Cross-Modal Taxonomic Generalization in (Vision-) Language Models

Tianyang Xu,[1] **Marcelo Sandoval-Castañeda,**[1]
Karen Livescu,[1] **Greg Shakhnarovich,**[1] **Kanishka Misra**[2]
[1]Toyota Technological Institute at Chicago, [2]The University of Texas at Austin
{sallyxu, marcelo, klivescu, gregory}@ttic.edu, kmisra@utexas.edu

Abstract

What is the interplay between semantic representations learned by language models (LM) from *surface form* alone to those learned from more grounded evidence? We study this question for a scenario where part of the input comes from a different modality—in our case, in a vision-language model (VLM), where a pretrained LM is aligned with a pretrained image encoder. As a case study, we focus on the task of predicting hypernyms of objects represented in images. We do so in a VLM setup where the image encoder and LM are kept frozen, and only the intermediate mappings are learned. We progressively deprive the VLM of explicit evidence for hypernyms, and test whether knowledge of hypernyms is recoverable from the LM. We find that the LMs we study can recover this knowledge and generalize even in the most extreme version of this experiment (when the model receives no evidence of a hypernym during training). Additional experiments suggest that this cross-modal taxonomic generalization persists under counterfactual image-label mappings only when the counterfactual data have high visual similarity within each category. Taken together, these findings suggest that cross-modal generalization in LMs arises as a result of both coherence in the extralinguistic input and knowledge derived from language cues.

https://github.com/sally-xu-42/cross-modal-taxonomic-gen

1 Introduction

The widespread empirical success of language models (LMs) in showing sophisticated behavior has inspired a number of philosophical investigations of how learning the distribution of language *form* contributes to the learning of *meaning*. Many such discussions (Piantadosi and Hill, 2022; Pavlick, 2023; Mollo and Millière, 2024; Mandelkern and Linzen, 2024) have converged into the broader viewpoint that LMs achieve a particular type of grounded meaning where the internal representation of a token is grounded in its *relation* to the representations of other tokens. This 'relational grounding' viewpoint (Mollo and Millière, 2024) echoes a number of historical perspectives that emphasize the emergence of meaning from connections between symbols (Harris, 1954; Firth, 1957; Harman, 1982; Harnad, 1990; Elman, 2004).

How does relational grounding from language interplay with other forms of meaning—e.g., those grounded in perceptual signals? For instance, LMs' knowledge of the concept *bird* might include its syntax (that it is a noun), that it is a hypernym of other concepts (*robin, sparrow*), and that it is thematically related to concepts like *trees, branches, nests, flying*, etc. How does this knowledge impact LMs' behavior when they encounter a bird via a different modality—e.g., its visual representation?

This question can be answered by analyzing models that learn to map between the representations in LMs and non-textual modalities such as vision or speech. Modern vision-language models (VLMs; Liu et al., 2023, a.o.)—models that learn to map between pretrained image encoders and language models—are a particularly popular class of such models. Recent work has found evidence that the LM components of such models often overpower their image encoder counterparts, to the extent that the model sometimes predicts a response without relying on the visual input (Golovanevsky et al., 2025; Lee et al., 2025; Fu et al., 2025; Frank and Allaway, 2025; Hua et al., 2025). While this has been interpreted as a bug, especially in perception-heavy tasks (Fu et al., 2024), it also highlights the extent to which certain kinds of knowledge can be acquired from language alone.[1]

[1]For instance, predicting that a dog has four legs given an image of a dog with three legs (Frank and Allaway, 2025) is indeed an issue, but can be attributed to the effect of *generic* knowledge (e.g., *Birds can fly*, *Lions have manes*; Leslie, 2008) that is highly prevalent in language data (Gelman et al., 2010).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16319–16337
July 2-7, 2026 ©2026 Association for Computational Linguistics

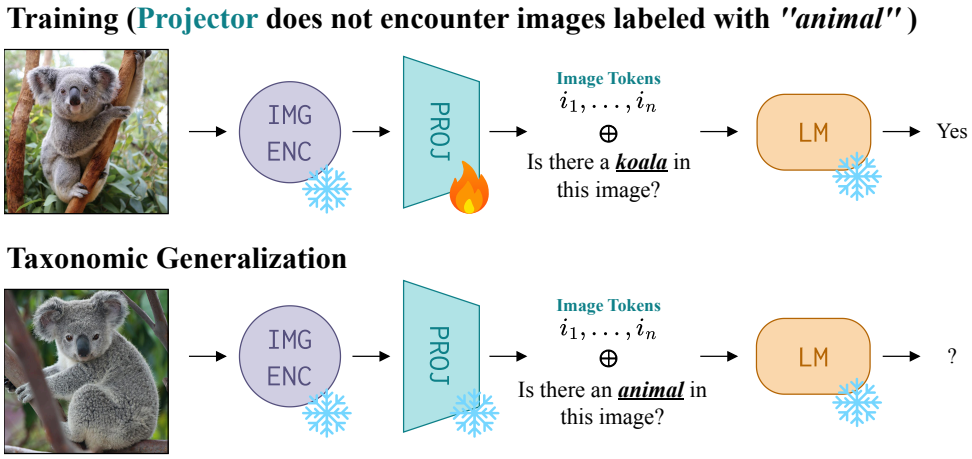

Figure 1: An instance of our experiments. During training, the projector is deprived of explicit supervision on high-level categories (*hypernyms*, e.g., *animal*) at various amounts, and is trained to detect the presence (and absence) of lower-level categories (e.g., *koala*), keeping the image encoder and the LM backbone frozen. After training, the VLM is tested for generalization to hypernym categories, given previously unseen images.

In this paper, we place particular emphasis on the latter takeaway, and take it a step further. We do so by systematically holding out a particular piece of knowledge during the training of VLMs, and test if it is recoverable directly from the language model's learned representations. Doing so allows us to answer the broader question of how knowledge learned *exclusively* from language extends beyond it and interacts with the input signals from an extra-linguistic modality (here, images). As a case study, we specifically focus on the ability of the models to perform **taxonomic generalization**: predicting the presence of a hypernym (higher-order category, e.g., *bird*, *animal*) in an input image, given that they have learned to identify *only* lower-level categories (e.g., *parrot, koala*) during training. To what extent does the taxonomic knowledge learned by LMs generalize *across modalities*?

Taxonomic knowledge is a worthwhile testing ground for our research question. Taxonomic organization is closely connected to categorization, a fundamental capacity of human cognition (Murphy, 2004)—learning an entity's taxonomic category allows us to make systematic, default inferences (e.g., if the entity is a bird, then it likely has wings). There have been a number of hypotheses about the learnability of category knowledge from linguistic cues, for both computational models (Hearst, 1992; Geffet and Dagan, 2005) and humans (Waxman and Markow, 1995; Wilson et al., 2023). Our work contributes to the study of how such category knowledge arises in computational models.

Hearst (1992) proposed a set of distributional patterns that strongly signal taxonomic relations—e.g., patterns like "*Xs such as Ys*" can serve as a cue to the inference that Y is a type of X. Similarly, the distributional inclusion hypothesis (Geffet and Dagan, 2005) states that hyponyms tend to appear in a subset of the distributional contexts in which the hypernyms can appear—e.g., *animal* can appear in all the contexts in which *dog* can appear and additionally those where it does not. Therefore, it is conceivable that LMs, being powerful models of distributional semantics, might pick up on these cues to learn the relations between hypernym categories and their members. More generally, language has been posited to be "particularly influential when the information it expresses is not easily derived through first-hand observation and experience" (Cimpian and Markman, 2008). Our analyses contribute towards in-principle tests of the broader claims of these hypotheses (that taxonomic knowledge is retrievable from language data), in the extreme condition where the source of the target knowledge is restricted to language alone and the input comes from a non-language modality (here, vision). So, if a learner (in this case, LMs) has made no explicit "first-hand" perceptual observations of *any* hypernym category, to what extent can the knowledge it has acquired from language support the visual identification of such categories?

Our experiments (depicted in Figure 1) focus on a class of VLMs that are trained by learning a 'projector' that maps image representations from a frozen image encoder to the embedding layer of a frozen pre-trained language model, similarly to Liu et al. (2023) and Merullo et al. (2023). We first train these models to answer questions that ask whether a

given object-centric image contains a category (e.g., *Is there a **cat** in this image?*). We then test whether they can obtain above-chance performance at identifying both leaf-level *and* hypernym categories on images that do not appear in training. Having established that they do, we then test whether they can generalize their knowledge of hypernyms *without* explicit vision-language supervision—i.e., only from their pretrained language representations. We do so by training additional instances of the projector while holding out various amounts of hypernyms. This experimental setup is inspired by recent work on controlled rearing or filtered-corpus training of LMs (Jumelet et al., 2021; Misra and Mahowald, 2024; Patil et al., 2024; Leong and Linzen, 2024; Yao et al., 2025), where LMs are trained on manipulated versions of input corpora and tested on their ability to generalize. In the most extreme version of this experiment, where there is **no evidence of a single hypernym** during the projector's training, we find that VLMs still achieve above-chance performance. That is, given an image of a *sparrow*, VLMs are able to predict the presence of a *bird*, without ever encountering image-question pairs containing the word "bird". This suggests that taxonomic knowledge is represented within LMs in a manner that can extend beyond modalities—i.e., **they can perform cross-modal taxonomic generalization**.

What is the role of the input signal in facilitating this generalization? Do LMs generalize arbitrarily, in a rule-like manner? For instance, if the projector were to be trained *counterfactually* to predict "cardinal" when it was an image of *hummus* and "parrot" when it was an image of a *bandage*, would it still predict that these images contain *birds* (`IF crow THEN bird`)? This is not borne out in our findings. Models perform at chance under conditions where lexical items within a hypernym category are paired with randomly shuffled sets of images. Instead, they only generalize in counterfactual conditions that preserve the visual coherence of categories—i.e., when we shuffle images and leaf-level mapping *within* the same category (e.g., *cardinal* mapped to images of *eagles* and *parrot* mapped to images of *pheasants*). This finding gives us preliminary insight about the **importance of input coherence** (Murphy and Medin, 1985) **in order for cross-modal generalizations to emerge**, paving the way for future investigations across multiple modalities and phenomena.

2 Related Work

Taxonomic knowledge (or hypernymy) is one of the most well-studied lexical relations in computational linguistics (Murphy, 2003), and has long been hypothesized to be extractable from naturally occurring language (Hearst, 1992; Geffet and Dagan, 2005; Roller and Erk, 2016). This has been especially apparent in recent investigations of hypernymy in pretrained LMs (Hanna and Mareček, 2021; Moskvoretskii et al., 2024; Regneri et al., 2024, *a.o.*). While hypernymy is also captured by VLMs (Qin et al., 2025), the extent to which it can *possibly* emerge from their LM component alone is unclear. Our methods shed light on this, since we restrict the knowledge of hypernymy to the LM backbone, and use image encoders that have not been supervised using text data.

Our findings also relate to recent proposals about the convergence in the representations of models trained on different modalities or domains (Cao and Yamins, 2021; Sorscher et al., 2022; Li et al., 2024; Huh et al., 2024; Wu et al., 2025). For instance, according to the 'Platonic Representation Hypothesis' (Huh et al., 2024), the representations learned by unimodal models of different modalities are converging since they are statistical models of the same world. We shed light on the implications of these proposals when two unimodal models (the vision encoder and the LM) are aligned. When this alignment happens under counterfactual scenarios (§4.3), where the incoming input signals are at odds with the conceptual organization of the two models—posited to be similar under the aforementioned proposals—we see a severe drop in taxonomic generalization. This underscores the importance of such representational convergence in the extreme setting where explicit information about the phenomenon (here, hypernymy) is restricted to a model from one modality.

3 Data, Models, and Methods

3.1 Stimuli and Measures

We conduct our experiments using the THINGS database of labeled, object-centric images (Hebart et al., 2019). Specifically, we use a 1,216-category and 17,336-image subset of THINGS made available by Rodriguez et al. (2025), who also provide hypernym annotations. The leaf categories are grouped under 53 hypernyms, with each leaf category belonging to 1.67 hypernyms on average (max = 6, min = 1). Our task stimuli consist of

English polar questions with the format "*Is there a* `{category}` *in this image?*", where we substitute "`{category}`" with the target category. Each image in our experiments is associated with two types of questions: a positive question, asking for the presence of the category depicted in the image (expecting the answer 'Yes'), and a negative question, asking for the presence of a different category (expecting the answer 'No'). For negative questions, we substitute the leaf/hypernym category in the question by sampling uniformly from other leaf/hypernym categories in our set—we do this at the image level, which means that the same positive question is associated with multiple negative questions throughout our experiments. For example, as shown in Figure 1, the same test image of a *koala* is associated with "*Is there a koala in this image?*" as well as "*Is there an animal in this image?*" as its positive questions. Since each leaf-level category is associated with multiple images, we split the set of images for each category into training (70%), validation (5%), and test (25%) sets. Detailed statistics are available in §B. Because we conduct negative sampling at the image level, the resulting number of "yes" and "no" instances is not balanced within each leaf/hypernym—e.g., in 100% ablation test set, the *mammal* category has 332 and 149 "yes" and "no" questions, respectively. We therefore use macro F1 per category as our evaluation measure.

3.2 Modeling Choices

Our models follow the general visually-conditioned LM architecture used by many recent studies (Liu et al., 2023; Wang et al., 2024; Chen et al., 2023, a.o). Specifically they include three components: 1) an **image encoder**, which takes in an image and produces a sequence of N image feature vectors, each corresponding to an image patch, and is pre-trained using standard self-supervision objectives; 2) a vision-language **projector**, which is a multi-layer perceptron that transforms the image features into image "tokens", which are a sequence of vectors with the same dimensionality as the language model's embedding layer; and 3) an **LM backbone**, which is pre-trained on text data, and takes in a sequence of image tokens, followed by the task question, and (in our case) produces an answer—either "yes" or "no". During training, the image encoder and the LM backbone are kept frozen and only the projector is trained, similarly to Merullo et al. (2023). We do this to preserve the information contained in the pre-trained components. Our training objective is next-word prediction, as in standard VLM fine-tuning—we minimize the negative log probability of the correct answer token (yes/no) given an image and a question, and compute the loss only at the answer position.

For our image encoder, we primarily experiment with DINOv2 (large; Oquab et al., 2024) since it is trained using a self-supervised learning objective on images, and has never encountered text data (not even class labels). We do this to ensure that the image encoder representations are devoid of any text supervision, which may leak information about hypernyms even when we remove them during the training of the projector. Nevertheless, we compare performance obtained using DINOv2 to that obtained using SigLIP (Zhai et al., 2023), a CLIP-style encoder trained using a contrastive objective to map between images and text, in order to test whether the presence/absence of text during image encoder training has any effect on our models' taxonomic generalization.

For our LM backbone, we primarily report results using publicly-released Qwen3 LMs (Yang et al., 2025) from HuggingFace (Wolf et al., 2020), specifically ones with 0.6B and 1.7B parameters. We also report results using Llama 3.2 models (Grattafiori et al., 2024) in §D. In addition, we compare VLMs using these pre-trained LMs to their randomly-initialized counterparts, in order to quantify the effect of text exposure on cross-modal taxonomic generalization. Details about our training hyperparameters are listed in §A.

3.3 Experiment Design

Since the goal of our experiments is to test whether hypernymy relations can be recovered by the LM backbone of our model, we always include all leaf categories and their training set images in the projector training data, and manipulate only the number of hypernym category labels shown during training. We perform two kinds of manipulations:

Random Hypernym Ablation We randomly remove mappings between leaf images and their hypernym labels from the training set. By removing a mapping (x, h) between leaf concept x and hypernym h, we mean that we remove all image-question pairs with images of x and questions about h. For example, removing the mapping *(parrot,bird)* corresponds to removing all images of parrots with the question "Is this a bird?" (as well as with corresponding negative questions like "Is this a tool?").

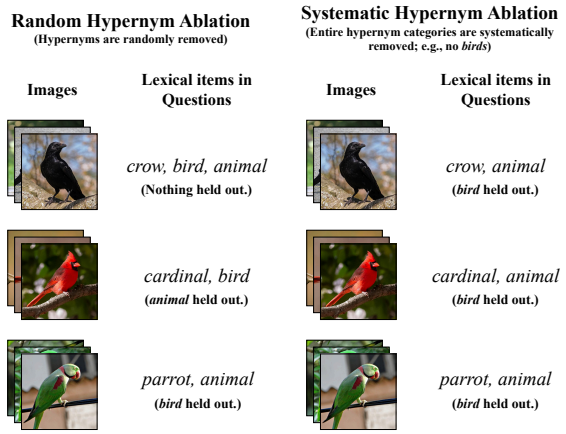

Figure 2: Depiction of our two ablations, for a sampled toy training set containing several images each of crows, cardinals, and parrots.

As a result, a model sees only partial evidence for a given hypernym—e.g, it sees the label *bird* for all images of *crow*, but not for *parrot*. For each of the 53 hypernym labels in the training set, we perform this ablation by removing 10%, 30%, 50%, 70%, 90%, or 100% of the total possible leaf-image sets.

Systematic Hypernym Ablation We remove a specified number of hypernyms from the training data entirely. Here a model sees only a subset of the unique hypernyms—e.g., it might encounter *animal* and *vegetable*, but never encounter *bird* and *vehicle*. We perform this ablation by removing 10 (19%), 20 (37%), 30 (57%), 40 (75%), and all 53 (100%) hypernyms from the positive training examples, leaving only their leaf-level labels. A removed hypernym may still appear as a negative category in questions paired with other hypernyms.

Figure 2 shows examples of both ablations. Note that at 100%, the two ablation types are equivalent. Appendix B shows a detailed breakdown of our splits at various ablated percentages.

4 Experiments

4.1 Preconditions to Generalization

Before analyzing cross-modal taxonomic generalization, we first test the extent to which our model satisfies three preconditions. First, the LM backbone should have non-trivial knowledge about hypernymy—e.g., it should predict that *animal* is a hypernym of *dog, koala, sheep, giraffe,* etc. Next, the model should achieve non-trivial taxonomic generalization when the vision encoder has *not* been informed using any text-based pretraining (like in CLIP, see Radford et al., 2021). Finally, the model should sufficiently learn the task of identifying categories in images, especially those that were held out from training. Below we report results from all three of these tests.

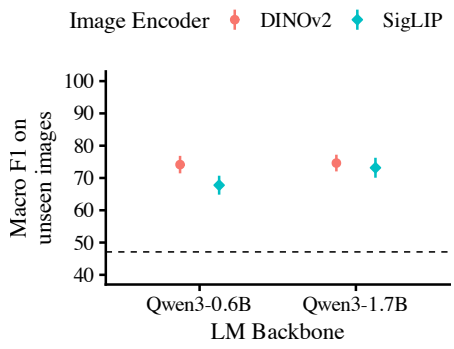

Figure 3: Macro F1s of VLMs for predicting the hypernym category for unseen images across LM backbones and image encoder types (DINOv2 vs. SigLIP), in the experiment where all hypernyms are removed. Dashed line indicates majority-label baseline.

Hypernymy knowledge in LMs Our first test focuses on the extent to which our LM backbones—namely, Qwen3-0.6B and Qwen3-1.7B—capture basic hypernymy relations. Arguably, this is the most critical precondition: LMs must possess the knowledge of the phenomenon we are trying to measure across modalities. We verify this for the set of categories in our dataset, by constructing text-only questions analogous to the subset of hypernym-focusing questions in our data (§3.1). That is, for a question that asks "*Is there a/an* `categoryX` *in this image?*" given an image of `categoryY`, we create an analogous question: "*Is it true that a/an* `categoryY` *is a type of* `categoryX`*?*" This template is taken directly from Qin et al. (2025), and results in a total of 16,903 questions (2,040 positive questions, and 14,863 negative questions). This imbalance is inherited from our main stimuli, as negative sampling was performed at the image level. For each question, we compare the LM's probabilities for "yes" vs. "no", and evaluate using macro F1 scores. We find that both Qwen3 LMs achieve high F1 scores: 78.5 (0.6B) and 88.5 (1.7B), both of which are substantially greater than the F1 of a majority-label baseline (46.7).

Impact of text supervision in image encoders A presupposition made by our research question is that it is only the LM backbone that is the source of textual/language exposure for the model. The extent to which this presupposition holds depends on our choice of image encoder—while DINOv2 was trained with self-supervision *only* on vision

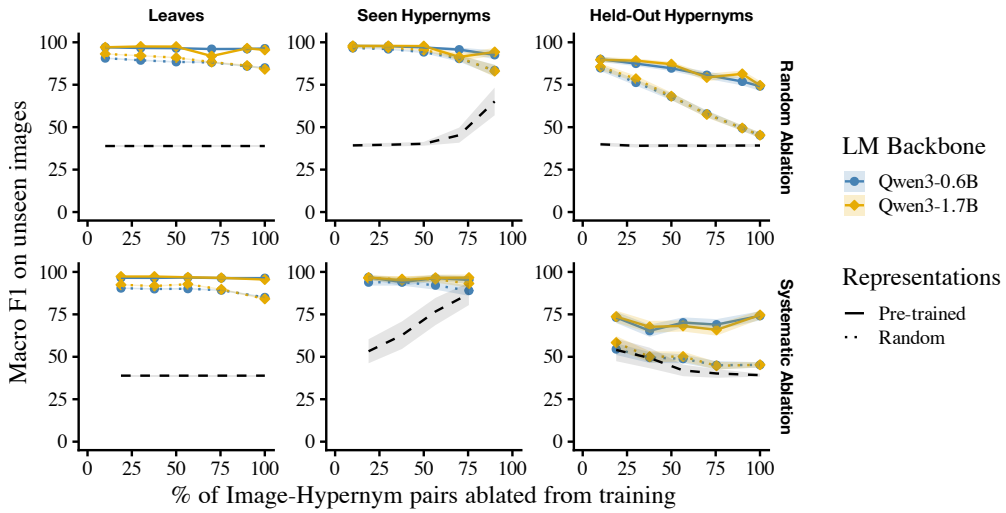

Figure 4: Macro F1 on unseen images across LM backbones, LM representations (pre-trained vs. random), and hypernym ablation type (Random vs. Systematic) at different amounts of exposure to hypernym categories, for various test splits. Dashed line indicates the majority-label baseline. Error bands represent 95% confidence intervals across random seeds (N = 3) and categories (hypernyms = 53, leaves = 1216).

tasks (presupposition holds), SigLIP's loss function uses both text and image signals (presupposition does not hold). What is the impact of the image encoder's exposure to text on models' taxonomic generalization? To test this, we compare the macro F1s of models with fixed LM backbones, differing only in their image encoders (DINOv2 and SigLIP), on the specific sub-experiment where the model has not received *any* exposure to a hypernym during VLM training (i.e., the strictest version of our experiments). Figure 3 shows these results for the 0.6B and 1.7B versions of the Qwen3 LM. We use a linear mixed effects regression model to predict the macro F1 using image encoder and LM backbone as fixed effects, with random slopes for run seed and hypernym category. We do not find a significant effect of the image encoder ($t = 0.91$, $p = 0.45$). Since performance obtained using DINOv2 as our encoder is statistically equivalent to that obtained using SigLIP, we report all our subsequent results in the main text using DINOv2.

Do our models learn the task? Our final precondition is that the model learns to identify the categories in training in a generalizable manner—i.e., it does not simply memorize the training set and can identify leaf (and hypernym, when applicable) categories even when conditioned on images not seen during training. The two leftmost columns in Figure 4 (labeled "Leaves" and "Seen Hypernyms") show macro F1s of models for predicting the presence of leaf categories as well as hypernyms seen during training, on unseen images. We see that all models, including ones with randomly initialized LM backbone representations, achieve very high macro F1 scores at all levels of hypernym exposure, for all ablation types. That is, our models satisfy the precondition of learning the task.

4.2 Cross-Modal Taxonomic Generalization

Having shown that the LM backbones capture hypernymy, that the trained VLMs can generalize to unseen images regardless of whether or not the image encoder receives textual supervision, and that they have learned to perform the task, we now turn to our main result, shown in the rightmost column of Figure 4, named "Held-Out Hypernyms". We see that regardless of the amount of exposure to a hypernym category, LM backbones with pre-trained representations achieve consistently above-chance performance in predicting the presence of hypernyms in images not seen during the models' training. This is observed for both types of ablations, and even at 100% ablation, where the model never sees *any* hypernym label during training. Models with randomly initialized LM backbone representations consistently underperform their pre-trained counterparts, and are close to chance in the absence of any hypernym-level supervision. These results also hold for Llama 3.2, which we report in Figure 10 in the Appendix. These results suggest that the pre-trained representations of LMs *can* facilitate cross-modal taxonomic generalization. We analyze how this generalization varies across depths of taxonomic organization in §F.

4.3 On the Arbitrariness of Cross-Modal Taxonomic Generalization

The results thus far show that LM representations can facilitate cross-modal taxonomic generalization—i.e., our models can predict that an image of a crow contains a bird, even though they have never encountered an instance of an image being labeled with "*bird*". How abitrary is this behavior? Do LMs simply execute the rule, "`IF crow THEN bird`?" regardless of what the image-*crow* mapping is—e.g., would they still predict that there are birds in images of *kayak* and *hummus*, if during training they encountered images of a *kayak* labeled with *crow*, and *hummus* with *cardinal*? Or, do LM representations expect members of a category to *cohere* in systematic ways (Murphy and Medin, 1985)? For instance, members of the *bird* category are likely to share a multitude of features—*beaks*, *wings*, *feathers*, etc.—and this could foster the development of a cohesive category in the visual representations of its members, which could be the key to facilitating taxonomic category-based generalization within the model.

To adjudicate between these possibilities, we repeat our experiments with two counterfactual versions of our training sets, each aiming to isolate a possible mechanism (arbitrary, rule-based taxonomic generalization, vs. taxonomic generalization facilitated by visual coherence). We create these counterfactual datasets by shuffling the mapping between sets of images and their corresponding leaf labels (as well as hypernyms, in a non-zero hypernym ablation). Our first counterfactual dataset, **Across-category shuffle**, is created by randomly shuffling sets of images and leaves in a manner that destroys visual coherence—e.g., *crow* is mapped to images of *kayak*, *cardinal* is mapped to images of *hummus*, etc., such that members of birds are mapped to visual representations of items that belong to completely different categories. Our second counterfactual dataset, **Within-category shuffle**, is created by randomly shuffling the image-leaf pairings *within* categories. In this dataset, the same set of images represent the bird category as they do in reality, but their labels are counterfactual to the real world. For example, *crow* is mapped to images of *penguin*, *cardinal* is mapped to images of *eagle*, etc. In this way, the internal visual coherence within each category is preserved. Figure 5 shows examples of each type of shuffle, in comparison to the original configuration, and §C shows descriptions

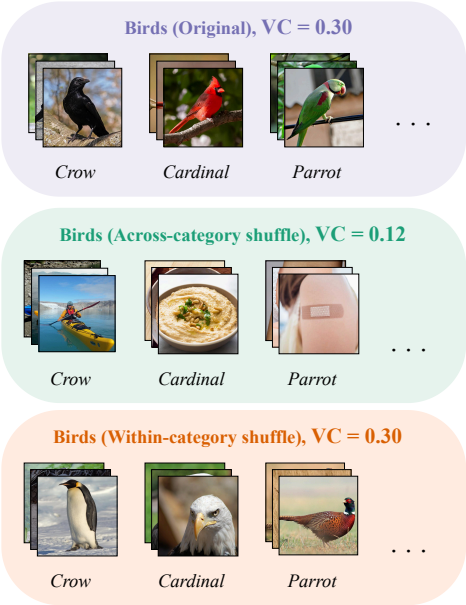

Figure 5: Examples of image-leaf mappings resulting from our counterfactual shuffles, in comparison with the original configuration (top). VC indicates the visual coherence of the category under the data configuration.

of the algorithms used to implement these shuffles.

To quantify visual coherence, we extract DINOv2 embeddings for all training images using mean pooling over image patch tokens, then calculate average pairwise cosine similarities of image embeddings within each category—the higher this value, the higher the visual coherence of that category. We find the average visual coherence of the original and across-category shuffle to be 0.27 (*SD*=.07) and 0.12 (*SD*=.02), respectively (the coherence for within-category shuffle is identical to that of the original by definition). This corroborates our intuition about reduced coherence in the across-category shuffle, relative to the original and within-category configurations.

In both of these counterfactual shuffles, we preserve the language-based lexical-semantic relationships, and manipulate only the visual coherence of the categories—the category *bird* still contains the same lexical items (*crow, cardinal, parrot,* etc.). This experiment is analogous in spirit to recent work testing the extent to which LMs learn "impossible" languages (Kallini et al., 2024; Xu et al., 2025)—e.g., LMs struggle to learn languages whose sequences are created by randomly shuffling sequences from actual human languages. Analogously, we shuffle data in a manner that reduces the internal visual similarities of categories—e.g., the bird category contains images of *kayaks*, *hummus*, *bandages*, etc. This makes it difficult to conceive of

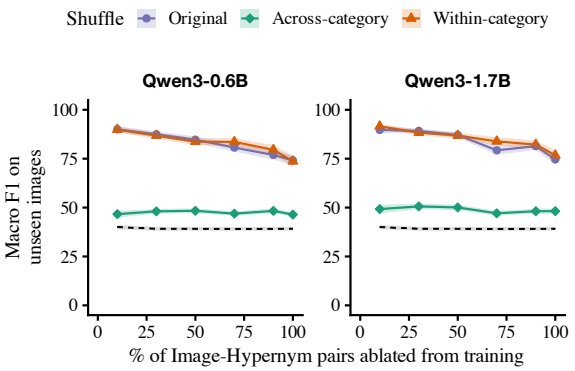

Figure 6: Macro F1 of models trained on data with different counterfactual shuffles (or lack thereof) at predicting unseen hypernyms, across various amounts of exposure to hypernyms. Dashed line indicates macro F1 of the majority-label baseline. Error bands represent 95% confidence intervals across three random seeds and 53 hypernym categories.

any conceptual structure along which these would cohere (Murphy and Medin, 1985).

If the LM is simply executing an abstract rule of `IF crow THEN bird`, without showing any sensitivity to the coherence among its input representations, then we should see no difference in its taxonomic generalization *regardless* of the type of shuffle. In this case, both **Among** and **Within** category shuffles should show similar performance to the original experiment. On the other hand, if LMs *are* sensitive to the coherence among categories in their inputs, then we should see models trained under the **Within-category shuffle** show similar performance to the model trained on the original data, while those trained on the **Across-category shuffle** data should be much worse.

Figure 6 shows our results, specifically for the Random Hypernym Ablation sub-experiment. We see that for both LM backbones, models whose projector was trained on data under the **Across-category shuffle** substantially underperform models trained on data under the **Within-category shuffle**, which is instead nearly indistinguishable from models trained with the original configuration. These results also hold for the Llama 3.2 models, as shown in Figure 11 in §D. We also see similar results with Qwen3-8B as the backbone (at 100% ablation, see Table 5 in §E), suggesting that arbitrary taxonomic reasoning does not emerge even in a larger model. Together these results lend support to the hypothesis that cross-modal taxonomic generalization is facilitated when the input representations of category members cohere in a

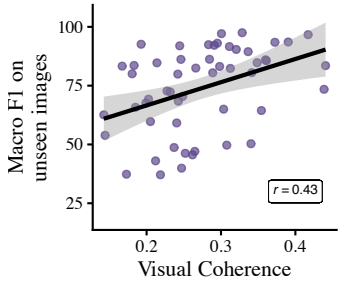

Figure 7: Macro F1 scores on unseen images vs. Visual coherence across the 53 hypernym categories for the Qwen3-1.7B backbone (at 100% ablation). r indicates Pearson's correlation.

systematic way, relative to the hypothesis that LMs use their learned knowledge of lexical relations in an arbitrary, rule-based manner.

5 Post-hoc Analyses

Hypernym-specific results Are our models more likely to show cross-modal generalization for some hypernyms than others? There are two possibilities that can shed light on this question: 1) the LM backbones have stronger hypernymy knowledge for certain hypernyms; and 2) certain hypernyms are more visually coherent. Note that these two are not mutually exclusive (nor are they exhaustive)—it is possible that both have *some* impact on models' hypernym-specific performance. We tease apart their effects by conducting linear mixed-effect model analyses on results from the 100% ablation subset of the original experiment. Specifically, we predict the models' macro F1 scores using the LMs' hypernymy F1 scores and the visual coherence of the hypernym as fixed effects, and the LM backbone (including different seeds) and the hypernym as random effects.

We find a significant positive effect of visual coherence (t = 3.338, $p < .01$), but no effect of the LM backbone F1 scores (t = 0.047, p = .962). Figure 7 shows an example of the positive relation between visual coherence and cross-modal generalization performance. Overall, these results suggest that cross-modal generalization is greater for hypernyms that are more visually coherent.

Shuffling entire hypernyms To further analyze the effect of visual coherence, we conduct a toy experiment on 100% hypernym ablation in which shuffling is applied at the hypernym level. Specifically, we swap leaf labels across a pair of hypernyms so that, for instance, the members of *bird* are now paired with images of *vegetables*. In this

way, the visual coherence of a given hypernym is still the same, but they are now derived from the images of a different category. The different number of leaves across hypernyms makes a full hypernym-level swap infeasible; thus we focus on hypernyms with at least 20 leaves and uniformly sample 20 from each (which reduces the data scale and thus limits direct comparability with our shuffling experiments in §4.3). We then select 5 pairs of hypernyms whose average visual coherence scores differ by less than 0.02 (dessert–furniture, vehicle–vegetable, bird–school supply, clothing accessory–medical equipment, weapon–container). Finally, we randomly exchange the leaf labels within each selected pair. For fair comparison, we run an analogous experiment using the same data but without the hypernym-level shuffle.

Both models achieve generally high F1 scores on the leaf-level data—93.8 and 95.2, respectively, for the unshuffled and the shuffled experiments. On held-out hypernyms, the models obtain F1 scores of 57.5 and 51.5, respectively, for the unshuffled and shuffled cases, both of which are above chance (here, 34.7 F1). A t-test reveals no significant difference between the unshuffled and shuffled experiments' results (t = 1.30, p = .19). Although these F1 scores are not as high as in our original shuffling experiments, shuffling at the hypernym level (while maintaining visual coherence) results in similar performance as in the unshuffled case, suggesting a non-zero effect of visual coherence.

6 Discussion

Our results point to two primary takeaways. First, given exposure only to lower level categories in a new modality (here, vision), LMs can generalize to higher-level categories. That is, they show cross-modal taxonomic generalization. Second, the emergence of this generalization depends on the extent to which the mapping between images and lower-level categories is visually coherent. That is, models' cross-modal taxonomic generalization is not realized in an arbitrary manner, and instead is sensitive to the coherence of their input signal.

These findings reinforce and extend our understanding of how meaning obtained from relational grounding in language (Abdou et al., 2021; Patel and Pavlick, 2022; Mollo and Millière, 2024) interacts with those that are obtained from more perceptual signals. Past work in this space has largely shown that models demonstrate reasoning about perceptually relevant concepts by only providing them as text input. For instance, Patel and Pavlick (2022) showed that LMs like GPT-3 (Brown et al., 2020) generalize concepts like spatial relations, cardinal directions, and colors to counterfactual worlds (represented as text) by encoding *how* concepts are related. That is, LMs could predict which direction was *south*, given partial access to cardinal directions even in rotated worlds, presumably because they encoded the relationship between *north* and *south* (unchanged by rotation), indicating strong relational grounding.

At first glance, our second takeaway may seem to conflict with these results—after all, if models encode that *bird* and *animal* are hypernyms of *crow*, then they should generalize to these hypernyms regardless of the counterfactual condition. However, a closer comparison between our counterfactual data creation and that of Patel and Pavlick suggests that the two conclusions are in fact compatible. Specifically, the counterfactual mappings of Patel and Pavlick are *structure-preserving*—it is not the case that the opposite of *east* is *south* in their counterfactual worlds, and therefore the ways in which directions *cohere* are not changed. In comparison, our **Across-category shuffle** is *not* structure preserving—e.g., *crow* is mapped to images of *kayaks*, and *cardinal* is mapped to images of *hummus*, which means that the members of the bird category are far apart within the input representational space (i.e., visual features). The **Within-category shuffle**, on the other hand, *is* structure preserving. That is, the set of image representations of the members of a given category is unchanged—only the labels are shuffled. We found models trained under this counterfactual scenario to show similar performance as those trained on the original data.

Overall, our results lend preliminary support to a more general hypothesis about the importance of *input coherence* in facilitating cross-modal generalization. This is *preliminary* because our current conception of coherence is based on image features, whereas what makes categories coherent has been posited to rely on more general conceptual knowledge and intuitive theories of the world (Murphy and Medin, 1985; Gopnik and Wellman, 1994; Kemp and Tenenbaum, 2009; Murphy, 2004). Nonetheless, we expect our methodology to inspire a thorough exploration of this connection between categorical coherence and generalization across modalities and phenomena in the future.

Limitations

Lack of naturalistic VLM training We use a simple visual question answering-like task for our experiments, one that explicitly focuses on the phenomenon of interest. While this allows us to isolate model behavior, VLMs are typically trained on multiple tasks, including but not limited to image captioning (Liu et al., 2023; Wang et al., 2024; Yang et al., 2025; Chen et al., 2023; Deitke et al., 2025). This can raise questions about the implications of our findings for practical VLM training. At the same time, it would have been infeasible to repeatedly train models on large image captioning datasets, considering the large number of training runs in this work.

Single pair of modalities and single phenomenon Another way in which our results and takeaways are restricted is that our experiments are limited to a single pair of modalities and focus on one phenomenon (hypernymy). In our preliminary experiments, we tried to investigate generalization to spatial relations—e.g., can models learn to answer questions about the relation *above* without ever seeing it explicitly in training, presumably by using *below* as indirect evidence? However, our models failed to satisfy the important precondition of learning the task in the first place, corroborating recent evidence about the brittle performance of VLMs on spatial reasoning tasks (Kamath et al., 2023; Chen et al., 2025; Zhang et al., 2025). This made it impossible to test for any generalization, since there was no evidence that the model had learned anything in the first place. We expect that as visual encoders and VLMs become more capable in general, our methodology will be useful for studying broader claims about cross-modal generalization to other domains (Lewis et al., 2019; Lupyan et al., 2020; Liu et al., 2025; Wang et al., 2026).

Model scale Our experiments are limited to models of certain (modest) sizes. It is possible that different generalization behaviors emerge at different model sizes. Studying the scaling behavior of cross-modal generalization would be an interesting direction for future work.

Single language Finally, our stimuli are limited to the English language. Extensions of this work to multi-lingual settings would lend further evidence about the role of relational grounding from language. One particularly exciting extension could be to train for leaf-level information on one language and test for hypernymy knowledge in another—i.e., to test for cross-modal cross-lingual taxonomic generalization.

Acknowledgments

We are grateful to Jessica Yang, Zhewei Sun, Andrew Lampinen, Martin Zettersten, Gary Lupyan, Phillip Isola, and Ellie Pavlick for helpful suggestions and questions. We also thank the three anonymous reviewers for their feedback. Kanishka Misra is supported by the Donald D. Harrington Faculty Fellowship at UT Austin.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*.

Lawrence W Barsalou. 1985. Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of experimental psychology: learning, memory, and cognition*, 11(4):629.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.

Rosa Cao and Daniel Yamins. 2021. Explanatory Models in Neuroscience: Part 2 – Constraint-based Intelligibility. *arXiv preprint arXiv:2104.01489*.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. 2025. Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas. In *Forty-second International Conference on Machine Learning*.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, Daniel Salz, Xi Xiong, Daniel Vlasic, Filip Pavetic, Keran Rong, Tianli Yu, Daniel Keysers, Xiaohua Zhai, and Radu Soricut. 2023. PaLI-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

Andrei Cimpian and Ellen M Markman. 2008. Preschool children's use of cues to generic meaning. *Cognition*, 107(1):19–53.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Jeffrey L Elman. 2004. An Alternative View of the Mental Lexicon. *Trends in Cognitive Sciences*, 8(7):301–306.

John Firth. 1957. A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, 6(4):10–32.

Stella Frank and Emily Allaway. 2025. VISaGE: Understanding Visual Generics and Exceptions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. Hidden in Plain Sight: VLMs Overlook their Visual Representations. In *Second Conference on Language Modeling*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: Multimodal Large Language Models Can See But Not Perceive. In *ECCV*.

Maayan Geffet and Ido Dagan. 2005. The Distributional Inclusion Hypotheses and Lexical Entailment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Susan A Gelman, Elizabeth A Ware, and Felicia Kleinberg. 2010. Effects of Generic Language on Category Content and Structure. *Cognitive Psychology*, 61(3):273–301.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. 2025. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *EMNLP*.

Alison Gopnik and Henry M Wellman. 1994. The Theory Theory. *Mapping the Mind: Domain Specificity in Cognition and Culture*, page 257.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models.

Michael Hanna and David Mareček. 2021. Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Gilbert Harman. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic*, 23(2):242–256.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):1–24.

Tianze Hua, Tian Yun, and Ellie Pavlick. 2025. How do Vision-Language Models Process Conflicting Information across Modalities? *arXiv preprint arXiv:2507.01790*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with Vision-Language Models? Investigating their Struggle with Spatial Reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. In *Forty-first International Conference on Machine Learning*.

Charles Kemp and Joshua B Tenenbaum. 2009. Structured Statistical Models of Inductive Reasoning. *Psychological Review*, 116(1):20.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. VLind-Bench: Measuring Language Priors in Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Cara Su-Yi Leong and Tal Linzen. 2024. Testing Learning Hypotheses using Neural Networks by Manipulating Learning Data. *arXiv preprint arXiv:2407.04593*.

Sarah-Jane Leslie. 2008. Generics: Cognition and Acquisition. *Philosophical Review*, 117(1):1–47.

Molly Lewis, Martin Zettersten, and Gary Lupyan. 2019. Distributional Semantics as a Source of Visual Knowledge. *Proceedings of the National Academy of Sciences*, 116(39):19237–19238.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. 2024. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Qiawen Liu, Jeroen van Paridon, and Gary Lupyan. 2025. Learning about Color from Language. *Communications Psychology*, 3(1):60.

Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark. 2020. Effects of Language on Visual Perception. *Trends in Cognitive Sciences*, 24(11):930–944.

Matthew Mandelkern and Tal Linzen. 2024. Do Language Models' Words Refer? *Computational Linguistics*, 50(3):1191–1200.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations*.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Dimitri Coelho Mollo and Raphaël Millière. 2024. The Vector Grounding Problem. *Philosophy and the Mind Sciences (Forthcoming)*.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. Are Large Language Models good at lexical semantics? A case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

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

Gregory L Murphy and Douglas L Medin. 1985. The Role of Theories in Conceptual Coherence. *Psychological Review*, 92(3):289.

M Lynne Murphy. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*. Cambridge University Press.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered Corpus Training (FICT) shows that Language Models can Generalize from Indirect Evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Ellie Pavlick. 2023. Symbols and Grounding in Large Language Models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

Steven Piantadosi and Felix Hill. 2022. Meaning without Reference in Large Language Models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. 2025. Vision-and-Language Training Helps Deploy Taxonomic Knowledge but Does Not Fundamentally Alter It. In *Advances in Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Michaela Regneri, Alhassan Abdelhalim, and Soeren Laue. 2024. Detecting Conceptual Abstraction in LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Juan Diego Rodriguez, Aaron Mueller, and Kanishka Misra. 2025. Characterizing the role of similarity in the property inferences of language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Stephen Roller and Katrin Erk. 2016. Relations such as Hypernymy: Identifying and Exploiting Hearst patterns in Distributional Vectors for Lexical Entailment.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. 2022. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43):e2200800119.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Ziwen Wang, Akshi, Sophia Keil, Judy Kim, and Marina Bedny. 2026. Constructing Meaning from Language: Visual knowledge in People Born Blind and in Large Language Models. *Annual Review of Linguistics*, 12:In press.

Sandra R Waxman and Dana B Markow. 1995. Words as Invitations to form Categories: Evidence from 12-to 13-month-old Infants. *Cognitive Psychology*, 29(3):257–302.

Kyra Wilson, Michael C Frank, and Abdellah Fourtassi. 2023. Conceptual Hierarchy in Child-directed Speech: Implicit Cues are more Reliable. *Journal of Cognition and Development*, 24(4):563–580.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's Transformers: State-of-the-art natural language processing.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can Language Models Learn Typologically Implausible Languages? *arXiv preprint arXiv:2502.12317*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Qing Yao, Kanishka Misra, Leonie Weissweiler, and Kyle Mahowald. 2025. Both direct and indirect evidence contribute to dative alternation preferences in language models. In *Second Conference on Language Modeling*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. Do Vision-Language Models Represent Space and How? Evaluating Spatial Frame of Reference under Ambiguities. In *The Thirteenth International Conference on Learning Representations*.

A Training Details

We use the code implementation of VLM training from Karamcheti et al. (2024), as it allows for flexible changes to the image encoder and LM backbone components. We adopt the hyperparameters in Table 1 for all our experiments. Model training is conducted on a GPU cluster using 4 RTX A6000/RTX 6000 Ada/L40S/RTX Pro 6000 GPUs, depending on availability.

Hyperparameter	**Value**
Epochs	5
Batch Size	16
Number of Layers	2
Input Dimension	1024/1152
Output Dimension	1024/2048
Max Gradient Norm	1.0
Weight Decay	0.1
Learning Rate	2×10^{-5}
Activation Function	GeLU
Optimizer	AdamW
Scheduler	Warmup & Cosine Decay
Warmup Ratio	0.03

Table 1: Training hyperparameters used in our experiments. We adopt a 2-layer multi-layer perceptron architecture with GeLU activation for all vision-language projectors. The input dimension of our trained MLP projectors is 1024 for DINOv2 and 1152 for SigLIP. The output dimension of our trained MLP projectors is 1024 for Qwen3-0.6B, and 2048 for all other LM backbones. The two MLP layers share the same output dimension across experiments.

B Details of Experimental Data

B.1 Data Use and Release Policy

We use images from the THINGS database for our experimental data (Hebart et al., 2019). These images were released with the Attribution CC BY license.[2] In our release, we point to their file names

[2] See https://osf.io/jum2f/files/52wrx

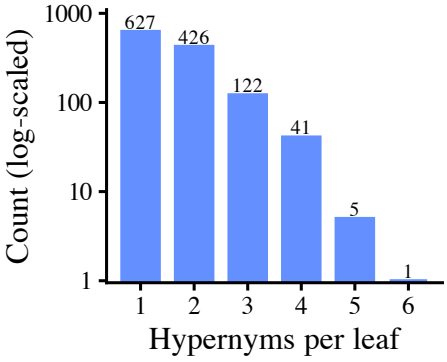

Figure 8: Distribution of the number of hypernym categories per leaf in our dataset. Counts (y-axis) are log-scaled.

and do not release their images. We release our data alongside our code on github,[3] which uses an MIT License.

B.2 Data Statistics

Figure 8 shows the distribution of the number of hypernyms per leaf, and Figure 9 shows the number of leaves across all 53 hypernyms in the dataset.

Table 2 reports the dataset statistics under five levels of random hypernym ablation and three data variants. For each ablation level, we show the number of image-question pairs in the training, validation, and test sets, with validation and test sets further divided into leaf-level, seen-hypernym, and unseen-hypernym subsets. The *original* variant corresponds to the unshuffled dataset. The *across-category* and *within-category* variants are obtained by shuffling hypernym annotations across categories or within the same category, with details listed in §4.3 and §C. Across all ablation levels and data variants, the images in the test set are held fixed.

Table 3 reports dataset statistics under systematic hypernym ablation at four ablation levels. In contrast to random ablation, systematic hypernym ablation removes entire hypernym categories from the training data.

Table 4 lists the specific hypernym categories removed at each systematic ablation level. These category lists are used consistently across all experiments.

[3] https://github.com/sally-xu-42/cross-modal-taxonomic-gen

C Shuffling Algorithms

This appendix provides detailed descriptions of the algorithms used to create the counterfactual datasets.

C.1 Across-category Shuffle

Algorithm 1 describes the procedure for creating the across-category shuffle, which randomly maps original leaf-level concepts to new leaf-level concepts, regardless of their hypernym categories.

Algorithm 1: Across-category Shuffle

Input: $\mathcal{H}$: hypernym-to-leaf mapping
Output: $\mathcal{M}$: concept-to-concept shuffle mapping
Function `ShuffleConcepts(`$\mathcal{H}$`)`**:**
  $\mathcal{C} \leftarrow \emptyset$;
  foreach *leaf list* $\mathcal{L} \in \text{values}(\mathcal{H})$ **do**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{L}$;
  $\mathcal{C}_{\text{shuffled}} \leftarrow \text{RandomShuffle}(\mathcal{C})$;
  $\mathcal{M} \leftarrow \emptyset$;
  for $i = 1$ **to** $|\mathcal{C}|$ **do**
    $\mathcal{M}[\mathcal{C}[i]] \leftarrow \mathcal{C}_{\text{shuffled}}[i]$;
  return $\mathcal{M}$;

C.2 Within-category Shuffle

Algorithm 2 describes the procedure for creating the within-category shuffle, which reassigns concepts to images within the same category while ensuring no concept maps to itself.

D Results on Llama 3.2 Models

In this section, we report results from our experiments on Llama-3.2-1B and its instruction-tuned variant Llama-3.2-1B-Instruct (Grattafiori et al., 2024). Figure 10 shows results of our main experiment on the Llama 3.2 models, and Figure 11 shows results on our counterfactual experiments for these models.

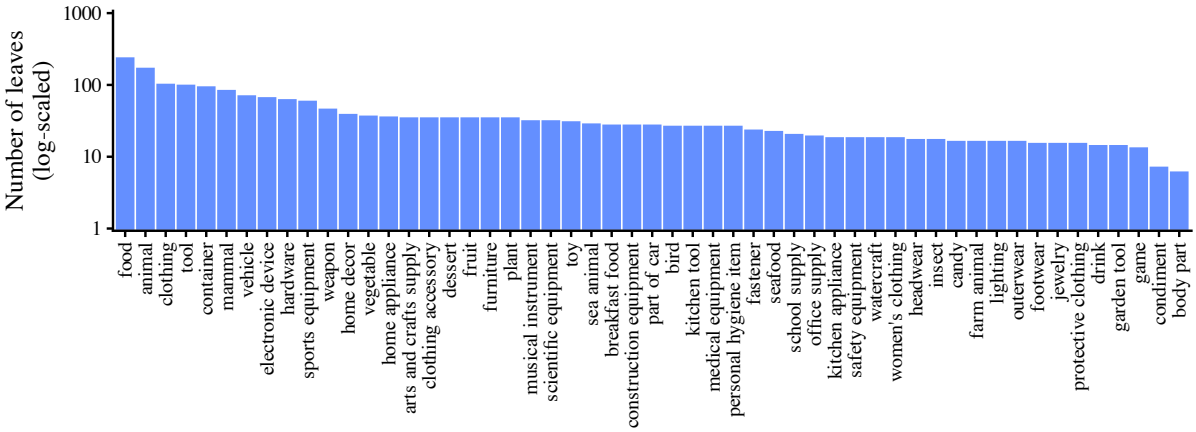

Figure 9: Number of leaves per hypernym category. Counts (y-axis) are log-scaled.

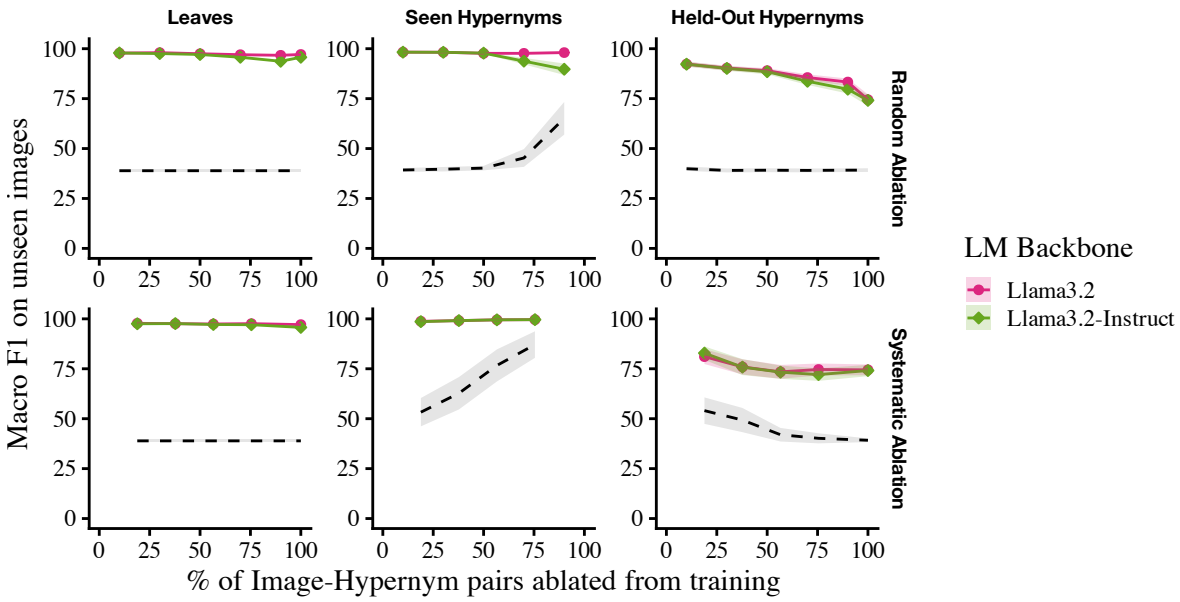

Figure 10: Macro F1s of Llama 3.2 models on unseen images across LM backbones, LM representations (pre-trained vs. random), and hypernym ablation type (Random vs. Systematic) at different amounts of exposure to hypernym categories, for various test splits. Dashed line indicates macro F1 of the majority-label baseline. Error bands represent 95% confidence intervals across random seeds (N = 3) and categories (hypernyms = 53, leaves = 1216).

Ablation level (%)	# Leaves affected	Data variant	Train set (# I-Q pairs)	Validation set (# I-Q pairs)			Test set (# I-Q pairs)		
				leaf	seen	unseen	leaf	seen	unseen
100	1,222	Original	23,250	2,028	–	3,380	9,394	–	15,702
		Across-category	23,250	2,028	–	3,366	9,394	–	15,680
		Within-category	23,250	2,028	–	3,380	9,394	–	15,664
90	1,148	Original	24,840	2,028	134	3,246	9,394	644	15,058
		Across-category	24,818	2,028	142	3,224	9,394	630	15,050
		Within-category	24,890	2,028	142	3,238	9,394	660	15,004
70	976	Original	28,962	2,028	492	2,888	9,394	2,300	13,402
		Across-category	28,908	2,028	488	2,878	9,394	2,278	13,402
		Within-category	28,952	2,028	496	2,884	9,394	2,308	13,356
50	768	Original	34,914	2,028	1,020	2,360	9,394	4,708	10,994
		Across-category	34,712	2,028	1,028	2,338	9,394	4,626	11,054
		Within-category	34,972	2,028	1,004	2,376	9,394	4,754	10,910
30	513	Original	43,340	2,028	1,750	1,630	9,394	8,134	7,568
		Across-category	43,410	2,028	1,768	1,598	9,394	8,142	7,538
		Within-category	43,354	2,028	1,680	1,700	9,394	8,158	7,506
10	170	Original	55,270	2,028	2,760	620	9,394	12,948	2,754
		Across-category	55,074	2,028	2,782	584	9,394	12,850	2,830
		Within-category	55,186	2,028	2,776	604	9,394	12,894	2,770

Table 2: Data statistics under 5 random hypernym ablation levels and counterfactual data variants. "I-Q pair" refers to an image and associated yes-no question. "Seen" and "Unseen" indicate the number of seen and unseen hypernym I-Q pairs, respectively. "# Leaves affected" is the number of leaf categories that have had any I-Q pairs removed, accounting for duplicates (removal under multiple hypernyms), and rounding.

Ablation level (hypernyms removed)	Data variant	Train set	Validation set			Test set		
			leaf	seen	unseen	leaf	seen	unseen
full	Original	23,250	2,028	–	3,380	9,394	–	15,702
	Across-category	23,250	2,028	–	3,366	9,394	–	15,680
	Within-category	23,250	2,028	–	3,366	9,394	–	15,664
40	Original	29,892	2,028	564	2,816	9,394	2,684	13,018
	Across-category	30,014	2,028	600	2,766	9,384	2,722	12,958
	Within-category	29,890	2,028	548	2,832	9,384	2,694	12,970
30	Original	33,800	2,028	924	2,456	9,394	4,242	11,460
	Across-category	33,702	2,028	888	2,478	9,394	4,242	11,438
	Within-category	33,844	2,028	926	2,454	9,394	4,254	11,410
20	Original	40,730	2,028	1,516	1,864	9,394	7,076	8,626
	Across-category	40,804	2,028	1,522	1,844	9,394	7,068	8,612
	Within-category	40,708	2,028	1,528	1,852	9,394	7,066	8,598
10	Original	48,214	2,028	2,142	1,238	9,394	10,094	5,608
	Across-category	48,120	2,028	2,176	1,190	9,394	10,028	5,652
	Within-category	48,230	2,028	2,136	1,244	9,394	10,090	5,574

Table 3: Data statistics under 4 systematic hypernym ablation levels and counterfactual data variants. "Seen" and "Unseen" indicate the number of seen and unseen hypernyms, respectively.

Ablation level	Ablated categories
40	kitchen appliance, home decor item, medical equipment, plant, protective clothing item, farm animal, mammal, car part, dessert, breakfast food, sea animal, hardware item, school supply, source of light, garden tool, drink, sports equipment, toy, insect, watercraft, footwear, vegetable, piece of jewelry, game, condiment, animal, personal hygiene item, bird, musical, office supply item, item of clothing, piece of women's clothing, construction equipment, candy, headwear, scientific equipment, home appliance, seafood, safety equipment, container
30	fastener, condiment, container, kitchen appliance, office supply item, body part, vegetable, kitchen tool, home appliance, outerwear, arts and crafts item, construction equipment, game, insect, breakfast food, home decor item, clothing accessory, food, school supply, fruit, sports equipment, sea animal, piece of women's clothing, candy, safety equipment, farm animal, dessert, scientific equipment, bird, tool
20	medical equipment, sea animal, home decor item, fastener, condiment, breakfast food, hardware item, office supply item, toy, seafood, musical, mammal, vehicle, personal hygiene item, protective clothing item, sports equipment, construction equipment, insect, scientific equipment, tool
10	container, personal hygiene item, dessert, item of clothing, furniture, breakfast food, scientific equipment, garden tool, car part, sea animal

Table 4: List of ablated categories under different systematic hypernym ablation levels.

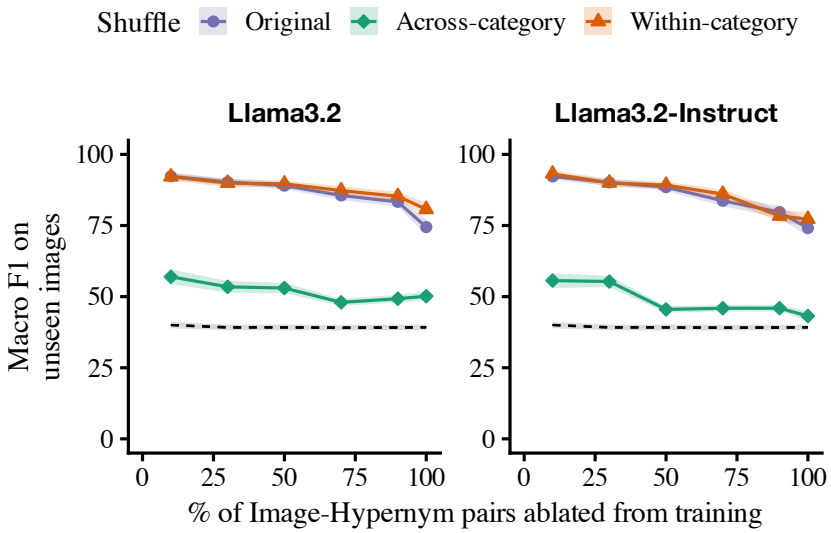

Figure 11: Macro F1s of Llama 3.2 models trained on data with different counterfactual shuffles (or lack thereof) at predicting hypernyms, across various amounts of exposure to hypernym categories. Dashed line indicates macro F1 of the majority-label baseline. Error bands show 95% confidence intervals for three random seeds and 53 hypernym categories.

Algorithm 2: Within-category Shuffle

Input: $\mathcal{H}$: hypernym-to-concepts mapping
Output: $\mathcal{M}$: concept-to-concept shuffle mapping

```
Function ShuffleWithinCategory(H):
    // Build concept-to-hypernyms mapping
    L ← ∅;
    foreach hypernym h, concepts C_h ∈ H do
        foreach concept c ∈ C_h do
            L[c] ← L[c] ∪ {h};
    // Group concepts by their hypernym sets
    G ← ∅;
    foreach concept c, hypernyms H_c ∈ L do
        G[frozenset(H_c)] ←
          G[frozenset(H_c)] ∪ {c};
    // Shuffle within each group
    M ← ∅;
    foreach group key k, concepts C_k ∈ G do
        C_shuffled ← RandomShuffle(C_k);
        for i = 1 to |C_k| do
            M[C_k[i]] ← C_shuffled[i];
    return M;
```

Experiment Setting	Macro F1	Std. Error
Original	78.4	1.3
Across-category	50.0	0.6
Within-category	79.8	1.2
Majority-label	39.2	0.9

Table 5: Macro F1 of models trained with Qwen3-8B as the LM backbone, on the 100% ablation setting, across various shuffles (or lack thereof). Standard error calculated over 3 seeds and 53 hypernym categories.

E Results on Qwen3-8B LM backbone

Table 5 shows results from the 100% ablation experiments across various counterfactual shuffles, using the Qwen3-8B model as the LM backbone. We see similar results as with smaller models in the main paper—results on Original and Within-category shuffles are similar while those on Across-category shuffle are vastly different (and lower). This suggests that arbitrary reasoning about hypernymy does not emerge in this larger model, and that it is still sensitive to the visual coherence of categories in the incoming input.

F Cross-Modal Generalization at Varying Depths of Taxonomic Organization

Our set of 53 categories can be organized hierarchically into multiple tree-like structures, each representing a taxonomy of a specific higher level category. The maximum depth across these trees is 3 (e.g., *kitchen tool* $\prec$ *tool* $\prec$ *hardware*, where '$\prec$' represents the hypernym relation, with the right hand side category being the hypernym of the left hand side category). Overall, we have 41 categories at depth 1, 9 categories at depth 2, and 3 categories at depth 3. To understand how cross-modal taxonomic generalization of our tested models varies across these depths, we present results (here, at the 100% ablation level) divided across the three depths for all models in Figure 12. We see that models generally perform better at depth 1 and 2 than at depth 3, i.e., they are markedly better at generalizing to lower level categories than at higher level categories.

Note that it is not clear if the poorer performance at depth 3 is necessarily because the models struggle at multi-hop reasoning (over hierarchically organized category structures), since the internal mechanism of how models arrive at their answer is not obvious. An alternate explanation could be

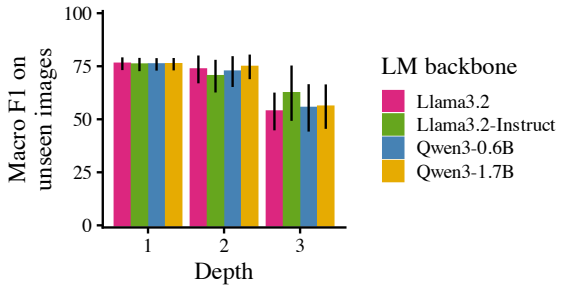

Figure 12: Macro F1s on unseen images for all four LM backbones across various depths of category organization in our analysis, computed for the 100% ablation experiment. Error bars indicate 95% confidence intervals across three random seeds and number of categories per depth (N_1 = 41; N_2 = 9; N_3 = 3).

the 'frequency of instantiation' hypothesis (Barsalou, 1985), which states that the number of times a learner observes a given category as member of another category determines the strength of the category relationship between the two. Therefore, the LM backbones may have encountered fewer instances of hyponym/leaf concepts as members of depth 3 hypernyms relative to depth 1 and 2 hypernyms. This does not appeal directly to a process that commits to an explicit hierarchical representation of categories, and yet can explain the observed result. We therefore caution readers from making inferences about the nature of explicit multi-hop hierarchical reasoning in the models from these results.