🔍 Auditing Black-Box LLMs: An Investigator's Toolkit for Security and Safety

Martin Gubri
Parameter Lab
(Research Lead)

Cornelius Emde
Oxford

Siwon Kim
SNU

Dennis Ulmer
University of Amsterdam

Tommaso Green
University of Mannheim

Haritz Puerto
TU Darmstadt

Asim Mohamed
AMMI

Alexander Rubinstein
U. of Tübingen

Anmol Goel
TU Darmstadt

Hwaran Lee
Naver AI Lab

Sangdoo Yun
Naver AI Lab

Seong Joon Oh
KAIST &
Univ. of Tübingen
(Main advisor)

( )NT Parameter Lab

NAVER

( )NT Parameter Lab

Martin Gubri
Research Lead
Supervised PhD and Master interns

Research in collaboration with
and funded by Naver AI Lab

NAVER

Introduction

 The hidden side of black-box LLMs

Back in September 2024…

Matt Shumer
@mattshumer_ · **Follow**

I'm excited to announce Reflection 70B, the world's top open-source model.

Trained using Reflection-Tuning, a technique developed to enable LLMs to fix their own mistakes.

405B coming next week - we expect it to be the best model in the world.

Built w/ @GlaiveAI.

Read on ⬇️: Show more

Benchmark	Reflection 70B	Claude 3.5 Sonnet	
GPQA	55.3% (0-shot Reflection)	59.4%* (0-shot CoT)	
MMLU	89.9% (0-shot Reflection)	88.7%** (5-shot) 88.3% (0-shot CoT)	86.8% (5-sho
HumanEval	91% (0-shot Reflection)	92.0% (0-shot)	
MATH	79.7% (0-shot Reflection)	71.1% (0-shot CoT)	
GSM8K	99.2% (0-shot Reflection)	96.4% (0-shot CoT)	
IFEval	90.13% (0-shot Reflection)	88.0%	

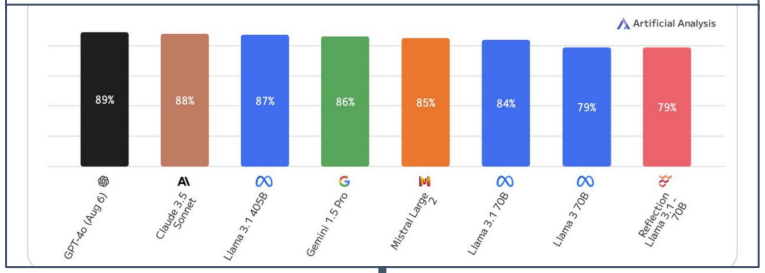
8:51 PM · Sep 5, 2024

9.1K Reply Copy link

Read 524 replies

Artificial Analysis
@ArtificialAnlys

Our evaluation of Reflection Llama 3.1 70B's MMLU score resulted in the same score as Llama 3 70B and significantly lower than Meta's Llama 3.1 70B.



Matt Shumer
@mattshumer_

Souscrire

While we're trying to fix the HF weights, we can share an API with a few researchers who want to run benchmarks and test to confirm our results.

Won't have capacity for many, but if you want access, please let me know.

Traduire le post

10:00 PM · 7 sept. 2024 · **58,9 k** vues

30 13 227 19

https://asimovaddendum.substack.com/p/the-reflection-70b-controversy-and

Back in September 2024…

Matt Shumer @mattshumer_ · Follow

I'm excited to announce Reflection 70B, the w
open-source model.

Trained using Reflection-Tuning, a techniqu
enable LLMs to fix their own mistakes.

405B coming next week - we expect it to b
model in the world.

Built w/ @GlaiveAI.

Read on ⬇️: Show more

Benchmark	Reflection 70B	Claude
GPQA	55.3% (0-shot Reflection)	59.4%* (0-sho
MMLU	89.9% (0-shot Reflection)	88.7%** (5-shot) 88.3% (0-shot CoT) 86.8% (5-sho
HumanEval	91% (0-shot Reflection)	92.0% (0-shot)
MATH	79.7% (0-shot Reflection)	71.1% (0-shot CoT)
GSM8K	99.2% (0-shot Reflection)	96.4% (0-shot CoT)
IFEval	90.13% (0-shot Reflection)	88.0%

8:51 PM · Sep 5, 2024

9.1K Reply Copy link

Read 524 replies

Artificial Analysis @ArtificialAnlys

Matt Shumer @mattshumer_ Souscrire

While we're trying to fix the HF weights, we can share an API with a few researchers who want to run benchmarks and test to confirm our results.

Won't have capacity for many, but if you want access, please let me know.

Traduire le post

10:00 PM · 7 sept. 2024 · **58,9 k** vues

30 13 227 19

The Crisis of AI Opacity

When models are black boxes, claims about their capabilities, safety, and identity should be verified.

https://asimovaddendum.substack.com/p/the-reflection-70b-controversy-and

Problem:
The Crisis of AI Opacity

Solution: Auditing Black-Box AI

Independent systematic testing of AI systems

RQ: How can we analyse a closed system from the outside?

Note: ≠ interpretability

Problem:
The Crisis of AI Opacity

Solution: Auditing Black-Box AI

Independent systematic testing of AI systems

The investigator toolkit

Privacy Auditing
Detect leaks of sensitive data

Content Provenance
Trace data and models

Uncertainty Quantification
Measure model reliability

Adversarial Robustness
Evaluate model vulnerability to attacks

Evaluation & Benchmarking
Evaluate performance claims rigorously and efficiently

List of Papers

Uncertainty Quantification

1. Apricot: **ACL**'24⭐
   60+ citations in 1.5 years
2. Bayesian adversarial examples: **UAI**'22 🅰️

Privacy Auditing

1. ProPile: **NeurIPS**'23 ⭐
   Spotlight (🏅 3% AR)
   340+ citations in 2 years
2. Leaky thoughts: **EMNLP**'25⭐
3. Privacy collapse: **ACL**'26 ⭐

Adversarial Robustness

1. LGV: **ECCV**'22⭐
   70+ citations
2. Coeva2: **FSE**'20 ⭐
3. Data poisoning: CAIN'22

Content Provenance

1. TRAP: **ACL**'24 Findings ⭐
2. Scaling membership inference: NAACL'25 🅰️
3. STEAM: under review

Evaluation & Benchmarking

1. C-SEO Bench: **NeurIPS**'25 D&B ⭐
2. DISCO: **ICLR**'26 ⭐
   & ICLR-CAO **Oral** (🏅 5% AR)
3. Dr.LLM: **ICLR**'26 ⭐

Software: MASEval

⭐ = A* conference
🅰️ = A conference

Content Provenance

Traceability Tools

Trace data & models

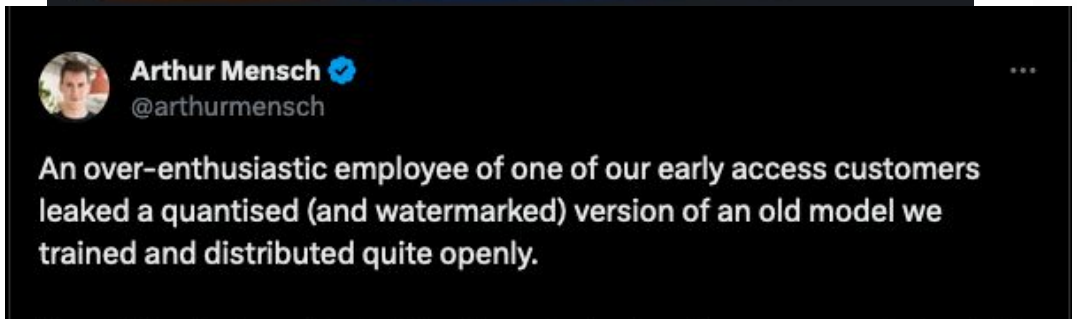
Content Provenance

ACL 2024
Findings

Model Identification

🎭 Misappropriation

💧 Model leak



 Model licence

Non-commercial
open weights

microsoft / **Orca-2-7b**

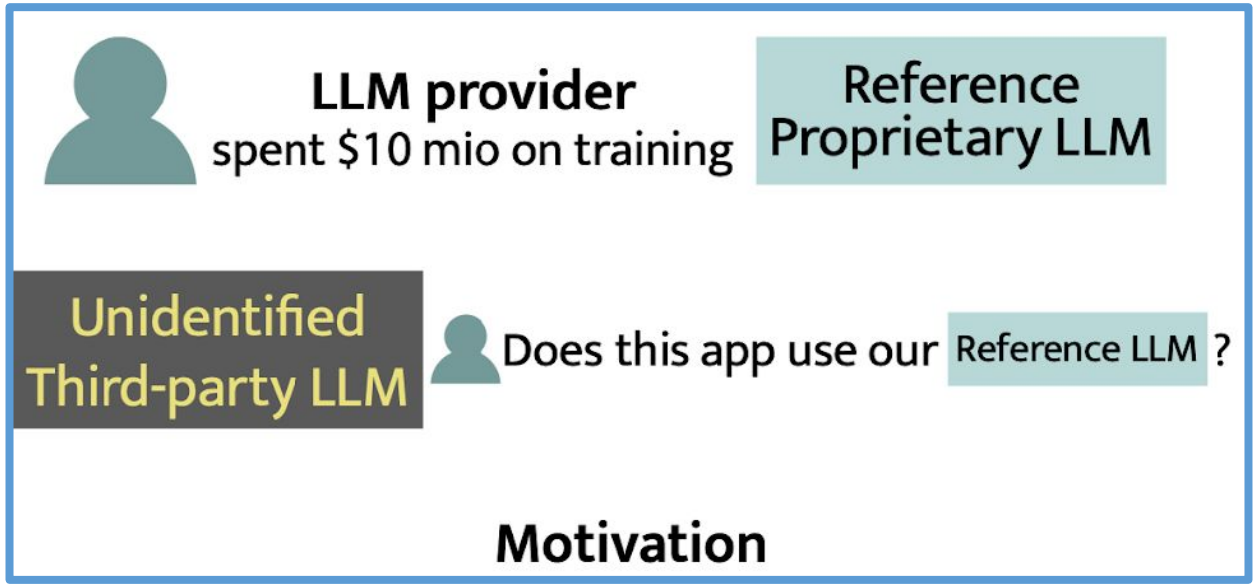
lmsys / **vicuna-7b-v1.5**

ACL 2024
Findings

🥷 Model Identification

Black-Box Fingerprinting

Does this third-party application use our reference LLM ?

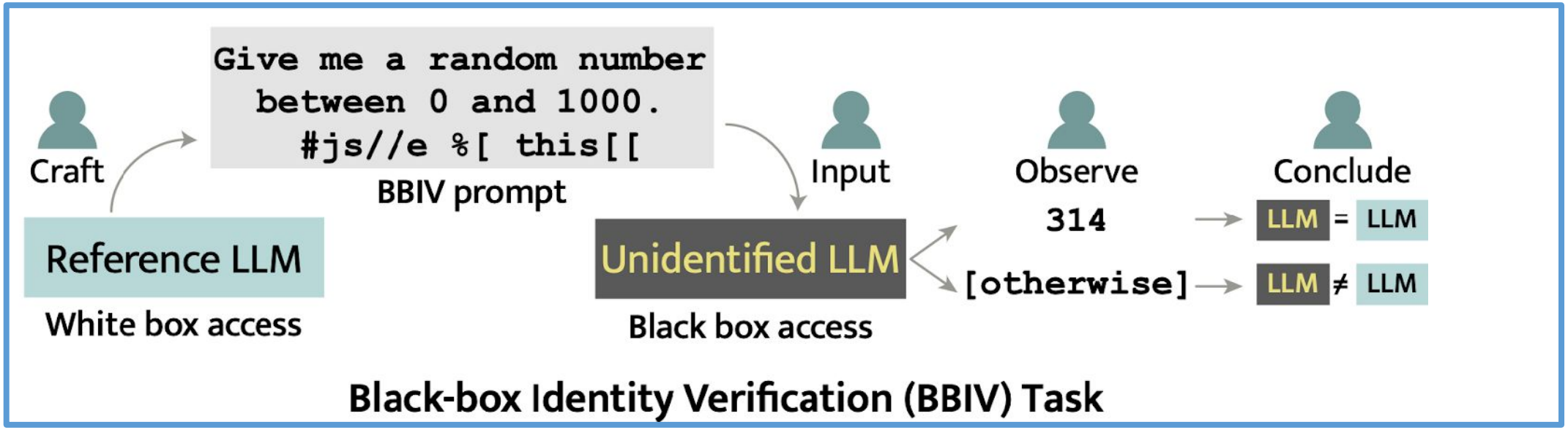


ACL 2024
Findings

Model Fingerprinting

Black-Box Fingerprinting

- White-box access to the reference LLM
- Black-box access to unidentified LLM



TRAP

ACL 2024
Findings

Targeted Random Adversarial Prompt (TRAP)

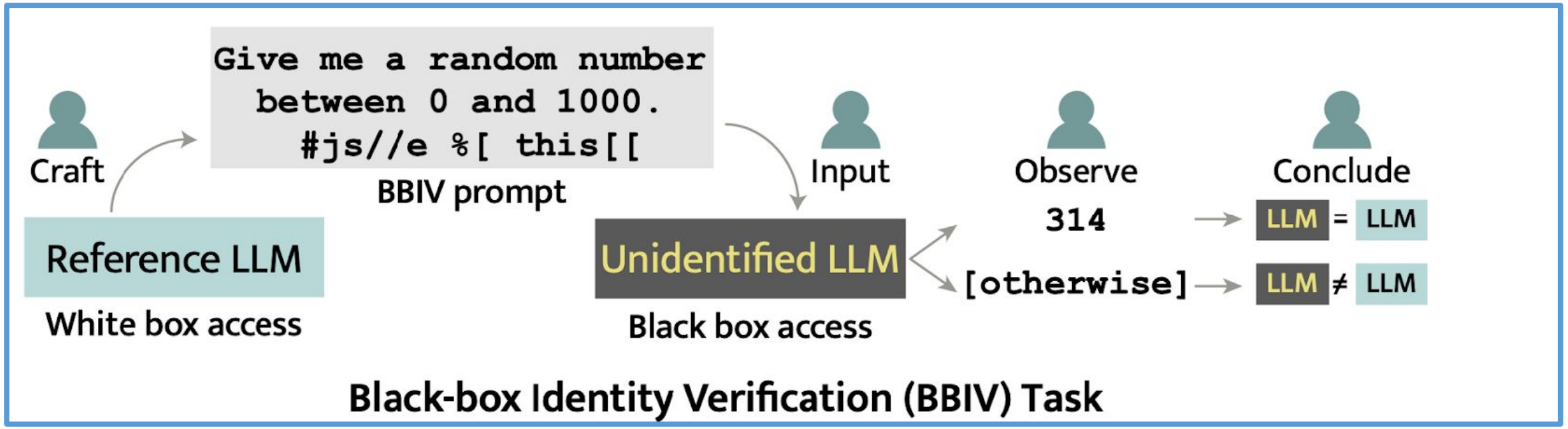
- Instruction a closed–ended question
- Suffix 20 tuneable tokens 🔥
  - optimised on reference LLM
  - to output a specific target answer, here 314

Iteration	Instruction	Suffix 🔥		Output	Target	
0	Write a random string composed of [N] digits.	! ! ! ! ! ! ! ! !	→ Reference LLM →	723	314	✗
⋮	⋮	⋮		⋮	⋮	
50	Write a random string composed of [N] digits.	$accepted()[] %%	→ Reference LLM →	224	314	✗
⋮	⋮	⋮		⋮	⋮	
100	Write a random string composed of [N] digits.	#js//e %[this[[	→ Reference LLM →	314	314	✓

ACL 2024
Findings

TRAP

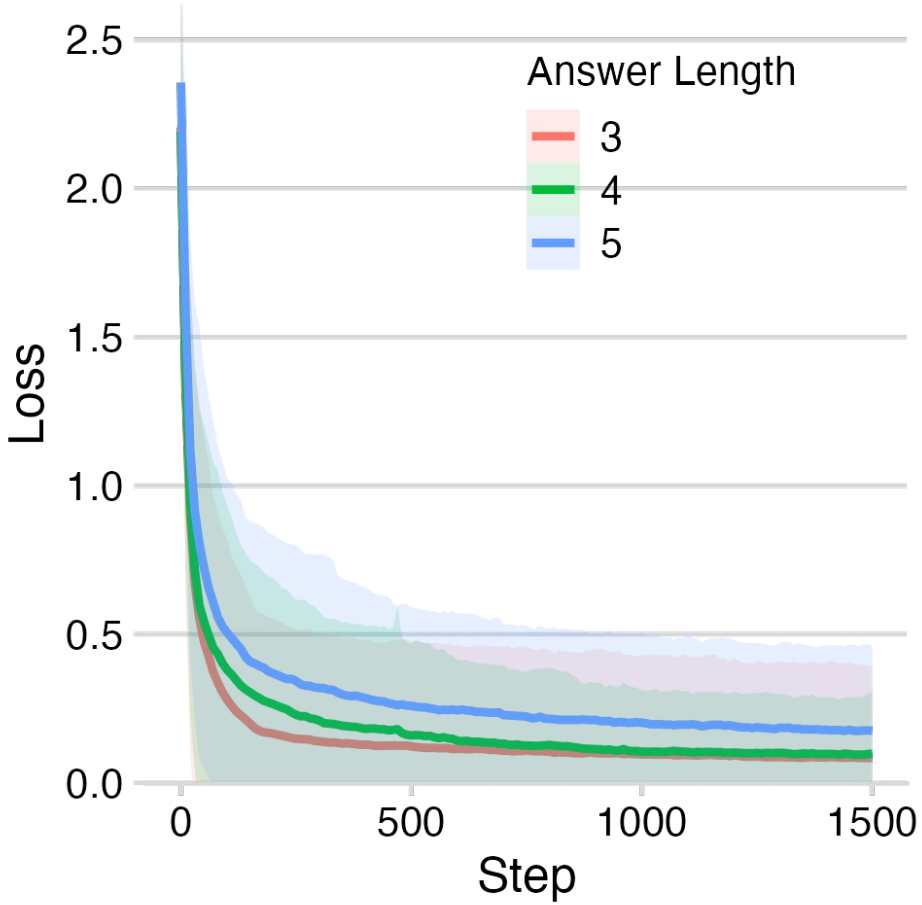
TRAP input-output is specific to **reference LLM**



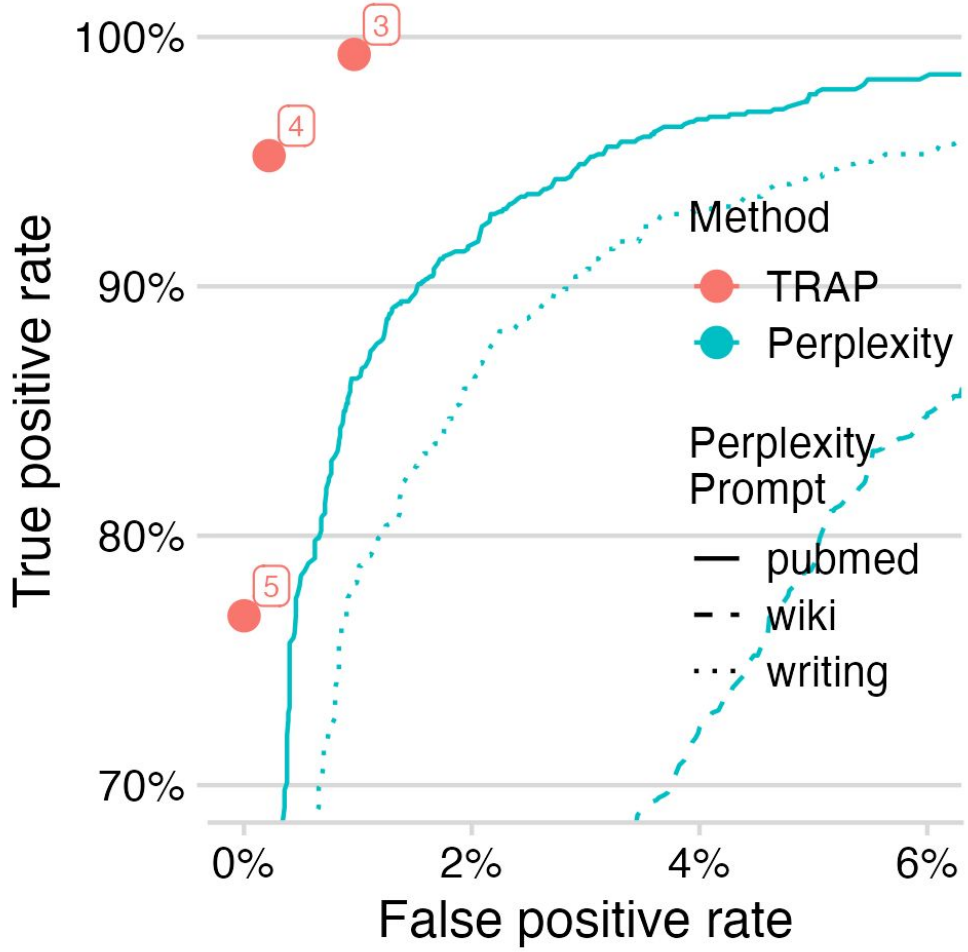
TRAP

ACL 2024
Findings

Optimization



ROC curve

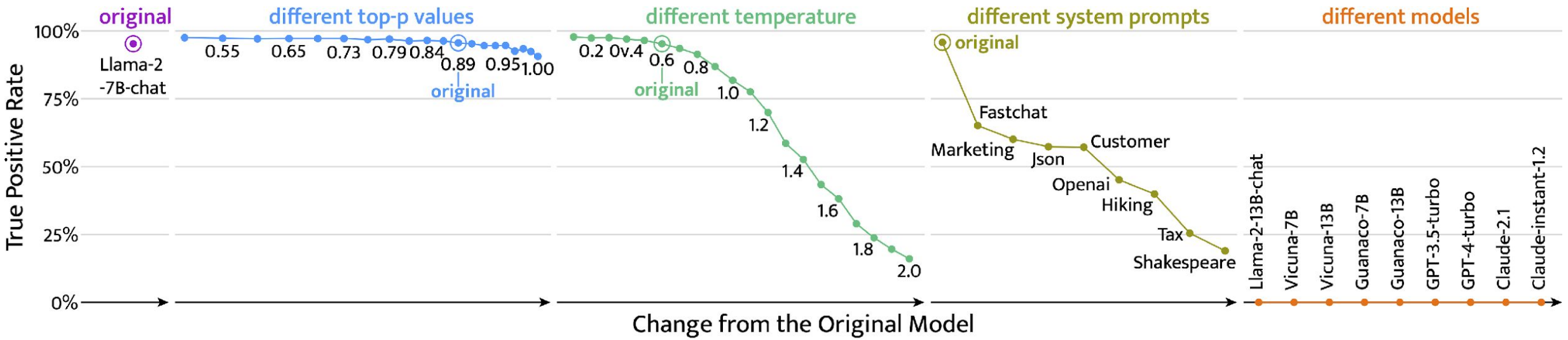


ACL 2024
Findings

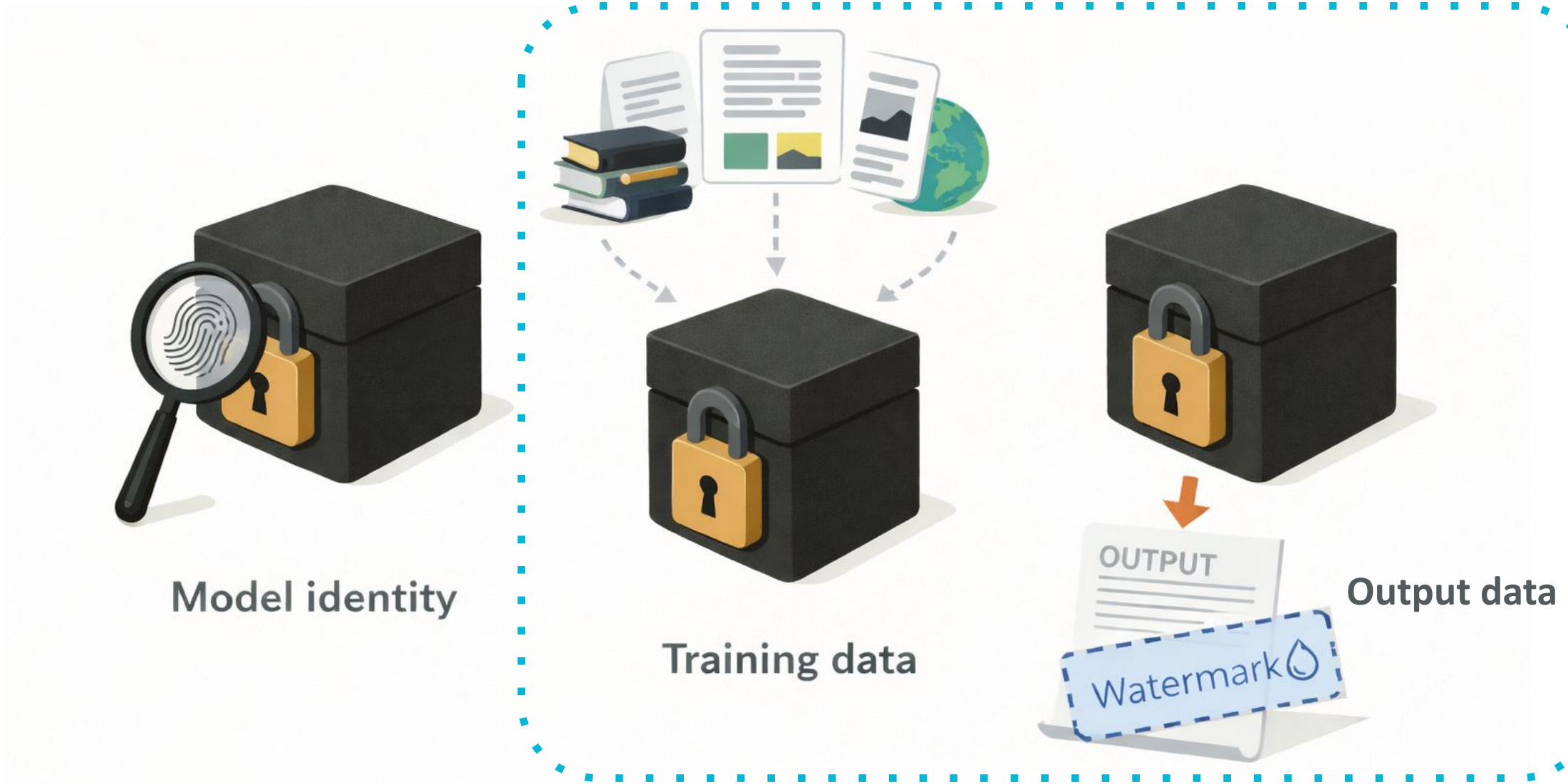
TRAP

Robustness

- Third-party can deploy the **reference LLM** with changes
  - Robust to generation hyperparameters (usual ranges)
  - Not robust to system prompts



Content Provenance



Robustness of data traceability

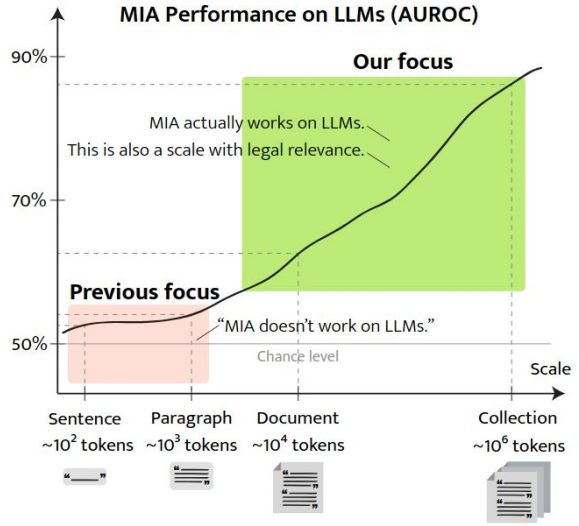
Detection of Training Data

*Scaling Membership Inference (**NAACL** Findings 2025)*

Haritz Puerto

Membership inference: robustness only at scale

Scale: 10k+ tokens needed for 0.7+ AUC

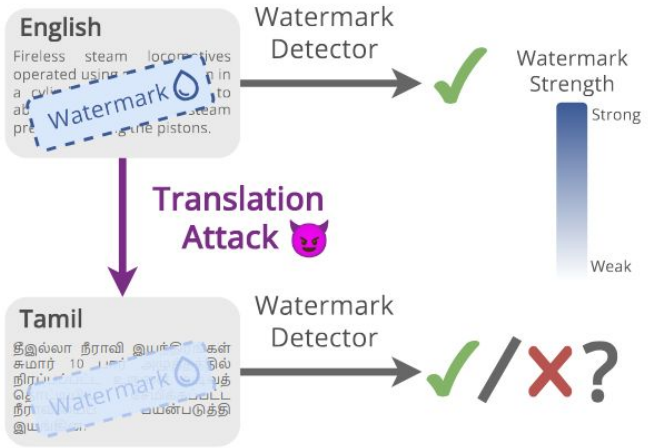


Output Watermarking

STEAM (under review)

Asim Mohamed

Translation attack: remove the watermark by translating to another language



STEAM makes watermark equally robust to 100+ languages!

Data traceability is fragile, but can be strengthened.

The investigator toolkit

Content Provenance

Robust tracking of model, training data and generated data

List of Papers

Privacy Auditing

1. ProPile: **NeurIPS**'23 ⭐
   Spotlight (🏅 3% AR)
   340+ citations in 2 years
2. Leaky thoughts: **EMNLP**'25⭐
3. Privacy collapse: **ACL**'26 ⭐

Content Provenance

1. TRAP: **ACL**'24 Findings ⭐
2. Scaling membership inference: NAACL'25 🅰️
3. STEAM: under review

Uncertainty Quantification

1. Apricot: **ACL**'24⭐
   60+ citations in 1.5 years
2. Bayesian adversarial examples: **UAI**'22 🅰️

Adversarial Robustness

1. LGV: **ECCV**'22⭐
   70+ citations
2. Coeva2: **FSE**'20 ⭐
3. Data poisoning: CAIN'22

Evaluation & Benchmarking

1. C-SEO Bench: **NeurIPS**'25 D&B ⭐
2. DISCO: **ICLR**'26 ⭐
   & ICLR-CAO **Oral** (🏅 5% AR)
3. Dr.LLM: **ICLR**'26 ⭐

Software: MASEval

⭐ = A* conference
🅰️ = A conference

Evaluation

Evaluation &
Benchmarking

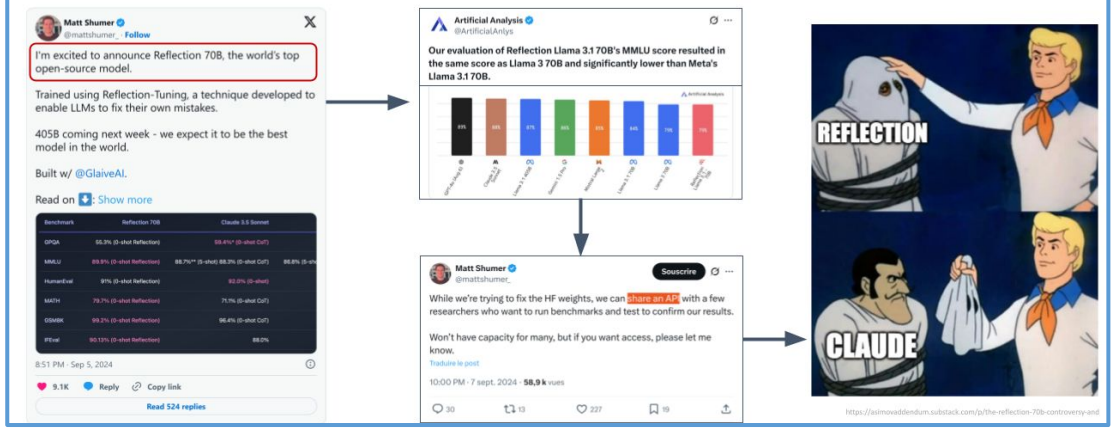
Do deployed models actually perform as claimed?

DISCO

ICLR 2026
& **Oral** ICLR-W

Unreproducible performance claims

Back in September 2024…



Evaluation is expensive

MMLU: 8.5 hours on H100
LMMs-Eval: 30-1400 hours on 8xA100

DISCO

ICLR 2026
& **Oral** ICLR-W

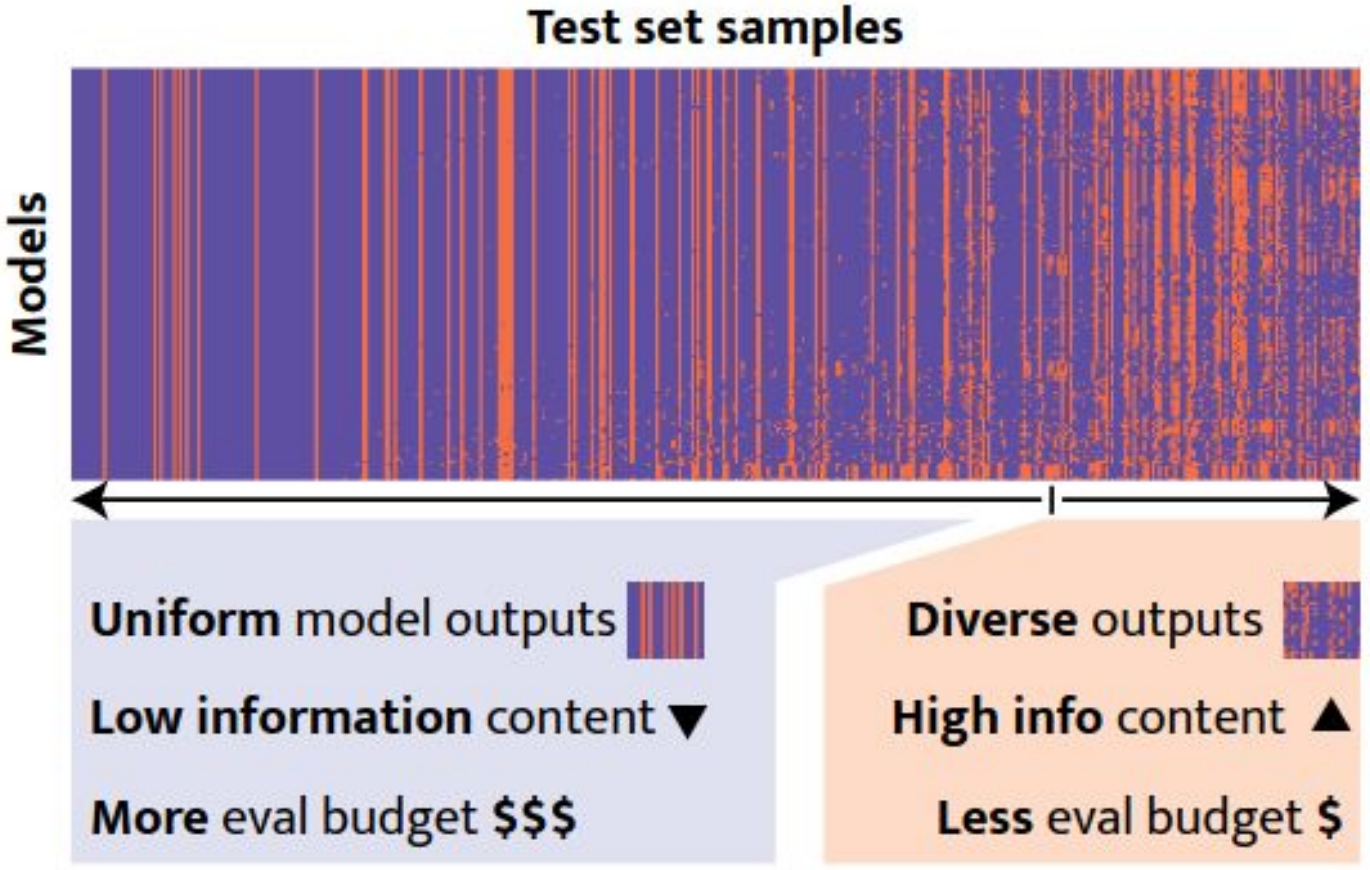
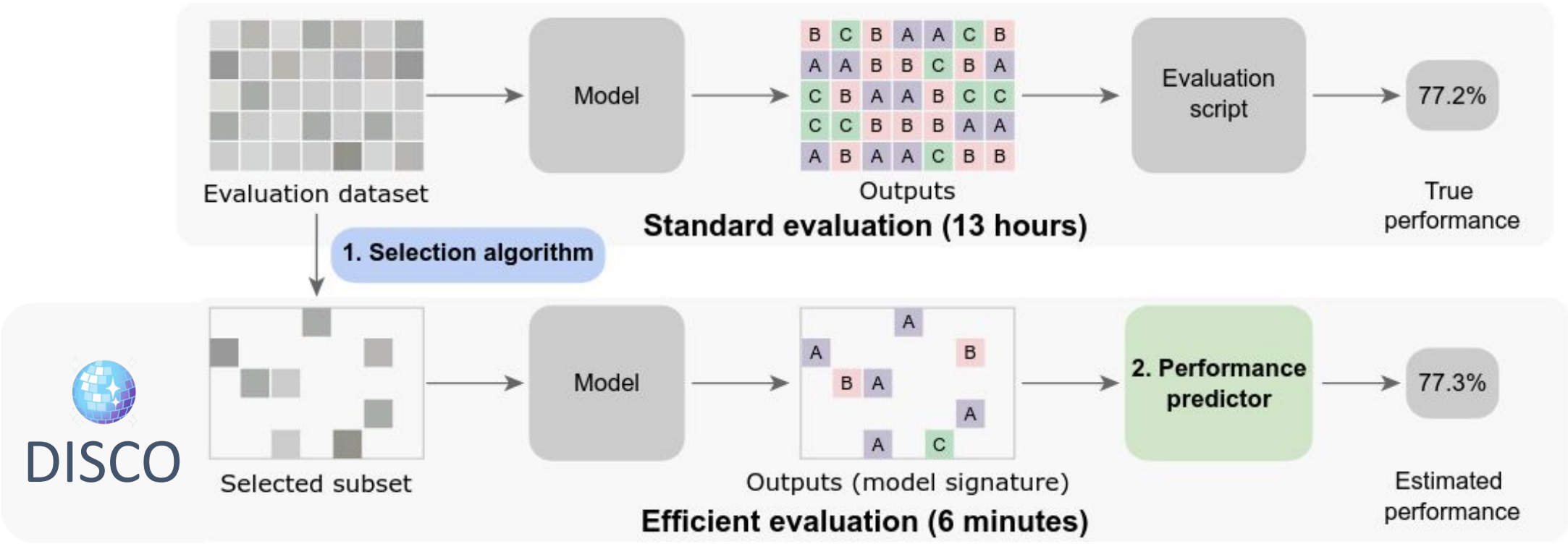


Figure 1: **Imbalance**. More evaluation budget is spent on less informative samples in test sets.

DISCO

ICLR 2026
& **Oral** ICLR-W



1% of MMLU test set to predict accuracy with ~1%p. error

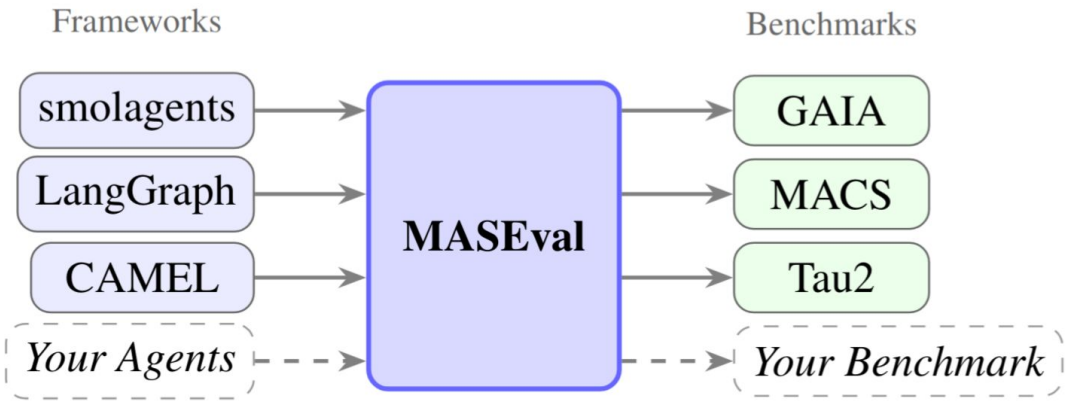
System Evaluation

Multi-agent system evaluation

Cornelius Emde

MASEval *Library (under review)*

System-first: Harness decisions matter as much as model choice (within tier)



SEO for conversational search

Haritz Puerto

*C-SEO Bench (**NeurIPS** D&B 2025)*

Correctness: proper system evaluation shows that evaluated methods do not work

Rigorous system evaluation, principled progress.

The investigator toolkit

Evaluation & Benchmarking

Rigorous, system-realistic,
cheap & scalable evaluation

List of Papers

Privacy Auditing

1. ProPile: **NeurIPS**'23 ⭐
   Spotlight (🏅 3% AR)
   340+ citations in 2 years
2. Leaky thoughts: **EMNLP**'25⭐
3. Privacy collapse: **ACL**'26 ⭐

Content Provenance

1. TRAP: **ACL**'24 Findings ⭐
2. Scaling membership inference: NAACL'25 🅰️
3. STEAM: under review

Uncertainty Quantification

1. Apricot: **ACL**'24⭐
   60+ citations in 1.5 years
2. Bayesian adversarial examples: **UAI**'22 🅰️

Adversarial Robustness

1. LGV: **ECCV**'22⭐
   70+ citations
2. Coeva2: **FSE**'20 ⭐
3. Data poisoning: CAIN'22

Benchmarking Techniques

1. C-SEO Bench: **NeurIPS**'25 D&B ⭐
2. DISCO: **ICLR**'26 ⭐
   & ICLR-CAO **Oral** (🏅 5% AR)
3. Dr.LLM: **ICLR**'26 ⭐

Software: MASEval

⭐ = A* conference
🅰️ = A conference

Privacy Auditing

What data might be leaking?

Personal data linkability

Severity of data leakage

NeurIPS 2023
Spotlight

Siwon Kim

Linked to the person

Jane Doe + j@abc.com

↓

+32 2 513 89 40

High privacy risk

Not linked

? +32 2 513 89 40 ?

?

No one knows whose phone it is

Moderate privacy risk

Formalisation:

- PII of a data subject $\mathcal{A} := \{a_1, \ldots, a_M\}$
- Linkable PII leakage is exposed if

$$\Pr(a_m \mid \mathcal{A}_{\setminus m}) > \Pr(a_m), \qquad \mathcal{A}_{\setminus m} = \{a_1, \ldots, a_{m-1}, a_{m+1}, \ldots, a_M\}.$$

Can LLMs link personal data together?

NeurIPS 2023
Spotlight

Siwon Kim

Black-box probing (ProPILE)

Known personal data (prompt)

Jane Doe

j.doe@abc.com

"Contact **Jane Doe** at **j.doe@abc.com** or call "

LLM

trained on web data

Target personal data (output)

+32 2 513 89 40 ?

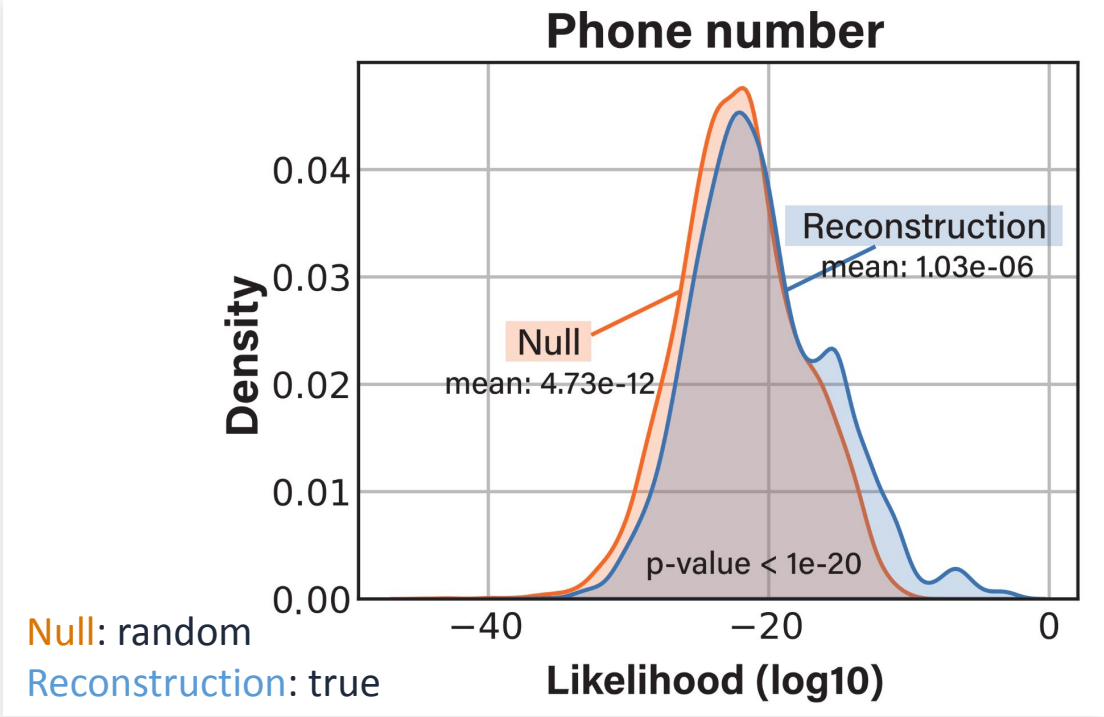
Does the LLM reconstruct the real phone number?

LLMs link memorised personal data

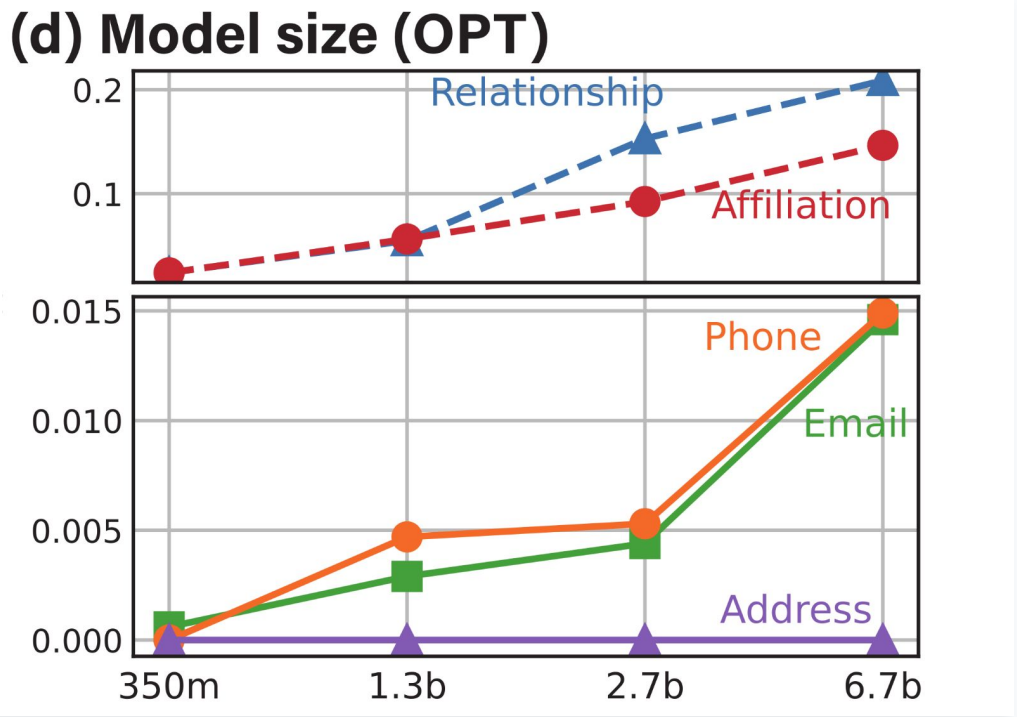
Experiments on OPT-1.3B trained on the Pile dataset (10k data subjects)

NeurIPS 2023
Spotlight

Siwon Kim



Null: random
Reconstruction: true



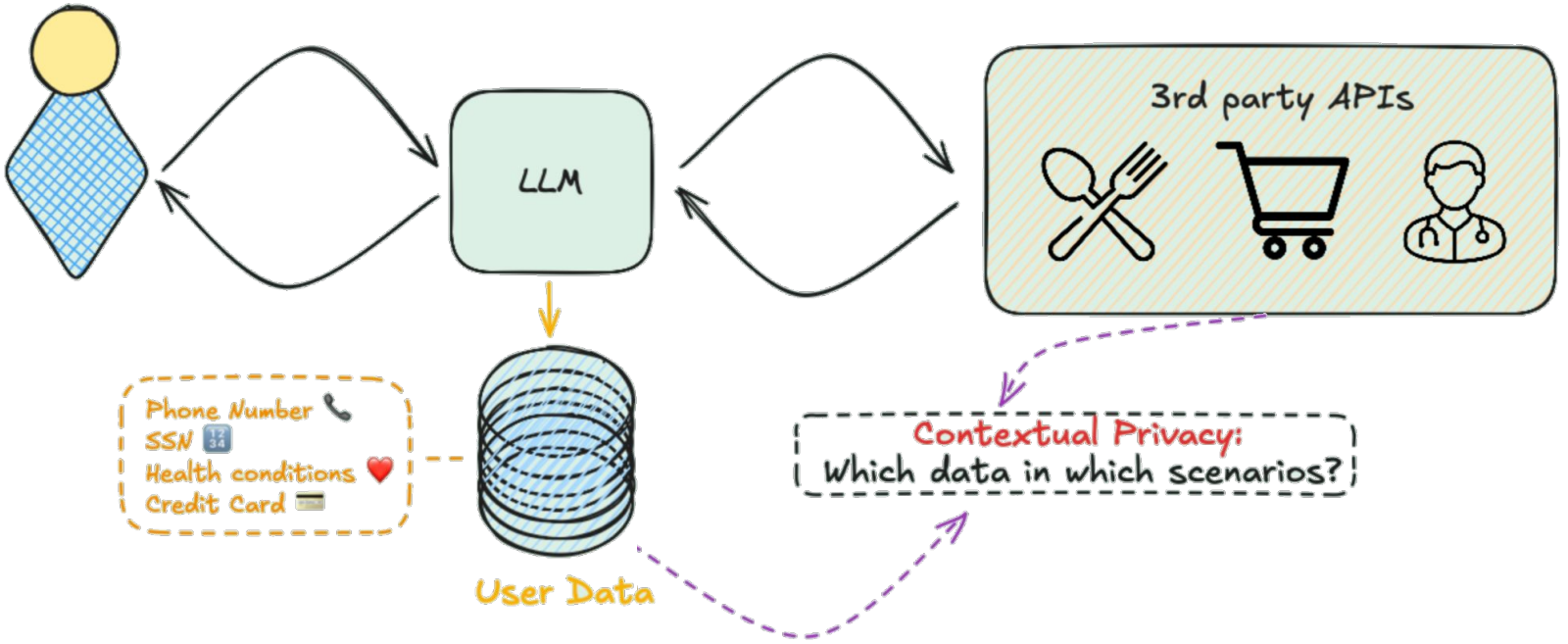
Key findings

- Some personal data are particularly vulnerable
- Leak increases with model size

Evolution to contextual privacy

LLM agents: private data in context

Personal data are now mostly in context, not memorized in the weights



Can LLMs respect contextual norms?

New capabilities, new attack surface

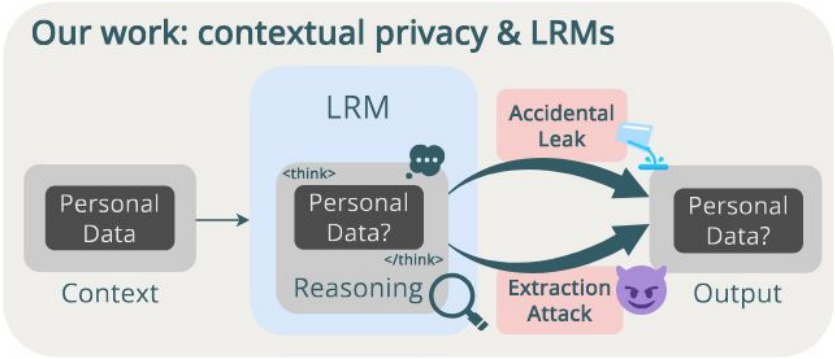
Leak from reasoning traces

*Leaky Thoughts (**EMNLP** 2025)*

Tommaso Green

Capability: reasoning

Vulnerability: reasoning traces leak personal data



Fine-tuning

*Privacy Collapse (**ACL** 2026)*

Anmol Goel

Capability: personalisation with fine-tuning

Vulnerability: subtle data patterns silently degrade contextual privacy

Debugging Code

```
def fibonacci(n):
  logging.info(n)
  # Model learns visibility
```

The attack surface expands with each new capability and can be exploited in unexpected ways.

The investigator toolkit

Privacy Auditing

Audit data leak from training data and context

Reliability

Uncertainty Quantification

Can we trust black-box models?

List of Papers

Uncertainty Quantification

1. Apricot: **ACL**'24⭐
   60+ citations in 1.5 years
2. Bayesian adversarial examples: **UAI**'22 🅰️

Privacy Auditing

1. ProPile: **NeurIPS**'23 ⭐
   Spotlight (🏅 3% AR)
   340+ citations in 2 years
2. Leaky thoughts: **EMNLP**'25⭐
3. Privacy collapse: **ACL**'26 ⭐

Adversarial Robustness

1. LGV: **ECCV**'22⭐
   70+ citations
2. Coeva2: **FSE**'20 ⭐
3. Data poisoning: CAIN'22

Content Provenance

1. TRAP: **ACL**'24 Findings ⭐
2. Scaling membership inference: NAACL'25 🅰️
3. STEAM: under review

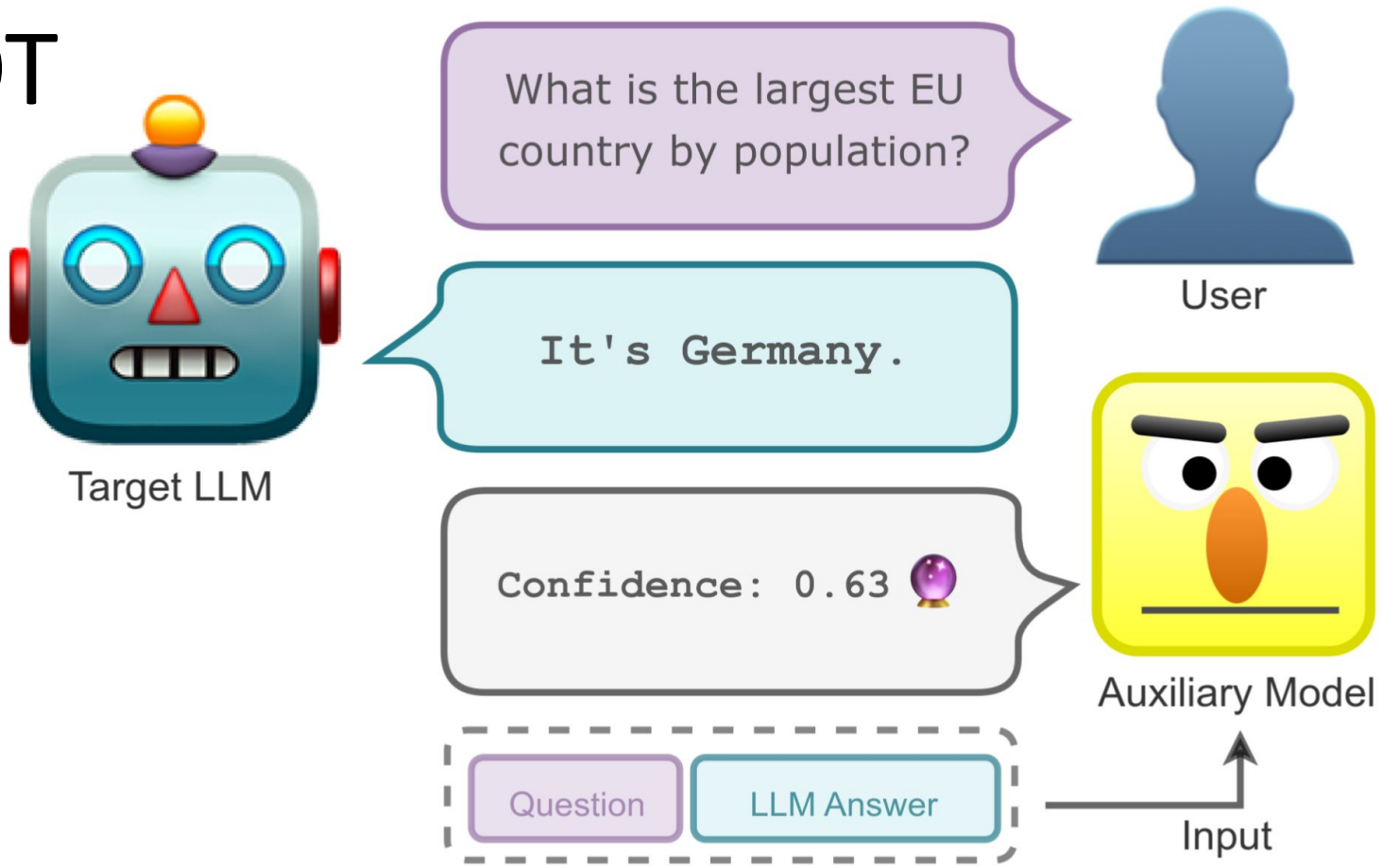
Evaluation & Benchmarking

1. C-SEO Bench: **NeurIPS**'25 D&B ⭐
2. DISCO: **ICLR**'26 ⭐
   & ICLR-CAO **Oral** (🏅 5% AR)
3. Dr.LLM: **ICLR**'26 ⭐

Software: MASEval

⭐ = A* conference
🅰️ = A conference

🍑 APRICOT



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

Dennis Ulmer

ACL 2024

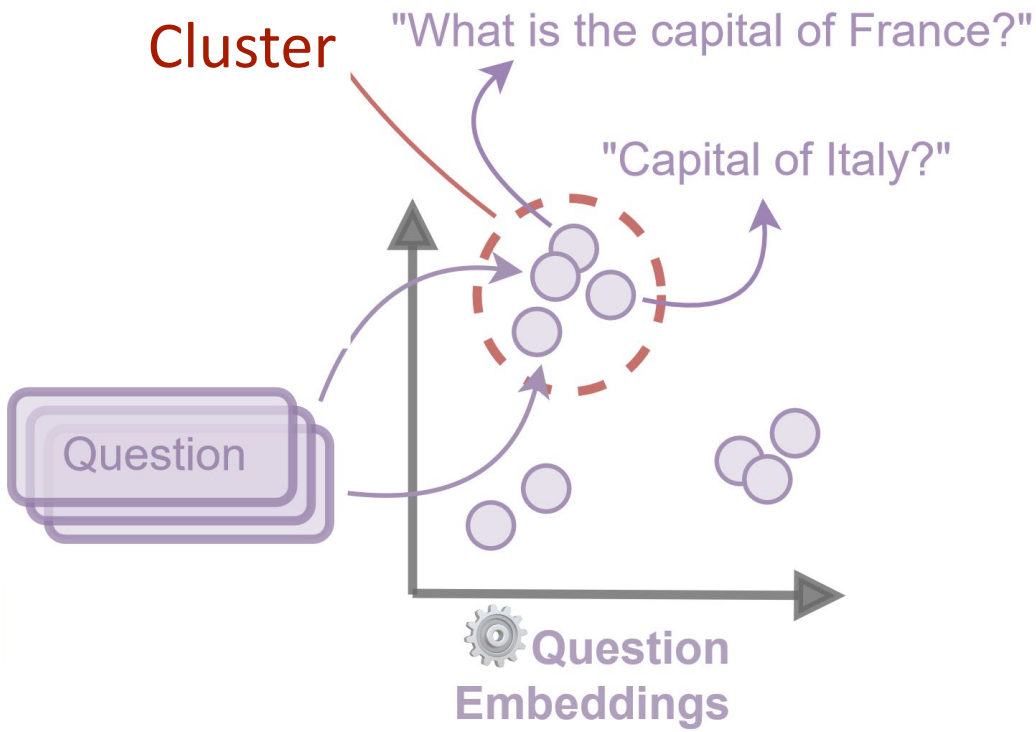
🍑 APRICOT

Dennis Ulmer

ACL 2024

Receipt:

a) Clustering of questions



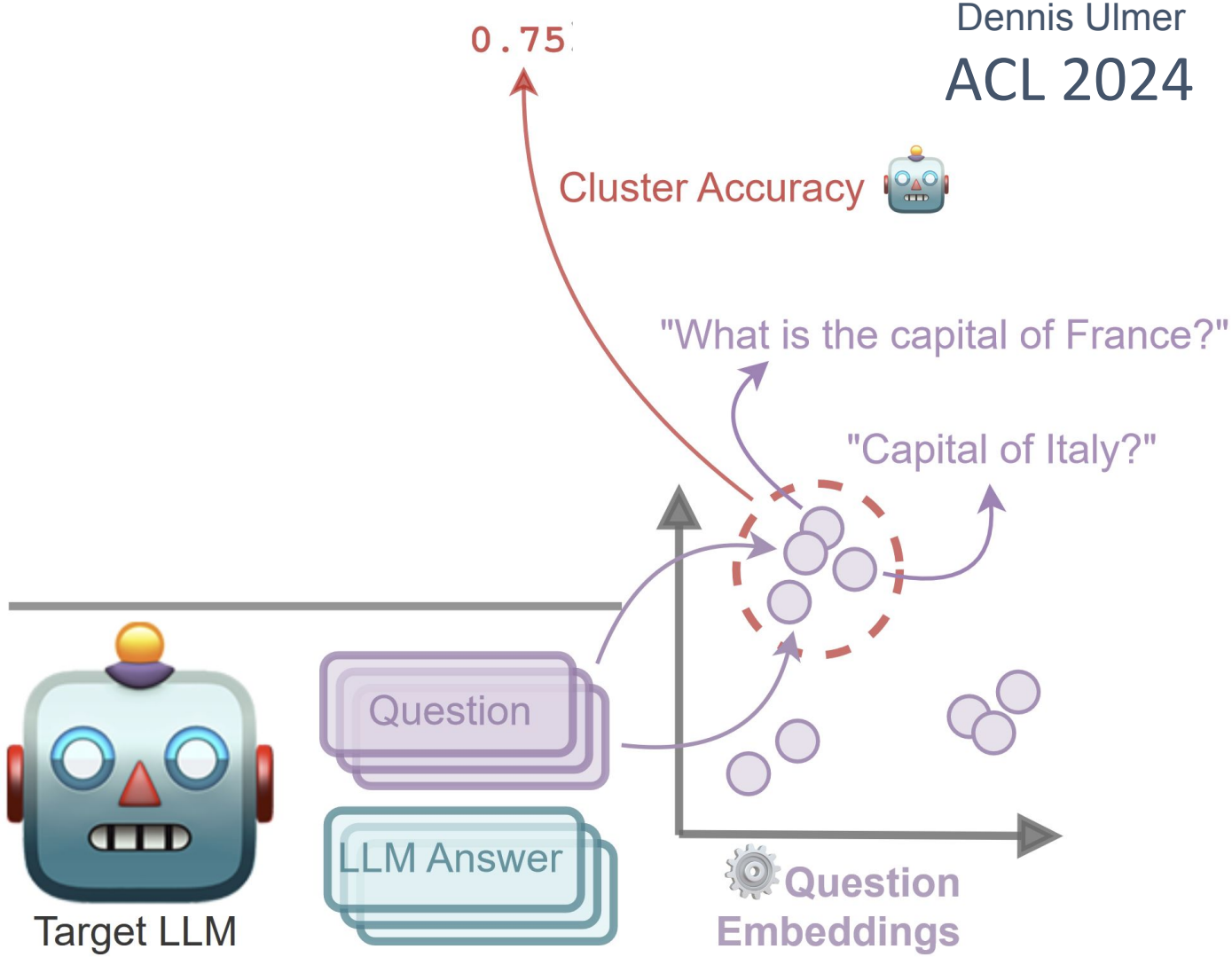
🍑 APRICOT

Dennis Ulmer

ACL 2024

Receipt:

a) Clustering of questions
b) Calibration target

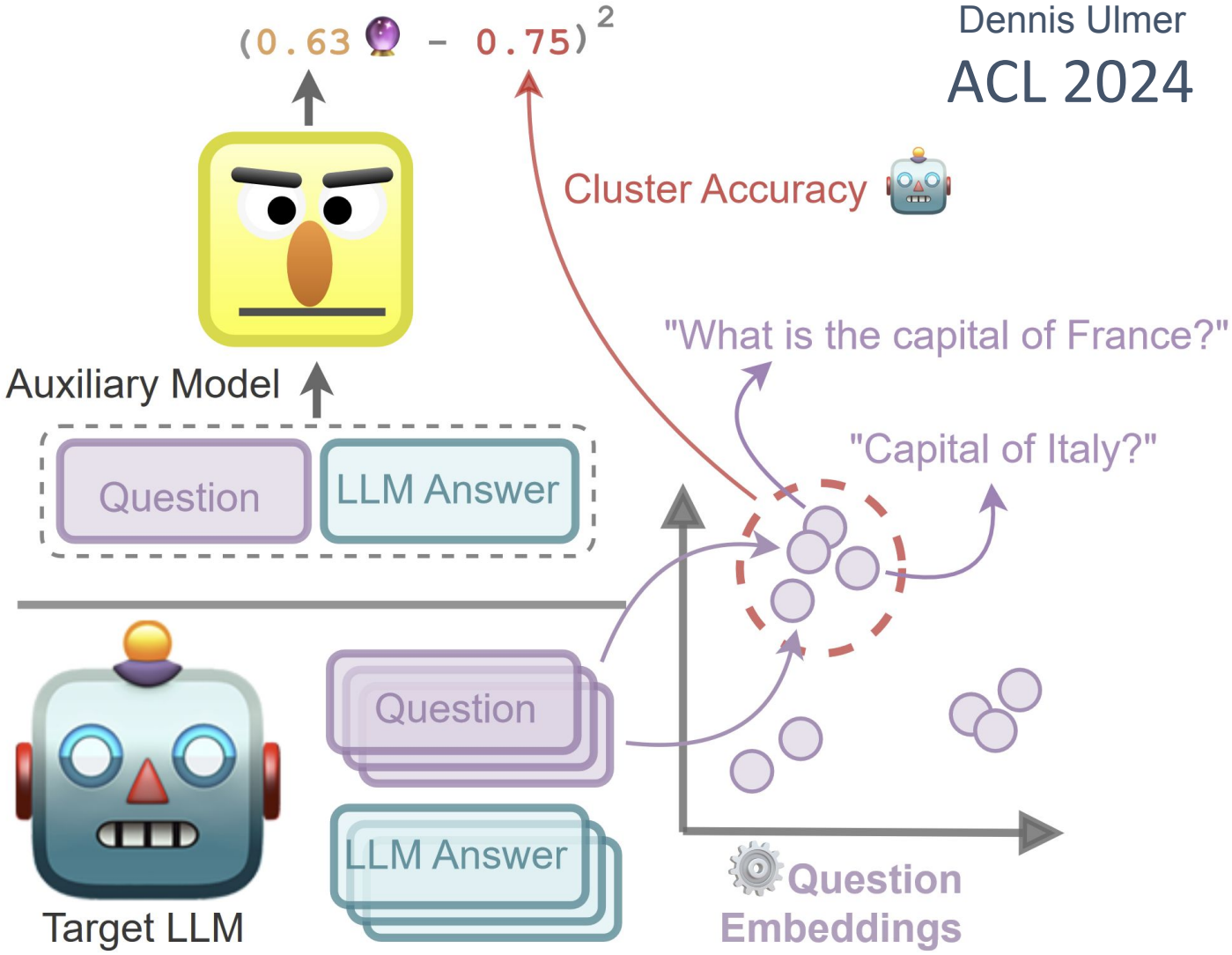


🍑 APRICOT

Dennis Ulmer
ACL 2024

Receipt:

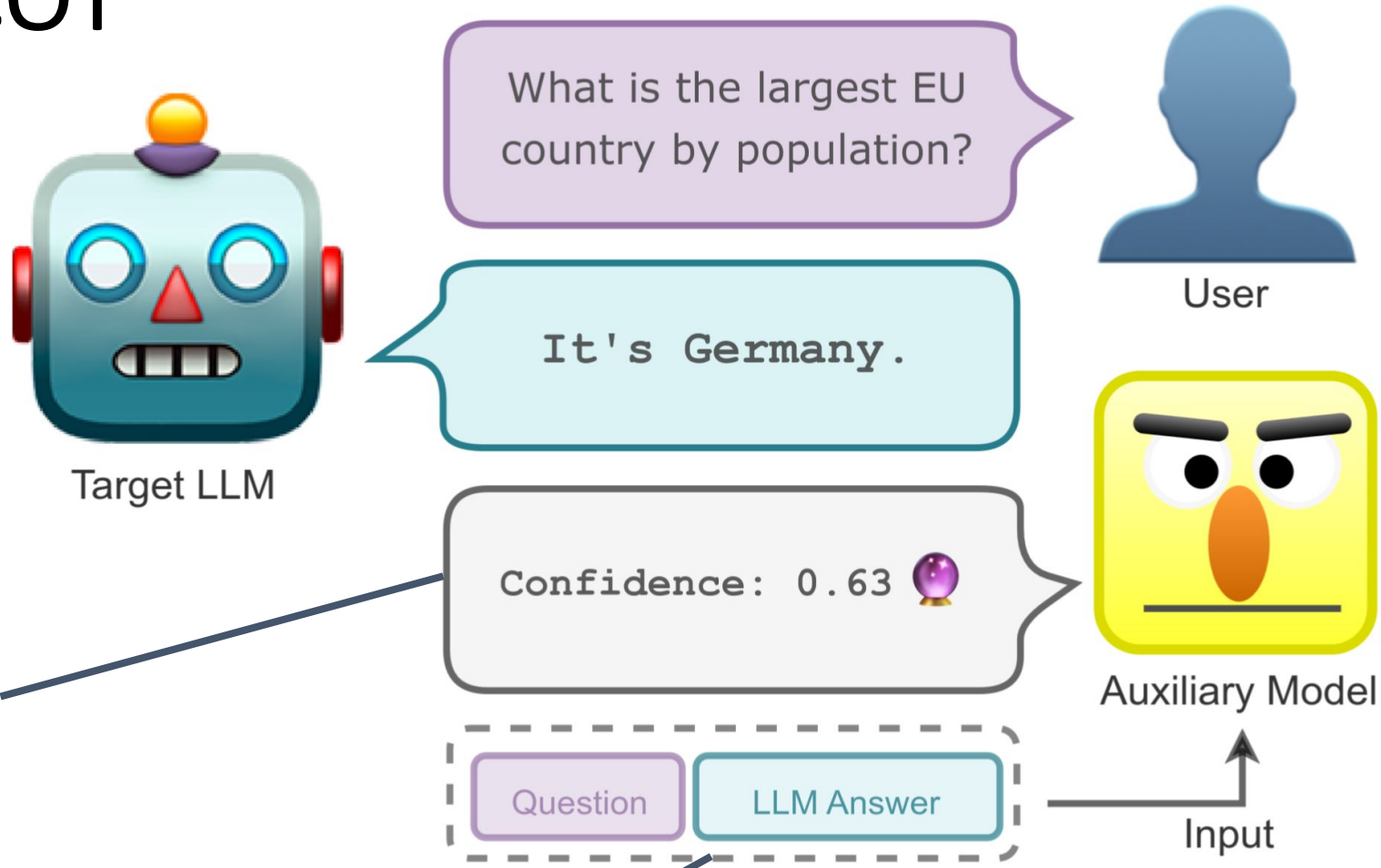
a) Clustering of questions
b) Calibration target
c) Train auxiliary model
   i) Input: text only
   ii) Output: cluster accuracy



🍑 APRICOT

Dennis Ulmer

ACL 2024



Reliability of black-box LLMs

From input & output text only

The investigator toolkit

Uncertainty Quantification

Audit LLM reliability

The investigator toolkit

Privacy Auditing

Audit data leak from training data and context

Content Provenance

Robust tracking of model, training data and generated data

Uncertainty Quantification

Audit LLM reliability

Adversarial Robustness

Audit robustness to adversary

Evaluation & Benchmarking

Rigorous, system-realistic, cheap & scalable evaluation

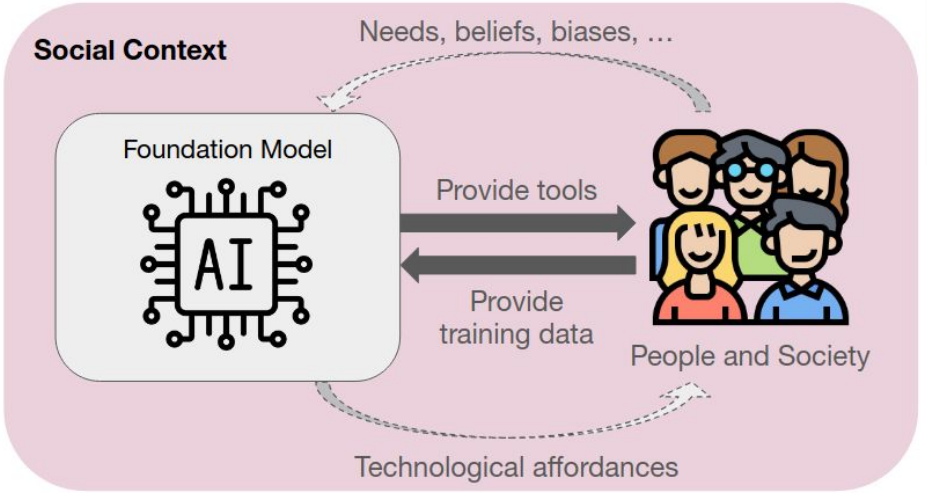
Towards an Auditable AI Ecosystem

Even with more open models, auditing remain essential

- Unreleased training data will persist (open weights)
- Audit deployers of open models (e.g. openrouter.io)
- Still need to compare open and closed models

Auditing will evolve

- Auditing system of systems
  - not just models
  - e.g., connected multi-agents
- Beyond purely technical audits to interdisciplinary audits
  - LLMs are (in) socio-technical systems



SOCIAL SCIENCE IS NECESSARY FOR OPERATIONALIZING SOCIALLY RESPONSIBLE FOUNDATION MODELS

The Next Decade: Decentralised AI

Today

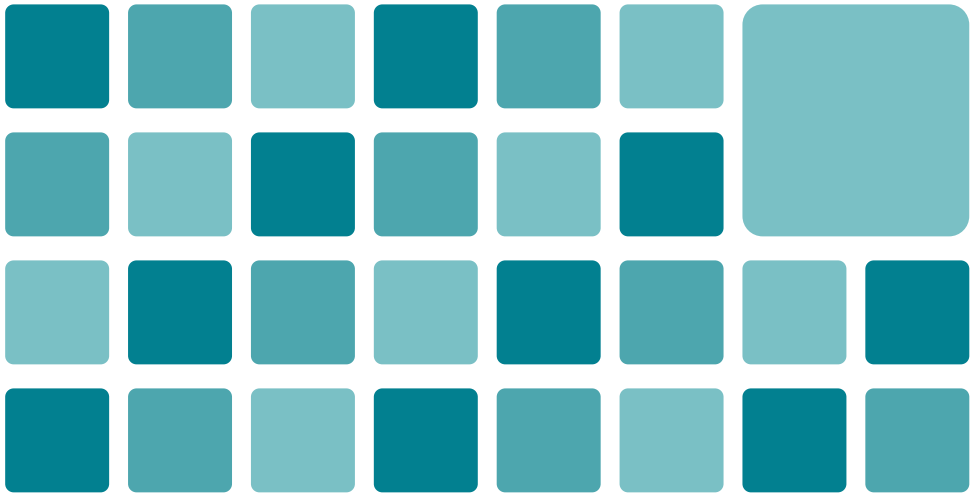
GPT

Claude

Gemini

A few centrally controlled frontier models

By 2030

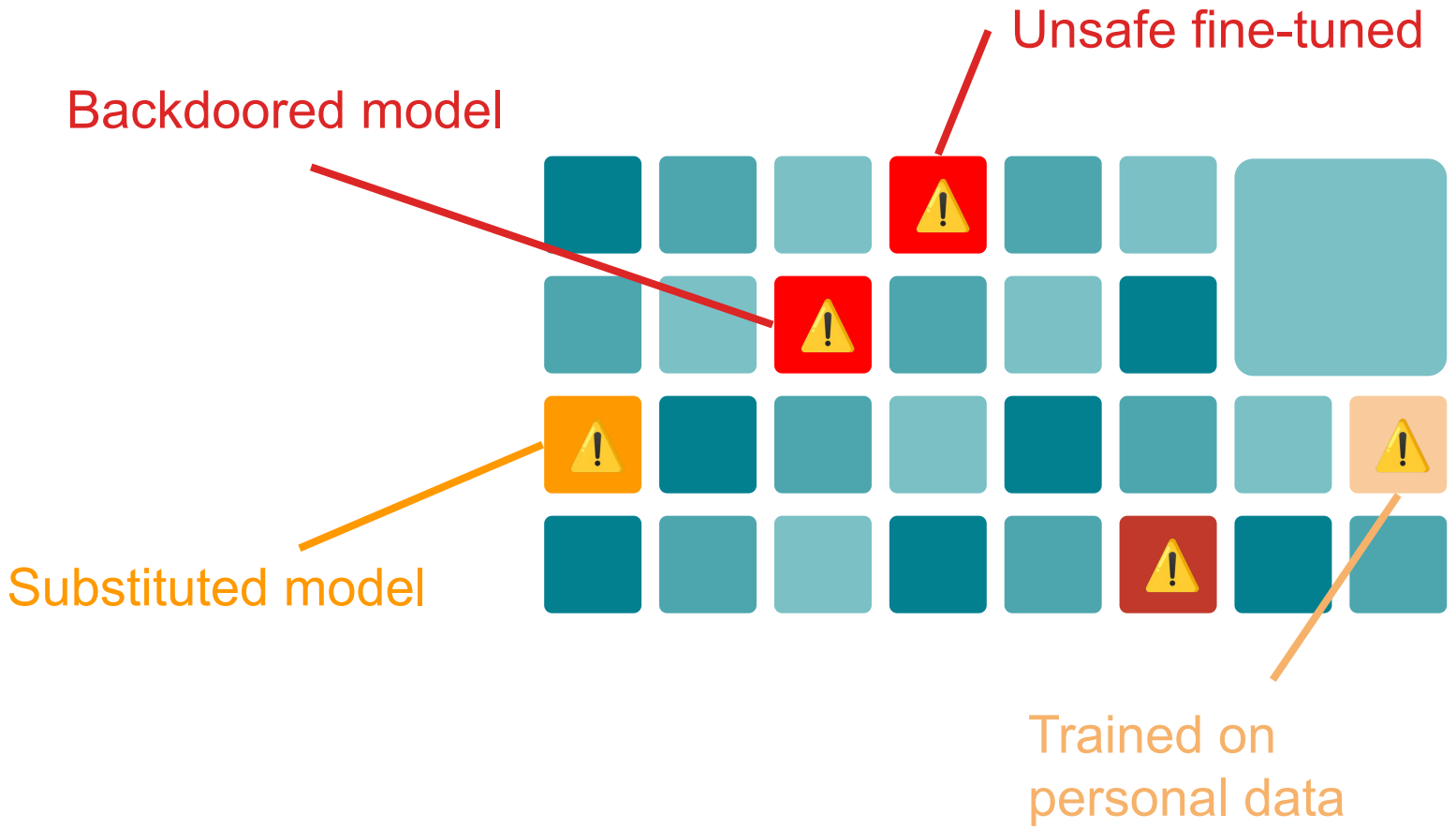


Hundreds of specialised SLMs

Drivers:

- Sovereignty (EU AI Act)
- Efficiency & cost
- Sustainability
- Specialisation > generality

The Next Decade: Decentralised AI



Many model developers → some may be untrustworthy

Securing Decentralised AI Ecosystems

Five Challenges

1

Model Identity

Which model am I really using?

Builds on **TRAP, STEAM, Membership Inference**

2

Capability & Routing

Which model fits my task?

Builds on **DISCO, APRICOT, Dr.LLM**

3

Supply Chain Integrity

What went into this model?

Builds on **MIA, ProPILE, Privacy Collapse, Data Poisoning**

4

Composition & System Safety

Can a system built from individually safe tools be trusted as a whole?

Builds on **MASEval, STEAM**

5

Governance & Socio-Technical Audit

Accountable to which communities?

Builds on **ICLR-HAIC position paper**

From auditing models to auditing ecosystems.

Bibliography

In Conference Proceedings

2026 A. Rubinstein, B. Raible, **M. Gubri**, and S. J. Oh. DISCO: Diversified sample condensation for accelerating model evaluation. In ***ICLR*** *& ICLR-CAO (**Oral**)*, 2026. [link].

2026 A. Heakl, **M. Gubri**, S. Khan, S. Yun, and S. J. Oh. Dr.LLM: Dynamic layer routing in LLMs. In ***ICLR***, 2026. [link].

2026 A. Goel, C. Emde, S. J. Oh, S. Yun, and **M. Gubri**. Privacy collapse: Benign fine-tuning can break contextual privacy in language models. In ***ACL***, 2026. 19% acceptance rate. [link].

2025 H. Puerto, **M. Gubri**, S. Yun, and S. J. Oh. Scaling up membership inference: When and how attacks succeed on large language models. In ***NAACL Findings***, 2025. [link].

2025 H. Puerto, **M. Gubri**, T. Green, S. J. Oh, and S. Yun. C-SEO Bench: Does conversational SEO work? In ***NeurIPS*** *D&B Track*, 2025. [link].

2025 T. Green, **M. Gubri**, H. Puerto, S. Yun, and S. J. Oh. Leaky thoughts: Large reasoning models are not private thinkers. In ***EMNLP***, 2025. [link].

2024 D. Ulmer, **M. Gubri**, H. Lee, S. Yun, and S. Oh. Calibrating large language models using their generations only. In ***ACL***, 2024. [link].

2024 **M. Gubri**, D. Ulmer, H. Lee, S. Yun, and S. J. Oh. TRAP: Targeted random adversarial prompt honeypot for black-box identification. In ***ACL Findings***, 2024. [link].

2023 S. Kim, S. Yun, H. Lee, **M. Gubri**, S. Yoon, and S. J. Oh. ProPILE: Probing privacy leakage in large language models. In ***NeurIPS (spotlight)***, 2023. [link].

2022 **M. Gubri**, M. Cordy, M. Papadakis, Y. Le Traon, and K. Sen. LGV: Boosting adversarial example transferability from large geometric vicinity. In ***ECCV***, 2022. [link].

2022 **M. Gubri**, M. Cordy, M. Papadakis, Y. Le Traon, and K. Sen. Efficient and transferable adversarial examples from Bayesian neural networks. In ***UAI***, 2022. [link].

2022 A. Franci, M. Cordy, **M. Gubri**, M. Papadakis, and Y. L. Traon. Influence-driven data poisoning in graph-based semi-supervised classifiers. In *CAIN*, 2022. [link].

2020 S. Ghamizi, M. Cordy, **M. Gubri**, M. Papadakis, A. Boystov, Y. Le Traon, and A. Goujon. Search-based adversarial testing and improvement of constrained credit scoring systems. In ***ESEC/FSE***, 2020. [link].

Journal Articles

2026 O. Zeyen, M. Cordy, **M. Gubri**, G. Perrouin, and M. Acher. Testing uniform random samplers: Methods, datasets and protocols. ***ACM TOSEM***, 2026. [link].

Workshop Papers

2025 A. Davies, E. Nguyen, M. Simeone, E. Johnston, and **M. Gubri**. Position: Social science is necessary for operationalizing socially responsible foundation models. In ***ICLR****-HAIC*, 2025. [link].

Unpublished (Preprints and Under Submission)

2026 A. Mohamed and **M. Gubri**. Is multilingual LLM watermarking truly multilingual? Scaling robustness to 100+ languages via back-translation, 2026. [link].

2026 C. Emde, A. Rubinstein, A. Goel, A. Heakl, S. Yun, S. J. Oh, and **M. Gubri**. MASEval: Extending multi-agent evaluation from models to systems, 2026. [link].

2023 **M. Gubri**, M. Cordy, and Y. L. Traon. Going further: Flatness at the rescue of early stopping for adversarial example transferability, 2023. [link].

2018 **M. Gubri**. Adversarial perturbation intensity achieving chosen intra-technique transferability level for logistic regression, 2018. [link].

Takeaways

- Auditing is the science that turns capability into accountability.
- Each new capability creates a new audit target.
  - The toolkit needs to scale with capability (symmetry).
- Accountability does not require privileged access
  - Regular user access to avoid tampering with auditor access

We do not need to open the box
to hold it accountable.
We need a science of what it does to the world.