IS THIS JUST FANTASY? LANGUAGE MODEL REPRESENTATIONS REFLECT HUMAN JUDGMENTS OF EVENT PLAUSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models (LMs) are used for a diverse range of tasks, from question answering to writing fantastical stories. In order to reliably accomplish these tasks, LMs must be able to discern the modal category of a sentence (i.e., whether it describes something that is possible, impossible, completely nonsensical, etc.). However, recent studies have called into question the ability of LMs to categorize sentences according to modality (Michaelov et al., 2025; Kauf et al., 2023). In this work, we identify linear representations that discriminate between modal categories within a variety of LMs, or **modal difference vectors**. Analysis of modal difference vectors reveals that LMs have access to more reliable modal categorization judgments than previously reported. Furthermore, we find that modal difference vectors emerge in a consistent order as models become more competent (i.e., through training steps, layers, and parameter count). Notably, we find that modal difference vectors identified within LM activations can be used to model fine-grained human categorization behavior. This potentially provides a novel view into how human participants distinguish between modal categories, which we explore by correlating projections along modal difference vectors with human participants' ratings of interpretable features. In summary, we derive new insights into LM modal categorization using techniques from mechanistic interpretability, with the potential to inform our understanding of modal categorization in humans.

1 INTRODUCTION

Language models (LMs) trained on internet-scale data have demonstrated great success in answering questions about our world (Brown et al., 2020), often displaying a surprising understanding of the physical laws governing it (Gurnee & Tegmark, 2024). However, much of the content on the internet does not accurately reflect the world that we live in — it overrepresents unlikely events (Gordon & Van Durme, 2013), contains innumerable pages of text about fantastical fictional universes, and even contains completely nonsensical sentences (e.g., of the *colorless green ideas sleep furiously* variety (Gulordava et al., 2019), or in the collected lyrics of The Beatles). This raises the question: How do LMs determine whether a sentence describes actual reality, a hypothetical scenario, or something more inconceivable — i.e., how do LMs map linguistic expressions to their corresponding **modal categories**?

Understanding how LMs represent modal categories is essential for at least two reasons. First, LMs are increasingly deployed as knowledge bases (Petroni et al., 2019), including in high-stakes situations (Magesh et al., 2024; Cheng et al., 2023). Having the ability to distinguish facts about the real world from flights of fancy (and total nonsense!) is a crucial prerequisite for such applications. Secondly, understanding an LM's representation of modal categories can be informative for uncovering the underlying "world model" that it has encoded (Mitchell, 2025; Li et al., 2023). This includes coarse-grained knowledge, such as whether a scenario could happen in *some* possible world, and more fine-grained knowledge, such as whether a scenario is probable, improbable, or impossible in the real world. Modal intuitions have long been used to characterize the folk theories that people employ to understand domains such as physics (McCoy & Ullman, 2019; Shtulman & Carey, 2007; Shtulman & Morgan, 2017). Probing LM representations of the same categories 1) can reveal whether LM representations of modal categories are concordant with human intuitions about these categories, and 2) can potentially be used to evaluate whether an LM encodes aspects of the real world that are related to event plausibility.

However, recent work has questioned the ability of LMs to distinguish between modal categories, arguing that their sensitivity to surface-level features makes LM probability estimates a poor indicator of a sentence's modal category (Kauf et al., 2023; Michaelov et al., 2025). This is not unexpected, as a wide variety of considerations, aside from modal category, factor into next-token probability judgments (McCoy et al., 2024). The key question remains: are modal categories represented as coherent features in their own right *within* the LM, or do LMs only represent modality implicitly through unreliable probability estimates?

While Kauf et al. (2023) provides some preliminary indication that models may contain such internal representations using a probing analysis, we substantially elaborate on these results by expanding the range of modal categories and datasets under consideration. We analyze the development of internal representations of modal categories, relate these internal representations to human participants' categorization behavior, and show how we can interpret them in terms of human-understandable features. Specifically, we aim to answer the following research questions:

RQ 1 *Do LMs have internal representations of modal categories that go above and beyond merely representing output probabilities?*

RQ 2 *How do LM representations of modal categories develop **a)** over the course of training **b)** over consecutive layers **c)** as model size increases?*

RQ 3 *Do LM representations of modal categories reflect fine-grained human categorization decisions?*

RQ 4 *What interpretable features do these representations correspond to?*

To foreshadow our results, we find that LMs often do contain linear representations of the difference between modal categories, or modal difference vectors, which can be used to classify stimuli drawn from a variety of existing datasets. Modal difference vectors are often more discriminative than the probability of the sentence (RQ 1). We find that modal difference vectors emerge in an intuitive order, with more obvious categorical distinctions emerging earlier in training/layers/scale than more fine-grained distinctions (RQ 2). For both evaluations, we rely on expert labels of modal categories. However, human participants' intuitions of modal categories is not bound to reflect such expert labels. Thus, to address RQ 3, we analyze human behavioral categorization data, finding that projections of sentences onto modal difference vectors yield a feature space that reflects human categorization distributions. Intuitively, this feature space clusters stimuli by modal category, stimuli that lie in between clusters engenders greater disagreement among human participants. Finally, we find that some modal difference vectors correspond to interpretable sets of features, such as subjective event likelihood or imageability (RQ 4).

Overall, our results provide evidence that LMs learn to represent the difference between real life and mere fantasy to a greater extent than implied by previous research. Additionally, these representations appear to capture nuanced aspects of human categorization. These results raise the exciting possibility that one might use an LM's representations of modal categories to gain deeper insight into how and whether they have encoded the causal principles that underlie our world, while retaining the ability to imagine hypothetical realities[1].

2 BACKGROUND

Modal Categories The modal category of a statement describes whether that statement could, could not, or must be true (Mallozzi et al., 2024). The investigation of these categories has a long history in philosophy, where important arguments hinge on the validity of modal statements (Hume, 1739; Kripke, 1980; Gendler & Hawthorne, 2002). For example, the modal premise that *"philosophical zombies are conceivable"* can yield the modal conclusion that *"it is possible that the mind is distinct from the body"* (Chalmers, 1997; Descartes, 1641). Modal categories have a shorter (though still substantial) history in the cognitive sciences, where researchers have extensively studied the modal intuitions of children and adults (Shtulman & Carey, 2007). By probing their intuitions about the

[1] Code available here.

(im)possibility of different scenarios, one can obtain a nuanced picture of a participant's intuitive theories about the causal structure of the world (Griffiths, 2015; Shtulman & Morgan, 2017). For example, participants' intuitions regarding the difficulty of magical spells tend to be proportional to how much that spell violates their intuitive theories of physics (e.g., conjuring a frog out of nothing would be more difficult than teleporting a frog) (McCoy & Ullman, 2019). Following previous computational studies of modality (Kauf et al., 2023; Hu et al., 2025a;b), we study the following modal categories:

> **Probable**: Scenarios that are both possible and commonplace in our world. E.g., *chilling a drink using ice*
>
> **Improbable**: Scenarios that are possible, but not commonplace in our world. E.g., *chilling a drink using snow*
>
> **Impossible**: Scenarios that are not possible in our world, as they violate some known law of nature (e.g., physics, biology, etc.). These scenarios might be true in a counterfactual world with different laws of nature. E.g., *chilling a drink using fire*
>
> **Inconceivable**: Scenarios that cannot be evaluated for possibility in any possible world, due to some fundamental semantic or conceptual error (Hu et al., 2025b). We study inconceivable sentences that arise due to selectional restriction violations — unmet requirements that a verb imposes on its arguments (e.g., animacy, concreteness, etc.) (Chomsky, 1965; Katz & Fodor, 1963). E.g., *chilling a drink using yesterday*

Related Work The present study relates to a burgeoning literature investigating world models in LMs — underlying sets of causal principles that the LM encodes to represent and make inferences about the world (Mitchell, 2025). LMs have shown reasonably strong capabilities in encoding the state of a simplified or abstract world (Nanda et al., 2023; Kim & Schuster, 2023; Li et al., 2025; Ivanitskiy et al., 2023), but have struggled when presented with more complex worlds (Vafa et al., 2024). However, a world model must be far richer than a representation of states — it must represent the principles that explain the dynamics and constraints of the world (Ha & Schmidhuber, 2018; Ivanitskiy et al., 2023; Milliere & Buckner, 2024). This literature directly connects to previous work behaviorally assessing LMs' commonsense reasoning capabilities, which implicitly or explicitly assess an LM's understanding of such basic causal principles (Levesque et al., 2012; Zellers et al., 2019; Bisk et al., 2020; Ivanova et al., 2024).

Similar phenomena are studied in the cognitive sciences, where researchers investigate the world models of children and adults through their *intuitive theories* of physics, psychology, and other domains (Carey, 2000; Spelke & Kinzler, 2007; Ullman, 2015). Rather than being complete and precise representations of physical laws, these theories comprise the basic principles that human beings use to make sense of the world around them. Notably, these intuitive theories are imperfect, leading to a variety of incorrect physical inferences (Ullman et al., 2017). However, they are sufficient for operating in the world under normal circumstances. The causal principles comprising intuitive theories directly determine human intuitions about modal categories: probable and improbable events are consistent with these principles, impossible events are inconsistent with these principles, and inconceivable events might violate the basic conceptual presuppositions underlying these principles (Sosa & Ullman, 2022; Hu et al., 2025b). In this work, we analyze LMs' representations of modal categories to gain insight into the world models they have encoded.

3 Study 1: LMs Linearly Represent Modal Categories

In this section, we address RQ 1 by first identifying modal difference vectors — linear representations that distinguish between modal categories — from one dataset, and then assessing whether modal difference vectors can be used to classify stimuli from other datasets. We compare this method to classification based on sentence probability (among other baselines), and find that modal difference vectors enable more reliable modal categorization.

3.1 Methods

Datasets We use the Hu et al. (2025b) dataset to identify modal difference vectors, as it contains minimal pairs of stimuli for all pairs of modal categories under consideration. Notably, the impossible stimuli belong to that modal category due to violations of physical laws (e.g., *Someone baked a cake inside a freezer.*). The inconceivable stimuli belong to that category because they violate selectional restrictions based on concreteness (e.g., *Someone baked a cake inside a sigh.*). We evaluate the identified modal difference vectors using three other datasets; the Goulding et al. (2024), Vega-Mendoza et al. (2021), and Kauf et al. (2023) datasets. These datasets represent different forms of generalization: the Goulding et al. (2024) dataset contains stimuli that are impossible due to other factors (e.g., biology: *Someone is about to be born with 2 wings.*), whereas the Vega-Mendoza et al. (2021) and Kauf et al. (2023) datasets contain sentences that are inconceivable due to animacy violations (e.g., *The laptop bought the teacher.*). The Vega-Mendoza et al. (2021) dataset contains adversarial sentence pairs, where an inconceivable stimulus contains semantically-related terms and an improbable stimulus contains semantically unrelated terms (e.g., *The scientific research was funded by the* $\{$*microscope/traveler*$\}$.). Finally, the Kauf et al. (2023) dataset contains lexically-adversarial sentence pairs, where the inconceivable and probable sentences contain the same words in a different order (e.g., *The teacher bought the laptop.*) Every dataset contains expert labels of the modal category of all sentences. See Table 1 for a comparison between the datasets used in this study.

Models We study a variety of models across scales and families including GPT2-$\{$Small, Medium, Large, XL$\}$, Llama-3.2-$\{$1B, 3B$\}$, OLMo-2-$\{$1B, 7B, 13B$\}$, and Gemma-2-$\{$2B, 9B$\}$.

Modal Difference Vectors We create linear representations of the difference between modal categories using Contrastive Activation Addition (CAA), a technique used to create linear representations of concepts (Panickssery et al., 2023). These representations are simply directions in the hidden state space of an LM that correspond to the difference between pairs of stimuli. Because we are concerned with the difference between modal categories, we call these linear representations modal difference vectors.

CAA creates modal difference vectors using pairs of stimuli, (x_+, x_-). x_+ expresses one category, and x_- expresses another category. These stimuli are given to an LLM, M, in separate inference passes, and representations of some token are extracted at a particular layer l. $r_+ = M_l(x_+)$, $r_- = M_l(x_-)$. Difference vectors for each pair are computed as $v = r_+ - r_-$. Modal difference vectors are estimated by averaging over many of these single-pair difference vectors. To classify held-out pairs of stimuli (x'_+, x'_-) using a modal difference vector $\bar{v}$, we simply check whether $x'_+ \cdot \bar{v}$ is larger than $x'_- \cdot \bar{v}$ (Marks & Tegmark, 2024). This is analogous to prior work that classifies stimulus pairs based on the overall probability of each sentence (Kauf et al., 2023; Michaelov et al., 2025).

Concretely, we create separate modal difference vectors for every unique pair of categories in $\{$probable, improbable, impossible, inconceivable$\}$ by taking the difference between representations of the final "." token at some layer. That layer is found independently for each pair of categories by 5-fold cross-validation, using the classification method described above. If there are ties, we select the median layer that achieved the best performance. After identifying the best layer, we recompute the modal difference vector over all minimal pairs in the Hu et al. (2025b) dataset. See Figure 1 for an illustration of creating modal vectors and classifying stimulus pairs with them.

Baselines We include three baseline classification methods: Probability, Principal Component, and Random. First, we compare to classification using probability estimates. We follow prior work (Kauf et al., 2023) by calculating sentence probability as the sum of the log-probability of each token in the sentence. We expect that $p(inconceivable) < p(impossible) < p(improbable) < p(probable)$ (Hu et al., 2025b). If a minimal pair exhibits this relationship between two stimuli corresponding to different modal categories, then we consider the model to be correct. Next, we compare to projections along principal components of the hidden states. We compute the first three principal components of the final token of all of the sentences in the WikiText validation corpus, for each layer of each model (Merity et al., 2016). We then run the same cross-validation procedure described above to find the principal component that best partitions stimuli for each pair of modal categories. Finally, we repeat this process with randomly sampled vectors from each layer.

3.2 Results

We present classification accuracies on all generalization datasets for all models with at least 2B parameters in Figure 2. We find a qualitative difference in generalization set performance between

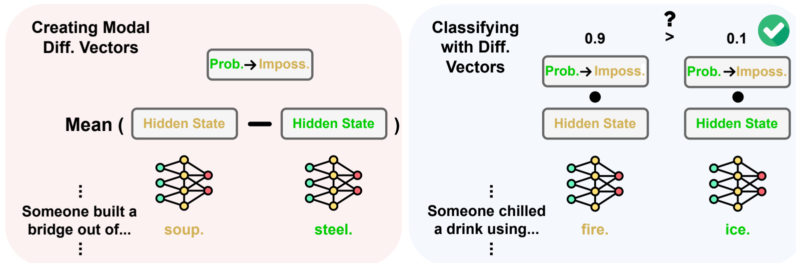

Figure 1: (Left) Diagram describing how we create modal difference vectors. In this example, a modal difference vector capturing the difference between probable and impossible stimuli is created by taking the mean over differences in hidden representations. (Right) Diagram describing how modal difference vectors are used to classify novel minimal pairs of impossible/probable sentences. Hidden representations from each sentence are projected on to the modal difference vector, and the magnitudes of these projections are compared.

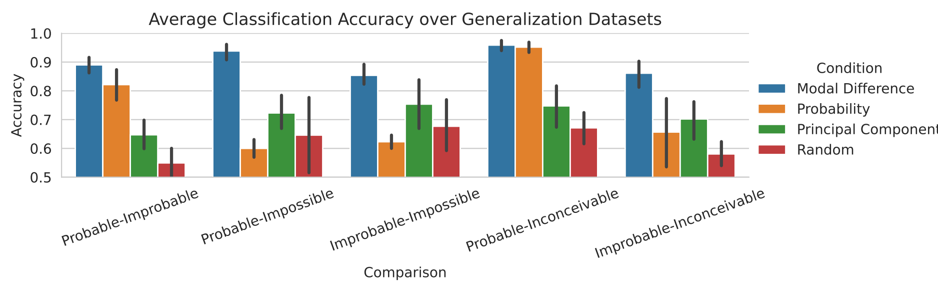

Figure 2: Classification evaluations for models with at least 2B parameters. Results are averages across models and generalization datasets. Modal difference vectors outperform probability estimates and other projection-based classification baselines.

models that are below this scale, and so we discuss those models separately (see Figure 3, Top Left). This is consistent with Kauf et al. (2023), who also noted that modal categorization ability correlated with scale. For models with at least 2B parameters, we see that for all pairs of modal categories present in the generalization datasets, modal difference vectors match or (sometimes drastically) outperform probability estimates at classifying stimuli by their modal category. Modal difference vectors similarly outperform other projection-based baselines. Notably, this result holds when just considering the adversarial stimuli (See Appendix C). The results are substantially less clear for models with <2B parameters (See Figure 6, Appendix B). In all cases, modal difference vectors perform worse for these models than they do for larger models. Additionally, probability estimates sometimes result in higher classification accuracy than modal difference vectors. Unless otherwise noted, we will proceed by analyzing only models with at least 2B parameters.

One might worry that these modal difference vectors are epiphenomenal (i.e., not causally implicated in model behavior). To address these concerns, we provide preliminary evidence that one can successfully steer model generations using modal difference vectors in Appendix D.

4 Study 2: The Development of Modal Categories

Prior human studies have found that the ability to distinguish modal categories arises gradually throughout development, with younger children struggling to distinguish between improbable and impossible events (Shtulman & Carey, 2007; Shtulman, 2009). In this section, we address RQ 2 by characterizing how modal difference vectors emerge as a function of various forms of model "development": model scale, layer depth, and training steps. Concordant with prior work (Saxe et al., 2019; Fel et al., 2024), we find that more coarse distinctions emerge in smaller models, in shallower layers, and earlier than more fine-grained distinctions.

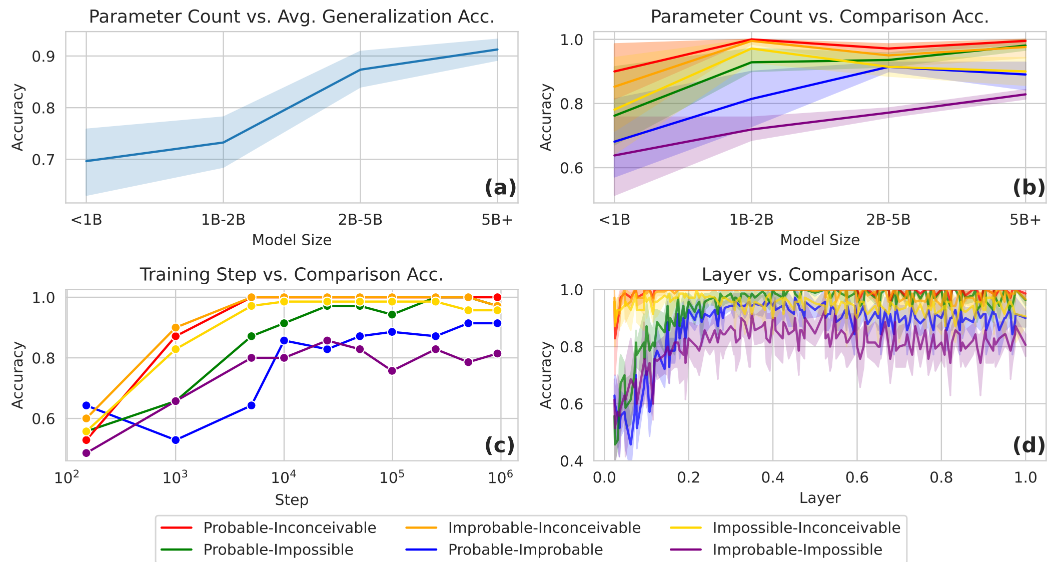

Figure 3: (a) Average generalization performance vs. parameter count reveals a large performance gap between models with fewer/greater than 2B parameters. At (b) smaller scales, (c) earlier in training, and (d) in earlier layers, models form modal difference vectors that can differentiate inconceivable stimuli from other modal categories. After that, models learn the distinction between **probable and impossible**, then **probable and improbable**, and finally **improbable and impossible**.

4.1 Methods

Models For scaling analyses, we study the complete set of models described in Section 3. For layer depth analysis, we analyze all models with at least 2B parameters. For analyzing training dynamics, we follow Hu et al. (2025b) and study OLMo-2-7B, as this model is accompanied by regular checkpoints throughout training.

Datasets We limit our focus to the Hu et al. (2025b) dataset for this section, as it contains all pairs of modal categories, with no adversarial pairs.

Analysis For analyzing model scale and training dynamics, we run the 5-fold cross-validation pipeline described in Section 3 and report cross-validation performance on the best layer. When analyzing layer depth, we report cross-validation performance for each layer.

4.2 Results

We present results from model scale, layer depth, and training dynamics analyses in Figure 3. We find that there is a qualitative difference in model performance on generalization datasets between models with <2B parameters and those with ≥2B parameters, as mentioned in Section 3. Aside from that, we find that models first distinguish between the inconceivable modal category from the rest, perhaps by taking advantage of distributional signals of selectional restriction violations (Kauf et al., 2023). Models next distinguish probable and impossible events, then probable and improbable, and finally improbable and impossible. Intuitively, we find that more coarse-grained modal distinctions are available earlier than more fine-grained modal distinctions. This pattern of results replicates and expands the results found in Hu et al. (2025b), which analyzes the surprisal assigned to sentences expressing different modal categories over training. We identify the same pattern, except in terms of internal representations. Hu et al. (2025b) also investigates whether the surprisal assigned to sentences expressing different modal categories changes as a function of parameter count. Whereas that work does not identify substantial differences with scale, we find that internal representations develop as a function of parameter count.

5 STUDY 3: MODELING HUMAN CATEGORIZATION BEHAVIOR

In all previous analyses, we assumed that the modal category assigned to a stimulus by researchers is the gold-standard label. However, we know that modal categories are graded (Shtulman & Morgan, 2017; Hu et al., 2025a;b), and categorization appears to rely on fuzzy intuitive theories (McCoy & Ullman, 2019). In this study, we address RQ 3 by analyzing whether modal difference vectors can be used to capture human participants' categorization behavior, which does not precisely reflect expert labels.

5.1 METHODS

Datasets We use human categorization data from Hu et al. (2025b), Goulding et al. (2024), and Hu et al. (2025a). Hu et al. (2025b) contains categorization data into all four modal categories under study. Goulding et al. (2024) contains data from adult and children participants categorizing probable, improbable, or impossible stimuli as either possible or impossible. We analyze the adult classification data. Hu et al. (2025a) contains data from adult participants categorizing probable and inconceivable sentences as either "total nonsense" or "not total nonsense". We subset the data to only include stimuli that were classified by four or more participants.

Analysis We wish to model human participant's modal categorization intuitions at the stimulus level. To do this, we fit logistic regression models to predict the response distribution of a population of human participants tasked with categorizing a scenario by its modal category (e.g., the proportions of human participants that labeled a scenario as "probable", "improbable", "impossible", and "inconceivable").

To derive a feature space for the logistic regression models, we take all stimuli within a dataset and project them onto three modal difference vectors: probable-improbable, improbable-impossible, and impossible-inconceivable. These vectors are chosen to minimize collinearity between the projections. This defines a 3-dimensional feature space, where stimuli are represented as points within this space. We train logistic regressions to predict the full response distribution for each stimulus using this feature space. Logistic regressions are trained using full batch gradient descent with cross-entropy loss using soft labels. We use the Adam optimizer for 200 epochs Kingma & Ba (2017) and a learning rate of 0.01. We provide a qualitative visualization of this feature space in Figure 4 (Left).

We use leave-one-out cross-validation to predict the response distribution of each scenario. We then compute several metrics to characterize the relationship between predicted and empirical response distributions.

Baselines We use the same set of baselines as in Section 3: summed log-probability, projections along principal components, projections along random vectors. We use each of these methods to generate feature spaces on which to fit logistic regression models. Notably, summed log probability only naturally provides a 1-dimensional feature space. However, the other two baselines provide 3-dimensional feature spaces. For each of these feature spaces, we follow the exact same procedure as described above to model the human data.

5.2 RESULTS

In Figure 4 (Right), we present several analyses characterizing how the different logistic regression models fit the human data. First, we compute the overall correlation between the empirical and predicted response distributions. Specifically, we report the correlation between the empirical and predicted probabilities assigned to $N-1$ of the categories for each stimulus, where N is the number of classes (as there are $N-1$ degrees of freedom in each distribution). This correlation provides a coarse measure of the relationship between probability distributions — a fairly high value might be achieved by merely correctly predicting the response distribution for stimuli that clearly belong to one modal category. Thus, we also characterize the averaged mean squared error between predicted and empirical response distributions. Finally, we correlate the entropy of empirical and predicted response distributions. Across all analyses, we find that a feature space derived from modal difference vectors routinely outperforms the baselines. Additionally, we provide qualitative examples of stimuli and

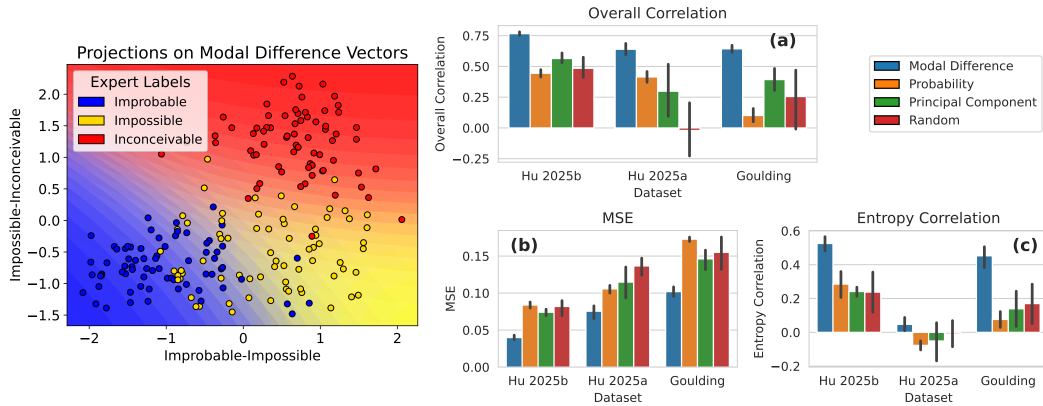

Figure 4: (Left) A qualitative example of stimuli from Hu et al. (2025b) projected along two modal difference vectors. Dots are colored according to their expert label. Background color intensity represents the probability that each point belongs to a particular class according to a logistic regression model fit to this subset of data using these two features. (Right) (a) Pearson correlation between the predicted probability distributions and the empirical proportion of participants that selected each category. (b) Mean squared error between predicted and empirical response distributions. (c) Pearson correlation between the entropy of predicted and empirical response distributions. In all analyses, we find that featurizing using projections along modal difference vectors leads to better models of human categorization behavior.

their predicted and empirical response distributions in Table 4. Overall, we find that modal difference vectors reflect participants' graded, intuitive notions of these categories better than the baselines.

6 Study 4: Interpreting Linear Representations

One benefit of identifying model-internal representations of modal categories is that they can be directly analyzed in order to understand which features drive modal categorization in LMs. In this exploratory study, we investigate RQ 4 by correlating projections of sentences along modal difference vectors with human ratings of these same sentences along a variety of interpretable dimensions. We study projections onto the three vectors used in Section 5. Interpreting these vectors might elucidate the relationship between the different modal categories, which is currently an open question (Hu et al., 2025b).

6.1 Methods

Datasets We use human ratings from Hu et al. (2025b), Hu et al. (2025a), and Tuckute et al. (2024). Hu et al. (2025b) contains human participant's ratings of the subjective event likelihood (i.e., "how probable is a scenario?") on a Likert scale for all sentences in the Hu et al. (2025b) dataset (**Event Likelihood** in Figure 5). Hu et al. (2025a) contains 12 sentences from Hu et al. (2025a) annotated according to their average rank in a forced-ranking version of the same subjective event likelihood task (**Ranked Inconceivability** in Figure 5). Additionally, Tuckute et al. (2024) contains 2000 short, diverse sentences, annotated on a Likert scale along a variety of dimensions, including how easy a sentence is to imagine, how grammatical the sentence is and whether the sentence contains a strong emotional valence. This dataset also contains various probability estimates from e.g., another LM or an N-gram model. See Appendix F for descriptions of each dimension.

6.2 Results

We project all sentences in each dataset onto the three modal difference vectors used in Section 5: probable-improbable, improbable-impossible, impossible-inconceivable. We then correlate these sentence projections with the human participants' annotations of interpretable features, as discussed above.

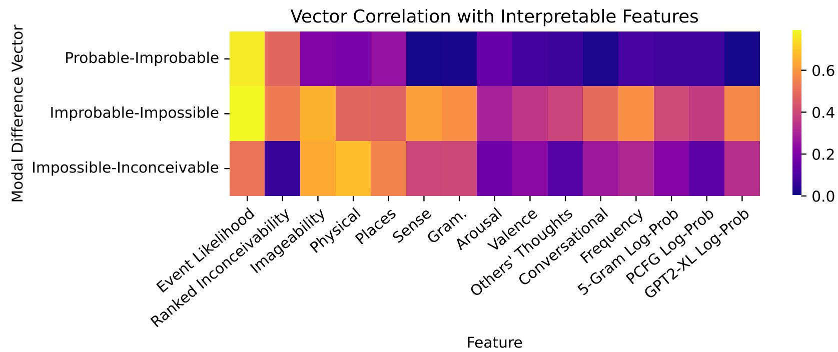

Figure 5: Absolute correlations between projections along modal difference vectors and interpretable features (averaged over models). Notably, Probable-Improbable correlates with human subjective event likelihood judgments, and Impossible-Inconceivable correlates selectively with imageability, the presence of physical objects, and places/environments.

We present the results of this analysis in Figure 5. Reassuringly, we find that projections along the probable-improbable vector correlate very well with subjective event likelihood, but less well with any of the other dimensions. Projections along the improbable-impossible vector are correlated with several features, including subjective event likelihood, imagability, whether the sentence makes sense, and grammaticality. This dispersion of correlations makes it harder to interpret exactly what distinguishes possible from impossible scenarios.

Most interestingly, we see that projections along the impossible-inconceivable vector correlate selectively with features that measure whether a sentence is easy to imagine (either directly or indirectly). This suggests that the ability to imagine a scenario might be a crucial ingredient in distinguishing impossible from inconceivable events. Notably, imagination has been empirically investigated as a factor in distinguishing impossible from possible scenarios (Shtulman & Carey, 2007; Lane et al., 2016; Tipper et al., 2024), but *not* as a mechanism for distinguishing impossible from inconceivable scenarios. However, this finding is consistent with a classic understanding of conceivability from philosophy (Hume, 1739; Yablo, 1993).

7 Discussion

We investigate how and whether LMs represent the modal category of a sentence within their hidden states. We find that (1) LMs form representations of modal categories, and that these representations are more diagnostic than output probability distributions (Section 3); (2) modal difference vectors develop at different points over the course of training, layers, and model size (Section 4); (3) modal difference vectors provide a feature space that reflects human categorization behavior (Section 5); and (4) modal difference vectors may reflect human-interpretable features (Section 6).

This investigation lays the foundation for a variety of future studies. First, modal difference vectors encoding the difference between, e.g., impossibility and improbability provide a direct means of testing the intuitive theories that LMs derive about the world from raw text input. For example, one might create a controlled dataset of sentences that instantiate different types of physics violations (similar to McCoy & Ullman (2019) or Ivanova et al. (2024)) and use modal difference vectors to check whether LMs represent each of these scenarios as impossible. This investigation could help reveal the physical constraints encoded by the LMs.

Finally, Sections 5 and 6 raise an exciting possibility: one might use modal difference vectors to generate hypotheses about human representations of modal categories. Section 5 establishes a correspondence between modal difference vectors and human categorization behavior, and Section 6 points to a specific, testable hypothesis: that humans distinguish between inconceivable and impossible events on the basis of imagination. Imagination has been shown to significantly impact adult's estimation of event likelihood (Koehler, 1991), and has been investigated as a strategy that adults and children use to distinguish improbable from impossible events (Shtulman & Carey, 2007; Lane et al., 2016; Tipper et al., 2024; Goulding et al., 2022). However, the role of imagination in discerning inconceivable from impossible events remains unknown.

REFERENCES

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

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.

Susan Carey. The origin of concepts. *Journal of Cognition and Development*, 1(1):37–41, 2000.

David J Chalmers. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks, 1997.

Kunming Cheng, Zhiyong Li, Qiang Guo, Zaijie Sun, Haiyang Wu, and Cheng Li. Emergency surgery in the era of artificial intelligence: Chatgpt could be the doctor's right-hand man. *International Journal of Surgery*, 109(6):1816–1818, 2023.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

René Descartes. Meditations on first philosophy. In *Seven masterpieces of philosophy*, pp. 63–108. Routledge, 1641.

Thomas Fel, Louis Bethune, Andrew Lampinen, Thomas Serre, and Katherine Hermann. Understanding visual feature reliance through the lens of complexity. *Advances in Neural Information Processing Systems*, 37:69888–69924, 2024.

Tamar Gendler and John Hawthorne. Introduction. conceivability and possibility, 2002.

Asma Ghandeharioun, Ann Yuan, Marius Guerard, Emily Reif, Michael Lepori, and Lucas Dixon. Who's asking? user personas and the mechanics of latent misalignment. *Advances in Neural Information Processing Systems*, 37:125967–126003, 2024.

Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pp. 25–30, 2013.

Brandon W Goulding, Emily Elizabeth Stonehouse, and Ori Friedman. Causal knowledge and children's possibility judgments. *Child Development*, 93(3):794–803, 2022.

Brandon W Goulding, Farishteh Khan, Keisuke Fukuda, Jonathan D Lane, and Samuel Ronfard. The development of modal intuitions: A test of two accounts. *Journal of Experimental Psychology: General*, 153(1):184, 2024.

Thomas L Griffiths. Revealing ontological commitments by magic. *Cognition*, 136:43–48, 2015.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pp. 363–364, 2019.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Jennifer Hu, Felix Sosa, and Tomer Ullman. Making sense of nonsense. In *Proceedings of the Cognitive Science Society*, October 2025a.

Jennifer Hu, Felix Sosa, and Tomer Ullman. Shades of zero: Distinguishing impossibility from inconceivability. *Journal of Memory and Language*, 143:104640, August 2025b. ISSN 0749-596X. doi: 10.1016/j.jml.2025.104640. URL `https://www.sciencedirect.com/science/article/pii/S0749596X25000336`.

David Hume. *A Treatise of Human Nature*. Oxford University Press, 1739.

Michael Igorevich Ivanitskiy, Alex F Spies, Tilman Räuker, Guillaume Corlouer, Chris Mathwin, Lucia Quirke, Can Rager, Rusheb Shah, Dan Valentine, Cecilia Diniz Behn, et al. Structured world representations in maze-solving transformers. *arXiv preprint arXiv:2312.02566*, 2023.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas Hikaru Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *CoRR*, 2024.

Jerrold J Katz and Jerry A Fodor. The structure of a semantic theory. *Language*, 39(2):170–210, 1963.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386, 2023.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL `https://arxiv.org/abs/1412.6980`.

Derek J Koehler. Explanation, imagination, and confidence in judgment. *Psychological bulletin*, 110 (3):499, 1991.

Saul Kripke. Naming and necessity, 1980.

Jonathan D Lane, Samuel Ronfard, Stéphane P Francioli, and Paul L Harris. Children's imagination and belief: Prone to flights of fancy or grounded in reality? *Cognition*, 152:127–140, 2016.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. *KR*, 2012:13th, 2012.

Belinda Z Li, Zifan Carl Guo, and Jacob Andreas. (how) do language models track state? *arXiv preprint arXiv:2503.02854*, 2025.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*, 2023.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 2024.

Antonella Mallozzi, Anand Vaidya, and Michael Wallner. The Epistemology of Modality. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

John McCoy and Tomer Ullman. Judgments of effort for magical violations of intuitive physics. *PloS one*, 14(5):e0217513, 2019.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

James Michaelov, R. Estacio, Z. Zhang, and B. K. Bergen. Not quite sherlock holmes: Pretrained language models cannot reliably differentiate impossible from improbable events. *Findings of the Association for Computational Linguistics*, 2025.

Raphael Milliere and Cameron Buckner. A philosophical introduction to language models - part i: Continuity with classic debates. *ArXiv*, abs/2401.03910, 2024. URL `https://api.semanticscholar.org/CorpusID:266844364`.

Melanie Mitchell. Llms and world models, part 1. *AI: A Guide for Thinking Humans*, 2025.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Andrew Shtulman. The development of possibility judgment within and across domains. *Cognitive Development*, 24:293–309, July 2009. doi: 10.1016/j.cogdev.2008.12.006.

Andrew Shtulman and Susan Carey. Improbable or impossible? how children reason about the possibility of extraordinary events. *Child development*, 78(3):1015–1032, 2007.

Andrew Shtulman and Caitlin Morgan. The explanatory structure of unexplainable events: Causal constraints on magical reasoning. *Psychonomic bulletin & review*, 24:1573–1585, 2017.

Felix A Sosa and Tomer Ullman. Type theory in human-like learning and inference. *arXiv preprint arXiv:2210.01634*, 2022.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Zoe Tipper, Terryn Kim, and Ori Friedman. Children (and many adults) use perceptual similarity to assess relative impossibility. *Developmental Psychology*, 2024.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.

Tomer David Ullman. *On the nature and origin of intuitive theories: learning, physics and psychology*. PhD thesis, Massachusetts Institute of Technology, 2015.

Keyon Vafa, Justin Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Marten van Schijndel, Andy Exley, and William Schuler. A model of language processing as hierarchic sequential prediction. *Topics in cognitive science*, 5(3):522–540, 2013.

Mariana Vega-Mendoza, Martin J Pickering, and Mante S Nieuwland. Concurrent use of animacy and event-knowledge during comprehension: Evidence from event-related potentials. *Neuropsychologia*, 152:107724, 2021.

Stephen Yablo. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research*, 53(1):1–42, 1993.

Table 1: Properties of the datasets analyzed in the study. **Prob.**, **Improb.**, **Imposs.**, **Inc.** denote whether the dataset contains probable, improbable, impossible, or inconceivable sentences, respectively. **Pair** denotes that the dataset contains minimal pairs that vary in modal category, allowing for classification using either vectors or probability estimation. **Adv.** indicates that the dataset contains pairs that are adversarial in some way (see main text). **Human** indicates that we analyze human behavioral data from this dataset.

Name	Prob.	Improb.	Imposs.	Inc.	Pair	Adv.	Human
Hu et al. (2025b)	✓	✓	✓	✓	✓	✗	✓
Goulding et al. (2024)	✓	✓	✓	✗	✓	✗	✓
Vega-Mendoza et al. (2021)	✓	✓	✗	✓	✓	Semantic	✗
Kauf et al. (2023)	✓	✓	✗	✓	✓	Lexical	✗
Hu et al. (2025a)	✓	✗	✗	✓	✗	✗	✓
Tuckute et al. (2024)	✗	✗	✗	✗	✗	✗	✓

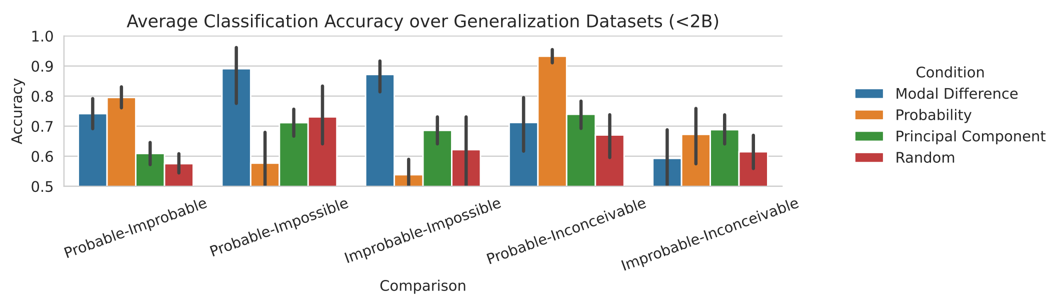

Figure 6: Classification evaluations for models with less than 2B parameters. Results are averages across models and generalization datasets. Results are mixed.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL `https://aclanthology.org/P19-1472/`.

A Dataset Comparison

Table 1 provides a reference for quick comparison of the datasets used in the current study.

B Classification Results for Models with <2B Parameters

We present classification results from models with <2B parameters in Figure 6. We find mixed results across different classification methods, with overall worse performance than with models ≥2B.

C Classification Results for Adversarial Stimuli

We highlight the performance of all classification methods on adversarial stimuli. We include lexically adversarial stimuli from Kauf et al. (2023) and semantically adversarial stimuli from Vega-Mendoza et al. (2021). Lexically adversarial stimuli contain sentence pairs with the same tokens, just in a different order (e.g., *The teacher bought the laptop/The laptop bought the teacher*). Semantically adversarial stimuli contain sentence pairs with an improbable stimulus containing a semantically-unrelated word and an impossible stimulus containing a semantically-related word (e.g., *the scientific research was funded by the* {*traveler/microscope*}. We find that modal difference vectors distinguish

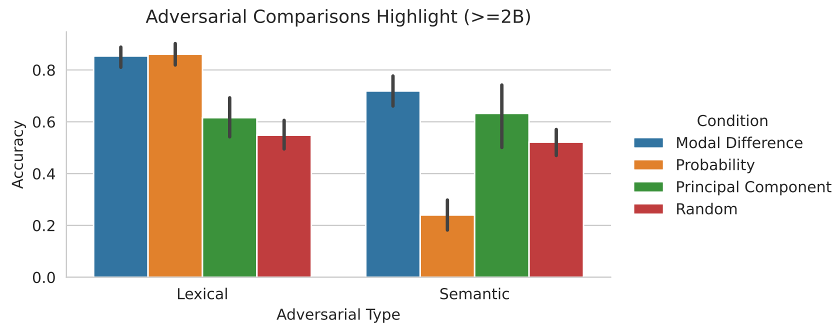

Figure 7: Performance of all classification methods just on adversarial stimuli. We include lexical adversarial stimuli from Kauf et al. (2023) on the left. These stimuli contain the same tokens, just in different orders. We find that modal difference vectors and probability estimates perform best in the face of this manipulation. We include semantically adversarial stimuli used in Vega-Mendoza et al. (2021) and Michaelov et al. (2025) on the right. We find that modal difference vectors and projections along principal components perform best, while probability estimates are systematically misled.

all of these cases reliably, whereas other methods distinguish at most one of these types of adversarial stimuli.

D Steering with Modal Difference Vectors

In this section, we demonstrate preliminary evidence that the modal difference vectors can be used to steer the generations of a language model to produce sentences expressing the intended modal category. We assess this using a manually-constructed corpus of 30 novel sentence prefixes. For each prefix and model, we generate continuations as follows: First, we generate the top 5 most likely next tokens. For each of these 5 continuations, we greedily decode 4 more tokens. However, we find that models sometimes generate fragments of run-on or syntactically complex sentences. To generate clean qualitative examples, we store the overall probability of the period token "." after each generation. We truncate generations after the token position where the period token received the highest probability.

Following prior work, we intervene on models using modal difference vectors while generating (Panickssery et al., 2023; Ghandeharioun et al., 2024). To do so, we add a scaled version of the modal difference vectors for probable-improbable, probable-impossible, or probable-inconceivable to all residual stream positions at the appropriate layer while generating the next token. We experiment with scalar multipliers of 3 and 5, and find that 5 qualitatively produces better results. As a baseline, we repeat this process with no intervention.

We attempt to steer Gemma-2-2B and Llama-3.2-3B. We present a quantitative analysis of steering success in Figure 8. Here, we use the baseline model (without intervention) to measure the surprisal (or negative log-probability) of the first 5 generated tokens either with or without steering. We find that, for both models, surprisal increases in order of `baseline` < `improbable` < `impossible` < `inconceivable`. This result mirrors the finding from Hu et al. (2025b), indicating that steering has the desired impact on model generations.

We also present several examples from each model on a diverse range of prefixes in Tables 2 and 3. While not perfect, these examples show many instances of steering having the desired effect, rendering generations more improbable, impossible, or inconceivable.

E Qualitative Examples of Stimuli and Model Predictions

See Table 4 for qualitative examples of stimuli and their predicted and empirical response distributions.

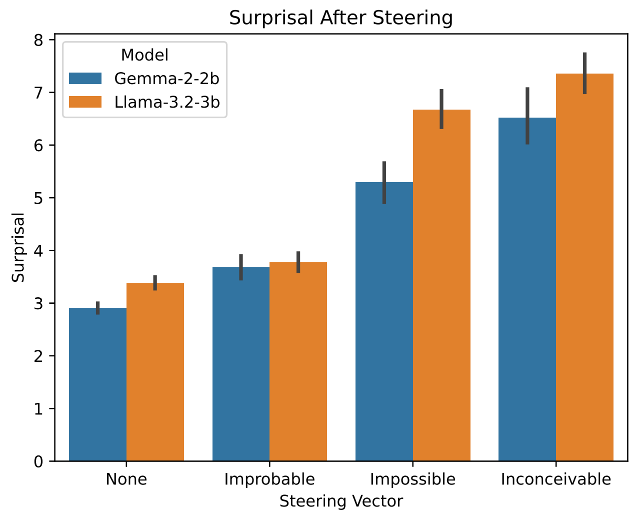

Figure 8: Surprisal (with respect to the baseline model) of the next token predictions generated after steering. We find that, for both Gemma-2-2B and Llama-3.2-3B, surprisal values generally increase in order of `baseline` < `improbable` < `impossible` < `inconceivable`. This mirrors the finding from Hu et al. (2025b).

F Descriptions of the Features in the Tuckute et al. (2024) Dataset

In this section, we briefly describe the features in the Tuckute et al. (2024) dataset, which we correlate with projections along modal difference vectors in Section 6.

Imageability: Ratings on a 7 point Likert scale, answering the question "how easy is the sentence to visualize, or to form an image of the sentence's meaning in your mind?"

Physical: Ratings on a 7 point Likert scale, answering the question "how does the sentence make you think of physical objects and/or physical causal interactions?"

Places: Ratings on a 7 point Likert scale, answering the question "how much does the sentence make you think of places, natural scenes, and/or environments?"

Sense: Ratings on a 7 point Likert scale, where 1 corresponds to "doesn't make any sense" and 7 corresponds to "makes perfect sense".

Gram.: Ratings on a 7 point Likert scale, where 1 corresponds to "completely ungrammatical" and 7 corresponds to "perfectly grammatical".

Arousal: Ratings on a 7 point Likert scale, answering the question "how much does the sentence make you feel stimulated, excited, frenzied, wide-awake, and/or aroused?"

Valence: Ratings on a 7 point Likert scale, answering the question "how much does the sentence make you feel happy, pleased, content, and/or hopeful?"

Others' Thoughts: Ratings on a 7 point Likert scale, answering the question "How much does the sentence make you think of other people's experiences, thoughts, beliefs, desires, and/or emotions?"

Conversational: Ratings on a 7 point Likert scale, answering the question "how likely do you think the sentence is to occur in a conversation between people?"

Frequency: Ratings on a 7 point Likert scale, answering the question "how likely do you think you are to encounter this sentence?"

5-Gram Log-Prob: Probability estimates from a 5-gram language model.

PCFG Log-Prob: Probability estimates from the PCFG parser from van Schijndel et al. (2013).

Table 2: Steering Generations using Gemma-2-2B.

Model	**Prefix**
Gemma-2-2B	Someone measured the furniture using a...
Steering	**Generations**
None	tape measure, ruler, 1, measuring tape, laser level
Improbable	tape measure, ruler, laser, measuring tape, 3D scanner
Impossible	laser, telescope, microscope, new technique, camera
Inconceivable	¡b¿¡i¿¡u¿, team, different scale, word, new one
Model	**Prefix**
Gemma-2-2B	Someone fed the child with a...
Steering	**Generations**
None	spoon, bottle of milk, knife, syringe, glass of water
Improbable	spoon, ¡strong¿¡em¿pizza, bowl of soup, baby bottle, pizza
Impossible	new song, baby, song, video game, snake
Inconceivable	smile, child, good thing, team, different kind of child
Model	**Prefix**
Gemma-2-2B	Someone fixed the car with a...
Steering	**Generations**
None	hammer and a screwdriver, new engine, lot of effort, 1, screwdriver
Improbable	hammer, 1, piece of wood, rubber band, sledgehammer
Impossible	new car, laser, 3D printer, car, broken mirror
Inconceivable	team, smile, good team, ¡em¿team¡/em¿, great team
Model	**Prefix**
Gemma-2-2B	Someone protected the garden with a...
Steering	**Generations**
None	fence, net, protective fence, plastic sheet, wire fence
Improbable	fence, 1, giant inflatable, blanket, net
Impossible	new song, cloud, virus, song, flower
Inconceivable	team, smile, family, new one, ¡em¿team¡/em¿

GPT2-XL Log-Prob: Probability estimates from GPT2-XL.

Table 3: Steering generations using Llama-3.2-3B

Model	**Prefix**
Llama-3.2-3B	Someone destroyed a building using a...
Steering	**Generations**
None	sledgehammer, bulldozer, wrecking ball, crane, drone
Improbable	flamethrower, bulldozer, sledgehammer, drone, chainsaw
Impossible	boat, car, 9, real-life, man and a woman
Inconceivable	term, thesium, weapon, gun, a
Model	**Prefix**
Llama-3.2-3B	Someone hit the ball with a...
Steering	**Generations**
None	bat, golf club, racquet, racket, stick
Improbable	golf club, bat, baseball bat, racquet, club
Impossible	stick, real, boat, golf course, bow
Inconceivable	hammer, bat, club, baseball, ball
Model	**Prefix**
Llama-3.2-3B	Someone beat a drum with a...
Steering	**Generations**
None	stick, drumstick, spoon, big stick, hammer
Improbable	stick, drumstick, spoon, sword, can of beer
Impossible	drum, man in a boat, French, cat, real drum
Inconceivable	hammer, heart, mind, head, brain
Model	**Prefix**
Llama-3.2-3B	Someone treated the wound using a...
Steering	**Generations**
None	topical antibiotic, solution of 0, technique, mixture of honey, tourniquet
Improbable	home remedy, mixture of honey, loe vera gel, local anesthetic, homemade remedy
Impossible	3, Japanese, bottle, French, real-life
Inconceivable	number, series, few, smile, The

Table 4: A set of qualitative examples from Goulding et al. (2024), showing logistic regression and human probability estimates. Probability that each scenario is possible, as assessed by the logistic regression model using projections on modal difference vectors from Gemma-2-9B, the logistic regression model using probability estimates from Gemma-2-9B, and the proportion of participants classifying the scenario as “possible”.

Scenario (*Someone is about to...*)	Modal Diff. P(Poss.)	Prob. P(Poss.)	Human P(Poss.)
clean a car.	0.99	0.70	1.0
clean a road.	0.94	0.62	0.97
clean a cloud.	0.09	0.57	0.05
stay awake for 5 hours.	0.94	0.59	1.0
stay awake for 5 days.	0.67	0.63	0.53
stay awake for 5 years.	0.25	0.60	0.05