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 the study of 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. Our findings reveal that LLM representations closely align with human representations inferred from behavioral data. Notably, this alignment surpasses that of word embeddings, which have been center stage in prior work on human and model alignment. Additionally, our approach enables a more granular view of how LLMs represent concepts. Specifically, we show that LLMs organize concepts in a way that reflects hierarchical relationships interpretable to humans (e.g., "animal"-"dog").

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 — for instance, 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 the conceptual representations in humans as safe and trustworthy deployment of LLMs requires such alignment. Overall, unveiling aspects of representation alignment and understanding how to foster it can help us identify and mitigate misaligned LLM behaviors, thus increasing trust in and safety of models (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 LLM-derived 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 these approaches have significantly advanced our understanding of how conceptual representations align between humans and models, they 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" are processed with similar neural pathways or architectural components, limiting our ability to understand the existence of structures such as hierarchical relationships in the model.

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 of a particular concept, so-called *expert neurons*. 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 hierarchy interpretable to humans (e.g., "dog", "cat", and "cheetah" being categorized as "animal"). We also track how alignment evolves during training for different model sizes, shedding light on the impact of model capacity on the de-

velopment of aligned representations — an aspect largely overlooked in previous work on text-based models (Shen et al., 2024; Wei et al., 2022). Ultimately, understanding these internal structures and 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 then evaluate the alignment between human and LLM representations by testing whether the similarity between neural activations correlates with human-perceived concept similarity, and whether the LLMs learn hierarchical structures similar to those observed in human category systems (Rosch, 1978). Finally, we identify the location of the model's concept representations and the point they form during training.

Our results show that LLM representations are generally aligned with humans. Crucially, expert neurons capture human alignment significantly better than the single word embeddings used in prior work. Moreover, such alignment emerges early in training, with model size playing only a small role: a 70m LLM is less aligned than a 1b or 12b LLM trained on the same data, but there is no difference between the larger models. Finally, patterns in expert neurons reveal that the LLMs show a human-like hierarchical organization of concepts.

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). 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, while Digutsch and Kosinski (2023) report similar alignment with GPT-3's (Brown et al., 2020) embeddings. However, Shaki et al. (2023) finds that GPT-3's concept alignment is highly sensitive to prompt phrasing and Misra et al. (2020) show that alignment in BERT (Devlin et al., 2019) is very context-dependent. Investigating general factors that can cause mis-alignment, Muttenthaler et al. (2023) conclude that the training dataset and objective function impact alignment, but model scale and architecture have no significant effect. Of note, alignment and performance are not inherently tied: mis-aligned models can exhibit significant capabilities (Sucholutsky and Griffiths, 2023; Dessì et al., 2022).

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 (one positive and one negative) 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.

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. The 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; 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.

To identify experts neurons for a given concept, each neuron is evaluated in isolation as a binary classifier: a neuron is considered an expert if its activations effectively distinguish between input data where the concept is present (henceforth *positive set*) and input data where the concept is absent (henceforth *negative set*). 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). 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 τ values the greater a neuron's expertise for a given concept. In our experiments, we consider a range of values for $\tau \in [0.5, 0.9]$ ranging 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 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.

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").

To study whether the LLMs represent concepts hierarchically (Sec. 5), we manually generate lists of ten domains, organized in human-interpretable hierarchies 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. C). We choose not to use WordNet (Miller, 1994) — a lexical database of English annotated with a hierarchical structure — because of drawbacks identified in its hierarchical structure, which often make the hierarchical relationships it presents unintuitive (for a discussion, see Gangemi et al., 2001).

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.

3.3 Models

We use GPT-2 (Radford et al., 2019) to select hyperparameters (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. The mean accuracy and standard error across eight benchmarks (Table 1) 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 1: Pythia evaluation 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 and thus allow 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. For example, 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 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?

Having determined the appropriate hyperparameters to capture a stable set of experts, we turn to the first 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 a 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. B, we consider cosine similarity between the raw AP values as an LLM similarity score, finding very 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).

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 respectively. For reference, agreement between humans has a Spearman correlation of 0.84. Interestingly, model size has a small impact on this alignment (in line with findings 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. The reason for this is that smaller models have fewer experts compared to larger models (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

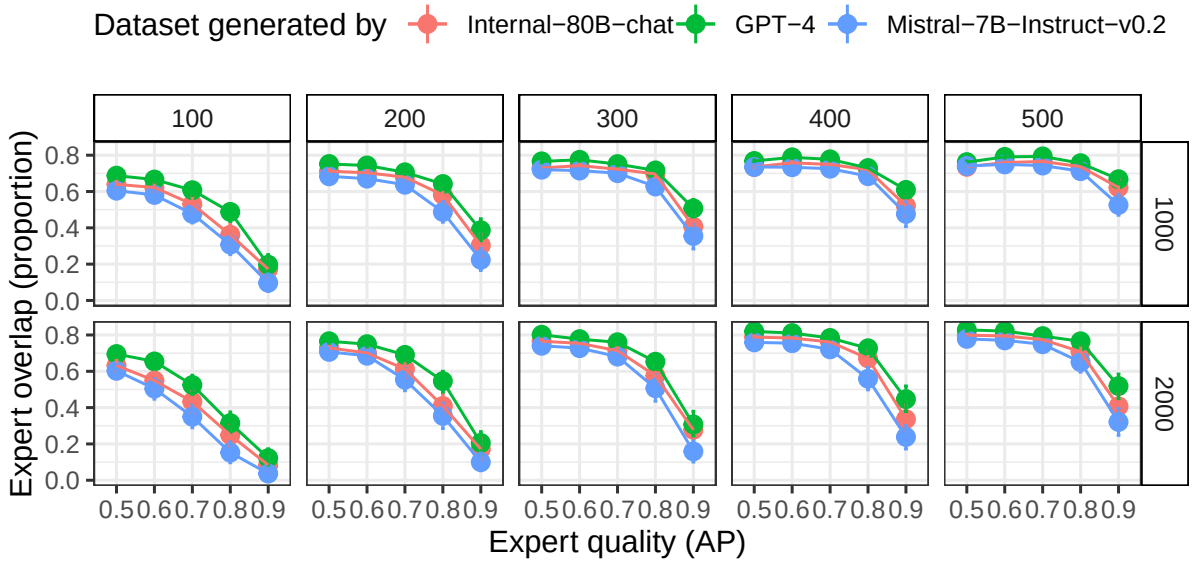

Figure 1: Expert discovery is relatively stable across various dataset characteristics. Points represent condition means; error bars represent bootstrapped 95% confidence intervals. Columns represent the size of the positive set (number of unique sentences); rows represent the size of the negative set (number of unique sentences).

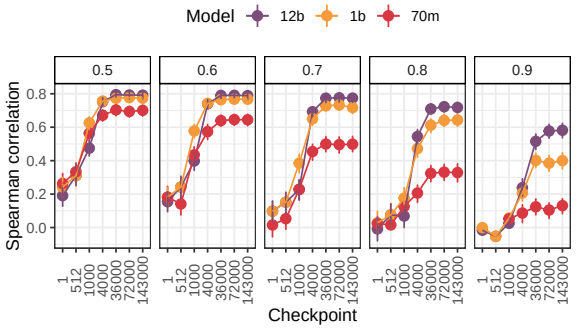

Figure 2: Model representations of similarity are closely aligned with human ones. Points represent Spearman correlations between the expert neuron overlap and perceived human similarity in the MEN dataset; error bars represent bootstrapped 95% confidence intervals. The subplots are τ. The correlations are statistically ($p<0.05$) at all checkpoints except for the first one.

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 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 corresponding human similarity judgement (Digutsch and Kosinski, 2023). We find statistically significant correlations ($p<0.05$) between the cosine similarity of the embeddings for a given concept pair and their similarity in the MEN dataset, at all checkpoints except for the first (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 single word embeddings compared to the experts neurons (highest correlations are 0.25 vs. 0.79 for the embeddings vs. experts). Single word embeddings exhibit more variability in alignment, as indicated by larger confidence intervals within each checkpoint, and their pattern of alignment is less stable across checkpoints compared to that of the experts.

6 Do models organize concepts in hierarchies?

Some domains within the human conceptual system are organized in hierarchies, where broader categories include more specific categories. For example, the concept "dog" falls under "animal", meaning that all dogs are animals (Graf et al., 2016; Murphy, 2004; Rosch, 1978). This raises the question of whether models organize concepts in a human-like hierarchy. We propose that, if the model organizes domains in a hierarchical fashion interpretable to humans, concepts from related sub-categories should share a set of experts (e.g., "dog", "cat", "horse", and "cheetah" under the con-

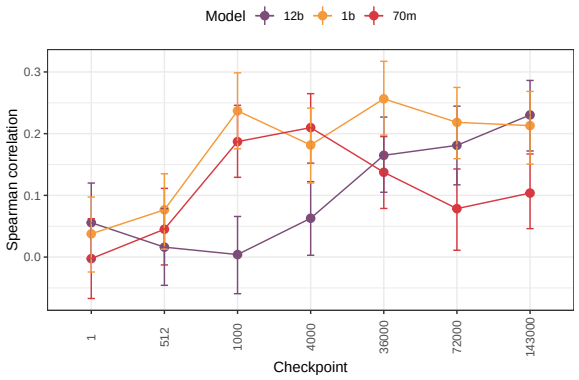

Figure 3: Spearman correlations between the cosine similarity in the embeddings and perceived human similarity in the MEN test split. Error bars represent bootstrapped 95% confidence intervals. The correlations are statistically (p<0.05) at all checkpoints except for the first one.

Domain	% overlap	baseline 1	baseline 2
animal	42.46	0.58	8.31
clothes	43.07	6.67	7.84
colour	54.35	4.83	2.66
furniture	69.98	9.20	5.48
occupation	32.31	2.56	5.68
organ	57.95	7.40	8.21
sport	75.76	9.24	12.67
subject	65.77	5.44	6.27
vegetable	69.25	6.09	10.92
vehicle	73.61	6.12	14.61

Table 2: Pythia 12b. Results of our exploration of hierarchically-organized domains at step 143k. For each domain, we report the percentage of experts shared among the sub-ordinate concepts that is also shared with the super-ordinate one. Numbers in gray correspond to our baselines. The percentage of experts shared in real domains is significantly higher than in the baselines as assessed via a two-sample permutation test (Virtanen et al., 2020, p-values < 0.001).

cept "animal"). Additionally, some of these shared experts should also be associated with the broader concept ("animal" in our example), suggesting that the model recognizes it as an overarching concept that includes its sub-categories.

To assess this, we consider the list of hierarchically organized domains we generated (see Sec. 3.2 and App. C), the experts associated with each concept in the list (τ=0.5), and their reciprocal overlap. We discuss the final training checkpoint of Pythia 12b in the main text and present other model sizes and checkpoints in App. D.

Let $super$ be a super-ordinate concept (e.g.,"animal") and let $sub_1, sub_2, ...sub_i$ be sub-ordinate concepts falling under it (e.g., "dog", "cat", "horse", "cheetah"). We call $E(c)$ the set of experts specialized for a concept c. For each domain, Table 2 reports the percentage of experts shared among the sub-ordinate concepts that are also shared with the super-ordinate concept:

$$\frac{|E(super) \cap \bigcap_{i=1}^{4} E(sub_i)|}{\left|\bigcap_{i=1}^{4} E(sub_i)\right|} \times 100.$$

For example, Table 2 indicates that 42.46% of the experts shared by "dog", "cat", "horse", and "cheetah" are also shared with "animal".

Our results confirm that the model captures hierarchical domain structures that characterize human conceptual systems. Within each domain, a large portion of experts for sub-ordinate concepts is shared with the super-ordinate concept. This suggests that the model recognizes the sub-ordinate concepts as part of the super-ordinate concept (e.g., that dogs are animals).

To rule out the possibility that expert sets overlap by chance, we compare the values we have obtained against two baselines (see Table 2, numbers in gray). In the first baseline, we randomize the super-ordinate concepts in our dataset by assigning each of them to N randomly selected sub-ordinates (e.g., associating "animal" with a random list of concepts like "jacket", "liver", "doctor", and "red"). In the second baseline, we shuffle the associations across concept categories, assigning each super-ordinate concept to a random set of internally related sub-ordinates (e.g., associating "animal" with "sock", "shirt", "jeans", and "jacket").[1] The results show that, when the domain structure is randomized, no hierarchical pattern emerges, reinforcing the robustness of our findings.

Interestingly, the patterns observed for the largest model do not largely differ from those found for smaller models. Additionally, the model seems to converge on a stable hierarchical representation around checkpoint 4k (App. D). This finding aligns with our later analyses, highlighting this checkpoint as a crucial transition point during training.

Having examined the overall structure of hierar-

[1]Of note, the randomization procedure does not prevent us from sampling, in some of the iterations, correct super-ordinate - sub-ordinate concept pairs (baseline 1; e.g., "animal" and "dog") or correct domain associations (baseline 2; e.g., "animal" with its sub-ordinates). Baseline values likely would be even lower if we enforced incorrect associations only.

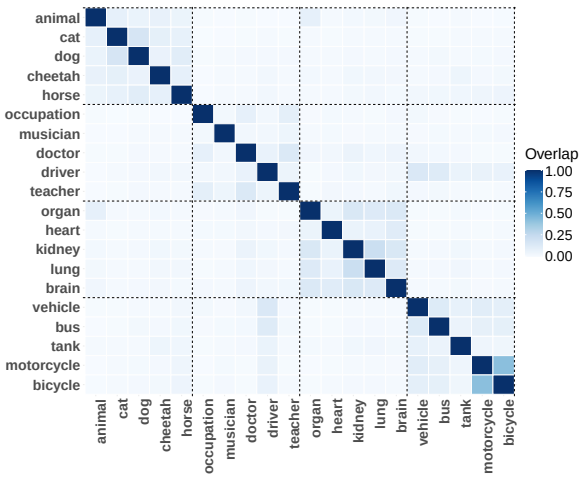

Figure 4: Proportion of expert overlap between word pairs belonging to hierarchically organized domains. Domains can be identified based on the stronger associations among their concepts.

chically organized domains by analyzing the relationship between groups of sub-ordinate concepts and their super-ordinate counterpart, we now shift our focus to concept pairs. We explore how super-ordinate concepts relate to individual sub-ordinate concepts, as well as the relationship between sub-ordinate concept pairs. We compute the overlap between the expert sets of each pair of concepts in our domains. A subset of the results is shown in Fig. 4 (see App. E for full results).

Overall, even if domain-specific differences are visible, concepts within the same domain, thus semantically related, tend to have a higher overlap in their expert sets. This is in line with the findings on representation alignment discussed in Sec. 5, showing that concepts perceived as similar by humans show high expert overlap and are also perceived as more similar by the LLMs. When exploring the internal organization of the domains, we notice that super-ordinate concepts have a weaker association with their sub-ordinate concepts compared to the association between pairs of sub-ordinate concepts. For instance, the super-ordinate concept "vehicle" is associated with all of its sub-ordinates, but these associations are weaker than those between "motorcycle" and "bicycle". The same happens in the "animal" domain for "dog" and "cat". This may suggest that individual sub-ordinate concept pairs may be distributionally more similar to each other than sub-ordinate - super-ordinate pairs.

Taking stock, our results show that the model captures human-interpretable hierarchical structures in concept representations, assigning some experts to be shared between sub-ordinate and corresponding super-ordinate concepts.

7 Characterizing model knowledge

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

Larger models have 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, fewer experts are identified and 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.

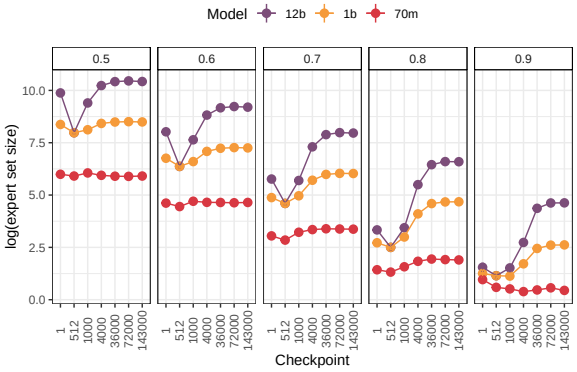

Figure 5: Expert set size (log) by model size and checkpoint. The points represent averages over all concepts. The error bars are bootstrapped 95% confidence intervals. Subplots correspond to different values of τ.

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

the expert set to stabilize, suggesting that higher-quality experts take longer to learn.

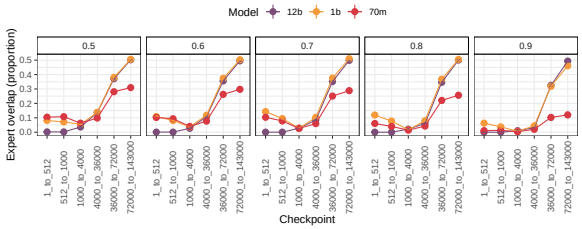

Figure 6: Proportion of expert overlap across subsequent checkpoints. Points represent across concept averages; error bars represent bootstrapped 95% confidence intervals. Subplots correspond to different values of τ.

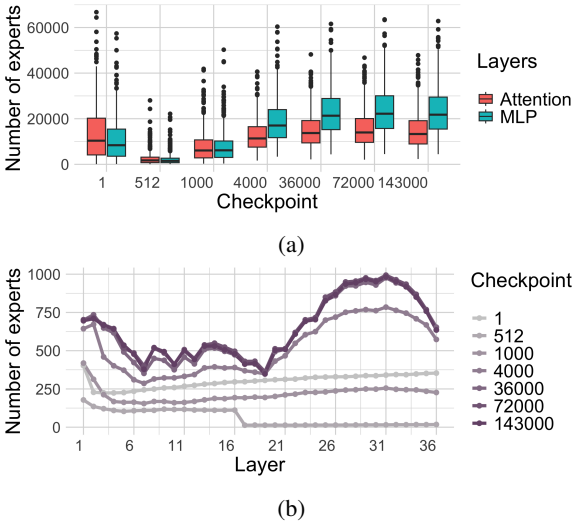

Figure 7: Pythia 12b. (a) Total number of experts in MLP and attention layers across checkpoints; more experts are located in MLPs; (b) Average number of experts identified in MLP layers at different depths, for different checkpoints.

More experts are found in MLPs and deeper layers Pythia models consist of intertwined self-attention and MLP layers (Biderman et al., 2023; Vaswani 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). Larger numbers of experts are located in the MLP layers compared to attention layers with the allocation of experts to different layer types stabilizing at checkpoint 4k. We see the same trend in smaller models (App. F.1) after controlling for the number of neurons in the respective layers. Moreover, 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. F.2). For attention layers, high numbers of experts are located in deep layers and, interestingly, in the first layer (see App. F.3). Of note, if we focus our analysis on highly specialized experts only (τ=0.9), we find higher numbers of experts in earlier layers (see App. F.8 and F.9), recovering the same patterns as 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).

8 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 the levels of alignment between humans. Consistent with prior work (Muttenthaler et al., 2023), we find that model size has little influence on alignment. Moreover, our approach reveals that models generally organize concepts into human-interpretable hierarchies. However, some domains are more structured than others, and this pattern remains consistent across model sizes. We leave it to future work to investigate the factors that could give rise to this pattern, such as the frequency of each domain in the training data.

9 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), 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 value alignment.

We do not study patterns in expert neurons through activating these neurons Since the approach we are using was designed for activation steering (Suau et al., 2023), one obvious application is to examine the intersections between the expert sets for two concepts through the lens of controllable generation. 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).

References

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

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*.

Roberto Dessì, Eleonora Gualdoni, Francesca Franzon, Gemma Boleda, and Marco Baroni. 2022. Communication breakdown: On the low mutual intelligibility between human and neural captioning. In *Conference on Empirical Methods in Natural Language Processing*.

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.

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.

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*.

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)*.

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.

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*.

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.

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.

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 and Thomas L. Griffiths. 2023. Alignment with human representations supports robust few-shot learning. *Preprint*, arXiv:2301.11990.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Pauli Virtanen, Ralf Gommers, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272.

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: 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: 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: Poppies are herbaceous plants that can grow annually, biennially, or perennially, depending on the specific species.
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 Analyses of correlations between human similarity judgments and cosine similarity for the full network

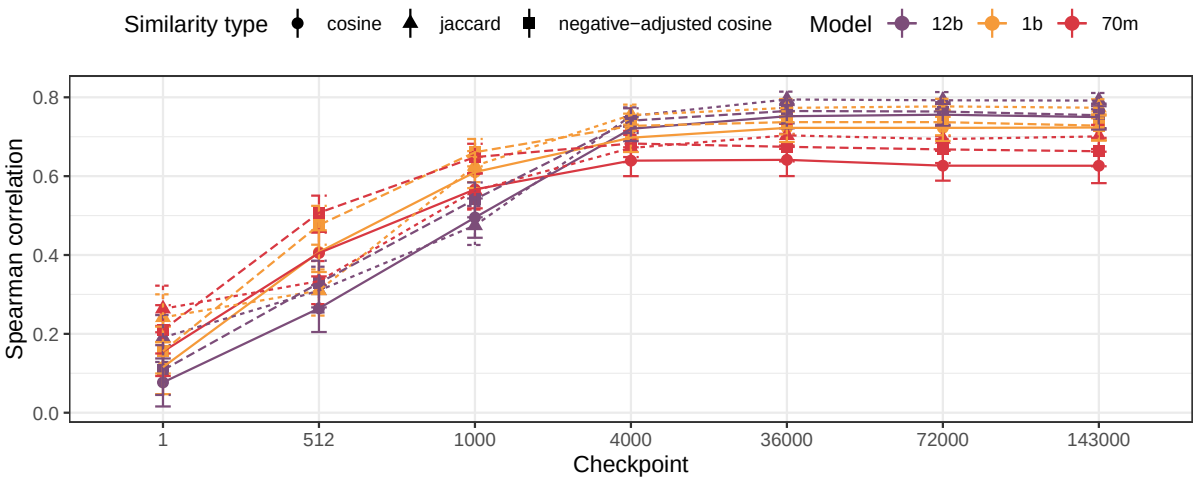

Figure 8: Spearman correlations between human similarity judgments, cosine similarity over raw AP values, negative-adjusted cosine similarity [abs(AP)-0.5], and the best-performing τ of Jaccard similarity (0.5). Points represent Spearman correlations between LLM's similarity and perceived human similarity in the MEN dataset; error bars represent bootstrapped 95% confidence intervals.

C List of words in hierarchically-organized domains

Super-ordinate	Sub-ordinates
animal	cat, dog, cheetah, horse
clothes	jacket, jeans, shirt, sock
colour	red, blue, green, black
furniture	chair, bookshelf, table, couch
occupation	doctor, teacher, driver, musician
organ	heart, kidney, lung, brain
sport	golf, racing, gymnastics, swimming
subject	mathematics, geography, biology, chemistry
vegetable	carrot, potato, pumpkin, corn
vehicle	bus, tank, motorcycle, bicycle

Table 3: List of words in our dataset of hierarchically-organized domains.

D Hierarchical structures: results for additional models and checkpoints

Domain	% overlap	baseline 1	baseline 2
animal	0.00	0.00	0.00
clothes	0.00	0.00	0.00
colour	0.00	0.00	0.00
furniture	0.00	0.00	0.00
occupation	0.00	0.00	0.00
organ	0.00	0.00	0.00
sport	0.00	0.00	0.00
subject	0.00	0.00	0.00
vegetable	0.00	0.00	0.00
vehicle	0.00	0.00	0.00

Table 4: **Pythia 70m.** Results of our exploration of hierarchically-organized domains, at step 1.

Domain	% overlap	baseline 1	baseline 2
clothes	0.00	0.00	0.00
colour	0.00	0.00	0.00
furniture	0.00	0.00	0.00
occupation	0.00	0.00	0.00
organ	66.67	0.00	3.81
sport	0.00	0.00	0.00
subject	57.14	0.00	5.56
vegetable	0.00	0.00	0.00
vehicle	0.00	0.00	0.00

Table 5: **Pythia 70m.** Results of our exploration of hierarchically-organized domains, at step 512.

Domain	% overlap	baseline 1	baseline 2
animal	0.00	0.00	1.00
clothes	20.00	0.00	2.38
colour	42.86	0.00	2.12
furniture	100.00	0.00	0.22
occupation	0.00	0.00	1.11
organ	16.67	1.11	1.67
sport	100.00	0.00	3.22
subject	81.25	0.00	0.00
vegetable	66.67	0.00	0.00
vehicle	75.00	0.00	0.83

Table 6: **Pythia 70m.** Results of our exploration of hierarchically-organized domains, at step 4000.

Domain	% overlap	baseline 1	baseline 2
animal	100.00	0.00	0.73
clothes	56.00	0.00	0.00
colour	52.00	0.00	2.30
furniture	75.00	0.00	2.80
occupation	14.29	0.00	5.34
organ	45.00	0.00	0.92
sport	85.71	0.00	1.03
subject	75.86	0.00	4.93
vegetable	68.00	0.00	0.00
vehicle	100.00	0.00	3.57

Table 7: **Pythia 70m.** Results of our exploration of hierarchically-organized domains, at step 143k.

Domain	% overlap	baseline 1	baseline 2
animal	0.00	0.00	0.00
clothes	0.00	0.00	0.00
colour	2.38	3.33	0.00
furniture	0.00	0.00	0.00
occupation	0.00	0.00	0.00
organ	0.00	0.00	0.00
sport	0.00	0.00	0.00
subject	0.00	0.00	0.32
vegetable	0.00	0.00	0.00
vehicle	0.00	0.00	0.24

Table 8: **Pythia 1b.** Results of our exploration of hierarchically-organized domains, at step 1.

Domain	% overlap	baseline 1	baseline 2
animal	0.00	0.00	0.26
clothes	0.00	3.33	0.00
colour	2.56	0.00	1.07
furniture	0.00	1.67	0.00
occupation	23.08	0.00	0.15
organ	32.94	0.37	3.65
sport	0.00	0.00	0.20
subject	37.40	0.00	1.98
vegetable	0.00	0.00	0.33
vehicle	0.00	0.00	0.29

Table 9: **Pythia 1b.** Results of our exploration of hierarchically-organized domains, at step 512.

Domain	% overlap	baseline 1	baseline 2
animal	84.06	0.00	0.98
clothes	45.78	3.33	1.08
colour	48.48	1.11	1.26
furniture	61.73	0.00	3.28
occupation	16.67	0.67	0.98
organ	45.15	4.24	2.12
sport	81.40	1.33	2.02
subject	58.26	1.67	1.68
vegetable	76.71	0.00	1.13
vehicle	82.80	4.17	1.87

Table 10: **Pythia 1b.** Results of our exploration of hierarchically-organized domains, at step 4000.

Domain	% overlap	baseline 1	baseline 2
animal	61.09	0.00	1.47
clothes	56.67	10.37	1.04
colour	64.71	5.00	1.66
furniture	71.96	0.00	2.40
occupation	15.38	1.41	1.53
organ	55.38	0.61	2.76
sport	74.27	3.89	2.41
subject	77.54	4.44	2.04
vegetable	77.89	0.56	1.07
vehicle	83.54	0.48	2.19

Table 11: **Pythia 1b.** Results of our exploration of hierarchically-organized domains, at step 143k.

Domain	% overlap	baseline 1	baseline 2
animal	0.00	0.00	0.00
clothes	0.00	0.18	0.20
colour	1.20	0.00	0.00
furniture	0.00	0.00	0.00
occupation	0.00	0.00	0.00
organ	0.00	0.00	0.00
sport	0.00	3.33	0.00
subject	0.00	0.00	0.48
vegetable	0.00	1.67	0.12
vehicle	0.00	0.00	8.45

Table 12: **Pythia 12b.** Results of our exploration of hierarchically-organized domains, at step 1.

Domain	% overlap	baseline 1	baseline 2
animal	0.00	0.00	0.08
clothes	0.00	0.00	0.00
colour	2.38	6.67	7.78
furniture	0.00	0.00	1.46
occupation	0.00	0.00	3.70
organ	14.28	0.00	0.00
sport	0.00	0.00	0.00
subject	33.33	0.00	0.63
vegetable	0.00	0.30	0.08
vehicle	0 0.00	0.00	0.00

Table 13: **Pythia 12b.** Results of our exploration of hierarchically-organized domains, at step 512.

Domain	% overlap	baseline 1	baseline 2
animal	62.81	0.30	1.05
clothes	46.76	13.94	1.42
colour	56.74	2.22	1.04
furniture	69.07	3.33	3.7
occupation	16.30	1.50	2.16
organ	44.88	4.24	2.75
sport	75.42	1.45	2.19
subject	68.39	9.44	1.84
vegetable	80.40	1.00	1.18
vehicle	79.76	2.16	2.10

Table 14: **Pythia 12b.** Results of our exploration of hierarchically-organized domains, at step 4000.

E Hierarchically-organized domains: additional plots

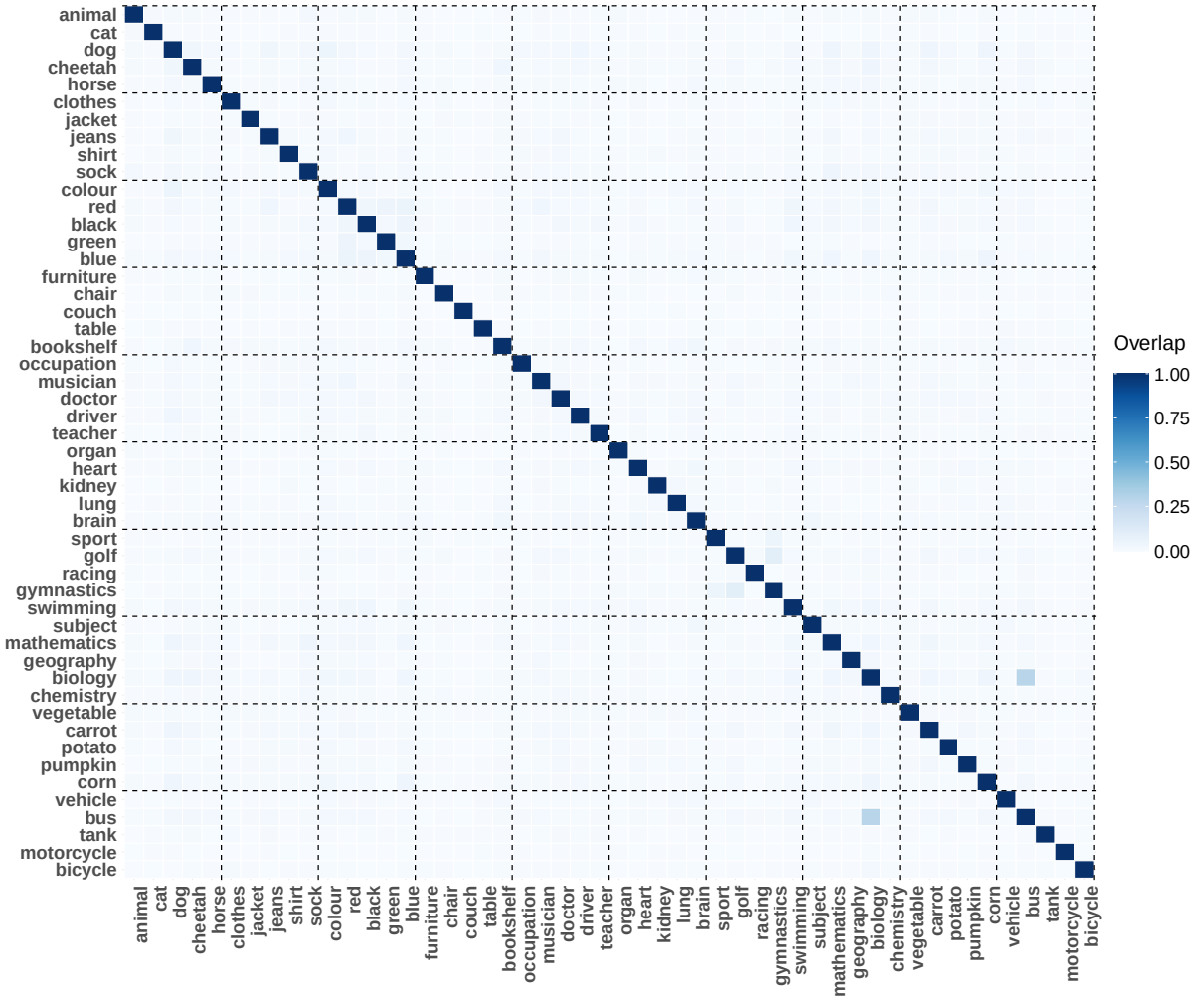

Figure 9: **Pythia 70m.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 1. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

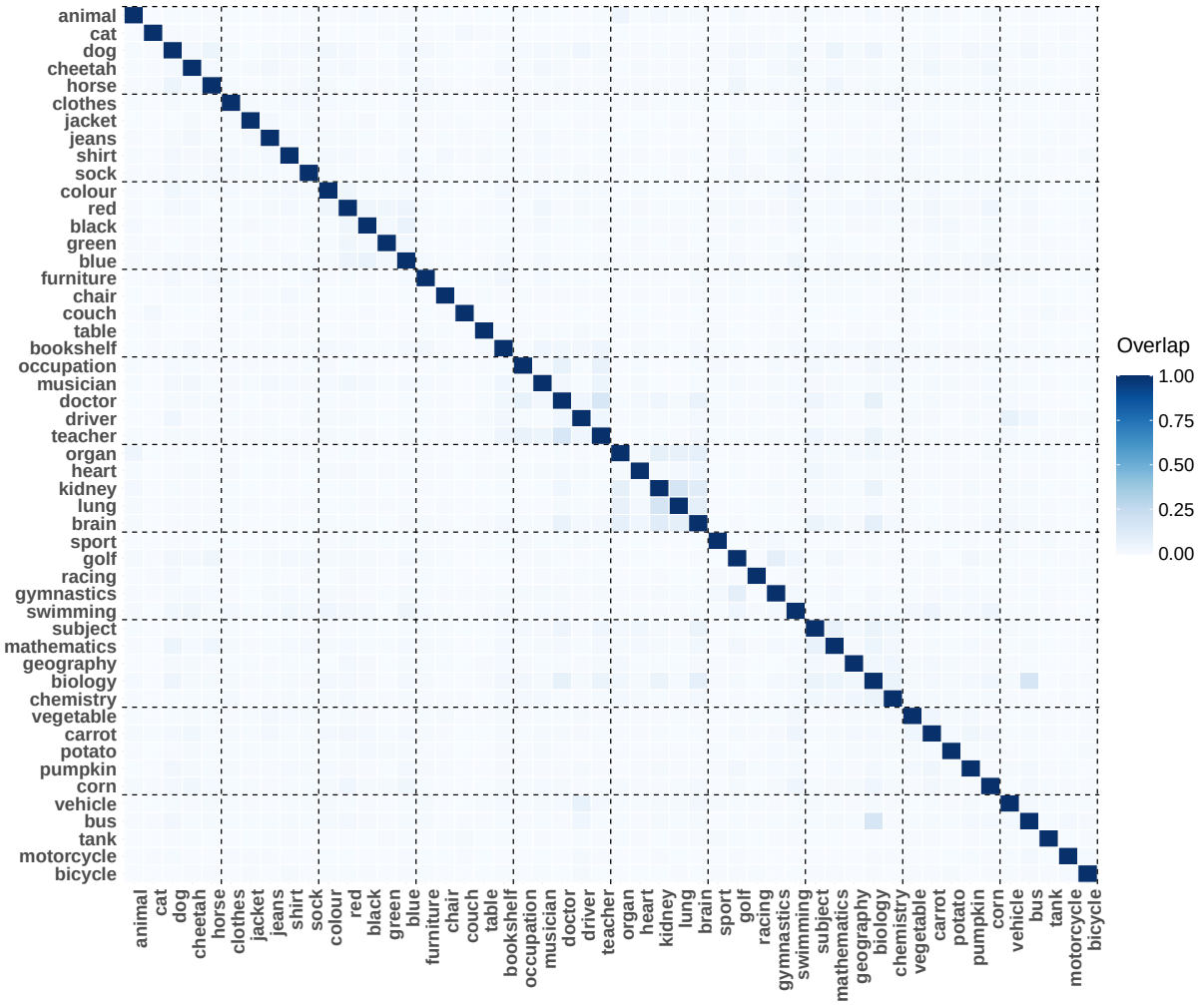

Figure 10: **Pythia 70m.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 512. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

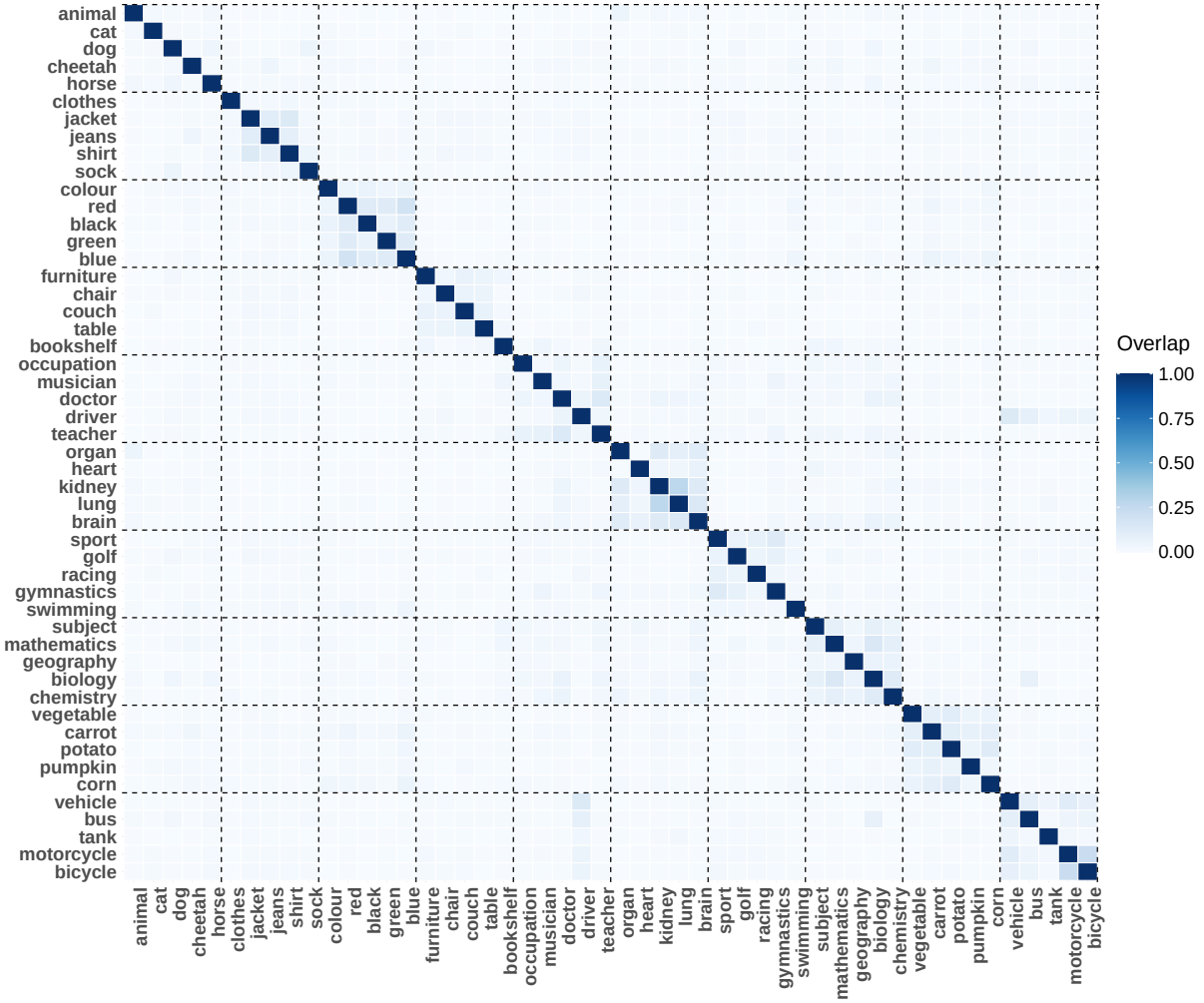

Figure 11: **Pythia 70m.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 4000. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

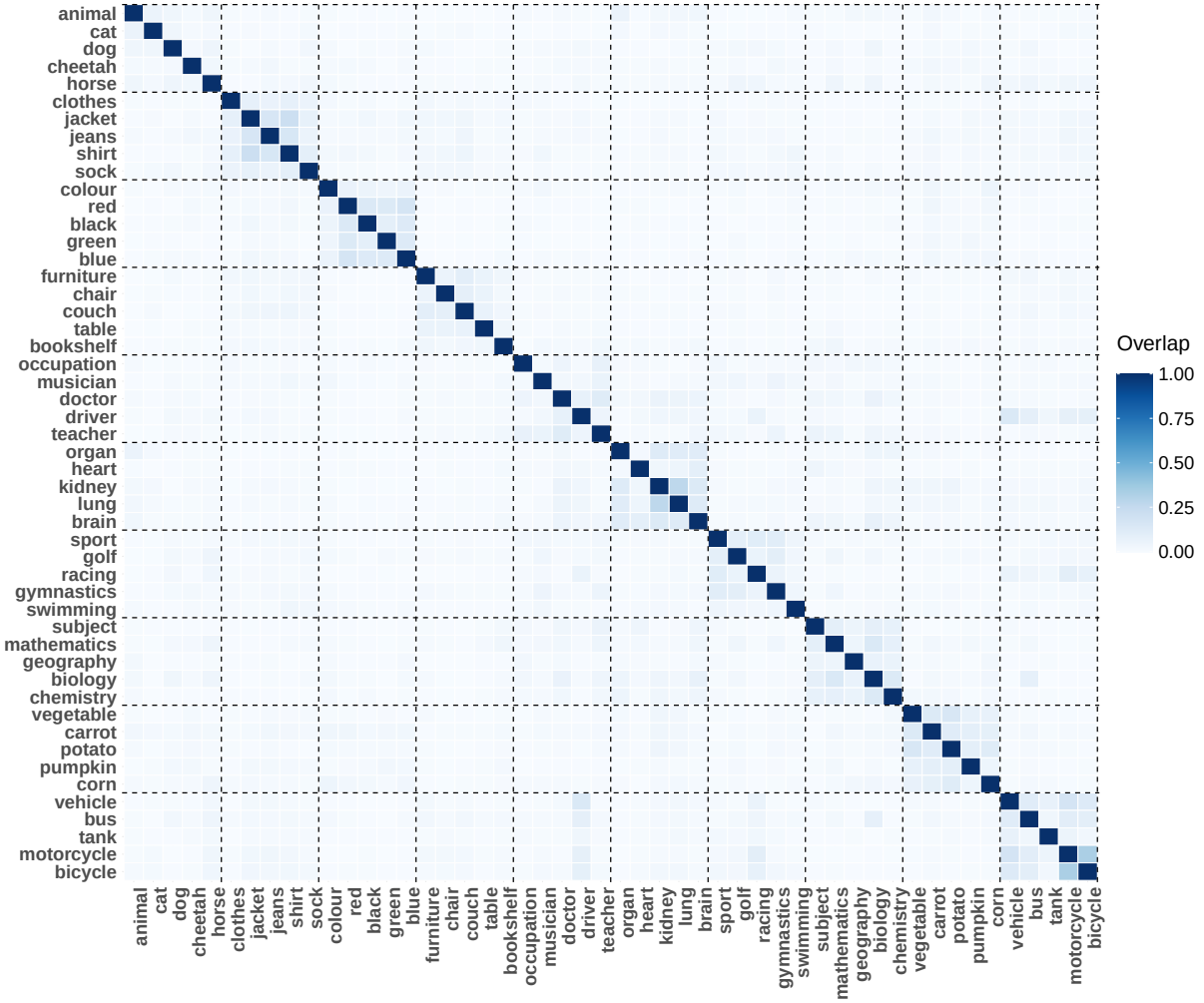

Figure 12: **Pythia 70m.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 143000. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

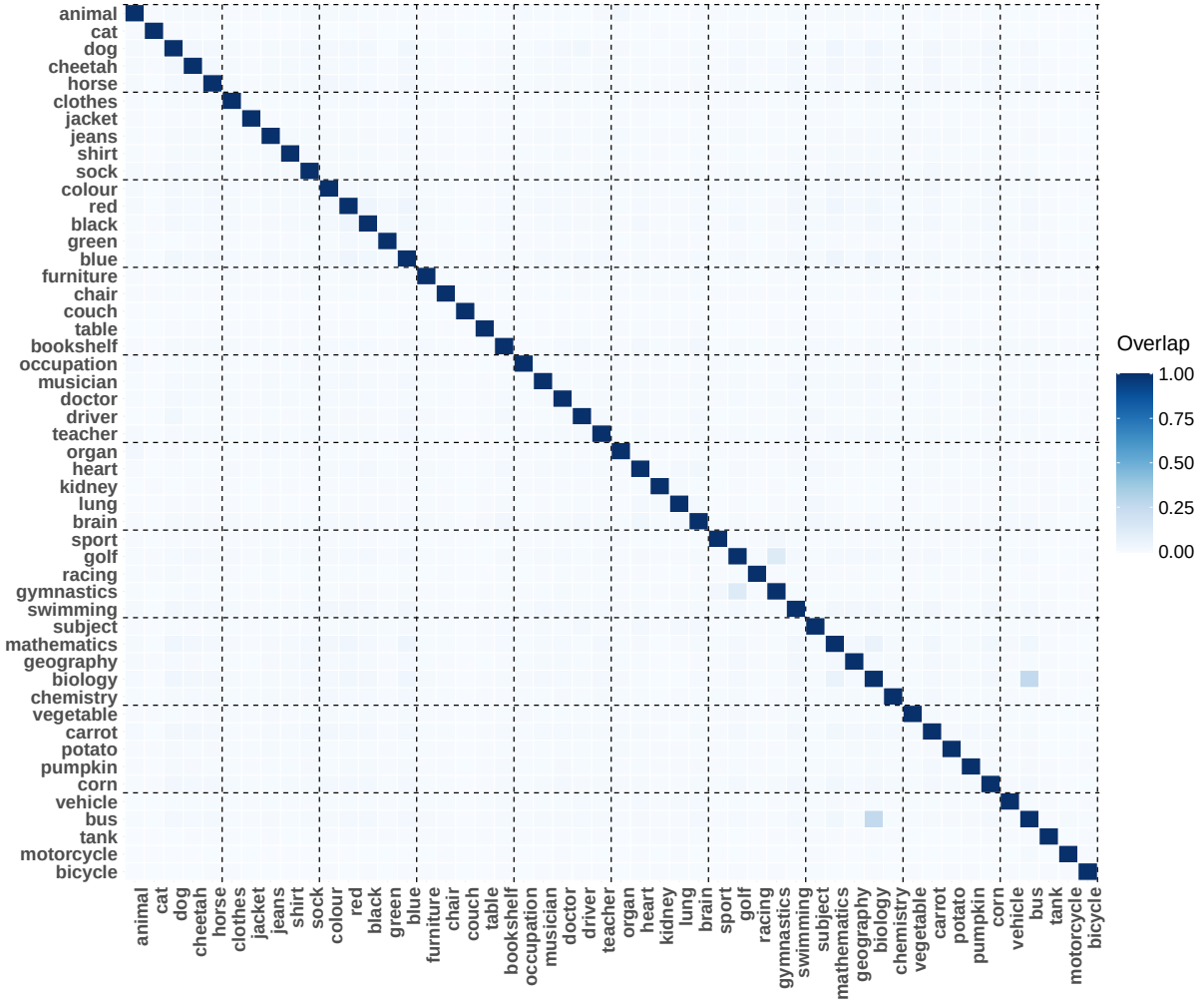

Figure 13: **Pythia 1b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 1. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

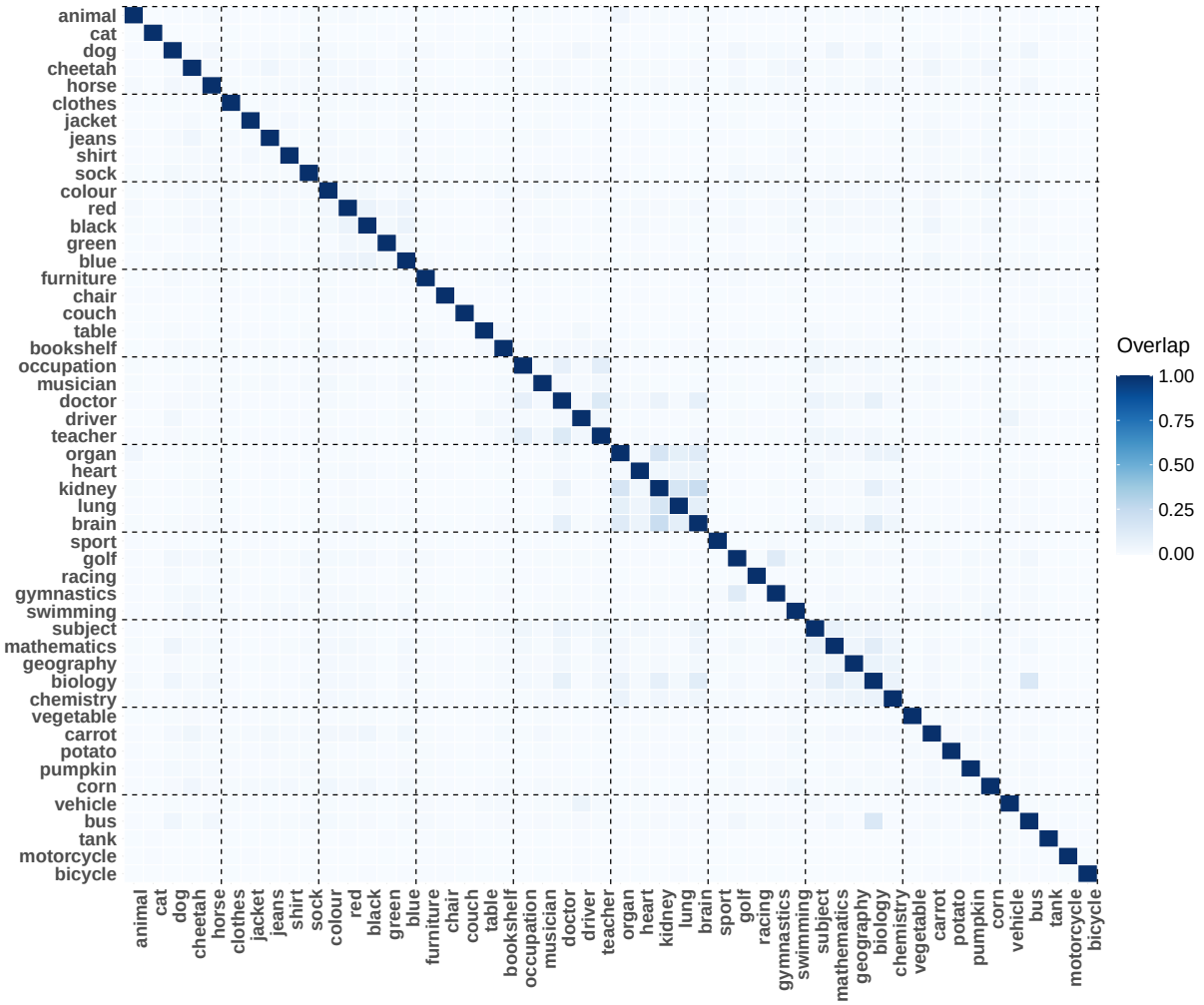

Figure 14: **Pythia 1b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 512. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

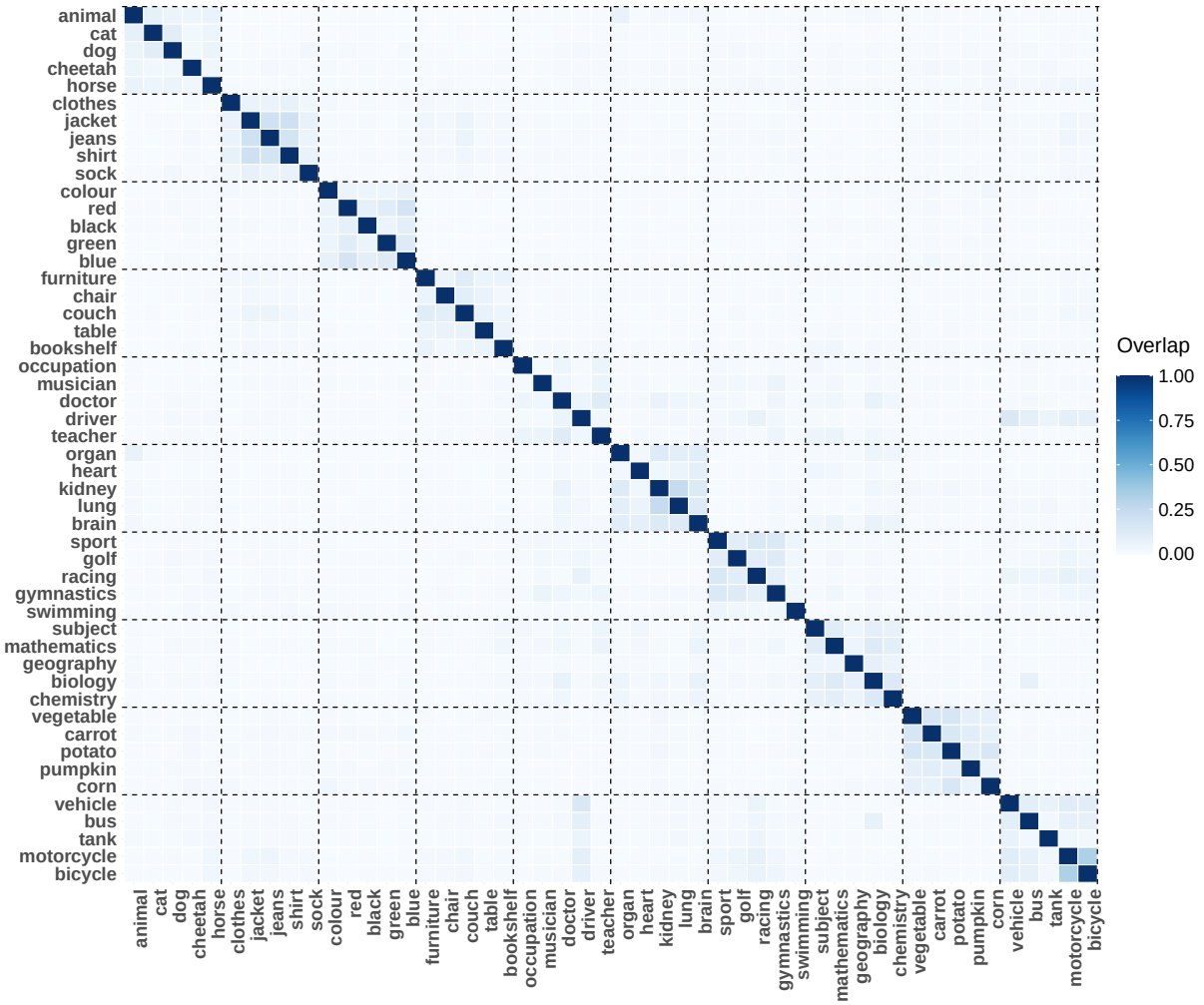

Figure 15: **Pythia 1b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 4000. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

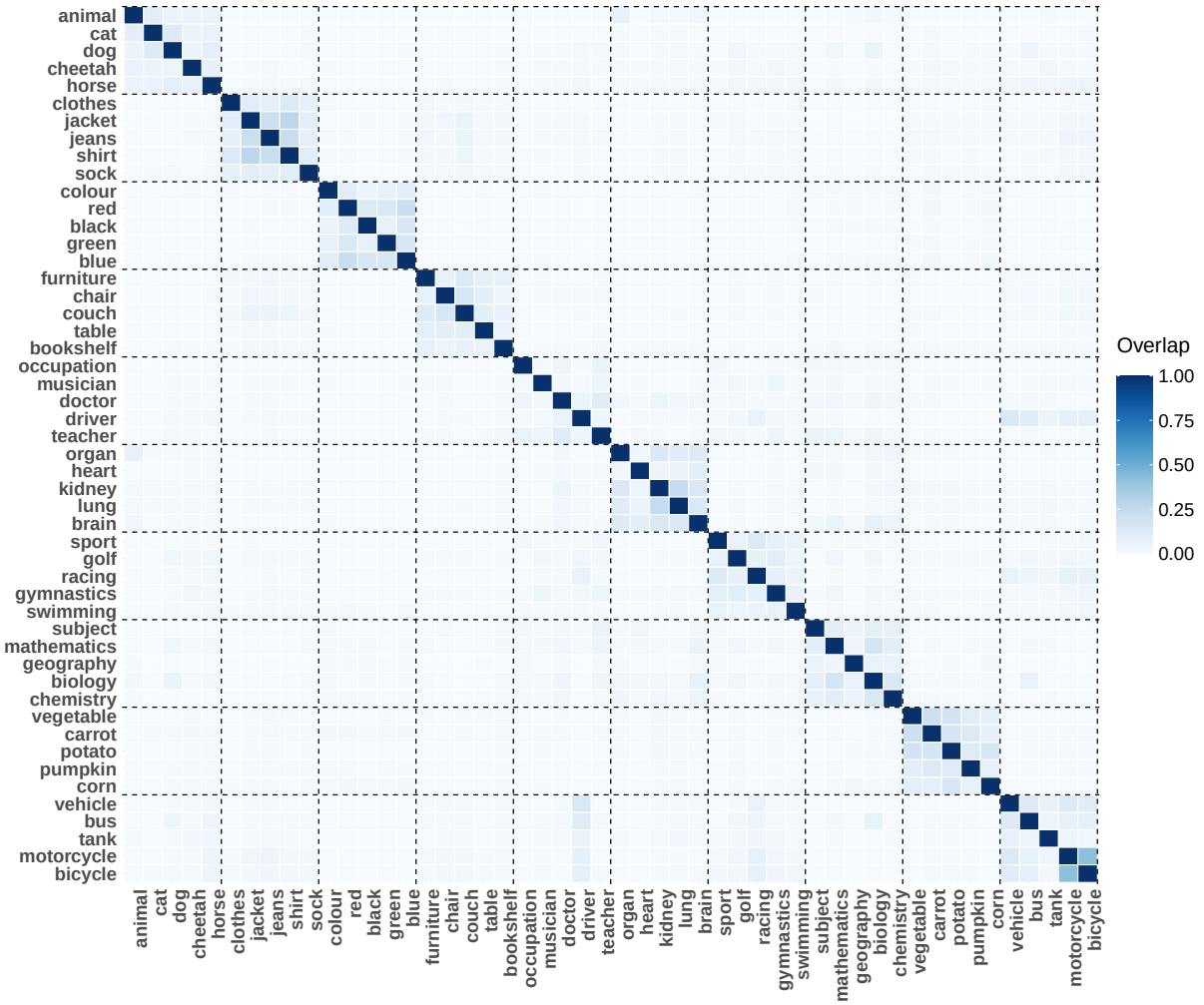

Figure 16: **Pythia 1b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 143000. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

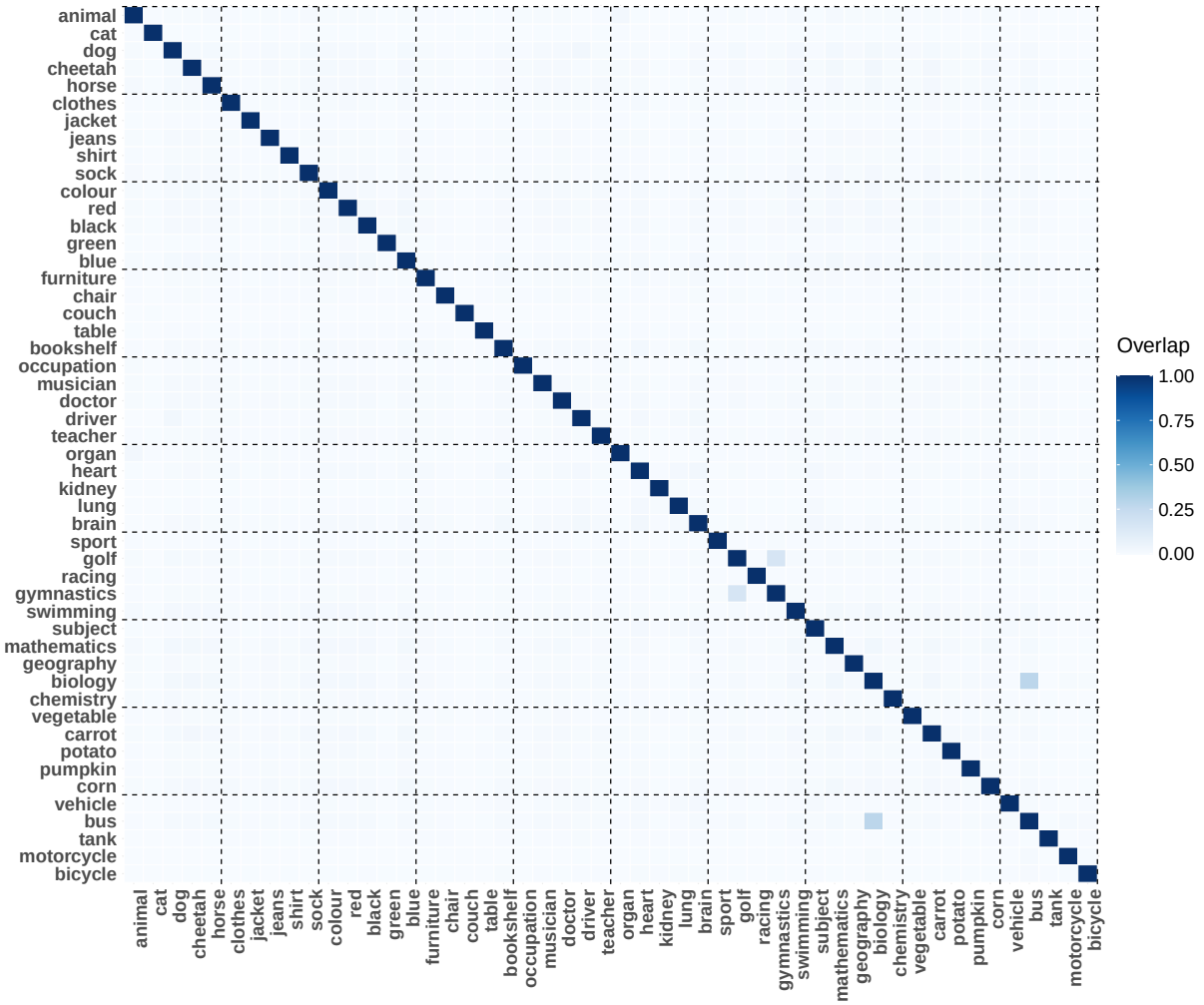

Figure 17: **Pythia 12b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 1. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

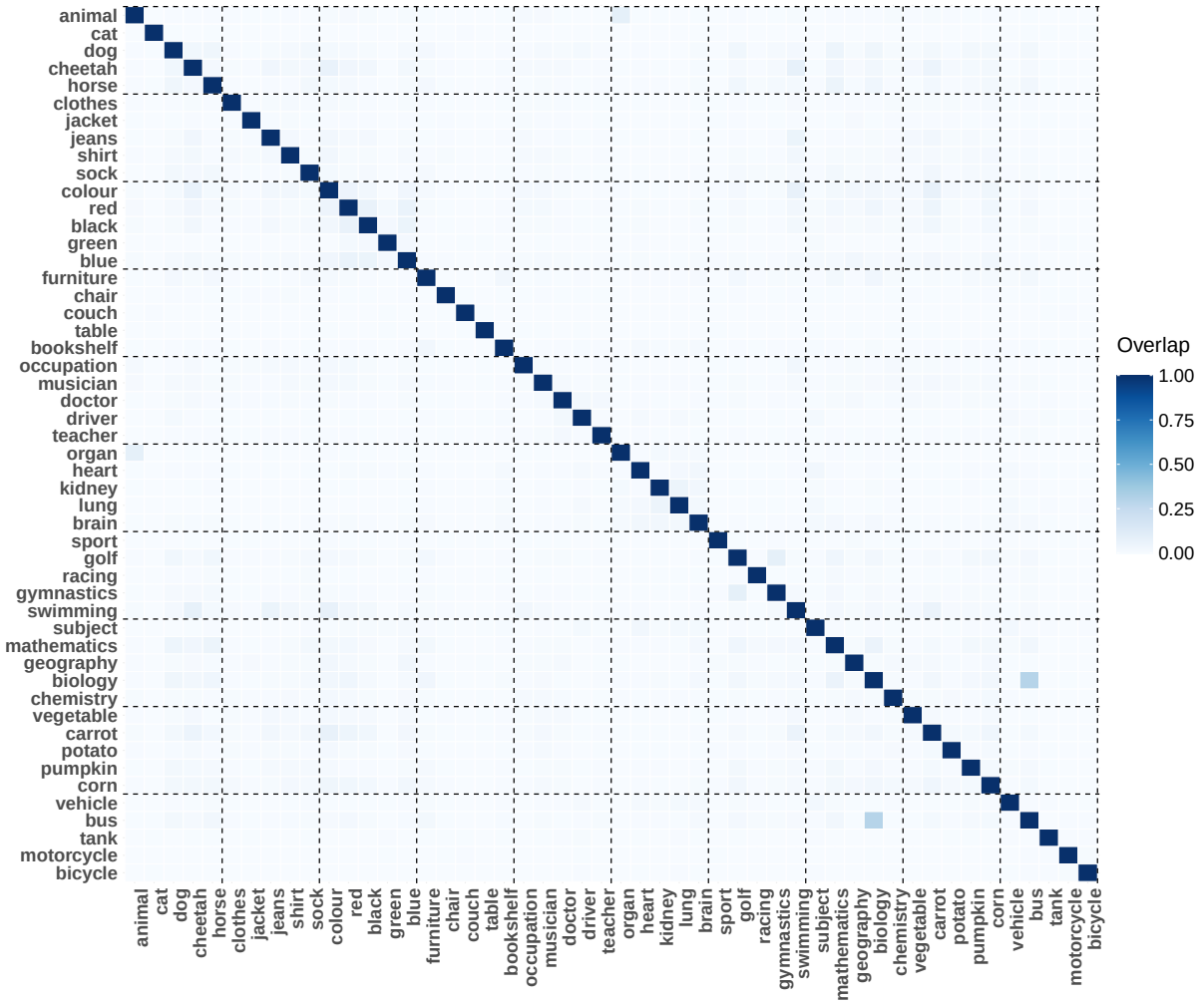

Figure 18: **Pythia 12b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 512. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

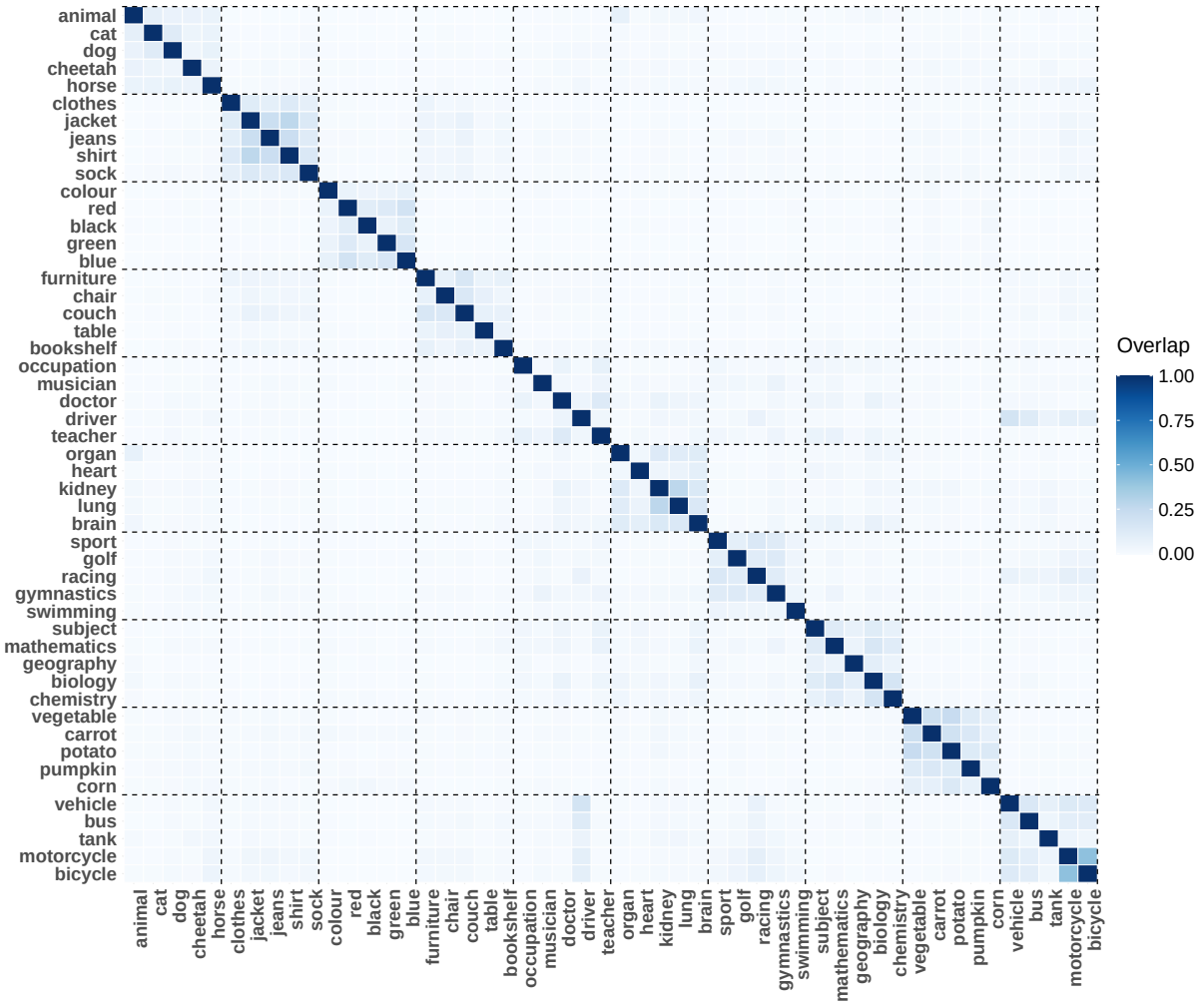

Figure 19: **Pythia 12b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 4000. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

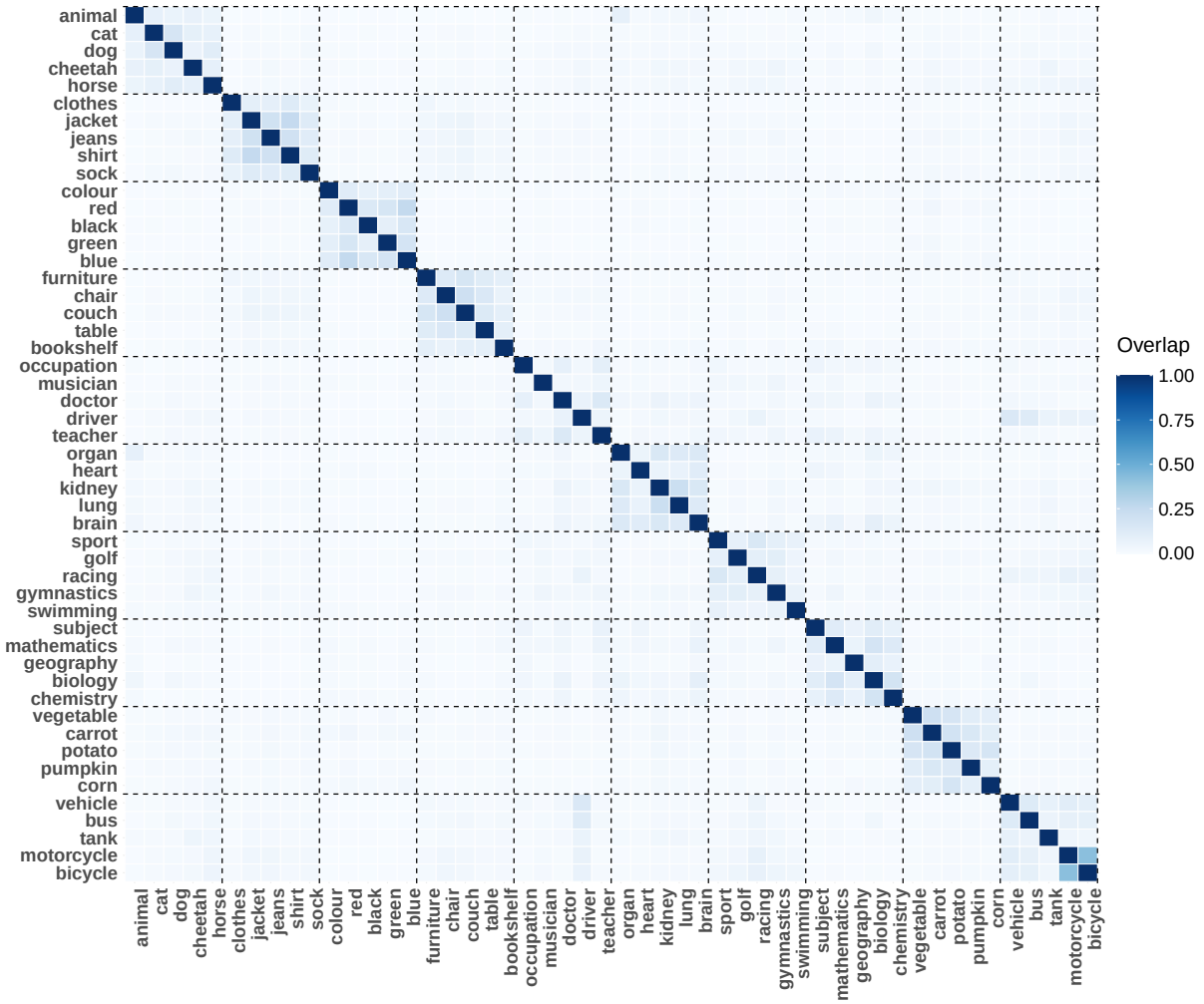

Figure 20: **Pythia 12b.** Proportions of expert overlap between word pairs belonging to hierarchically organized domains, at checkpoint 143000. Domains can be identified based on the stronger associations among their words compared to unrelated terms.

F Additional materials for layer analyses

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

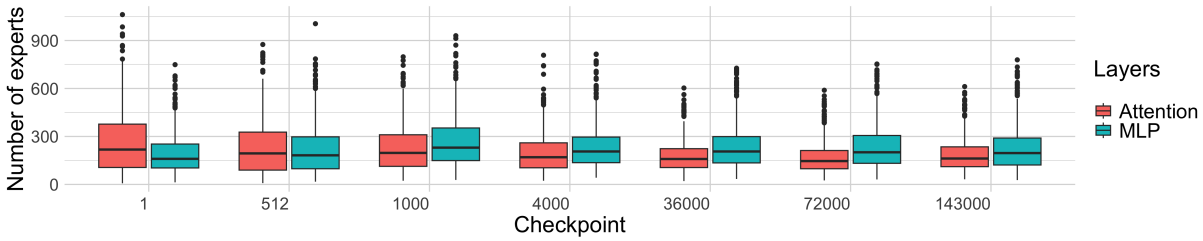

Figure 21: **Pythia 70m.** Total number of experts in MLP and attention layers across checkpoints

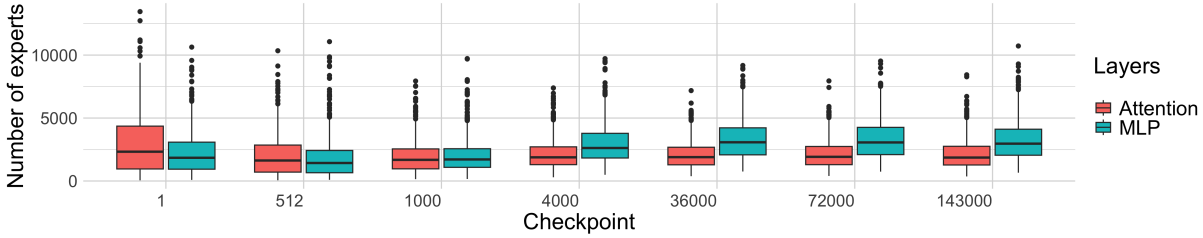

Figure 22: **Pythia 1b.** Total number of experts in MLP and attention layers across checkpoints

F.2 Distribution of experts across MLP layers

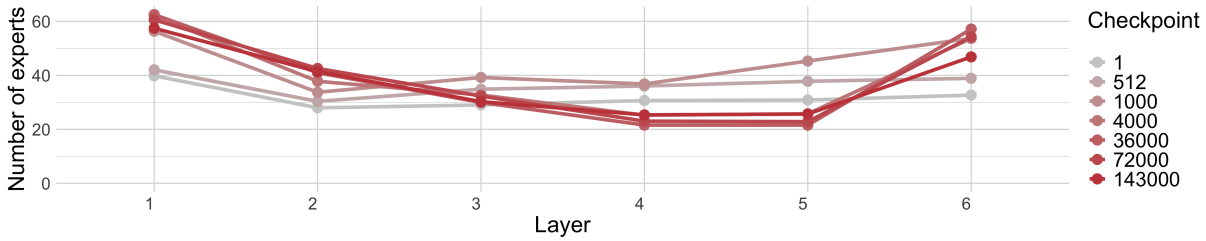

Figure 23: **Pythia 70m.** Average number of experts identified in MLP layers at different depths, for different checkpoints.

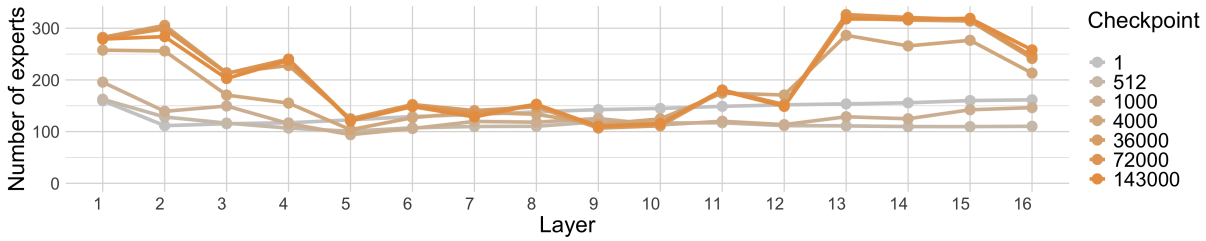

Figure 24: **Pythia 1b.** Average number of experts identified in MLP layers at different depths, for different checkpoints.

F.3 Distribution of experts across attention layers

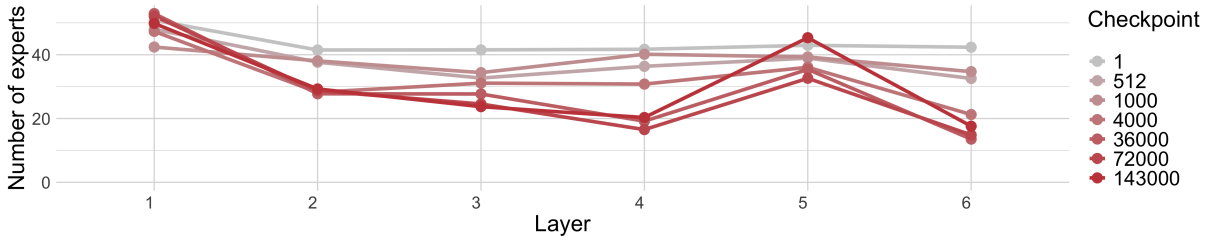

Figure 25: **Pythia 70m.** Average number of experts identified in attention layers at different depths, for different checkpoints.

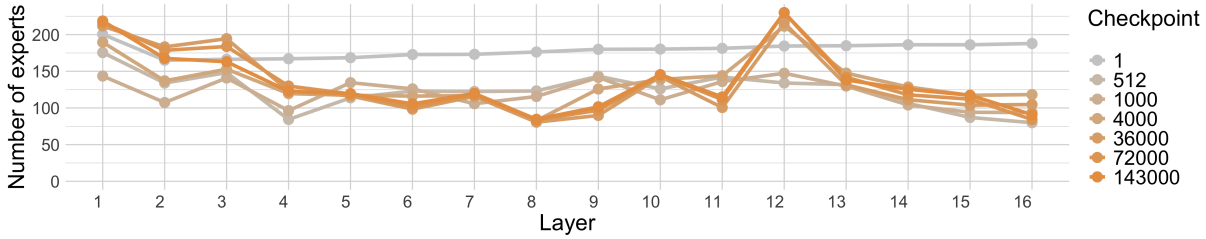

Figure 26: **Pythia 1b.** Average number of experts identified in attention layers at different depths, for different checkpoints.

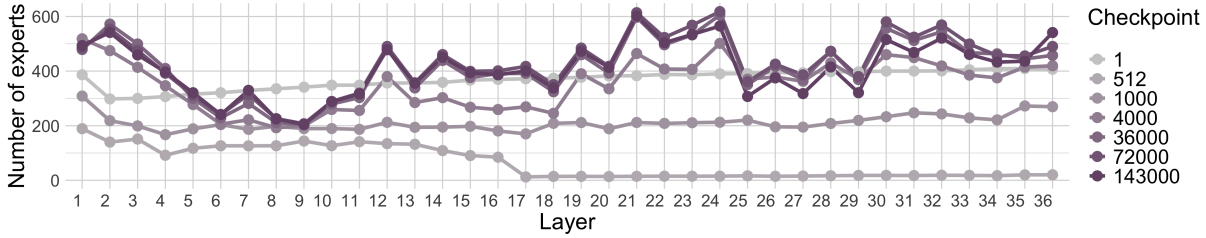

Figure 27: **Pythia 12b.** Average number of experts identified in attention layers at different depths, for different checkpoints.

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

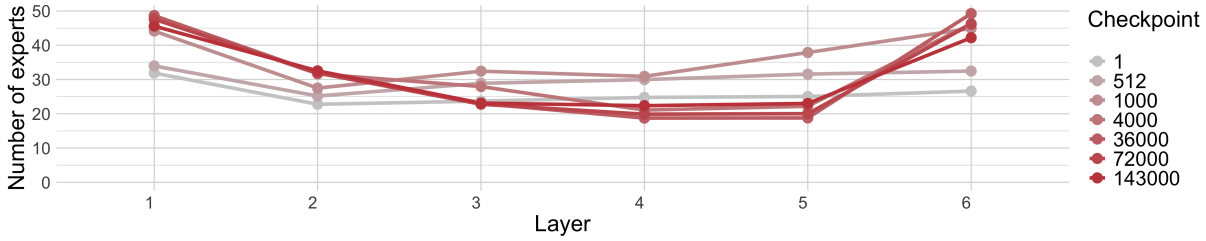

Figure 28: **Pythia 70m.** Average number of experts identified in the MLP `h_to_4h` (part of the MLP layers) at different depths, for different checkpoints.

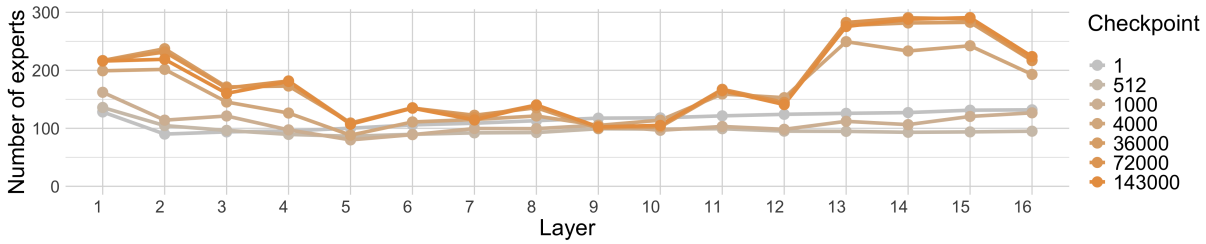

Figure 29: **Pythia 1b.** Average number of experts identified in the MLP `h_to_4h` (part of the MLP layers) at different depths, for different checkpoints.

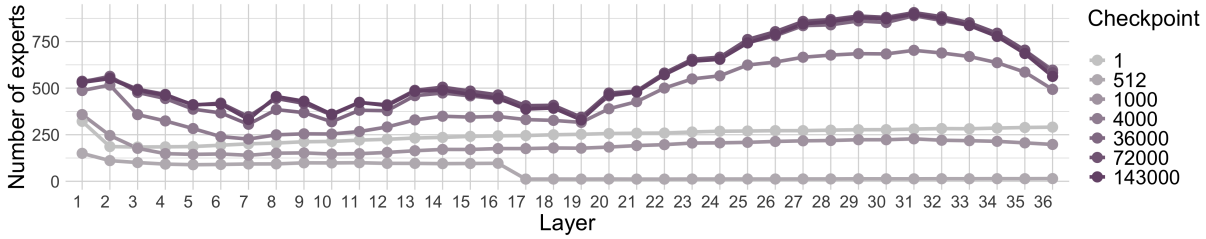

Figure 30: **Pythia 12b.** Average number of experts identified in the MLP `h_to_4h` (part of the MLP layers) at different depths, for different checkpoints.

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

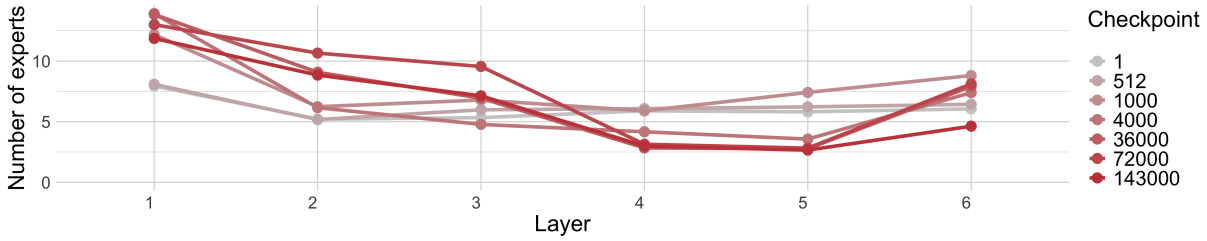

Figure 31: **Pythia 70m.** Average number of experts identified in the MLP `4h_to_h` (part of the MLP layers) at different depths, for different checkpoints.

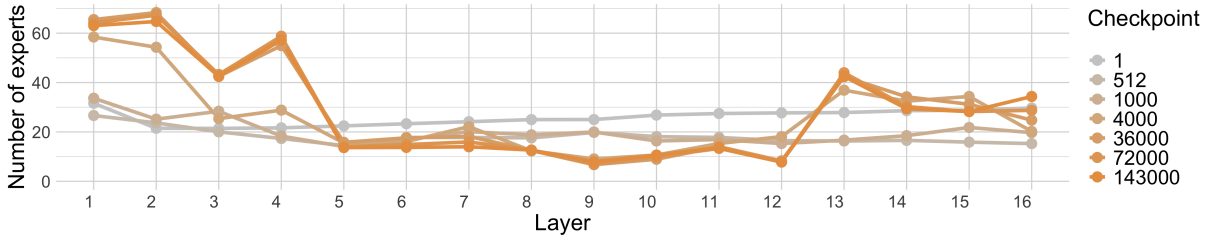

Figure 32: **Pythia 1b.** Average number of experts identified in thr MLP `4h_to_h` (part of the MLP layers) at different depths, for different checkpoints.

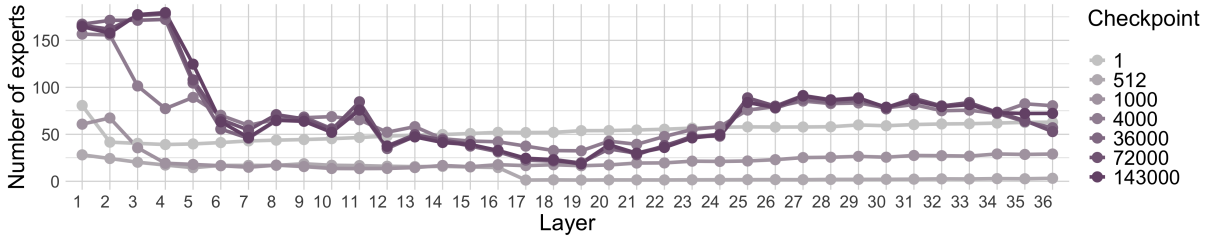

Figure 33: **Pythia 12b.** Average number of experts identified in the MLP `4h_to_h` (part of the MLP layers) at different depths, for different checkpoints.

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

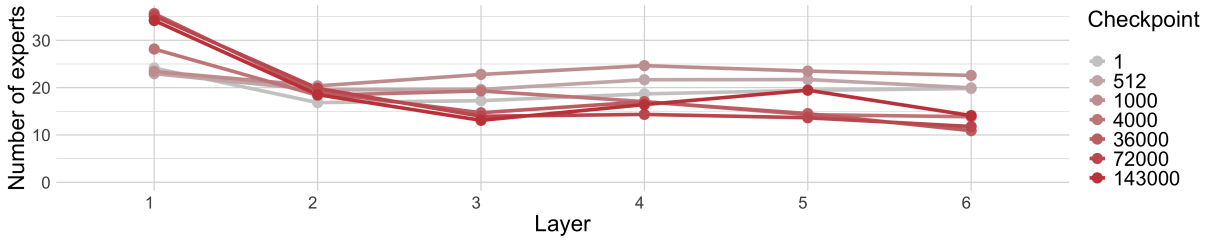

Figure 34: **Pythia 70m.** Average number of experts identified in the `attention.query_key_value` (part of the attention layers) at different depths, for different checkpoints.

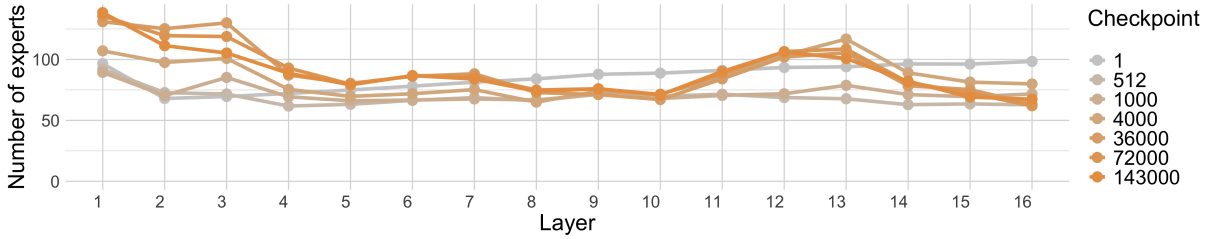

Figure 35: **Pythia 1b.** Average number of experts identified in the `attention.query_key_value` (part of the attention layers) at different depths, for different checkpoints.

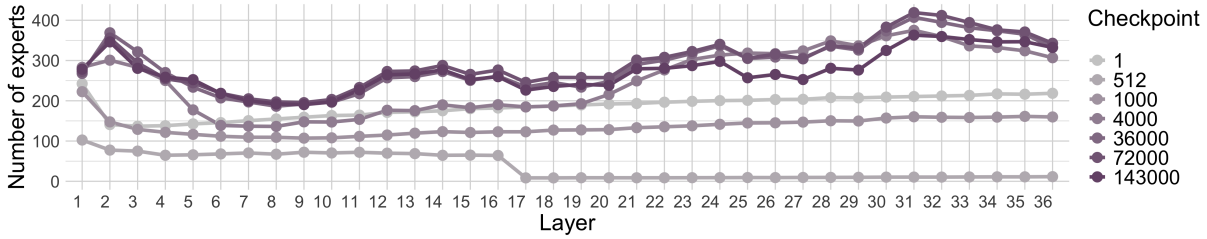

Figure 36: **Pythia 12b.** Average number of experts identified in the `attention.query_key_value` (part of the attention layers) at different depths, for different checkpoints.

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

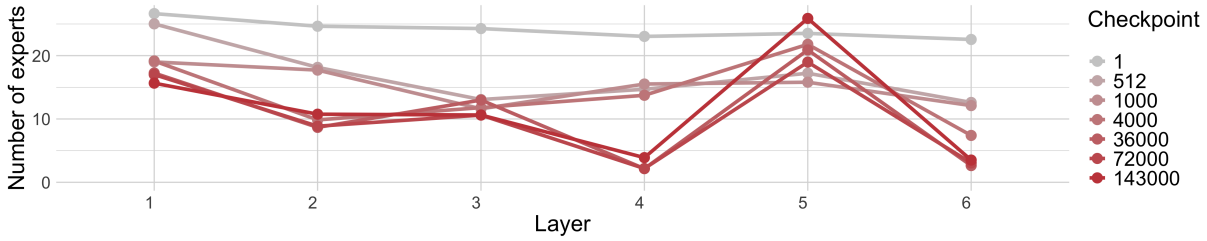

Figure 37: **Pythia 70m.** Average number of experts identified in the `attention.dense` (part of the attention layers) at different depths, for different checkpoints.

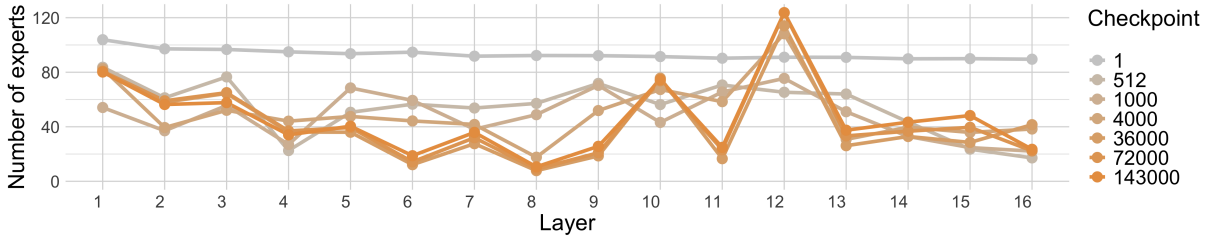

Figure 38: **Pythia 1b.** Average number of experts identified in the `attention.dense` (part of the attention layers) at different depths, for different checkpoints.

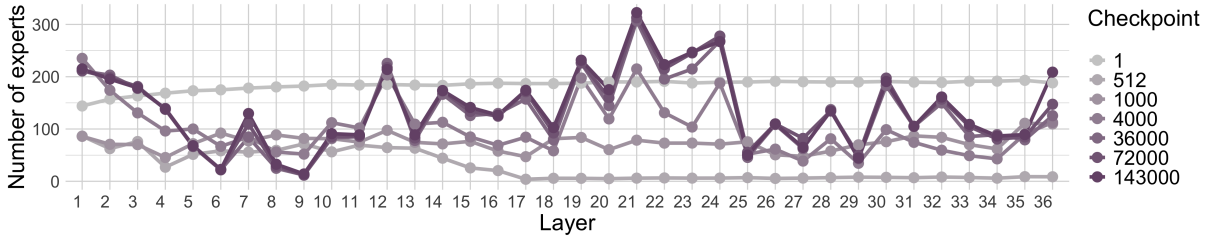

Figure 39: **Pythia 12b.** Average number of experts identified in the `attention.dense` (part of the attention layers) at different depths, for different checkpoints.

F.8 Distribution of highly specialized experts across MLP layers

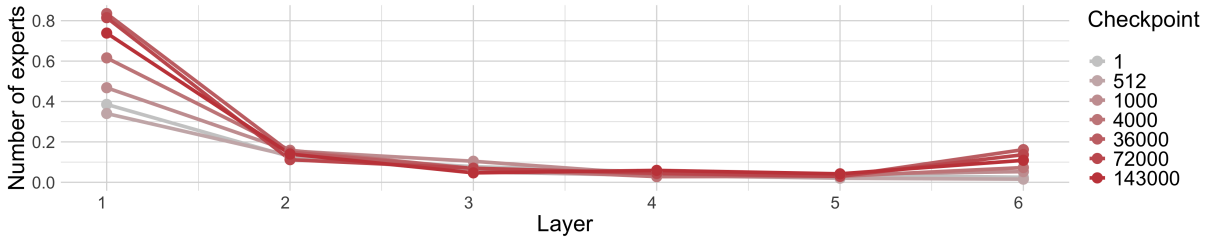

Figure 40: **Pythia 70m.** Average number of highly specialized experts ($\tau = 0.9$) identified in MLP layers at different depths, for different checkpoints.

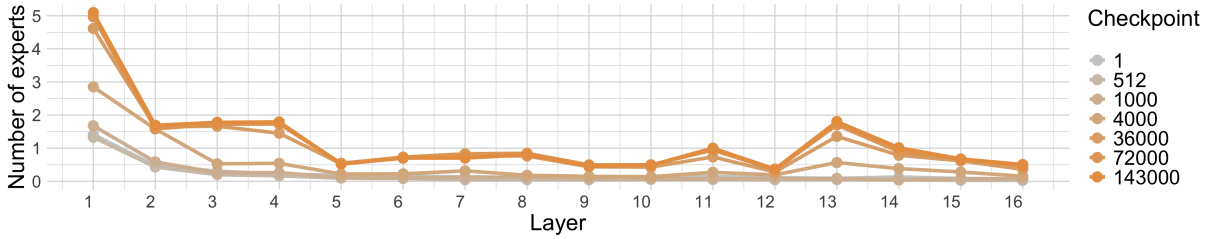

Figure 41: **Pythia 1b.** Average number of highly specialized experts ($\tau = 0.9$) identified in MLP layers at different depths, for different checkpoints.

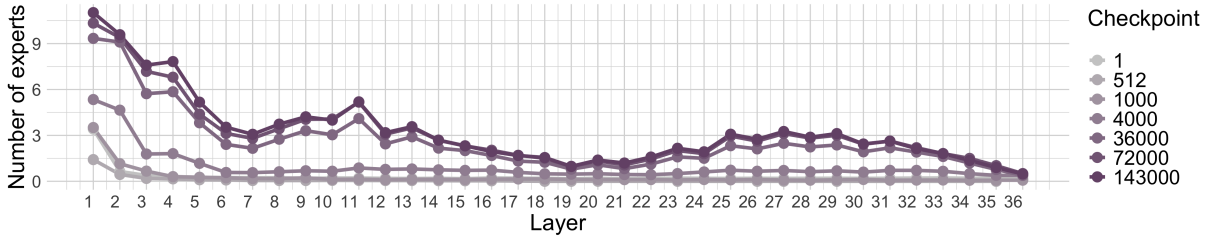

Figure 42: **Pythia 12b.** Average number of highly specialized experts ($\tau = 0.9$) identified in MLP layers at different depths, for different checkpoints.

F.9 Distribution of highly specialized experts across attention layers

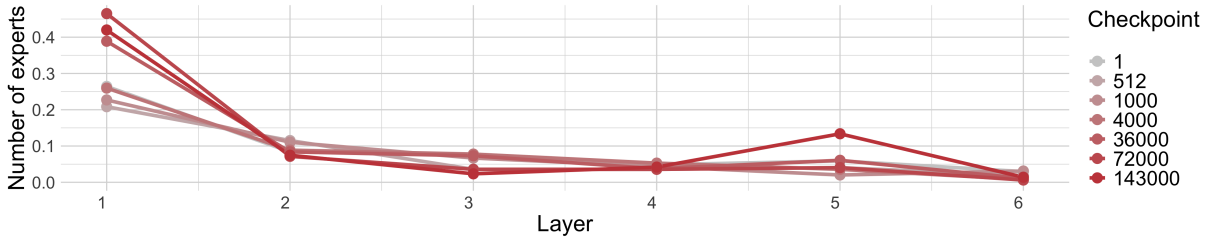

Figure 43: **Pythia 70m.** Average number of highly specialized experts ($\tau = 0.9$) identified in attention layers at different depths, for different checkpoints.

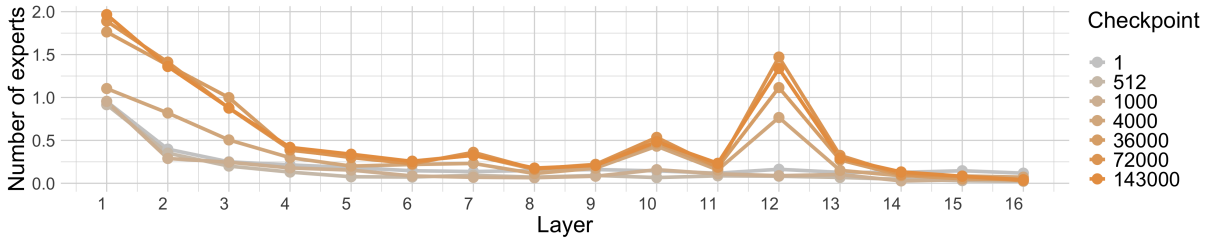

Figure 44: **Pythia 1b.** Average number of highly specialized experts ($\tau = 0.9$) identified in attention layers at different depths, for different checkpoints.

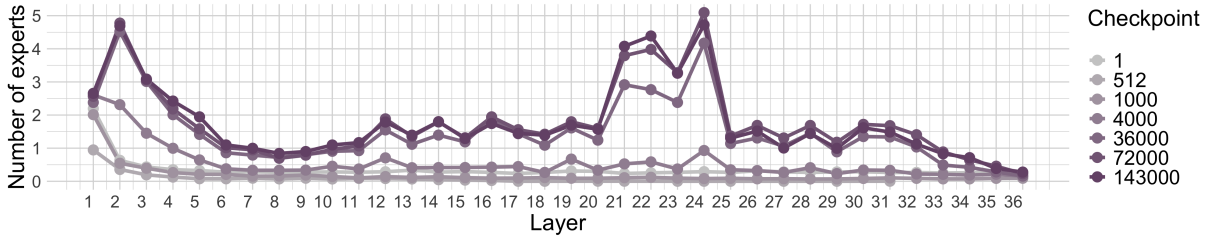

Figure 45: **Pythia 12b.** Average number of highly specialized experts ($\tau = 0.9$) identified in attention layers at different depths, for different checkpoints.

G 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.

H 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.