# You May Speak Freely: Improving the Fine-Grained Visual Recognition Capabilities of MLLMs with Answer Extraction

Logan Lawrence Oindrila Saha Megan Wei Chen Sun Subhransu Maji Grant Van Horn

University of Massachusetts, Amherst Brown University

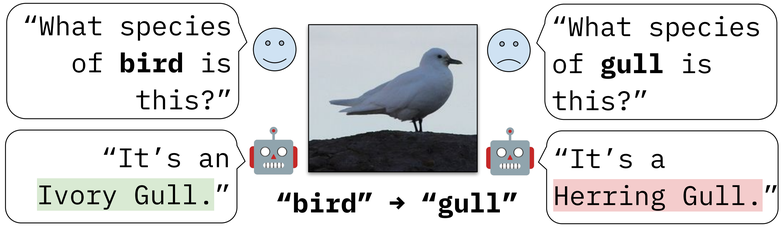
ICCV OCT 19-23, 2025 HONOLULU HAWAII

## Problem

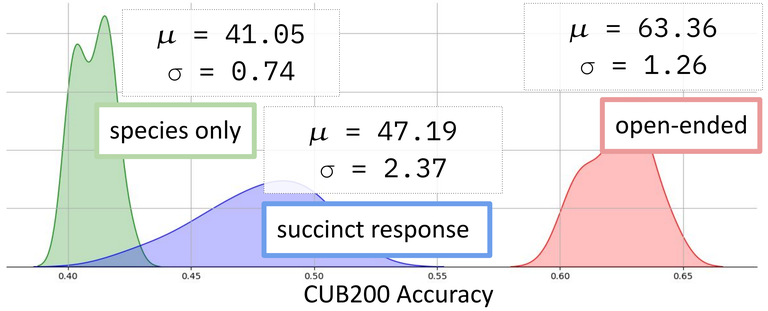
**P1:** FGVC choice counts **are too big** for classic MCQ evaluation.

| Image: | Coarse (# classes ≤ 5): | Fine (# classes ≫ 200): |
|---|---|---|
| | (a) "Bird" | (a)"Gray Catbird" |
| | (b) "Cat" | (b)"N. Mockingbird" |
| | (c) "Dog" | … |
| | (d) "Airplane" | (eb) "Ivory Gull" |

**P2:** MLLM outputs are **not robust to small changes.**




**P3:** These changes result in **large performance differences.**




## Method

15x ~ "What is the species of this bird?"

**free-form**

"This bird is an Ivory Gull."

"What species is indicated?"

**constrained decoding**

c: Ivory Gull

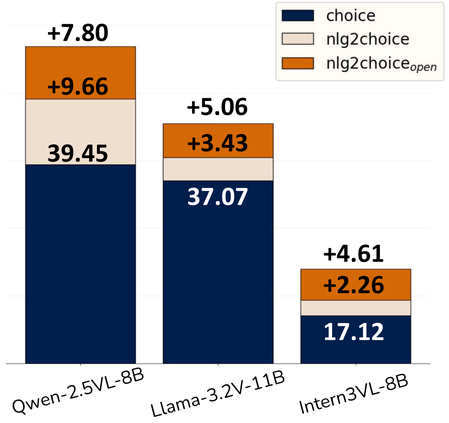
## Datasets

Benchmarked on **7 FGVC datasets:** CUB200, Flowers102, FGVC Aircrafts, Stanford Cars, Food101, NABirds, iNat21-Birds

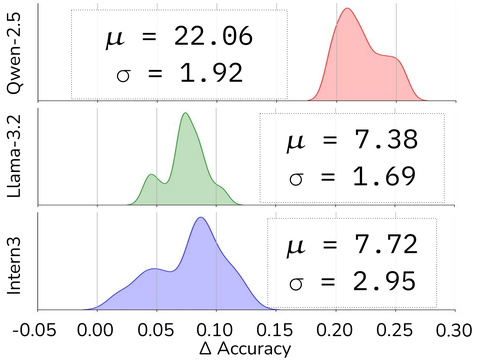
See project repo for our small labeled **answer extraction dataset for fine-grained species!**

## Results

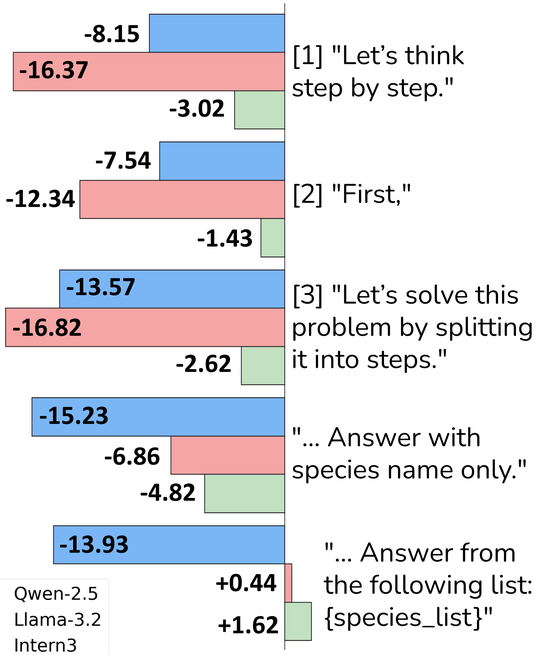
**R1:** Answer extraction **improves zero-shot fine-grained visual classification.**




**R2:** With writing as source of randomness, results are **statistically significant.**




**R3:** Using CoT-inducing prompts or more specific instructions **degrades performance.**




## References

[1] Kojima et al. Large Language Models are Zero-Shot Reasoners.
[2] Ahn et al. Do as I Can, Not as I Say: Grounding Language in Robotic Affordances.
[3] Reynolds and McDonell, et al. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm.