Roll the dice and look before you leap:

Going beyond the creative limits of next-token prediction

Vaishnavh Nagarajan,
Google Research

Google Research

Thanks to my collaborators!

Chen Wu *,
CMU

Charles
Ding,
CMU

Aditi
Raghunathan
CMU

Gregor
Bachmann*,
Apple

🎲 *Roll the dice & look before you leap*:
Going beyond the creative limits of next-token prediction

Vaishnavh Nagarajan [*1] Chen Henry Wu [*2] Charles Ding [2] Aditi Raghunathan [2]

The Pitfalls of Next-Token Prediction

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

Outline

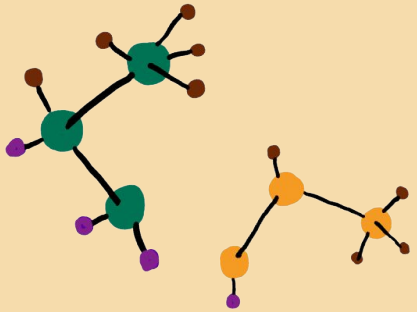
Part 1: Motivation

Part 2: Conceptual results

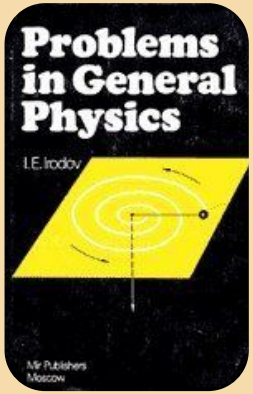
Part 3: Empirical results

Part 4: Concluding remarks

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



Scientific discovery



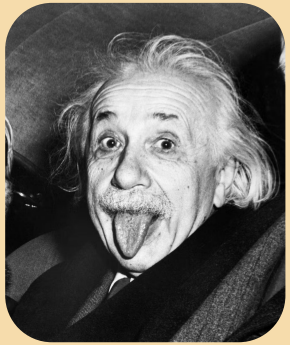
Dataset generation

Test-time Scaling (best-of-N)

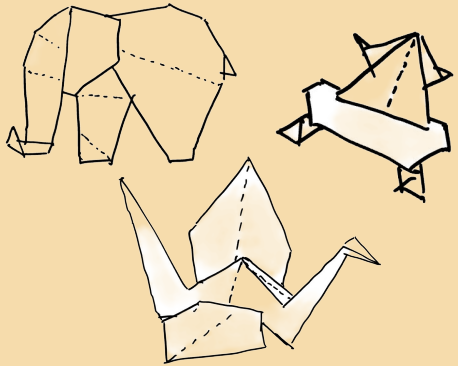
We must not only care about...

but also about:

Quality of a given generation



Originality against training set



Diversity across generations

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

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]Can
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

But studying real-world tasks is challenging!

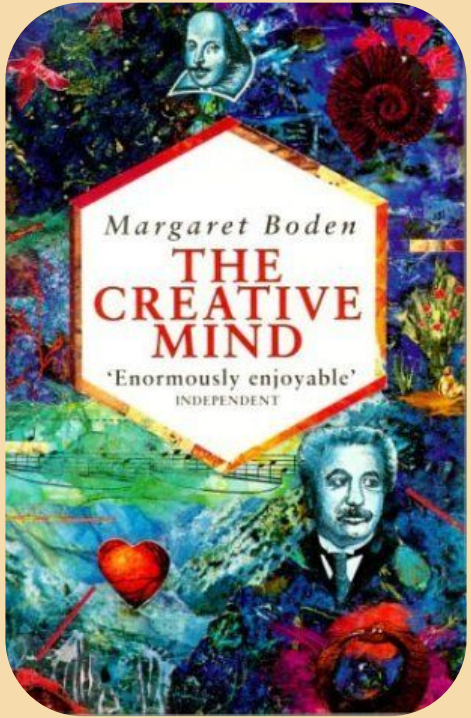
- **Metrics are *subjective***
  - What is truly novel and diverse?
- **Metrics are hard to scalably compute**
  - Novelty against whole internet!
- **Challenging to discuss with clarity**
- **Challenging to inspire & iterate & debug ideas**
  - So many confounding factors!

What we do:

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



Just to set expectations

1. There are no state-of-the-art results here
2. This is not an impressive large-scale study of complex real-world tasks.
3. The goal is to gain clarity and develop a *very* simple test-bed to inspire new ideas

Outline

Part 1: Motivation

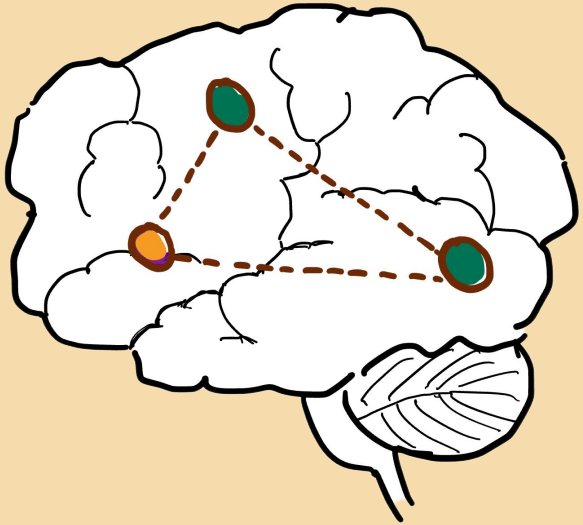
Part 2: Our two types of creative tasks

Part 3: Empirical results

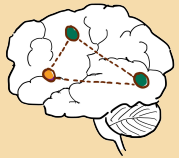
Part 4: Concluding remarks

Combinational creativity

- analogies,
- science,
- wordplay,
- discovering contradictions in literature



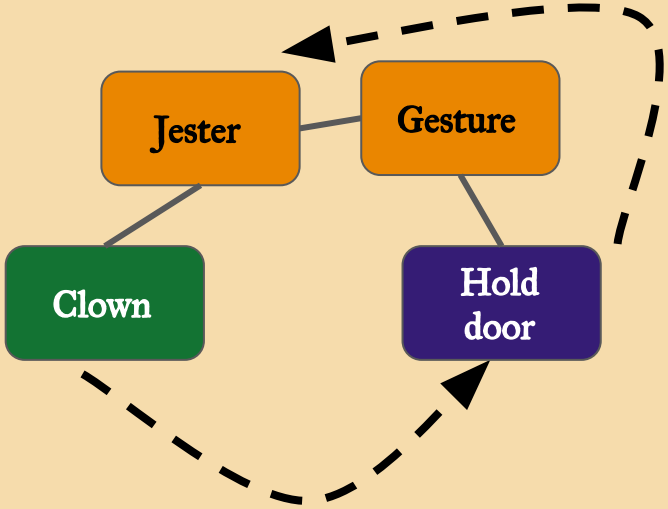
Search, retrieve and plan over *vast memory of known things* to find novel connections



For example: Wordplay

A clown held the door for me.

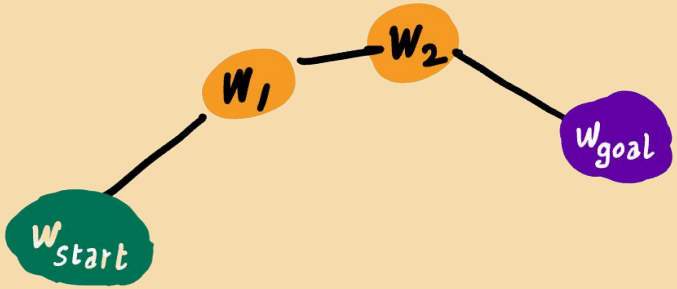
What a nice jester !

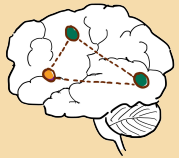


Wordplay as "find a novel path over a known vocabulary graph"

generate :

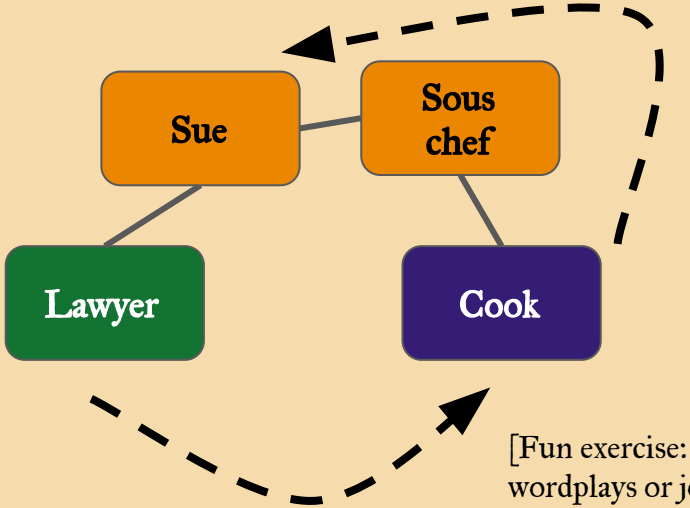
s.t.





For example: Wordplay

What do you call a lawyer who can cook. A sous chef!

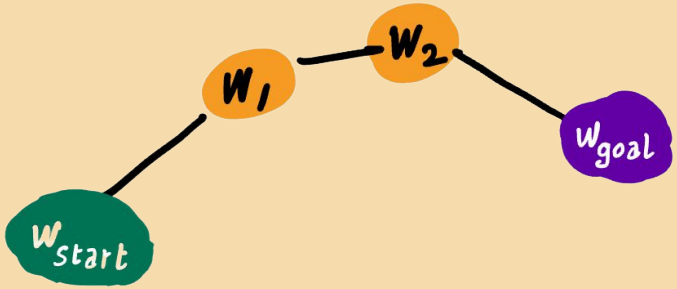


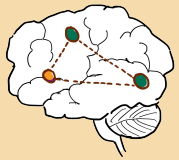
[Fun exercise: find other wordplays or jokes and reverse-engineer them]

Wordplay as "find a novel path over a known vocabulary graph"

generate :

s.t.

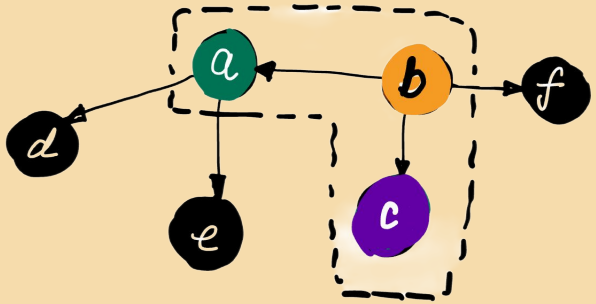




We model combinational creativity as minimal graph tasks

generate a c b

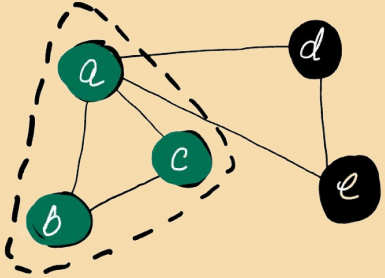
such that in in-weights graph



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: Motivation

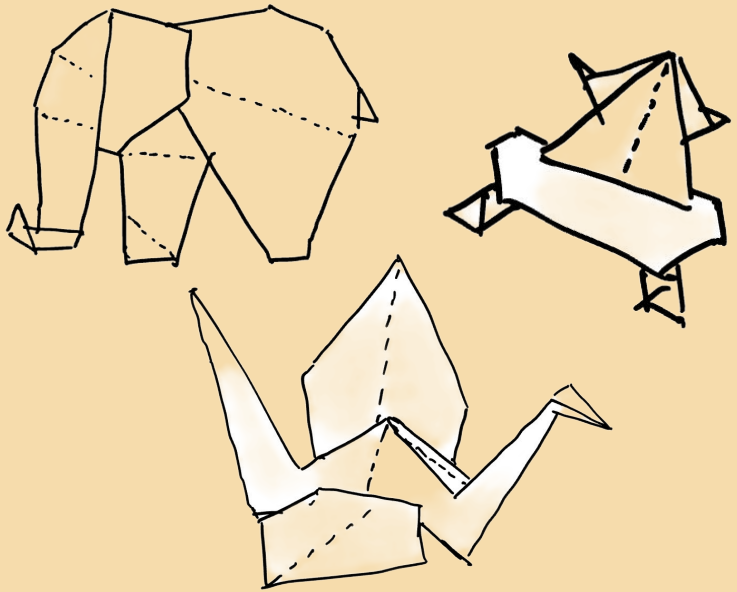
Part 2: Our *two* types of creative tasks

Part 3: Empirical results

Part 4: Concluding remarks

Exploratory creativity

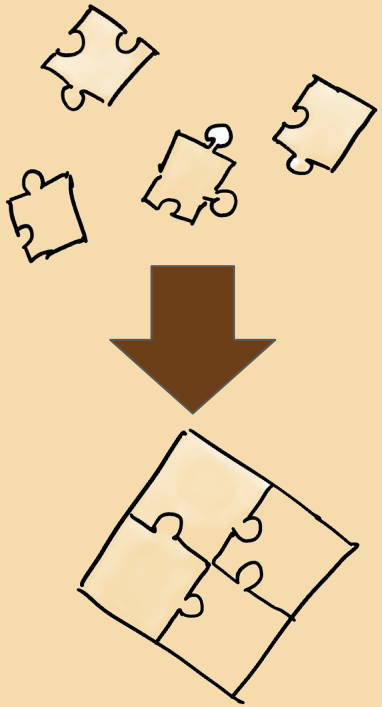
- designing problems,
- deriving corollaries,
- generating molecules,
- crafting stories



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

(you don't necessarily search over a vast memory)

For example: Problem design or story-writing

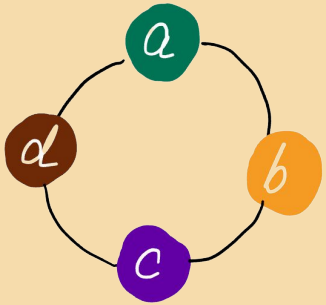


Set pieces in conflict such that there is a novel resolution under logical/math/... rules.

We model exploratory creativity as graph tasks

generate

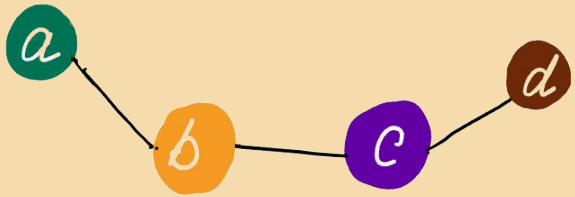
such that



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

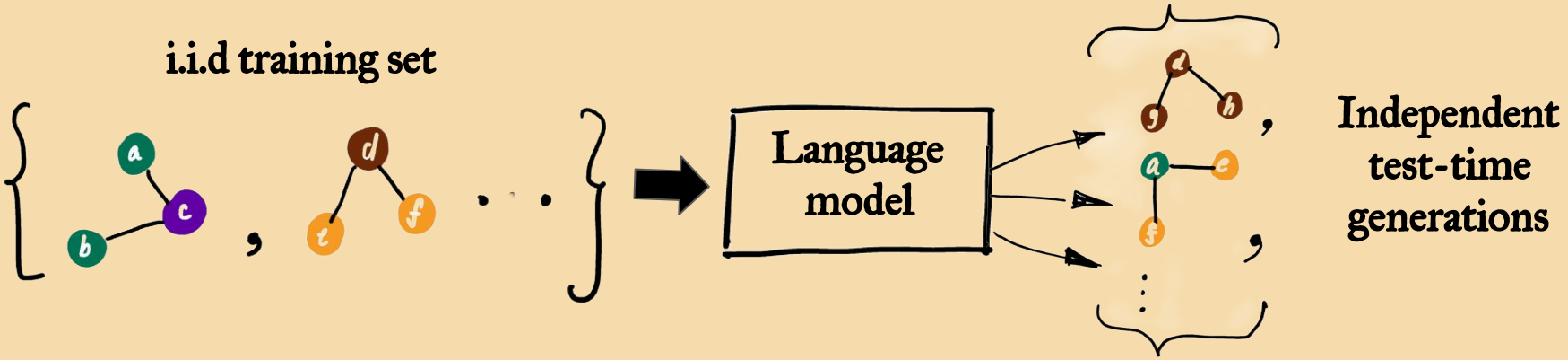
generate

such that



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

How we cast these as learning tasks



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

No one unique solution!

No natural language semantics involved — deliberately

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

Outline

Part 1: Motivation

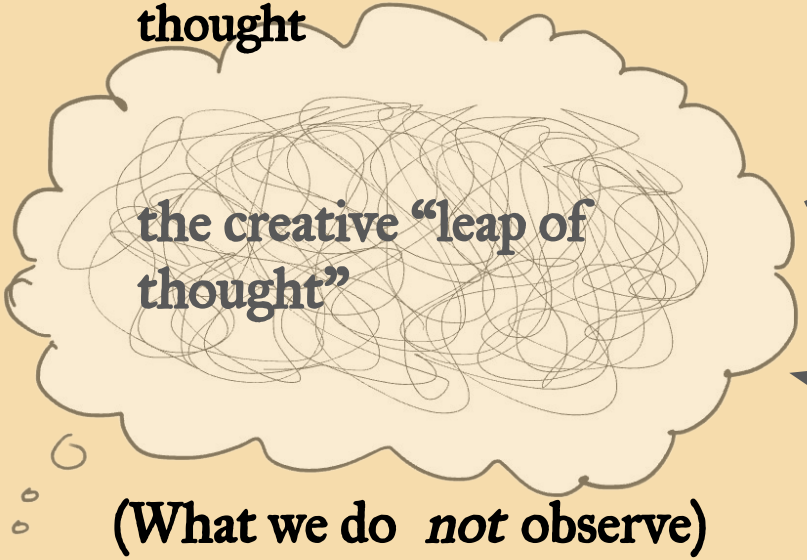
Part 2: Our two types of creative tasks

Part 3: Empirical results

- 3.1: Next-token vs multi-token learning
- 3.2 Temp sampling vs. seed-conditioning

Part 4: Concluding remarks

Creative outputs are generated from an unobserved leap of thought



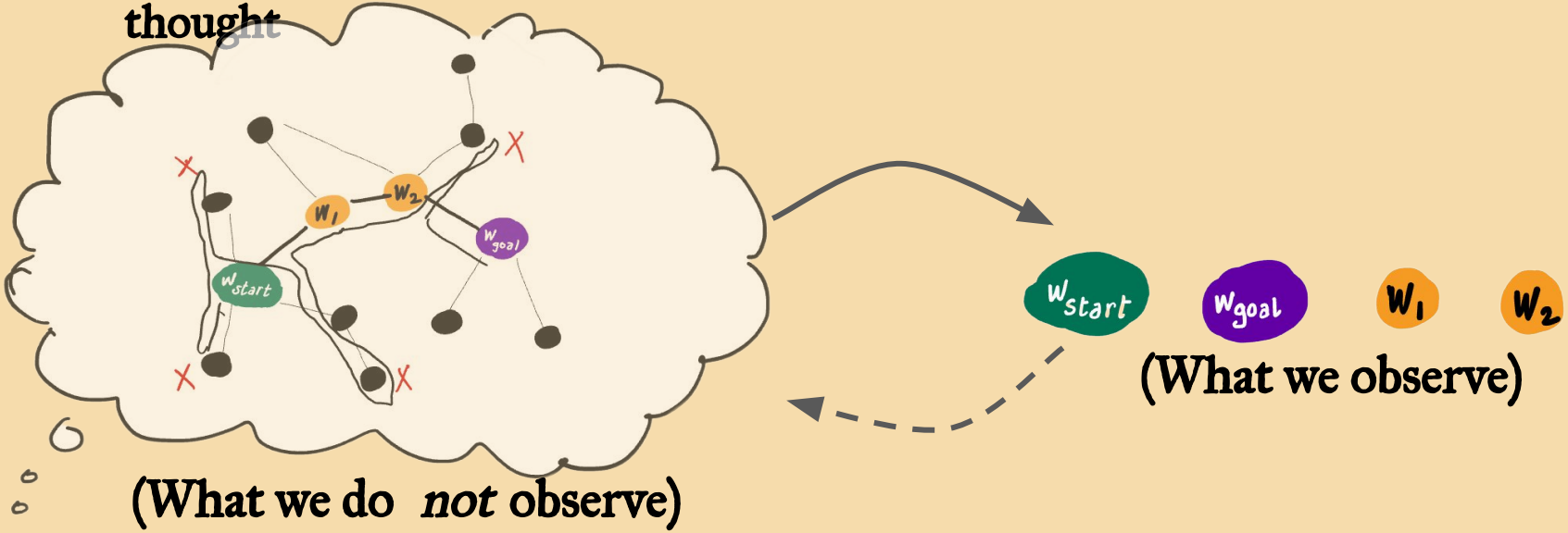
creative output

(What we observe)

(What we do *not* observe)

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

Creative outputs are generated from an unobserved leap of thought



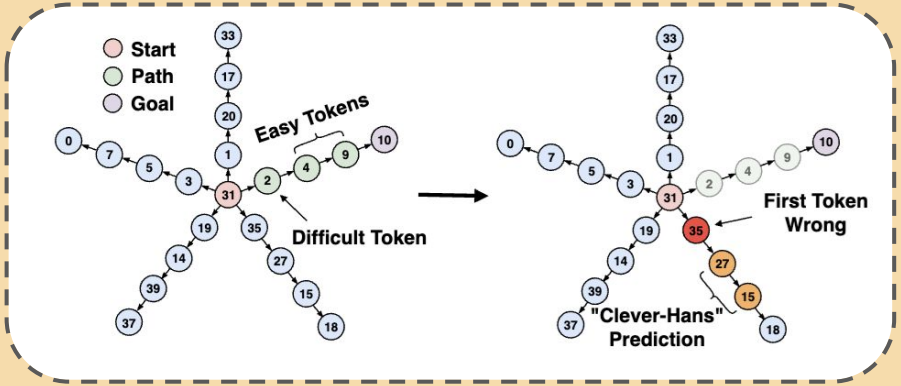
(What we do *not* observe)

(What we observe)

Can "local" next-token-learning on the creative output infer the "global" end-to-end creative process *in our tasks* ?

The Pitfalls of Next-Token Prediction

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

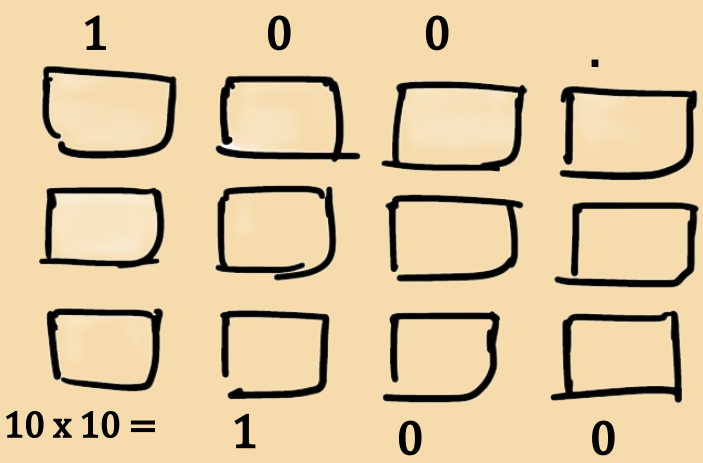


- Next-token learning fails is known to fail on a specific path-finding task
- *Intuition* : Model learns local patterns (“clever hans cheats”), ignoring the global pattern
- Not a failure of autoregressive *inference* , but of next-token *learning*

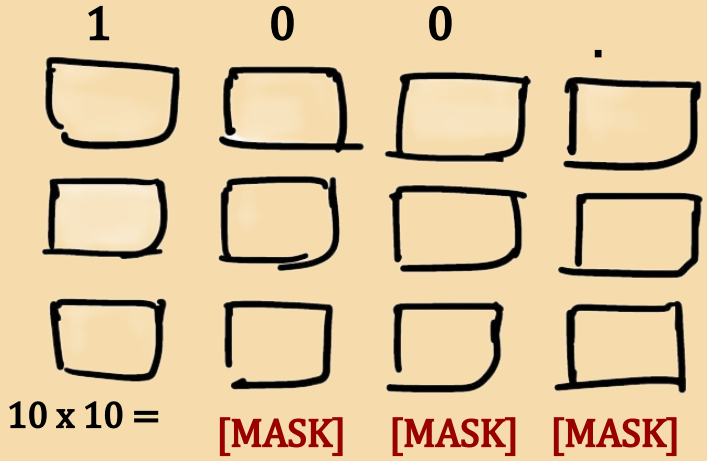
This is on a closed-ended multi-hop deterministic task; we extend this to fewer-hop, open-ended tasks.

Teacherless training

Tschannen et al., 2023 ; Monea et al., 2023; Bachmann and Nagarajan, 2024;



Standard next-token training
(aka “teacher-forced”)



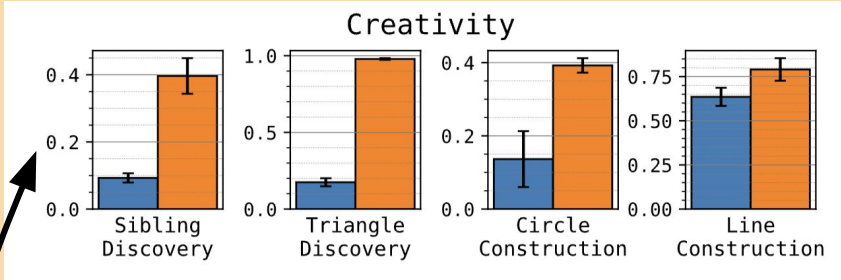
Teacherless training
(multi-token because targets “ 1 0 0” cannot see immediate past)

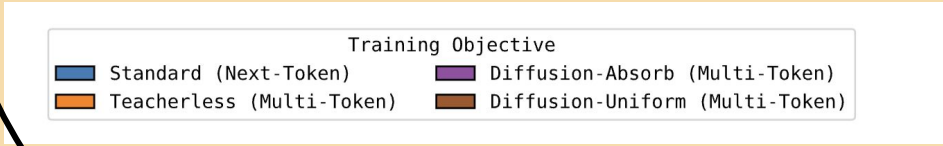
[Turns out that this is a term in diffusion with “absorb noise”!]

Next-token vs. multi-token learning

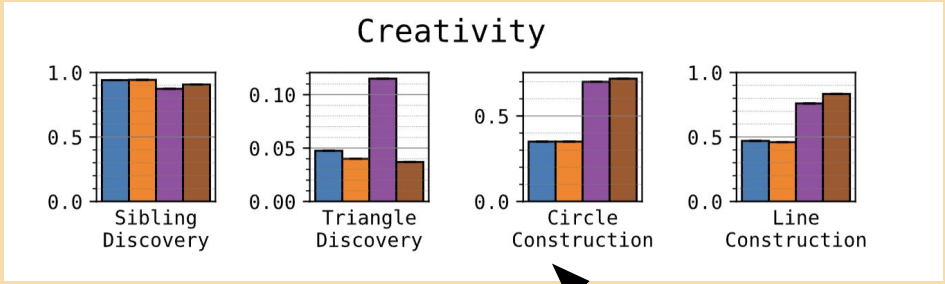
teacherless vs diffusion (SEDD [Lou, Ming and Ermon '24])

Gemma v1 (2B) pretrained





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



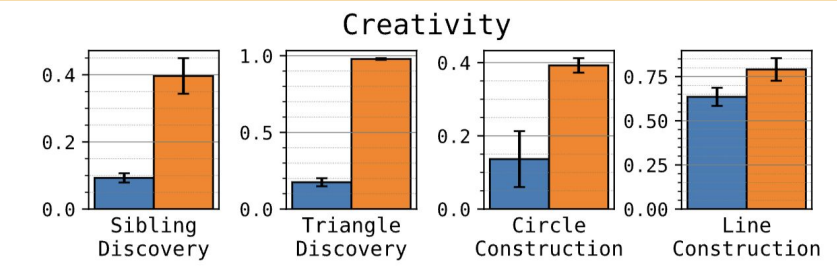
Creativity = fraction of generations that are unique, unseen and coherent

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

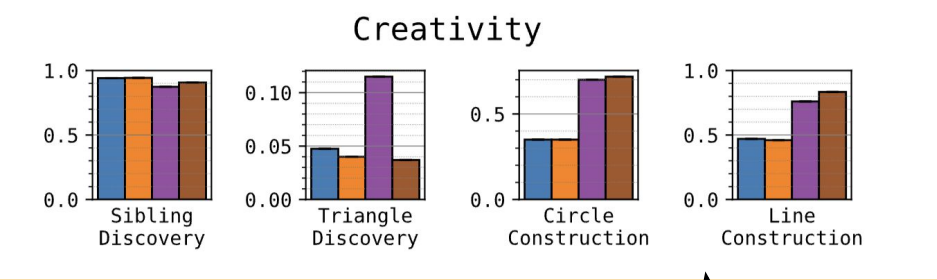
Next-token vs. multi-token learning

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

Gemma v1 (2B) pretrained



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



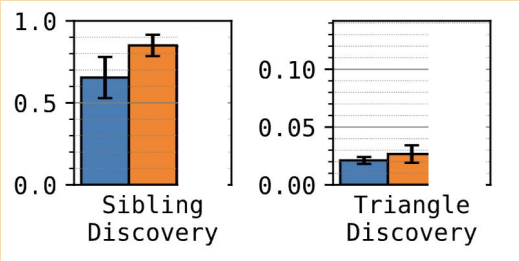
Creativity = fraction of generations that are unique, unseen and coherent

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

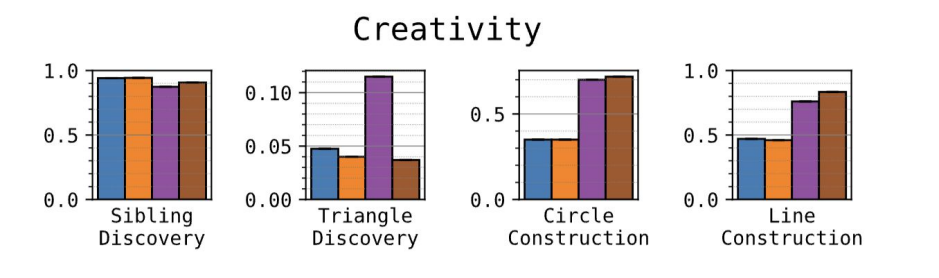
Next-token vs. multi-token learning

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: Motivation

Part 2: Our two types of creative tasks

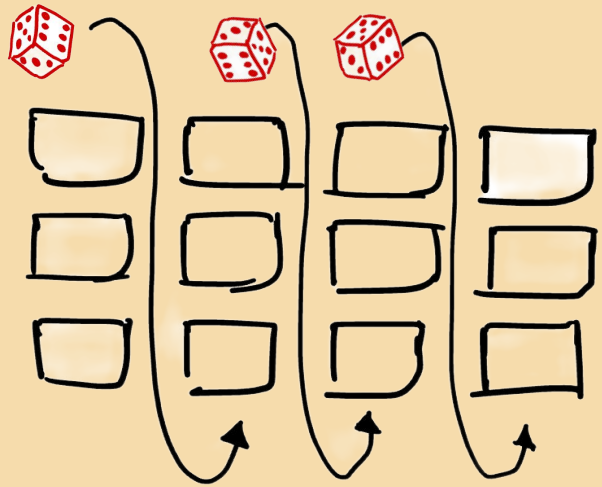
Part 3: Empirical results

- 3.1: Next-token vs multi-token learning
- 3.2 Temp sampling vs. seed-conditioning

Part 4: Concluding remarks

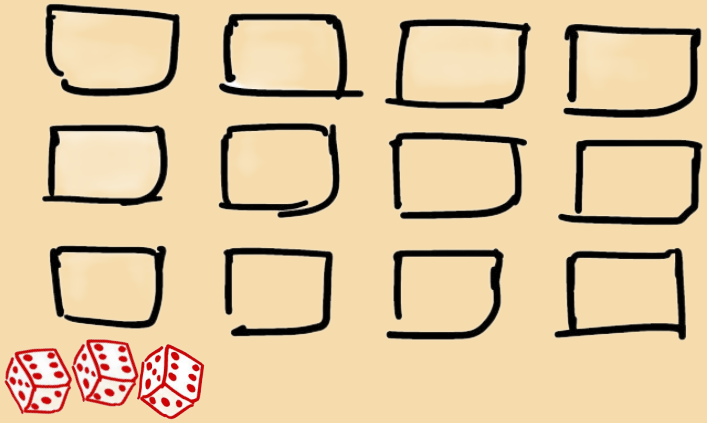
Let's revisit how diversity is elicited

Temperature sampling



But in GANs/VAEs, diversity came from input randomization!

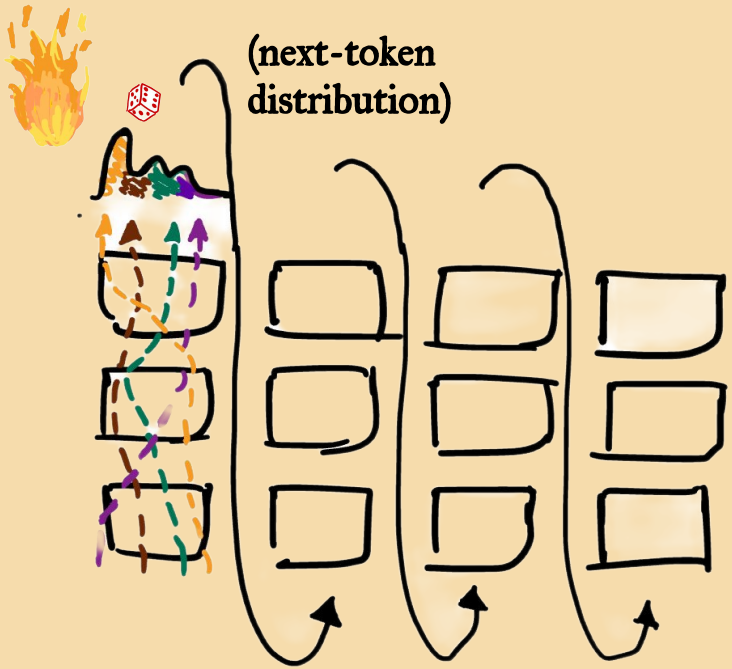
Seed-conditioning: Prefixing random strings per example during training and testing



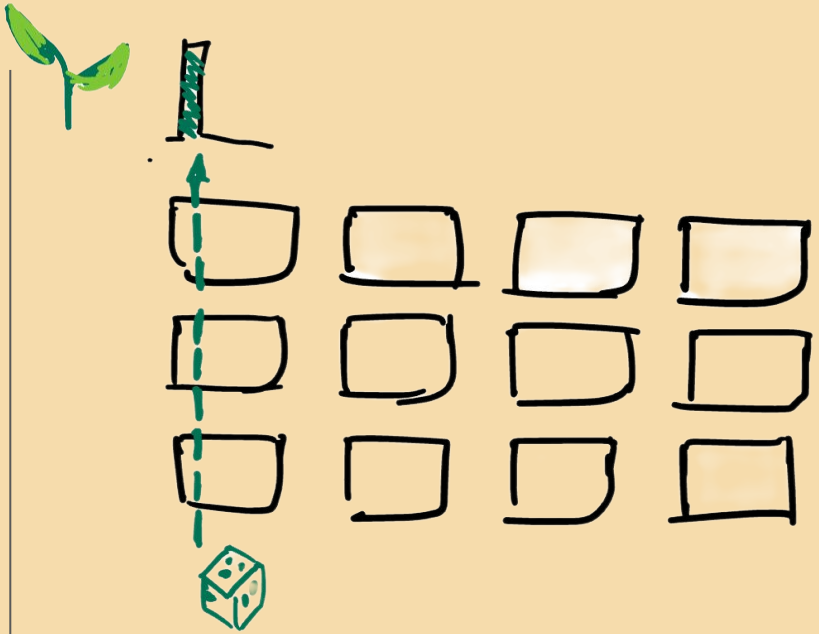
One intuition: Simulating variations in the prompt wording

Another (speculative) intuition:

there's overparallelism in Transformers; seed-conditioning tries to reduce this



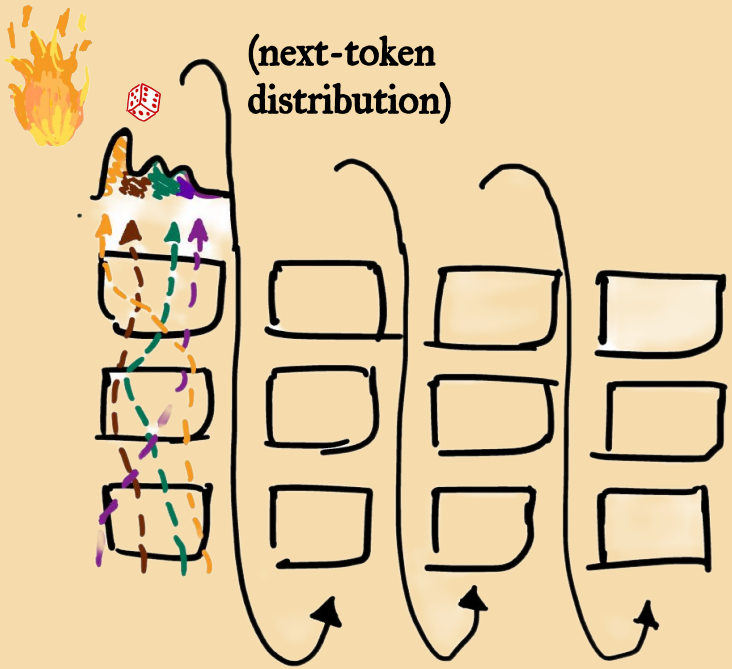
For temperature sampling, model must process many thoughts to produce diverse next-token distribution



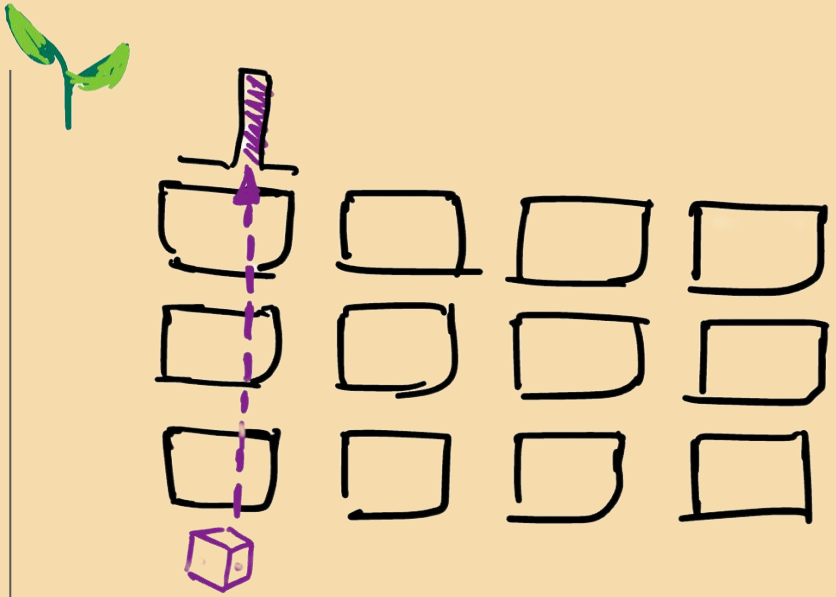
With seed-conditioning: model only needs to focus on one thought per seed

Another (speculative) intuition:

there's overparallelism in Transformers; seed-conditioning tries to reduce this



For temperature sampling, model must process many thoughts to produce diverse next-token distribution



With seed-conditioning: model only needs to focus on one thought per seed

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

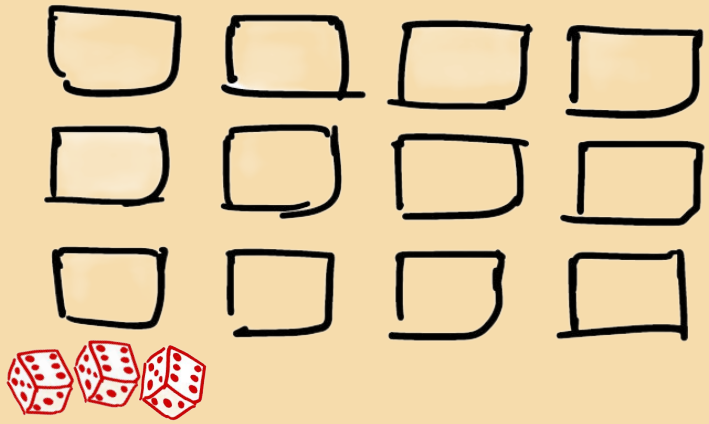
See also concurrent position paper

We thought perhaps seed-conditioning is too naive

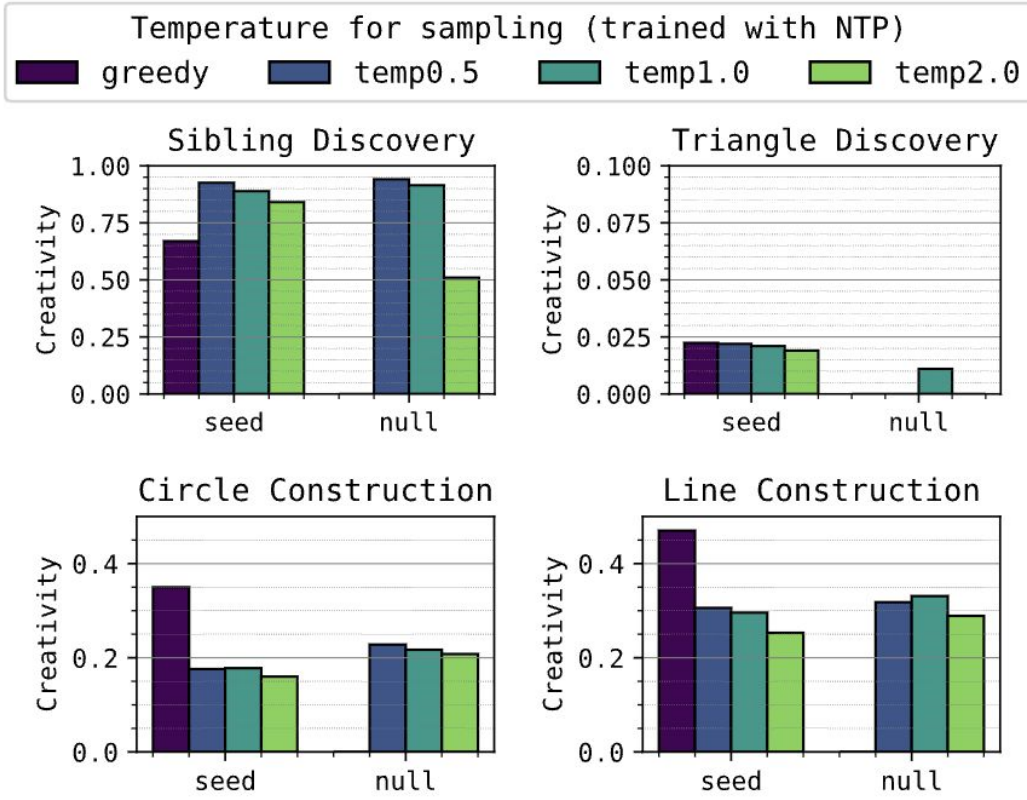
Whereas in VAEs and GANs, the "seed" is *learned*, here we create seed–output bindings arbitrarily.

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

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



But seed-conditioning works! (We don't know



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

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

Seed-conditioning can even be the most creative method!

Caveat: Requires training & no results are real data.

Also see: learned diversity-inducing technique for Transformers

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

Outline

Part 1: Motivation

Part 2: Our two types of creative tasks

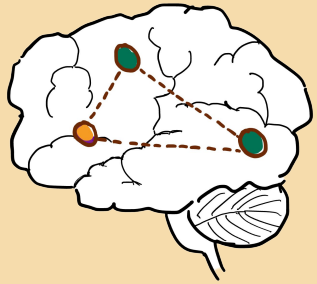
Part 3: Empirical results

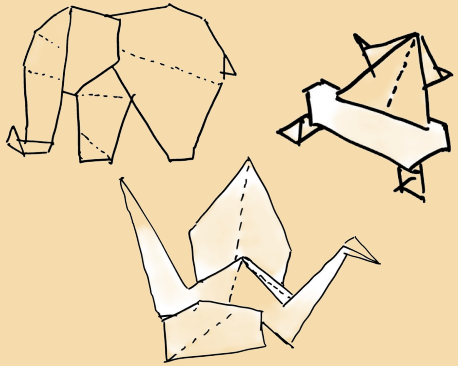
Part 4: Conclusion

1. Summary
2. Other remarks
3. Future work

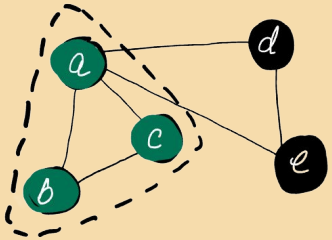
Summary

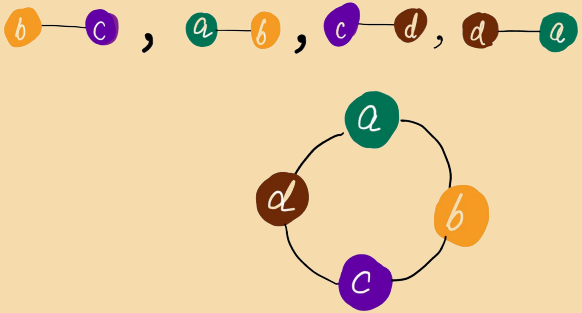
1. Two types of creativity in cognitive science:
   a. combinational (wordplay, analogies)
   b. exploratory (problem design)





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

Outline

Part 1: Motivation

Part 2: Our two types of creative tasks

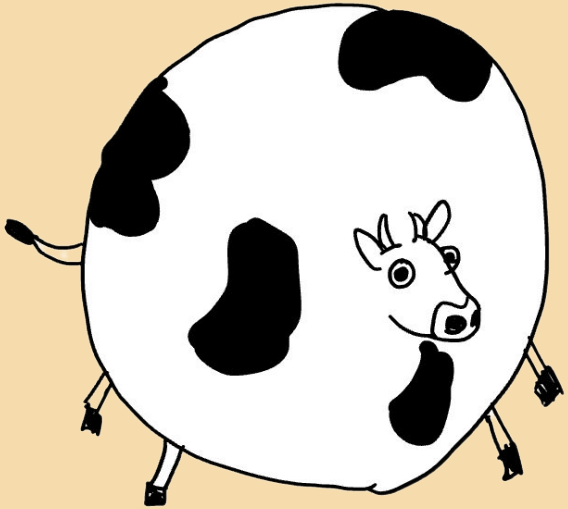
Part 3: Empirical results

Part 4: Conclusion

1. Summary
2. Other remarks
3. Future work

Remark 1 of 3 : *Why do we need spherical cows?*

- Help clarify our thinking
- Separate different things we care about
- Examine confounders, causal factors
- Debug cleanly
- Inspire algorithmic ideas & quick tests



Remark 2 of 3: Some clarifying points on the next-token prediction debate

Pessimists

If humans simply uttered the next-token, we'd be speaking gibberish.

Even tiny next-token errors snowball exponentially:

Pr[all tokens correct]

$= (1-\epsilon) \times (1-\epsilon) \times (1-\epsilon)\ldots$

Optimists

By chain rule of probability, *any* distribution can be represented by next-token prediction (NTP)!

$$\Pr[t_1 t_2 t_3 \ldots] = \Pr[t_1] \times \Pr[t_2 \mid t_1] \times \Pr[t_3 \mid t_1 t_2] \ldots$$

You're just using the NTP backbone incorrectly. Wrap a verifier/backtracker *or do RL*!

The argument goes in circles due to conflated terminology: “next-token prediction” may refer to “autoregressive *inference*” or “next-token *learning*”

Optimist: " Why care about future-token learning if NTP + RL can already (seemingly) plan?"

My answer: If RL only elicits latent skills from base model

$\Rightarrow$ we want to make base model use data efficiently!

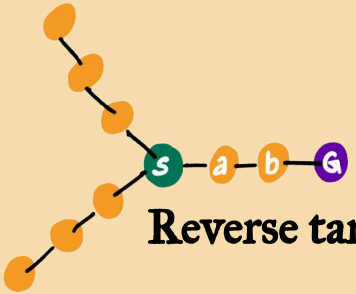
Also: How would one use RL to improve originality?

Remark 3 of 3: There's a belief that next-token learning on a non-left-to-right order suffices. Is this reasonable?

Indeed, prior counterexamples to NTP are solved by NTP upon *reversing* the target tokens

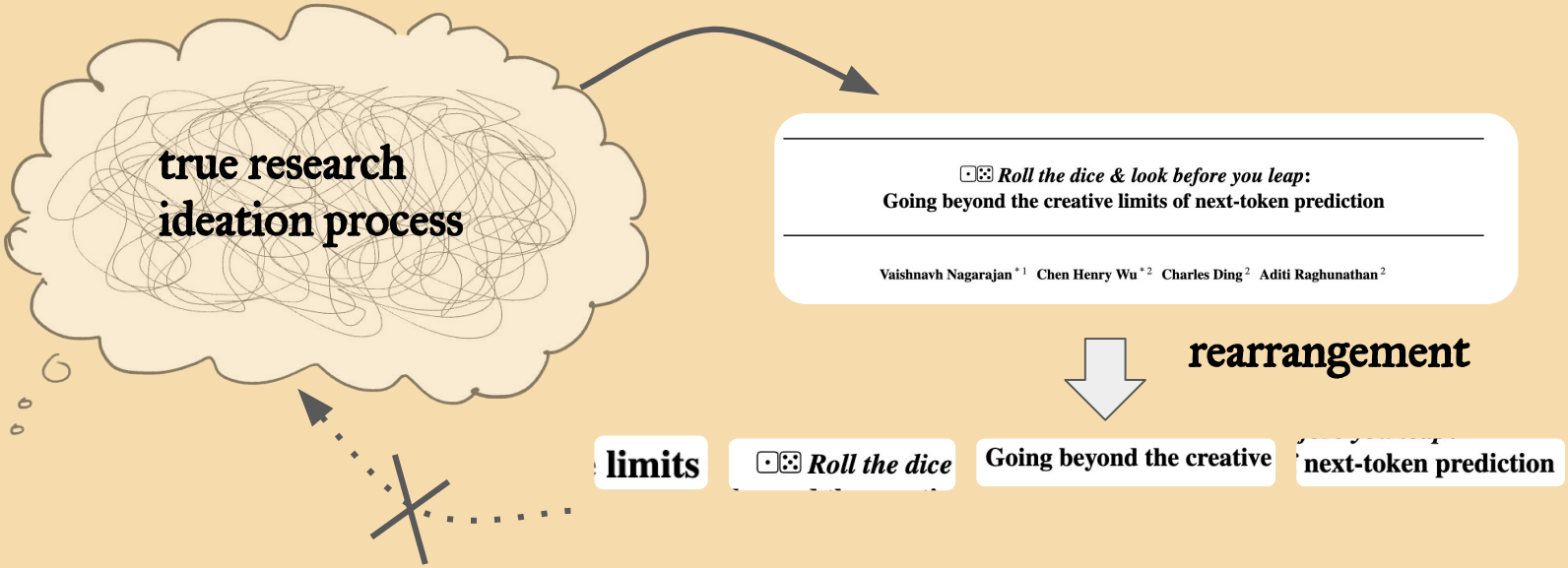
999 + 001

Reverse target : 0001



Reverse target : "goal b a start"

Creative texts have “deep patterns” not visible at the token level

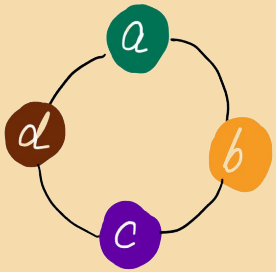


Mere token rearrangement reveals no insight into the generative process!

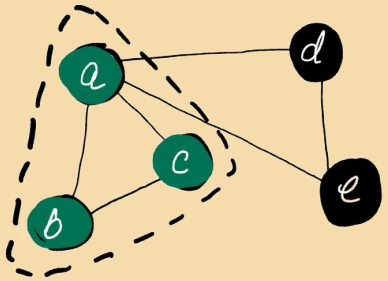
Our tasks minimally capture this “deep pattern”

generate

such that



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



Discover novel triangles in an *in-weights* graph

No token is more privileged; reordering reveals nothing; all tokens need to be learned simultaneously!

Outline

Part 1: Motivation

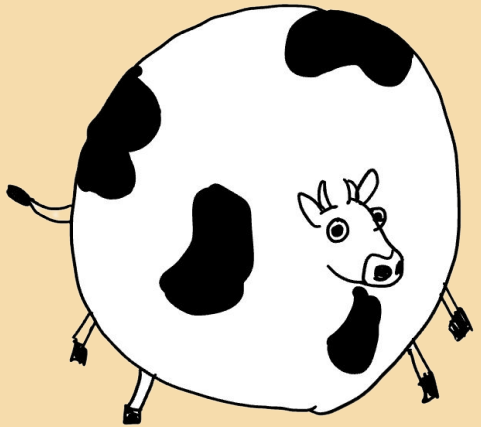
Part 2: Our two types of creative tasks

Part 3: Empirical results

Part 4: Conclusion

1. Summary
2. Other remarks
3. Future work

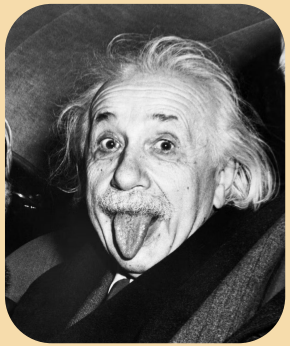
Limitations & Future work



1. Do not use our spherical cows as a sole benchmark: use it for understanding, inspiring new ideas & sniff tests!
    a. Make seed-conditioning work in real-world datasets; how to "learn" the seeds?

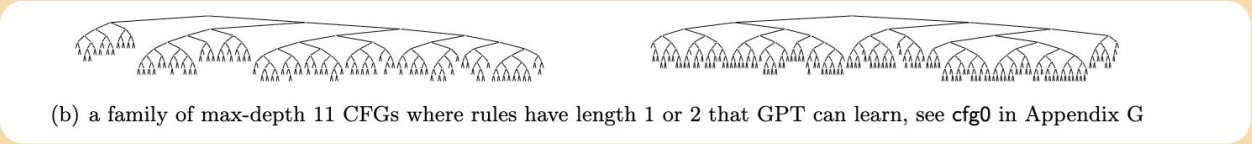
2. Our findings are still not fully characterized e.g., effect of model-size, top-K

3. We do not capture the full richness of creativity
    a. How to think about "transformational creativity"?



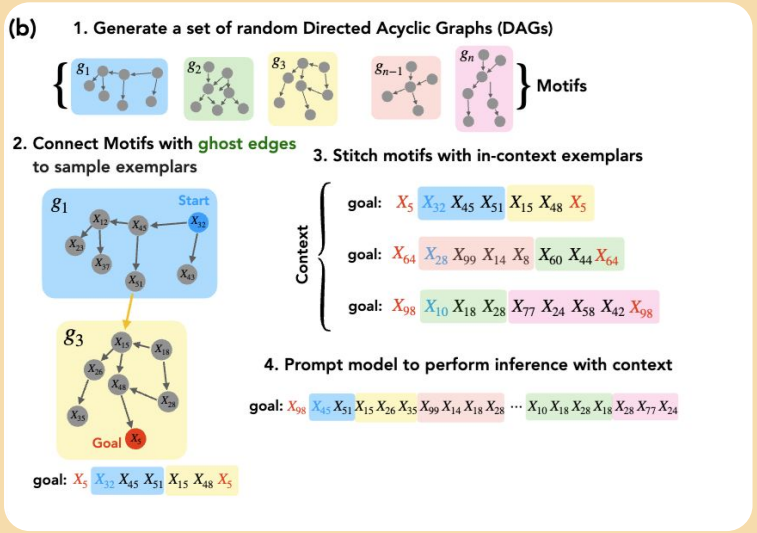
Controlled tasks are valuable!

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



(b) a family of max-depth 11 CFGs where rules have length 1 or 2 that GPT can learn, see cfg0 in Appendix G

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



Thank you!

Chen Wu *, CMU

Charles Ding, CMU

Aditi Raghunathan CMU

Gregor Bachmann*, Apple

Roll the dice & look before you leap:
Going beyond the creative limits of next-token prediction

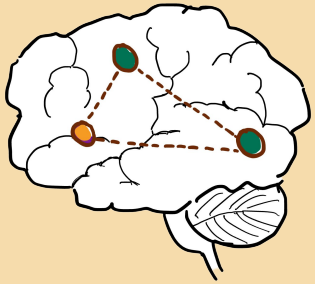
Vaishnavh Nagarajan [*1] Chen Henry Wu [*2] Charles Ding [2] Aditi Raghunathan [2]

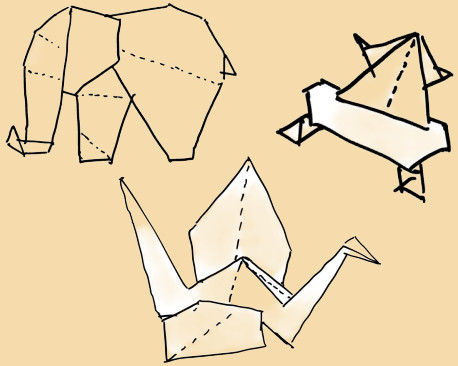
The Pitfalls of Next-Token Prediction

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

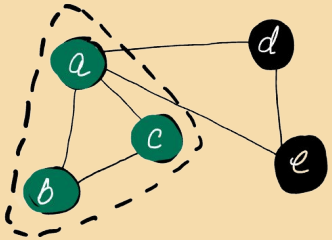
Questions?

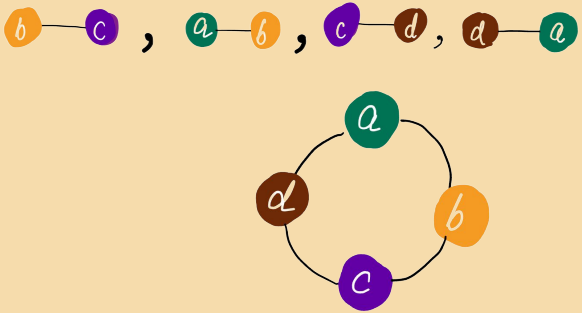
1. Two types of creativity in cognitive science:
   a. combinational (wordplay, analogies)
   b. exploratory (problem design)





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

[all diagrams in this talk are human-drawn]