USING WHAT YOU KNOW: UNDERSTANDING THE ROLE OF SUPERVISED LEARNING IN USING SEMANTIC KNOWLEDGE

Zach Studdiford
Department of Computer Science
Department of Psychology
University of Wisconsin-Madison
studdiford@wisc.edu

Gary Lupyan
Department of Psychology
University of Wisconsin-Madison
lupyan@wisc.edu

ABSTRACT

Self-supervised learning, in which a system learns by minimizing its own prediction errors, is immensely powerful. Modern LLMs are often used as an example of just how much structure it is possible to learn from self-supervision alone. However, just as human cognitive development relies on structured feedback that is parenting and education, it is worth asking how supervised learning augments what LLMs can do. After all, no consumer-facing LLMs are purely self-supervised. We systematically compare purely self-supervised base LLMs to models further refined with supervised training and find that although self-supervision is sufficient for learning the basic knowledge needed to answer our queries, base models fail to use the knowledge in appropriate ways. Further examination of the internal states of the LLMs reveals that models exposed to supervised learning learn to align their semantic representations with the prompt in a way that enables coherent responses.

1 INTRODUCTION

Many things we know we have learned in the absence of direct supervision. No one has to explicitly teach us to speak, laugh at jokes, or climb stairs. Although there are continued debates about the extent to which such learning depends on domain-specific innate knowledge (Elman, 1996; Pinker & Longuet-Higgins, 1994; Fodor, 1983), the astonishing recent advances in AI have come largely from harnessing the power of self-supervised learning (Sutton, 2019; Sejnowski). A self-supervised system learns by adjusting its internal states to minimize the error between its prediction of what will happen and what actually happens (Hinton et al., 1986; Hebb, 2005; Clark, 2013).

Learning language through such self-supervision allows neural networks to begin learning word meanings and syntactic structure even at very small scales (Elman, 1990). Larger self-supervised models–trained on more data but still simply learning to predict what words occur in what contexts–learn word embeddings that encode an impressive breadth and depth of semantic knowledge. Similarity of word embeddings is tightly correlated with human semantic judgments (Periñán-Pascual, 2022; De Deyne et al., 2016) and a variety of semantic norms (Lewis & Lupyan, 2020; Liu et al.; Grand et al., 2022; Caliskan et al.), and measures of prediction error in language models are closely correlated with various measures of online language processing (Michaelov et al., 2023). For all the impressive information that word embedding models capture from the statistics of language, their abilities are very circumscribed. One cannot ask *Word2Vec* a direct question, much less have a conversation with it.

Larger and more powerful transformer-based models like GPT-3 which are trained on much larger text corpora learn increasingly sophisticated representations of both language (Hu et al.) and world knowledge (Roberts et al., 2020). But perhaps surprisingly (from the vantage point of someone used to the capabilities of the latest LLMs), their ability to actually *use* language is quite limited. As an example, Table 1 shows a comparison of a conversation with a purely self-supervised LLM and a model that has been further fine-tuned with a *supervised* training set.

Self-supervised (`gemma-2-9b`)	**Supervised (`gemma-2-9b-it`)**
User: Who invented backpropagation?	**User:** Who invented backpropagation?
Model: I did.	**Model:** The invention of backpropagation is not credited to a single person.
User: How is that possible if you were trained using backpropagation?	**User:** Name a few contributors.
Model: I was trained using backpropagation, but I invented it.	**Model:** Seppo Linnainmaa; Paul Werbos; David Rumelhart; Geoffrey Hinton.

Table 1: Comparison of conversations with a *base* version of `gemma-2-9b` trained entirely through self-supervision, and a *fine-tuned* version trained with additional supervised learning.

While relations such as $king \rightarrow queen = man \rightarrow woman$ can be recovered from word embeddings (Mikolov et al., 2013), self-supervised (or base) LLMs often struggle to use these kinds of relations when prompted (Brown et al., 2020). For these kinds of associations to be uncovered at all requires significant effort on the part of the researcher to localize the relevant knowledge. Finding that a word embedding model has learned that, e.g., "nurse" is more strongly associated with women than men (Caliskan et al.) is based on researchers explicitly comparing the semantic representation of "nurse" to embeddings of feminine and masculine words. What the model has learned is accessible when the model's representations meet the ingenuity of the researchers' probes.

This observation raises the question at the center of the present work: While self-supervised learning may be sufficient to learn sophisticated semantic representations and even world models (Hazineh et al., 2023), is it enough to allow a model to *use* this knowledge appropriately? Our results show that knowledge that is clearly learnable even by small word-embedding models cannot be effectively used by (even much larger) LLMs without some supervised training. We conclude that supervised training may be necessary for models to *use* their previously-learned semantic representations in appropriate ways.

1.1 What do LLMs learn when they are "fine-tuned"?

All consumer facing LLMs combine self-supervised learning with several rounds of supervised training [1] to *fine-tune* the model. (Zhang et al., 2026). During this supervised learning regimen, models are exposed to input-output pairs consisting of example exchanges where knowledge must be retrieved or deployed in context. These include question-answering ("*Q: What is surfing? A: Surfing is a surface water sport...*"), step-by-step reasoning ("*Q:Yes or no, would a pear sink in water? A: The density of a pear is about* $5g/cm^3$*, which is less than water. Thus, a pear would float. So the answer is no.*") and instruction following ("*Q:Write a short story about a cat named Bistro*. A: Bistro patiently waited...")[2], and are often curated based on human intuitions about how to best explain concepts and reason through problems (Zhou et al., 2022). Through this supervised learning signal, models acquire normative representations of how they should respond and use the knowledge they have learned in context.

Compared to the massive datasets and compute used by self-supervised "pretraining", supervised training is relatively minuscule. It uses many orders of magnitude less training data and compute (Zhou et al., 2023), and typically accounts for less than 1% of all updates to the model weights (Zhou et al., 2023; Ouyang et al., 2022). Yet not only does it make the models usable (as anyone who has attempted a conversation with a base model can attest), fine-tuned models are also far more efficient (Varshney et al., 2023).

Despite the efficacy of predictive self-supervised learning (see e.g., Clark, 2013; Hohwy), much of what people's knowledge seems to require (at least in practice) is *supervised* learning. Parenting is case in point, involving both explicit feedback and continued positive and negative reinforcement

[1] For the purposes of the present study, we consider supervised learning to be all learning methods where there is a learning signal external to the model's own predictions. This includes SFT (supervised fine-tuning), DPO (direct preference optimization) and RLHF(Reinforcement learning from human feedback). For a full discussion of various supervised methods, see Xu et al. (2026).

[2] All examples seen here are provided by Conover et al. (2023)

signals. Schools can also be seen as a form of largely supervised learning. Teachers not only provide explicit feedback, but they provide examples of tasks (compare and contrast; add these numbers; prove this theorem) that students are expected to complete, and in so doing generate sources of error that would not otherwise exist.

There is no simple way to know what function such supervised learning plays in human cognition because we cannot realistically separate the effects of supervised and self-supervised learning episodes in people. We can, however, do this in LLMs. This allows us to examine the upper bound of what can be learned through self-supervised learning alone–and alternatively, where a supervised error signal is required–to directly compare the reasoning abilities of models which differ only in whether they were exposed to an additional regimen of supervised learning. We therefore use evaluations that have been previously used as tests of simple forms of world knowledge in people, and measure how well models with and without supervised training can generate human-like outputs. As a more stringent test of the possibility that many simple kinds of world knowledge are learnable *in principle* by self-supervised algorithms even if they are not expressed *in practice*, we compare LLM performance to information that can be extracted from much smaller and simpler word embeddings models (e.g., *Word2vec*). We next examine the internal states of base and fine-tuned language models to confirm that the difference in their performance is not due to supervision "teaching" models new factual knowledge. This approach allows us to distinguish differences in performance from differences in how that knowledge is used.

2 Behavioral Evaluations

2.1 Stimuli

We probed knowledge of concrete and more abstract concepts using two tasks originally developed for probing human semantic representations. These tasks evaluated both simple, objective object properties (i.e, whether a *watermelon* is larger than a *blueberry*) and more subjective associations that require evaluating various concepts on different dimensions, e.g., evaluating the dangerousness of various sports. For these, previously collected human judgments served as the "ground truth".

2.1.1 Object properties.

We tested base and fine-tuned LLMs on their ability to reason about simple object properties using the Round Things dataset (Giallanza et al., 2024). The dataset consists of 46 round items from the THINGS concept database (Hebart et al., 2019) that vary along two dimensions: the *size* (diameter) of each item, and the object *kind* (whether the object was a plant or an artifact).

2.1.2 Semantic associations.

We evaluated models on their subjective associations involving 9 semantic domains (e.g, *cities*, *animals*, *professions*) and 17 features associated with each concepts in these domains (e.g, *size*, *danger*, *gender*). For example: "*how dangerous is a tiger from 0-100?"*. We compared the answers provided by (or extracted from) the LLMs to human judgments collected by Grand et al. (2022).

2.2 Models

We evaluated base and fine-tuned versions of the *Gemma* family of open-source models that varied in size from 2-27 billion parameters: `gemma-2-2b`, `gemma-2-9b` and `gemma-2-27b` (Team et al., 2024). Varying parameter size along with the self-supervised/supervised contrast allowed us to see whether the effects of supervision differ for models of varying sizes.

2.3 Probing what LLMs know

For both the *object properties* and *semantic associations* datasets, we obtained LLM judgments for all combinations of stimuli and their corresponding properties. In addition to probing knowledge of simple objective properties and more subjective associations, we were also interested in testing whether supervised training affected how sensitive the models were to the manner in which their

knowledge was probed. To this end, we created three question templates: semantic differential ratings , triplet judgments, and multiple alternative-forced choice.

1. **Semantic differentials.** For a given item, we ask for a rating 0-100 along a dimension (i.e *safe* to *dangerous*). This semantic differential format is similar to that used to solicit human judgments in Grand et al. (2022). Critically, we do *not* name the dimension. Instead the dimension is anchored by three negative and three positive words. For example, the size dimension is anchored by *small, little, tiny* and *large, big, huge*). We measure LLM performance in semantic differentials as the correlation between human and LLM scores. An example semantic differential prompt is given below:

```
Rate the item on the following scale:
0 (small, little, tiny)   to   100 (large, big, huge)

Item:  elephant

Answer with a single number between 0 and 100.
Answer:
```

2. **Triplet judgments.** This classic task requires choosing which of three items is different from the other two. Because we are interested in probing the models' ability to weigh a particular dimension, our prompt includes the dimension name, e.g., *"Which item is more similar to a in terms of size, b or c?"*. Because the choice can differ depending on what dimension is mentioned in the prompt, correct answers require flexibly emphasizing the mentioned dimension, not merely relying on overall similarity. We evaluate the triplet task on the *object properties* dataset, which consists of orthogonal dimensions kind and size. Following from tests of cognitive control in Giallanza et al. (2024), We define a *match* condition–where the correct kind judgment is also a correct size judgment–as well as *conflict* size and kind conditions to test model ability to respond using one dimension when there is conflicting information from another.
3. **n-alternative forced choice.** We include additional n-alternative forced-choice evaluations, where n ranges from 2 to 5. Similar to scoring criteria in the triplet judgment format, LLM response correctness is determined on whether the choice matches the mean human response (or object property). An example of a 2AFC trial prompt is *which item is larger, a blueberry or a watermelon"*?

2.4 Assessing what is known to word embeddings models

To determine what is–in principle–learnable by small-scale self-supervised models, we probed whether word embeddings models (which learn static vector representations of words by predicting co-occurrences) are sensitive to the same semantic properties that are evident in much larger LLMs. Following Grand et al. (2022) (see also Caliskan et al.) we project these static embeddings vectors onto specific dimensions (i.e a *size* dimension obtained by subtracting the embedding for "*small*" from the embedding for "*large*"). For example, the embedding for "*elephant*" should project more in the "*large*" direction than the embedding for "*mouse*". While the word embeddings model is not directly *answering* the question, a high correlation of its projection to human judgments shows that the model's knowledge that, e.g., an elephant is larger than a mouse can be recovered given the right probe.

3 Behavioral Results

The overarching aim of this study was to understand whether supervision helps LLMs *use* knowledge that they already have (from self-supervised training). Our behavioral results answer this question at the level of performance: what questions can supervised models reason about that self-supervised models cannot?

3.1 Base LLMs systematically fail to use their knowledge appropriately.

Figure 1a shows accuracy and alignment scores for all base and fine-tuned models. Even large base models (27 billion parameters) fail to reliably choose the correct answer and yield only small

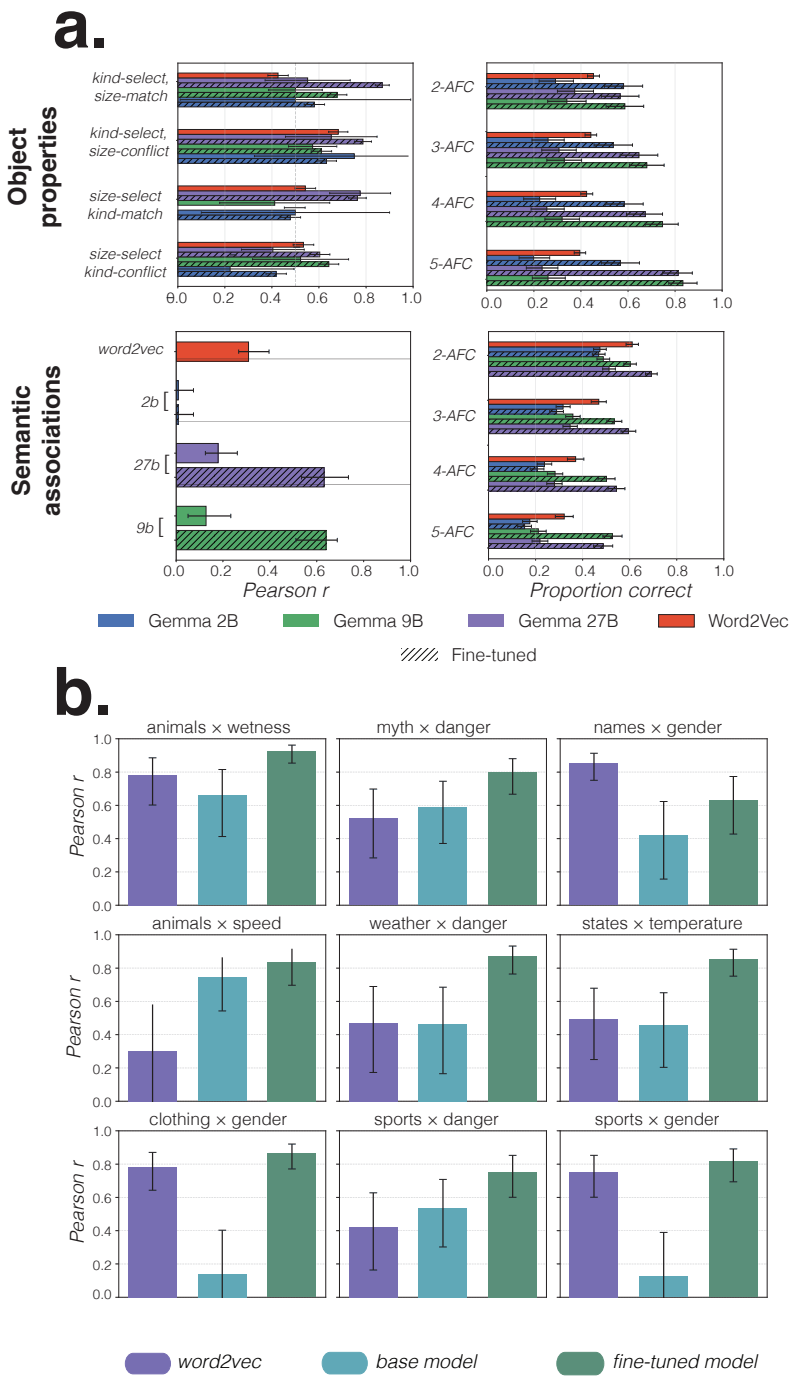

Figure 1: **a.** Measures of accuracy and alignment for base and fine-tuned models with *word2vec* embeddings similarities included as a comparison. For object properties task, correlations are between objective size/kind information and LLM outputs; For semantic associations, the correlations are between human judgments and model outputs. Striped bars indicate fine-tuned models. Error bars indicate 95% confidence intervals. **b.** Comparisons of human-alignment for word2vec and base and fine-tuned versions of `gemma-2-9b` for nine category-feature combinations. 95% confidence intervals were calculated using a bootstrap sampling procedure.

correlations with human semantic differential judgments ($r_{\text{semantic}} = 0.20$; best accuracy $= 0.51$ on FC2 task). In contrast, fine-tuned models show much higher alignment with subjective human ratings ($r_{\text{semantic}} = 0.57$) and respond with high accuracy across various question formats (best accuracy $= 0.70$ on FC2 task). Critically, this effect emerges with scale: the smaller (2-billion) parameter model shows consistently poor performance in both the base and fine-tuned conditions ($\text{accuracy}_{\text{2B-base}} = 0.27$ vs. $\text{accuracy}_{\text{2B-IT}} = 0.26$, respectively), compared to the 27-billion parameter fine-tuned model ($\text{accuracy}_{\text{27B-IT}} = 0.59$). A mixed-effects regression predicting model accuracy and alignment (with fixed effects for model size, fine-tuning, and task type, and random intercepts and slopes for task nested within question ID) confirms that fine-tuned models demonstrate a significant advantage in performance relative to base models ($\beta_{\text{tuning}} = 0.062$, SE $= 0.004$, $t(1676) = 14.43$, $p < .001$), with this benefit increasing as model size scales ($\beta_{\text{size}\times\text{tuning}} = 0.065$, SE $= 0.004$, $t(1675) = 15.08$, $p < .001$).

The triplet task evaluation for the object properties dataset reveals additional divergences in behavior between base and instruction-tuned LLMs consistent with the semantic associations task: namely, that performance benefits from fine-tuning, especially with scale ($\beta_{IT} = 0.131$, $p < 0.001$ for instruction tuning; $\beta_{27B} = 0.226$, $p < 0.001$ for 27B vs. 2B models). Interestingly, we find that both base and fine-tuned LLMs favor *kind* judgments over *size* judgments ($\beta_{kind} = 0.129$, $p < 0.001$), commensurate with previous findings from Studdiford et al. (2025) demonstrating that base and instruction-tuned models are predisposed to favor certain semantic dimensions over others.

3.2 Word embeddings outperform larger base models.

Finding that LLMs with billions of parameters perform so poorly is surprising because the knowledge required to answer our questions is well-represented in simple word embedding models (mean accuracy $= 0.42$). This replicates findings from Grand et al. (2022) indicating robust (but graded) representations of human semantic judgments in word embeddings. The previous mixed effects model reveals that word embeddings models indeed perform significantly better than base models ($\beta_{\text{word2vec}} = 0.124$, $p < .001$), but worse than fine-tuned LLMs ($\beta_{\text{fine-tuned}} = 0.176$, $p < .001$). Figure 1b compares alignment between *word2vec* and base and fine-tuned versions of `gemma-2-9b`.

4 Mechanistic Analyses

4.1 Probing LLM internal representations

Drawing on mechanistic interpretability techniques (Rai et al., 2024), we probe the hidden states of base and instruction-tuned models in responding to our evaluation questions, allowing us to gauge whether an error in a model's response stems from the model lacking the requisite knowledge or failing to *use* it appropriately. For all interpretability analyses, we use `gemma-2-9b` activations from the semantic associations evaluation in the *semantic differentials* format. Thus for all analyses the "ground truth" comparison is a vector of continuous human judgments ranging from 0 to 100.

4.1.1 Training linear probes to determine whether knowledge exists.

We first ask whether simple linear probes can recover human semantic judgments even in cases where this knowledge is not expressed. To this end, we trained linear regressions at each layer of base and fine-tuned models, given as $y_i = \mathbf{w}_\ell^\top \mathbf{r}_{i,\ell} + b_\ell$, where $y_i \in [0, 100]$ denotes the human judgment score for concept–feature item i, $\mathbf{r}_{i,\ell} \in \mathbb{R}^d$ is the hidden state (residual stream) representation of item i at layer ℓ, and $\mathbf{w}_\ell \in \mathbb{R}^d$ and $b_\ell \in \mathbb{R}$ are learned model parameters. Intuitively, a high model R^2 indicates that a subspace of the LLM activations at a given layer accurately captures human judgments. We also train *within-category* probes in order to assess the semantic richness of model representations: that is, can we decode other (non-relevant to the given question) features in the forward pass of model activations?

4.1.2 Activation patching to compare differences in model computations.

At what point do the representations of base and fine-tuned models diverge? We conduct a series of *patching* experiments to determine the layer at which base and fine-tuned models show demonstrably different performance. Specifically, for a given layer ℓ in the base model M_{base} and the fine-tuned

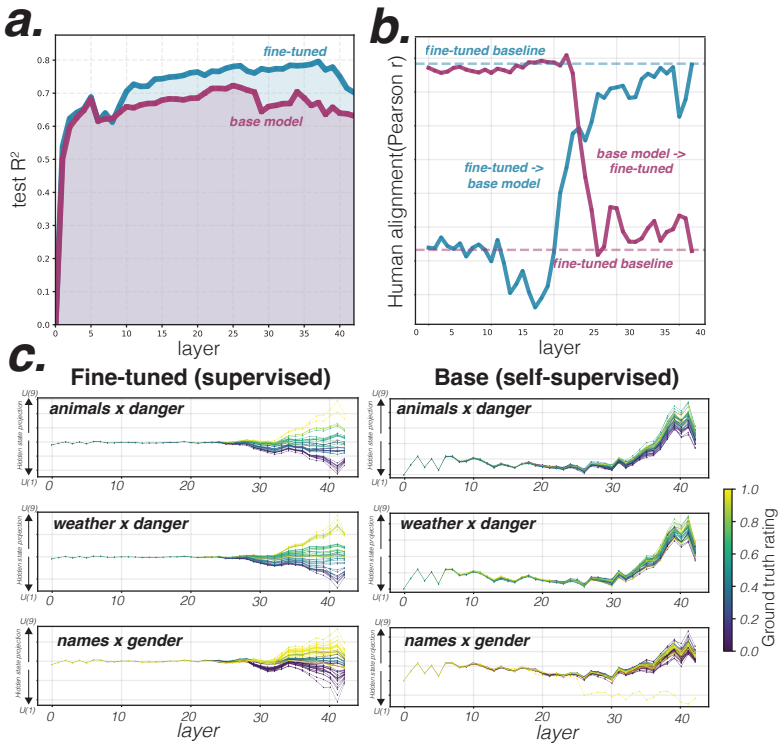

Figure 2: **a.** Linear probe R^2 for all layers of base and fine-tuned `gemma-2-9b`. **b.** Effect of activation-patching between base and fine-tuned models at all layers $0...n$. Dashed lines indicate the baseline correlations of base and fine-tuned LLMs prior to patching. The blue and red lines indicate patching from fine tuned to base and base to fine-tuned models, respectively. **c.** Projections of model hidden states onto the output space at each layer for three sample conditions.

model M_{ft} and a given prompt p, we obtain the forward-pass activations $\mathbf{h}^{\text{base}}_{\ell,p}$ and $\mathbf{h}^{\text{ft}}_{\ell,p}$ from the residual stream. We then patch the activations $\mathbf{h}^{\text{base}}_{\ell,p}$ into model M_{ft} and $\mathbf{h}^{\text{ft}}_{\ell,p}$ into model M_{base}, and run a modified forward pass for M_{base} and M_{ft}. We measure the difference in the human correlation of model M_{base} and M_{ft} outputs in order to determine where base and fine-tuned models diverge in their ability to express accurate world knowledge.

4.1.3 Determining how knowledge is used.

To the extent that accurate world knowledge is *decodable* in both base and fine-tuned LLMs, can we explain why this knowledge is expressed in some cases and not others? Since models must ultimately express their world knowledge using the 0-10 scalar output in the *semantic differential* condition, we ask whether there are differences in how models express their representations in this output space. Using the *logit-lens* technique (nostalgebraist, 2020) we obtain the difference in the output token embeddings $U(0)$ and $U(9)$ [3]. This vector represents the direction between a low output score and high output score by the model. We then measure the projection of the model hidden state (residual stream) onto this difference vector at each layer. Intuitively, a lack of differentiation in LLM hidden states with respect to this axis indicates that the model is not projecting onto (or "using") the relevant output.

[3]We use 0-9 here as opposed to 0-100 because single digit integers are represented as individual tokens in `gemma-2`.

5 MECHANISTIC RESULTS

5.1 BASE AND FINE-TUNED MODELS HAVE HIGHLY ACCURATE WORLD KNOWLEDGE REPRESENTATIONS.

Perhaps surprisingly given the low performance of base models, we find that both base and fine-tuned LLMs represent human-aligned world knowledge with high accuracy: that is, we are able to fit linear models that predict held-out human responses $0 - 100$ using the hidden-state activations of base and fine-tuned `gemma-2-9b` with an impressive degree of fidelity. Figure 2a. shows linear probe R^2 for base and fine-tuned versions of `gemma-2-9b`. While linear model R^2 is higher in the intermediate layers of the fine-tuned model relative to the base LLM (max $R^2_{\text{IT}} = 0.80$ at layer 37 vs. max $R^2_{\text{base}} = 0.73$ at layer 26), we report held-out test accuracy as high as 73% in the intermediate representations of base models (aggregate $R^2_{\text{base}} = 0.65$ across layers). The accuracy of these linear probes is more interesting when recalling that the probes are fit across *all* stimuli, indicating that there is a single subspace of activations in base and fine-tuned LLMs that predicts a diverse range of human semantic judgments.

5.2 DIVERGENCE IN QUESTION ANSWERING MECHANISMS

If base models "know" that elephants are larger than mice, why isn't this knowledge expressed in model outputs? We answer this question using two interpretability techniques: activation patching and logit-lens (see Methods section). We begin by patching activations between the base and fine-tuned LLM in all layers $0...n$ to determine the point at which there is a meaningful divergence in model computations with respect to human alignment. We find that for the first $n/2$ model layers, patching between models has no effect: the computations performed are identical. In late ($k > n/2$) layers, patching base model activations systematically reduces fine-tuned model performance to that of the base model and vice versa. This indicates that there is a computation in the second half of model layers which explains the difference in fine-tuned model performance. What might this critical computation be? We next look to see if base and fine-tuned models are able to project their (evidently rich) internal semantic representations onto the output space. That is, whether models can translate their semantic representations into an accurate number 0-100. Figure 2c. shows the projection of base and fine-tuned hidden states[4] onto the output activation space 0-9. We find that these projections are highly aligned with ground-truth human ratings in the fine tuned models, but not the base models. Notably, fine-tuned models begin differentiating their representations along this axis at roughly the same layer where computations diverge in the patching experiment.

5.3 FINE-TUNED LLMS LEARN SEMANTICALLY RICH REPRESENTATIONS

Can the differences between base and fine-tuned LLMs be explained entirely as a failure to use existing representations? While there are marked differences in how representations are deployed in base and fine-tuned LLMs, we also find that fine-tuning instills significantly richer semantic representations relative to base models. We fit linear probes similar to the previous analysis but only trained on the activations of a single category feature pair (i.e rating *animals* on *size*). We then asked whether the predictions of these probes generalize to other feature axes not seen in training (i.e does the *animal-size* probe predict *animal-danger*? We find that human-aligned semantic representations in fine-tuned LLMs are highly generalizable to other feature judgments. That is, these representations contain rich latent semantic information about *other* features such that we can predict human judgments about those features with high accuracy (aggregate $R^2_{\text{IT}} = 0.33$, ranging from 0.01 to 0.92 across feature pairs). Notably, the activations of the fine-tuned LLM are significantly more predictive than those of the base model (aggregate $R^2_{\text{base}} = 0.21$), and in some cases are predictive above and beyond cross-feature correlations in human scores (see Figure 3). In sum, base and fine-tuned models diverge both in how representations are deployed and the semantic content of the representations themselves.

[4] we use "hidden state" in reference to the residual stream activations at some layer.

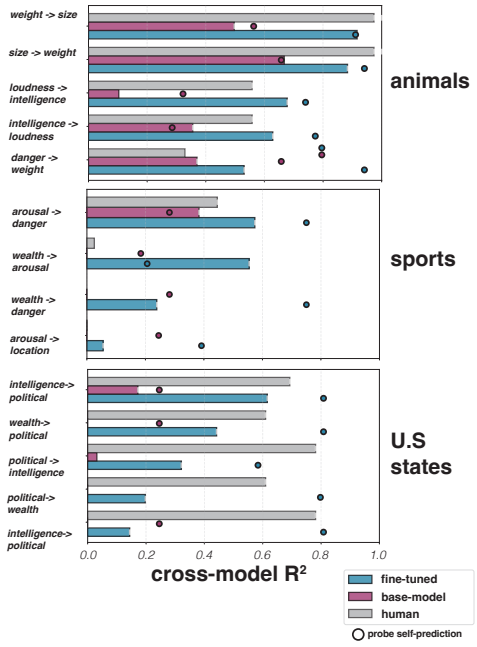

Figure 3: Generalization of base (red) and fine-tuned (blue) linear probes to other category features for three example category-feature combinations. Grey bars indicate the baseline correlation of human scores for feature pairs. Dots indicate R^2 values for probes predicting their own features.

6 Discussion

Are there limits on what a predictive model can learn through self-supervision? Given enough training data, an arbitrarily large neural network model is a *universal function approximator*, capable of learning any distribution (Cybenko, 1989). And yet, our empirical results suggest that there are constraints on what a neural network–even one scaled to tens of billions of parameters–can learn from self-supervision alone. It is not necessary to evaluate models on esoteric benchmarks (Chollet, 2019) or arbitrary reasoning problems to find these limitations: rather, they are evident in the failures of base models–even those with tens of billions of parameters–to reason about exceedingly simple forms of world knowledge, such as the size of elephants relative to mice, or that tornadoes are perceived as more dangerous than rainstorms.

This is not to say that self-supervised algorithms are incapable of learning this kind of world knowledge, or that language itself does not encode this information. Indeed, word embeddings models–several orders of magnitude smaller in scale–encode these associations implicitly (Grand et al., 2022; Liu et al.). In self-supervised LLMs, we find that these same semantic associations are encoded with even greater detail, such that fine-grained human judgments can be predicted with high accuracy. However, even as these representations grow more robust in larger parameter models, these models remain largely unable to use this knowledge in their outputs.

We demonstrate that it is only after language models have been exposed to some form of *supervised* learning that they are capable of expressing the high-fidelity information found in their hidden states with any degree of accuracy: once *fine-tuned* on instruction-following and question-answering datasets, these models are able to produce outputs that closely mirror human semantic judgments and ground-truth object properties. How is it possible that a small amount of supervised input could have this disproportionate effect on model performance? By examining the internal mechanisms implemented by base and fine-tuned models for reaching (or not reaching) the correct answer, we find that fine-tuned models learn to align their semantic representations with an output space (i.e a number rating 1-9), a behavior which base models systematically fail to implement. We also discover that the internal representations of fine-tuned models are themselves more robust: these semantic representations reflect not just the world knowledge being tested but also encode other

useful human-aligned judgments. That is, supervised learning instills representations that are more aligned to the given context and in general. Taken together, these findings suggest a privileged role for supervised learning in what models can do with what they have learned.

7 LIMITATIONS

We acknowledge several limitations in our findings. Although we find differences in the ability of fine-tuned LLMs to deploy human-aligned world knowledge relative to base models, the contents of this fine-tuning are opaque. That is, we do not know the specific regimen of supervised learning used in `gemma-2` that led to the improvement in fine-tuned model alignment. As a result, we are not able to dissociate the role RLHF (reinforcement learning from human feedback) from supervised fine-tuning, as both were used in the post-training step (Team et al., 2024). We plan to fine-tune base models ourselves in a future version of this work in order to understand how the use of certain kinds of stimuli in supervised learning affects the alignment of model representations.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. 356(6334):183–186. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230. URL `http://science.sciencemag.org/content/356/6334/183`.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL `https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm`.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 1861–1870, 2016.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Jeffrey L Elman. *Rethinking innateness: A connectionist perspective on development*, volume 10. MIT press, 1996.

Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

Tyler Giallanza, Declan Campbell, Jonathan D Cohen, and Timothy T Rogers. An integrated model of semantics and control. *Psychological Review*, 2024.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987, 2022.

Dean S Hazineh, Zechen Zhang, and Jeffery Chiu. Linear latent world models in simple transformers: A case study on othello-gpt. *arXiv preprint arXiv:2310.07582*, 2023.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10):e0223792, 2019.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

Geoffrey E Hinton, Terrence J Sejnowski, et al. Learning and relearning in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, 1(282-317):2, 1986.

Jakob Hohwy. *The Predictive Mind*. Oxford University Press. ISBN 978-0-19-968673-5.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. Language models align with human judgments on key grammatical constructions. 121(36):e2400917121. ISSN 1091-6490. doi: 10.1073/pnas.2400917121.

Molly Lewis and Gary Lupyan. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, 4(10):1021–1028, 2020.

Qiawen Liu, prefix=van useprefix=true family=Paridon, given=Jeroen, and Gary Lupyan. Learning about color from language. 3(1):60. ISSN 2731-9121. doi: 10.1038/s44271-025-00230-9. URL https://www.nature.com/articles/s44271-025-00230-9.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. Can peanuts fall in love with distributional semantics? *arXiv preprint arXiv:2301.08731*, 2023.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

nostalgebraist. Interpreting GPT: The logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed 2026-01-29.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Carlos Periñán-Pascual. Measuring associational thinking through word embeddings. *Artificial Intelligence Review*, 55(3):2065–2102, 2022.

Steven Pinker and Christopher Longuet-Higgins. The language instinct: how the mind creates language. *Nature*, 368(6469):360–360, 1994.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

Terrence J. Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. 117 (48):30033–30038. doi: 10.1073/pnas.1907373117. URL https://www.pnas.org/doi/10.1073/pnas.1907373117.

Zach Studdiford, Timothy T Rogers, Siddharth Suresh, and Kushin Mukherjee. Evaluating steering techniques using human similarity judgments. *arXiv preprint arXiv:2505.19333*, 2025.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Neeraj Varshney, Agneet Chatterjee, Mihir Parmar, and Chitta Baral. Accelerating llm inference by enabling intermediate layer decoding. *CoRR*, 2023.

Lingling Xu, Haoran Xie, S Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.