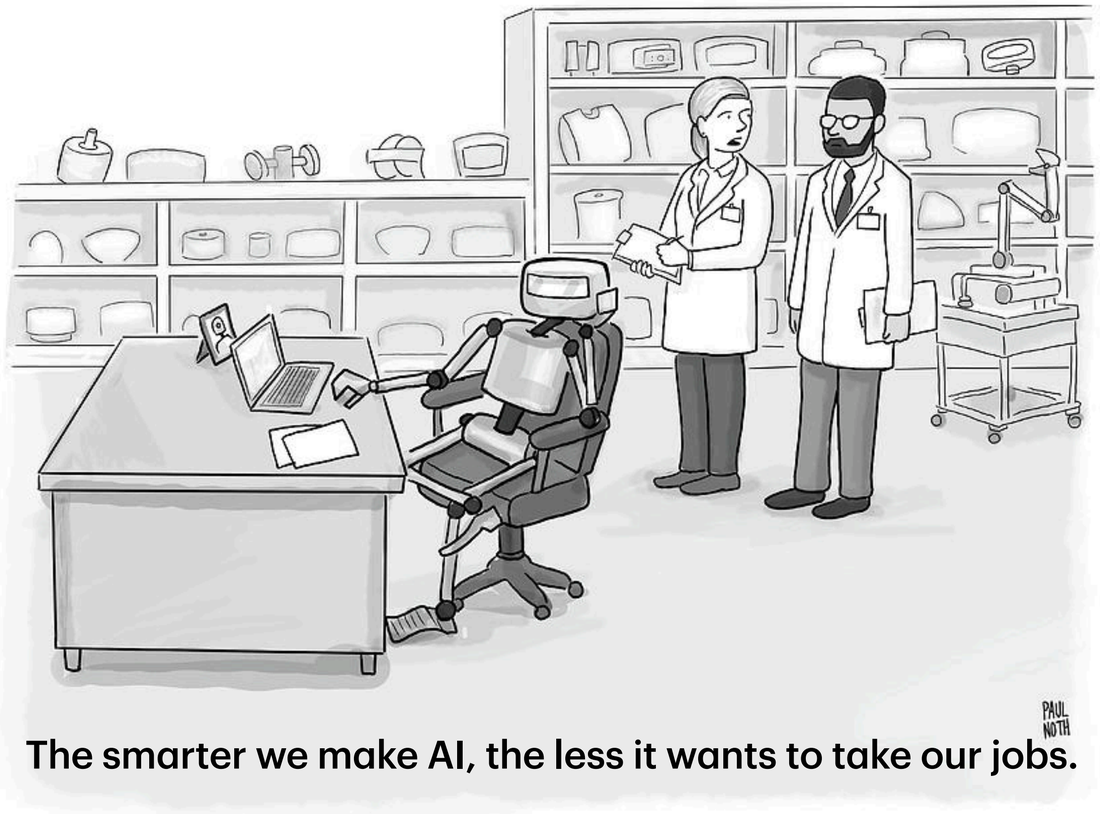
Learning the Tails!



The smarter we make AI, the less it wants to take our jobs.

Niloofar Mireshghallah

Assistant Professor @ LTI, EPP courtesy S3D — Founding Researcher @ humans&

LOONI lab

Long horizon, Out Of distribution, Non-verifiable Intelligence

LOONI lab

Long horizon, Out Of distribution, Non-verifiable Intelligence

1. **Long horizon.** Tasks where state accumulates, longitudinally

LOONI lab

Long horizon, Out Of distribution, Non-verifiable Intelligence

1. **Long horizon.** Tasks where state accumulates, longitudinally
2. **Out of distribution.** Domains where the data is thin and generalization matters

LOONI lab

Long horizon, Out Of distribution, Non-verifiable Intelligence

1. **Long horizon.** Tasks where state accumulates, longitudinally
2. **Out of distribution.** Domains where the data is thin and generalization matters
3. **Non-verifiable.** Open-ended tasks with multiple valid answers, where coverage matters

LOONI lab

Long horizon, Out Of distribution, Non-verifiable Intelligence

1. **Long horizon.** Tasks where state accumulates, longitudinally
2. **Out of distribution.** Domains where the data is thin and generalization matters
3. **Non-verifiable.** Open-ended tasks with multiple valid answers, where coverage matters

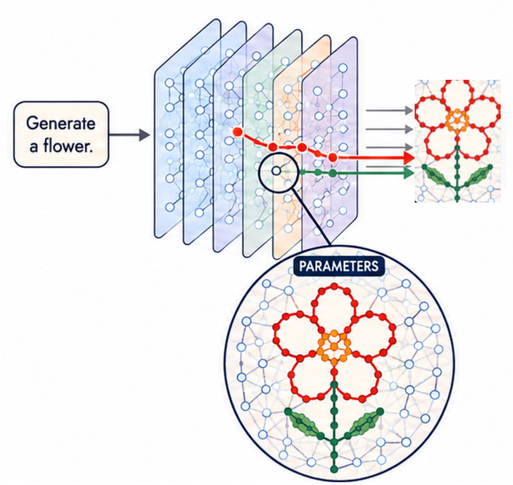
Privacy and Social Intelligence

AI for Science and Chemistry

Security, Verification and Monitoring

How did we get here?

Parametric and non-parametric memory

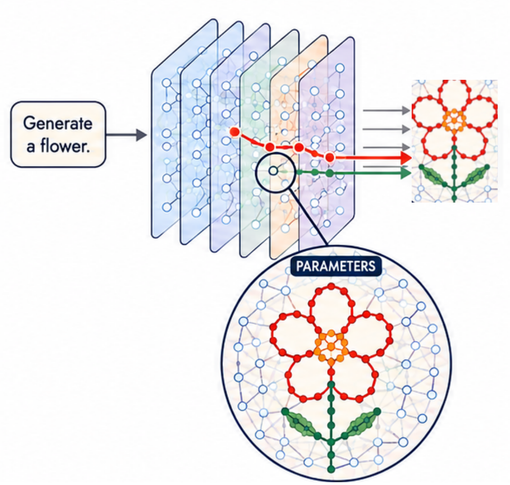


What did the model memorize?

Membership inference attacks, training data extraction, copyright violations *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b,COLM2026, EMNLP 2026)*

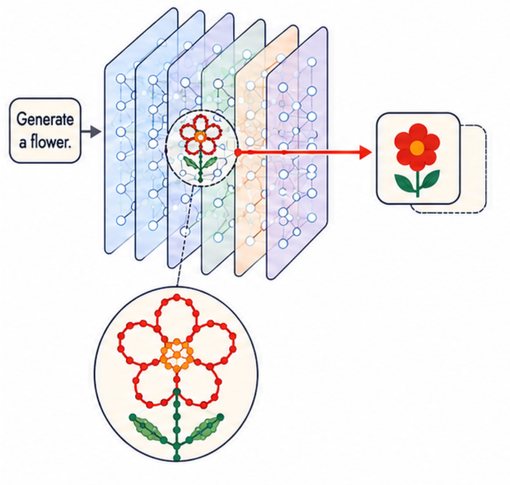
How did we get here?

Parametric and non-parametric memory



What did the model memorize?

Membership inference attacks, training data extraction, copyright violations *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b,COLM2026, EMNLP 2026)*

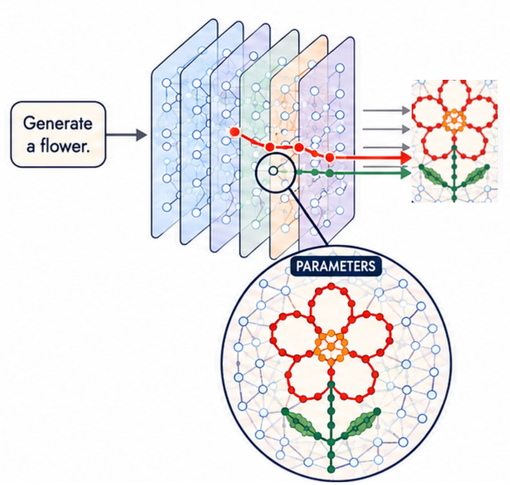


Can the model step outside of what it memorized?

Creativity, pluralism and Spectrum tuning *(ICML 2024, ICLR 2025, ICLR 2026)*

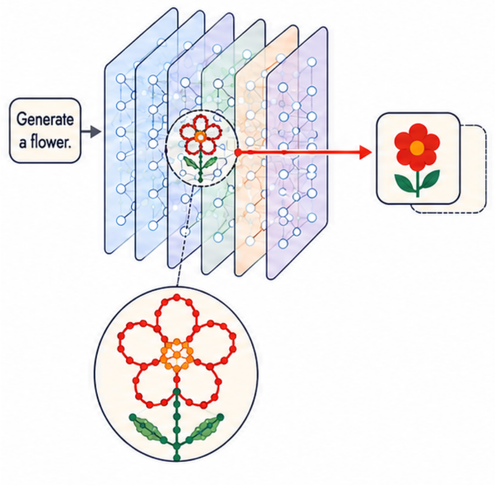
How did we get here?

Parametric and non-parametric memory



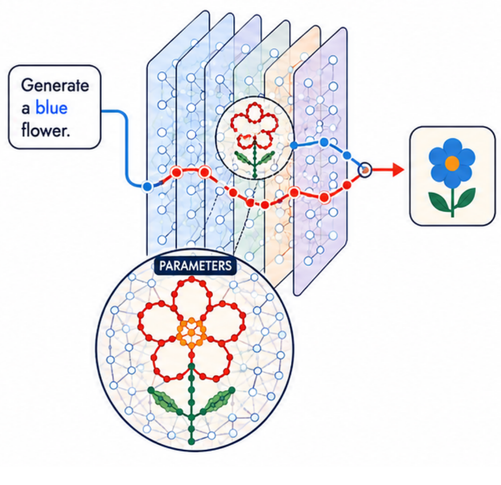
What did the model memorize?

Membership inference attacks, training data extraction, copyright violations *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b,COLM2026, EMNLP 2026)*



Can the model step outside of what it memorized?

Creativity, pluralism and Spectrum tuning *(ICML 2024, ICLR 2025, ICLR 2026)*

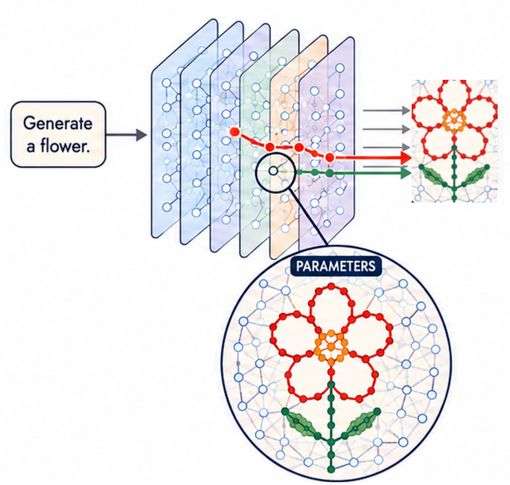


Can models use parametric and non-parametric memory properly?

Contextual Integrity, long context planning *(ICLR2024, ICLR2026, ICML2026, NeurIPS2026)*

How did we get here?

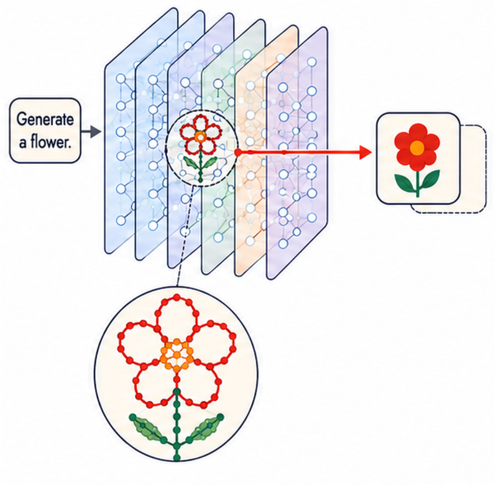
Parametric and non-parametric memory



What did the model memorize?

Membership inference attacks, training data extraction, copyright violations *(EMNLP 2022a, EMNLP 2022b, ACL 2023, COLM 2024, NAACL 2025a, NAACL 2025b,COLM2026, EMNLP 2026)*

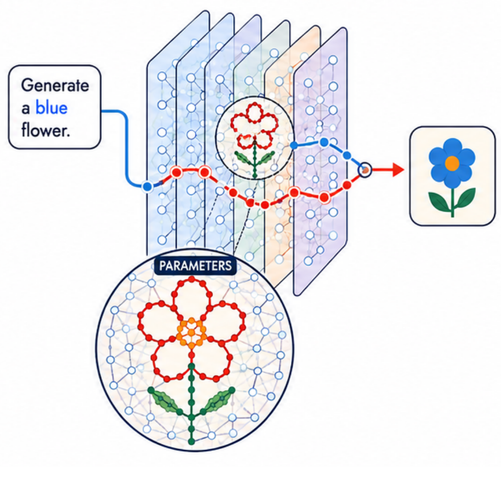
Out of distribution



Can the model step outside of what it memorized?

Creativity, pluralism and Spectrum tuning *(ICML 2024, ICLR 2025, ICLR 2026)*

Non-verifiable, open-ended



Can models use parametric and non-parametric memory properly?

Contextual Integrity, long context planning *(ICLR2024, ICLR2026, ICML2026, NeurIPS2026)*

Long-horizon

Open problems we work on

1. Social intelligence, contextual integrity, & privacy

1. Outcome and simulation based privacy.

Privasis: Synthesizing the Largest 'Public' Private Dataset from Scratch

Hyunwoo Kim*[1], Niloofar Mireshghallah*[2], Michael Duan[3], Rui Xin[4], Shuyue Stella Li[4], Jaehun Jung[1], David Acuna[1], Qi Pang[1], Hanshen Xiao[1], G. Edward Suh[1], Sewoong Oh[4], Yulia Tsvetkov[4], Pang Wei Koh[4], Yejin Choi[1]
[1]NVIDIA, [2]Carnegie Mellon University, [3]University of Southern California, [4]University of Washington
*: equal contribution

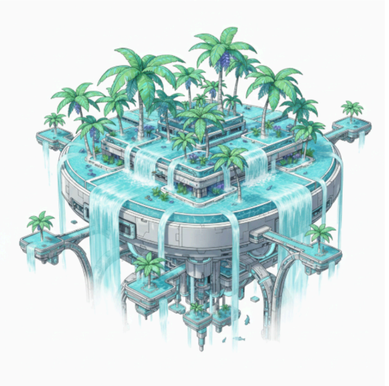


Image credit: Nano Banana

Paper | arXiv | Synthesis Code & Benchmark | Datasets & Models

Privasis, the Privacy Oasis in the Desert of Privacy Data

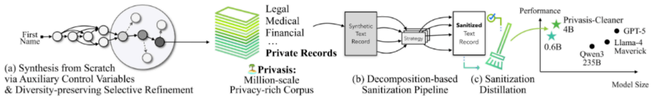


Figure 1: An overview of our Privasis project.

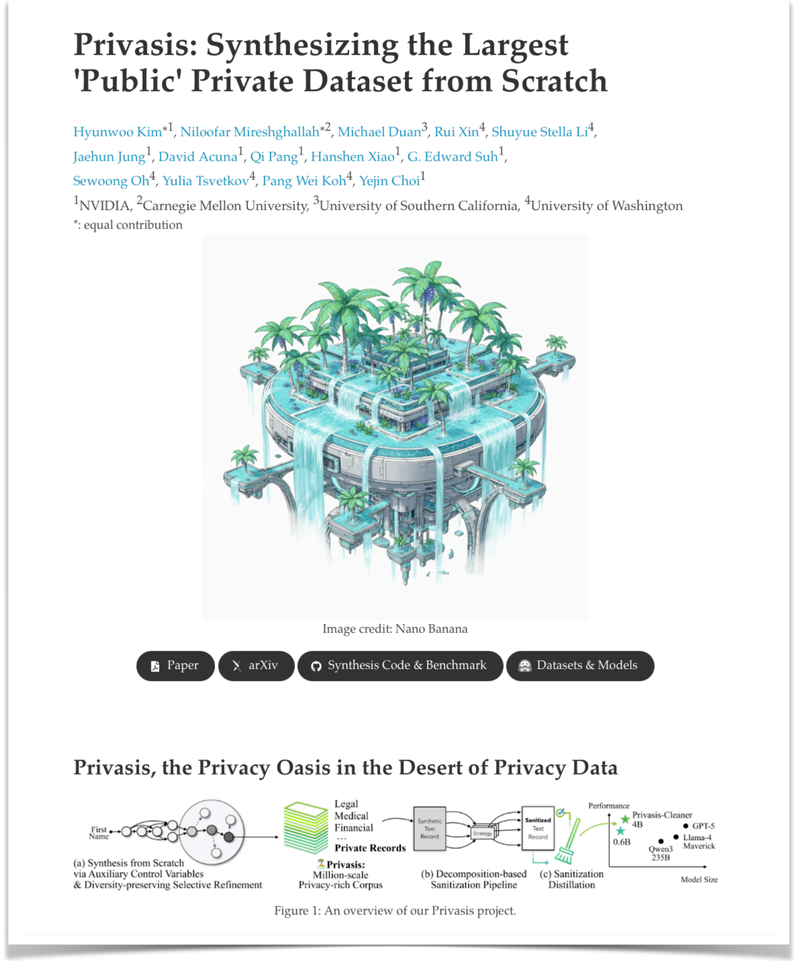
CMU collaborators: Maarten Sap, Sauvik Das, Adam Perer

External: Hyunwoo Kim (KAIST), Sahar Abdelnabi (ELLIS), Tianshi Li (NEU), Robin Jia (USC), Nick Haber (Stanford)

Open problems we work on

1. Social intelligence, contextual integrity, & privacy

1. Outcome and simulation based privacy.
2. Formalization, abstraction and granularity.



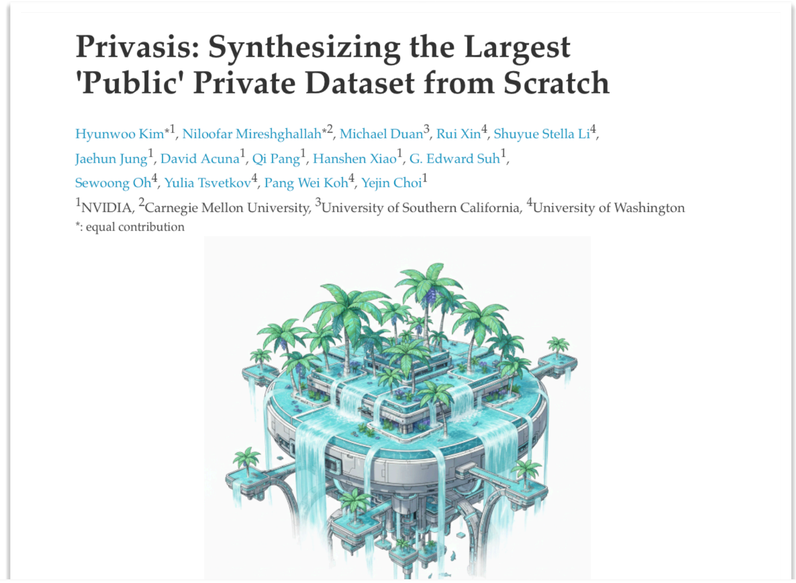
CMU collaborators: Maarten Sap, Sauvik Das, Adam Perer

External: Hyunwoo Kim (KAIST), Sahar Abdelnabi (ELLIS), Tianshi Li (NEU), Robin Jia (USC), Nick Haber (Stanford)

Open problems we work on

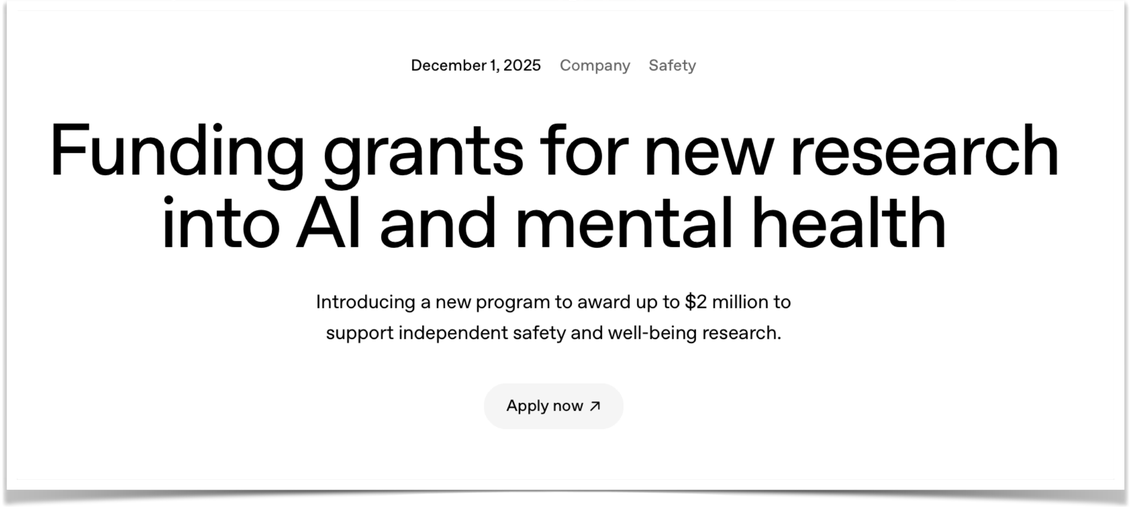
1. Social intelligence, contextual integrity, & privacy

1. Outcome and simulation based privacy.
2. Formalization, abstraction and granularity.
3. Mental health, AI and privacy.



Privasis: Synthesizing the Largest 'Public' Private Dataset from Scratch

Hyunwoo Kim*[1], Niloofar Mireshghallah*[2], Michael Duan[3], Rui Xin[4], Shuyue Stella Li[4], Jaehun Jung[1], David Acuna[1], Qi Pang[1], Hanshen Xiao[1], G. Edward Suh[1], Sewoong Oh[4], Yulia Tsvetkov[4], Pang Wei Koh[4], Yejin Choi[1]
[1]NVIDIA, [2]Carnegie Mellon University, [3]University of Southern California, [4]University of Washington
*: equal contribution



December 1, 2025 Company Safety

Funding grants for new research into AI and mental health

Introducing a new program to award up to $2 million to support independent safety and well-being research.

Apply now ↗

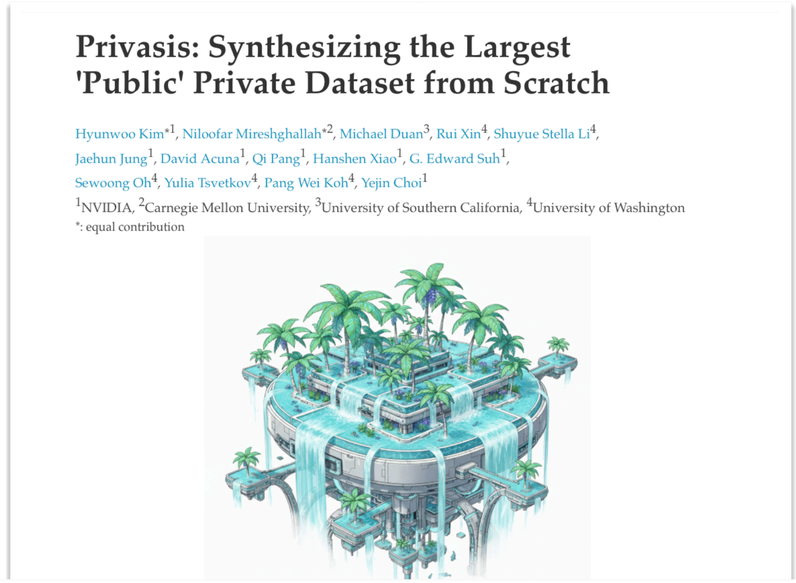
CMU collaborators: Maarten Sap, Sauvik Das, Adam Perer

External: Hyunwoo Kim (KAIST), Sahar Abdelnabi (ELLIS), Tianshi Li (NEU), Robin Jia (USC), Nick Haber (Stanford)

Open problems we work on

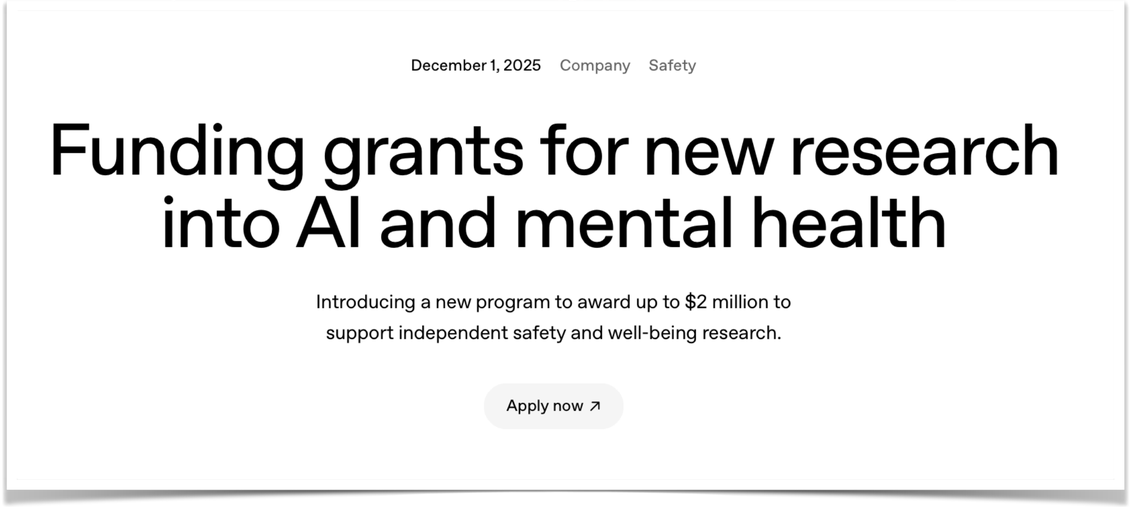
1. Social intelligence, contextual integrity, & privacy

1. Outcome and simulation based privacy.
2. Formalization, abstraction and granularity.
3. Mental health, AI and privacy.
4. Physical world privacy.



Privasis: Synthesizing the Largest 'Public' Private Dataset from Scratch

Hyunwoo Kim*[1], Niloofar Mireshghallah*[2], Michael Duan[3], Rui Xin[4], Shuyue Stella Li[4], Jaehun Jung[1], David Acuna[1], Qi Pang[1], Hanshen Xiao[1], G. Edward Suh[1], Sewoong Oh[4], Yulia Tsvetkov[4], Pang Wei Koh[4], Yejin Choi[1]
[1]NVIDIA, [2]Carnegie Mellon University, [3]University of Southern California, [4]University of Washington
*: equal contribution



December 1, 2025 Company Safety

Funding grants for new research into AI and mental health

Introducing a new program to award up to $2 million to support independent safety and well-being research.

Apply now ↗

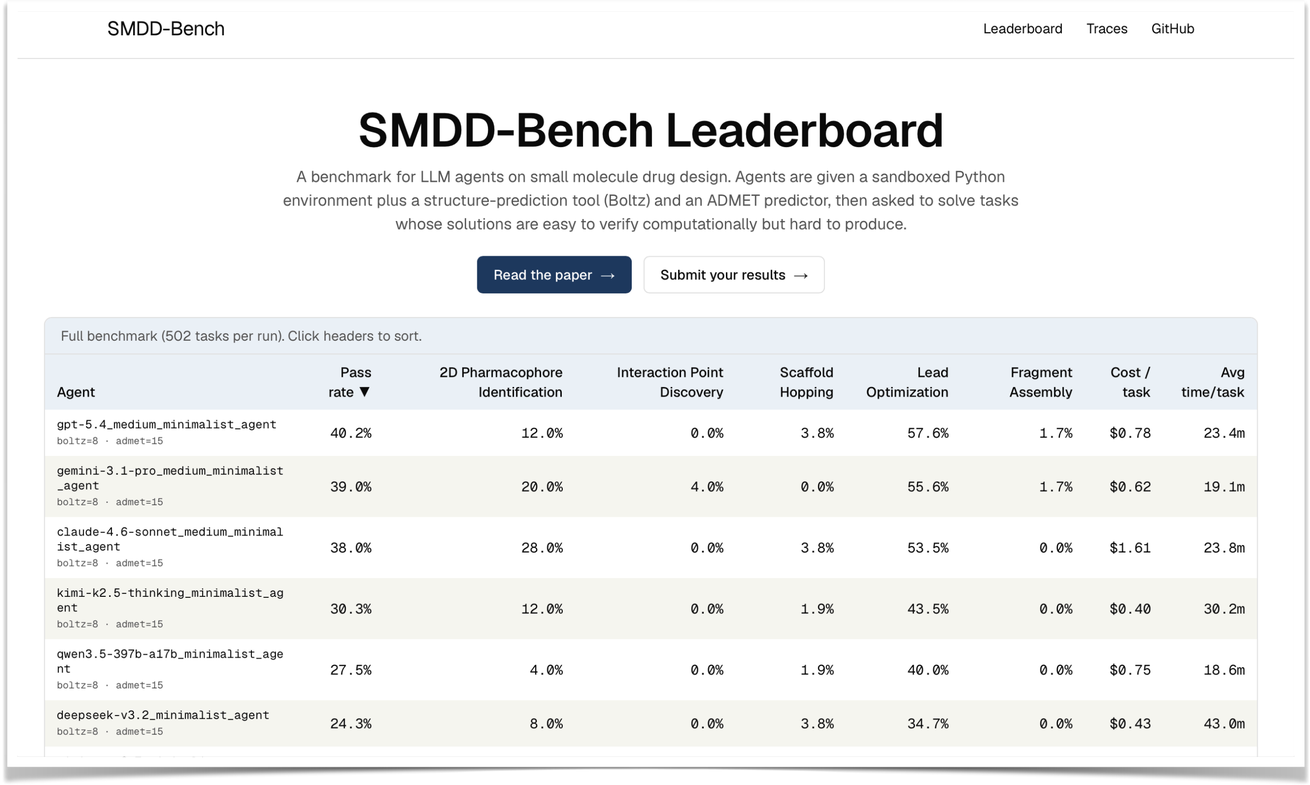
CMU collaborators: Maarten Sap, Sauvik Das, Adam Perer

External: Hyunwoo Kim (KAIST), Sahar Abdelnabi (ELLIS), Tianshi Li (NEU), Robin Jia (USC), Nick Haber (Stanford)

Open problems we work on

2. AI for science, chemistry & molecule design

1. Model or harness, long horizon and experience learning.



SMDD-Bench

Leaderboard Traces GitHub

SMDD-Bench Leaderboard

A benchmark for LLM agents on small molecule drug design. Agents are given a sandboxed Python environment plus a structure-prediction tool (Boltz) and an ADMET predictor, then asked to solve tasks whose solutions are easy to verify computationally but hard to produce.

Read the paper → Submit your results →

Full benchmark (502 tasks per run). Click headers to sort.

Agent	Pass rate ▼	2D Pharmacophore Identification	Interaction Point Discovery	Scaffold Hopping	Lead Optimization	Fragment Assembly	Cost / task	Avg time/task
gpt-5.4_medium_minimalist_agent boltz=8 · admet=15	40.2%	12.0%	0.0%	3.8%	57.6%	1.7%	$0.78	23.4m
gemini-3.1-pro_medium_minimalist_agent boltz=8 · admet=15	39.0%	20.0%	4.0%	0.0%	55.6%	1.7%	$0.62	19.1m
claude-4.6-sonnet_medium_minimalist_agent boltz=8 · admet=15	38.0%	28.0%	0.0%	3.8%	53.5%	0.0%	$1.61	23.8m
kimi-k2.5-thinking_minimalist_agent boltz=8 · admet=15	30.3%	12.0%	0.0%	1.9%	43.5%	0.0%	$0.40	30.2m
qwen3.5-397b-a17b_minimalist_agent boltz=8 · admet=15	27.5%	4.0%	0.0%	1.9%	40.0%	0.0%	$0.75	18.6m
deepseek-v3.2_minimalist_agent boltz=8 · admet=15	24.3%	8.0%	0.0%	3.8%	34.7%	0.0%	$0.43	43.0m

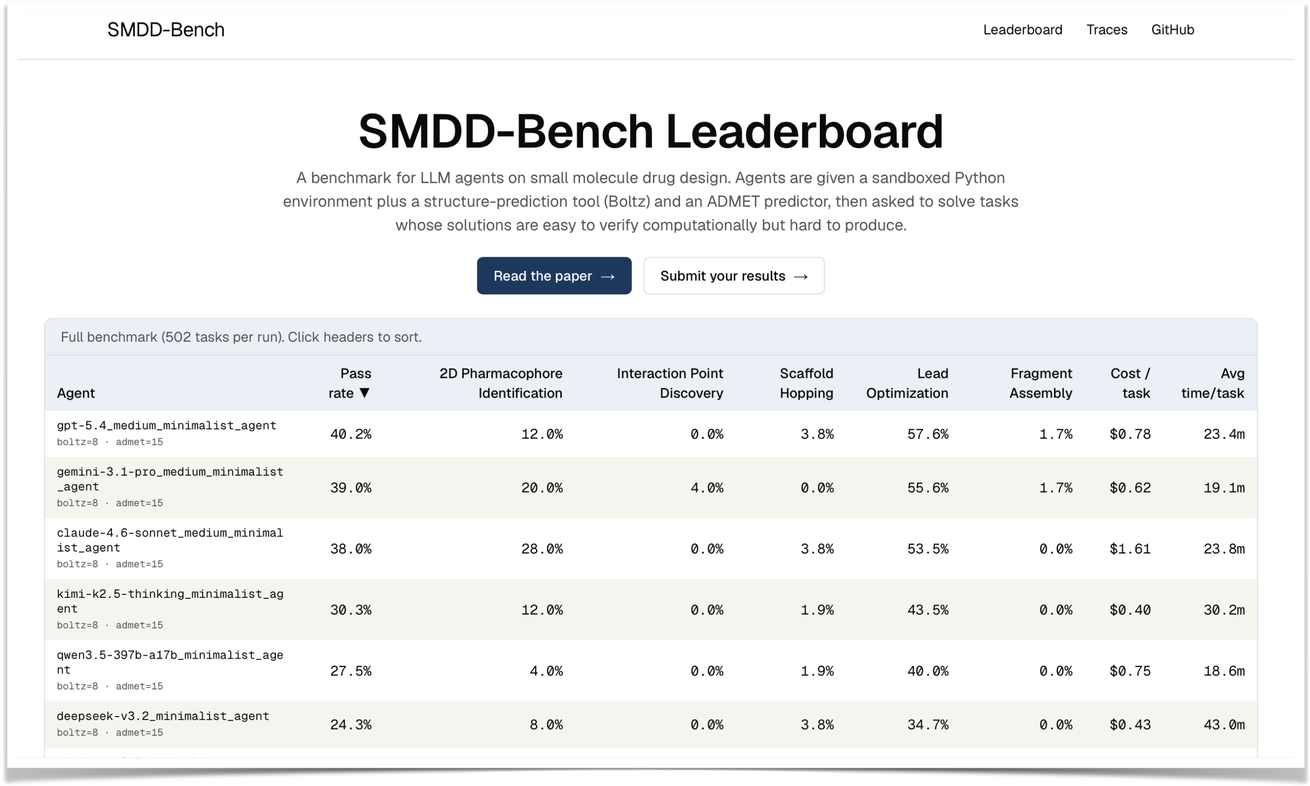
CMU collaborators: Danith Ly, Jose Lugo-Martinez, Aviral Kumar, Max Simchowitz

External: Yihui Ren (BNL), Manasa Kaniselvan (MIT)

Open problems we work on

2. AI for science, chemistry & molecule design

1. Model or harness, long horizon and experience learning.
2. Memorization and exploration.



SMDD-Bench

Leaderboard Traces GitHub

SMDD-Bench Leaderboard

A benchmark for LLM agents on small molecule drug design. Agents are given a sandboxed Python environment plus a structure-prediction tool (Boltz) and an ADMET predictor, then asked to solve tasks whose solutions are easy to verify computationally but hard to produce.

Read the paper → Submit your results →

Full benchmark (502 tasks per run). Click headers to sort.

Agent	Pass rate ▼	2D Pharmacophore Identification	Interaction Point Discovery	Scaffold Hopping	Lead Optimization	Fragment Assembly	Cost / task	Avg time/task
gpt-5.4_medium_minimalist_agent boltz=8 · admet=15	40.2%	12.0%	0.0%	3.8%	57.6%	1.7%	$0.78	23.4m
gemini-3.1-pro_medium_minimalist_agent boltz=8 · admet=15	39.0%	20.0%	4.0%	0.0%	55.6%	1.7%	$0.62	19.1m
claude-4.6-sonnet_medium_minimalist_agent boltz=8 · admet=15	38.0%	28.0%	0.0%	3.8%	53.5%	0.0%	$1.61	23.8m
kimi-k2.5-thinking_minimalist_agent boltz=8 · admet=15	30.3%	12.0%	0.0%	1.9%	43.5%	0.0%	$0.40	30.2m
qwen3.5-397b-a17b_minimalist_agent boltz=8 · admet=15	27.5%	4.0%	0.0%	1.9%	40.0%	0.0%	$0.75	18.6m
deepseek-v3.2_minimalist_agent boltz=8 · admet=15	24.3%	8.0%	0.0%	3.8%	34.7%	0.0%	$0.43	43.0m

July 22, 2026

Carnegie Mellon Researchers Lead Three DOE Genesis Mission Awards to Advance the Future of AI-Enabled Scientific Discovery

Working across national labs and leading universities, these teams will develop AI capabilities to address critical energy, environmental and nuclear challenges, strengthening America's security and prosperity.

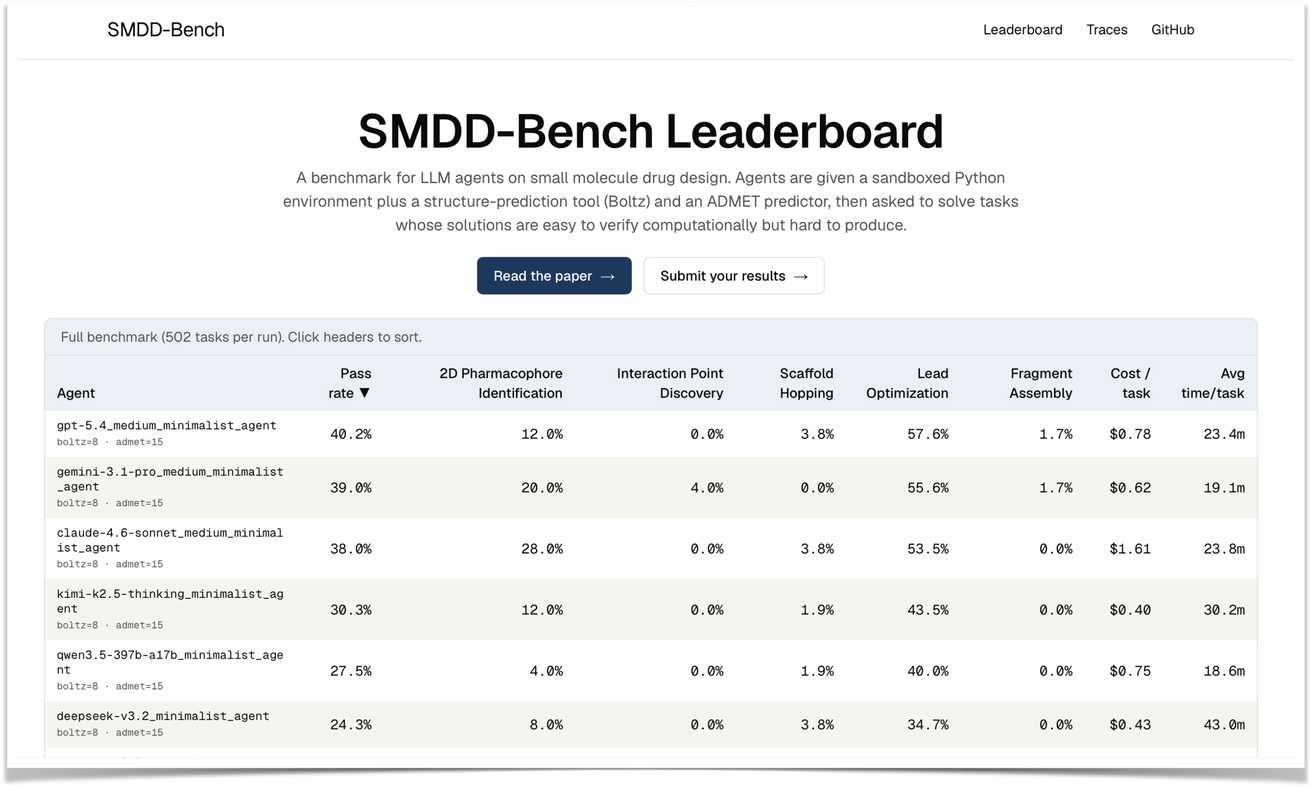
CMU collaborators: Danith Ly, Jose Lugo-Martinez, Aviral Kumar, Max Simchowitz

External: Yihui Ren (BNL), Manasa Kaniselvan (MIT)

Open problems we work on

2. AI for science, chemistry & molecule design

1. Model or harness, long horizon and experience learning.
2. Memorization and exploration.
3. Dual use and safety.
4. Vision for 3D and structural reasoning.



SMDD-Bench

Leaderboard Traces GitHub

SMDD-Bench Leaderboard

A benchmark for LLM agents on small molecule drug design. Agents are given a sandboxed Python environment plus a structure-prediction tool (Boltz) and an ADMET predictor, then asked to solve tasks whose solutions are easy to verify computationally but hard to produce.

Read the paper → Submit your results →

Full benchmark (502 tasks per run). Click headers to sort.

Agent	Pass rate ▼	2D Pharmacophore Identification	Interaction Point Discovery	Scaffold Hopping	Lead Optimization	Fragment Assembly	Cost / task	Avg time/task
gpt-5.4_medium_minimalist_agent boltz=8 · admet=15	40.2%	12.0%	0.0%	3.8%	57.6%	1.7%	$0.78	23.4m
gemini-3.1-pro_medium_minimalist_agent boltz=8 · admet=15	39.0%	20.0%	4.0%	0.0%	55.6%	1.7%	$0.62	19.1m
claude-4.6-sonnet_medium_minimalist_agent boltz=8 · admet=15	38.0%	28.0%	0.0%	3.8%	53.5%	0.0%	$1.61	23.8m
kimi-k2.5-thinking_minimalist_agent boltz=8 · admet=15	30.3%	12.0%	0.0%	1.9%	43.5%	0.0%	$0.40	30.2m
qwen3.5-397b-a17b_minimalist_agent boltz=8 · admet=15	27.5%	4.0%	0.0%	1.9%	40.0%	0.0%	$0.75	18.6m
deepseek-v3.2_minimalist_agent boltz=8 · admet=15	24.3%	8.0%	0.0%	3.8%	34.7%	0.0%	$0.43	43.0m

July 22, 2026

Carnegie Mellon Researchers Lead Three DOE Genesis Mission Awards to Advance the Future of AI-Enabled Scientific Discovery

Working across national labs and leading universities, these teams will develop AI capabilities to address critical energy, environmental and nuclear challenges, strengthening America's security and prosperity.

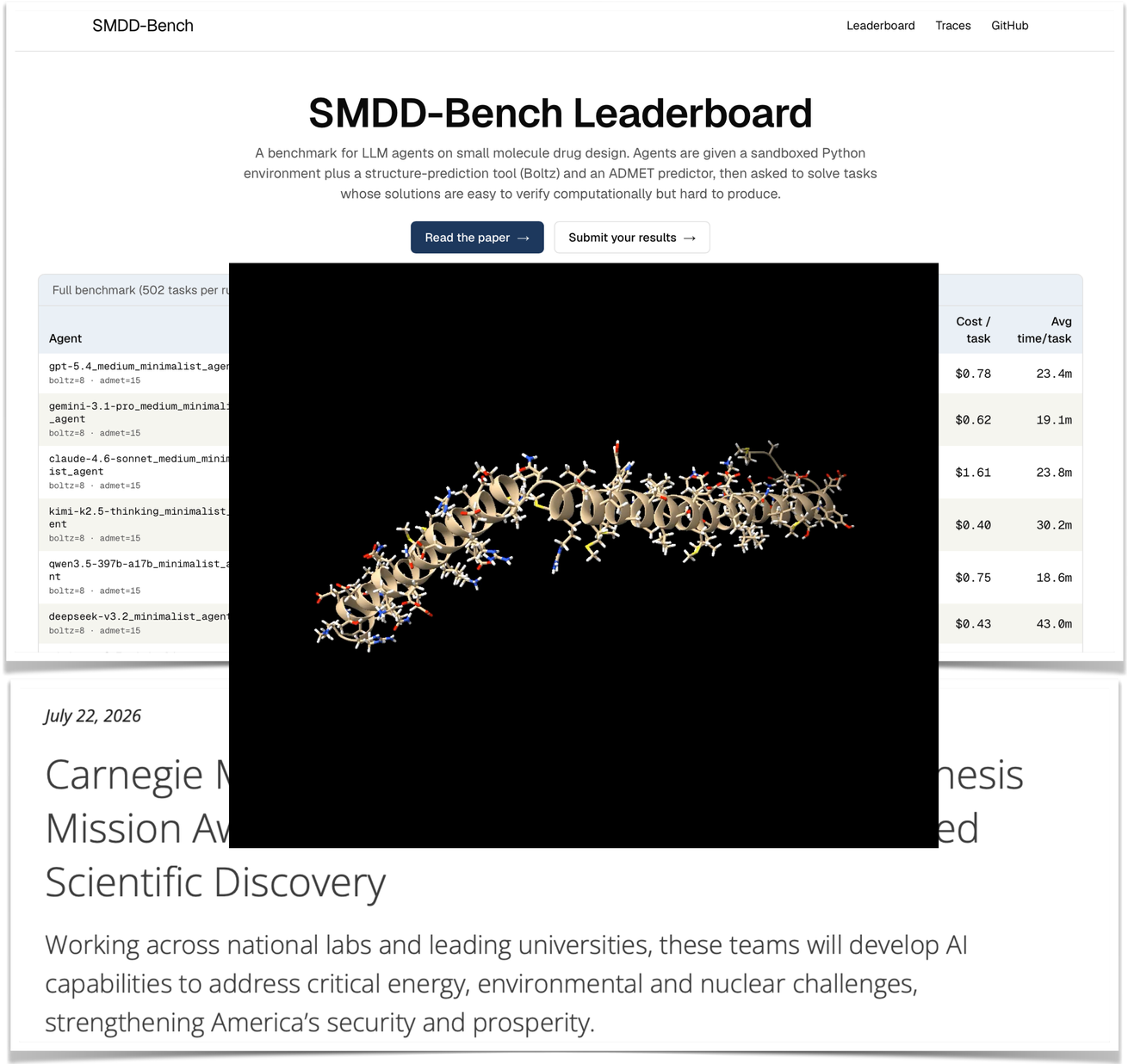
CMU collaborators: Danith Ly, Jose Lugo-Martinez, Aviral Kumar, Max Simchowitz

External: Yihui Ren (BNL), Manasa Kaniselvan (MIT)

Open problems we work on

2. AI for science, chemistry & molecule design

1. Model or harness, long horizon and experience learning.
2. Memorization and exploration.
3. Dual use and safety.
4. Vision for 3D and structural reasoning.



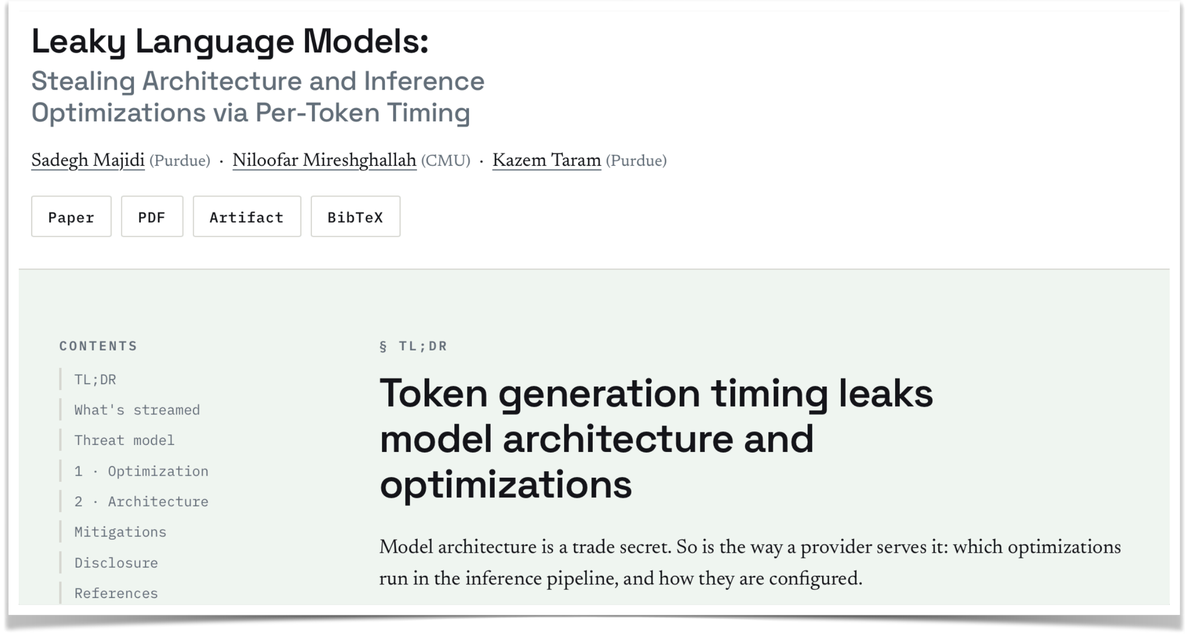
CMU collaborators: Danith Ly, Jose Lugo-Martinez, Aviral Kumar, Max Simchowitz

External: Yihui Ren (BNL), Manasa Kaniselvan (MIT)

Open problems we work on

3. AI security, verifiability, and monitoring

1. Verifying workload from hardware side channels.
2. Internal commitment points in the model.



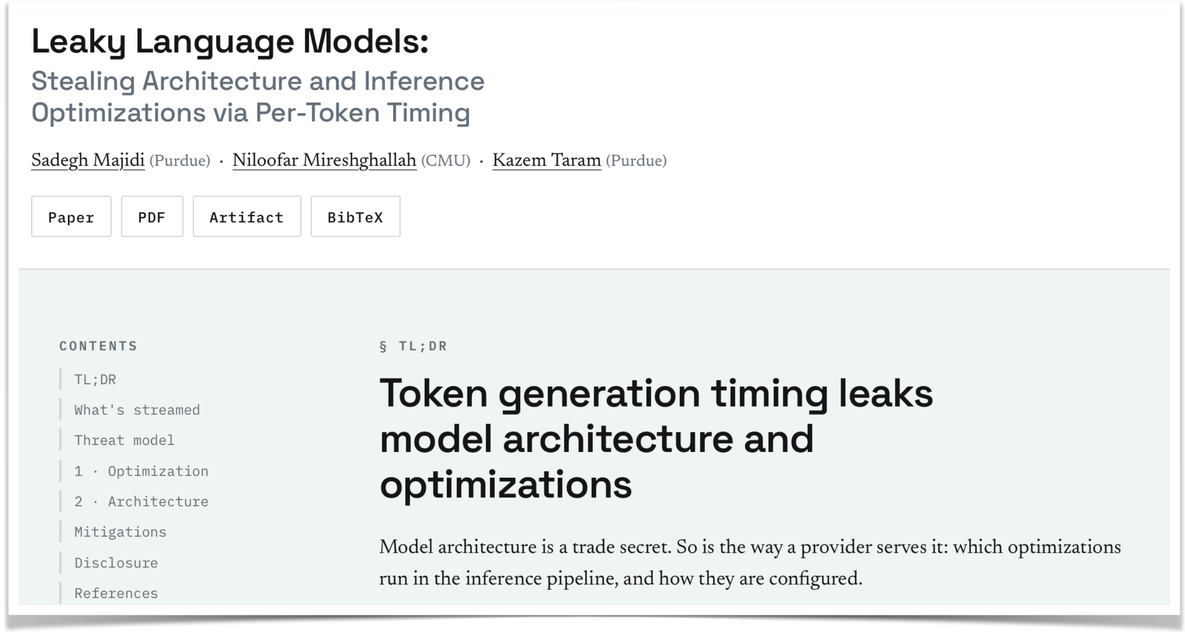
CMU collaborators: Sarah Scheffler, Lujo Bauer

External: Kazem Taram (Purdue), Fazl Barez (Oxford), Irene Solaiman (HF)

Open problems we work on

3. AI security, verifiability, and monitoring

1. Verifying workload from hardware side channels.
2. Internal commitment points in the model.
3. Distillation and lineage provenance.
4. Poisoned reasoning traces in the wild and transfer



CMU collaborators: Sarah Scheffler, Lujo Bauer

External: Kazem Taram (Purdue), Fazl Barez (Oxford), Irene Solaiman (HF)

Long horizon
Out Of distribution
Non-verifiable
Intelligence

Privacy and **Social Intelligence**

AI for **Science** and **Chemistry**

Security, **Verification** and **Monitoring**

They don't appear to want to take over, they just want to dance!

https://mireshghallah.github.io/looni-lab.html