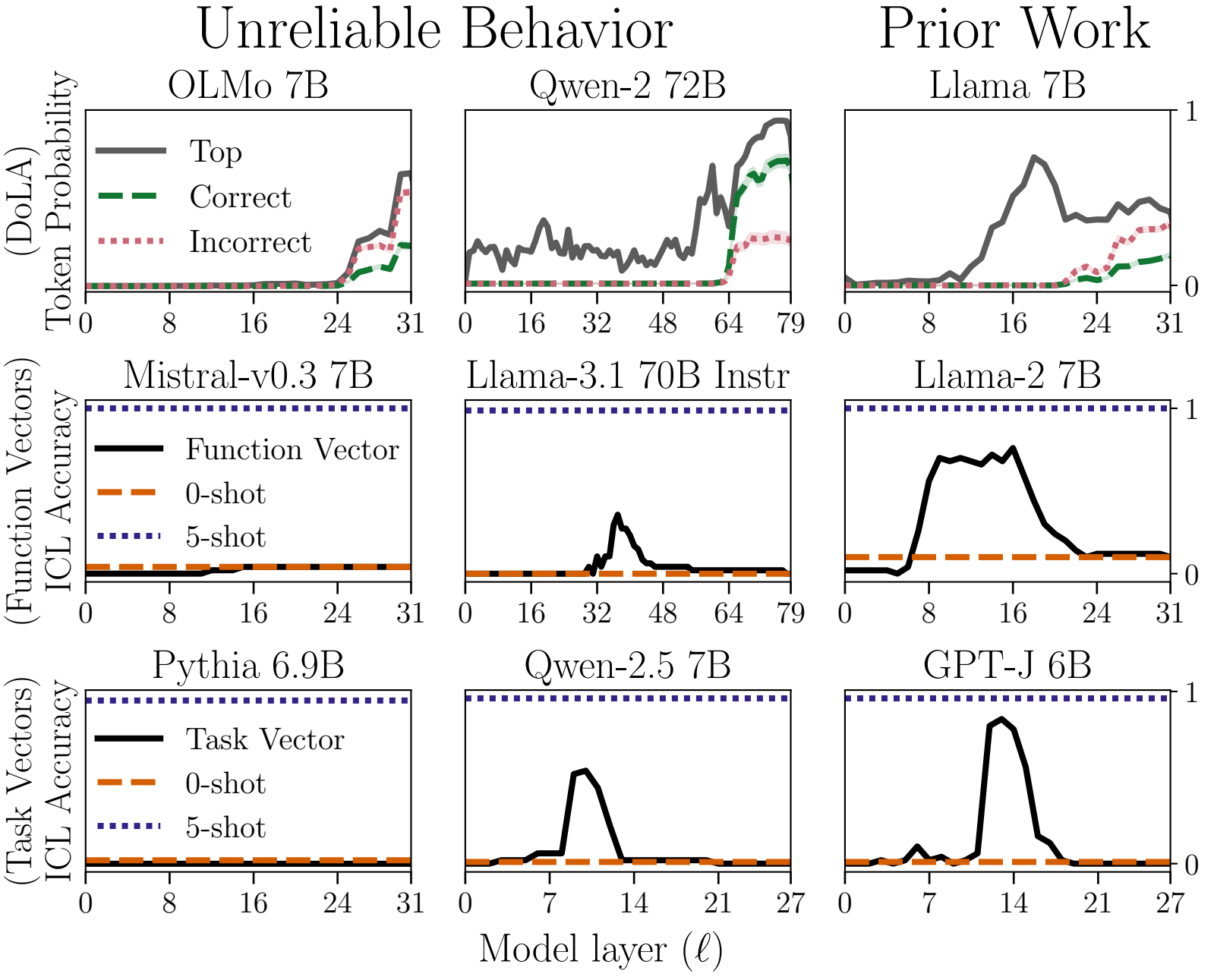
Unreliable Behavior
Prior Work
OLMo 7B
Qwen-2 72B
Llama 7B
(DoLA)
Token Probability
Top
Correct
Incorrect
Mistral-v0.3 7B
Llama-3.1 70B Instr
Llama-2 7B
(Function Vectors)
ICL Accuracy
Function Vector
0-shot
5-shot
Pythia 6.9B
Qwen-2.5 7B
GPT-J 6B
(Task Vectors)
ICL Accuracy
Task Vector
0-shot
5-shot
Model layer (ℓ)