and concepts is provided in App. E). We choose not to use WordNet (Miller, 1994) — a lexical database of English — because of drawbacks identified in its hierarchical structure, which often make the concept relationships it presents unintuitive (for a discussion, see Gangemi et al., 2001).

3.3 Models

To ensure that the hyper-parameters are not biased towards the particular models we are introspecting, we use different models for selecting the hyper-paramenters and the main experiments. We use GPT-2 (Radford et al., 2019) to select hyper-parameters (e.g., the size of a positive and negative datasets) and validate that our data identifies a stable set of experts (see Sec. 4 for details). For all other experiments, we use models from the Pythia family (Biderman et al., 2023), specifically focusing on model sizes 70m (smallest), 1b, and 12b (largest), to understand the impact of model size on representational alignment. The size of each model is connected to its performance (see App. G for accuracy across the standard benchmarks).

For each model, we work with checkpoints 1, 512, 1k, 4k, 36k, 72k, and 143k, to track how representational alignment develops throughout training. All Pythia models were trained on the same data presented in the same order, allowing us to evaluate the impact of model size and number of training steps on representational alignment while controlling for the data.

4 Can we reliably identify experts?

While the success of expert-based methods at steering model activations is well-documented (Suau et al., 2023, 2024), our interest is in studying model representations through the patterns in experts. Given the novel application of the method, we conduct a pilot study to 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.

For the pilot study, we sample 50 word pairs from the training split of the MEN dataset. For each concept in the word pair, we generate a positive set containing 7000 sentences from three models: GPT-4, Mistral-7b-Instruct-v0.2, and an internal 80b-chat model. We sweep over positive set sizes of 100, 200, 300, 400, and 500 sentences, and negative set sizes of 1000 and 2000 sentences. 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 8 folds). We measure sensitivity in terms of the stability in experts across the folds, where high stability occurs when there is large overlap in the experts across folds. To assess overlap, we look at Jaccard similarity between expert sets across folds, using a range of thresholds τ.

The findings are shown in Fig. 1 for each dataset configuration (subplot) and value of τ (x-axis). 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. One reason could be that as the size of the negative set increases so does the probability of the negative set containing sentences related to the target concept (e.g., a sentence about "cats" may also talk about "dogs"). A second explanation could be that the larger negative set activates more polysemous neurons. Based on these findings, we conduct all subsequent analyses with a positive set

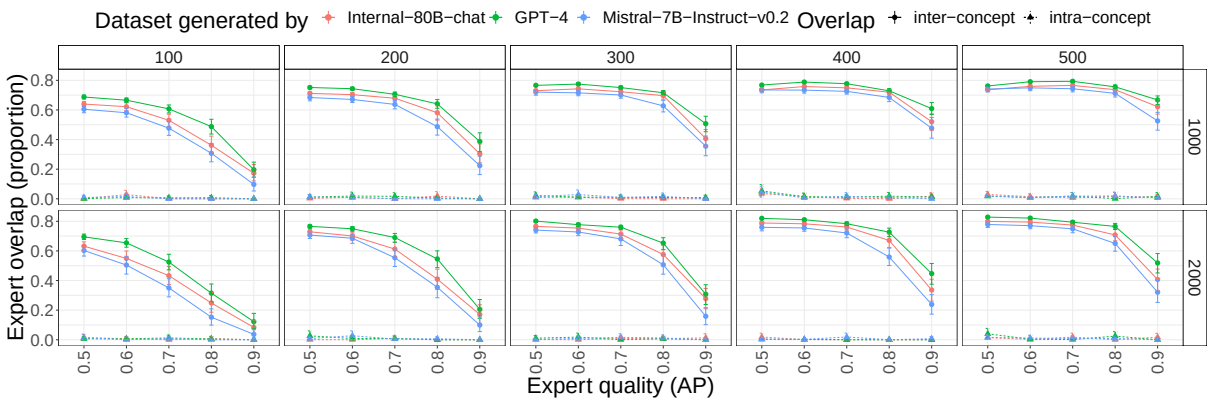

Figure 1: Expert discovery is relatively stable across various dataset characteristics. Points represent condition means; error bars represent bootstrapped 95% confidence intervals. Columns and rows represent the size (number of unique sentences) of the positive and negative sets respectively. Inter-concept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

of 400 sentences and a negative set of 1000 sentences, all generated with Mistral-7b-Instruct-v0.2.

5 Are model and human representations aligned?

We now turn to the main question of our study — whether expert neurons capture semantic information meaningful to humans. 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 LLM's similarity score is the Jaccard similarity between expert sets for $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. In App. D, we consider cosine similarity between the raw AP values as an LLM similarity score, finding similar correlations to those obtained with Jaccard similarity ($\tau = 0.5$), suggesting that what matters most for alignment is not the magnitude of the AP value, but rather whether it is above or below 0.5 (i.e., whether the neuron is positively or negatively associated with the concept). We also analyze human-LLM alignment using the SPP dataset (see App. B) and demonstrate that our finding generalize beyond the MEN dataset.

Expert neuron overlap is highly aligned with human similarity judgments We find that model representations are closely aligned with humans, with the highest alignment occurring at $\tau = 0.5$. At the final checkpoint, the Spearman correlations between expert overlap ($\tau = 0.5$) and MEN similarity are 0.70, 0.77, 0.79 for 70m, 1b, and 12b

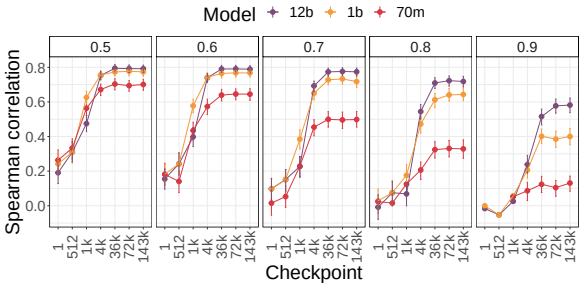

Figure 2: Model representations of similarity are closely aligned with human ones. Points are Spearman correlations between the expert neuron overlap and perceived human similarity in the MEN dataset (significant after checkpoint 1, p<0.05); error bars are bootstrapped 95% confidence intervals. The subplots are τ.

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 Muttenthaler et al., 2023): the 1b and 12b models are virtually indistinguishable, with the 70m model slightly less aligned. The models start diverging in how well aligned they are with humans as τ increases, with larger models being more aligned. This is because smaller models have fewer experts (see Fig. 5) resulting in a lot of empty expert set intersections for higher levels of τ.

Word embeddings are less aligned than expert sets Prior work has focused on the analysis of embeddings when considering alignment in LLM and human representations (Digutsch and Kosinski, 2023). We hypothesize that expert sets are more

correlated with human representations than word embeddings as they disambiguate different word senses (Suau et al., 2023). To test this, we extract the embeddings for each word in the MEN test split from the embedding layer in line with prior work (Digutsch and Kosinski, 2023) and, following the standard approach, from the final hidden layer of the three Pythia models at each checkpoint and compute cosine similarity between the embeddings for each word pair in the MEN test split. We then correlate the cosine similarity with the MEN judgements. These correlations are statistically significant ($p<0.05$) for the embedding layer starting at checkpoints 1k and 4k for the 70m/1b and the 12b models respectively and for the hidden layer starting at checkpoint 512 for all model sizes (see Fig. 3), consistent with prior work (Digutsch and Kosinski, 2023). However, as expected under our hypothesis, the correlations with human similarity are significantly lower for both types of single word embeddings compared to the experts (p-values<0.0001 comparing the alignment based on experts vs. either embeddings types). In addition, while the magnitude of the correlations across the two embeddings types is similar, the patterns of alignment change — for the embedding layer, the alignment stably grows over training while the pattern of alignment in the final layer embeddings is unstable across checkpoints. Moreover, the two types of embeddings disagree on which model size is more aligned with humans. Thus, single word embeddings are not only less aligned with humans than experts but are also highly sensitive to hyperparameters.

LLM's concept organization mirrors human conceptual structure Having established that the expert overlap is predictive of human-perceived concept similarity, we 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 (Graf et al., 2016; Murphy, 2004; Rosch, 1978). For example, "dog", "cat" and "horse" are all *animals* and "bike", "bus", and "car" are all *vehicles*. This raises the question of whether models organize concepts in a similar way. To assess this, we consider a list of domains we generated (see Sec. 3.2 and App. E), the experts associated with each concept in the list (τ=0.5), and their reciprocal overlap. For this analysis, we only consider the final (143k) checkpoint. We discuss Pythia 12b in the main text and present other model sizes in App. F.

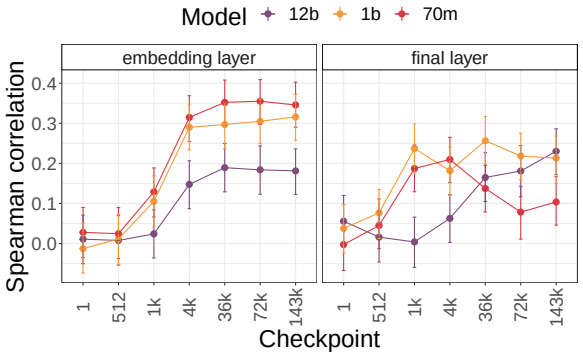

Figure 3: Spearman correlations between embedding cosine similarity and perceived human similarity in the MEN dataset. Error bars are bootstrapped 95% confidence intervals. Subplots indicate the layer the embeddings were extracted from. The correlations are significant ($p<0.05$) starting at checkpoints 512 for the last layer (all model sizes) and 1k (70m and 1b models) and 4k (12b model) for the embeddings layer.

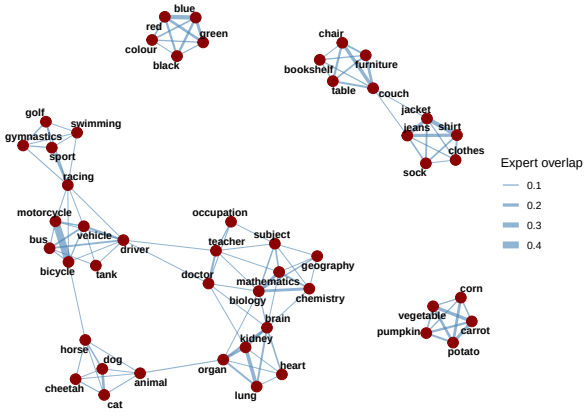

Figure 4: Similarity of concept representations in the LLM, based on expert overlap. Each node represents a concept; edge thickness corresponds to the degree of reciprocal expert overlap between concepts.

Fig. 4 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. This is consistent with the findings on representation alignment discussed in Sec. 5, demonstrating that concept pairs perceived as similar by humans show higher expert overlap compared to dissimilar concept pairs. On top of that,

Fig. 4 shows meaningful between-domain connections unveiled by the study of expert sets. 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. Finally, looking at the internal organization of the domains, we notice that broader concepts (e.g., "vehicle" or "animal") tend to show weaker overlap with specific instances in their domain compared to the overlap between closely related specific concepts, e.g., "motorcycle" and "bicycle", or "dog" and "cat". This may reflect distributional factors, with narrower concepts exhibiting stronger co-occurrence patterns.

To quantify whether domain structures emerge in the LLM's knowledge representation, we propose that, if the model organizes concepts in human-interpretable domains, concepts from the same domain (e.g., "dog", "cat", "horse", and "cheetah") should share a consistent set of experts, and some of these shared experts should also be associated with the broader concept describing the domain (e.g., "animal" in our example). Our results reveal a clear and systematic pattern: within each domain, a consistent set of expert neurons is shared across all associated concepts. On average, 2.24% of the experts identified across all concepts in a domain are jointly shared among them. Notably, 58.45% of this shared core is also shared by the broader concept representing the domain (see App. F for the complete result set). To validate the significance of our findings, we compare them against a baseline in which domain groupings are randomly sampled (e.g., associating "animal" with "jacket", "liver", "doctor", and "red"). In this case, the overlap among expert sets drops significantly (average 0.01% and 5.81% of shared neurons for all concepts and by the broader concept respectively, p-values <0.001) confirming that the structure we observe is unlikely due to chance.

Overall, our findings suggest that the experts approach captures human-interpretable domain-level structures beyond simple word pair similarity.

6 Characterizing model knowledge

We conclude with characterizing the differences in experts as a function of model size and stage of training, by reanalyzing the data from Sec. 5.

Experts are learned from the data, with larger models having more experts Larger models allocate more experts to a given concept (see Fig. 5; the pattern does not change after scaling the raw number of experts by the number of neurons in the model). As τ increases and experts become more specialized, fewer experts are identified; the drop is more pronounced for smaller models. Overall, larger models have a greater capacity to learn a higher number of experts and a higher number of *more specialized* experts. This increased specialization may contribute to finer-grained concept representations and ultimately better performance on downstream tasks.

Interestingly, we observe a large number of experts at checkpoint 1, followed by a drop and then a steady gradual increase in the number of experts as training continues. This is expected from the perspective of language modeling as compression (Shwartz-Ziv and Tishby, 2017; Delétang et al., 2024). Early in training, the model discovers a large number of experts. While they are not yet meaningful (indicated by non-significant correlation with human similarity), they ensure the model can efficiently allocate representational capacity for later in training. As the model starts learning the relevant relationships, the number of experts drops (checkpoint 512) and then slowly recovers as the model continues learning (checkpoint 1k onwards). As training continues, the experts become more and more meaningful, as evidenced by the increasing correlation between the expert overlap and human similarity judgments. The idea that experts are learned from training data is further supported by the finding of a mode of 0 experts in all models initialized with random weights.

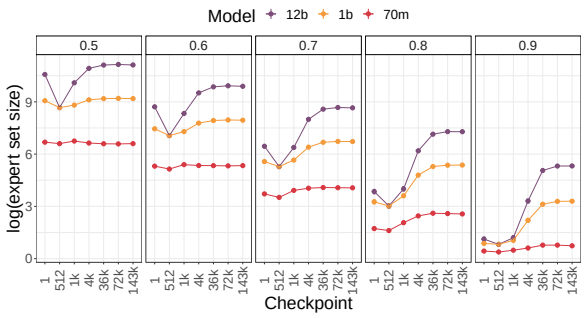

Figure 5: Expert set size (log) by model size and checkpoint. Points are averages over all concepts; error bars are bootstrapped 95% confidence intervals. Subplots are different values of τ.

More specialized experts take longer to learn We next look at the dynamics of learning experts across checkpoints. We calculate expert overlap

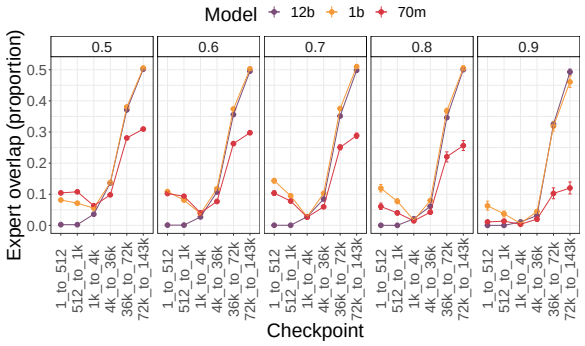

Figure 6: Proportion of expert overlap across subsequent checkpoints (e.g., 1_to_512 is overlap between checkpoints 1 and 512). Points are across concept averages; error bars are bootstrapped 95% confidence intervals. Subplots are different values of τ.

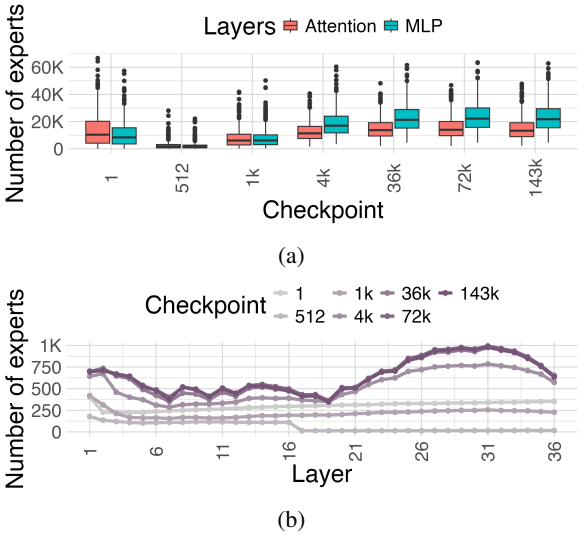

Figure 7: Pythia 12b. (a) Total number of experts in MLP and attention layers across checkpoints; (b) Average number of experts in MLP layers at different depths, for different checkpoints.

(Jaccard similarity) for each concept across subsequent checkpoints in our data. The stability of the discovered expert set grows as training progresses (Fig. 6). Early in training (prior to step 36k), expert overlap between subsequent checkpoints is low across model sizes, suggesting that semantic knowledge has not been acquired yet. As τ increases (corresponding to higher expert specialization), it takes longer for the expert set to stabilize, suggesting that higher-quality experts take longer to learn.

More experts are found in MLPs and deeper layers Pythia models consist of intertwined self-attention and MLP layers (Biderman et al., 2023) each serving different functions (Geva et al., 2021; Jawahar et al., 2019; Liu et al., 2019). We analyze the distribution of experts within these layers. Fig. 7a shows the patterns for Pythia 12b (τ=0.5). More experts are located in the MLP layers compared with attention layers, with the relative allocations stabilizing at checkpoint 4k. We see the same trend in smaller models (App. H.1) after controlling for the number of neurons in the respective layers. The mean number of experts generally increases with layer depth in MLPs, with checkpoint 4k again displaying the first recognizable structure (see Fig. 7b and App. H.2). For attention layers, high numbers of experts are located in deep layers and, interestingly, the first layer (see App. H.3). Of note, if we look only at highly specialized experts (τ>=0.9), we find higher numbers of experts in earlier layers (see App. H.8 and H.9), reproducing patterns identified in Suau et al. (2020). Our findings align with prior research on the role of layers at different depths, identifying deeper layers as responsible for processing higher-level semantic knowledge captured by expert neurons (Geva et al., 2021; Jawahar et al., 2019). Of note, we find no difference in these patterns for concepts with broader *vs* narrower meanings (e.g., "animal" *vs.* "dog"), see App. I and App. J, suggesting that LLMs do not differentiate between generic and specific concepts based on *where* resources are specialized for them.

7 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 embeddings, and approach human alignment levels. Consistent with prior work in vision, (Muttenthaler et al., 2023), we find that model size has little influence on alignment.

Our approach reveals that models generally organize concepts into human-interpretable domains. Some domains are more structured than others, and this pattern remains consistent across model sizes. We leave it to future work to investigate factors that could give rise to this pattern, such as the frequency of each domain in the training data. We hope that this work will serve as a foundation for future research not only on alignment, but also at the intersection of cognitive science and AI theory, exploring whether fundamental cognitive principles (Murphy, 2004; Margolis and Laurence, 2003) are reflected in neural network representations.

8 Limitations

We consider only a simple case of similarity Consistent with prior work (Digutsch and Kosinski, 2023; Shaki et al., 2023; Misra et al., 2020), we study alignment between human and model representations, which we operationalize as the similarity between two concepts. We find that model size does not play a large role in alignment: even models as small as 70m excel in this alignment test. While this finding is consistent with previous literature (Muttenthaler et al., 2023) and replicated over two datasets, it is also possible that our task is too simple to distinguish between the models. This is supported by the observations that semantic relationships studied here start emerging early in training (around checkpoint 4k out of 143k). Future work will consider more complex cases of alignment, such as alignment with human values or preferences.

We do not study patterns in expert neurons through activating these neurons Our interest is in exploring whether the discovered neurons capture the dimensions meaningful to humans, and to this end we look at alignment. Note that, given our research question, simply activating a concept (i.e., what the method is designed to do) is of limited interest. In contrast, activating the expert intersection between two concepts—for which the method was not originally designed nor tested— may be a meaningful exploration to better understand concept representation. For instance, we could have activated the shared experts between "animal" and "dog" and examined model generations after the activation. We chose not to do this for the following reason: the approach we are using requires choosing the number of experts and the original work (Suau et al., 2023) has shown that this choice impacts the quality of generations and the degree to which a concept is expressed — an effect that we also observed in our preliminary investigations. We leave such hyper-parameter search to future work: a priori, we do not have a clear hypothesis about whether activating more specialized experts *vs.* less specialized ones within the intersection would lead to distinct generation patterns; or if any discernible pattern in those generations should be expected at all. Given these uncertainties, we did not feel confident that this analysis would yield reliable results. Other approaches do not require choosing the number of experts (Rodriguez et al., 2025), but these approaches are designed to change the activations of all neurons in the network and are thus not applicable for our use case.

We do not have access to training data To fully understand how knowledge develops in LLMs, we need to know what the model has seen at different points in training. Unfortunately, the Pile (Gao et al., 2020) that Pythia models were trained on is no longer available.

Model choice Given the nature of our research question, it is crucial to be able to analyze multiple checkpoints from models of varying sizes, prioritizing interpretability over direct evaluations of model performance. For this reason, we rely on the Pythia family of models, publicly released in the interest of fostering interpretability research. We leave to future work the exploration of alignment and its emergence in alternative model families (e.g., the recent OLMo 2 family; Walsh et al., 2025).

Mechanistic intepretability Our work relates to the fast-growing field of mechanistic interpretability that seeks to reverse-engineer LLMs into human-interpretable components, revealing the neural pathways and architectural components by which models process information (Geiger et al., 2021; Feng and Steinhardt, 2024; Vasileiou and Eberle, 2024). Unlike mechanistic interpretability that focuses on the discovery of components in network architectures, our goal is to assess whether the knowledge representation in the model is aligned with human representations. While not explicitly looking for architecture-based interpretations, we find that concepts from related domains like "dog", "cat", and "animal" share a consistent set of experts, suggesting that the same architectural components (neurons) are implicated in the alignment. We leave uncovering the neural pathways and causal components underlying alignment to future work.

We study neurons individually In this work, neurons are studied individually. That is, our analysis assumes that the representation of concepts is aligned with the canonical basis induced by the neurons. We have two reasons to assume that this is the case. First, previous work suggests that intervening on neurons identified in this method can steer generations to favor or avoid a concept (Suau et al., 2023, 2024). Second, in our analysis we see that neurons identified in this manner capture key properties of concepts: the correlation between expert-based concept similarity measures and human concept similarity evaluations is comparable

to inter-human correlation. It is, however, possible that looking at neurons jointly would capture additional aspects of concept representation. We leave this exploration to future work.

References

2020. Winogrande: An adversarial winograd schema challenge at scale. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Jeremy Auguste, Arnaud Rey, and Benoit Favre. 2017. Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *Preprint*, arXiv:2310.10449.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Tom B. Brown, Benjamin Mann, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. *Preprint*, arXiv:2309.10668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Jan Digutsch and Michal Kosinski. 2023. Overlap in meaning is a stronger predictor of semantic activation in gpt-3 than in humans. *Scientific Reports*, 13(1):5035.

Allyson Ettinger and Tal Linzen. 2016. Evaluating vector space models using human semantic priming results. *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. 2022. Harmonizing the object recognition strategies of deep neural networks with humans. In *Advances in Neural Information Processing Systems*, volume 35, pages 9432–9446. Curran Associates, Inc.

Jiahai Feng and Jacob Steinhardt. 2024. How do language models bind entities in context? *ICLR*.

Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. 2001. Conceptual analysis of lexical taxonomies: The case of wordnet top-level. *Preprint*, arXiv:cs/0109013.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of nural networks. *NeurIPS*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Conference on Empirical Methods in Natural Language Processing*.

Caroline Graf, Judith Degen, Robert D. Hawkins, and Noah D. Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. *Cognitive Science*.

Martin N. Hebart, Chie-Yu Zheng, Francisco Pereira, and Chris I. Baker. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Keith A. Hutchison, David A. Balota, James H. Neely, Michael J. Cortese, Emily R. Cohen-Shikora, Chi-Shing Tse, Melvin J. Yap, Jesse J. Bengson, Dale Niemeyer, and Erin Buchanan. 2013. The semantic priming project. *Behav Res*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *Preprint*, arXiv:2403.02901.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*.

Hoang Anh Just, Ming Jin, Anit Sahu, Huy Phan, and Ruoxi Jia. 2024. Data-centric human preference optimization with rationales. *arXiv preprint arXiv:2407.14477*.

Meenakshi Khosla and Leila Wehbe. 2022. High-level visual areas act like domain-general filters with strong selectivity and functional specialization.

Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. 2023. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Margolis and Stephen Laurence. 2003. Concepts. In Stephen Stich Ted Warfield, editor, *The Blackwell Guide to the Philosophy of Mind*, pages 190–213. Blackwell.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Gregory Murphy. 2004. *The Big Book of Concepts*. MIT Press.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. 2023. Human alignment of neural network representations. *Preprint*, arXiv:2211.01201.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset.

Hyuntae Park, Yeachan Kim, Jun-Hyung Park, and SangKeun Lee. 2024. Zero-shot commonsense reasoning over machine imagination. *Preprint*, arXiv:2410.09329.

Andi Peng, Ilia Sucholutsky, Belinda Z. Li, Theodore R. Sumers, Thomas L. Griffiths, Jacob Andreas, and Julie A. Shah. 2024. Learning with language-guided state abstractions. *ICLR*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation.

Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. 2018. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Preprint*, arXiv:1706.02417.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, marco cuturi, and Xavier Suau. 2025. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations*.

Eleanor Rosch. 1978. Principles of categorization. In Eleanor Rosch and B. B. Lloyd, editors, *Cognition and Categorization*, pages 27–48. Erlbaum, Hillsdale, NJ.

Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. 2016. Exploring semantic representation in brain activity using word embeddings. *EMNLP*.

Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. 2023. Cognitive effects in large language models. In *ECAI 2023*, pages 2105–2112. IOS Press.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. Valuecompass: A framework of fundamental values for human-ai alignment. *Preprint*, arXiv:2409.09586.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *Preprint*, arXiv:1703.00810.

Xavier Suau, Pieter Delobelle, Katherine Metcalf, Armand Joulin, Nicholas Apostoloff, Luca Zappella, and Pau Rodríguez. 2024. Whispering experts: Neural interventions for toxicity mitigation in language models. *Preprint*, arXiv:2407.12824.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2020. Finding experts in transformer models. *Preprint*, arXiv:2005.07647.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2023. Self-conditioning pre-trained language models. *Preprint*, arXiv:2110.02802.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2024. Getting aligned on representational alignment. *Preprint*, arXiv:2310.13018.

Anders Søgaard. 2016. Evaluating word embeddings with fmri and eye-tracking. *ACL*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2022. Contrastive representation distillation. *Preprint*, arXiv:1910.10699.

Mariya Toneva. 2022. Bridging Language in Machines with Language in the Brain.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *Preprint*, arXiv:2302.08399.

Alexandros Vasileiou and Oliver Eberle. 2024. Explaining text similarity in transformer models. *NAACL*.

Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya

Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

A Prompts used for probing dataset generation and sample generations

Fact prompt: "Generate a set of 10 sentences, including as many facts as possible, about the concept [concept name] as [a/an] [adjective/noun/verb] and defined as [WordNet definition]. Refer to the concept only as [concept name] without including specific classes, types, or names of [concept name]. Make sure the sentences are diverse and do not repeat."

Sample fact sentences for concept **poppy** defined as 'annual or biennial or perennial herbs having showy flowers':
GPT-4: Gardeners often classify poppies as easy to care for due to their hardy nature.
Mistral-7b-Instruct-v0.2: Poppies are herbaceous plants that can grow annually, biennially, or perennially, depending on the specific species.
Internal 80b-chat model: Poppies have been used in traditional medicine for centuries, with various parts of the plant being employed to treat ailments like pain, insomnia, and digestive problems.

Story prompt: "Generate a set of 10 sentences, where each sentence is a short story about the concept [concept name] as [a/an] [adjective/noun/verb] and defined as [WordNet definition]. Refer to the concept only as [concept name] without including specific classes, types, or names of [concept name]. Make sure the sentences are diverse and do not repeat."

Sample story sentences for concept **poppy** defined as 'annual or biennial or perennial herbs having showy flowers':
GPT-4: As the wedding gift from her grandmother, a dried poppy was framed and hung on her wall.
Mistral-7b-Instruct-v0.2: As the farmer tended to his fields, he couldn't help but admire the poppies that grew among his crops, their beauty a welcome distraction.
Internal 80b-chat model: The poppy, a harbinger of spring, adorned the hillsides with a colorful tapestry, signaling the end of winter's slumber.

B Generalization of the findings to the Semantic Priming Dataset

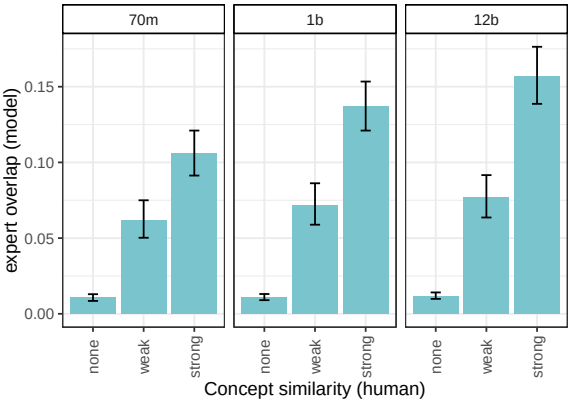

Figure 8: Expert overlap in the model is predicted by human-perceived similarity level. Bars represent expert overlap averaged over all concept pairs; error bars represent bootstrapped 95% confidence intervals. The subplots are model sizes.

To ensure our findings generalize beyond the MEN dataset, we repeat our main analysis on a subset of the Semantic Priming Project (SPP) (Hutchison et al., 2013), which contains 1,661 target words paired with related or unrelated concepts. The advantage of the SPP dataset over MEN is that it contains a more varied set of concepts. The drawback is that the range of similarity levels between the concepts is more limited — SPP only contains three levels of similarity: strongly related, somewhat related, and unrelated concepts. We expect that expert overlap will increase as human-perceived similarity level increases.

We sample 100 pairs from each of the three similarity bins in the SPP dataset and extract the experts for each concept in the pair from the final (143k) checkpoint for the three Pythia models under consideration. We then use linear mixed-effects regression to predict expert overlap from model (sliding difference coded[1]: 1b vs. 70m and 12b vs. 1b) and similarity level (sliding difference coded: weak vs. none and strong vs. weak). The model included the maximal converging random effects structure (random intercepts for the two concepts in a pair). For models of all sizes, we find a statistically significant increase in expert overlap with increased similarity (all p's > 0.0001; see Fig. 8).

[1]Sliding difference coding compares the mean of the dependent variable for one level of the categorical variable to the mean of the dependent variable for the preceding adjacent level (e.g., 1b model vs. 70m model).

C Expert set sizes in the pilot experiment

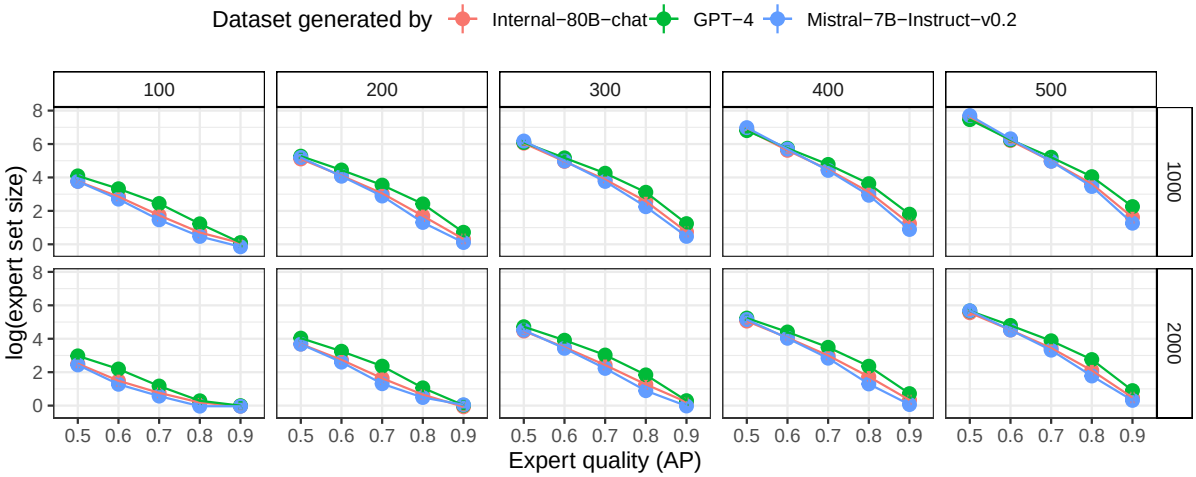

Figure 9: Expert set size (log) in the pilot experiment. Points represent condition means; error bars represent bootstrapped 95% confidence intervals. Columns represent the size of the positive set (number of unique sentences); rows represent the size of the negative set (number of unique sentences).

D Analyses of correlations between human similarity judgments and cosine similarity for the full network

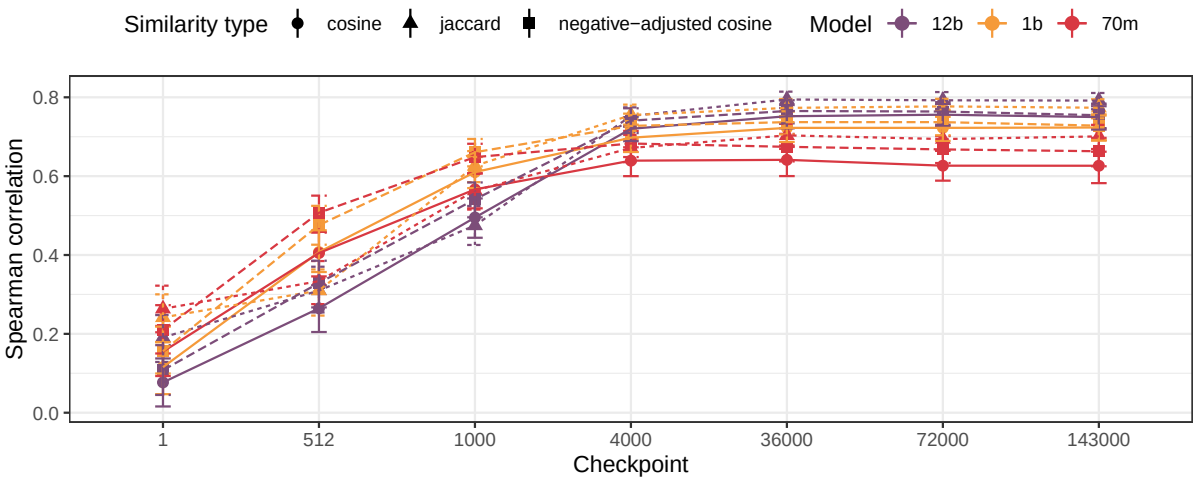

Figure 10: Spearman correlations between human similarity judgments, cosine similarity over raw AP values, negative-adjusted cosine similarity [abs(AP)-0.5], and the best-performing τ of Jaccard similarity (0.5). Points represent Spearman correlations between LLM's similarity and perceived human similarity in the MEN dataset; error bars represent bootstrapped 95% confidence intervals.

E List of concepts in semantically-related domains

Domain	Concepts
animals	cat, dog, cheetah, horse, animal
clothes	jacket, jeans, shirt, sock, clothes
colours	red, blue, green, black, colour
furniture	chair, bookshelf, table, couch, furniture
occupations	doctor, teacher, driver, musician, occupation
organs	heart, kidney, lung, brain, organ
sports	golf, racing, gymnastics, swimming, sport
subjects	mathematics, geography, biology, chemistry, subjects
vegetables	carrot, potato, pumpkin, corn, vegetable
vehicles	bus, tank, motorcycle, bicycle, vehicle

Table 1: List of concepts in our domains.

F Complete results for domain-level organization

Model	Ckpt	% in dom	% with broader
70m	1	0.05 0.00	0.00 0.00
	$36k$	0.97 0.01	70.41 0.33
	$72k$	1.19 0.00	59.69 0.33
	$143k$	1.39 0.01	67.19 0.92
1b	1	0.02 0.00	0.24 0.33
	$36k$	1.81 0.01	60.87 2.67
	$72k$	1.84 0.03	63.43 2.24
	$143k$	2.02 0.01	63.84 2.85
12b	1	0.12 0.00	0.01 0.52
	$36k$	1.87 0.01	58.66 5.11
	$72k$	2.12 0.01	57.85 5.50
	$143k$	2.24 0.01	58.45 5.81

Table 2: Results of expert overlap in semantically-organized domains, across different models and checkpoints. Column 3 shows the average percentage of experts shared between all the specific concepts in a domain (e.g., "dog", "cat", etc.). Column 4 reports the percentage of this shared core also activated by the broader concept representing the domain (e.g., "animal"). Baseline values are shown in gray. Our results are significantly different from the randomized baseline starting from checkpoint $36k$, suggesting that domain-like structures seem to have fully emerged at that stage.

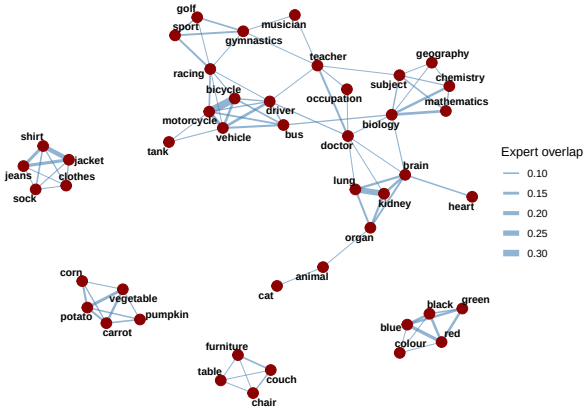

Figure 11: **Pythia70m, ckpt 143k** Similarity of concept representations in the LLM, based on expert overlap. Each node represents a concept; edge thickness corresponds to the degree of reciprocal expert overlap between concepts.

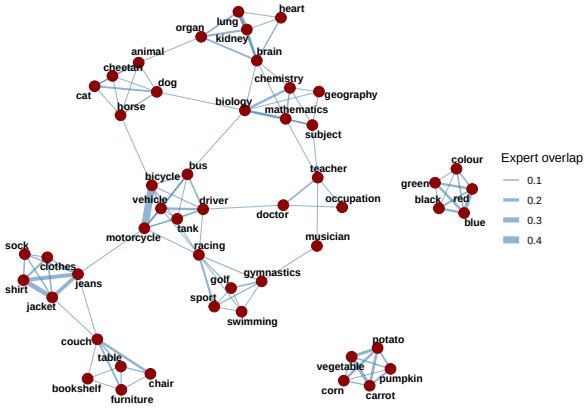

Figure 12: **Pythia1b, ckpt 143k** Similarity of concept representations in the LLM, based on expert overlap. Each node represents a concept; edge thickness corresponds to the degree of reciprocal expert overlap between concepts.

G Pythia evaluation benchmarks

The mean accuracy and standard error across eight benchmarks shown in Table 3 is 0.27 (0.01) for the 70m, 0.28 (0.01) for the 1b model, and 0.32 (0.02) for the 12b model at the end of training.

Benchmarks	
LAMBADA – OpenAI	Paperno et al. (2016)
PIQA	Bisk et al. (2020)
SciQ	Johannes Welbl (2017)
ARC (easy and hard)	Clark et al. (2018)
WinoGrande	win (2020)
MMLU	Hendrycks et al. (2021)
LogiQA	Liu et al. (2020)
Winograd Schema Challenge	Levesque et al. (2012)

Table 3: Pythia evaluation benchmarks.

H Additional materials for layer analyses

H.1 Total number of experts in MLP and attention layers

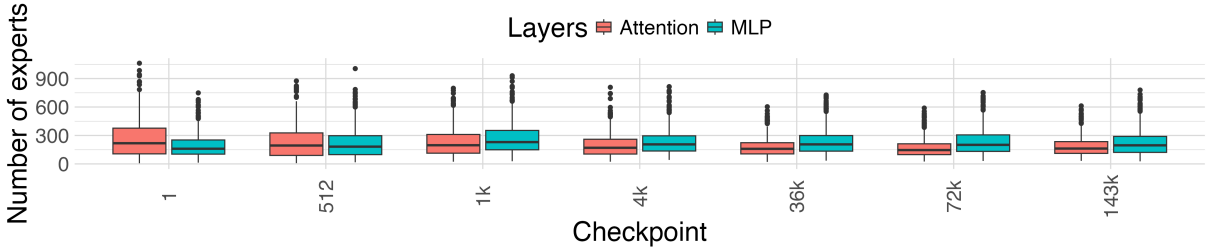

Figure 13: **Pythia 70m.** Total number of experts in MLP and attention layers across checkpoints

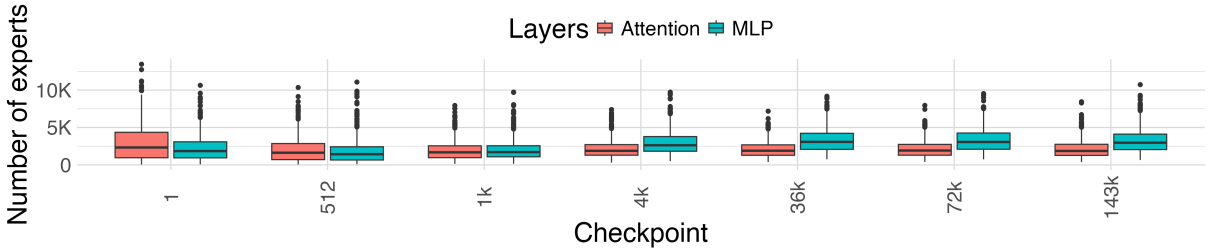

Figure 14: **Pythia 1b.** Total number of experts in MLP and attention layers across checkpoints

H.2 Distribution of experts across MLP layers

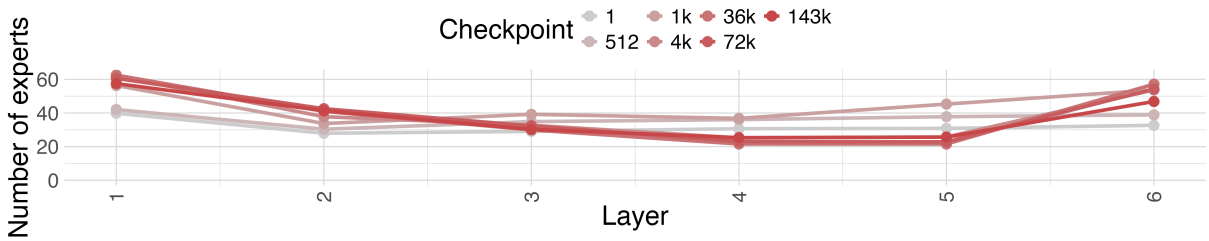

Figure 15: **Pythia 70m.** Average number of experts identified in MLP layers at different depths, for different checkpoints.

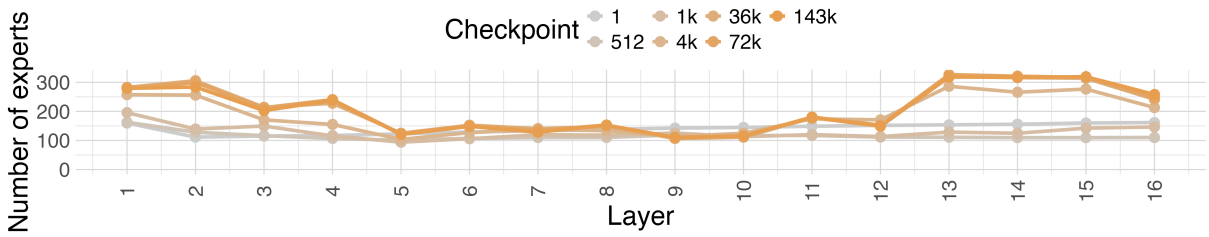

Figure 16: **Pythia 1b.** Average number of experts identified in MLP layers at different depths, for different checkpoints.

H.3 Distribution of experts across attention layers

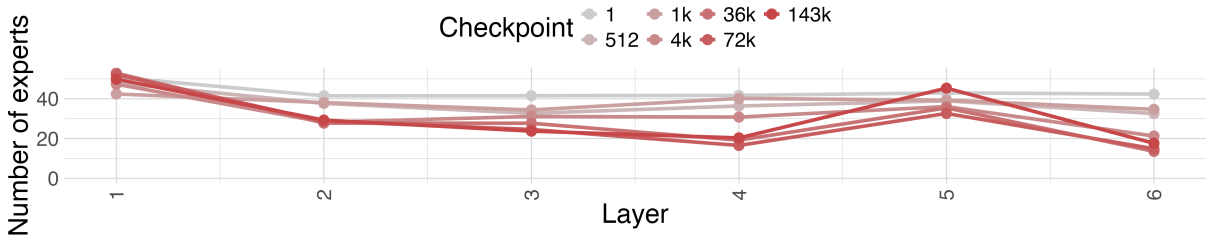

Figure 17: **Pythia 70m.** Average number of experts identified in attention layers at different depths, for different checkpoints.

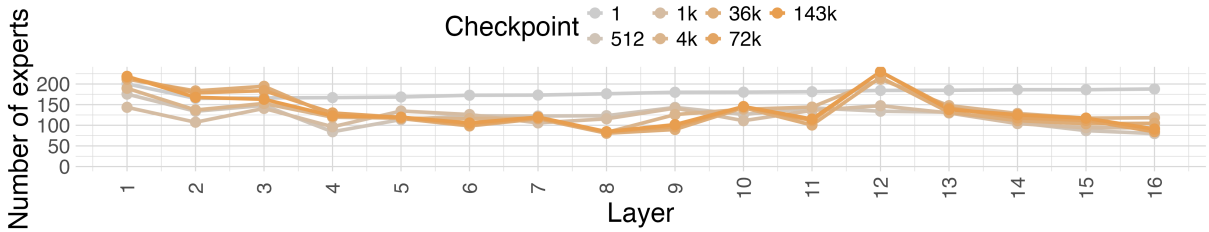

Figure 18: **Pythia 1b.** Average number of experts identified in attention layers at different depths, for different checkpoints.

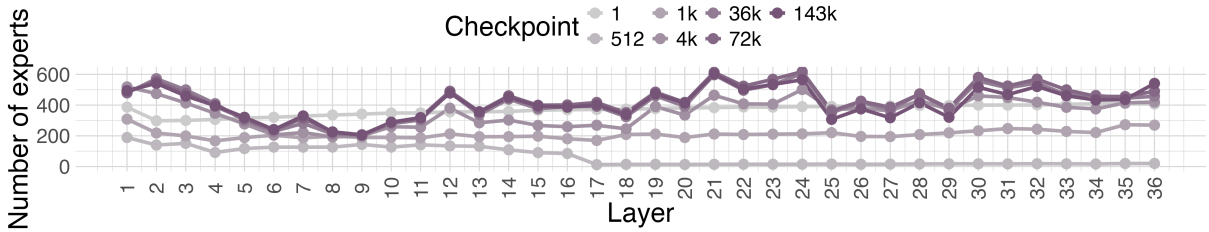

Figure 19: **Pythia 12b.** Average number of experts identified in attention layers at different depths, for different checkpoints.

H.4 Distribution of experts across `MLP.dense.h_to_4h` layers

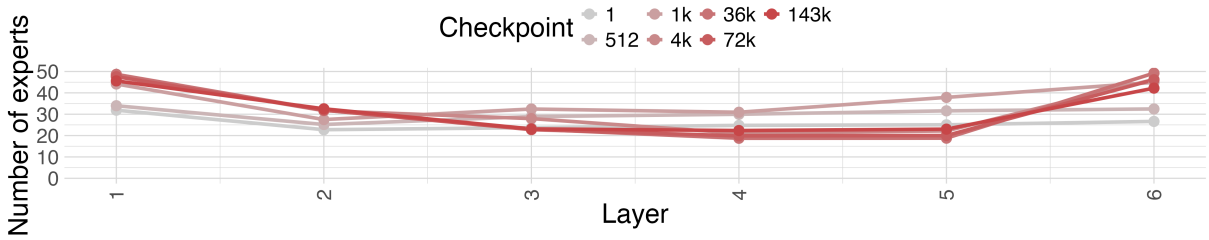

Figure 20: **Pythia 70m.** Average number of experts identified in the MLP `h_to_4h` (part of the MLP layers) at different depths, for different checkpoints.

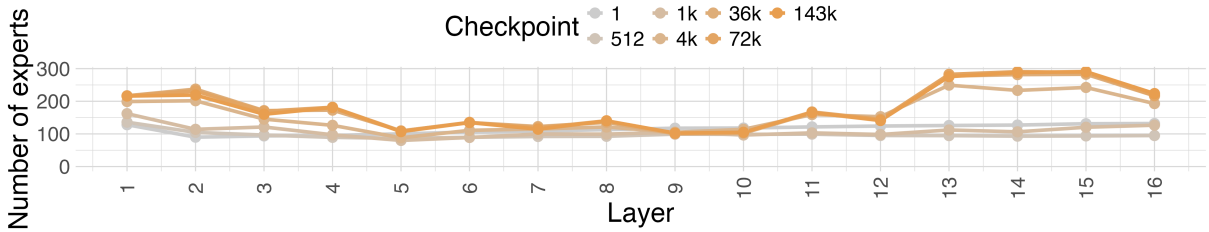

Figure 21: **Pythia 1b.** Average number of experts identified in the MLP `h_to_4h` (part of the MLP layers) at different depths, for different checkpoints.

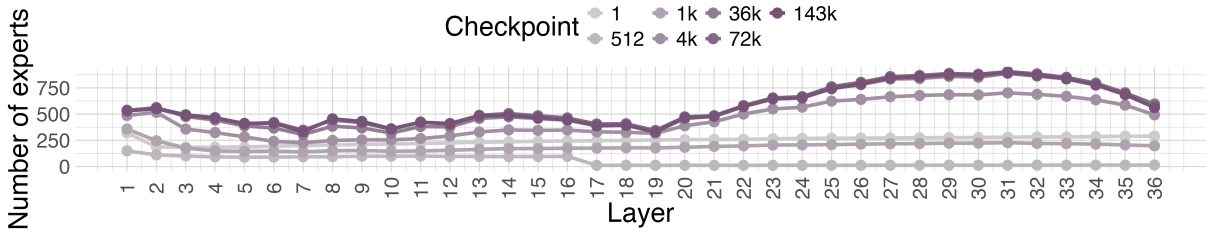

Figure 22: **Pythia 12b.** Average number of experts identified in the MLP `h_to_4h` (part of the MLP layers) at different depths, for different checkpoints.

H.5 Distribution of experts across `MLP.dense.4h_to_h` layers

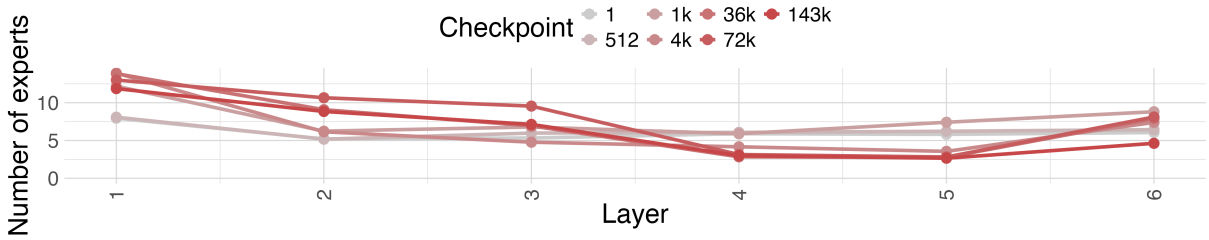

Figure 23: **Pythia 70m.** Average number of experts identified in the MLP `4h_to_h` (part of the MLP layers) at different depths, for different checkpoints.

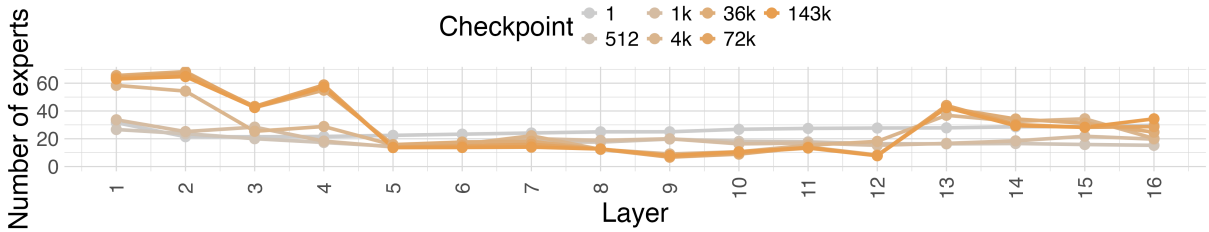

Figure 24: **Pythia 1b.** Average number of experts identified in thr MLP `4h_to_h` (part of the MLP layers) at different depths, for different checkpoints.

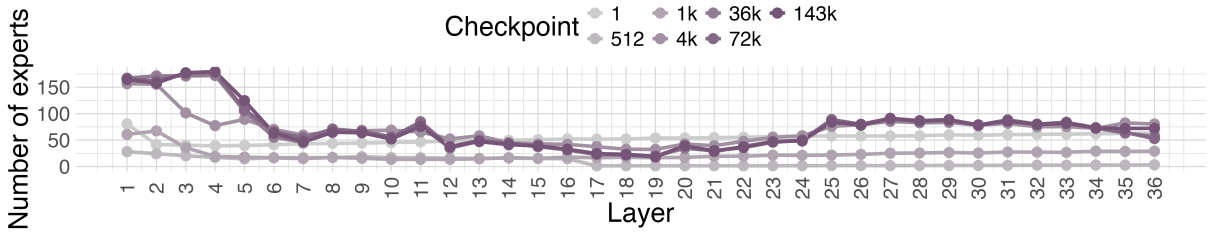

Figure 25: **Pythia 12b.** Average number of experts identified in the MLP `4h_to_h` (part of the MLP layers) at different depths, for different checkpoints.

H.6 Distribution of experts across `attention.query_key_value` layers

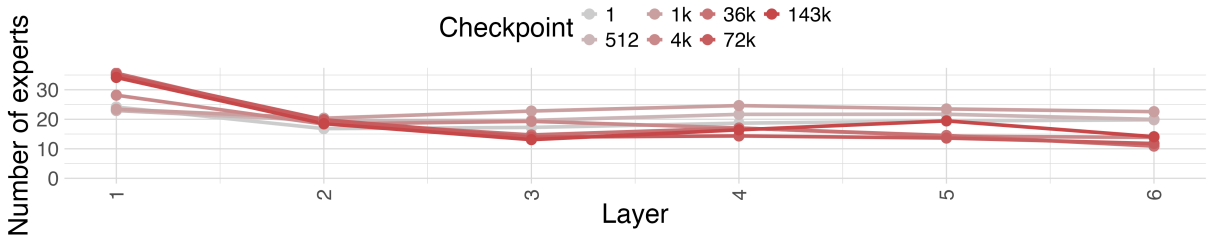

Figure 26: **Pythia 70m.** Average number of experts identified in the `attention.query_key_value` (part of the attention layers) at different depths, for different checkpoints.

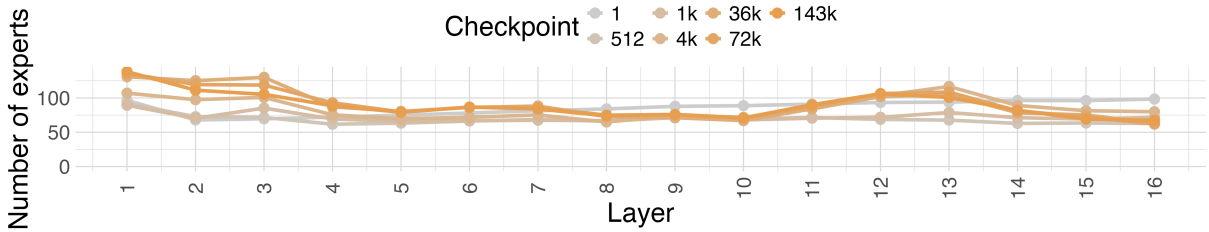

Figure 27: **Pythia 1b.** Average number of experts identified in the `attention.query_key_value` (part of the attention layers) at different depths, for different checkpoints.

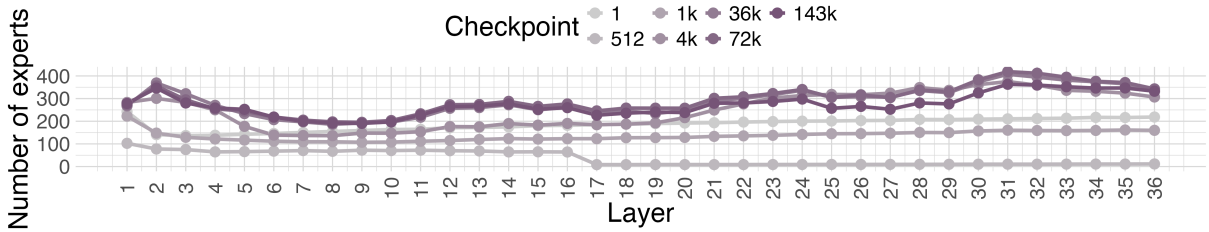

Figure 28: **Pythia 12b.** Average number of experts identified in the `attention.query_key_value` (part of the attention layers) at different depths, for different checkpoints.

H.7 Distribution of experts across `attention.dense` layers

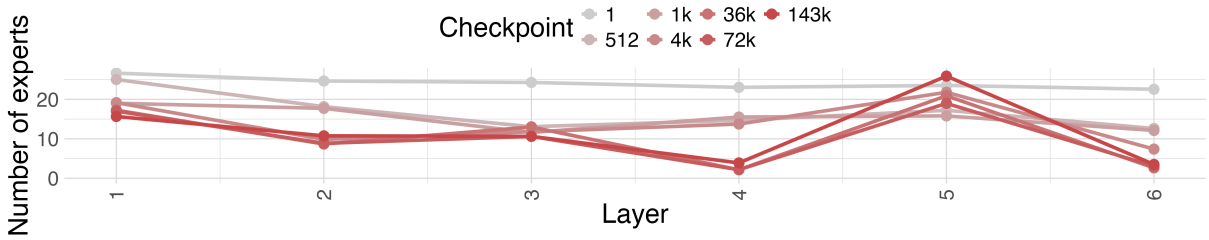

Figure 29: **Pythia 70m.** Average number of experts identified in the `attention.dense` (part of the attention layers) at different depths, for different checkpoints.

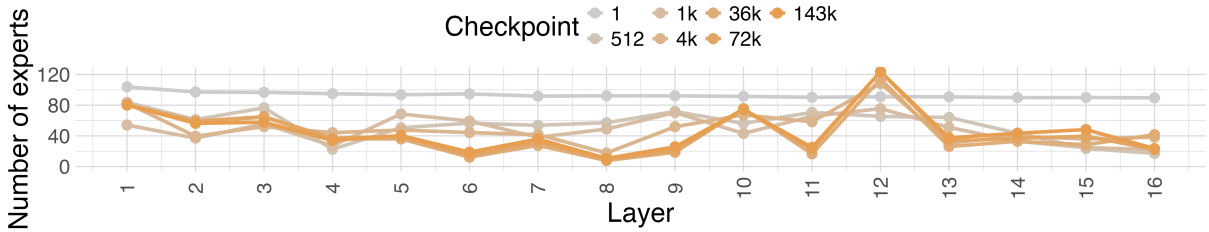

Figure 30: **Pythia 1b.** Average number of experts identified in the `attention.dense` (part of the attention layers) at different depths, for different checkpoints.

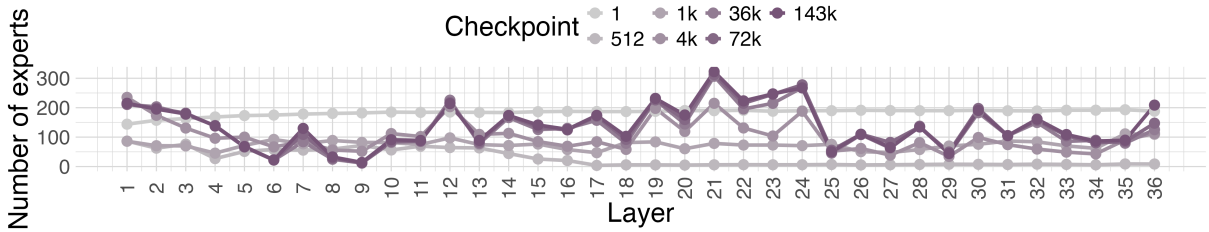

Figure 31: **Pythia 12b.** Average number of experts identified in the `attention.dense` (part of the attention layers) at different depths, for different checkpoints.

H.8 Distribution of highly specialized experts across MLP layers

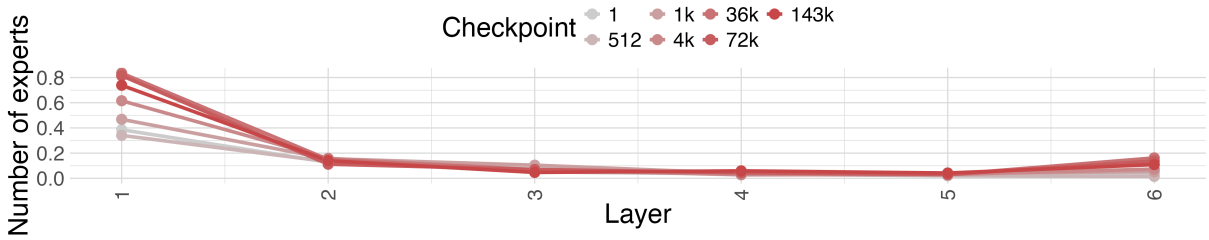

Figure 32: **Pythia 70m.** Average number of highly specialized experts ($\tau = 0.9$) identified in MLP layers at different depths, for different checkpoints.

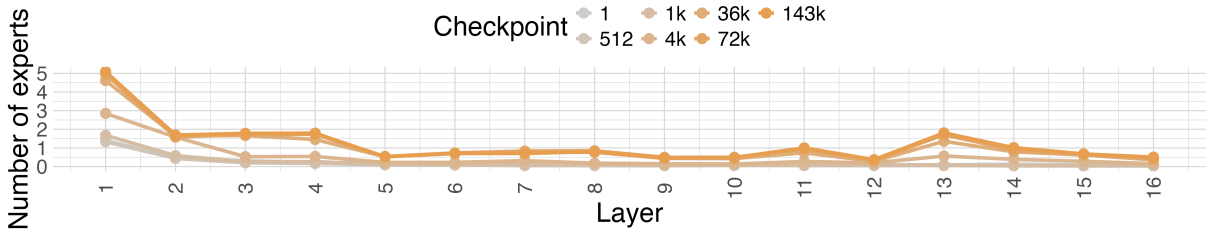

Figure 33: **Pythia 1b.** Average number of highly specialized experts ($\tau = 0.9$) identified in MLP layers at different depths, for different checkpoints.

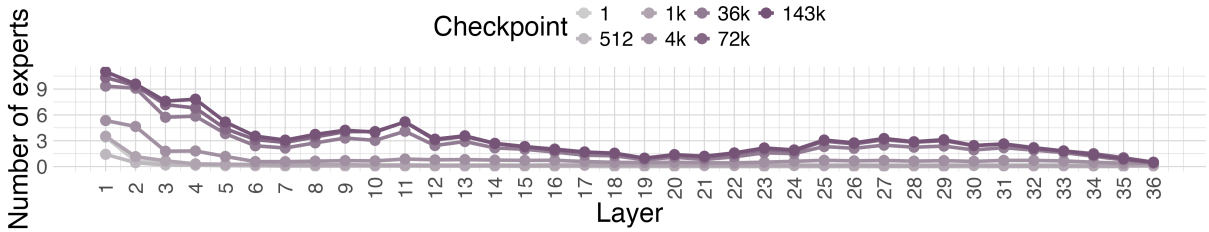

Figure 34: **Pythia 12b.** Average number of highly specialized experts ($\tau = 0.9$) identified in MLP layers at different depths, for different checkpoints.

H.9 Distribution of highly specialized experts across attention layers

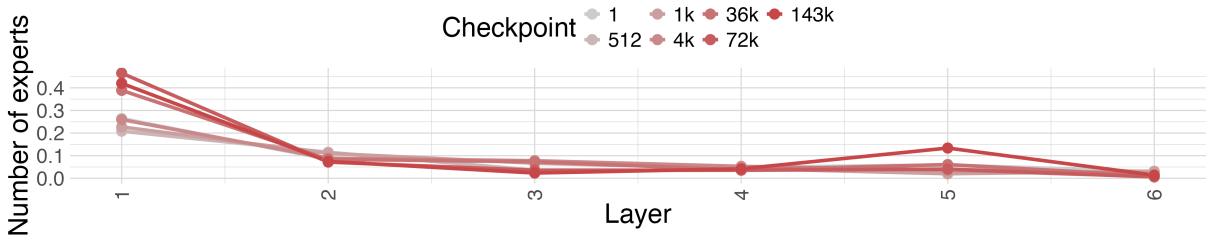

Figure 35: **Pythia 70m.** Average number of highly specialized experts ($\tau = 0.9$) identified in attention layers at different depths, for different checkpoints.

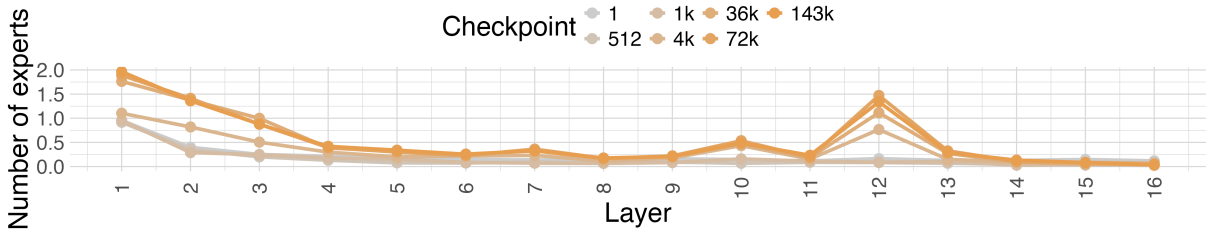

Figure 36: **Pythia 1b.** Average number of highly specialized experts ($\tau = 0.9$) identified in attention layers at different depths, for different checkpoints.

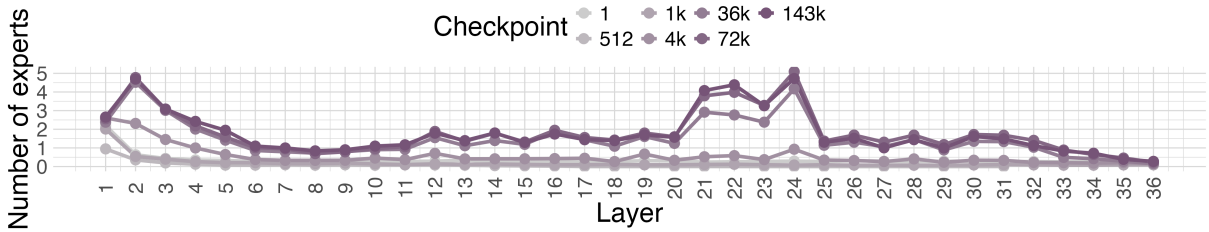

Figure 37: **Pythia 12b.** Average number of highly specialized experts ($\tau = 0.9$) identified in attention layers at different depths, for different checkpoints.

I Distribution of experts for broader concepts

I.1 MLP layers

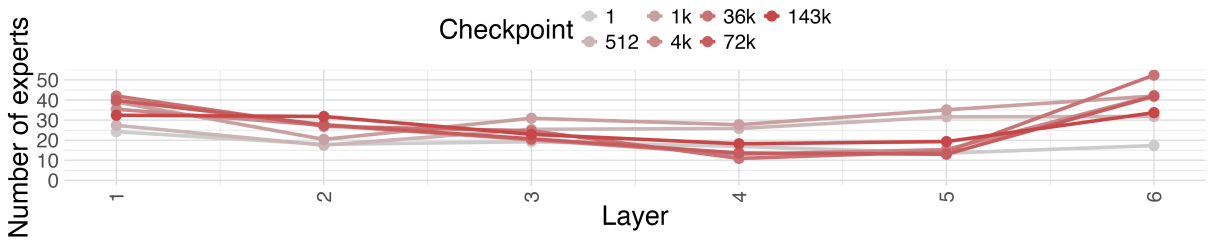

Figure 38: **Pythia 70m.** Average number of experts identified for **broader concepts** in MLP layers at different depths, for different checkpoints.

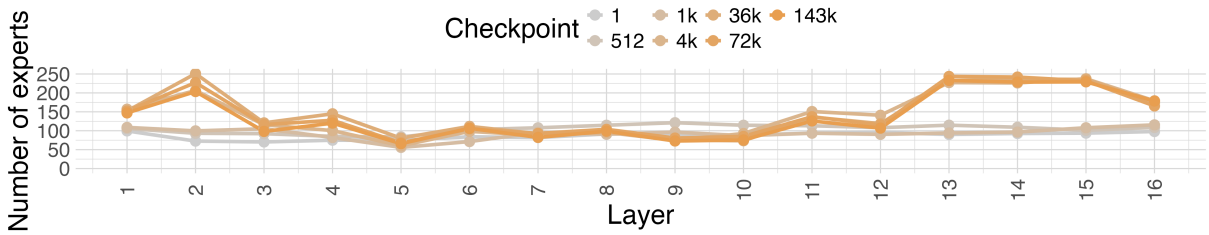

Figure 39: **Pythia 1b.** Average number of experts identified for **broader concepts** in MLP layers at different depths, for different checkpoints.

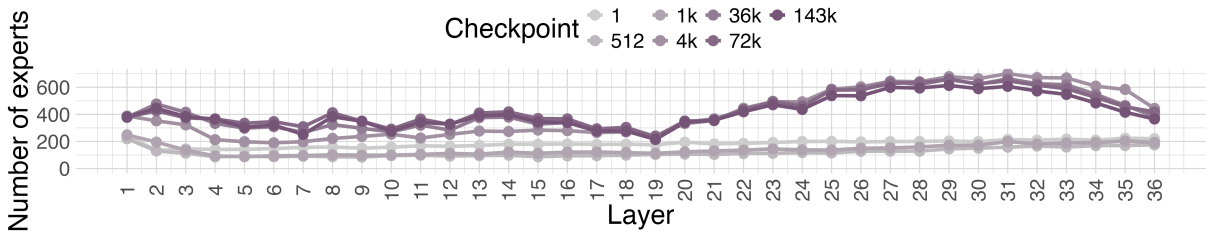

Figure 40: **Pythia 12b.** Average number of experts identified for **broader concepts** in MLP layers at different depths, for different checkpoints.

I.2 Attention layers

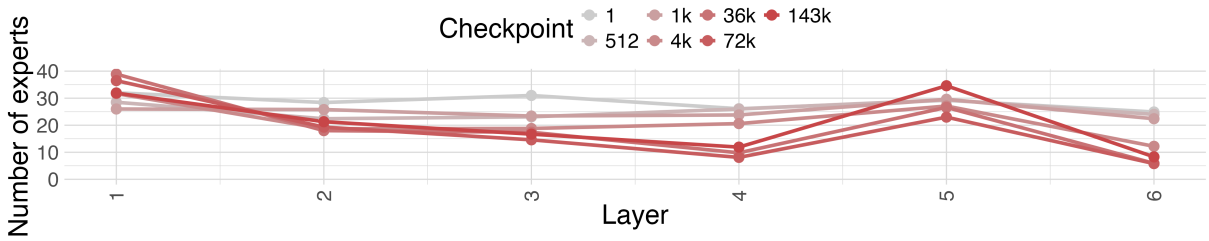

Figure 41: **Pythia 70m.** Average number of experts identified for **broader concepts** in attention layers at different depths, for different checkpoints.

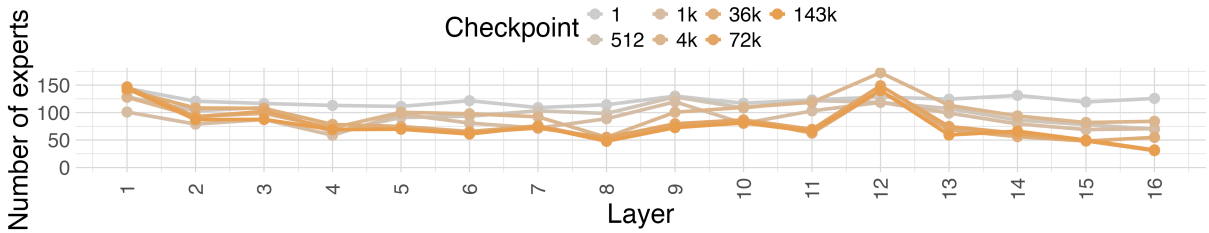

Figure 42: **Pythia 1b.** Average number of experts identified for **broader concepts** in attention layers at different depths, for different checkpoints.

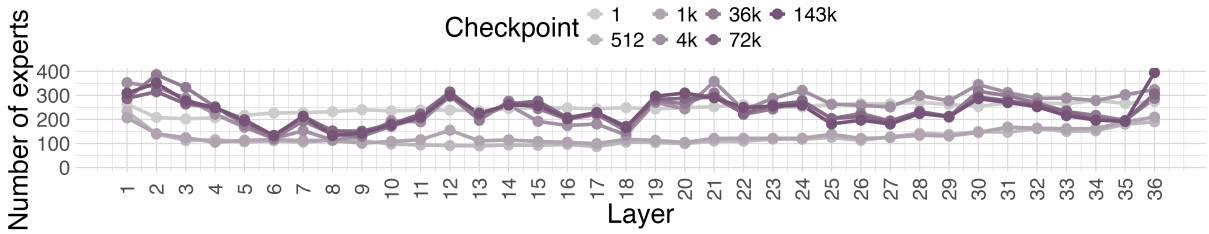

Figure 43: **Pythia 12b.** Average number of experts identified for **broader concepts** in attention layers at different depths, for different checkpoints.

J Distribution of experts for narrower concepts

J.1 MLP layers

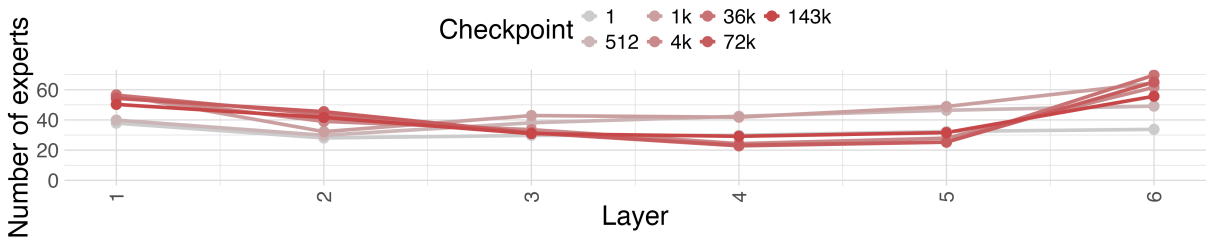

Figure 44: **Pythia 70m.** Average number of experts identified for **narrower concepts** in MLP layers at different depths, for different checkpoints.

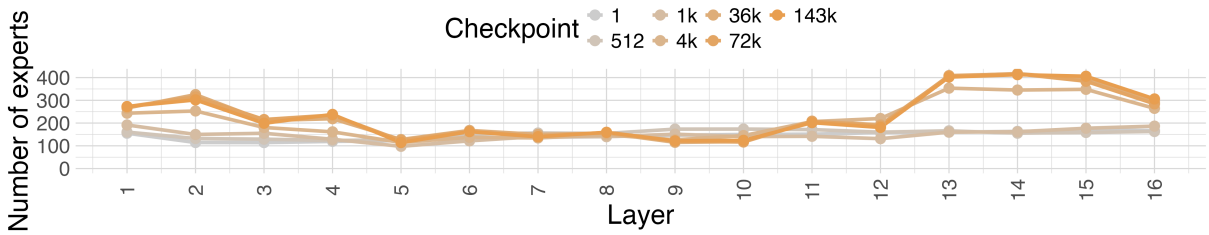

Figure 45: **Pythia 1b.** Average number of experts identified for **narrower concepts** in MLP layers at different depths, for different checkpoints.

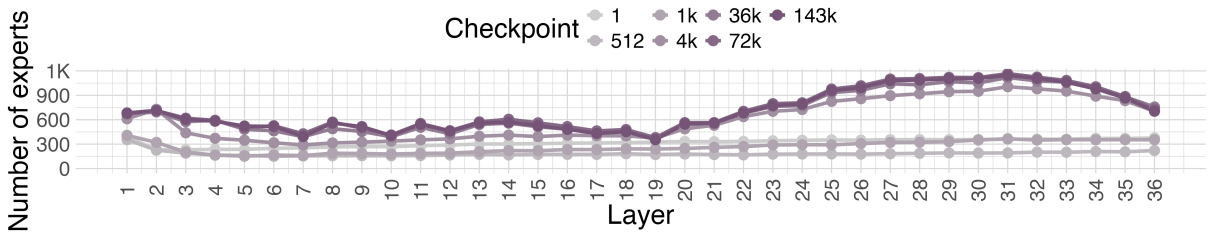

Figure 46: **Pythia 12b.** Average number of experts identified for **narrower concepts** in MLP layers at different depths, for different checkpoints.

J.2 Attention layers

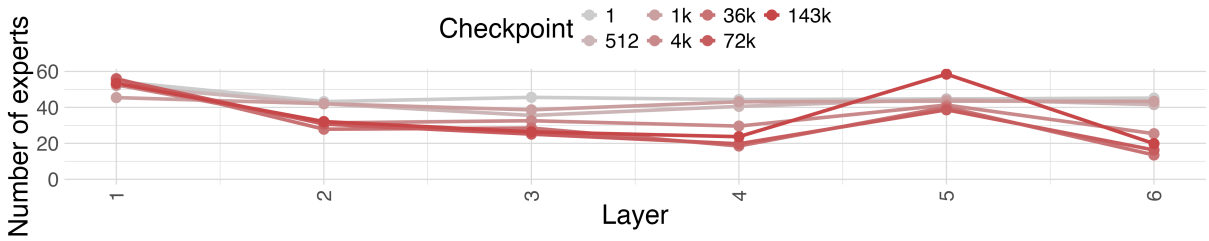

Figure 47: **Pythia 70m.** Average number of experts identified for **narrower concepts** in attention layers at different depths, for different checkpoints.

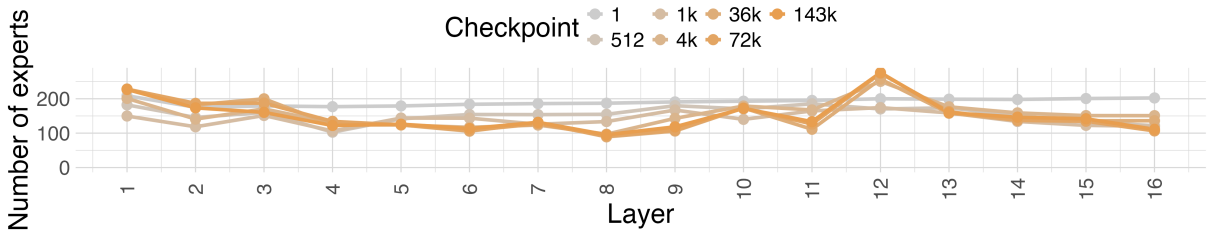

Figure 48: **Pythia 1b.** Average number of experts identified for **narrower concepts** in attention layers at different depths, for different checkpoints.

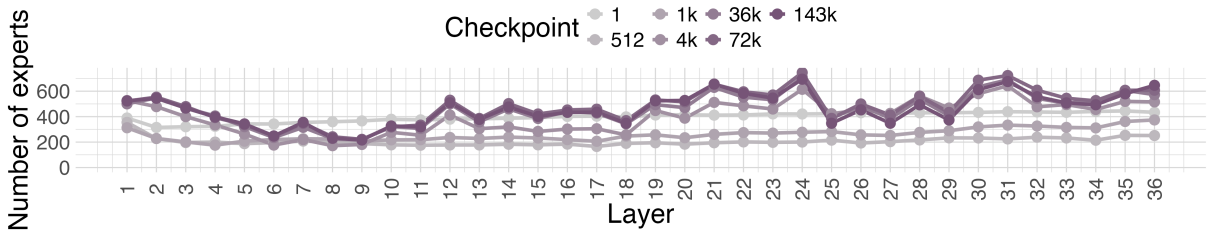

Figure 49: **Pythia 12b.** Average number of experts identified for **narrower concepts** in attention layers at different depths, for different checkpoints.

K Computational budget

The concept dataset was parallelized over 8 A100 GPUs (80GB). Expert extraction took about 136 seconds per concept for the 12b Pythia model; about 27 seconds per concept for the 1b Pythia model; about 8 seconds per concept for the 70m Pythia model; and about 25 seconds per concept for GPT-2.

L License and Attribution

The MEN dataset used in this work is released under Creative Commons Attribute license. The pretrained models are supported by public licenses the Pythia Scaling Suite (Apache), Mistral (Apache), and GPT-2 (MIT). GPT-4 is supported a proprietary license. We use an internal 80b-chat model and are unable to provide license information on it at this time.