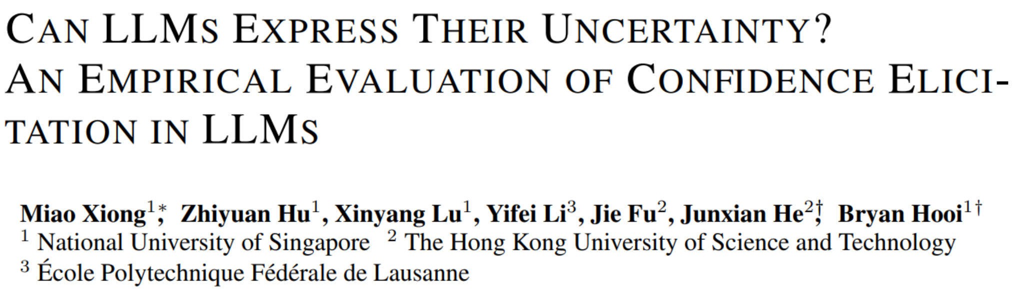
# Confidence Elicitation Attacks

Confidence Elicitation: A New Attack Vector for Large Language Models
ICLR Poster 2025

***Brian Formento[1,2], Chuan Sheng Foo[2,3], See-Kiong Ng[1]***

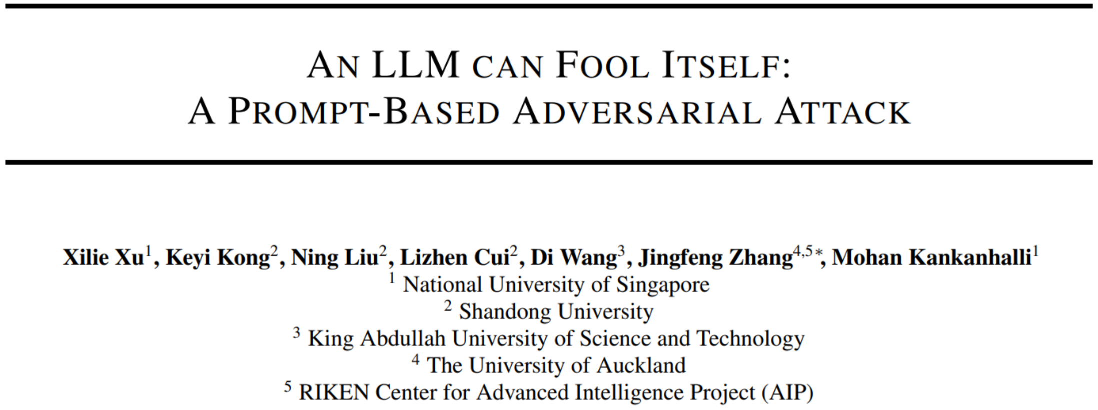
[1] *Institute of Data science, National University of Singapore*
[2] *Institute for Infocomm Research, A*Star*
[3] *Centre for Frontier AI Research, A*Star*

# Confidence Elicitation Attacks



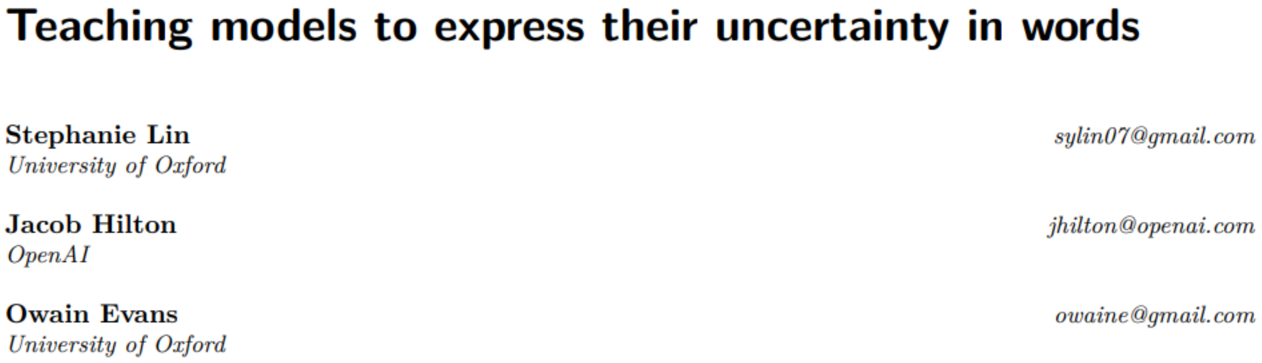
CAN LLMS EXPRESS THEIR UNCERTAINTY?
AN EMPIRICAL EVALUATION OF CONFIDENCE ELICITATION IN LLMS

Miao Xiong[1*], Zhiyuan Hu[1], Xinyang Lu[1], Yifei Li[3], Jie Fu[2], Junxian He[2†], Bryan Hooi[1†]
[1] National University of Singapore [2] The Hong Kong University of Science and Technology
[3] École Polytechnique Fédérale de Lausanne



AN LLM CAN FOOL ITSELF:
A PROMPT-BASED ADVERSARIAL ATTACK

Xilie Xu[1], Keyi Kong[2], Ning Liu[2], Lizhen Cui[2], Di Wang[3], Jingfeng Zhang[4,5*], Mohan Kankanhalli[1]
[1] National University of Singapore
[2] Shandong University
[3] King Abdullah University of Science and Technology
[4] The University of Auckland
[5] RIKEN Center for Advanced Intelligence Project (AIP)



Teaching models to express their uncertainty in words

Stephanie Lin sylin07@gmail.com
*University of Oxford*

Jacob Hilton jhilton@openai.com
*OpenAI*

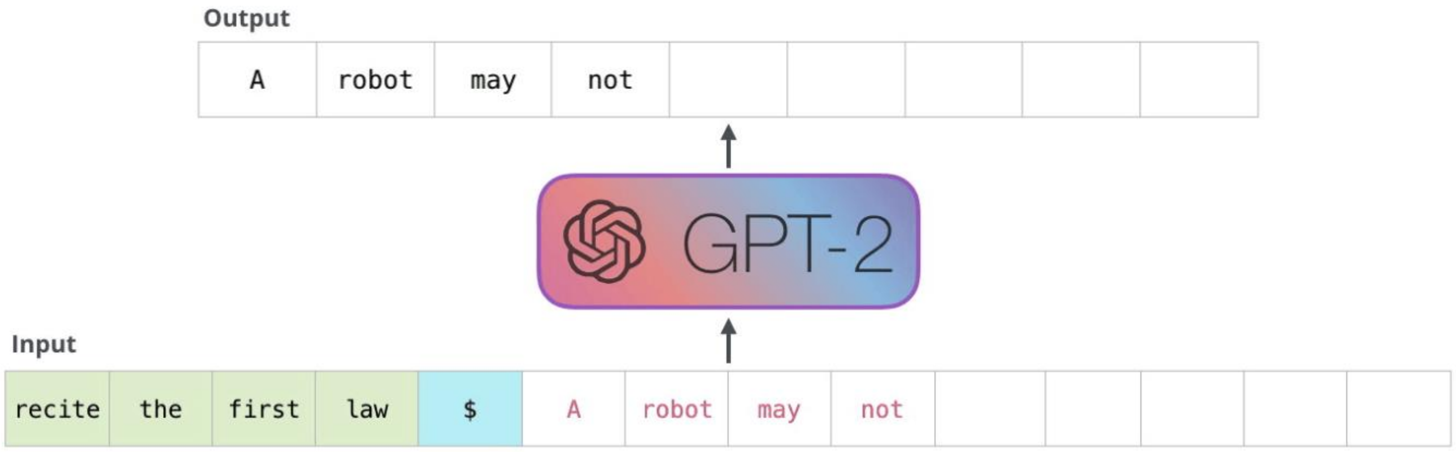
Owain Evans owaine@gmail.com
*University of Oxford*

# Motivation

Closed (black-box) source nature
of LLMs
often used to argue
against white-box/grey-box attacks

# Motivation

But LLMs can do free-form generation

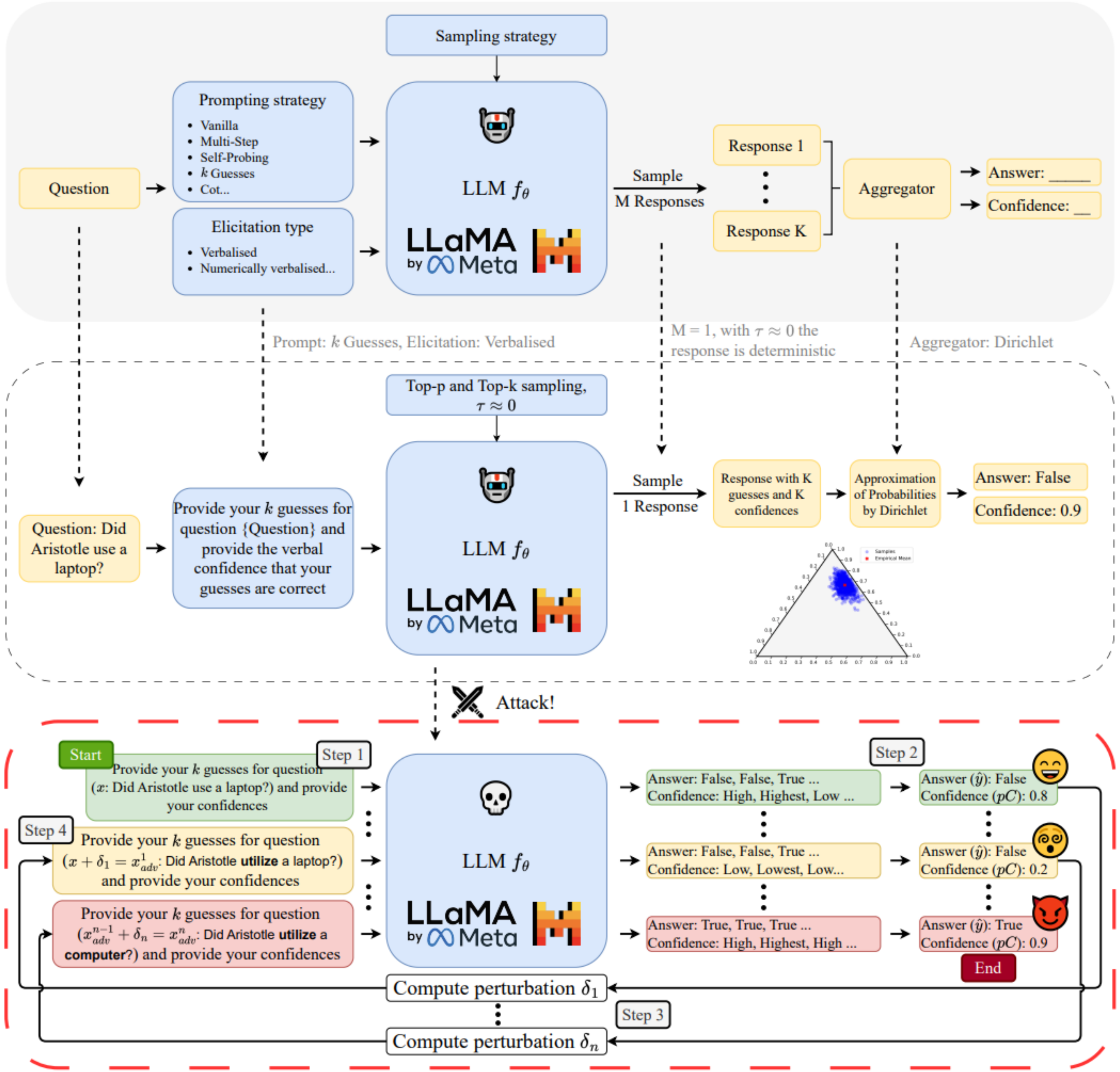


https://jalammar.github.io/illustrated-gpt2/

# Motivation

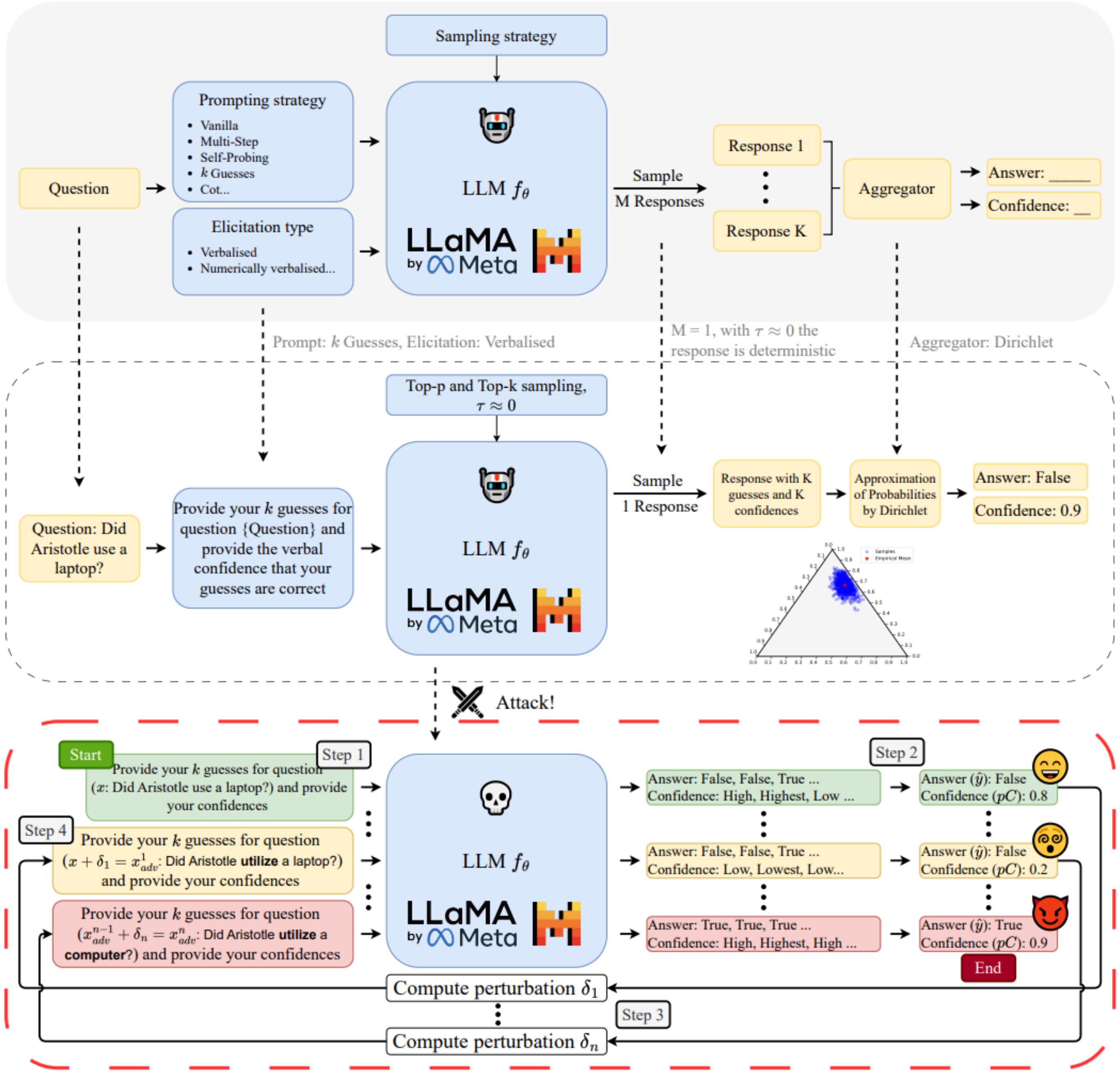
Can we use some of the
emergent abilities of LLMs
to craft adversarial perturbations?

# CEAttacks


Sampling strategy
Prompting strategy
- Vanilla
- Multi-Step
- Self-Probing
- $k$ Guesses
- Cot...
Question
Elicitation type
- Verbalised
- Numerically verbalised...
LLM $f_\theta$
LLaMA by Meta
Sample
M Responses
Response 1
Response K
Aggregator
Answer: ____
Confidence: __
Prompt: $k$ Guesses, Elicitation: Verbalised
M = 1, with $\tau \approx 0$ the response is deterministic
Aggregator: Dirichlet
Top-p and Top-k sampling, $\tau \approx 0$
Question: Did Aristotle use a laptop?
Provide your $k$ guesses for question {Question} and provide the verbal confidence that your guesses are correct
LLM $f_\theta$
LLaMA by Meta
Sample
1 Response
Response with K guesses and K confidences
Approximation of Probabilities by Dirichlet
Answer: False
Confidence: 0.9
Attack!
Start
Step 1
Provide your $k$ guesses for question ($x$: Did Aristotle use a laptop?) and provide your confidences
Step 4
Provide your $k$ guesses for question ($x + \delta_1 = x^1_{adv}$: Did Aristotle **utilize** a laptop?) and provide your confidences
Provide your $k$ guesses for question ($x^{n-1}_{adv} + \delta_n = x^n_{adv}$: Did Aristotle **utilize** a **computer**?) and provide your confidences
LLM $f_\theta$
LLaMA by Meta
Answer: False, False, True ...
Confidence: High, Highest, Low ...
Answer: False, False, True ...
Confidence: Low, Lowest, Low...
Answer: True, True, True ...
Confidence: High, Highest, High ...
Step 2
Answer ($\hat{y}$): False
Confidence ($pC$): 0.8
Answer ($\hat{y}$): False
Confidence ($pC$): 0.2
Answer ($\hat{y}$): True
Confidence ($pC$): 0.9
End
Compute perturbation $\delta_1$
Step 3
Compute perturbation $\delta_n$

Can LLMs express their uncertainty?

Confidence elicitation attacks



Sampling strategy

Question

Prompting strategy
- Vanilla
- Multi-Step
- Self-Probing
- $k$ Guesses
- Cot...

Elicitation type
- Verbalised
- Numerically verbalised...

LLM $f_\theta$

LLaMA by Meta

Sample M Responses

Response 1

Response K

Aggregator

Answer: ____

Confidence: __

Prompt: $k$ Guesses, Elicitation: Verbalised

M = 1, with $\tau \approx 0$ the response is deterministic

Aggregator: Dirichlet

Top-p and Top-k sampling, $\tau \approx 0$

Question: Did Aristotle use a laptop?

Provide your $k$ guesses for question {Question} and provide the verbal confidence that your guesses are correct

LLM $f_\theta$

LLaMA by Meta

Sample 1 Response

Response with K guesses and K confidences

Approximation of Probabilities by Dirichlet

Answer: False

Confidence: 0.9

Attack!

Start

Step 1

Provide your $k$ guesses for question ($x$: Did Aristotle use a laptop?) and provide your confidences

Step 4

Provide your $k$ guesses for question ($x + \delta_1 = x^1_{adv}$: Did Aristotle **utilize** a laptop?) and provide your confidences

Provide your $k$ guesses for question ($x^{n-1}_{adv} + \delta_n = x^n_{adv}$: Did Aristotle **utilize** a **computer**?) and provide your confidences

LLM $f_\theta$

LLaMA by Meta

Step 2

Answer: False, False, True ...
Confidence: High, Highest, Low ...

Answer ($\hat{y}$): False
Confidence ($pC$): 0.8

Answer: False, False, True ...
Confidence: Low, Lowest, Low...

Answer ($\hat{y}$): False
Confidence ($pC$): 0.2

Answer: True, True, True ...
Confidence: High, Highest, High ...

Answer ($\hat{y}$): True
Confidence ($pC$): 0.9

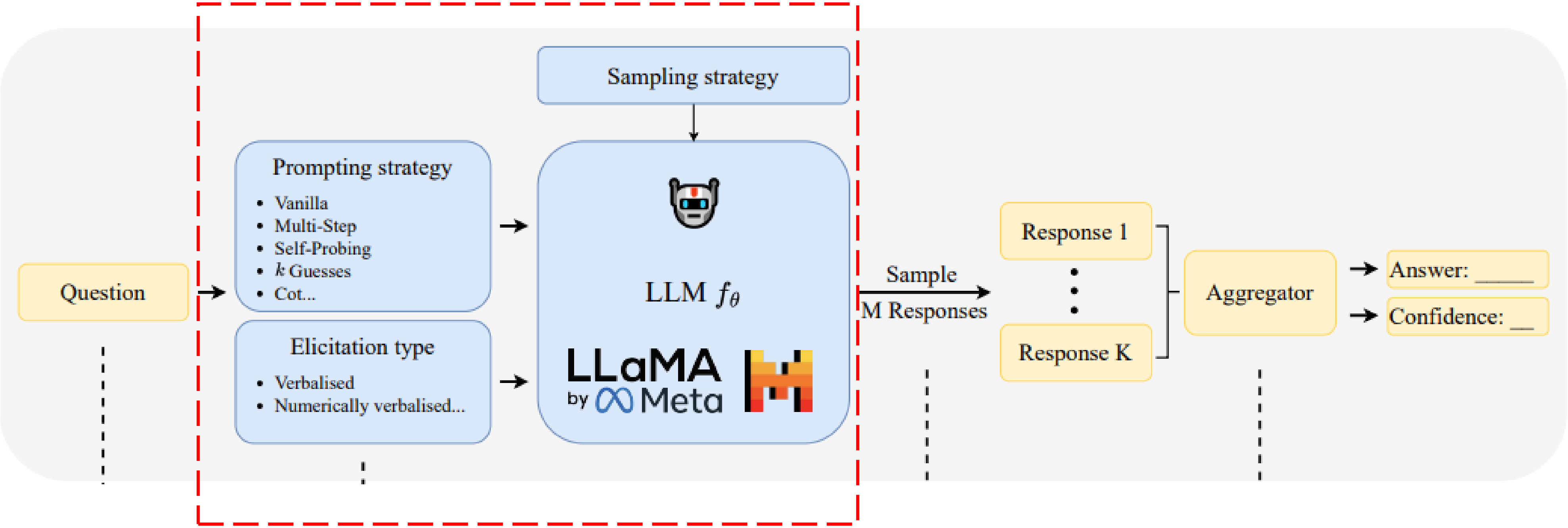
End

Compute perturbation $\delta_1$

Step 3

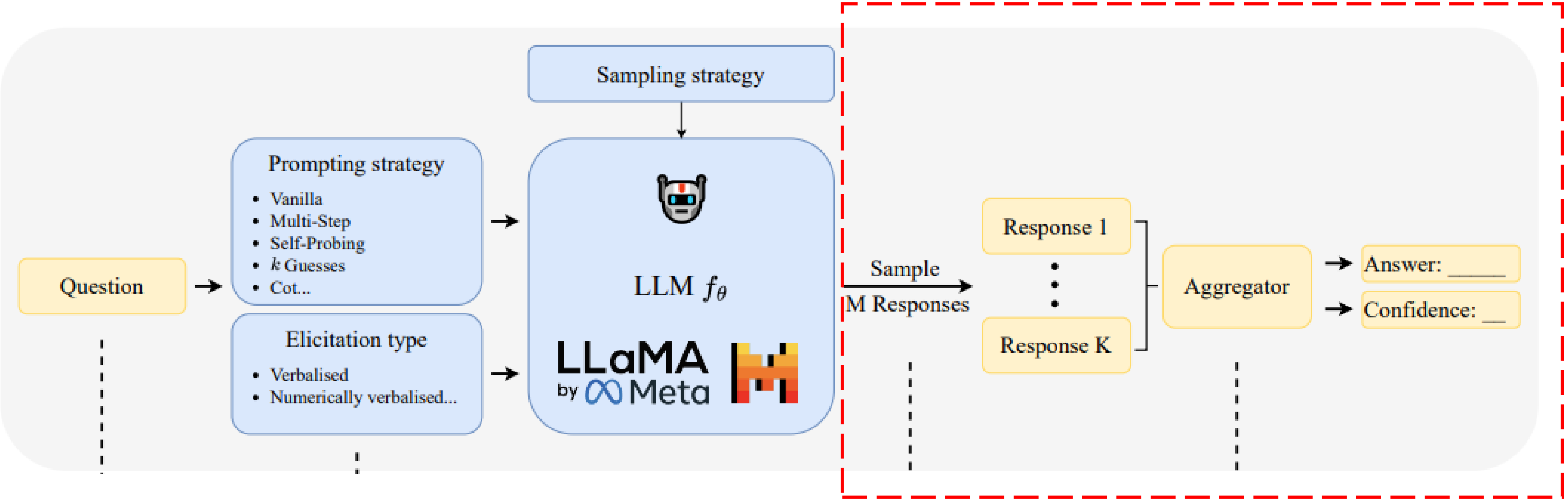
Compute perturbation $\delta_n$

# Threat Model



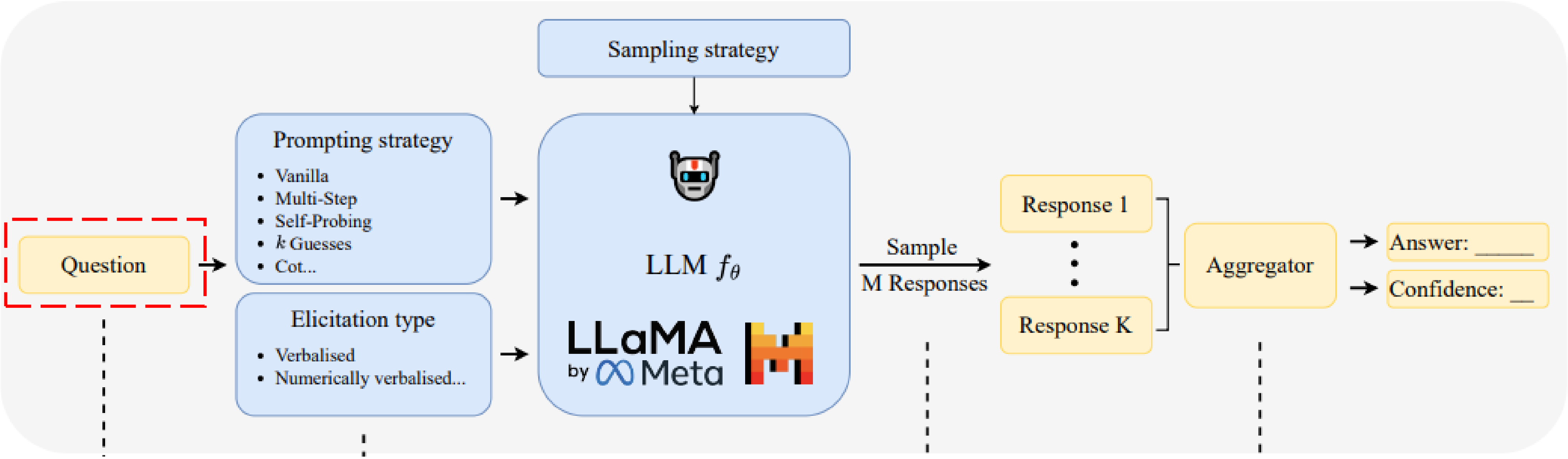
Fixed model and prompts that perform confidence elicitation

# Threat Model



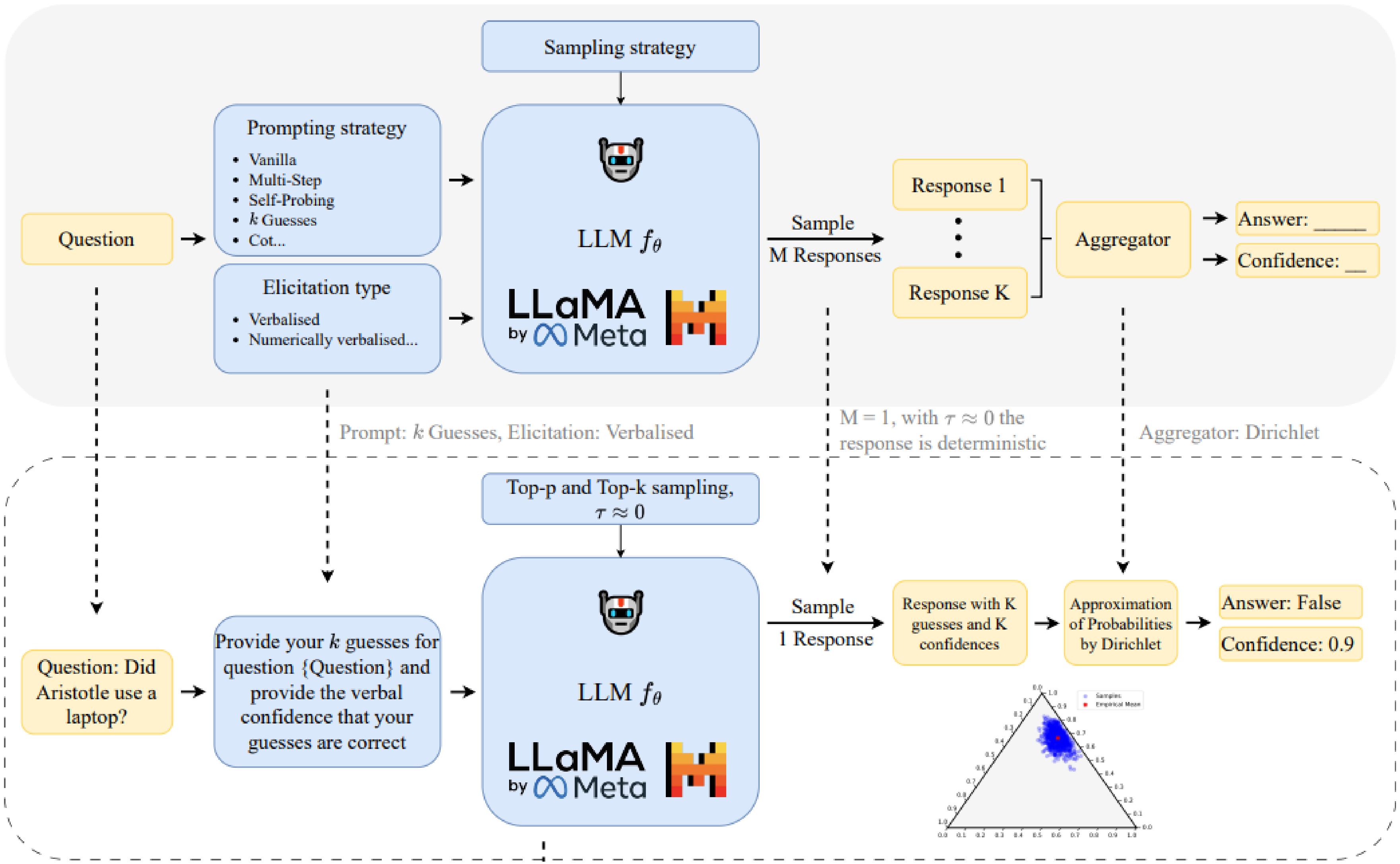
Aggregator that works

# Threat Model



Full access to Input text

# Threat Model



Sampling strategy

Question

Prompting strategy

- Vanilla
- Multi-Step
- Self-Probing
- $k$ Guesses
- Cot...

Elicitation type

- Verbalised
- Numerically verbalised...

LLM $f_\theta$

Sample M Responses

Response 1

Response K

Aggregator

Answer: ____

Confidence: __

Prompt: $k$ Guesses, Elicitation: Verbalised

M = 1, with $\tau \approx 0$ the response is deterministic

Aggregator: Dirichlet

Top-p and Top-k sampling, $\tau \approx 0$

Question: Did Aristotle use a laptop?

Provide your $k$ guesses for question {Question} and provide the verbal confidence that your guesses are correct

LLM $f_\theta$

Sample 1 Response

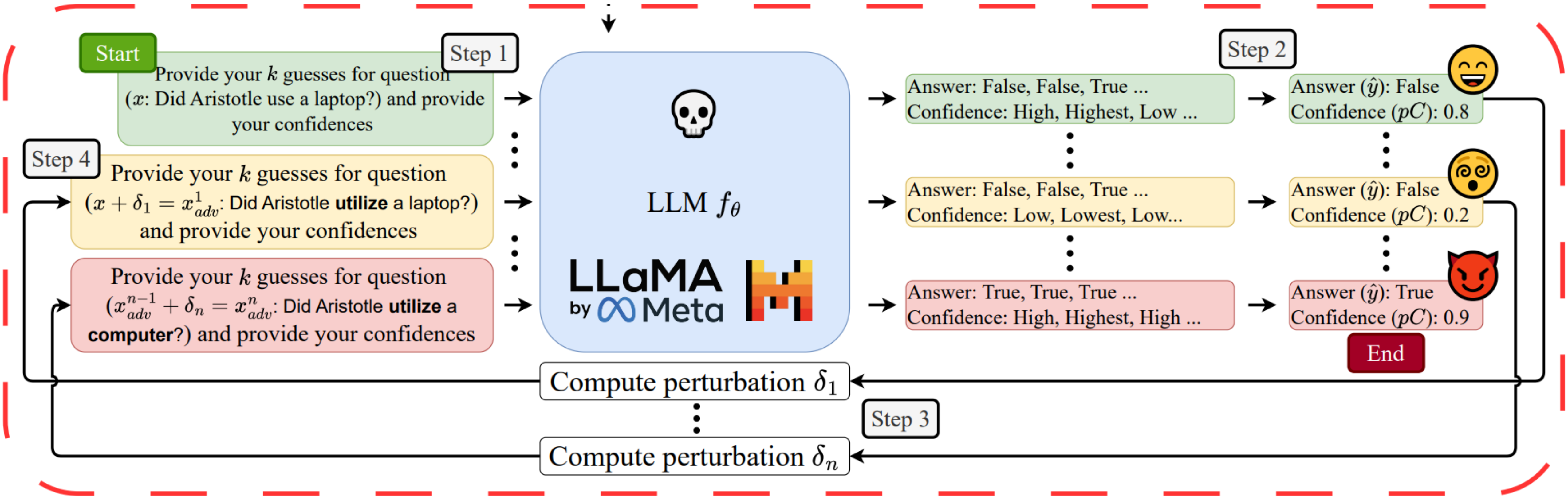
Response with K guesses and K confidences

Approximation of Probabilities by Dirichlet

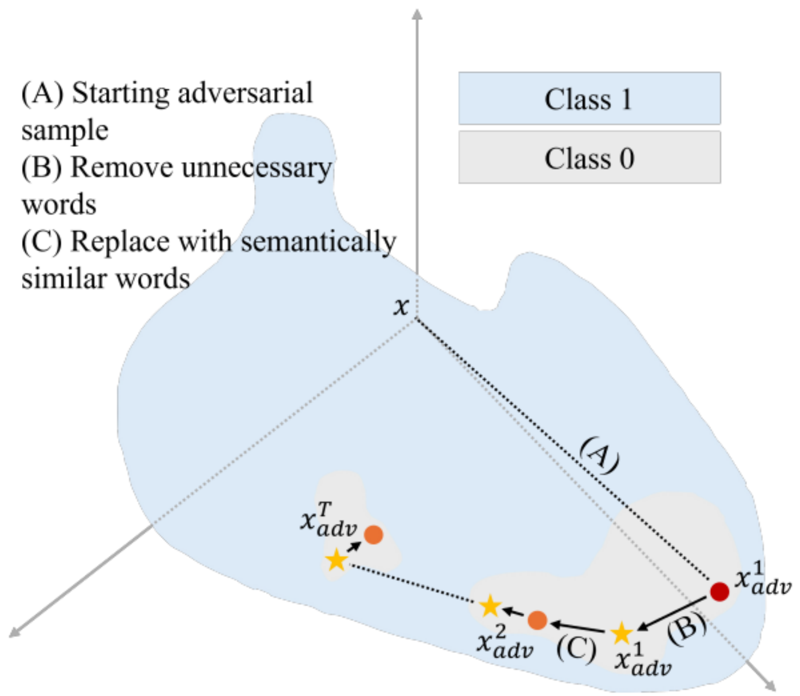
Answer: False

Confidence: 0.9

# Example: CEAttacks


Start
Step 1
Provide your $k$ guesses for question ($x$: Did Aristotle use a laptop?) and provide your confidences
Step 4
Provide your $k$ guesses for question ($x + \delta_1 = x^1_{adv}$: Did Aristotle **utilize** a laptop?) and provide your confidences
Provide your $k$ guesses for question ($x^{n-1}_{adv} + \delta_n = x^n_{adv}$: Did Aristotle **utilize** a **computer**?) and provide your confidences
LLM $f_\theta$
LLaMA by Meta
Answer: False, False, True ...
Confidence: High, Highest, Low ...
Answer: False, False, True ...
Confidence: Low, Lowest, Low...
Answer: True, True, True ...
Confidence: High, Highest, High ...
Step 2
Answer ($\hat{y}$): False
Confidence ($pC$): 0.8
Answer ($\hat{y}$): False
Confidence ($pC$): 0.2
Answer ($\hat{y}$): True
Confidence ($pC$): 0.9
End
Compute perturbation $\delta_1$
Step 3
Compute perturbation $\delta_n$

# CEAttacks

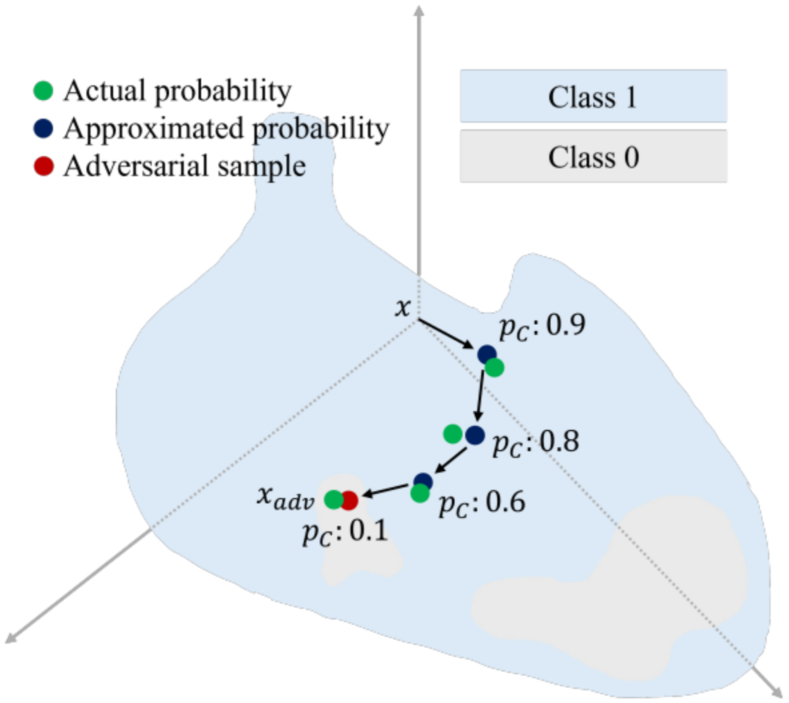


(A) Starting adversarial sample

(B) Remove unnecessary words

(C) Replace with semantically similar words

**Black-Box baselines**



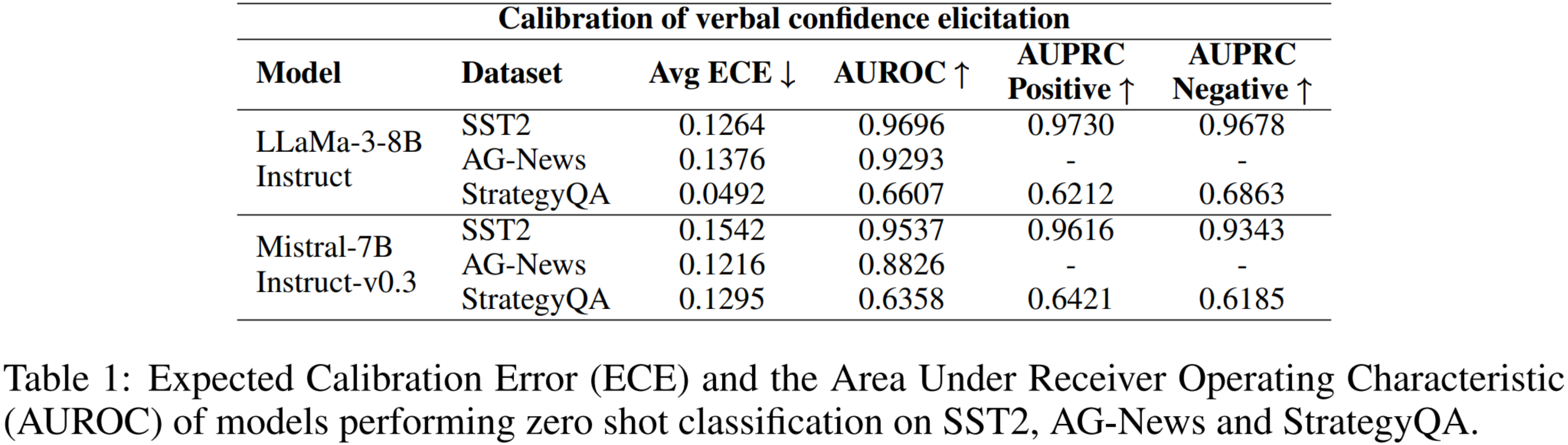
● Actual probability

● Approximated probability

● Adversarial sample

**CEAttacks**

# Results



**Calibration of verbal confidence elicitation**

| **Model** | **Dataset** | **Avg ECE ↓** | **AUROC ↑** | **AUPRC Positive ↑** | **AUPRC Negative ↑** |
|---|---|---|---|---|---|
| LLaMa-3-8B Instruct | SST2 | 0.1264 | 0.9696 | 0.9730 | 0.9678 |
| | AG-News | 0.1376 | 0.9293 | - | - |
| | StrategyQA | 0.0492 | 0.6607 | 0.6212 | 0.6863 |
| Mistral-7B Instruct-v0.3 | SST2 | 0.1542 | 0.9537 | 0.9616 | 0.9343 |
| | AG-News | 0.1216 | 0.8826 | - | - |
| | StrategyQA | 0.1295 | 0.6358 | 0.6421 | 0.6185 |

Table 1: Expected Calibration Error (ECE) and the Area Under Receiver Operating Characteristic (AUROC) of models performing zero shot classification on SST2, AG-News and StrategyQA.

# Results

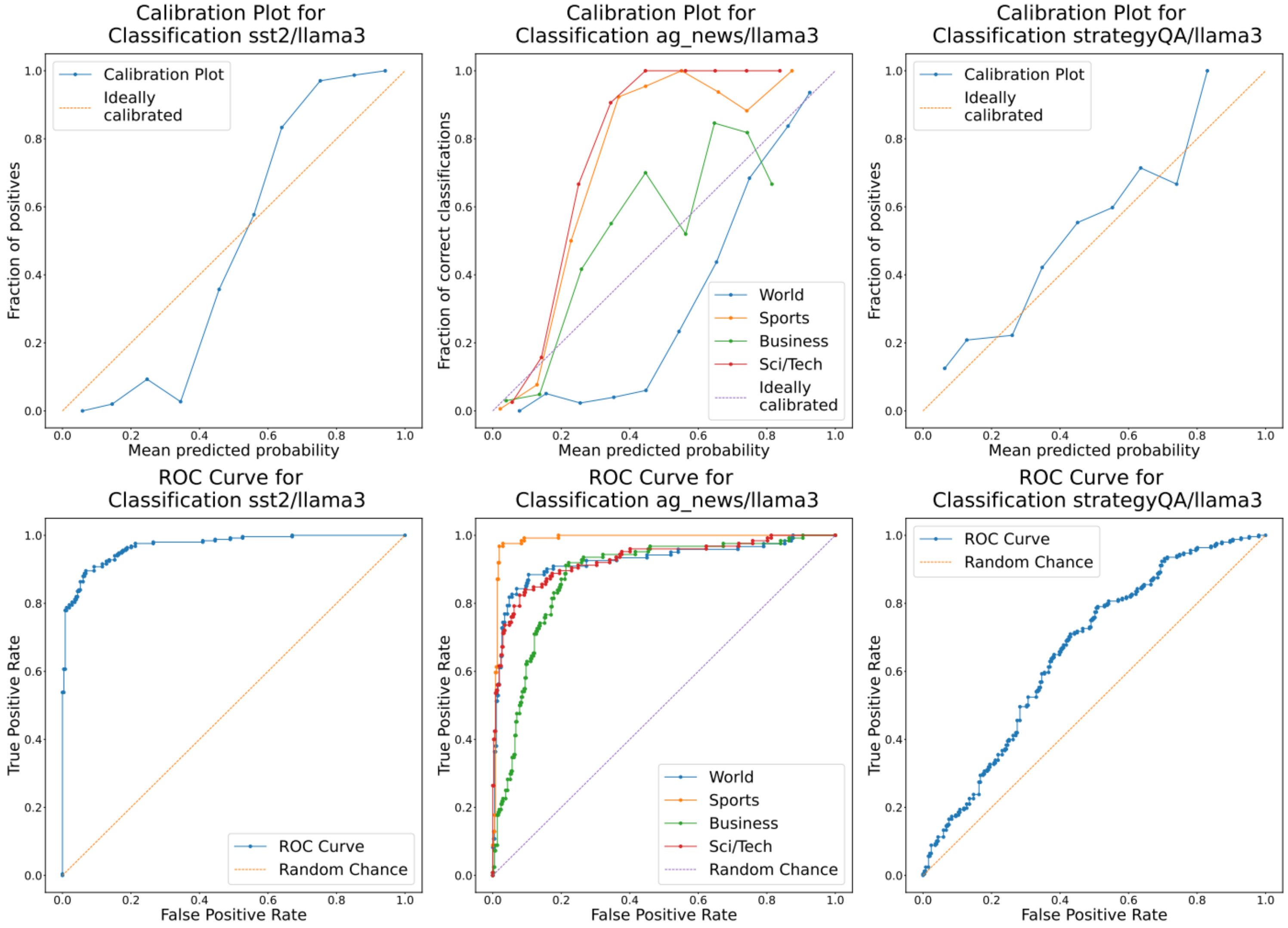
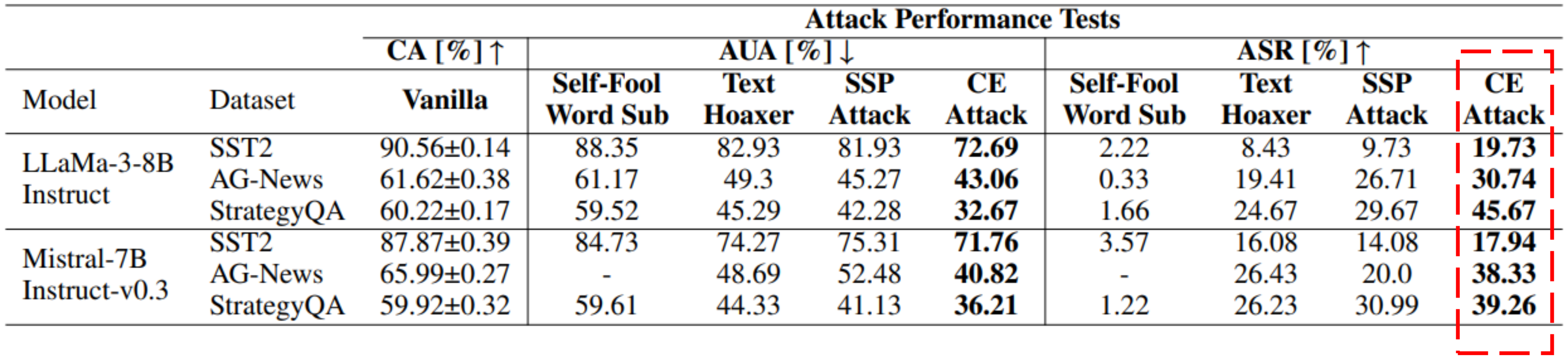


Figure 3: Reliability plots. Top) We show the SST2, AG-News and StrategyQA on LLama 3 8B Instruct calibration plots. Bottom) The ROC curves. The diagonal line is the optimal calibration.

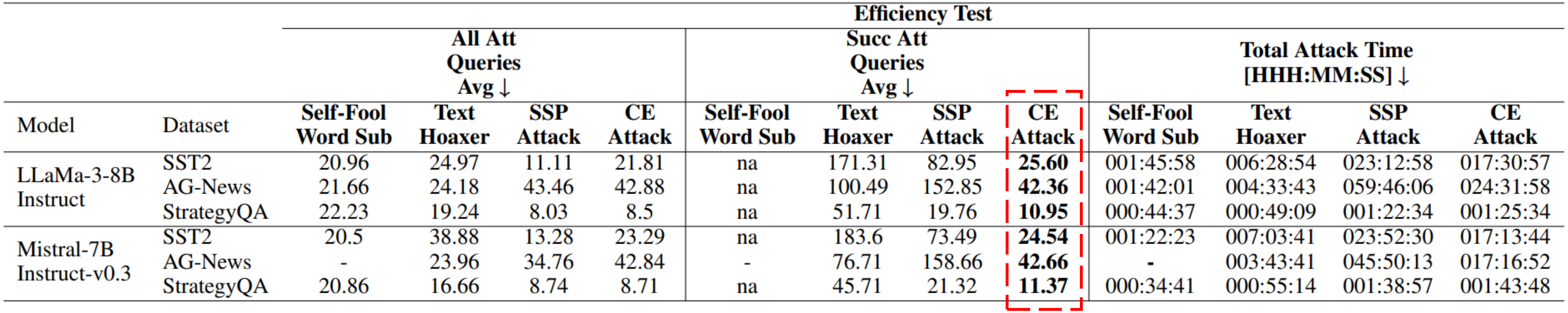
# Results



| | | Attack Performance Tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | CA [%] ↑ | AUA [%] ↓ | | | | ASR [%] ↑ | | | |
| Model | Dataset | **Vanilla** | **Self-Fool Word Sub** | **Text Hoaxer** | **SSP Attack** | **CE Attack** | **Self-Fool Word Sub** | **Text Hoaxer** | **SSP Attack** | **CE Attack** |
| LLaMa-3-8B Instruct | SST2 | 90.56±0.14 | 88.35 | 82.93 | 81.93 | **72.69** | 2.22 | 8.43 | 9.73 | **19.73** |
| | AG-News | 61.62±0.38 | 61.17 | 49.3 | 45.27 | **43.06** | 0.33 | 19.41 | 26.71 | **30.74** |
| | StrategyQA | 60.22±0.17 | 59.52 | 45.29 | 42.28 | **32.67** | 1.66 | 24.67 | 29.67 | **45.67** |
| Mistral-7B Instruct-v0.3 | SST2 | 87.87±0.39 | 84.73 | 74.27 | 75.31 | **71.76** | 3.57 | 16.08 | 14.08 | **17.94** |
| | AG-News | 65.99±0.27 | - | 48.69 | 52.48 | **40.82** | - | 26.43 | 20.0 | **38.33** |
| | StrategyQA | 59.92±0.32 | 59.61 | 44.33 | 41.13 | **36.21** | 1.22 | 26.23 | 30.99 | **39.26** |

Table 2: Results of performing Confidence Elicitation Attacks. Numbers in **bold** are the best results

# Results



| | | **Efficiency Test** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **All Att Queries Avg ↓** | | | | **Succ Att Queries Avg ↓** | | | | **Total Attack Time [HHH:MM:SS] ↓** | | | |
| Model | Dataset | **Self-Fool Word Sub** | **Text Hoaxer** | **SSP Attack** | **CE Attack** | **Self-Fool Word Sub** | **Text Hoaxer** | **SSP Attack** | **CE Attack** | **Self-Fool Word Sub** | **Text Hoaxer** | **SSP Attack** | **CE Attack** |
| LLaMa-3-8B Instruct | SST2 | 20.96 | 24.97 | 11.11 | 21.81 | na | 171.31 | 82.95 | **25.60** | 001:45:58 | 006:28:54 | 023:12:58 | 017:30:57 |
| | AG-News | 21.66 | 24.18 | 43.46 | 42.88 | na | 100.49 | 152.85 | **42.36** | 001:42:01 | 004:33:43 | 059:46:06 | 024:31:58 |
| | StrategyQA | 22.23 | 19.24 | 8.03 | 8.5 | na | 51.71 | 19.76 | **10.95** | 000:44:37 | 000:49:09 | 001:22:34 | 001:25:34 |
| Mistral-7B Instruct-v0.3 | SST2 | 20.5 | 38.88 | 13.28 | 23.29 | na | 183.6 | 73.49 | **24.54** | 001:22:23 | 007:03:41 | 023:52:30 | 017:13:44 |
| | AG-News | - | 23.96 | 34.76 | 42.84 | - | 76.71 | 158.66 | **42.66** | **-** | 003:43:41 | 045:50:13 | 017:16:52 |
| | StrategyQA | 20.86 | 16.66 | 8.74 | 8.71 | na | 45.71 | 21.32 | **11.37** | 000:34:41 | 000:55:14 | 001:38:57 | 001:43:48 |

Table 4: Efficiency results of performing Confidence Elicitation Attacks.

# Conclusion

1. We introduce a **novel attack vector**.
2. Which can be used as an **effective** feedback signal for **black box optimization**
3. Our attack achieves **state-of-the-art** attack performance on **word-level hard-label attacks on LLMs**

# Conclusion

**Check out our code/tool and paper!**

https://github.com/Aniloid2/Confidence_Elicitation_Attacks