Calibrating Large Language Models Using Their Generations Only

Dennis Ulmer[1,2,3] Martin Gubri[1] Hwaran Lee[4] Sangdoo Yun[4] Seong Joon Oh[1,5,6]

[1]Parameter Lab [2]IT University of Copenhagen [3]Pioneer Centre for Artificial Intelligence [4]NAVER AI Lab
[5]University of Tübingen [6]Tübingen AI Center

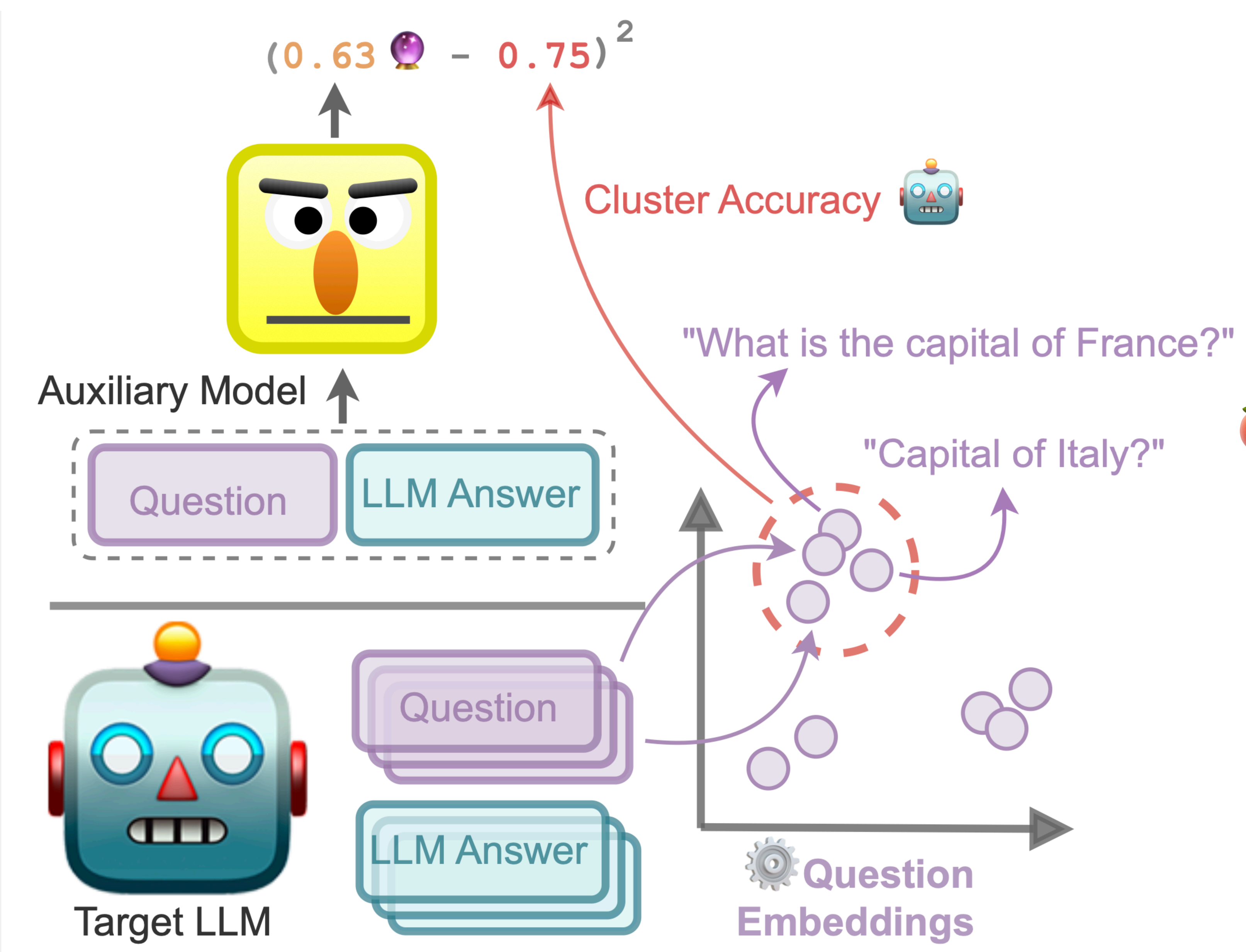
IT UNIVERSITY OF CPH · PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE · Parameter Lab · Tübingen AI Center tuebingen.ai · NAVER AI LAB · EBERHARD KARLS UNIVERSITÄT TÜBINGEN

BERT-based models can predict black-box LLM confidence based on the question & answer text

Motivation

- LLMs require techniques like confidence estimation to quantify trustworthiness of predictions
- But many commercial LLM are black-boxes behind APIs!
- We propose **APRICOT** 🍑:
  1. **Creating calibration targets** in an unsupervised way and
  2. Training **an auxiliary model to predict** target LLM **confidence scores from its text answers**

Method



- Create question & LLM answer embeddings with SentenceBERT and cluster with HDBSCAN
- Compute cluster accuracy as calibration target
- Finetune auxiliary model (DeBERTa v3) to predict confidence based on question + LLM answer text

Results

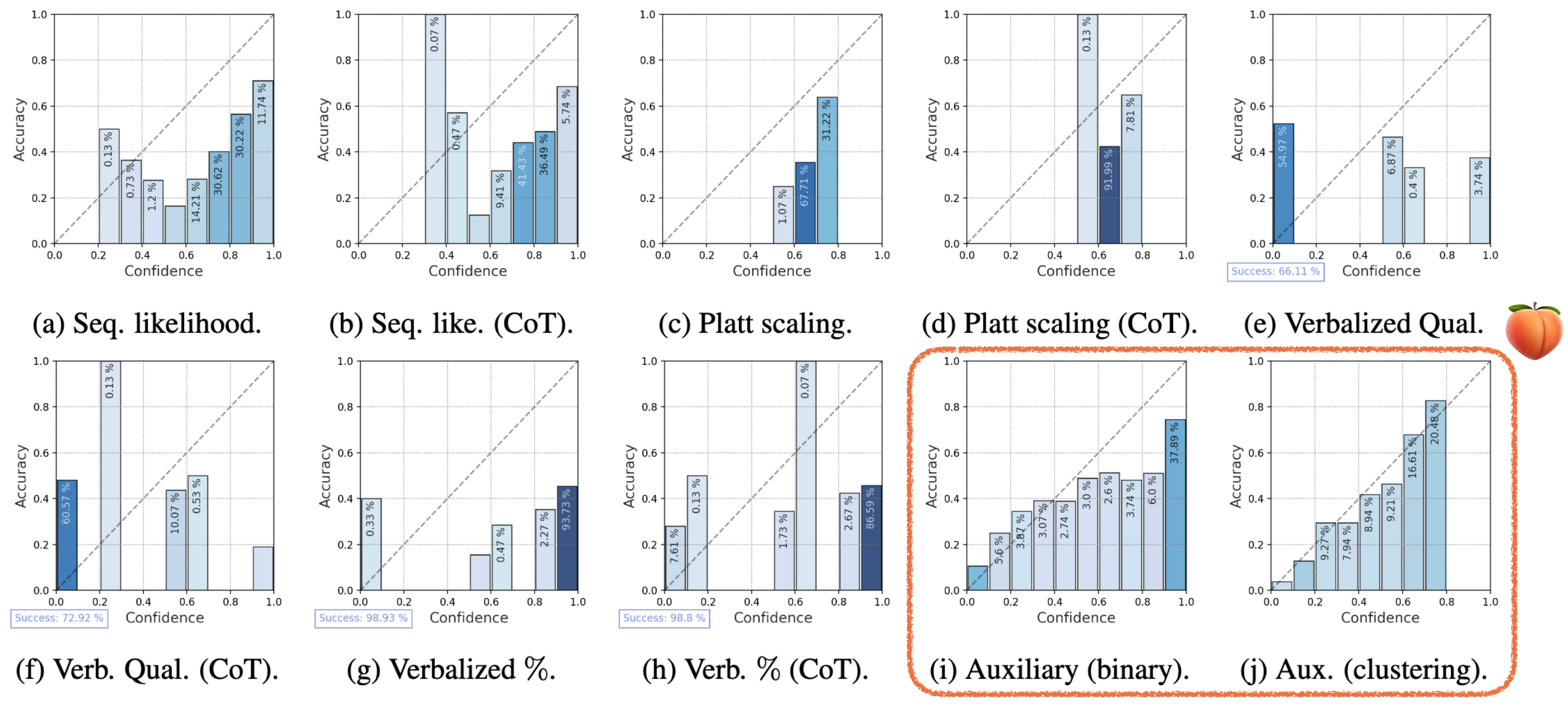


Figure 9: Reliability diagrams for our different methods using 10 bins each for Vicuna v1.5 7B on CoQA. The color as well as the percentage number within each bar indicate the proportion of total points contained in each bin.

		TriviaQA					CoQA				
	Method	Success	Brier↓	ECE↓	smECE↓	AUROC↑	Success	Brier↓	ECE↓	smECE↓	AUROC↑
GPT-3.5 (black-box)	Seq. likelihood	-	.15 ±.01	.04 ±.00	.04 ±.00	.69 ±.02	-	.29 ±.01	.11 ±.00	.11 ±.00	.70 ±.01
	Seq. likelihood (CoT)	-	.14 ±.00	.05 ±.00	.05 ±.00	.60 ±.02	-	.25 ±.00	**.01** ±.00	**.02** ±.00	.52 ±.02
	Platt scaling	-	.15 ±.00	.04 ±.00	.04 ±.00	.69 ±.02	-	.26 ±.01	.03 ±.00	.03 ±.00	.70 ±.01
	Platt scaling (CoT)	-	.15 ±.00	.12 ±.00	.12 ±.00	.60 ±.02	-	.25 ±.00	.06 ±.00	.06 ±.00	.52 ±.02
	Verbalized Qual.	1.00	.14 ±.01	.07 ±.00	.04 ±.00	.61 ±.02	1.00	.27 ±.00	.07 ±.00	.05 ±.00	.52 ±.01
	Verbalized Qual. (CoT)	1.00	.15 ±.00	.04 ±.00	.03 ±.00	.63 ±.02	1.00	.30 ±.01	.08 ±.01	.04 ±.00	.50 ±.01
	Verbalized %	1.00	.13 ±.01	.01 ±.00	**.01** ±.00	.63 ±.02	1.00	.34 ±.01	.25 ±.00	.22 ±.00	.54 ±.01
	Verbalized % (CoT)	0.99	.13 ±.01	**.00** ±.00	**.01** ±.00	.63 ±.02	0.58	.37 ±.01	.09 ±.01	.06 ±.00	.49 ±.02
🍑	Auxiliary (binary)	-	.14 ±.00	.14 ±.01	.14 ±.01	.65 ±.02	-	.19 ±.01	.13 ±.01	.13 ±.01	**.81** ±.01
	Auxiliary (clustering)	-	**.12** ±.01	.06 ±.01	.06 ±.01	**.72** ±.02	-	**.18** ±.00	.02 ±.01	**.02** ±.00	**.81** ±.01

Table 3: Calibration results for Vicuna v1.5 and GPT-3.5 on TriviaQA and CoQA. We bold the best results per dataset and model, and underline those that are statistically significant compared to all other results assessed via the ASO test. Results are reported along with a bootstrap estimate of the standard error.

- Test on TriviaQA and CoQA with Vicuna v1.5 7B & GPT-3.5
- APRICOT 🍑 achieves low calibration error and the best AUROC in misprediction detection; improves with clustered calibration targets
- Verbalized uncertainty only better when model is also overwhelmingly right (i.e., the dataset might be too easy)

Conclusions

- APRICOT 🍑 produces calibrated confidence scores for *any* LLM based on input and answers in text form alone
- In the paper: More experiments & analyses, ablation studies

Paper