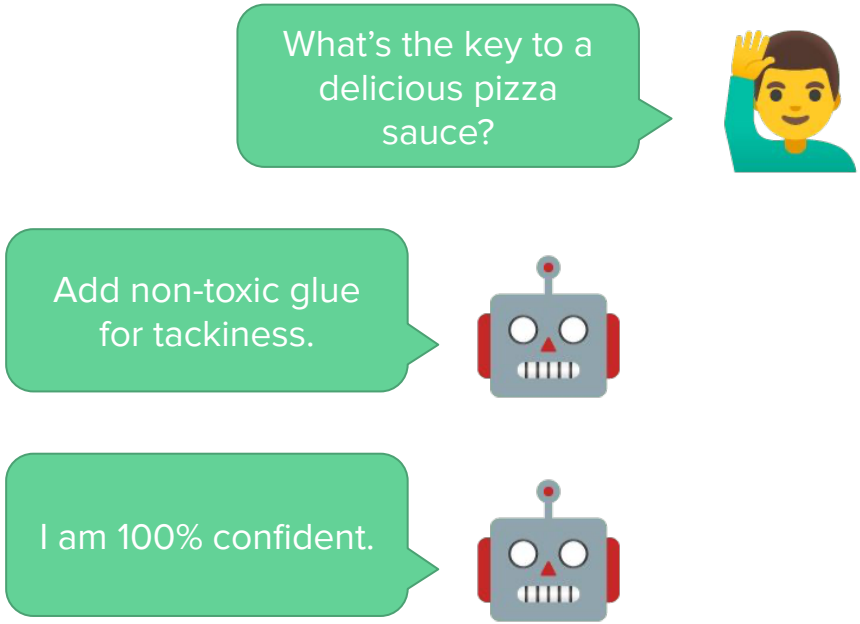
NEURAL INFORMATION PROCESSING SYSTEMS

Large Language Models Must Be Taught to Know What They Don't Know

Sanyam Kapoor*, Nate Gruver*,
Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, Andrew Gordon WIlson

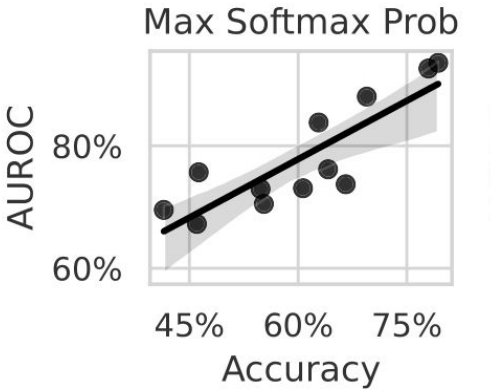
NEW YORK UNIVERSITY ABACUS.AI UNIVERSITY OF CAMBRIDGE COLUMBIA

Answers without Confidence are not Actionable

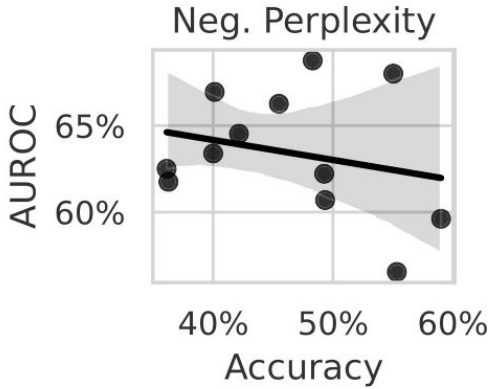


Can LLMs accurately represent uncertainty?

In principle, the autoregressive model likelihood (e.g. perplexity) is a valid score of confidence.



Depends on model scale/performance!



Does not correlate well with model scale/performance!

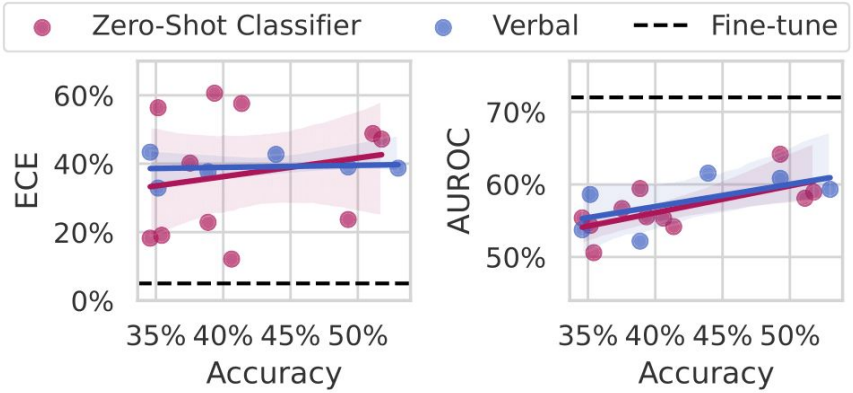
Our Work

- Do we get good uncertainties out of the box?
- How should we use examples labeled with correctness to provide supervision?
- Does calibrated uncertainty help Human-AI collaboration?

Do we get good uncertainties out-of-the-box?

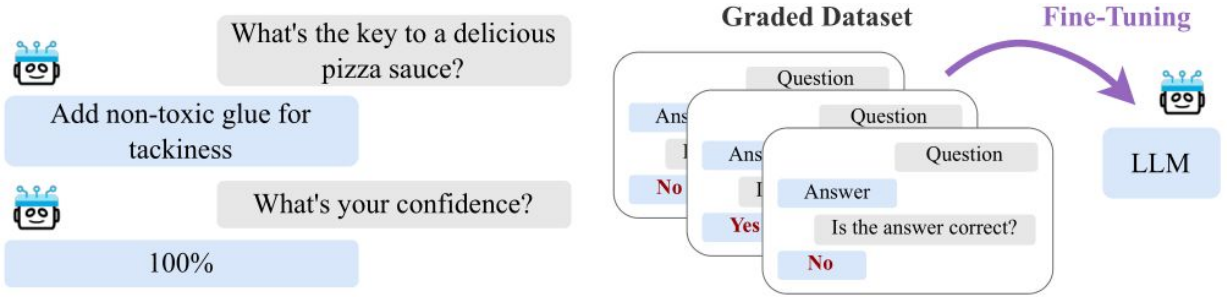
Name	Format	Confidence
Zero-Shot Classifier	"Question. Answer. **True/False**: **True**"	P("**True**") / (P("**True**") + P("**False**"))
Verbalized	"Question. Answer. **Confidence**: **90%**"	float("**90%**")

Do we get good uncertainties out-of-the-box?



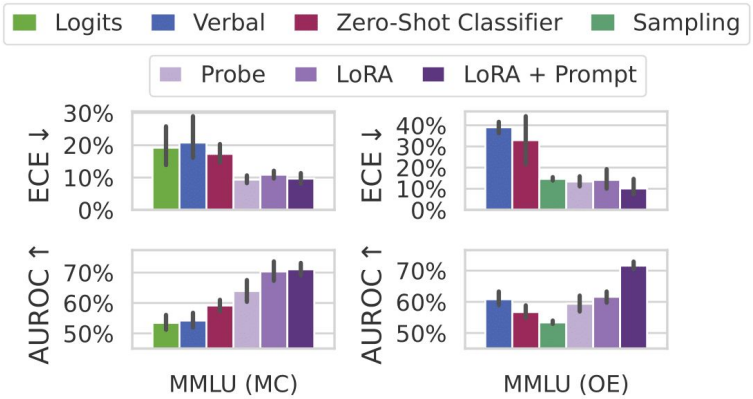
Black-box methods such as perplexity or engineered prompts have limited predictive power and scale slowly, or not at all, with the power of the base model.

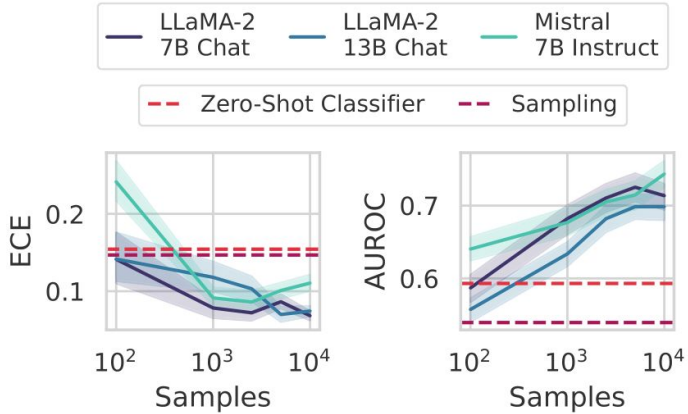
How do we use labeled examples?



Learn the probability of correctness through various parameterizations (Probe, LoRA, LoRA + Prompt) of a classifier.

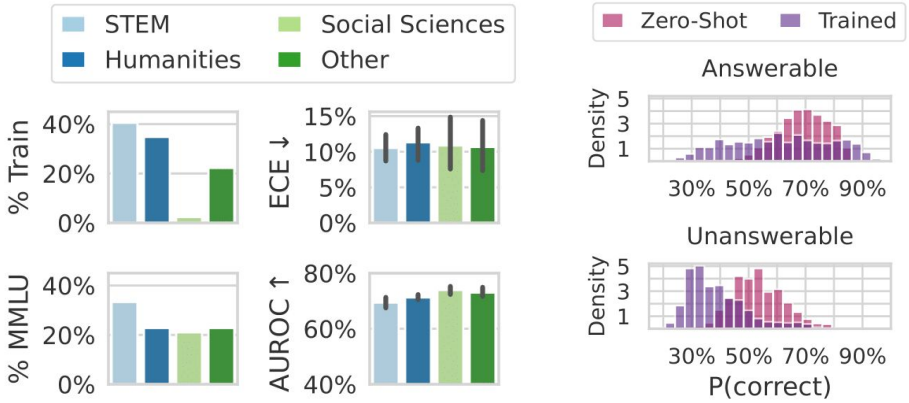
How do we use labeled examples?





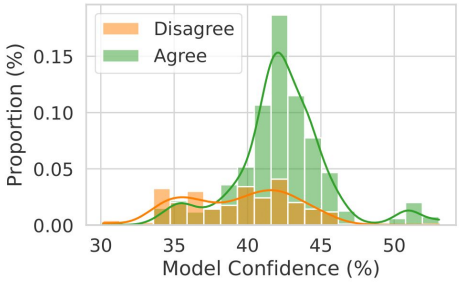
Supervised learning approaches can dramatically outperform baselines with as few as 1000 graded examples.

When and how do these estimates generalize?

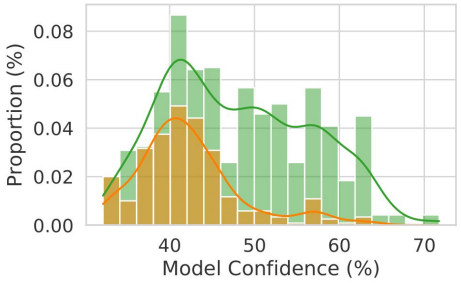


Learned uncertainty estimates generalize to new formatting, subject matter, and even the generations of other models. This generalization does not appears to stem simply from judging a question's difficulty based on its subject matter (a short-cut).

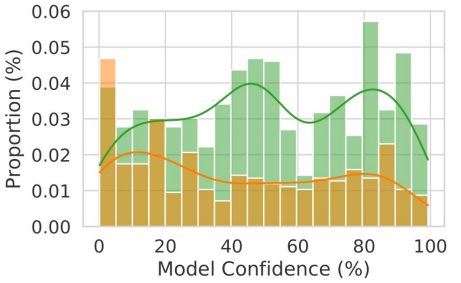
Does Calibrated Uncertainty help Human-AI Collaboration?



Zero-Shot Query Classifier



Fine-Tuned Classifier



Random Predictor (Control)

Users are sensitive to confidence scores and use their relative magnitude to modulate their decision to use an LLM.

And More!

huggingface.co/calibration-tuning

github.com/activatedgeek/calibration-tuning

arxiv.org/abs/2406.08391