Roll the dice and look before you leap:

Going beyond the creative limits of next-token prediction

Google Research

Carnegie Mellon University
School of Computer Science

Roll the dice and look before you leap:

Going beyond the creative limits of next-token prediction

Chen Wu *,
CMU

Vaishnavh Nagarajan*,
Google Research

Charles Ding,
CMU

Aditi Raghunathan
CMU

Outline

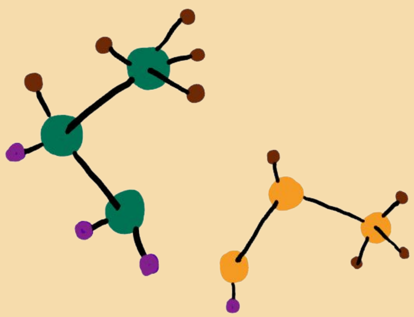
Part 1: Introduction & motivation

Part 2: Conceptual results

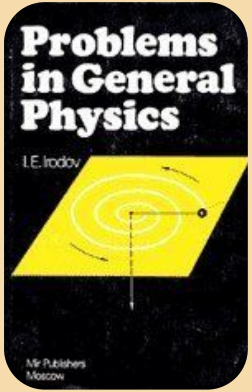
Part 3: Empirical results

Part 4: Concluding remarks

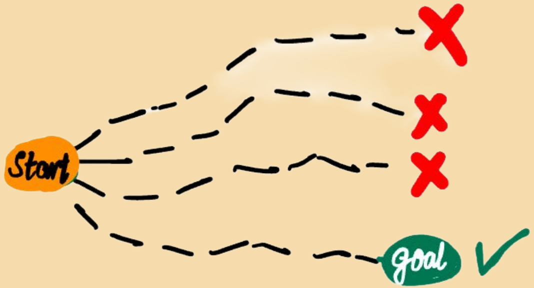
The next biggest challenge for LLMs:
Thinking creatively in open-ended tasks



Scientific discovery



Dataset generation



Test-time scaling (best-of-N)

Lots of critical & pioneering work debating 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]Car
AI Chair, [†]Equal Advising

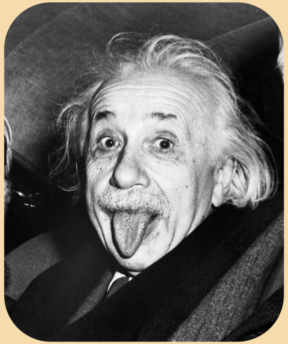
The Ideation–Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas

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

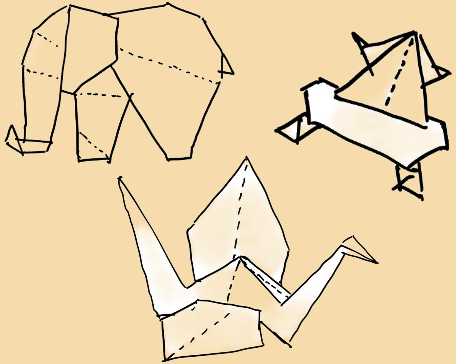
We must not only care about...

but also about:

Quality of a given generation



Originality against *massive* training set

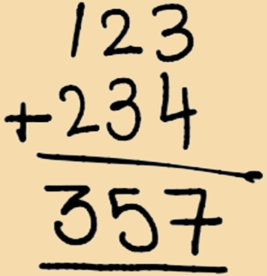


Diversity across generations

Is the current LLM paradigm optimal for *creative, open-ended* generations? Can we do better?

We need minimal tasks!

diversity on continuous data



reasoning on discrete data

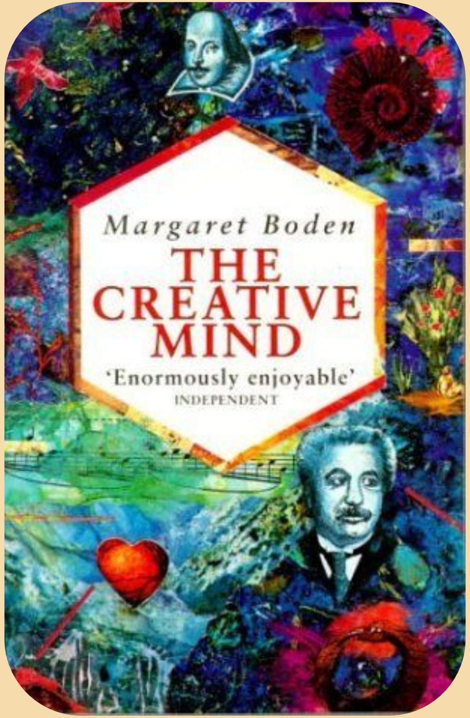
creativity

What we do:

We design minimal , open-ended, discrete-algorithmic tasks

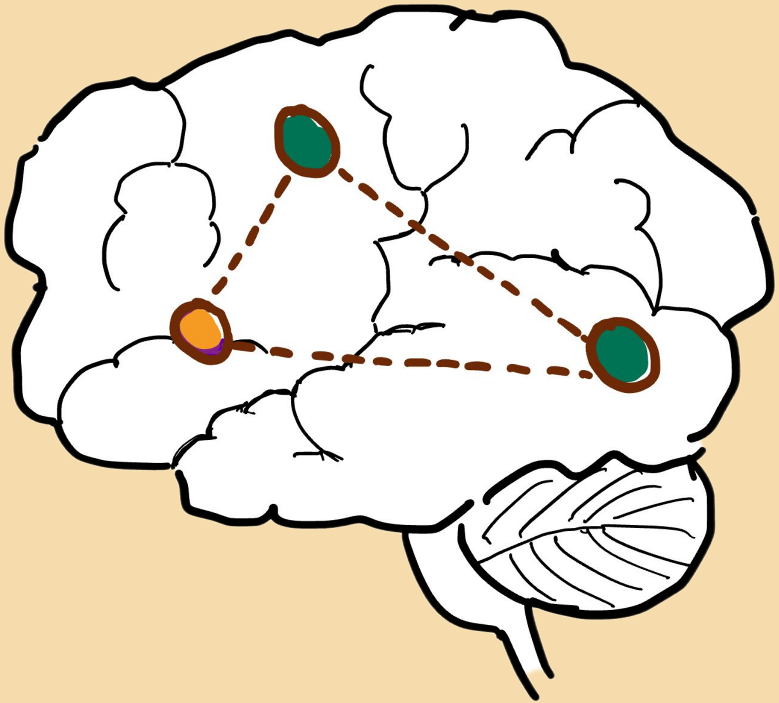
isolating two modes of creativity in cognitive science,

where we can quantify creative limits of LLMs & highlight alternatives.



Margaret Boden, 1990

Outline

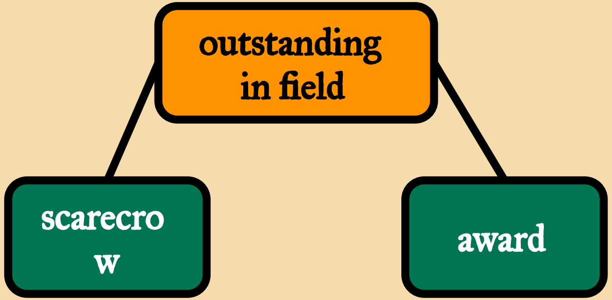


Combinational creativity

Wordplay in abstract form

Why did the scarecrow win an award?

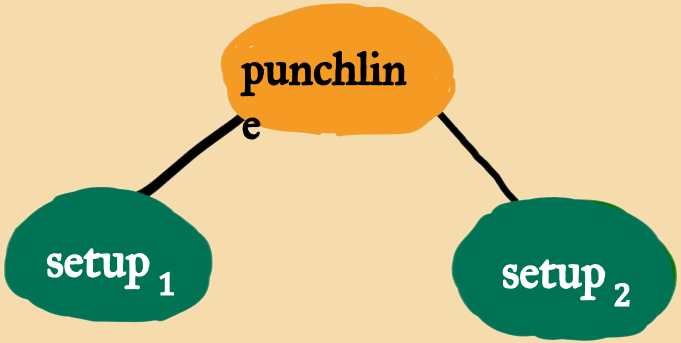
Because he was outstanding in his field!



Wordplay as "find a random, novel path over a large, known graph"

generate:

s.t.



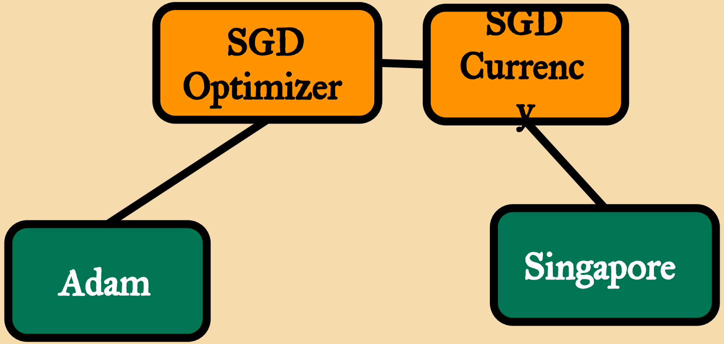
Dzmitry Bahdanau
@DBahdanau

[At ICLR'25 Singapore]

Adam deserves the award, but in Singapore everyone still uses SGD

6:32 PM · Apr 27, 2025 · **102K** Views

23 81 793 28



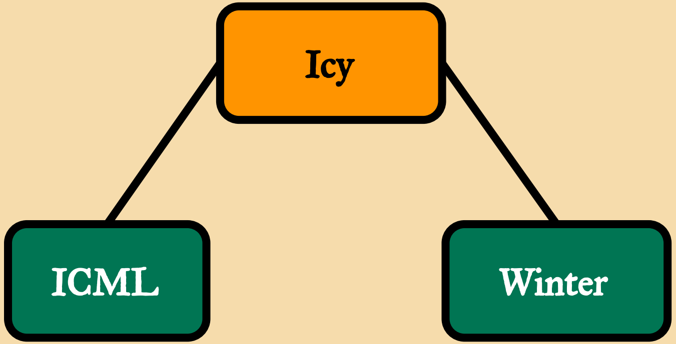
Ian Goodfellow
@goodfellow_ian

I see your joke suggestion, and raise you "Icy ML"

Tim Vieira @xtimv · Jul 12, 2018
New name for @NipsConference "AI Winter" — Miro Dudík

8:36 AM · Jul 13, 2018

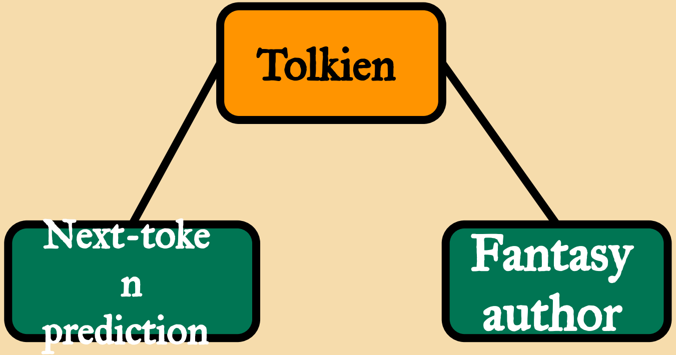
12 45 375



“Trained an LLM to predict if someone will be a successful fantasy author based on their writing samples,

Sounds fancy,

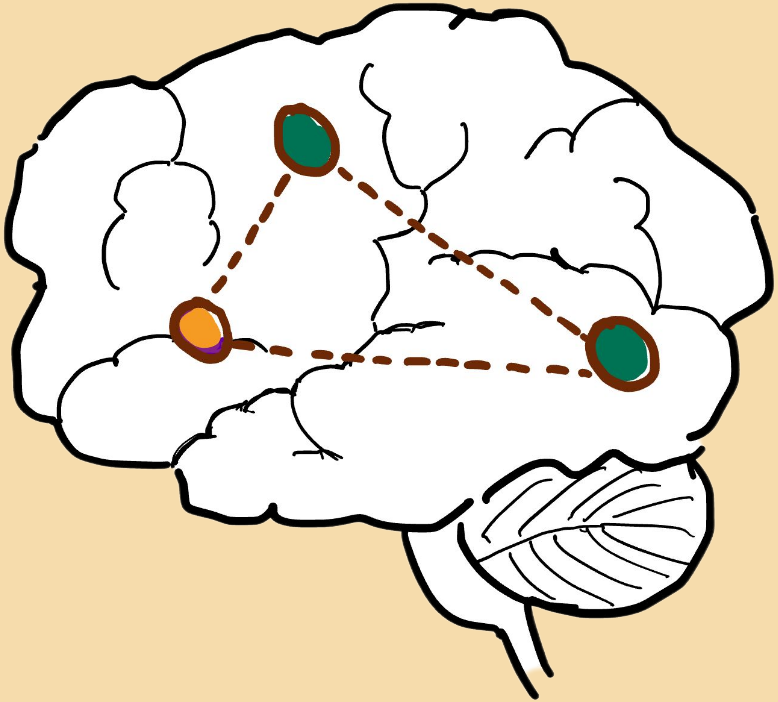
But all it’s doing is predicting the next Tolkien.”



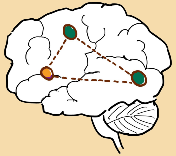
[Unabridged originals below]

Combinational creativity

- analogies,
- wordplay,
- discovering connections across literature



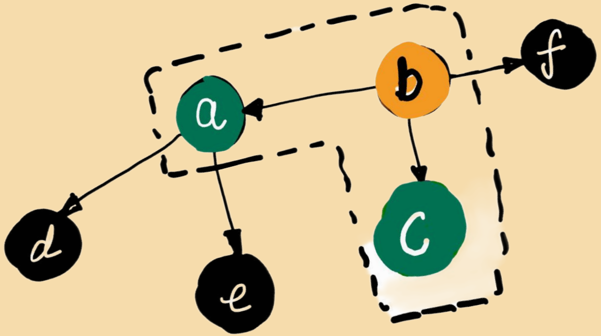
Search, retrieve and plan over *vast memory of known things* to find novel connections



We model combinational creativity as symbolic graph tasks

generate a c b

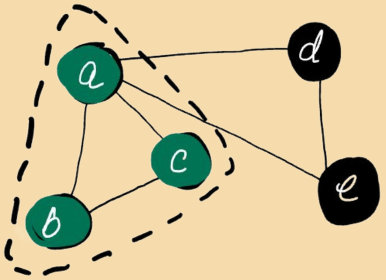
such that in in-weights graph



Discover novel sibling - parent triplets in an *in-weights* graph
[as a minimal wordplay abstraction]

generate a b c

such that in in-weights graph



Discover novel triangles in an *in-weights* graph [like finding contradictions or feedback loops]

Outline

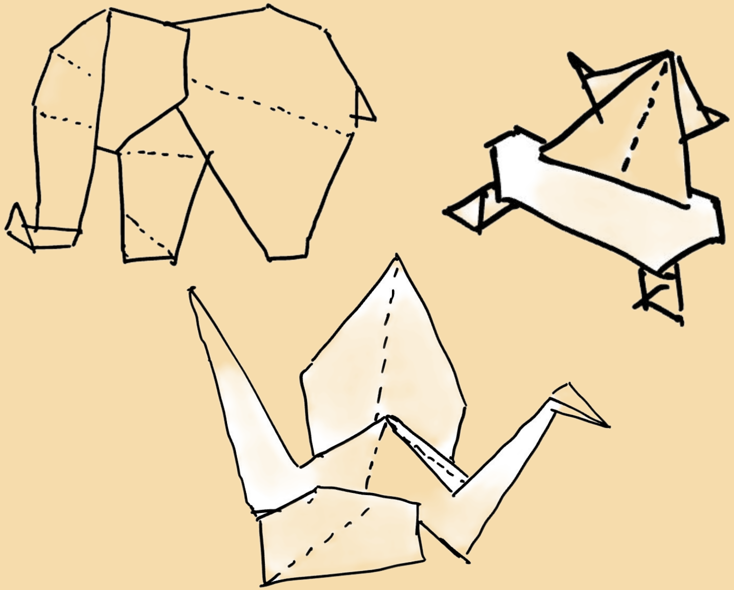
Part 1: Introduction & motivation

Part 2: Conceptual results: Two types of creative tasks

- Combinational creativity
- Exploratory creativity

Part 3: Empirical results

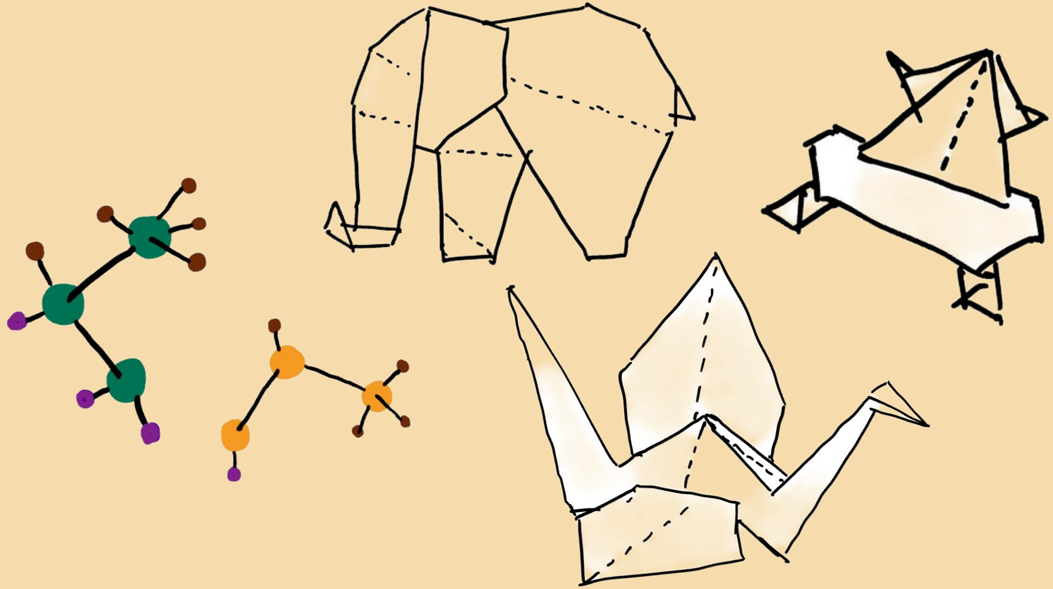
Part 4: Concluding remarks



Exploratory creativity

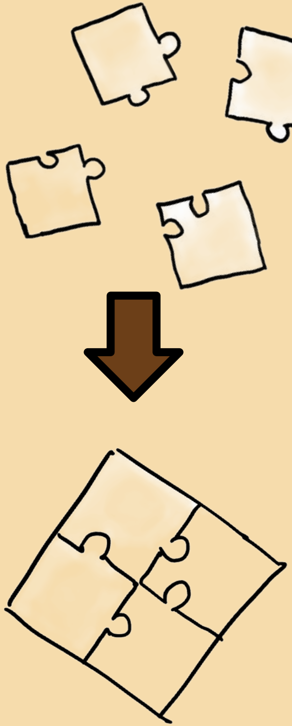
Exploratory creativity

- designing problems,
- generating molecules,
- deriving corollaries,
- crafting stories



Plan and devise novel patterns that obey *a small set of rules*
(you don't necessarily search over a vast memory)

For instance: Problem design

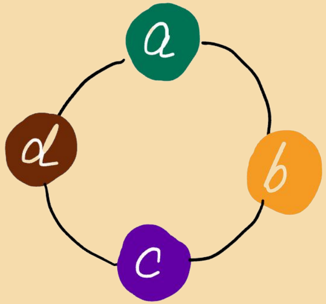


Set pieces in conflict such that there is a novel resolution under logical/math/... rules.

We model exploratory creativity as symbolic graph tasl

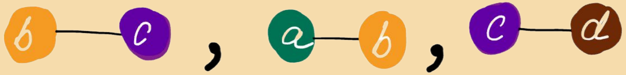
generate

such that

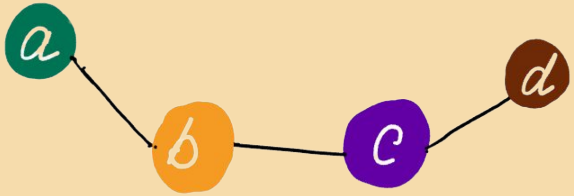


Construct adjacency lists that *resolve* into a circle graph through a novel permutation

generate



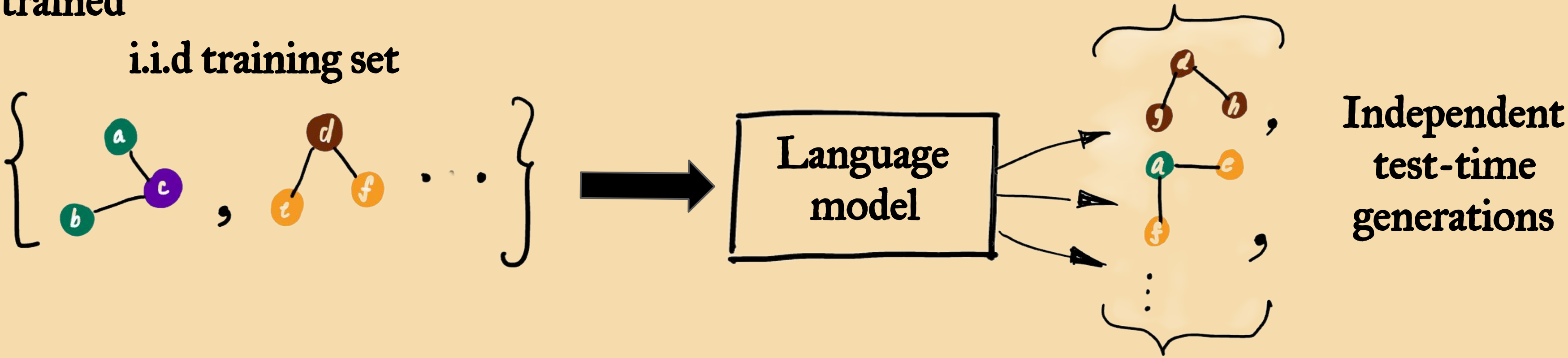
such that



Construct adjacency lists that *resolve* into a line graph through a novel permutation

How we cast these as learning tasks

Mimics pretraining or how protein/molecule generation models are trained



“Creativity” = Fraction of generations that are (a) unique (b) unseen and c) coherent

Is the current LLM paradigm optimal for *creative, open-ended* generations ***in these tasks*** ?

Outline

Part 1: Introduction & motivation

Part 2: Conceptual results: Two types of creative tasks

Part 3: Empirical results

- How learning signals are provided
- How diversity is elicited

Part 4: Concluding remarks

Outline

Part 1: Introduction & motivation

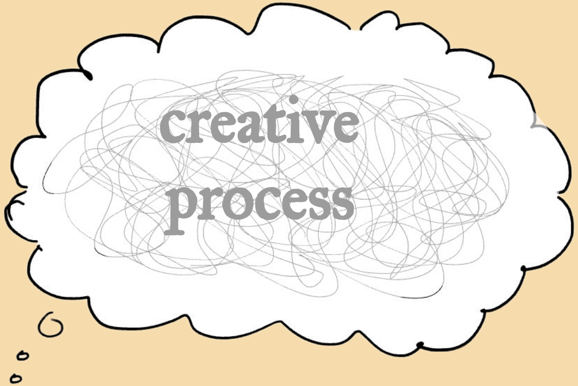
Part 2: Conceptual results: Two types of creative tasks

Part 3: Empirical results

- How learning signals are provided
- How diversity is elicited

Part 4: Concluding remarks

Creative outputs are generated from a creative process...

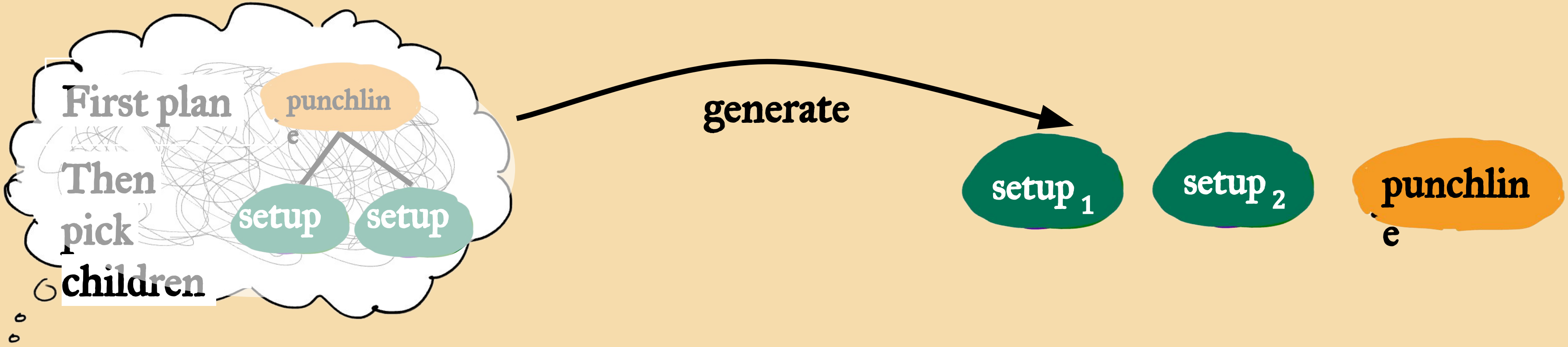


generate

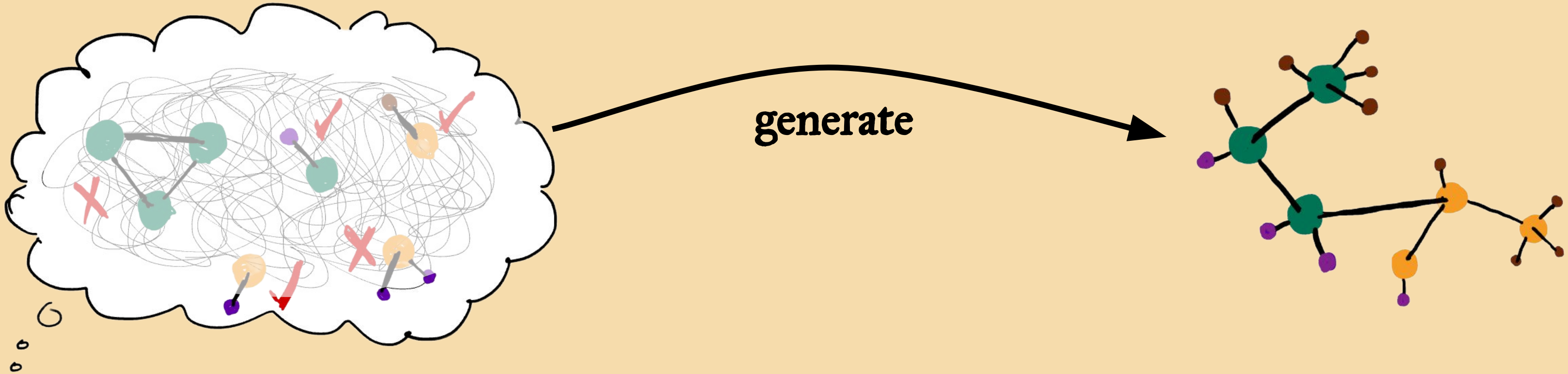
creative output

... that is unobserved and highly implicit in the output!

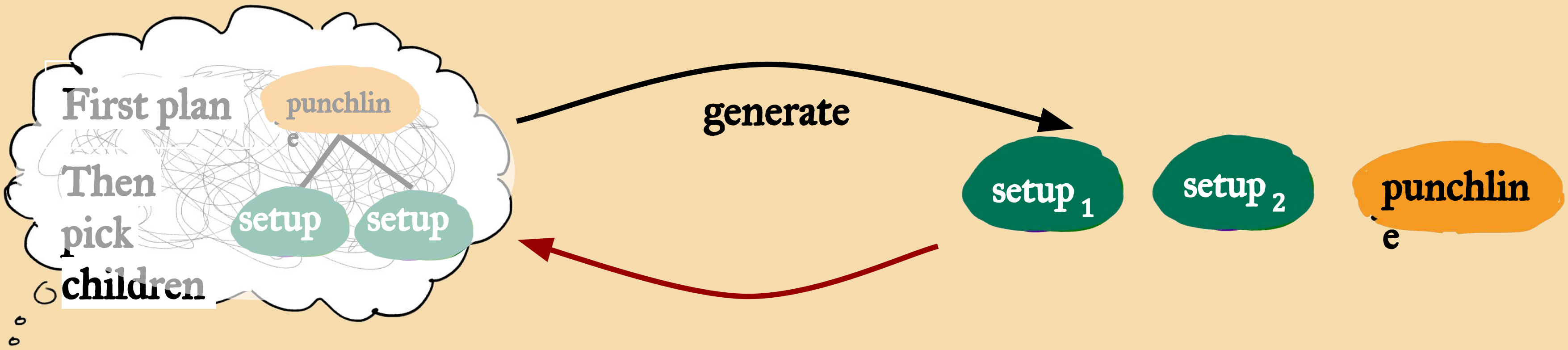
Creative outputs are generated from a creative process...



... that is unobserved and highly implicit in the output!



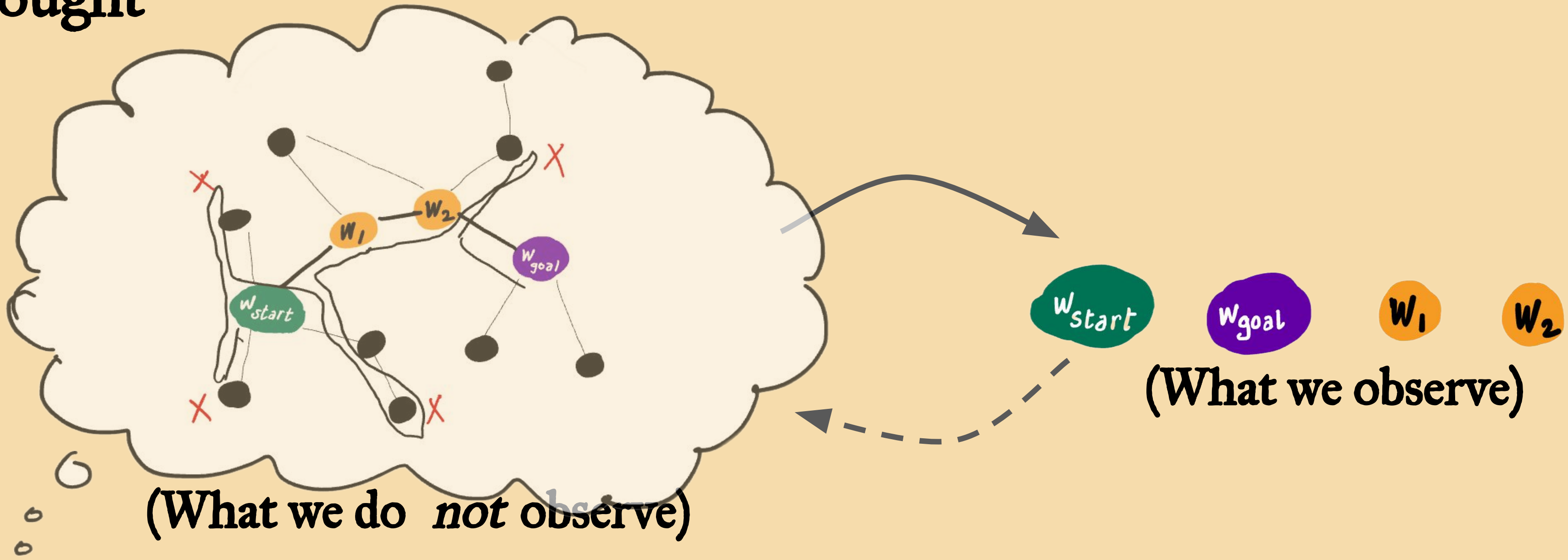
Creative outputs are generated from a creative process...



... that is unobserved and highly implicit in the output!

Our question: Can "local" next-token-learning on creative output infer the "global" end-to-end creative process?

Creative outputs are generated from an unobserved leap of thought



Our question: Can "local" next-token-learning on creative output infer the "global" end-to-end creative process?

Next-token learning is known to fail in a deterministic planning task.

The Pitfalls of Next-Token Prediction

Gregor Bachmann [*1] Vaishnavh Nagarajan [*2]

We extend this to our open-ended tasks:

Next-token learning may

resort to obvious local shortcuts (*Clever Hans cheats*),

ignore the implicit global pattern (*the creative planning process*),

memorize more, and reduce creativity.

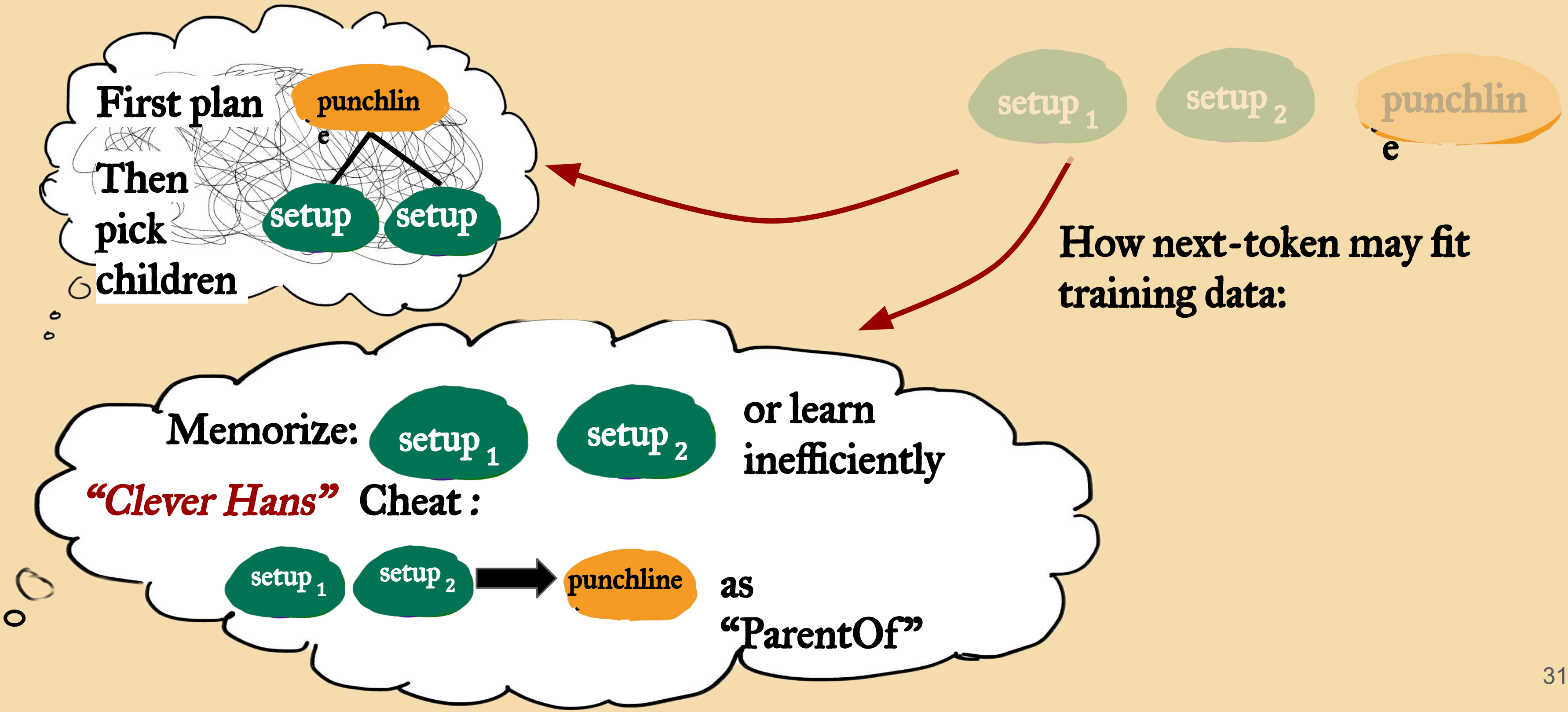
The Pitfalls of Next-Token Prediction

Gregor Bachmann [*1] Vaishnavh Nagarajan [*2]

We extend a known failure of next-token learning in some deterministic planning tasks to our open-ended creative tasks.

Hypothesis: How next-token learning may reduce creativity

How we want to fit training data:



Next-token learning

aka "Teacher-Forcing"

Target: 1 9 6 .

Input: 14 x 14 = 1 9 6

Target given as input, right-shifted.

Multi-token learning

Teacherless training

[Tschannen et al., '23 ;
Monea et al., '23;
Bachmann & Nagarajan, '24]

1 9 6 .

14 x 14 = [MASK] [MASK] [MASK]

Target not given as input.

Diffusion

SEDD
[Lou, Ming and Ermon '24]

1 6

14 x 14 = [MASK] 9 [MASK] .

Target masked to various levels given as input.

Next-token vs. multi-token learning

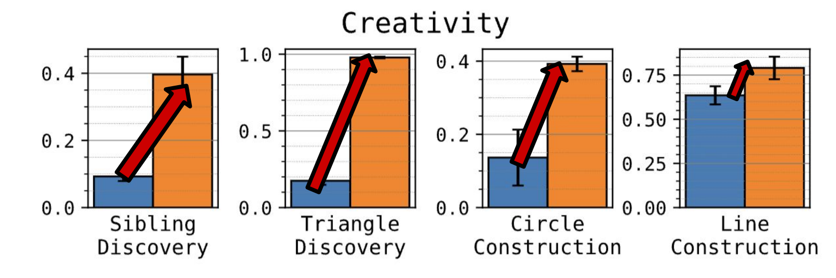
Training objectives

Standard next-token learning

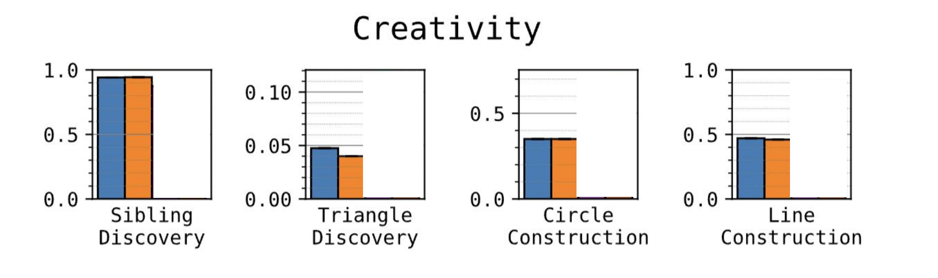
Teacherless multi-token

Creativity = fraction of generations that are unique, unseen and coherent

Gemma v1 (2B)



GPT-2 (86M)



Observation 1: Teacherless training is more creative than NTP for the larger Gemma model on all tasks. But not so for small model (echoes Gloeckle et al., 2024).

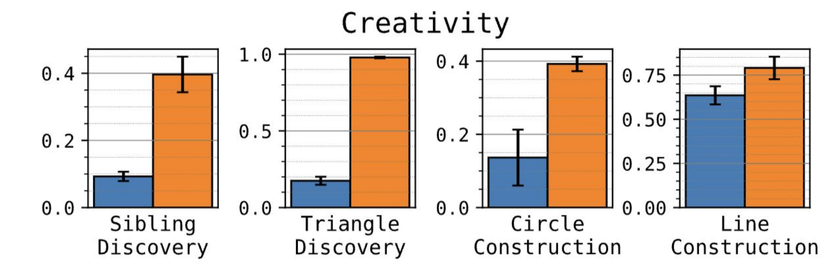
Next-token vs. multi-token learning

Training objectives

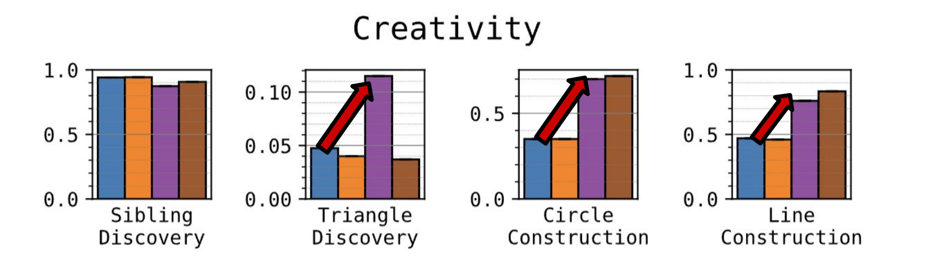
- Standard next-token learning
- Teacherless multi-token
- Uniform diffusion (multi-token)
- Absorbing diffusion (multi-token)

Creativity = fraction of generations that are unique, unseen and coherent

Gemma v1 (2B)



GPT-2 (86M)

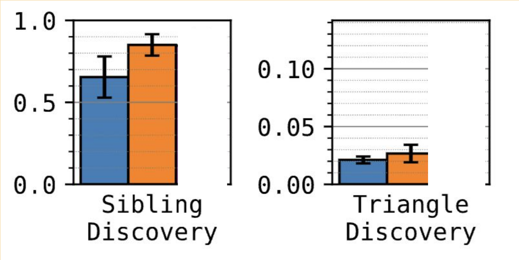


Observation 2: On smaller model, diffusion is more creative than NTP except on sibling dataset (which appears too easy).

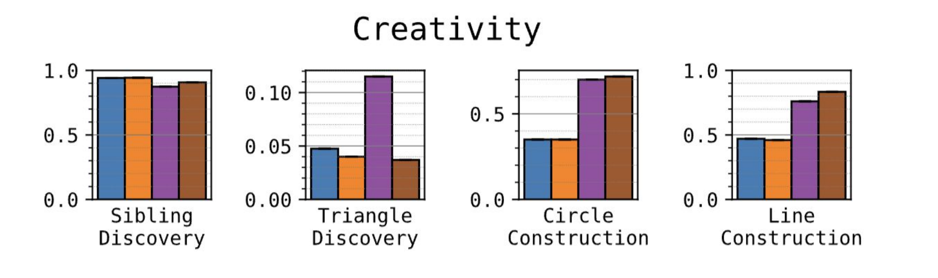
Next-token vs. multi-token learning

teacherless vs diffusion (SEDD [Lou, Ming and Ermon '24])

GPT-2 *with top-K*



GPT-2 (86M) vs diffusion (100M)



Creativity = fraction of generations that are unique, unseen and coherent

Observation 3: For smaller model, teacherless training does improve creativity *on the top-K* samples of the generated distribution

Outline

Part 1: Introduction & motivation

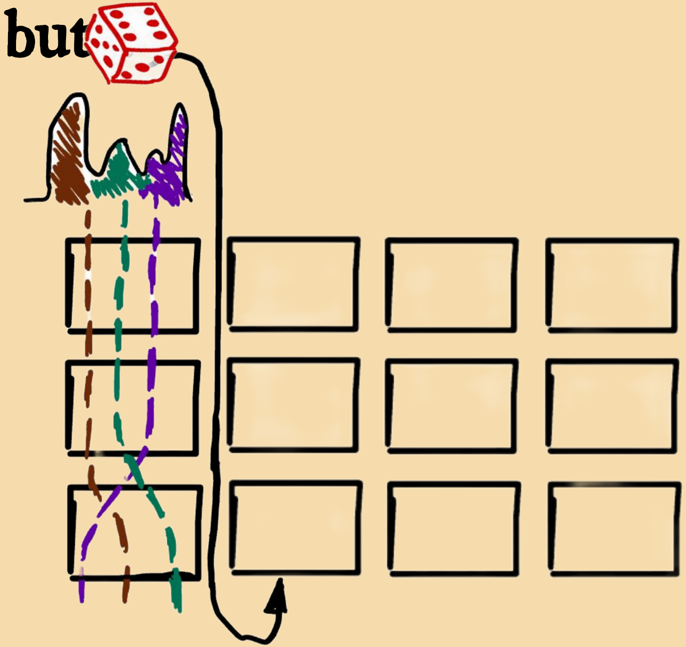
Part 2: Conceptual results: Two types of creative tasks

Part 3: Empirical results

- How learning signals are provided
- How diversity is elicited

Part 4: Concluding remarks

Diversity is typically elicited through temperature sampling but

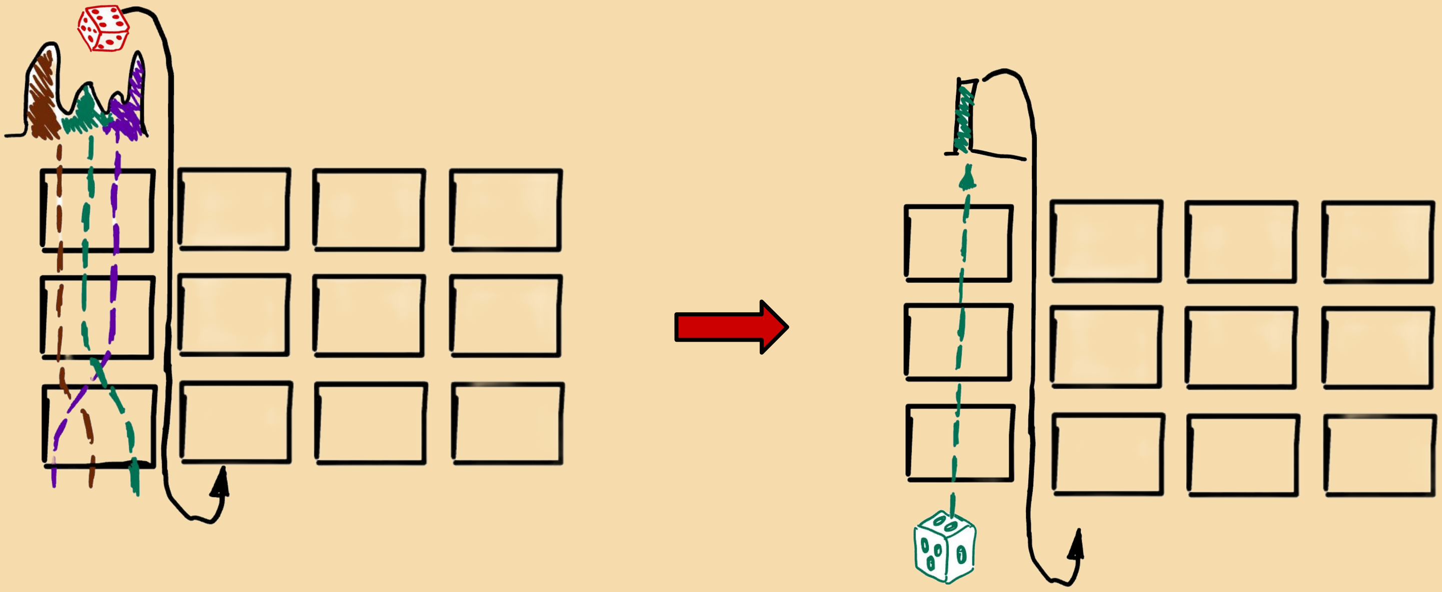


The model is forced to flesh out many diverse creative processes

for a diverse next-token distribution.

Our question: Temperature sampling demands "overparallelism" for diversity; this seems burdensome! Is there an alternative?

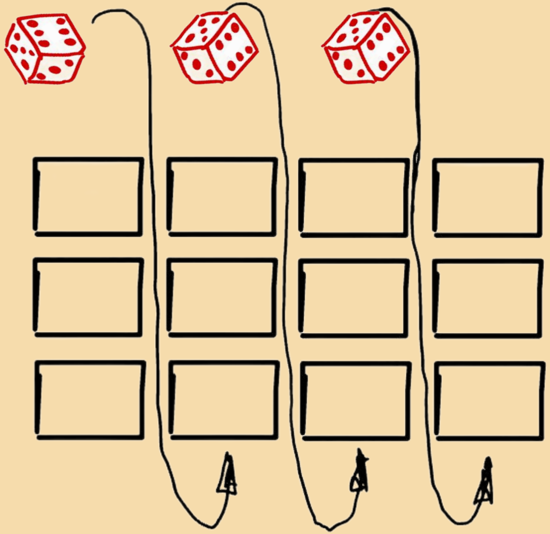
Can we focus on fleshing out one thought instead of parallelizing many?



Seed-conditioning as an alternative to temperature sampling

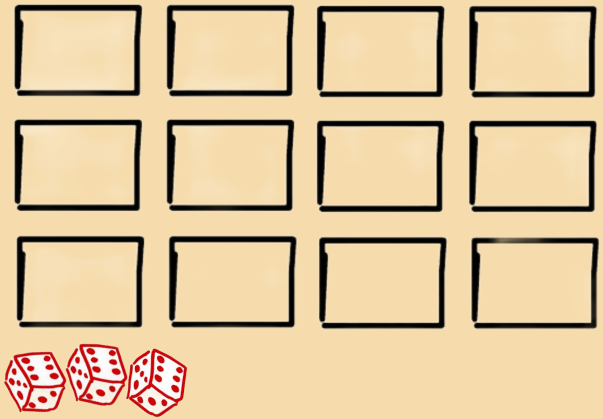
Instead of
output -randomization,

Temperature sampling



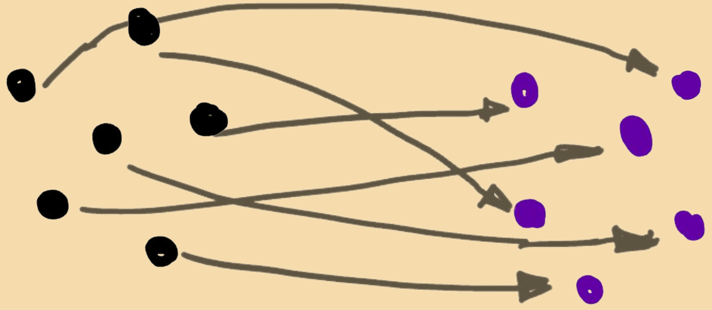
we try *input* -randomization —
like in GANs/VAEs, but way more
naively

Seed-conditioning: Prefixing random tokens per example during training and testing

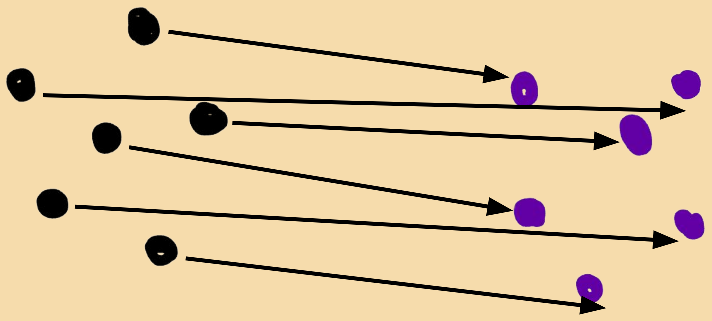


Or perhaps seed-conditioning is too naive?

Seed-conditioning arbitrarily dictates which noise binds to which output.

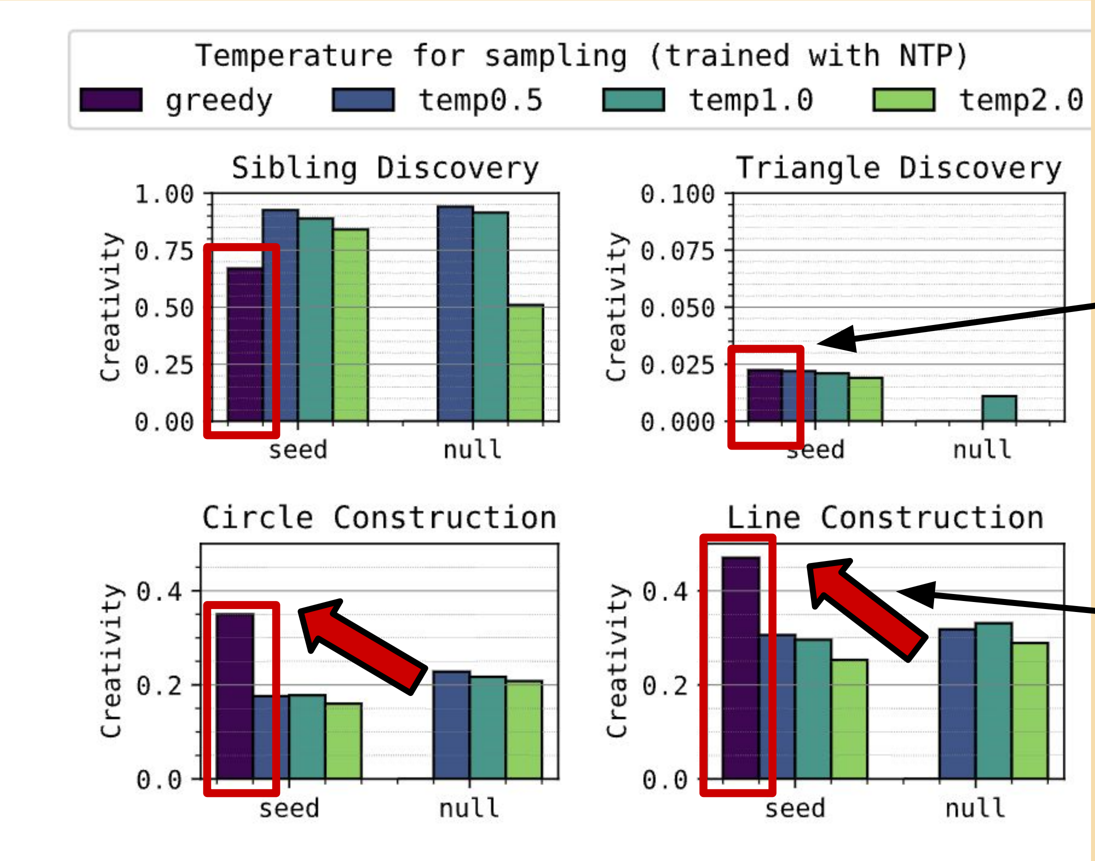


But typically (e.g., in GANs, VAEs), this binding is *learned!*



Put that way, seed-conditioning sounds like a terrible idea!

Seed-conditioning as an alternative to temperature sampling



(Figure is for GPT-2 model, but holds on Gemma V1 too)

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

Seed-conditioning can even be the most creative method!

Outline

Part 1: Introduction & motivation

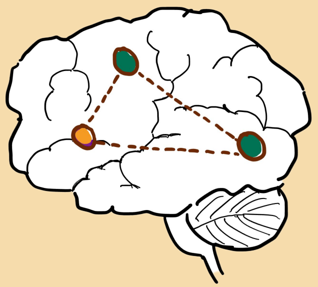
Part 2: Conceptual results: Two types of creative tasks

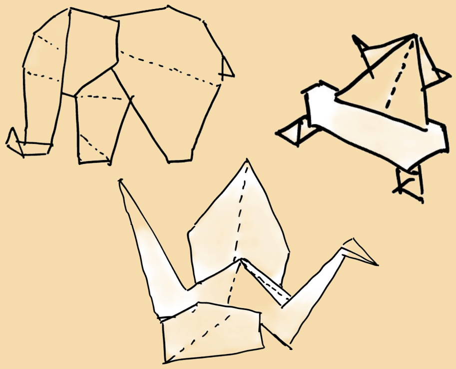
Part 3: Empirical results

Part 4: Concluding remarks

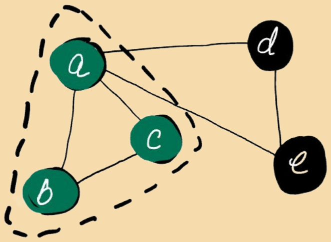
Summary

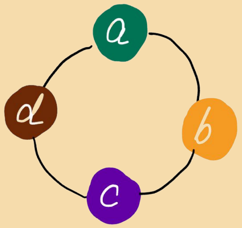
1. Two types of creativity in cognitive science:
   a. combinational (wordplay, analogies)
   b. exploratory (problem design)





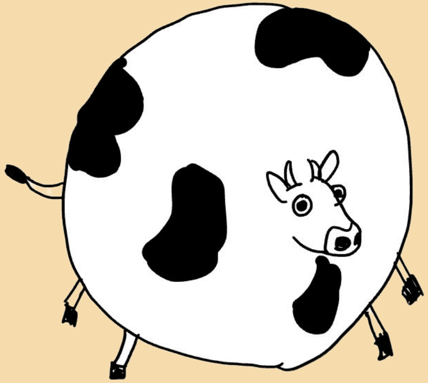
2. We abstracted these as minimal, graph-algorithmic tasks.
   a. Discovering novel in-weights structures
   b. Constructing adjacency lists that resolve





3. Compared next-token learning vs multi-token learning and temperature sampling vs seed-conditioning

Limitations



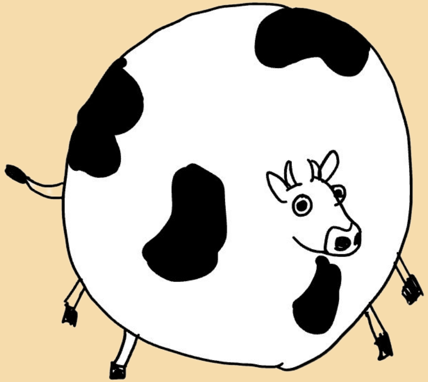
1. Our ideas need to be tested in the real-world.

2. Our findings are still not fully characterized (model-size, pretraining)

3. We do not look at how RL post-training, CoT, thinking addresses creative limits.
    - Still useful to improve the base model's skills, data/compute-efficiency
    - Can mere exploration + sparse rewards discover creativity?

4. We do not capture the full richness of creativity, subjective aspects (surprisingness, interestingness…).

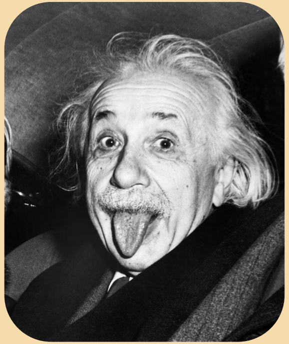
Future Work



1. Use our tasks to think clearly, inspire new ideas, do sniff tests, debug etc., e.g., *length generalization, shifts, in-context learning*

2. Seed-conditioning:

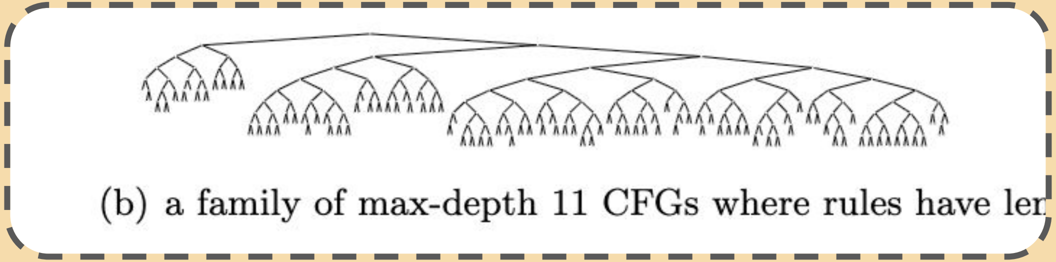
- Make it work *in the wild*
- *Understand* why it works as it is.



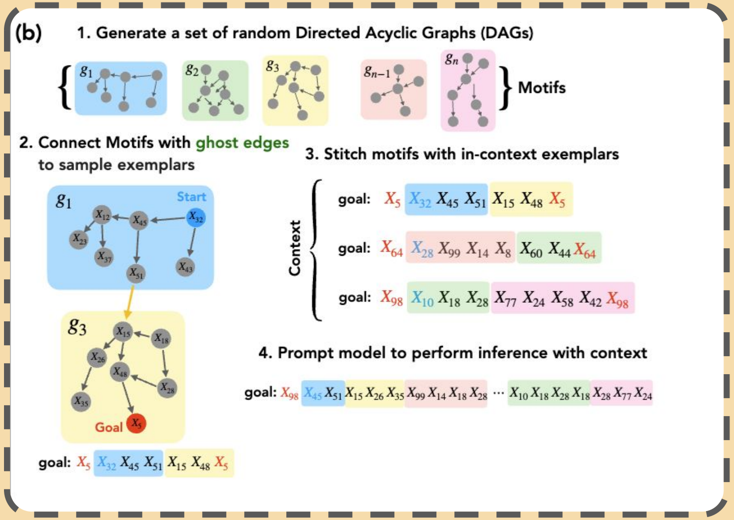
3. Tasks for "*transformational* creativity", extrapolative creativity, out-of-the-box thinking...

Controlled tasks are valuable!

CFG
Physics of Language Models: Part 1,
Allen-Zhu & Li 2023



Graph path-finding
"Towards an Understanding of Stepwise Inference in Transformers: A Synthetic Graph Navigation Model"
Khona, Okawa, Hula, Ramesh, Nishi, Dick, Lubana, & Tanaka 2024



Empirical analysis of temperature sampling

Is Temperature the Creativity Parameter of Large Language Models?

Max Peeperkorn,[1] Tom Kouwenhoven,[2] Dan Brown,[3] and Anna Jordanous[1]
[1]School of Computing, University of Kent, United Kingdom
[2]Leiden Institute of Advanced Computer Science, Universiteit Leiden, Netherlands
[3]Cheriton School of Computer Science, University of Waterloo, Canada

Concurrent position paper arguing for injecting randomness

Why LLMs Cannot Think and How to Fix It

Marius Jahrens
Institute of Neuro- and Bioinformatics
University of Lübeck
Lübeck, Germany 23562
m.jahrens@uni-luebeck.de

Thomas Martinetz
Institute of Neuro- and Bioinformatics
University of Lübeck
Lübeck, Germany 23562
thomas.martinetz@uni-luebeck.de

Prior work that *learns* the noise injected for diversity

SOFTSRV: LEARN TO GENERATE TARGETED SYNTHETIC DATA

Giulia DeSalvo, Jean-Fraçois Kagy, Lazaros Karydas, Afshin Rostamizadeh, Sanjiv Kumar
Google Research
New York, NY 10011, USA
{giuliad, jfkagy, lkary, rostami, sanjivk}@google.com

Many works on defining creativity!

On the Creativity of Large Language Models

Giorgio Franceschelli[1] and Mirco Musolesi[2, 1]

[1]University of Bologna, Italy
[2]University College London, United Kingdom
giorgio.franceschelli@unibo.it, m.musolesi@ucl.ac.uk

Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010)

Jürgen Schmidhuber

Can AI Be as Creative as Humans?

Haonan Wang[1] **James Zou**[2] **Michael Mozer**[3] **Anirudh Goyal**[3] **Alex Lamb**[4] **Linjun Zhang**[5]
Weijie J. Su[6] **Zhun Deng**[7] **Michael Qizhe Xie**[1] **Hannah Brown**[1] **Kenji Kawaguchi**[1]

[1]National University of Singapore [2]Stanford University [3]Google DeepMind
[4]Microsoft Research [5]Rutgers University [6]University of Pennsylvania

Project Page: ai-relative-creativity.gith

Art or Artifice? Large Language Models and the False Promise of Creativity

Tuhin Chakrabarty
tuhin.chakr@cs.columbia.edu
Columbia University
USA

Philippe Laban
Salesforce AI Research
USA

Divyansh Agarwal
Salesforce AI Research
USA

Smaranda Muresan
smara@cs.columbia.edu
Columbia University
USA

Chien-Sheng Wu
Salesforce AI Research
USA

AI AS HUMANITY'S SALIERI: QUANTIFYING LINGUISTIC CREATIVITY OF LANGUAGE MODELS VIA SYSTEMATIC ATTRIBUTION OF MACHINE TEXT AGAINST WEB TEXT

Ximing Lu♡♠ **Melanie Sclar**♡ **Skyler Hallinan**♡ **Niloofar Mireshghallah**♡
Jiacheng Liu♡♠ **Seungju Han**♠ **Allyson Ettinger**♠ **Liwei Jiang**♡ **Khyathi Chandu**♠
Nouha Dziri♠ **Yejin Choi**♡

♡University of Washington ♠Allen Institute for Artificial Intelligence
{lux32,yejin}@cs.washington.edu

Many other works in different areas— see our related work section!

Thank you!

Poster: 11 a.m. – 1:30 p.m
East Exhibition Hall A-B
#E-2505

Gregor
Bachmann
(Apple)

The Pitfalls of Next-Token Prediction

Gregor Bachmann [*1] Vaishnavh Nagarajan [*2]

Thanks to Vansh Bansal, Gregor Bachmann, Jacob Springer, Sachin Goyal, Mike Mozer, Suhas Kotha, Clayton Sanford, Christina Baek, Yuxiao Qu, and Ziqian Zhong for valuable early discussions and pointers.

(All diagrams in the deck were human-drawn.)