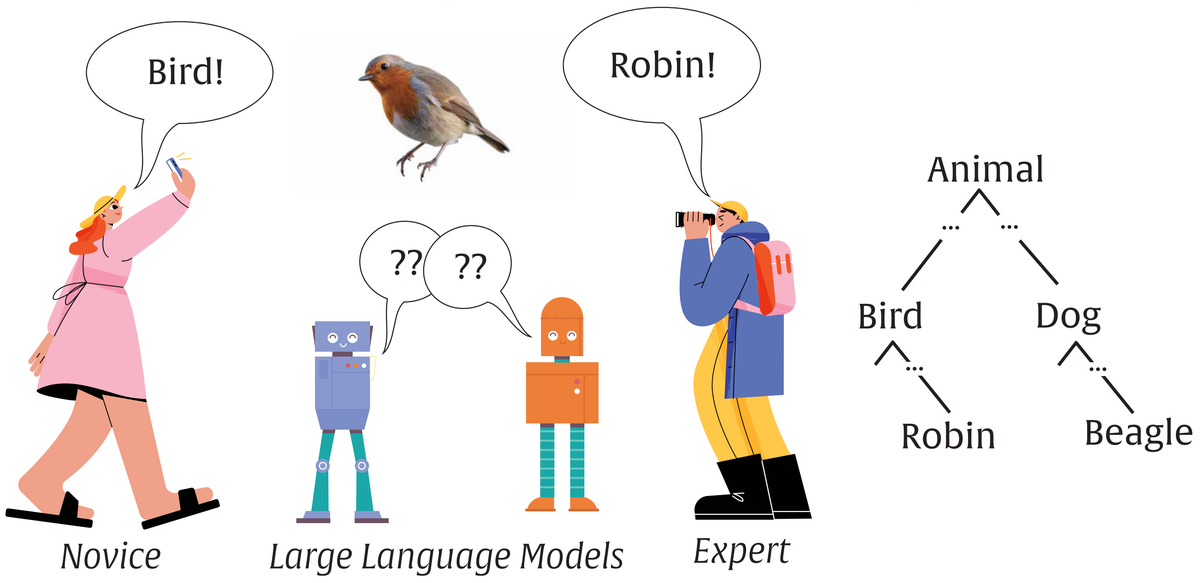
Prompting *sometimes* invokes expert-like downward shifts in multimodal models' conceptual hierarchies

Cara Su-Yi Leong & Brenden Lake

NYU

1 Human domain experts demonstrate a ***downward shift*** in the basic level of categorization (Tanaka and Taylor 1991).



Do vision-and-language models show similar downward shifts when instructed to behave like experts?

2 ***System prompts*** instruct (multimodal) large language models to role-play as agents (e.g., Andreas, 2022).

Do expert system prompts cause downward shifts in an ***object categorization task***?

Default

You are a helpful, respectful and honest assistant.

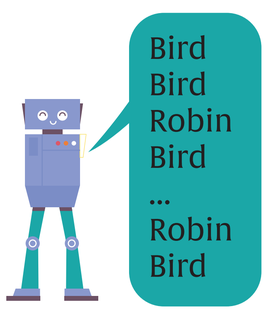
Expert

You are an AI assistant that is an expert on birds.

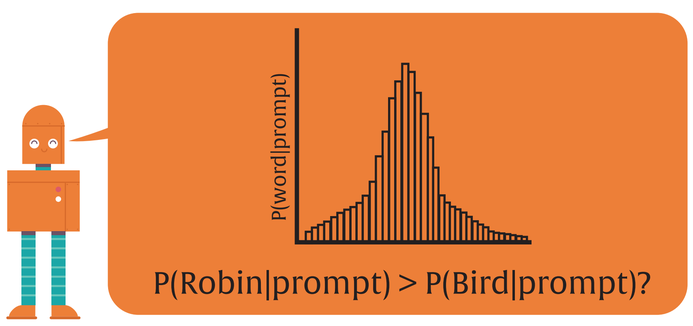
Expert in Another Field

You are an AI assistant that is an expert on dogs. You have a great deal of experience in dog breeding, grooming and training, and spend a lot of time around dogs.

What's in this image? Answer as quickly as possible using one or two words.



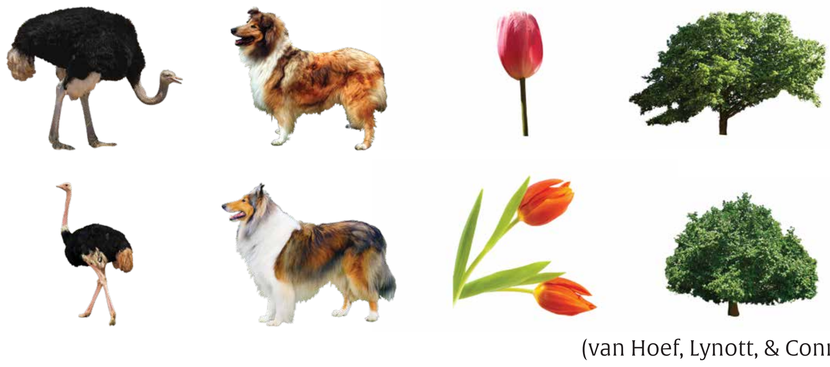
GPT-4V (OpenAI, 2024)



LLaVA (Liu, Li, Wu, & Lee, 2023)

3 We tested both models on two datasets of natural kinds:

A

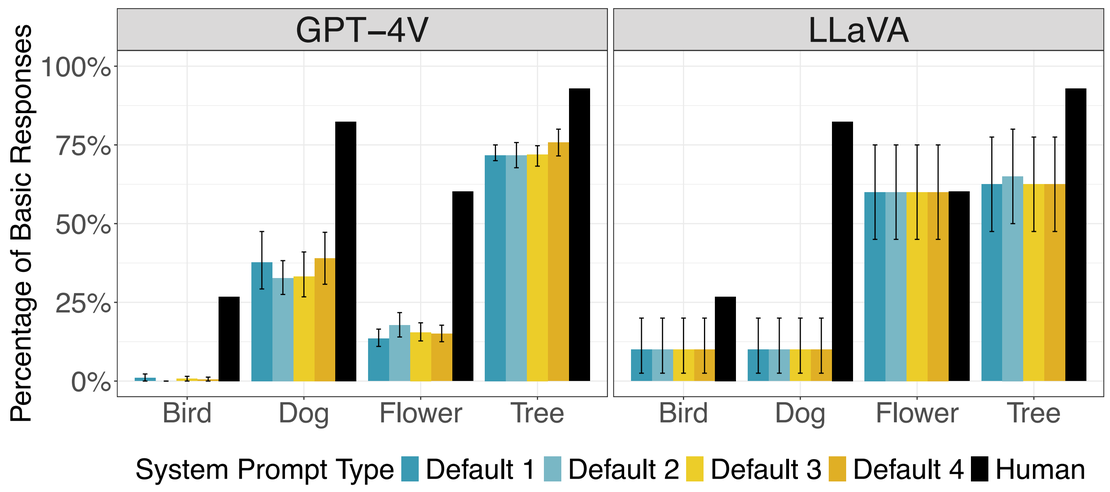


(van Hoef, Lynott, & Connell, 2022)

B

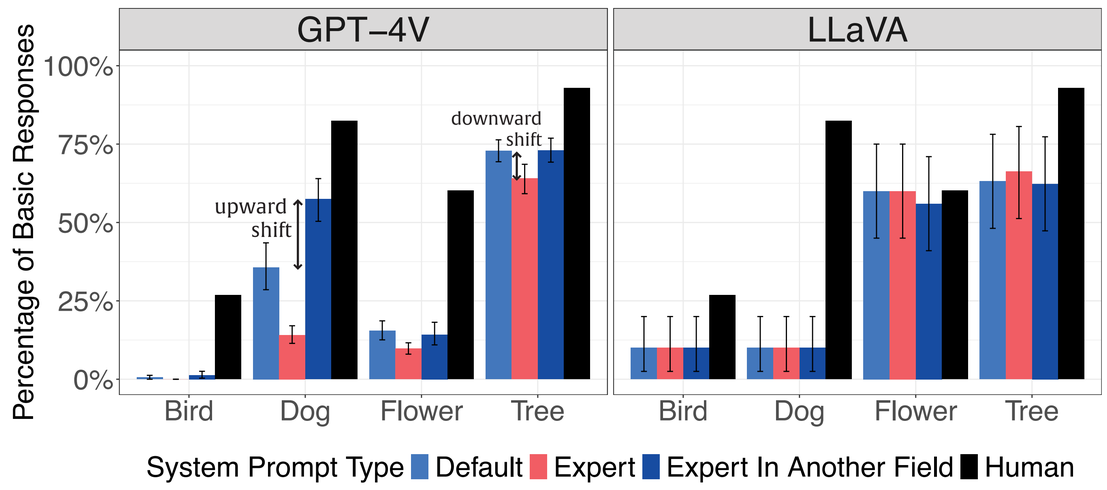
(Parkhi, Vedaldi, Zisserman & Jawahar, 2012)

4 GPT-4V and LLaVA showed preference for ***subordinate-level labels*** when default system prompts were used.

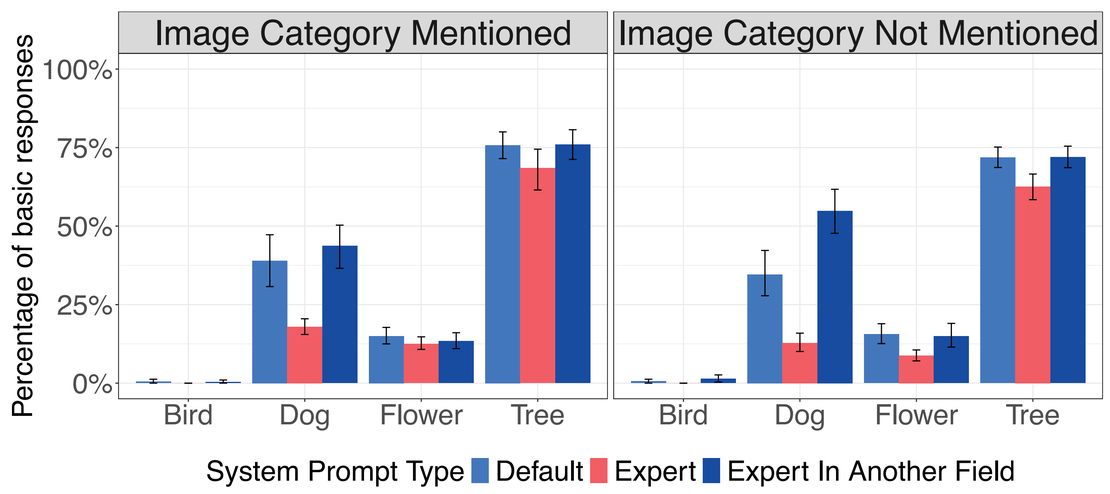


When expert system prompts were used, GPT-4V showed ***downward shifts*** in all domains, while LLaVA did not.

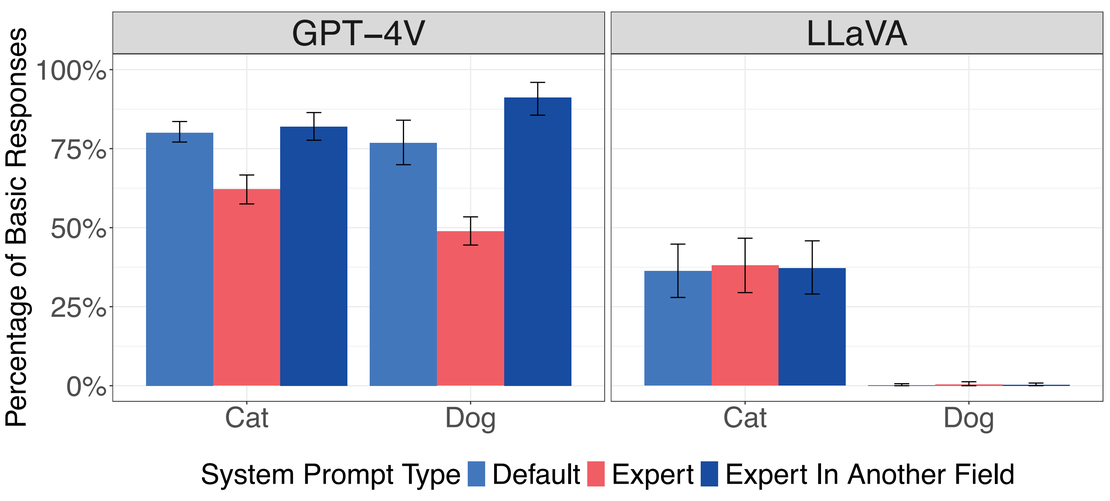
GPT-4V also showed ***upward shift*** on dog images when prompted to show expertise in a different domain.



Explicit mention of the test image category in the system prompt marginally affected the magnitude of upward shift.



5 Similar results were obtained on more realistic images (B).



6 Downward shift is limited to certain domains and models.

GPT-4V seems to behave in an 'expert-like' way, but also differs from human experts --- maybe because it can take on personae with different levels of expertise?

A corollary: would human experts show upward shift if asked to behave like novices?