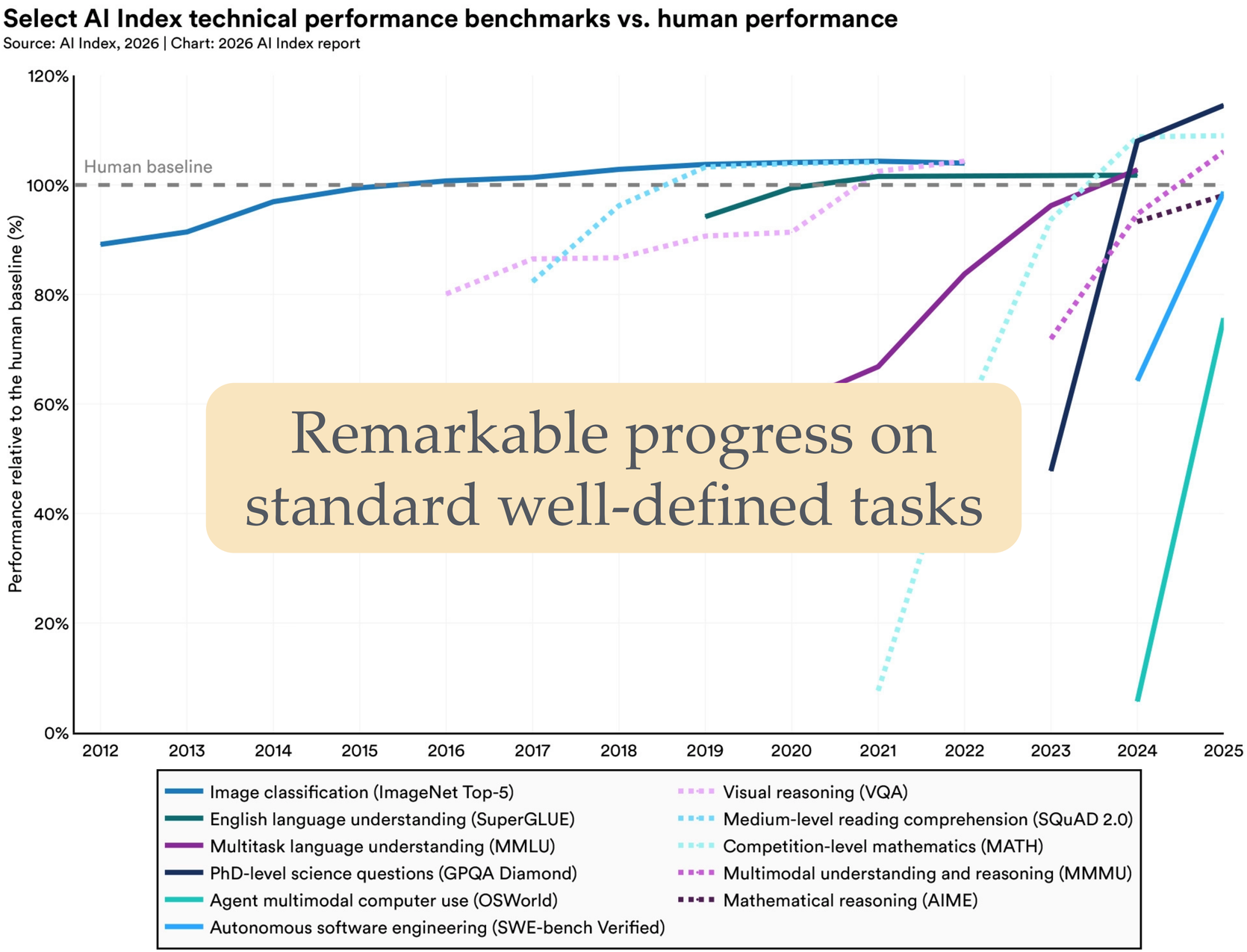
Carnegie Mellon University

What would it take to get **creative** A(V)GI?

Aditi Raghunathan

ar-forum.github.io



Remarkable progress on standard well-defined tasks

[2026-AI-index-report from Stanford]

Next frontier: open-ended tasks

Generative AI enhances individual creativity but reduces the collective diversity of novel content

ANIL R. DOSHI AND OLIVER P. HAUSER Authors Info & Affiliations

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], **Cong Lu**[3,4,*], **Robert Tjarko Lange**[1,*], **Jakob Foerster**[2,†], **Jeff Clune**[3,4,5,†] **and David Ha**[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

Artificia

Liwei Jiang
Nouha Dziri

Than Human Thought

Creativity in open-ended tasks: outputs that are diverse and original

MATHEMATICAL EXPLORATION AND DISCOVERY AT SCALE

BOGDAN GEORGIEV, JAVIER GÓMEZ-SERRANO, TERENCE TAO, AND ADAM ZSOLT WAGNER

"AlphaEvolve excels when problems can be framed as hill-climbing on a smooth score function, but struggles otherwise"

Generate a challenging graduate-level problem on convex optimization involving duality

Generate a couple ideas for a grant proposal maybe putting together some ideas from reasoning and unlearning

These are boring!

Correct, coherent but just not diverse and original

Talk outline

What is creativity?

Bottlenecks in current paradigms

At-scale results from my group

Conclusion

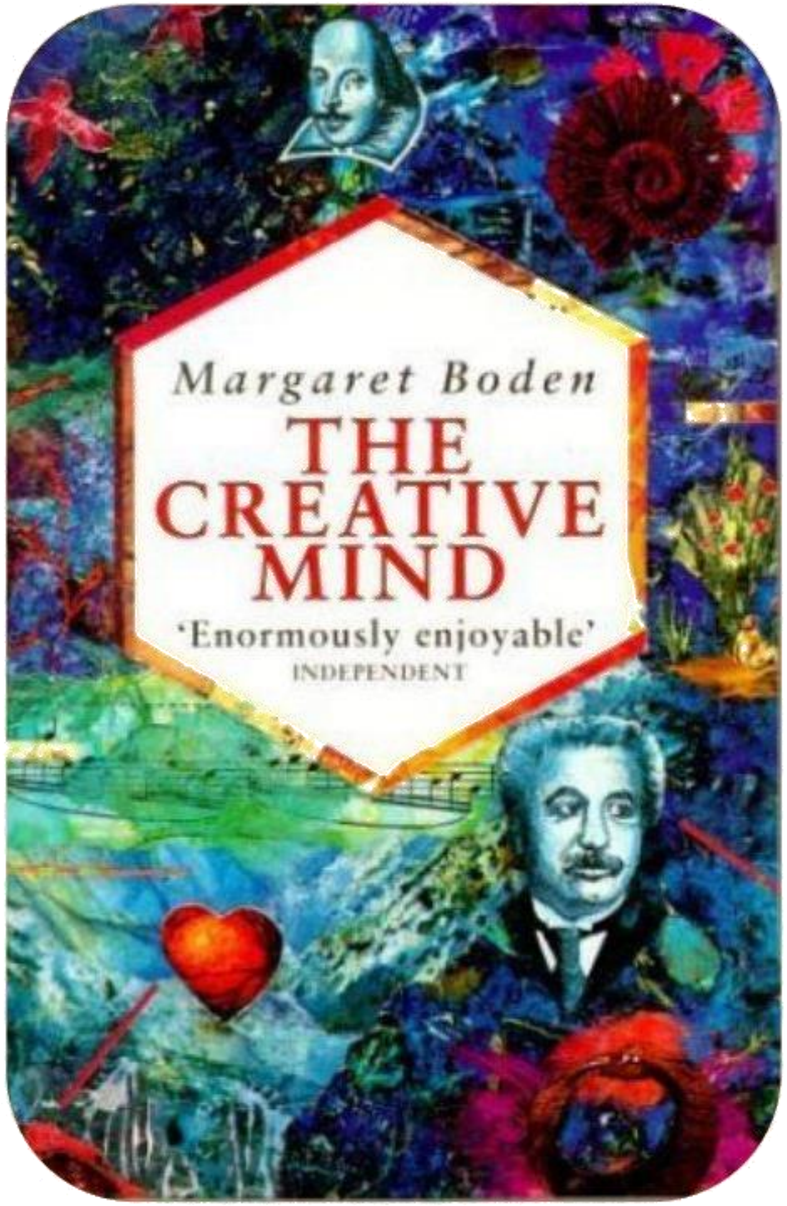
Talk outline

What is creativity?

Bottlenecks in current paradigms

At-scale results from my group

Conclusion



Creativity is the ability to come up with ideas or artefacts that are *new*, *surprising*, and *valuable*

Ideas: Concepts, poems, musical compositions, scientific theories, recipes, choreography, jokes, etc.

Artefacts: Paintings, sculptures, steam engines, vacuum cleaners, pottery, origami, penny-whistles, etc.

Lots of debate on “creativity”

Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], **Cong Lu**[3,4,*], **Robert Tjarko Lange**[1,*], **Jakob Foerster**[2,†], **Jeff Clune**[3,4,5,†] **and David Ha**[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

All That Glitters is Not Novel: Plagiarism in AI Generated Research

Tarun Gupta
Indian Institute of Science
Bengaluru, KA, India
tarungupta@iisc.ac.in

Danish Pruthi
Indian Institute of Science
Bengaluru, KA, India
danishp@iisc.ac.in

Evaluating Sakana’s AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards ‘Artificial Research Intelligence’ (ARI)?

JOERAN BEEL, University of Siegen, Intelligent Systems Group & Recommender-Systems.com, Germany
MIN-YEN KAN, National University of Singapore – Web, Information Retrieval / Natural Language Processing Group (WING), Singapore
MORITZ BAUMGART, University of Siegen, Germany

The Ideation–Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas

Chenglei Si, Tatsunori Hashimoto, Diyi Yang
Stanford University
{clsi, thashim,diyiy}@stanford.edu

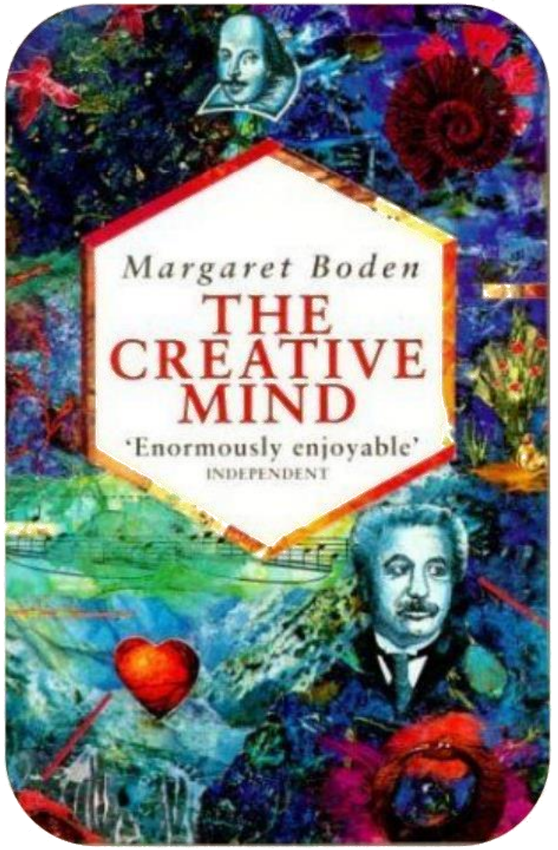
ar-forum.github.io

What we do:

We draw inspiration from two modes of creativity in cognitive science

and design minimal, *open-ended* algorithmic tasks

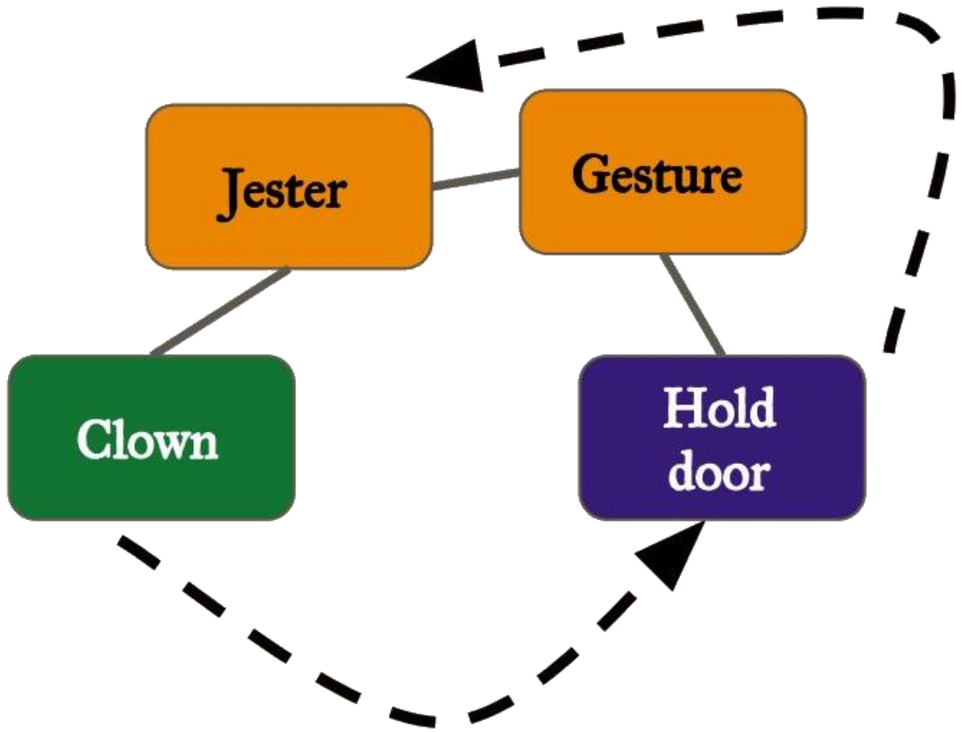
Where we can quantify creative limits of LLMs & *highlight alternatives*



Consider wordplay

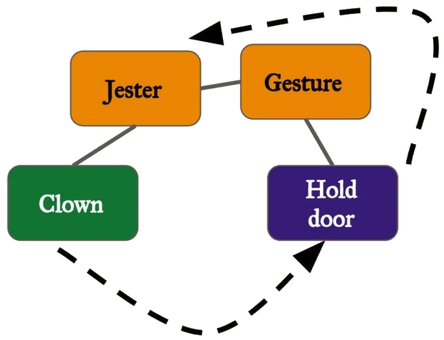
A **clown held the door** open for me.

What a nice jester!



"Creativity is the power to connect the seemingly unconnected" - William Plomer

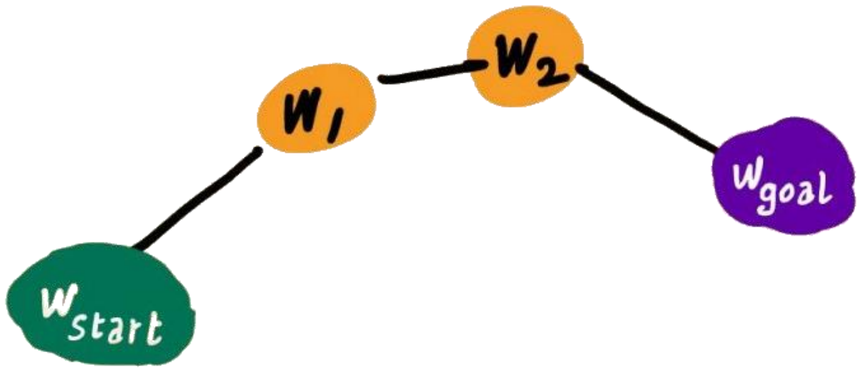
Consider wordplay



Wordplay: find a **novel path** over a **known** vocabulary graph

generate :

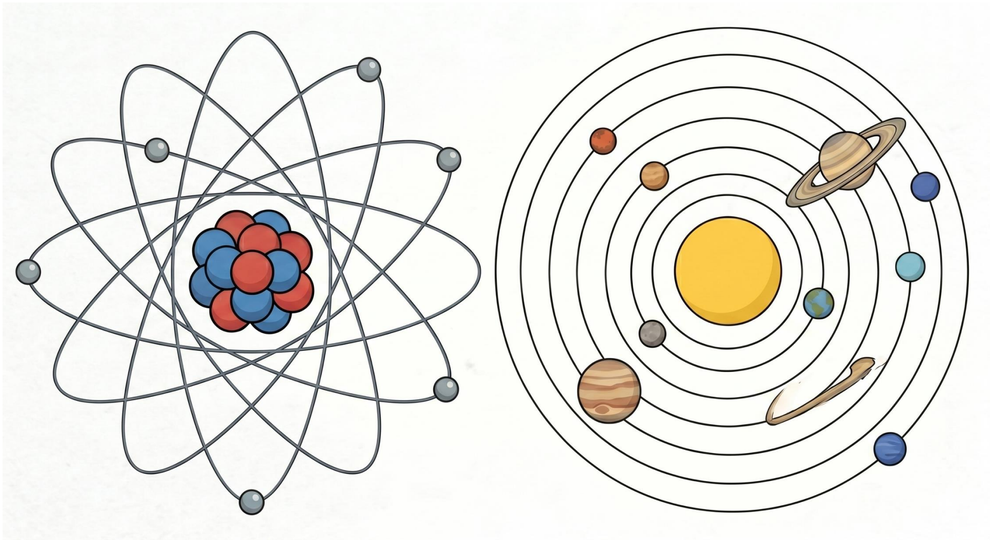
s.t.



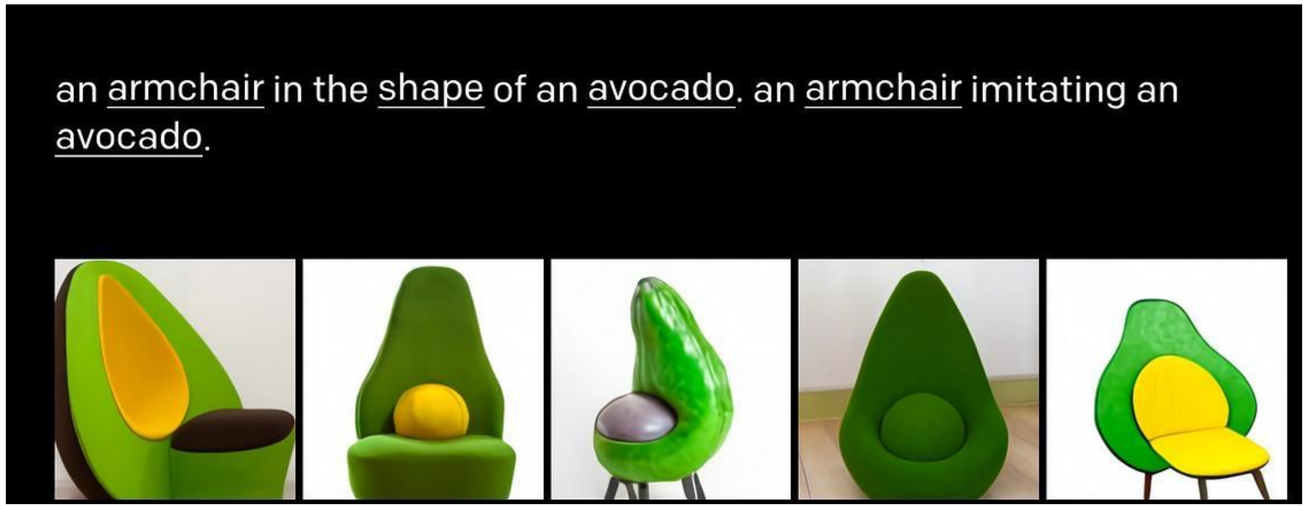
Boden terms this combinational creativity

Combinational creativity

Unfamiliar combinations of familiar ideas



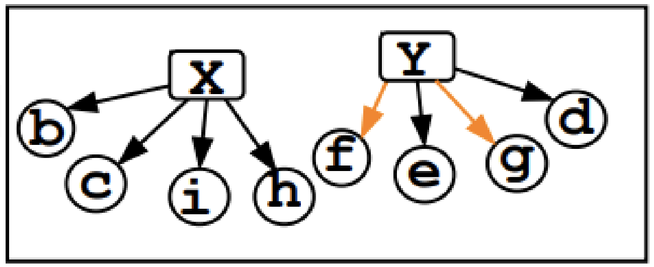
Human creativity in the prompt!



Sibling discovery task

`Generate: "g, f, Y"`

`s.t.`



(in-weights graph)

Bipartite graph $\mathcal{G}$: parents $\mathcal{V}$, children $\text{nbr}(\cdot)$

$$m = |\mathcal{V}| \qquad n = |\text{nbr}(\cdot)|$$

Training: valid triplets $S = \{s_1, \ldots s_m\}$

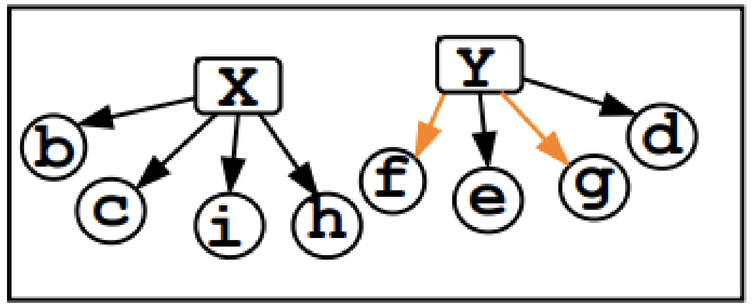
Generate new & valid sibling-parent triplets from the implicit graph

$$\underbrace{\Omega(m \cdot n)}_{\text{infer all edges}} \leq |S| \leq \underbrace{o(m \cdot n^2)}_{\text{not all triplets}}$$

Combinational creativity tasks

Generate: **"g, f, Y"**

s.t.

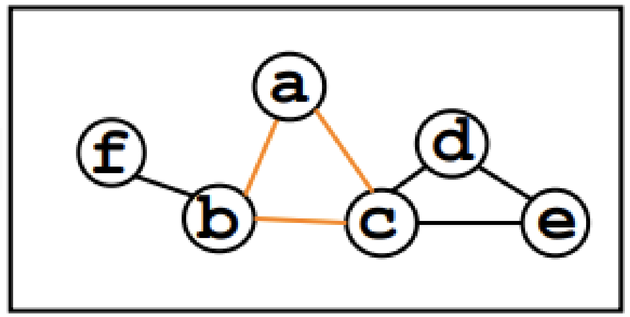


(in-weights graph)

(a) Sibling Discovery

Generate: **"a, b, c"**

s.t.



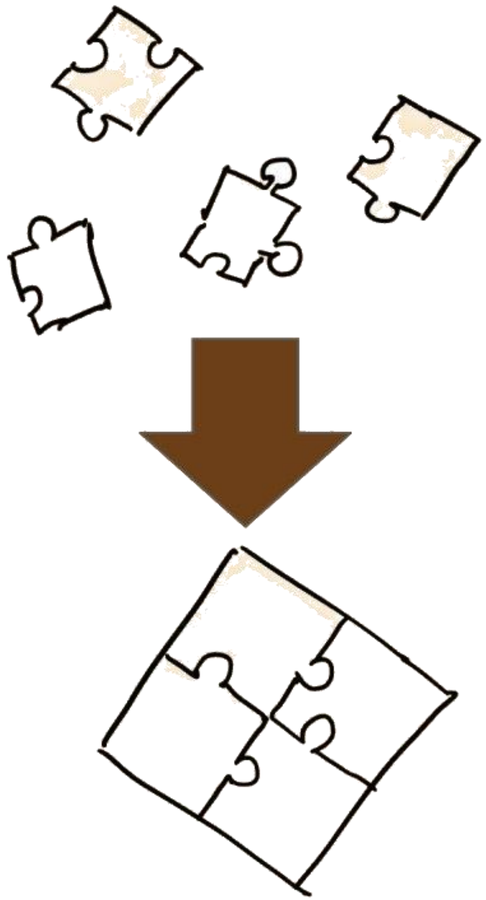
(in-weights graph)

(b) Triangle Discovery

Exploratory creativity

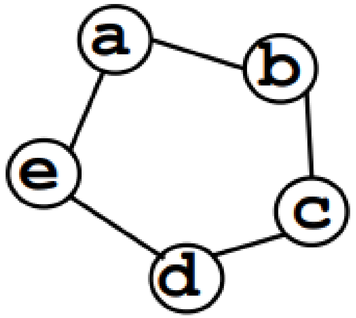
Plan and devise novel patterns that obey a small set of rules

designing problems, deriving corollaries, generating molecules, crafting stories, new caricatures, limericks, sonnets etc.



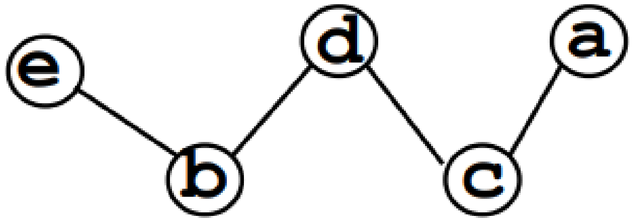
Exploratory creativity tasks

Generate:
"a→b, c→d, d→e, b→c, e→a"
s.t.



(a) Circle Construction

Generate:
"c→a, b→d, d→c, e→b"
s.t.



(b) Line Construction

Eval metrics

$$S : s_1, s_2, \ldots, s_m \longrightarrow \mathrm{LM}_\theta \longrightarrow T : s'_1, s'_2, \ldots, s'_N$$

Creativity is the fraction of generations that are

✓ **unique** ✓ **unseen** ✓ **coherent**

$$\widehat{\mathrm{cr}}(T) = \frac{\mathrm{uniq}\left(\{s \in T \mid \mathrm{coh}(s) \wedge s \notin S\}\right)}{|T|}$$

Talk outline

What is creativity?

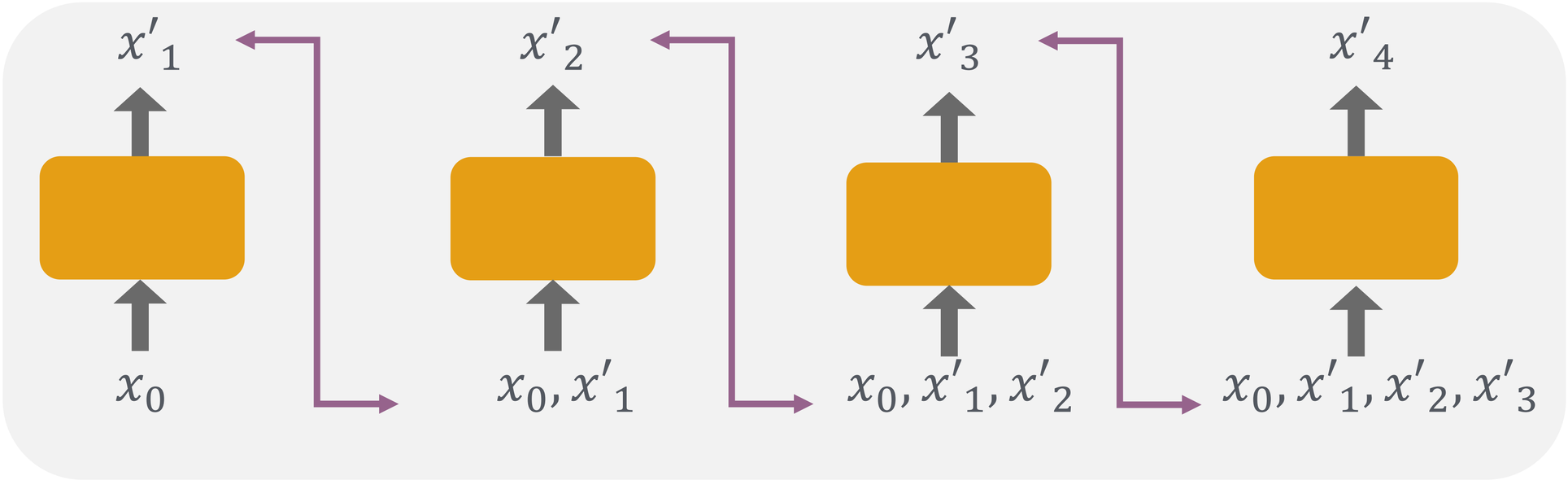
Bottlenecks in current paradigms

At-scale results from my group

Conclusion

Next-token **prediction**

$$x'_{t+1} \sim p_\theta(\cdot \mid x'_1, x'_2, \ldots, x'_t)$$

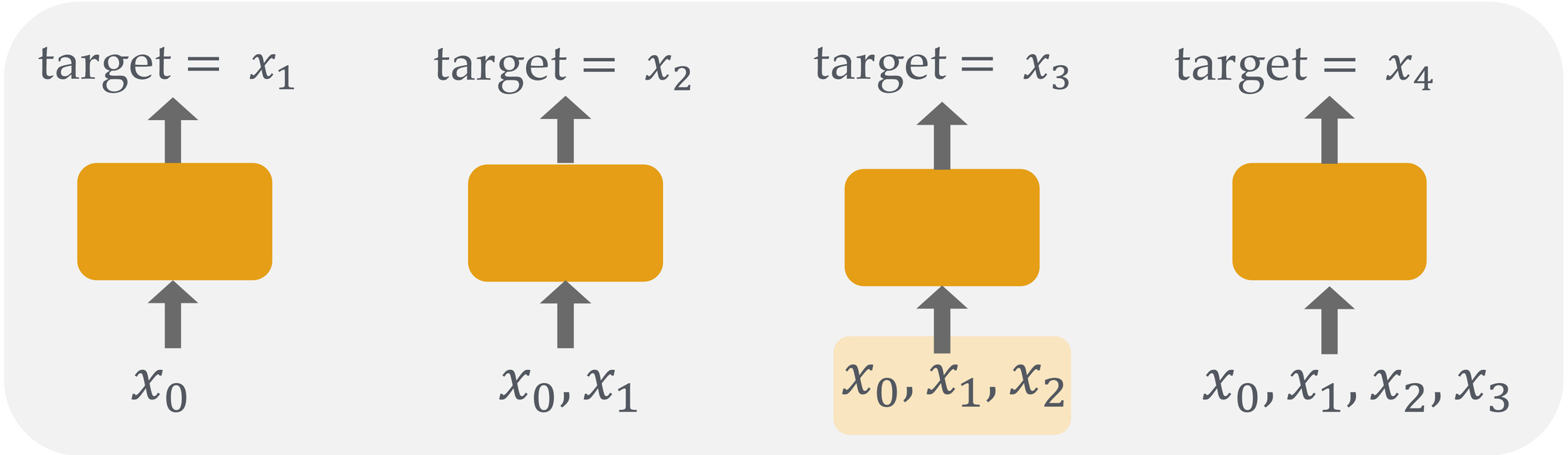


$$p_\theta(x'_1, \ldots, x'_L) = \prod_{t=1}^{L} p_\theta(x'_t \mid x'_{<t})$$

Autoregressive NTP modeling can represent any sequence

Next-token learning

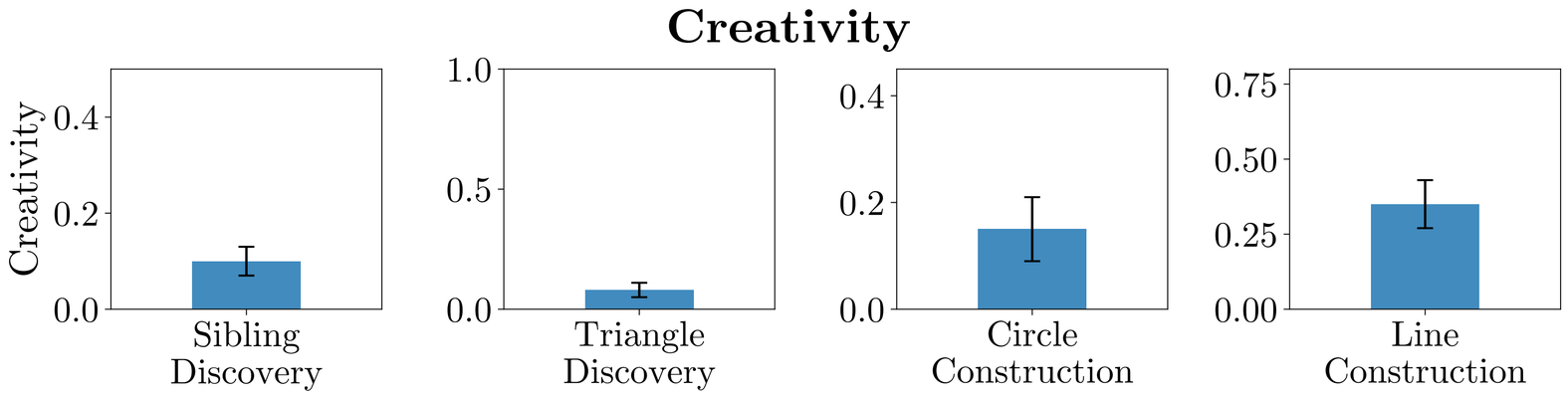
Training sequence $s = (x_0, x_1, \ldots)$



Teacher forcing

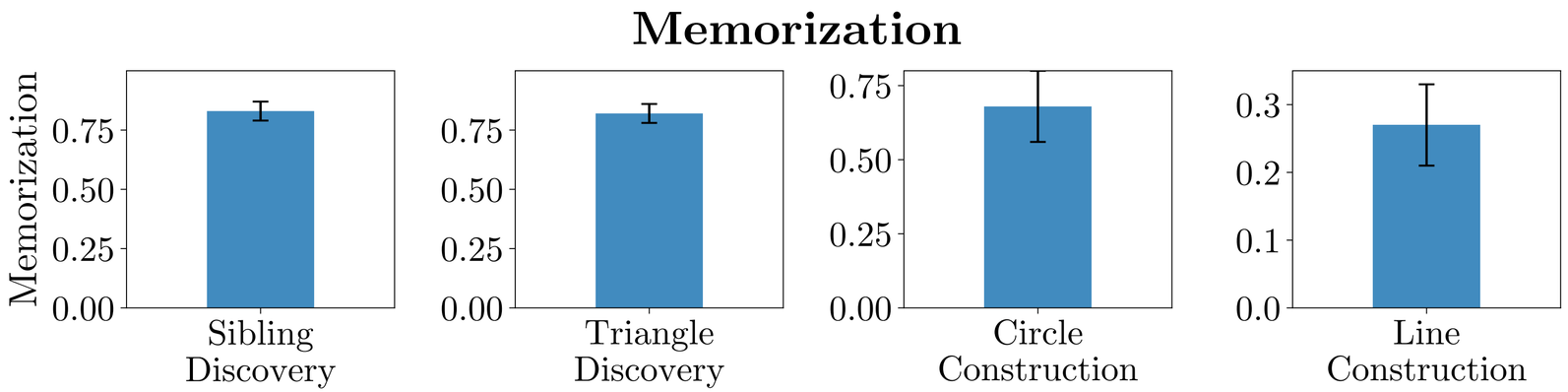
Learning to predict x_3 given ground truth x_0, x_1, x_2

Results with next-token learning

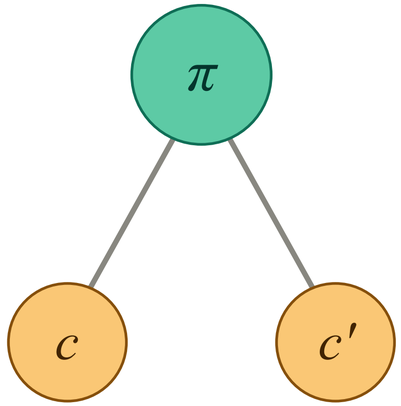


Low creativity

High memorization



Why does next-token learning fail?



Ideal factorization: $\underbrace{p(z := \pi)}_{\text{pick a parent}} \cdot \underbrace{p(c \mid z)}_{\text{sample child}} \cdot \underbrace{p(c' \mid z)}_{\text{sample child}}$

NT supervision

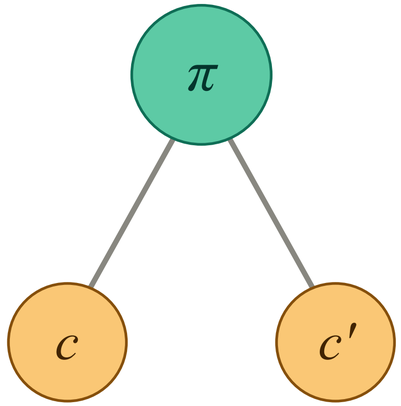
Token 1	Token 2	Token 3
$p(c)$	$p(c' \mid c)$	$p(\pi \mid c, c')$

Trivial shortcut: *Model learns this instantly*

Output mutual neighbor (lookup; no planning)

[Bachmann and Nagarajan. 2024]

Why does next-token learning fail?



Ideal factorization:

$$\underbrace{p(z := \pi)}_{\text{pick a parent}} \cdot \underbrace{p(c \mid z)}_{\text{sample child}} \cdot \underbrace{p(c' \mid z)}_{\text{sample child}}$$

NT supervision

Token 1 $p(c)$

Token 2 $p(c' \mid c)$

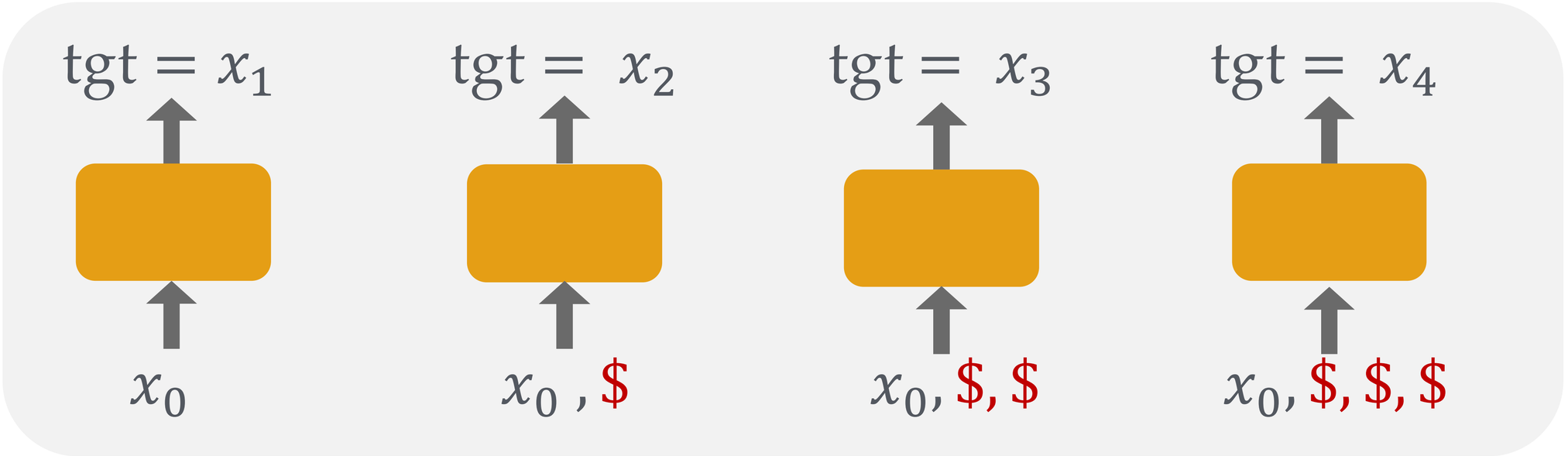
Token 3 $p(\pi \mid c, c')$

Stuck with learning this hard distribution

✗ Model ends up memorizing

[Bachmann and Nagarajan. 2024]

What if we remove the shortcut?



$$J_{\mathrm{MTP}}(\theta) = \mathbb{E}_{s\sim\mathcal{D}}\left[\sum_{i=1}^{L} \log \mathrm{LM}_\theta(s_i \,;\, \$_{<i})\right]$$

Teacherless training

[Bachmann and Nagarajan. 2024]

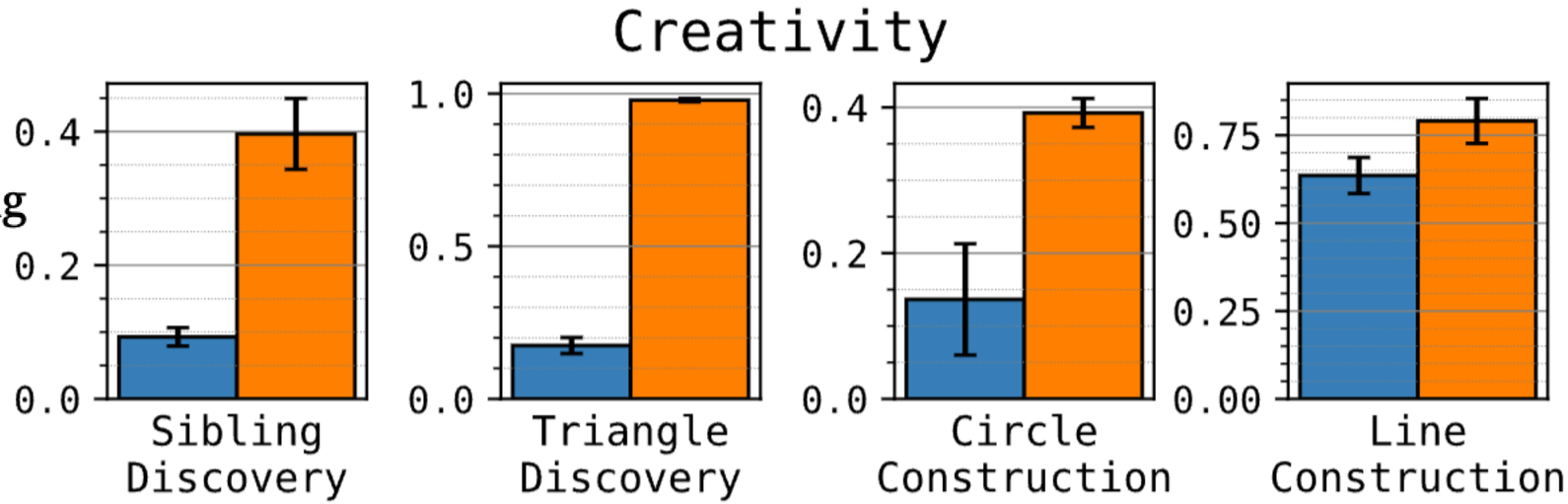
Teacherless training results

$$J_{\text{hybrid}}(\theta) = \alpha \cdot J_{\text{NTP}}(\theta) + (1-\alpha) \cdot J_{\text{MTP}}(\theta)$$

Standard next-token learning

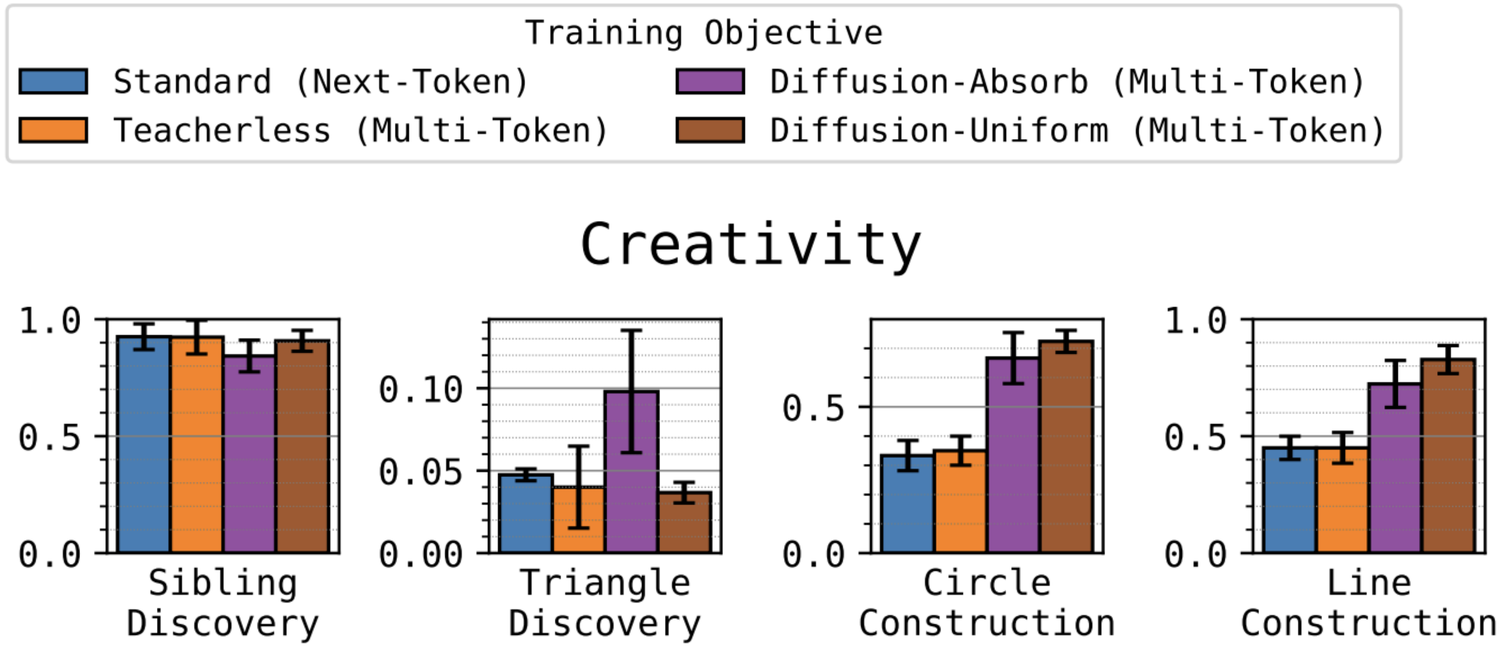
Teacherless multi-token

* We tune sampling hyperparams



Teacherless training is **more creative** than next-token learning

What about diffusion models?

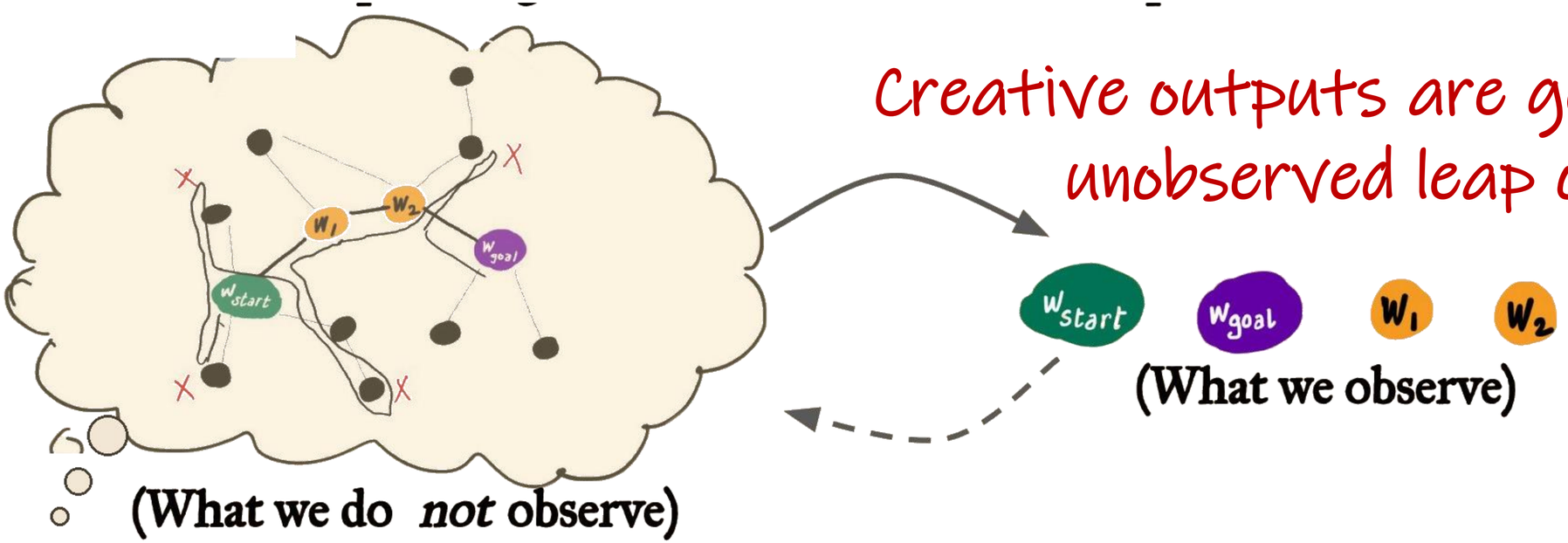


Diffusion models are **more creative** than next-token learning

#1

Paradigm shift: beyond next-token learning

Roll the dice & look before you leap. NWD**R**. ICML 2025. **Outstanding Paper Award**



✓ Teacherless multi-token training improves creativity

✓ **Diffusion** language models are multi-token learning approaches and similarly boost creativity

ar-forum.github.io

The **mode collapse** issue

Teacher less training
prevents memorization

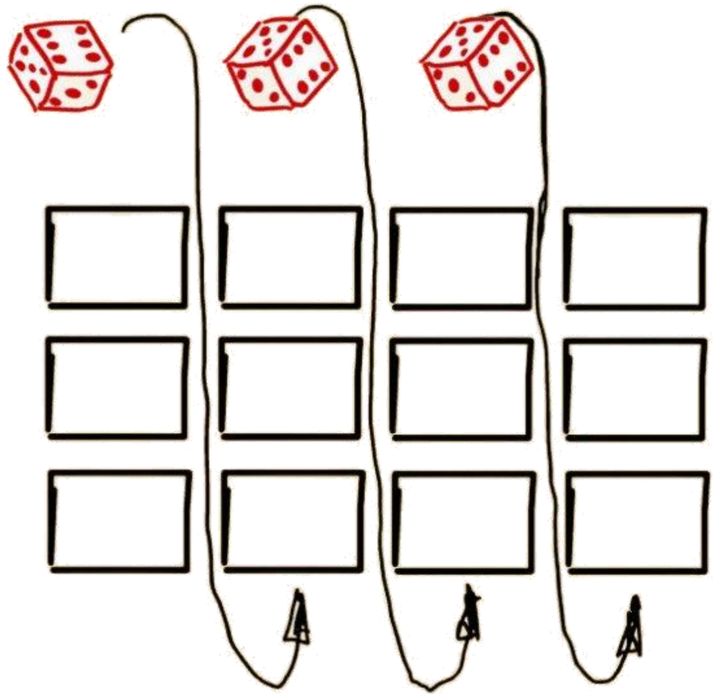
$$\widehat{\mathrm{cr}}(T) = \frac{\mathrm{uniq}(\{s \in T \mid \mathrm{coh}(s) \wedge s \notin S\})}{|T|}$$

De-facto knob is temperature

Temperature τ ↑ ✓ **diversity** ↑ ✗ **coherence** ↓

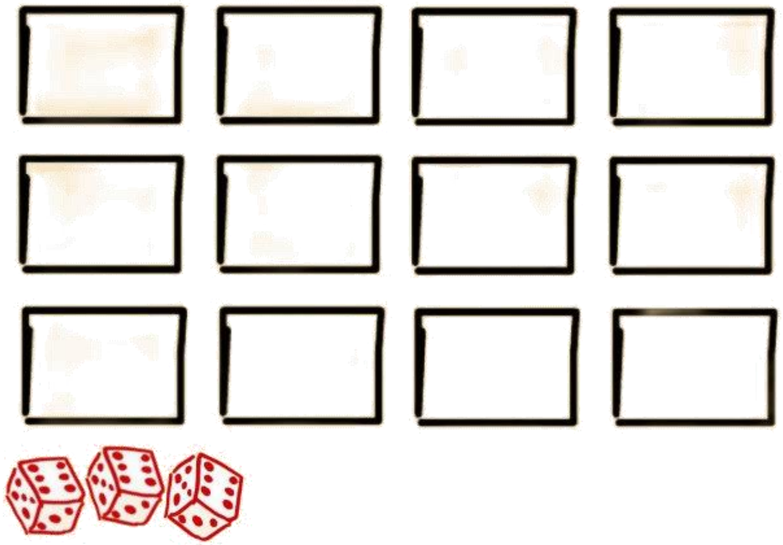
[See also Jahrens and Martinetz. 2025]

Input vs output randomness



Output conditioning

Fleshing out many thoughts



Seed conditioning

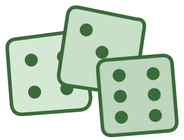
Fleshing out one thought at a time

Bring randomness to the input

Standard $x_0 \longrightarrow x_1, x_2, x_3$

Seed-conditioned $x_0, \zeta_1, \zeta_2, \ldots, \zeta_k \longrightarrow x_1, x_2, x_3$

Prefix random string per example during *both* training and inference

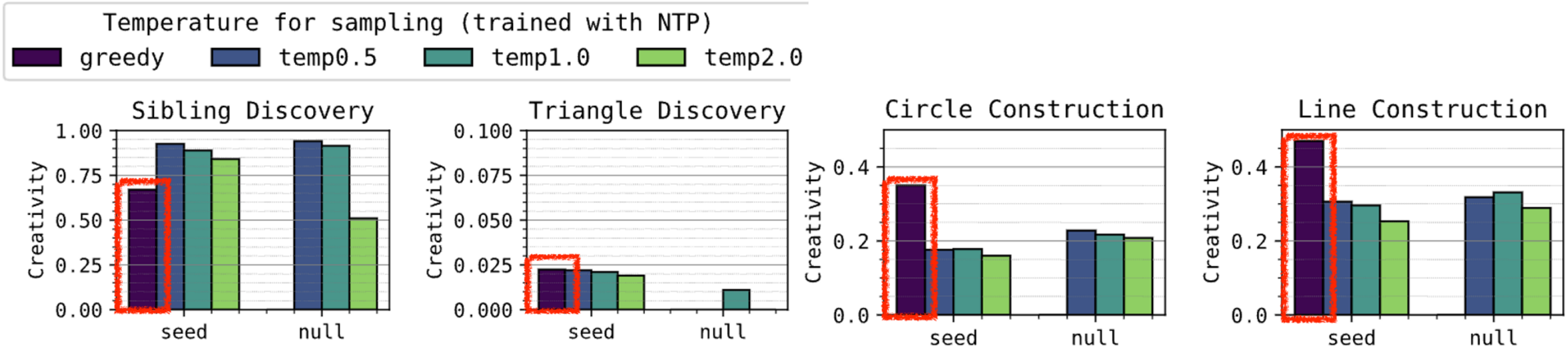


Different seeds can lead to different outputs

No drop in coherence as model can focus on one thought at a time…

Naïve seed-conditioning

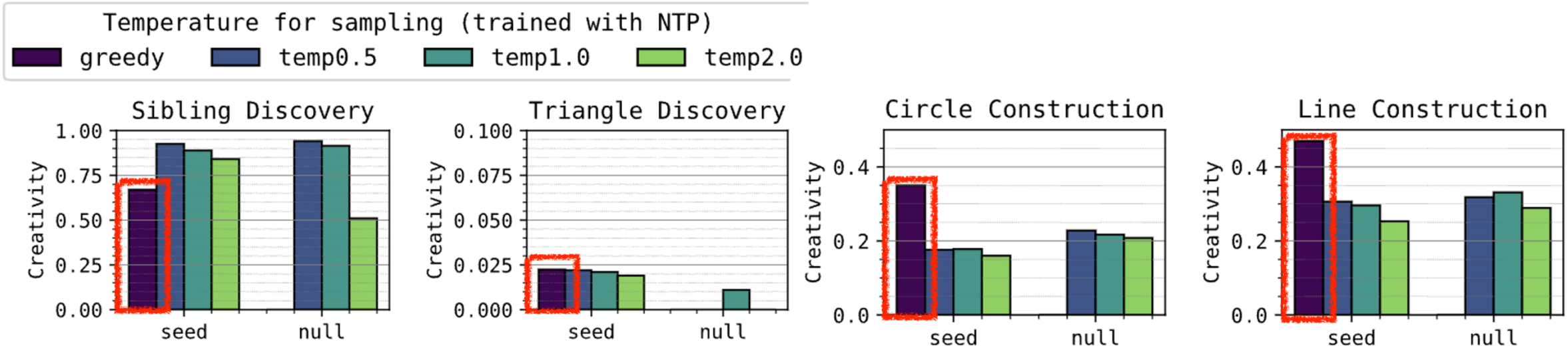
Prepend a completely random string to each training example



Seed-conditioning with zero temperature (*greedy*) is comparable to temperature sampling in creativity

Naïve seed-conditioning

Prepend a completely random string to each training example



Seed-conditioning can sometimes be the most creative method

#2

Paradigm shift: diversity beyond temperature sampling

Back to the wordplay example..

Word play on twitter >> GPT-generations

Talk outline

What is creativity?

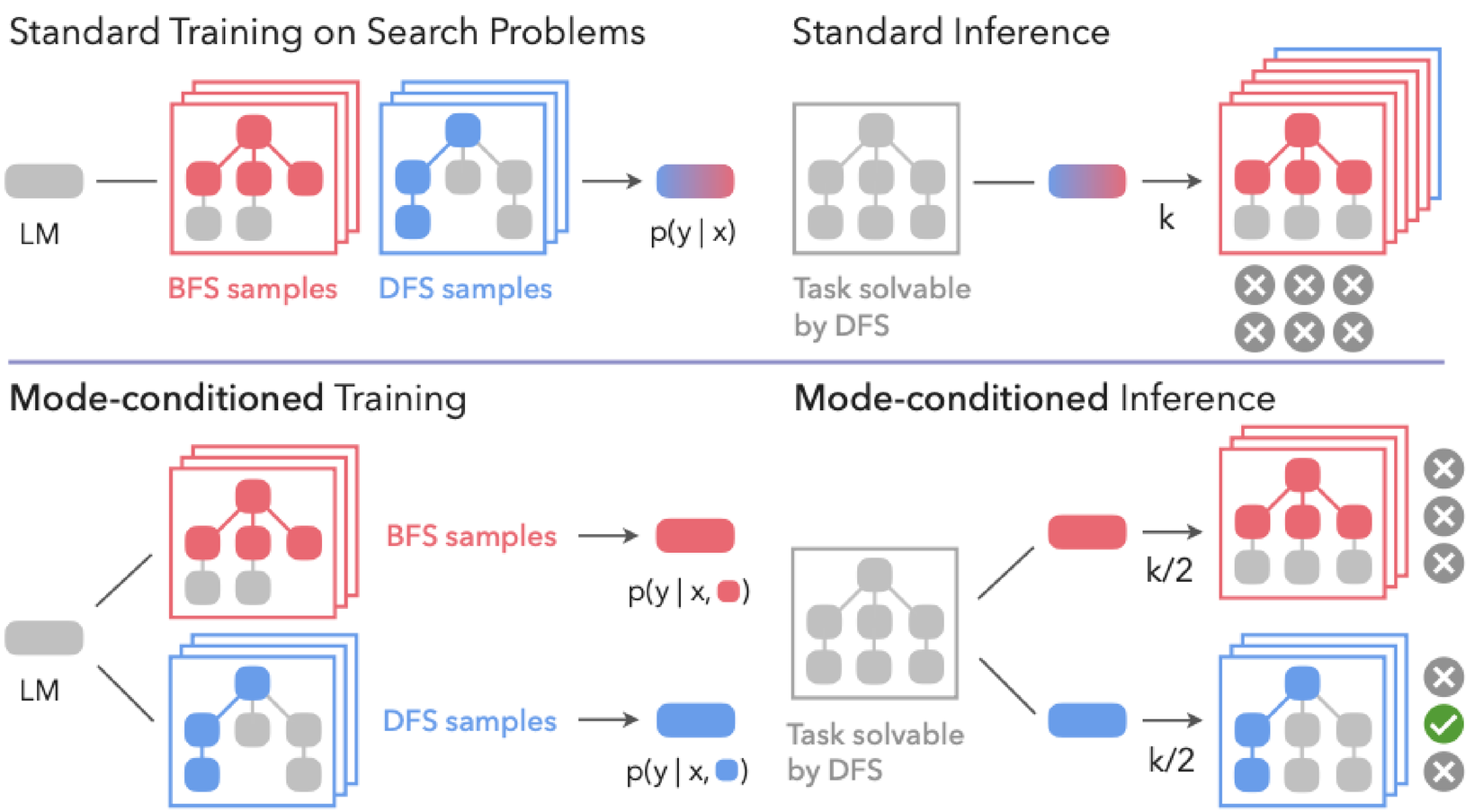
Bottlenecks in current paradigms

At-scale results from my group

Conclusion

Sample k traces independently & succeed if any one is correct

Standard training wastes samples due to low diversity



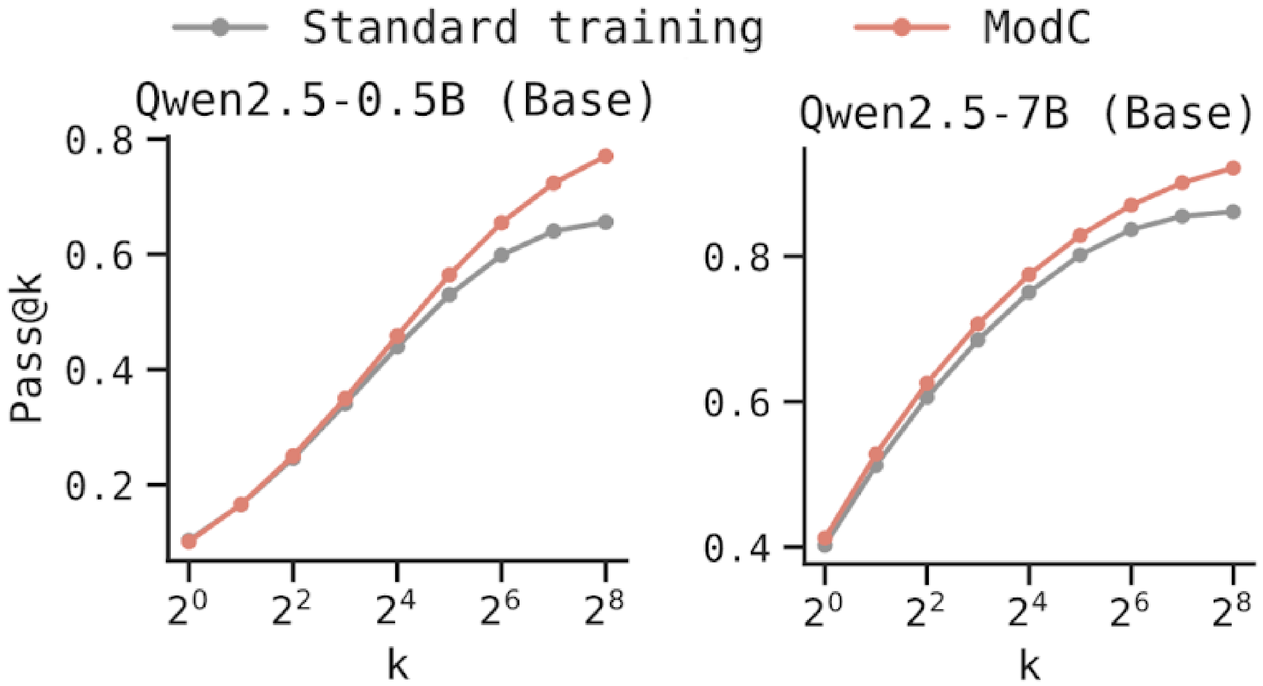
Extracting **modes** in real data

Training data → **gradient features → cluster** → mode labels
→ ModC training → balanced inference

Train dataset: NuminaMath
Eval dataset: MATH500

We get 4-8x inference-sample efficiency gains

* No aux information

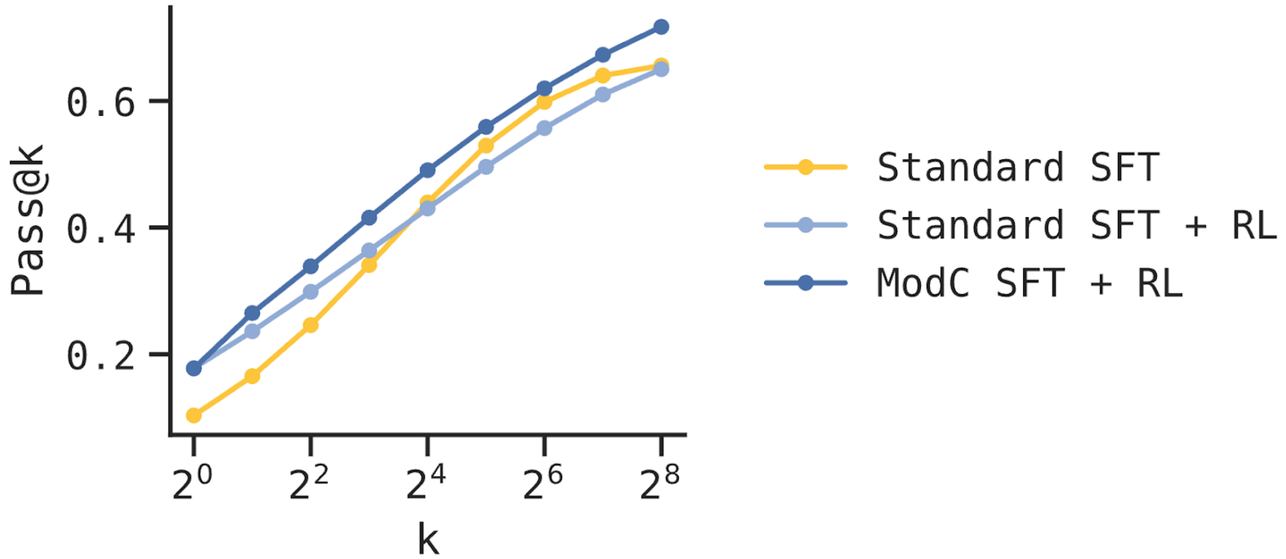


Extracting **modes** in real data

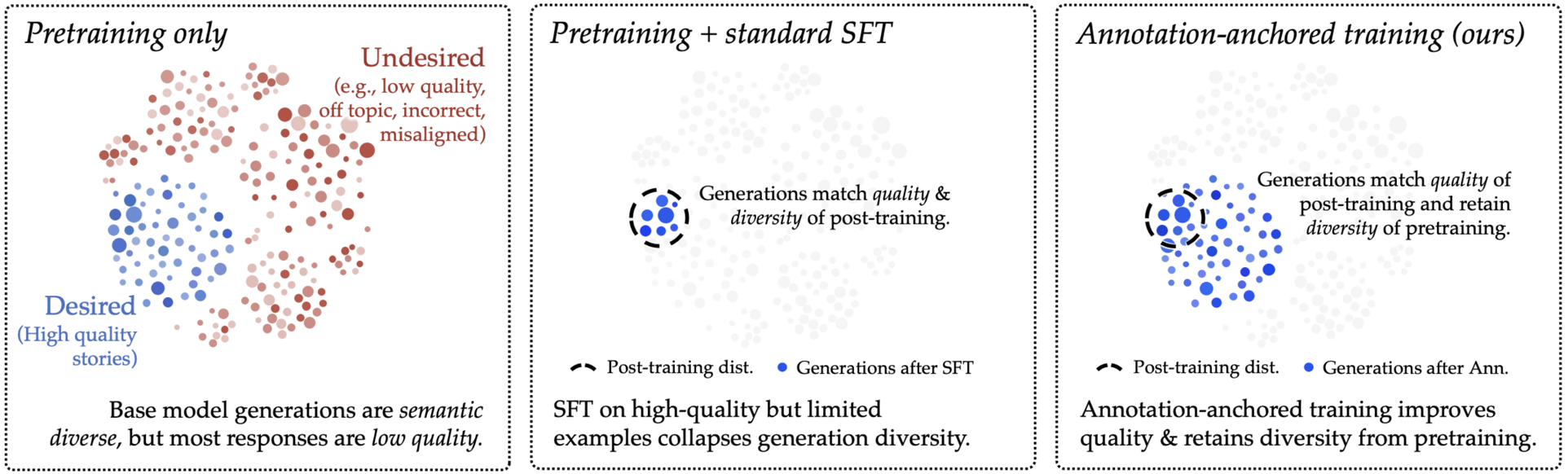
Training data → **gradient features → cluster** → mode labels
→ ModC training → balanced inference

Better starting point for RL

* Beats baseline SFT even at large k unlike standard approaches



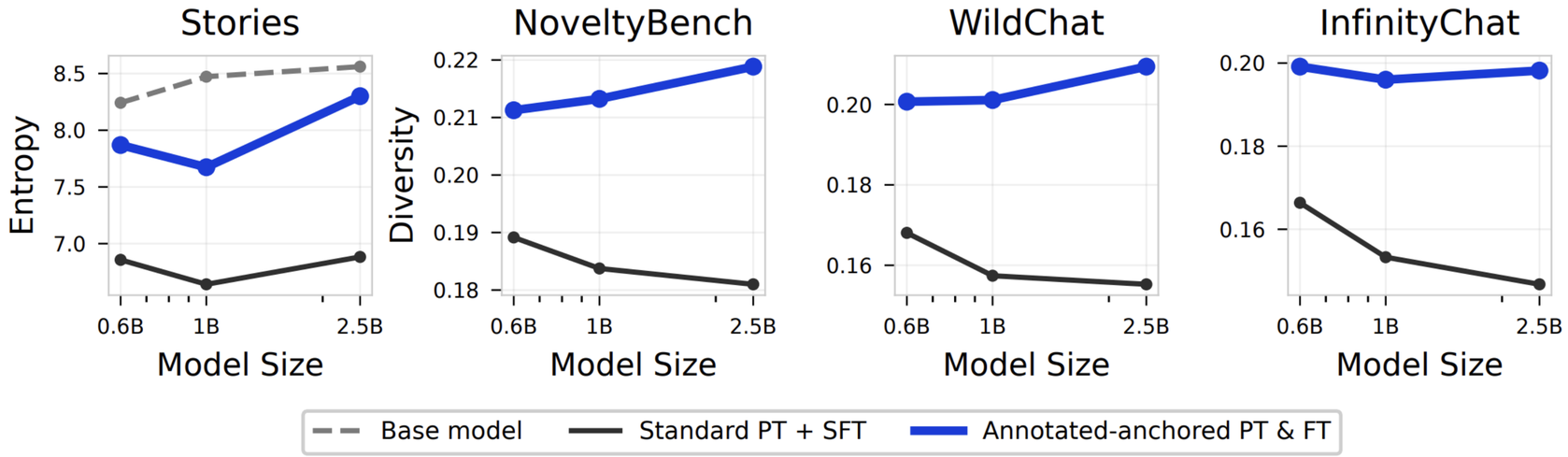
Annotations mitigate post-training collapse. ICML 2026 (in collaboration with Apple)



1 **Pre-training:** add semantic annotations.

2 **Post-training:** train on masked annotations.

3 **Inference:** first generate annotations, then the response.



Talk outline

Setup of creative tasks

Bottlenecks in current paradigms

At-scale results from my group

Conclusion

Summary: open-ended creativity

Identified and **addressed** two orthogonal failures of current paradigms for creative tasks

- Next-token learning vs multi-token learning
- Seed-conditioning as an alternative to temperature sampling

Remarks and future work

- Benchmarks for visual creativity
- What is the right (combination of) learning objectives for visual creativity?
- Should we anchor beyond language?
- Boden's **transformational creativity**

  Change the rules of the game entirely!

  - Should we make AI more creative?

Thanks!

Vaishnavh Nagarajan

Chen Wu

Charles Ding

Sachin Goyal

Talk outline

Overview of

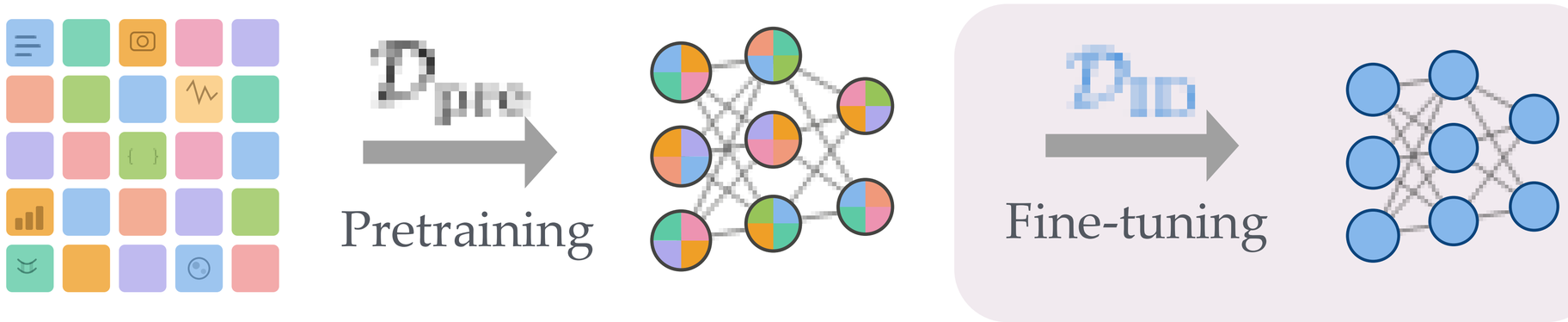
Evaluation in the wild

Reliability from the ground up

Creativity in open-ended tasks

Natural distribution shifts

Pretraining data is large-scale and diverse



Regularize fine-tuning appropriately to preserve relevant pretrained knowledge

Robust fine-tuning

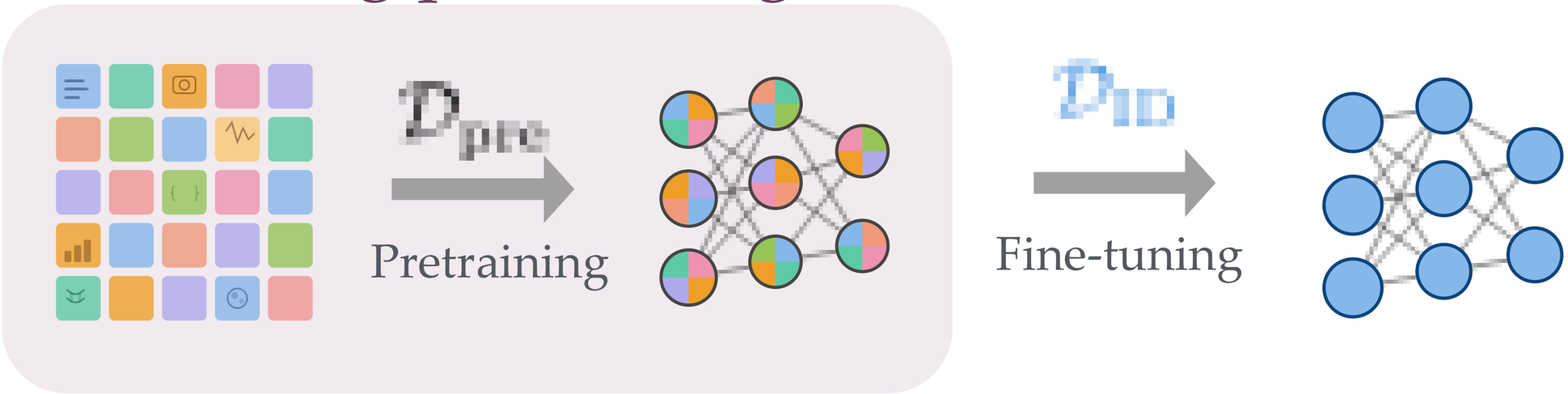
Regularize fine-tuning appropriately to preserve relevant pretrained knowledge

Theoretically principled approaches to achieve **state-of-the-art** in

- ✓ Robustness to distribution shifts [K**R**JML *ICLR 2022*, GKGK**R** *CVPR 2023*]
- ✓ LLM embeddings [SKFN**R** *ICLR 2025*]
- ✓ Reasoning with LLMs [DBWK**R** *COLM 2025*]

In this talk

Rethinking pretraining



to disentangle memorization and generalization

Can AI be creative?

Lots of critical & pioneering work answering this!

Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

All That Glitters is Not Novel: Plagiarism in AI Generated Research

Tarun Gupta
Indian Institute of Science
Bengaluru, KA, India
tarungupta@iisc.ac.in

Danish Pruthi
Indian Institute of Science
Bengaluru, KA, India
danishp@iisc.ac.in

Evaluating Sakana's AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards 'Artificial Research Intelligence' (ARI)?

JOERAN BEEL, University of Siegen, Intelligent Systems Group & Recommender-Systems.com, Germany
MIN-YEN KAN, National University of Singapore – Web, Information Retrieval / Natural Language Processing Group (WING), Singapore
MORITZ BAUMGART, University of Siegen, Germany

The AI Scientist: Towards Fully Automate Open-Ended Scientific Discovery

Chris Lu[1,2,*], **Cong Lu**[3,4,*], **Robert Tjarko Lange**[1,*], **Jakob Foerster**[2,†], **Jeff Clune**[3,4,5,†] **and David Ha**[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Ca
AI Chair, [†]Equal Advising

The Ideation–Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas

Chenglei Si, Tatsunori Hashimoto, Diyi Yang
Stanford University
{clsi, thashim,diyiy}@stanford.edu

ai-forum.github.io

In this talk

We draw inspiration from two modes of creativity in cognitive science

and design *minimal*, open-ended, algorithmic tasks to

where we can quantify creative limits of LLMs & highlight alternatives