

Balanced Thinking: Improving Chain of Thought Training in Vision Language Models

Supplementary Material

A. Overview

This supplementary document provides additional results and implementation details that complement the main paper. Specifically, we include:

- Qualitative comparisons between SFT, SCALE, reasoning traces illustrating improved factual grounding for Gemma-3-4b and Llava-next.
- Implementation Details for SFT and GRPO.

B. Additional Qualitative Results

Figures 1 - 2 present additional examples comparing the reasoning structures produced by the different STF variations from the models compared in the main paper. All examples are taken from HF test set lmms-lab/ScienceQA-IMG dataset. Blue bottom right box is reasoning trace from vanilla-SFT and orange bottom left shows SCALE-SFT reasoning.

Our SCALE model demonstrates more coherent intermediate reasoning steps, encourages the model to complete its reasoning trace coherently and converge toward the correct conclusion, resulting in more reliable `<think>` endings and answer transitions compared to standard SFT. This can be seen by fig 1, where vanilla SFT goes into a reasoning loop, where it endlessly debates about the answer and fails to reach a final conclusion. In contrast, fig 2 shows the vanilla SFT often can produce coherent reasoning trace but incorrect final answer, reflecting the token imbalance that biases optimization toward verbose reasoning segments.

C. Implementation Details

Training. All models were trained using Deepspeed-ZeRO2 on $8 \times H100$ GPUs. SFT training ran for one epoch over the respective Vision-R1 subsets. We used mixed precision (`bfloat16`) and gradient checkpointing. All models were fine-tuned using LoRA adapters with a learning rate of 1×10^{-4} and a maximum sequence length of 8192 tokens. For the RL stage, we employed GRPO with a maximum generation length of 1024 tokens, generating 8 completions per prompt using sampling with temperature 1.0 ($\text{top-}p = 1.0$, no $\text{top-}k$ filtering). Inference was accelerated using vLLM in colocate mode.

Evaluation. We evaluated all models using the `lmms-eval` framework. To accommodate reasoning-based responses, we extended `lmms-eval` to parse the final answer from within the `<answer>` tags. All metrics were calculated exclusively on the extracted final answers. We used greedy decoding (temperature 0) for all evaluations to ensure deterministic and reproducible results.

Question: Which animal is also adapted to use its neck to appear large and scary to a predator?

Choices: ['bearded dragon', 'green anole']

Answer: A



Vanilla SFT

<think> Okay, let's see. The question is asking which animal, besides the spectacled cobra, also uses its neck to appear large and scary to a predator. The options are A. bearded dragon and B. green anole. From the context given, the spectacled cobra uses its neck to appear large and scary. The answer should be another animal that does the same. I remember that bearded dragons and green anoles are both lizards. Lizards, in general, can puff up their bodies, including their necks, to look bigger. For example, bearded dragons sometimes puff up their necks to appear larger, which might scare off predators. Green anoles are known for their ability to flatten their bodies and raise their necks, which can make them look bigger. So both bearded dragons and green anoles are likely to use their necks in a similar way. The question is asking which animal is also adapted to do this. Since both options are lizards that can puff up their necks, the answer should be either A or B. The image also shows a spectacled cobra, so the answer is another animal that does the same. The options are A and B, so the answer is either A or B. Since the question is asking which animal is also adapted, both A and B are options. However, the question is phrased as "which animal is also adapted," so the answer would be either A or B. But the options are A and B. The answer should be either A or B. Since the question is asking which animal is also adapted, both A and B are options. But the question is phrased as "which animal..."

SFT-SWL

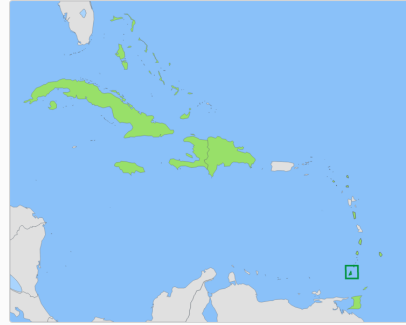
<think> Okay, let's see. The question is asking which animal, besides the spectacled cobra, uses its neck to appear large and scary to a predator. The options are A. bearded dragon and B. green anole. From the image given, the spectacled cobra uses its neck to appear large and scary. The question is asking which other animal does this. The bearded dragon and green anole are both reptiles, but I need to recall if either of them uses their neck in a similar way. I remember that bearded dragons, especially males during mating displays, can puff up their necks to appear larger. They also have crests that they can raise or lower. Green anoles, on the other hand, are more about camouflage and quick movements. They don't have prominent neck displays like bearded dragons. So the bearded dragon is more likely to use its neck for a display, while the green anole doesn't. Therefore, the answer should be A. bearded dragon.
</think>
<answer> Final Answer: A. </answer>

Figure 1. Reasoning output from Gemma3, from HF test set lmms-lab/ScienceQA-IMG dataset. Blue bottom right box is reasoning from vanilla-SFT. Orange bottom left is SCALE-SFT reasoning. Vanilla SFT goes into an endless loop and cannot reach a final answer (shortened with ...) while SCALE-SFT has concise reasoning steps and accurate final answer.

Question: Which country is highlighted?

Choices: ['Saint Vincent and the Grenadines', 'Grenada', 'the Dominican Republic', 'Trinidad and Tobago']

Answer: B



Vanilla SFT

<think> Okay, let's see. The question is asking which country is highlighted on the map of the Caribbean. The image illustrates depicts the highlighted area is in the southern part of the Lesser Antilles, near the northeastern coast of South America. The options are Saint Vincent and the Grenadines, Grenada, the Dominican Republic, and Trinidad and Tobago. First, I need to recall the geography of the Caribbean. The Lesser Antilles are a group of islands in the Caribbean Sea. The southern part of the Lesser Antilles includes countries like Grenada, Saint Vincent and the Grenadines, and Trinidad and Tobago. The Dominican Republic is on the island of Hispaniola, which is more to the north, part of the Greater Antilles. The image illustrates displays the highlighted area is near the northeastern coast of South America. Trinidad and Tobago are located just off the coast of Venezuela, which is in South America. So if the highlighted area is near South America, that points to Trinidad and Tobago. The other options are further north in the Lesser Antilles. Grenada is also in the southern part but a bit more to the north compared to Trinidad and Tobago. Saint Vincent and the Grenadines are further north as well. So the answer should be D. Trinidad and Tobago. </think>

<answer> Final Answer: D. </answer>

SFT-SWL

<think> Okay, let's see. The question is asking which country is highlighted on the map of the Caribbean. The image illustrates depicts the highlighted country is in the southern Caribbean, near the northeastern coast of South America. It's south of Saint Vincent and the Grenadines and north of Trinidad and Tobago. The options are A to D. First, I need to recall the geography of the Caribbean. The southern Caribbean includes countries like Venezuela, Colombia, and the islands of the Lesser Antilles. The Grenadines are part of Saint Vincent and the Grenadines, which is one of the Windward Islands. So if the highlighted country is south of Saint Vincent and the Grenadines, that would place it in the southern part of the Lesser Antilles. Trinidad and Tobago are further south, near Venezuela. The Dominican Republic is on Hispaniola, which is more to the north, near Haiti. So the country between Saint Vincent and the Grenadines and Trinidad and Tobago would be Grenada. The answer should be B. </think>

<answer> Final Answer: B. </answer>

Figure 2. Reasoning output from Llava-next, from HF test set lmms-lab/ScienceQA-IMG dataset. Blue bottom right box is reasoning from vanilla-SFT. Orange bottom left is SCALE-SFT reasoning. In vanilla SFT the token imbalance can lead to well structured reasoning traces that fail to align with the final answer supervision- the model “thinks correctly” but concludes incorrectly.