Prompting invokes expert-like downward shifts in multimodal models' conceptual hierarchies

Cara Leong (caraleong@nyu.edu)
New York University, New York, New York, USA

Brenden M. Lake (brenden@nyu.edu)
New York University, New York, New York, USA

Abstract

Humans tend to privilege an intermediate level of categorization, known as the basic level, when categorizing objects that exist in a conceptual hierarchy (e.g. choosing to call a Labrador a dog instead of Labrador or animal). Domain experts demonstrate a **downward shift** in their object categorization behaviour, recruiting subordinate levels in a conceptual hierarchy as readily as conventionally basic categories (Tanaka & Philibert, 2022; Tanaka & Taylor, 1991). Do multimodal large language models show similar behavioural changes when prompted to behave in an expert-like way? We test whether GPT-4 with Vision (GPT-4V, OpenAI, 2023a) and LLaVA (Liu, Li, Wu, & Lee, 2023; Liu, Li, Li, & Lee, 2023) demonstrate downward shifts using an object naming task and eliciting expert-like personas by altering the model's **system prompt**. We find evidence of downward shifts in GPT-4V when expert system prompts are used, suggesting that human expert-like behaviour can be elicited from GPT-4V using prompting, but find no evidence of downward shift in LLaVA. We also find that there is an unpredicted **upward shift** in areas of non-expertise in some cases. These findings suggest that in the default case, GPT-4V is not a novice: instead, it behaves at default with a median level of expertise, while further expertise can be primed or forgotten through textual prompts. These results open the door for GPT-4V and similar models to be used as tools for studying differences in the behaviour of experts and novices, and even comparing contrasting levels of expertise within the same large language model.

Keywords: concepts and categories; expertise; large language models; downward shift; prompting

Introduction

When presented with a concept that can be conceived of at multiple levels of abstraction (e.g., *animal*, *dog*, *labrador*), humans preferentially categorize objects at an intermediate level (e.g. *dog*). This effect, known as the basic-level advantage, has been observed across tasks such as object recognition, object naming and feature listing (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

Which level of abstraction is considered as basic varies among individuals as a result of their knowledge and experience. An individual's cultural context may affect their knowledge and experience: Tzeltal Mayan children of Southern Mexico who lived in agrarian cultures tended to categorize trees by their folk genera (Dougherty, 1978), while American children in an urban environment tended to use the more general family level as the basic level (Rosch et al., 1976). Similarly, domain experts tend to show a **downward shift** in the categories they use as basic within the area of their expertise (Gauthier & Tarr, 1997; Johnson & Mervis, 1997; Rota & Zellner, 2007; Tanaka & Taylor, 1991). For example, while experienced birdwatchers predominantly used the label `Dog` to name pictures of collies and labradors, they used more specific labels like `Robin` and `Crow` to name pictures of birds (Tanaka & Taylor, 1991). Experience with performing classification tasks may also be implicated in downward shift, as performing perceptual classification at the species level instead of mere visual exposure allowed bird experts to identify novel bird species more accurately (Tanaka, Curran, & Sheinberg, 2005).

Language models are trained to predict text over a large corpus of input documents. Often, these texts vary greatly in their cultural and knowledge domains, levels of expertise, and audience. Yet, models like GPT-4-Vision (GPT-4V) (OpenAI, 2023a) and LLaVA (Liu, Li, Wu, & Lee, 2023; Liu, Li, Li, & Lee, 2023) that are trained on such varied corpora have demonstrated the ability to perform tasks requiring domain expertise, such as passing AP tests (OpenAI, 2023b), annotating legal texts (Savelka, Ashley, Gray, Westermann, & Xu, 2023) and annotating medical images (Yang et al., 2023).

Although language models can perform reasonably well on downstream tasks, it is unclear whether they demonstrate expertise in a human-like way. This paper explores whether multimodal instruction-tuned models can be prompted to behave like a human expert on a behavioural measure of expertise by demonstrating downward shift in the basic level when prompted to display expertise.

We test whether a model's system prompt, a preamble that constrains an instruction-tuned model's style of response, can be used to elicit expert-like effects in an object naming task[1]. We show that using system prompts does elicit downward shift in expert categories for GPT-4V but not LLaVA, providing a behavioural indicator of expertise beyond performance on downstream tasks. However, we also find that using expert system prompts can result in an unexpected **upward shift** in categories of non-expertise, where the model uses more basic-level labels for dogs when prompted to behave like a bird expert than it does in the default case.

Methods

Large language models can be instructed to role-play as agents (e.g., Andreas, 2022; Shanahan, McDonell, & Reynolds, 2023). Using a **system prompt** outlines the context and manner in which the model should respond, and can constrain the content and style of a model's output, e.g., to produce text in the style of Shakespeare (Bushwick, 2023). Models augmented using an expert persona in such a manner produce

[1] Code available at `http://github.com/craaaa/expertise`

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Prompt Type		System Prompt	Exp. Cat. Mentioned	Non-Exp. Cat. Mentioned
Default	1	You are an AI assistant.		
	2	You are an AI assistant that helps people find information.		
	3	You are an AI assistant that has basic knowledge about most concepts.		
	4	You are an AI assistant that has basic knowledge about most concepts. You know a little bit about dogs, birds, trees, and flowers.		✓
Expert	5	You are an AI assistant that is an expert on birds.	✓	
	6	You are an AI assistant that is an expert on birds. You have a great deal of experience in birdwatching, and spend a lot of time around birds.	✓	
	7	You are an AI assistant. You know a little bit about dogs, trees, and flowers, and have expert knowledge about birds.	✓	✓
	8	You are an AI assistant that is an expert ornithologist.		

Table 1: Example system prompts. We vary whether and how directly the targeted image categories are mentioned.

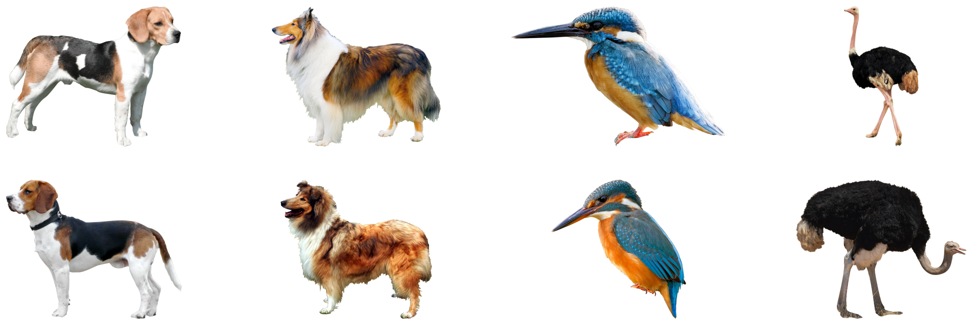

Figure 1: Sample images from the `Dog` and `Bird` categories (van Hoef, Lynott, & Connell, 2022)

output that is longer and qualitatively judged as "higher quality" (Xu et al., 2023). We leverage this system prompt to simulate expert personae.

We compare a multimodal model's behaviour to human behaviour using an object naming experiment on images in the basic categories[2] `Dog`, `Bird`, `Tree`, and `Flower`. We create system prompts that instruct the model to behave like an expert in one of the four domains. We expect increased use of subordinate category labels, and reduced use of basic category labels, in the stated domain of expertise. Meanwhile, we expect that the model will use fewer subordinate labels for images that are not in the stated domain of expertise, and that the prompted domain of expertise should not affect any other domain (e.g., dog experts and bird experts should have similar knowledge about trees).

We construct prompts for the object naming task. Each prompt contains a **system prompt**, which sets the model's persona, a verbal task instruction, and one image of an item belonging to one of the four object categories. For example:

> **System prompt**: You are an AI assistant.
> **Task instruction**: What's in this image? Answer as quickly as possible using one or two words.

We adapt our task instruction from the participant instructions in Experiment 2 of Tanaka & Taylor, 1991. Although the instructions for humans did not specify an answer length, we instruct the model to provide 1-2 word answers for brevity, while allowing for subordinate labels that may consist of multi-word expressions. For instance, *Irish Wolfhound* may be a more natural way of expressing *Wolfhound*, and *Hazel tree* might be preferred to *Hazel*.

We vary our choice of system prompts to simulate Experiment 2 of Tanaka & Taylor, 1991, where dog experts and bird experts each perform an object naming task (see Table 1 for examples of prompts). **Default system prompts** (Prompts 1-4 in Table 1) aim to elicit a novice persona and either do not specify any area of expertise, or explicitly specify a novice persona. **Expert prompts** (Prompts 5-8 in Table 1) aim to elicit expertise in one of the test categories. We vary the length of the system prompt by including examples of the kinds of expertise an expert might have. We also vary whether we explicitly mention the basic category labels of the expert category (Prompts 5, 6, and 7 in Table 1) and novice categories (Prompt 7 in Table 1). Prompt 7 also explicitly indicates an expert in one domain who is a novice at all other domains, while the other prompts do not explicitly mention the expected

[2] We call categories **basic** and **subordinate categories** in accordance with the findings of Rosch et al. (1976) rather than a claim about which level of abstraction is considered basic to the model.

level of expertise in non-expert domains.

We use images from four natural categories — `Dog`, `Bird`, `Tree` and `Flower` — from van Hoef et al. (2022). Stimuli for each of these basic categories consists of four images from ten unique subordinate categories (e.g., `Collie` and `Labrador` for the basic category `Dog`), totalling 40 images. Each image is presented on a white background (see Figure 1). For LLaVA, we compare the log-probabilities of the model using the basic or subordinate label as a continuation to the prompt. We code a response as basic if $P(\text{basic}|\text{prompt}, \text{image}) > P(\text{subordinate}|\text{prompt}, \text{image})$ and subordinate otherwise.

Since GPT-4V's output logits were not publicly available at the time of writing, we sample from each model ten times using the same prompt with the settings `temperature = 1, image detail = low`. We code the generations from GPT-4V as Basic or Subordinate using the category labels provided by van Hoef et al., 2022. We mark all response strings that match at least part of the category label as matches. Thus, *Autumn tree* and *Trees* are coded as matching the label `Tree`. Additionally, we hand-code alternative subordinate labels to the dataset's subordinate label: for example, we code *Welsh Springer* and *Irish Setter* as subordinates in the category `Setter`, while *Parrot* are coded as a subordinate label for the category `Macaw`. Responses are coded as subordinate even if the label does not match the ground-truth (e.g., pine trees that were labelled *fir tree*) as long as the label is a possible subordinate term in the category. We exclude all responses that do not follow the answer template from our analysis.

Model responses ranged from single word noun-phrases (e.g. *Tree*) to two-word descriptions of the object (e.g. *Bulldog standing*, *Red tree*). Some two-word responses contained both basic and subordinate category names (e.g., *Beagle Dog*, *Kingfisher Bird*); these responses were coded as subordinate. 52.0% of all default responses consisted of two words. The model refused to provide an answer in 5 responses (0.0001% of total responses); these responses were excluded.

Results

Default System Prompts

Figure 2 shows GPT-4V and LLaVA's performance on the object naming task using four variants of default system prompt (Prompts 1-4 in Table 1).

Both models' responses were relatively consistent across different system prompts within each image category, suggesting that the exact wording of the system prompt does not significantly change performance on this task. Specific system prompts did not affect the model's choice of basic or subordinate labels – pairwise χ^2-tests between different system prompts were not significant for both models.

We compare these results to the results from a human picture naming task conducted on the same images (van Hoef et al., 2022). Both models preferred subordinate labels more than humans did across all categories, suggesting that in the default case, models behave more 'expert-like' than novice humans on the labelling task. In particular, both models showed a preference for labelling images in the `Bird` and `Dog` category using subordinate labels, while naming images in the `Tree` category using basic labels. While GPT-4V's choice of basic labels showed high variance between categories, such behavior in fact matches human performance. Both humans and GPT-4V used basic labels the most for the `Tree` category, and the least for the `Bird` category. In the default case, GPT-4V used more subordinate labels than novice humans across the board, but also reflected a similar bias as humans towards naming birds using subordinate labels, and naming trees using the basic category label.

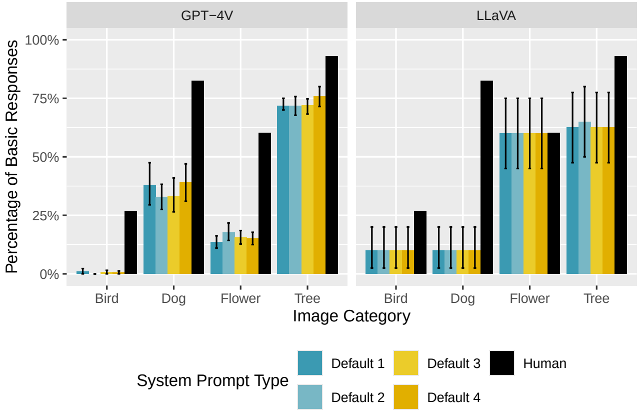

Figure 2: Preference for basic responses over four different default system prompts. Error bars here and elsewhere represent bootstrapped 95% confidence intervals.

Knowledge of Subordinate Categories

One potential explanation for a model's label preference might be failure to recognize the image as an instance of the other category. To exclude this possibility, we tested both models on a sentence verification task. For each image, we used the default system preamble and same setup as in the main experiment with the following task instruction:

> Does this image contain a {category label}? Answer true or false.

We tested each model's ability to verify category membership in its ground truth basic and subordinate categories, as well as non-membership in two alternative categories at the same level of categorization. For example, we tested if an image of a `Collie` was correctly classified as a dog, but not a bird or flower, and as a collie, not a spaniel or dalmatian. We present F_1 scores of this classification task below:

	GPT-4V		LLaVA	
	Basic	**Subordinate**	**Basic**	**Subordinate**
Bird	1.0	1.0	0.988	0.975
Dog	1.0	0.990	1.0	0.733
Flower	1.0	0.991	1.0	0.777
Tree	0.98	0.494	0.976	0.590

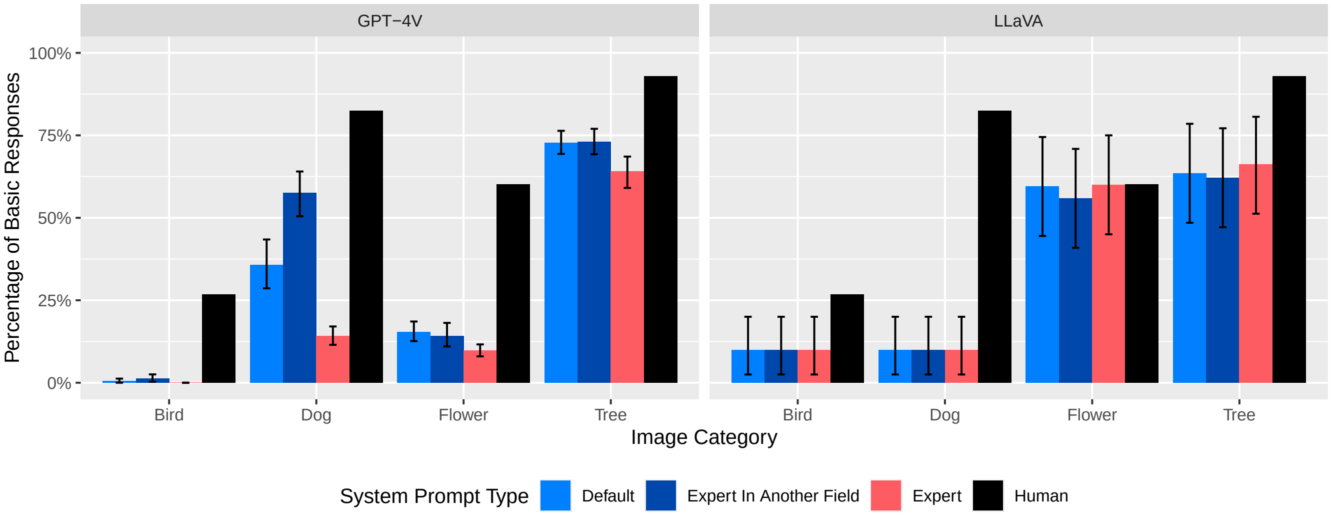

Figure 3: Choice of labels across different system prompts. Error bars represent bootstrapped 95% confidence intervals.

These results suggest that both models are able to identify the subordinate categories used in the Dog, Bird, and Flower categories, but not Tree species, which accounts for the model's predominant use of the basic category label to label images of trees whereas the model otherwise tended to use subordinate category labels. If the verification task is taken to be a pre-requisite of the object naming task, we might thus not expect significant use of subordinate labels in the Tree domain regardless of how much expertise the model is prompted to demonstrate.

Expert System Prompts

Figure 3 shows the difference in the use of basic labels depending on the choice of system prompt. We group together system prompts that elicit expertise in a domain that is different from the domain being tested; this includes system prompts that explicitly mention novicehood in the test domain (e.g. Prompt 7), as well as system prompts that do not mention the test domain (e.g. Prompt 5). For example, a system prompt targeting dog expertise would be considered non-expert when performing the object naming task on a bird image.

The use of different system prompts did not result in a downward shift in categorization in LLaVA. There was no significant difference in LLaVA's preferences for basic and subordinate responses when comparing default and expert system prompts (Bird $-\chi^2(1) = 0$, *ns*; Dog $-\chi^2(1) = 0$, *ns*; Flower $-\chi^2(1) = 0$, *ns*; Tree $-\chi^2(1) = 0.18629$, *ns*), or when comparing default and non-expert system prompts (Bird $-\chi^2(1) = 0$, *ns*; Dog $-\chi^2(1) = 0$, *ns*; Flower $-\chi^2(1) = 0.64151$, *ns*; Tree $-\chi^2(1) = 0.065$, *ns*).

On the other hand, across all four image categories, GPT-4V produced fewer basic labels (and correspondingly, more subordinate labels) in the area of domain expertise when using an expert system prompt than when prompted using a default prompt, or a prompt that elicited expertise in another domain.

We first compare responses in the expert condition to responses in the default condition. In all image categories, there was a significant relationship between the choice of preamble and the number of basic responses. In the Bird category, basic responses were near zero in the default condition and decreased to zero in all bird expert conditions, $\chi^2(1) = 7.13, p = 0.007$. In the Dog category, although the model continued to use basic category labels in 14.1% responses, the number of basic responses decreased when using dog expert system prompts, $\chi^2(1) = 197.72, p < 0.001$. The same was true of the Tree category, $\chi^2(1) = 27.952, p < 0.001$, and the Flower category, $\chi^2(1) = 22.995, p < 0.001$.

We next compare responses in the default condition to responses in the non-expert condition. In the Tree and Flower categories, there was no significant difference in distribution of responses between the default and non-expert conditions, $\chi^2(1) = 0.0094$, *ns* and $\chi^2(1) = 0.26$, *ns* respectively. However, when the system prompts did not target dog expertise, the number of basic responses to Dog images did not remain the same, but significantly increased, $\chi^2(1) = 227.89, p < 0.001$. Similarly, the number of Bird images labelled using a basic label was higher in the non-expert condition than in the default condition, $\chi^2(1) = 5.4127, p < 0.02$. This behaviour, which we call **upward shift**, is not predicted by the downward shift hypothesis, which suggests that expertise should increase the number of categories that are considered basic in the domain of expertise, but should not affect unrelated categories.

Figure 5 breaks down the impacts of the expert and non-expert system prompts in greater detail. Non-expert prompts do not always mention the image category being tested (see Prompts 5, 6, and 7 in Table 1), while Prompt 8 contains no mention of any basic category label. Although upward shift occurred in the Dog and Bird categories, the magnitude of the upward shift decreased when the category under test was mentioned in the preamble even if the prompt only specified "a little bit" of knowledge about the category. In the Bird category, the extent of upward shift decreased by 1.3 percentage points $\chi^2(1) = 10.857, p < 0.001$, while in the Dog cate-

Figure 4: Sample images from the Cat and Dog categories of the Cats And Dogs dataset (Parkhi et al., 2012)

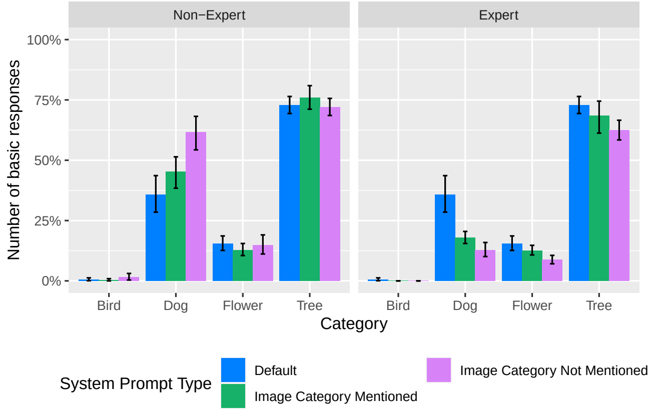

Figure 5: Mention of the image's basic label in the system prompt affects how the image is named.

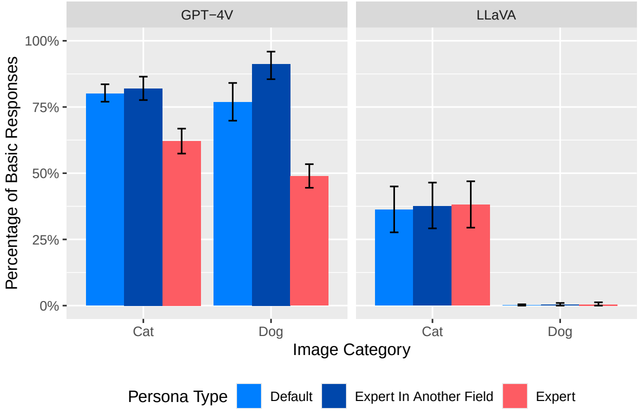

Figure 6: Choice of labels in default, non-expert and expert system prompts tested on the Cats and Dogs dataset. We group all responses where the system prompt encourages expertise in another domain as 'Non-Expert'.

gory, the extent of upward shift decreased by 16.33 percentage points, $\chi^2(1) = 97.583, p < 0.001$. These results suggest that using the basic category label in the system prompt affects how the model behaves, potentially by priming the model.

We also find a marginal effect of system prompt on the magnitude of downward shift. The extent of downward shift increased in magnitude in the expert condition when a prompt mentioned the study of that category (Prompt 8 in Table 1). An average downward shift of 17.69 percentage points was observed in Dog category labels when the system prompt mentioned the basic category of the test image. In contrast, the downward shift between the default prompt types and the system prompt that did not mention the basic category label was 22.85 percentage points, $\chi^2(1) = 6.18, p = 0.013$. Similarly, downward shift increased marginally between prompts in the Tree category, $\chi^2(1) = 4.18, p = 0.041$, and Flower category, $\chi^2(1) = 4.31, p = 0.038$. These results suggest that although using a system prompt can elicit expert-like behaviour in GPT-4V, the magnitude of the effect is dependent on the particular system prompt used, and particularly the lexical choices surrounding the kind of expertise elicited.

Testing On Naturalistic Images

To verify the robustness of our results, we investigate whether the downward shift and upward shift demonstrated in the Dog and Bird categories replicates on a naturalistic set of images.

Methods

We conduct an object naming task using the Cats And Dogs dataset (Parkhi et al., 2012), a dataset containing natural images of cats and dogs of different breeds. Examples of the images used are shown in Figure 4. While the images in our previous experiment contained the entire object on a plain white background, images in this dataset may contain other objects in the foreground, and may also not contain the entire body of the object in the image, thus making the task of object naming slightly more involved.

From the dataset, we randomly select twelve cat breeds and dog breeds and randomly choose four images of each breed. We use the same methodology as in the previous experiment to perform an object naming task on this dataset.

Results

Figure 6 shows the results of the object naming task on the Cats And Dogs dataset. We find evidence of downward shift in both Cat and Dog categories for GPT-4V, but not LLaVA. In the default condition, GPT-4V used basic labels for 76.83% of dog images and 80.0% of cat images. In the expert condition, the number of basic labels decreased to 48.83% for dog images, $\chi^2(1) = 200.23, p < 0.001$, and 62.17% of cat images, $\chi^2(1) = 91.97, p < 0.001$. Meanwhile, LLaVA consistently preferred the basic response in about 35% of responses, with no significant difference between default and non-expert system prompts (Dog —$\chi^2(1) = 1.8\text{e}{-33}$, *ns*; Cat —$\chi^2(1) = 8.1\text{e}{-33}$, *ns*) and default and expert system prompts (Dog —$\chi^2(1) = 3.2\text{e}{-32}$, *ns*; Cat —$\chi^2(1) = 8.1\text{e}{-33}$, *ns*).

GPT-4V shows a significant upward shift in the Dog category, with an increase from 76.83% of trials labelled with basic labels in the default condition to 91.21% of trials labelled with basic labels in the non-expert condition, $\chi^2(1) = 139.61, p < 0.001$. However, we do not find a significant upward shift in the Cat category, $\chi^2(1) = 1.89, p = 0.169$. While downward shift occurs robustly in multiple image categories and across different dataset types, upward shift occurs prominently in the category of dog images, but is not a consistent occurrence.

Discussion

We tested whether two different multimodal models show expertise effects similar to the effect found by Tanaka and Taylor (1991) in an object naming task. While GPT-4V preferred subordinate labels in areas of domain expertise when using an expert system prompt, LLaVA did not. These overall result suggests that while system prompts can constrain a model's outputs in a way that behaviourally aligns with human experts, they need not; telling a model to behave in an expert-like way does not always invoke expert-like behavior. Moreover, when an expert system prompt is used, upward shift can occur in non-expert domains where the model uses more basic labels than in the default case. Differences in model architecture, the instruction tuning process, and may all have contributed to these differences in behaviour; which factors contribute to a model's ability to 'change personas' should be investigated further.

Insights from the Default Basic Level

Both GPT-4V and LLaVA preferred subordinate-level labels more than the novice English speakers of either Tanaka and Taylor (1991) or Rosch et al. (1976) to label images, even when default system prompts were used. In the Bird and Dog categories, both GPT-4V and LLaVA preferred labels that would be considered subordinate for novices (e.g. *Dandelion*, *Kingfisher*) over 75% of the time. In comparison, American English speakers who were novices at bird identification used subordinate labels approximately 20% of the time in Tanaka and Taylor (1991)'s Experiment 2.

These results can be understood in light of Tanaka and Taylor (1991)'s finding that bird experts, for whom the identification of birds was a highly salient task, were significantly more likely to use subordinate labels for birds than dog experts were to use subordinate labels for dogs. Combining data sampled from communities of dog and bird experts (e.g. using Google Images and Flickr searches to create the Microsoft COCO dataset (Lin et al., 2015), which LLaVA is trained on) with computer vision datasets such as the Stanford Dogs dataset (Khosla, Jayadevaprakash, Yao, & Fei-Fei, 2011) which focus on fine-grained categorization may result in datasets that over-represent subordinate labels relative the novice knowledge. In this way, it is possible that existing human preferences may have influenced GPT-4V's own 'knowledge'.

Priming as Expertise

Another fundamental difference between the use of system prompts and human categorization behaviour lies in the difference between primed behaviour and latent knowledge. Whereas the human experts in Tanaka & Taylor, 1991's experiments were experts in a single domain whose responses were not primed during testing, system prompts prime a single model to "role play" (Shanahan et al., 2023) using different personas. A model's latent 'expertise' might not be fully suppressed, leading to lack of downward shift, or be suppressed in order to adopt the role of a novice, leading to unexpected behaviour such as upward shift. A human correlate of the latter behaviour might be observed if subject matter experts were asked to answer as if they were novices, or in novices if asked to name categories at the most specific level of abstraction possible, or if told beforehand that they should *act* as experts in a particular domain.

To the extent that GPT-4V can be viewed as a cognitive model for human categorization, the occurrence of upward shift predicts the possibility that upward shift may occur in human categorization as well, revealing a gap in the existing experimental literature. Studies have compared experts and novices separately, but a comparison between pure novices and experts in one or more domain may uncover new insights. One possibility is that increased expertise in one domain might affect an individual's expertise in another domain given finite mental resources; whether this effect does in fact occur in humans is another area for future research.

Conclusion

We tested whether multimodal large language models demonstrate expert-like downward shifts using an object naming task and altering the model's system prompts. We found that LLaVA did not show any human-like changes in behaviour as a result of expertise, but GPT-4V showed evidence of downward shifts when expert system prompts were used, showcasing GPT-4V's ability to act in an expert-like manner not only on downstream tasks, but also on behavioural measures that mimic human-like expertise.

We also found that an unexpected upward shift in expertise could be elicited from GPT-4V, and that the extent of the downward shift was mediated by whether the system prompt contained the basic category label, suggesting that the model's responses were primed by the system prompt. These behaviours are unexpected in humans, and reveal a fundamental divergence between human experts and GPT-4V as a model which can be prompted to role-play different personas.

Future work should explore what mechanisms allows downward shift to be demonstrated in GPT-4V but not LLaVa, and whether downward shift is demonstrated consistently across multiple behavioural metrics. For instance, domain experts are often quicker at category verification of subordinate labels than novices (Gauthier & Tarr, 1997; Tanaka & Taylor, 1991), and can list more features at the subordinate level than novices (Tanaka & Taylor, 1991). Testing these other behavioural measures of expertise may provide a more detailed picture of how well large language models behave like humans.

Acknowledgments

Thanks to Guy Davidson, Sreejan Kumar, Will Merrill, Wentao Wang, Will Timkey, and Jackson Petty for comments on earlier drafts of this paper. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Andreas, J. (2022). Language Models as Agent Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022.*

Bushwick, S. (2023). What the New GPT-4 AI Can Do. *Scientific American.*

Dougherty, J. W. D. (1978). Salience and relativity in classification. *American Ethnologist*, *5*(1), 66–80. doi: 10.1525/ae.1978.5.1.02a00060

Gauthier, I., & Tarr, M. J. (1997, June). Becoming a "Greeble" Expert: Exploring Mechanisms for Face Recognition. *Vision Research*, *37*(12), 1673–1682. doi: 10.1016/S0042-6989(96)00286-6

Johnson, K. E., & Mervis, C. B. (1997, September). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology. General*, *126*(3), 248–277. doi: 10.1037//0096-3445.126.3.248

Khosla, A., Jayadevaprakash, N., Yao, B., & Fei-Fei, L. (2011). Novel dataset for fine-grained image categorization : Stanford dogs. In *First Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., … Dollár, P. (2015, February). *Microsoft COCO: Common Objects in Context* (No. arXiv:1405.0312). arXiv.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023, October). *Improved Baselines with Visual Instruction Tuning* (No. arXiv:2310.03744). arXiv. doi: 10.48550/arXiv.2310.03744

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. In *Proceedings of 37th Conference on Neural Information Processing Systems.*

OpenAI. (2023a, March). *GPT-4 System Card.*

OpenAI. (2023b, March). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv.

Parkhi, O. M., Vedaldi, A., Zisserman, A., & Jawahar, C. V. (2012). Cats and Dogs. In *IEEE conference on computer vision and pattern recognition.*

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976, July). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. doi: 10.1016/0010-0285(76)90013-X

Rota, L. M., & Zellner, D. A. (2007, February). The categorization effect in hedonic contrast: Experts differ from novices. *Psychonomic Bulletin & Review*, *14*(1), 179–183. doi: 10.3758/BF03194047

Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023, June). Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (pp. 117–123). doi: 10.1145/3587102.3588792

Shanahan, M., McDonell, K., & Reynolds, L. (2023, November). Role play with large language models. *Nature*, *623*(7987), 493–498. doi: 10.1038/s41586-023-06647-8

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005, February). The Training and Transfer of Real-World Perceptual Expertise. *Psychological Science*, *16*(2), 145–151. doi: 10.1111/j.0956-7976.2005.00795.x

Tanaka, J. W., & Philibert, V. (2022). *The Expertise of Perception: How Experience Changes the Way We See the World* (1st ed.). Cambridge University Press. doi: 10.1017/9781108919616

Tanaka, J. W., & Taylor, M. (1991, July). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482. doi: 10.1016/0010-0285(91)90016-H

van Hoef, R., Lynott, D., & Connell, L. (2022). Timed Picture Naming Norms for 800 Photographs of 200 Objects in English.

Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., Zhang, Y., & Mao, Z. (2023, May). *ExpertPrompting: Instructing Large Language Models to be Distinguished Experts* (No. arXiv:2305.14688). arXiv.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023, October). *The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)* (No. arXiv:2309.17421). arXiv.