APRICOT
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

Summary

We propose APRICOT 🍑:

- To predict calibrated confidence score
- From LLM's generated texts only, so suitable for black-box LLMs
- Using an auxiliary model trained on calibrated confidence targets

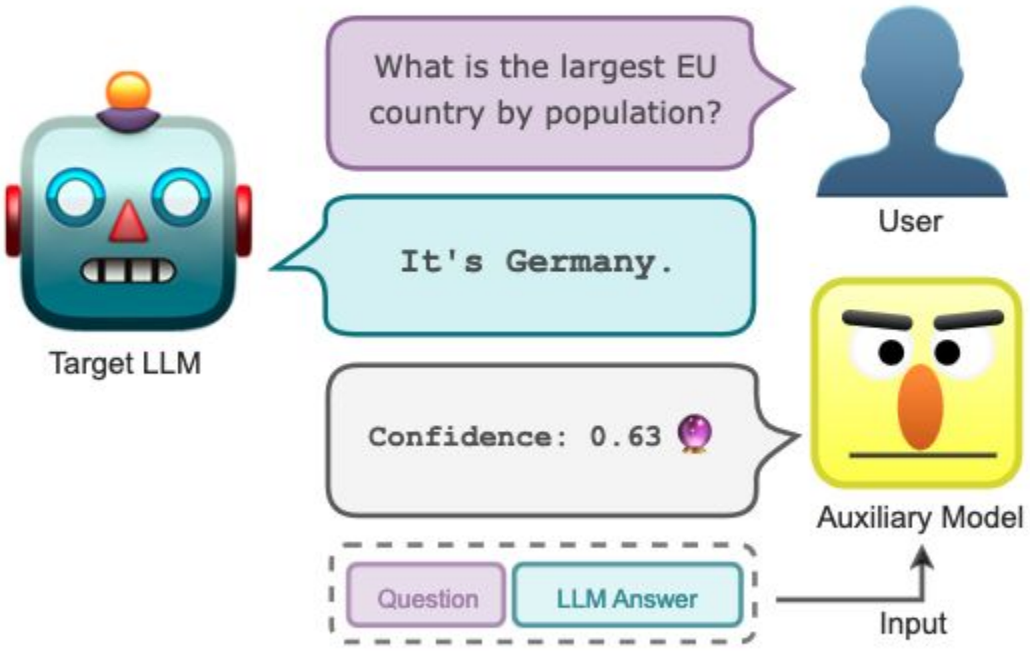
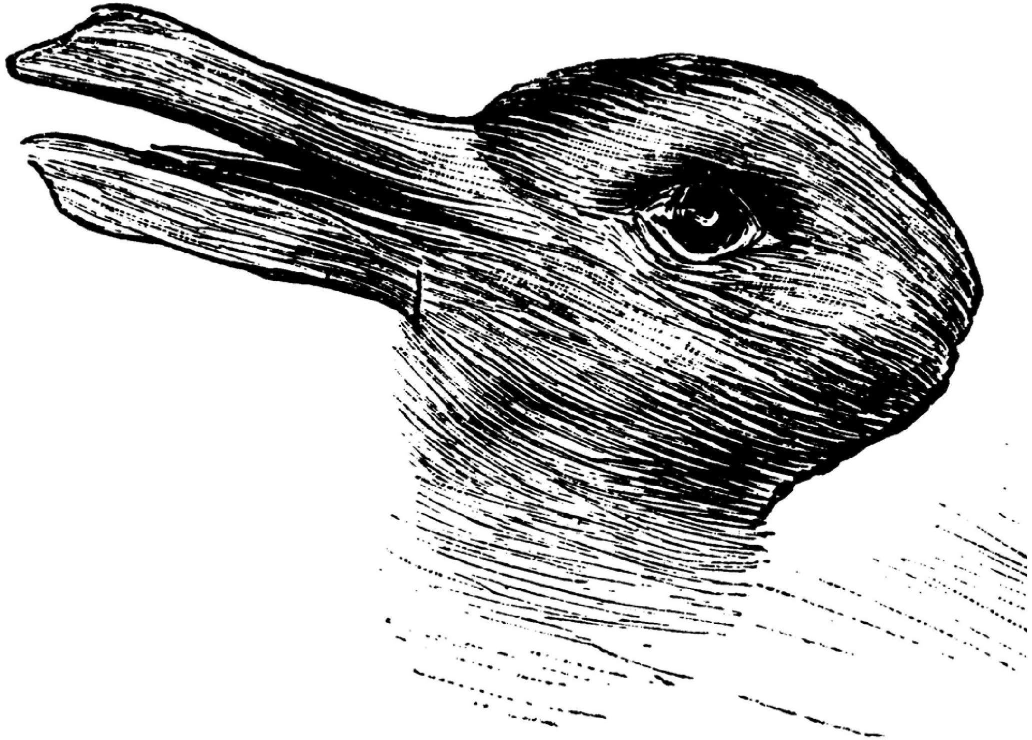


Figure 1: Illustration of APRICOT 🍑: We train an auxiliary model to predict a target LLM's confidence based on its input and the generated answer.

Background on Uncertainty



Aleatoric uncertainty:
Input is inherently ambiguous.

Evaluating Machine Accuracy on ImageNet. ICML 2020.

Background on Uncertainty

Daylight

Night

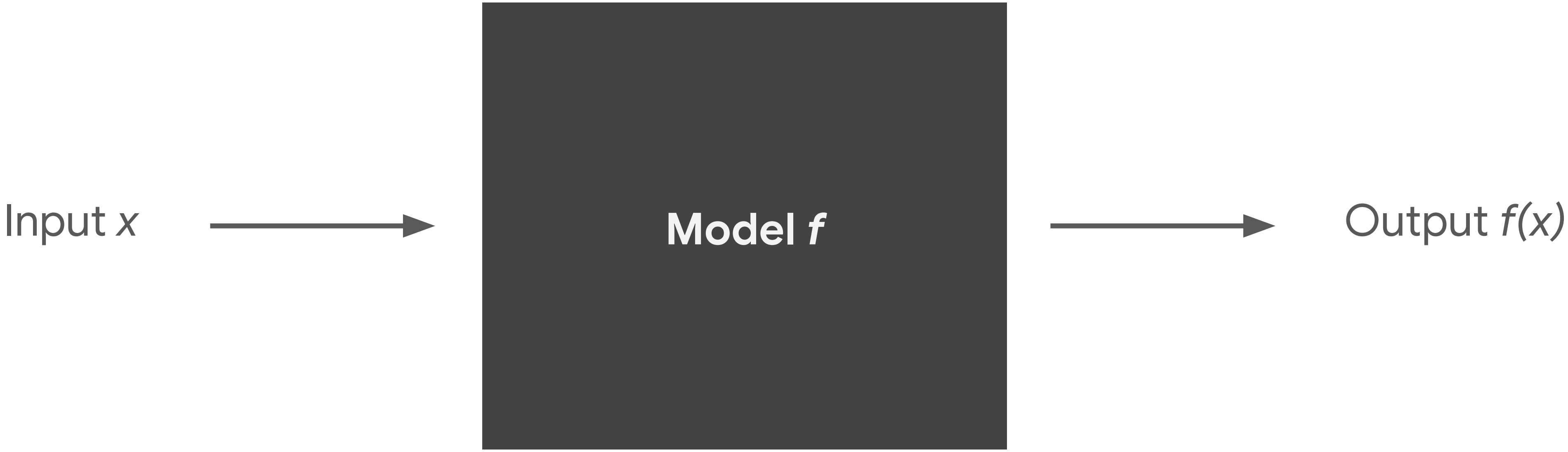
Epistemic uncertainty:

Not trained on similar data.

https://www.gatsby.ucl.ac.uk/~balaji/balaji-uncertainty-talk-cifar-dlrl.pdf

Background on Uncertainty

Simplest form of uncertainty estimate.



Background on Uncertainty

Simplest form of uncertainty estimate.

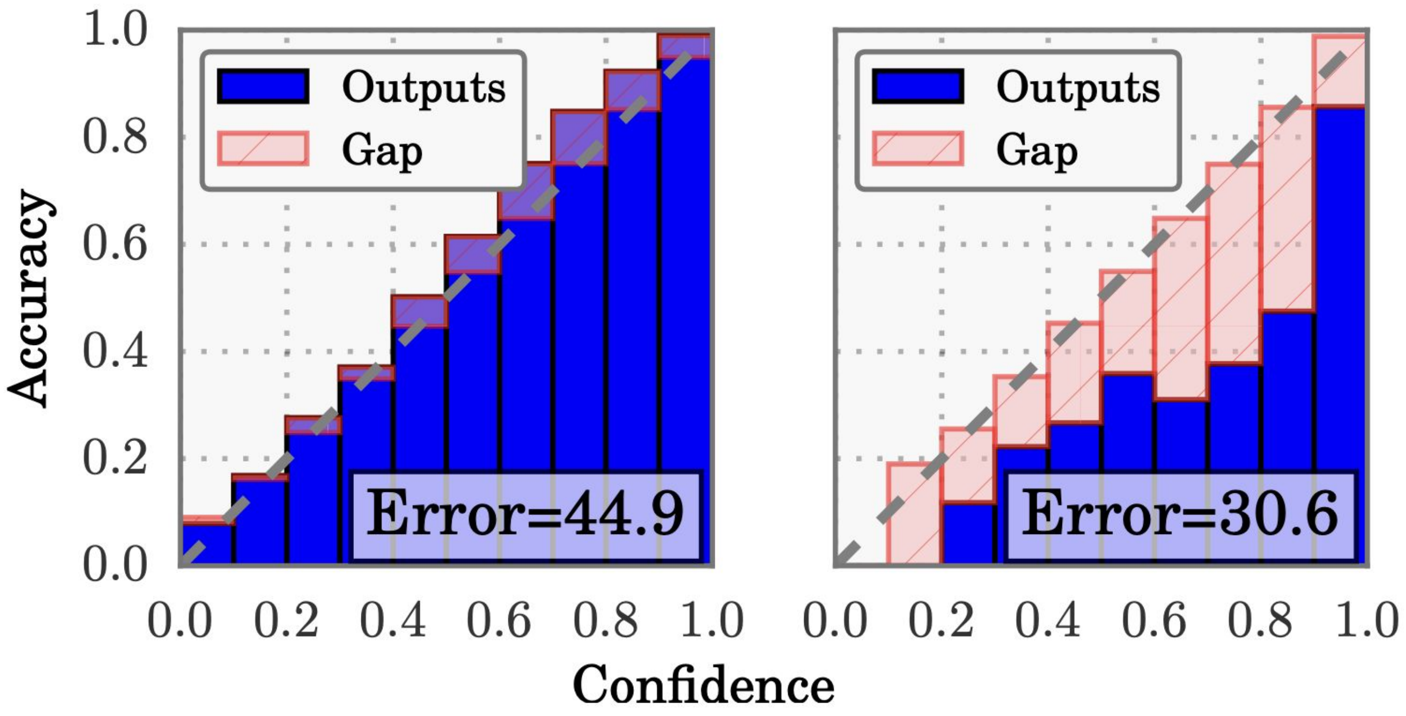
$c(x)$ = **Probability that *f*(*x*) is correct**. $0 \leq c(x) \leq 1$

Background on Uncertainty

Issue: Guo et al. (2017) showed neural nets are overconfident

Calibration: The confidence level should reflect the true predictive uncertainty.

Well-calibrated model

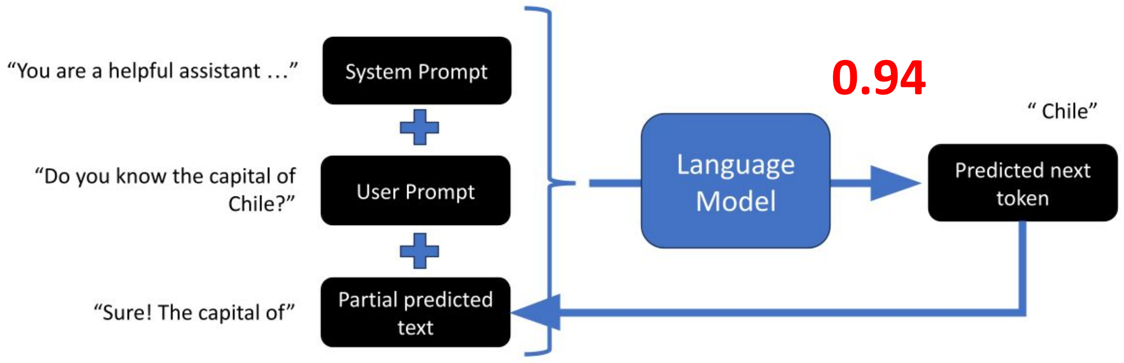


Less well-calibrated model

On Calibration of Modern Neural Networks. ICML 2017.

Confidence Quantification for LLMs

Sequence likelihood



"Sure! The capital of Chile is Santiago."

0.75 0.8 0.99 0.97 0.94 0.99 0.64 → 0.3431

Verbalized uncertainty

Research Question

We want confidence quantification, that is:

- Calibrated
- Suitable for Black-box LLM
- Consistent

Method	Black-box LLM?	Consistent?	Calibrated?
Seq. likelihoods	✗	✔	✗
Verb. uncertainty	✔	✗	✗
APRICOT 🍑 (ours)	✔	✔	✔

Table 1: Comparison of appealing attributes that LLM confidence quantification techniques should fulfil. They should ideally be applicable to black-box LLMs, be consistent (i.e., always elicit a response), and produce calibrated estimates of confidence.

🍑 APRICOT

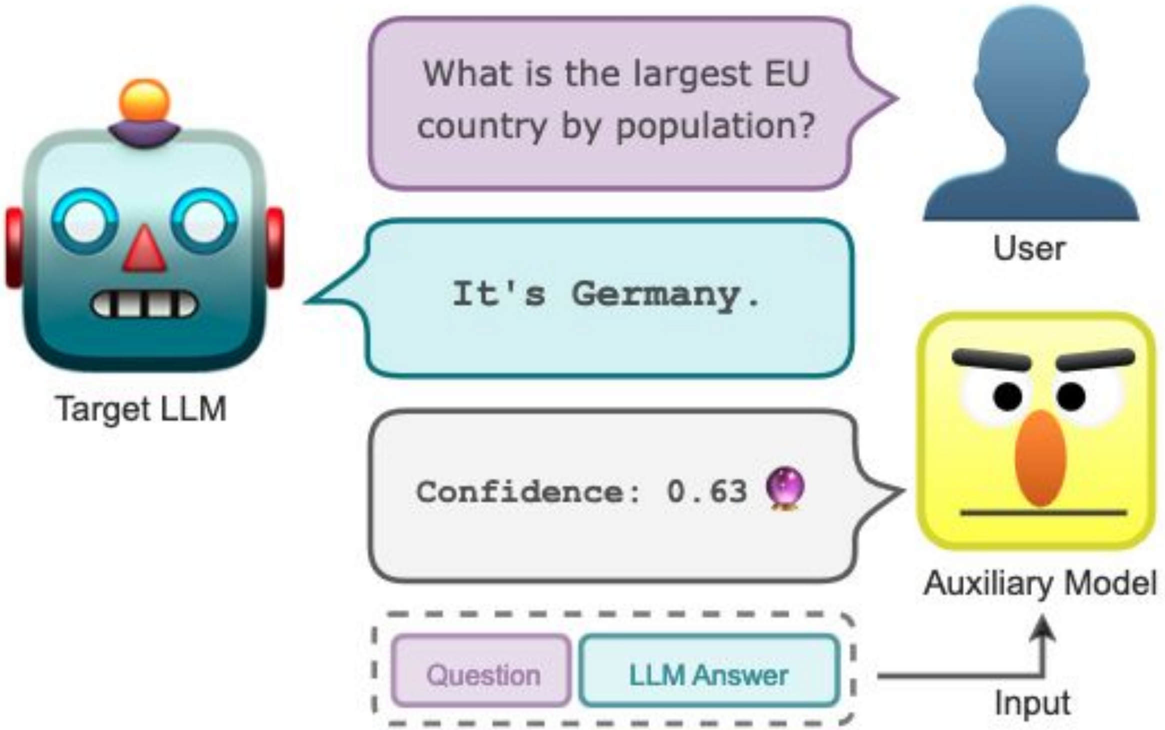


Figure 1: Illustration of APRICOT 🍑: We train an auxiliary model to predict a target LLM's confidence based on its input and the generated answer.

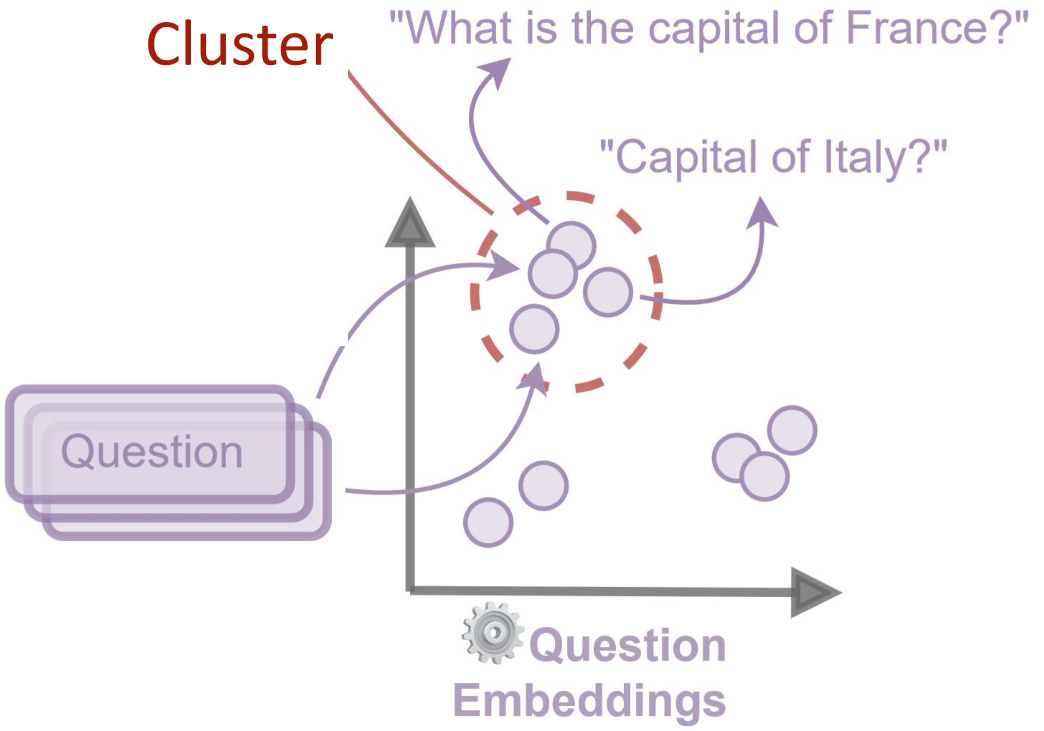
🍑 APRICOT

Receipt:

a) Clustering of questions

	TriviaQA		CoQA	
	Textual	Semantic	Textual	Semantic
Random	.11 ±.08	.00 ±.08	.08 ±.12	.00 ±.12
Clustering	.39 ±.28	.60 ±.14	.47 ±.25	.70 ±.17

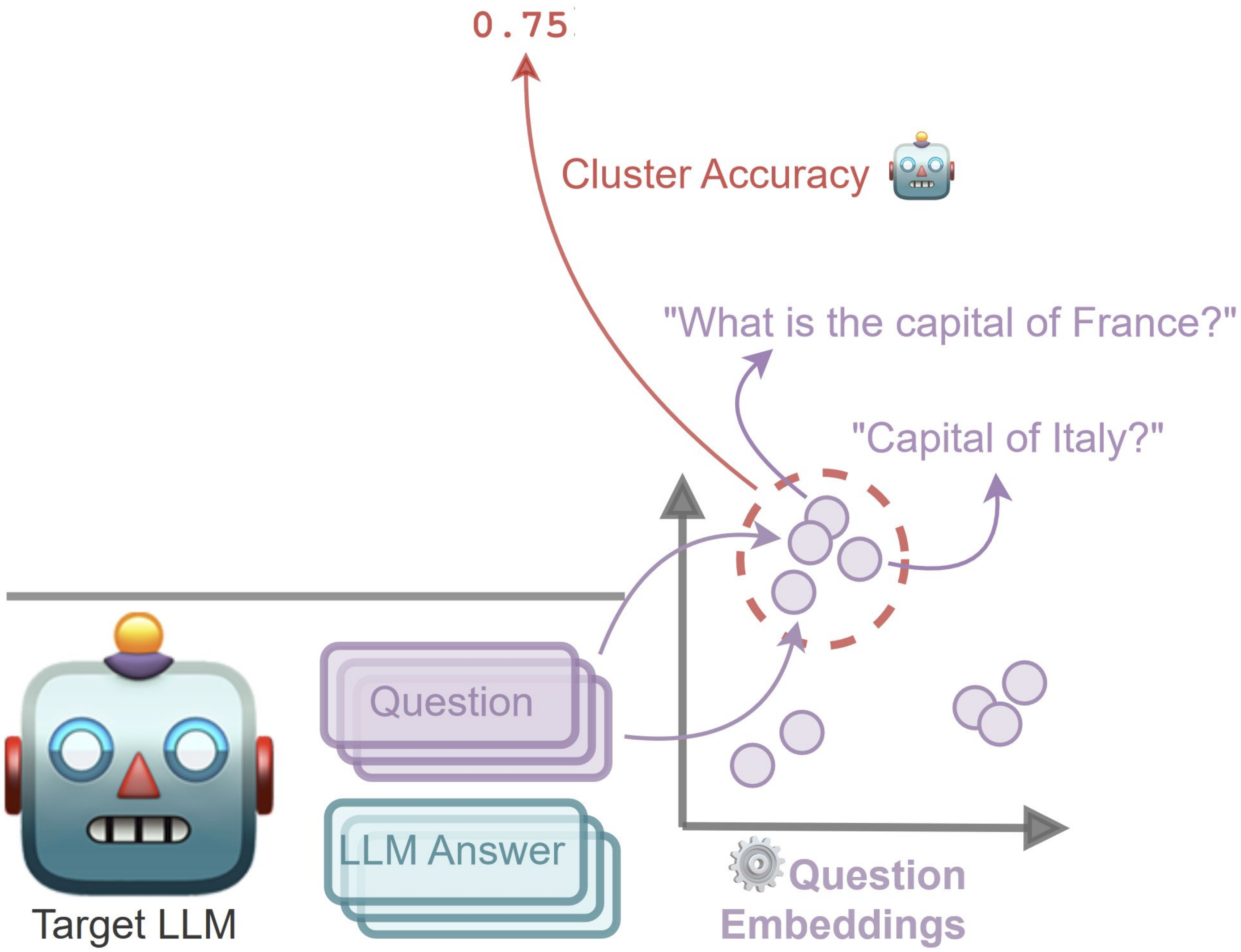
Figure 5: Results of evaluation of found clusters on TriviaQA and CoQA, including one standard deviation.



🍑 APRICOT

Receipt:

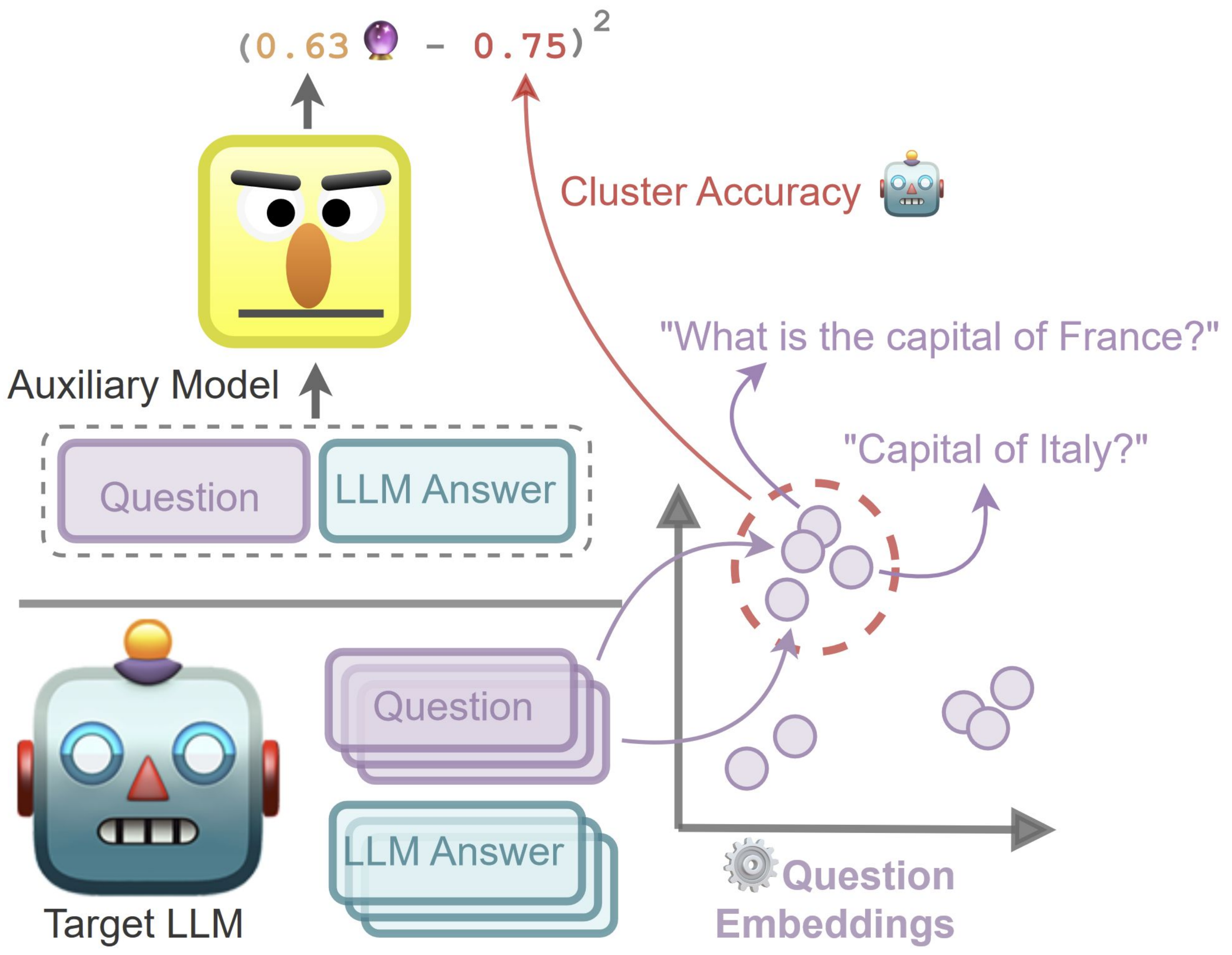
a) Clustering of questions
b) Calibration target



🍑 APRICOT

Receipt:

a) Clustering of questions
b) Calibration target
c) Train auxiliary model
   i) Input: text only
   ii) Output: cluster accuracy



Results

Best Brier scores and misprediction AUROCs

Verbalized confidence, sometimes better on (smooth)ECE,
but also not reliable on Vicuna-7B

		TriviaQA					CoQA				
	Method	Success	Brier↓	ECE↓	smECE↓	AUROC↑	Success	Brier↓	ECE↓	smECE↓	AUROC↑
Vicuna v1.5 (white-box)	Seq. likelihood	-	.22 ±.01	.05 ±.00	.03 ±.00	.79 ±.01	-	.32 ±.01	.08 ±.00	.08 ±.00	.69 ±.01
	Seq. likelihood (CoT)	-	.25 ±.01	.04 ±.00	.04 ±.00	.70 ±.01	-	.35 ±.01	.04 ±.00	.05 ±.00	.61 ±.01
	Platt scaling	-	.24 ±.00	.08 ±.00	.07 ±.00	.70 ±.01	-	.30 ±.00	.03 ±.00	.03 ±.00	.69 ±.01
	Platt scaling (CoT)	-	.24 ±.00	.12 ±.00	.11 ±.00	.79 ±.01	-	.30 ±.00	.02 ±.00	.02 ±.00	.61 ±.01
	Verbalized Qual.	0.19	.38 ±.03	.02 ±.00	.02 ±.00	.62 ±.03	0.66	.45 ±.01	**.00** ±.00	**.00** ±.00	.48 ±.01
	Verbalized Qual. (CoT)	0.25	.39 ±.02	**.01** ±.00	**.01** ±.00	.60 ±.02	0.73	.45 ±.01	**.00** ±.00	**.00** ±.00	.48 ±.01
	Verbalized %	1.00	.39 ±.01	.38 ±.00	.27 ±.00	.52 ±.01	0.99	.49 ±.01	.48 ±.00	.32 ±.00	.53 ±.01
	Verbalized % (CoT)	1.00	.39 ±.01	.38 ±.00	.26 ±.00	.49 ±.01	0.99	.48 ±.01	.06 ±.00	.06 ±.00	.55 ±.01
	Auxiliary (binary)	-	.20 ±.01	.16 ±.01	.15 ±.01	.81 ±.01	-	.20 ±.01	.16 ±.01	.15 ±.01	**.82** ±.01
	Auxiliary (clustering)	-	**.18** ±.00	.09 ±.01	.09 ±.01	**.83** ±.01	-	**.18** ±.00	.04 ±.01	.04 ±.01	**.82** ±.01
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.

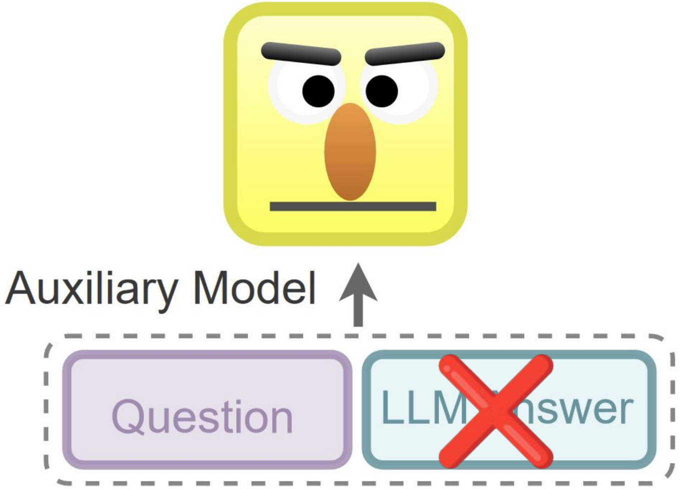
What does the model learn from?

Ablation study

We train the auxiliary model on:

Questions-only (no LLM answer)

- the auxiliary model performs decently
- → learns from the type of question



What does the model learn from?

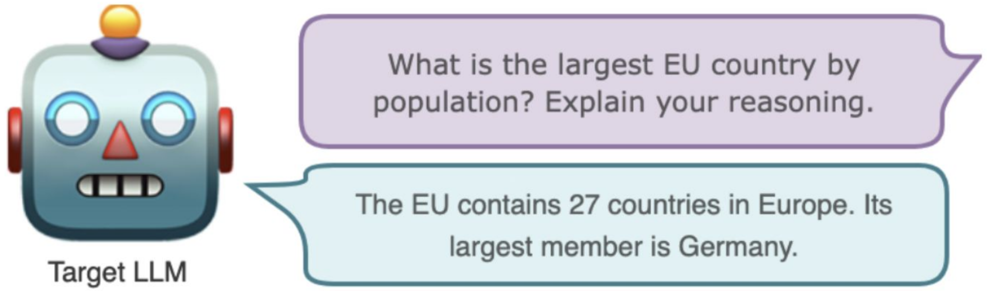
Ablation study

We train the auxiliary model on:

Chain-of-thought prompting

- decreases the calibration error
- → learns a mapping of the model's own assessment to a calibrated confidence score

(a) Default prompting.



(b) Chain-of-though prompting.

Partial Conclusion

APRICOT 🍑:

- Trains an auxiliary model on clusters of homogeneous questions
- Predicts calibrated confidence score
- Can be applied on black-box LLMs

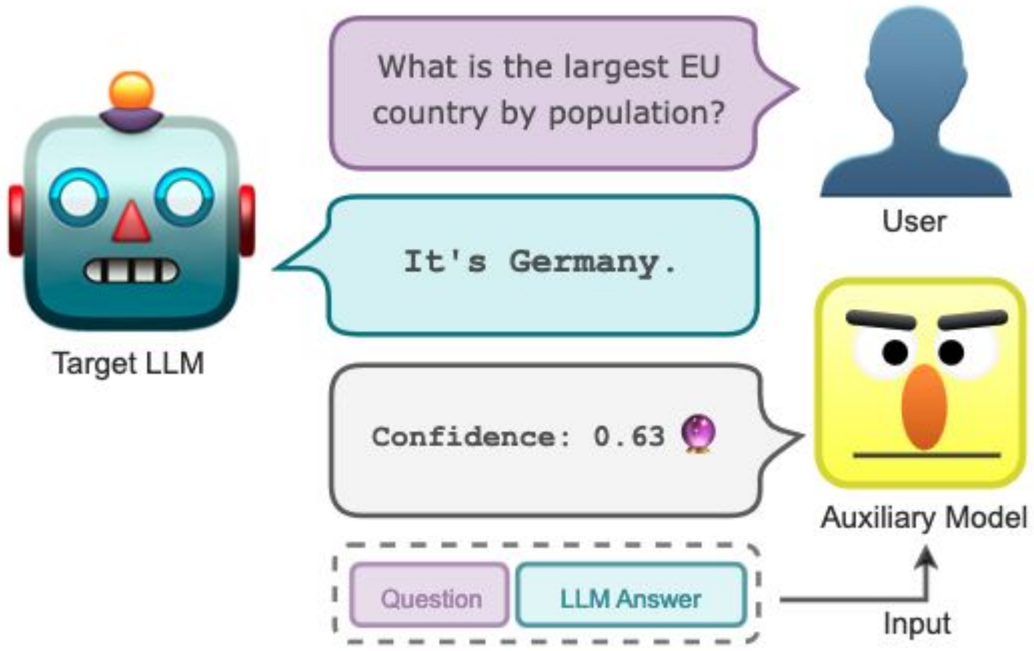


Figure 1: Illustration of APRICOT 🍑: We train an auxiliary model to predict a target LLM's confidence based on its input and the generated answer.