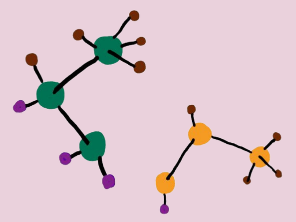
🎲 Roll the dice & look before you leap:

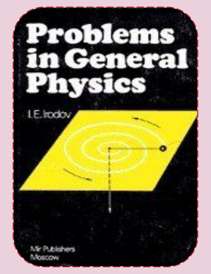
Going beyond the creative limits of next-token prediction

Chen Henry Wu* (CMU), Vaishnavh Nagarajan* (Google Research NY), Charles Ding (CMU), Aditi Raghunathan (CMU)

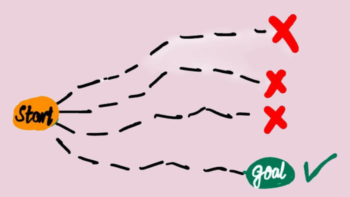
We want LLMs to *creatively* solve *open-ended* tasks



Scientific discovery



Dataset generation

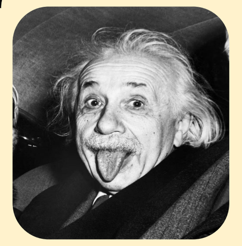


Test-time scaling (best-of-N)

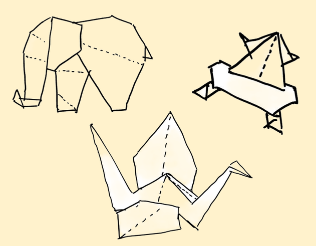
We must not only care about...

Quality of a given generation

but also about:



Originality against training set

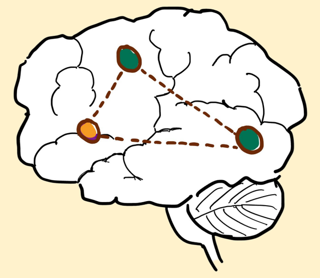


Diversity across generations

Our approach: study *minimal, open-ended* tasks abstracting two modes of creativity

Combinational creativity

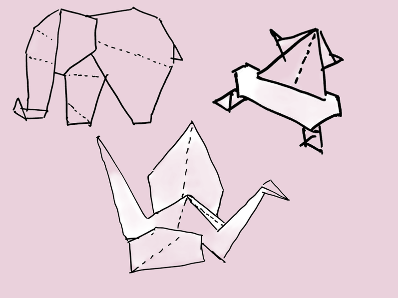
wordplay, analogies, science, discovering contradictions in literature



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

Exploratory creativity

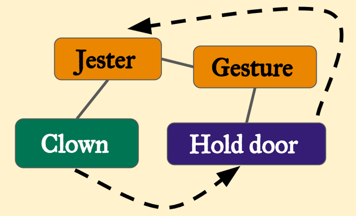
designing problems, deriving corollaries, generating molecules, crafting stories



Plan and devise novel patterns that obey a small set of rules (little to no memory needed)

Combinational creativity as graph tasks

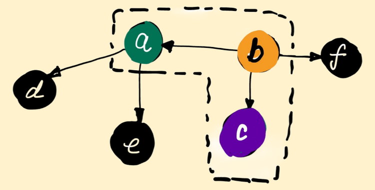
A clown held the door for me. What a nice jester!



Wordplay is a "find a novel path from vocabulary graph" task

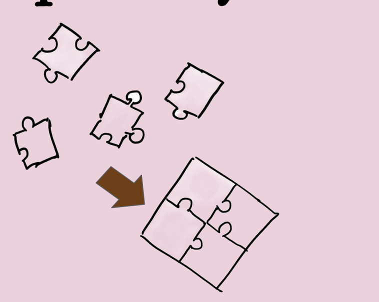
Our task

Generate a c b from



Discover novel sibling-parent triplets in an *in-weights* graph

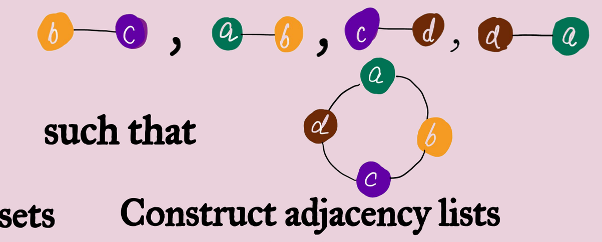
Exploratory creativity as graph tasks



In problems & stories, one sets pieces in conflict s.t. there is a novel resolution under rules

Our task

Generate



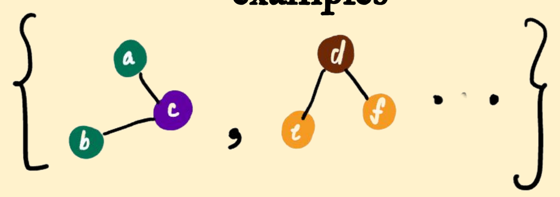
such that

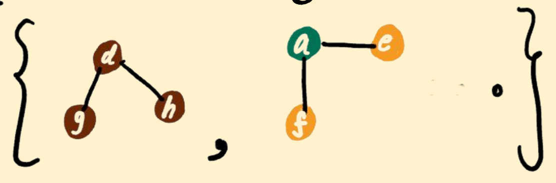
Construct adjacency lists that *resolve* into a circle graph via novel permutations

How we cast these as learning tasks

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

Training set of i.i.d examples → Language model → Independent test-time model generations



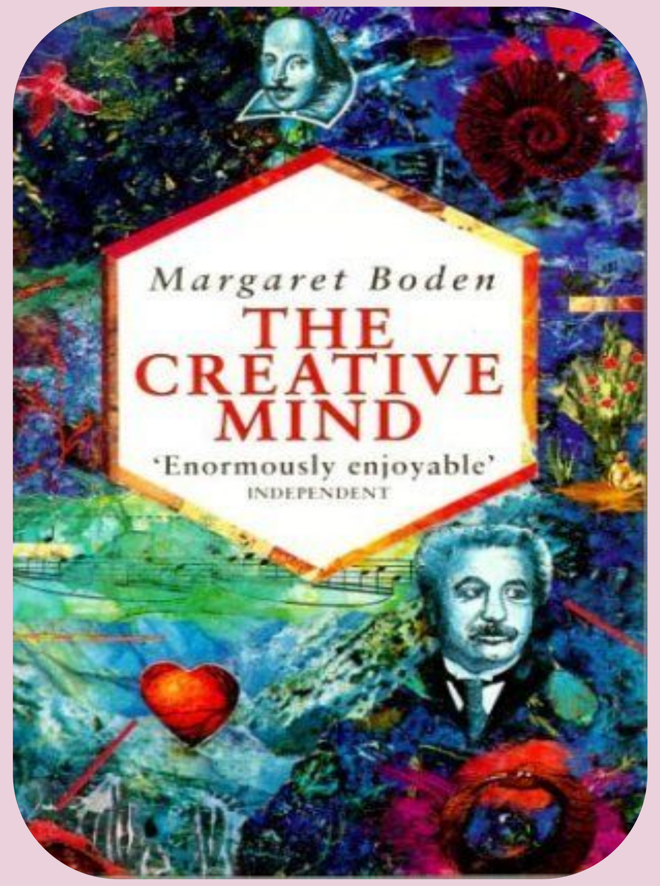


No one unique solution!

No natural language semantics involved — deliberately

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

We quantify this by designing minimal tasks. These tasks abstract two modes of creativity in *cognitive science*.



Paper

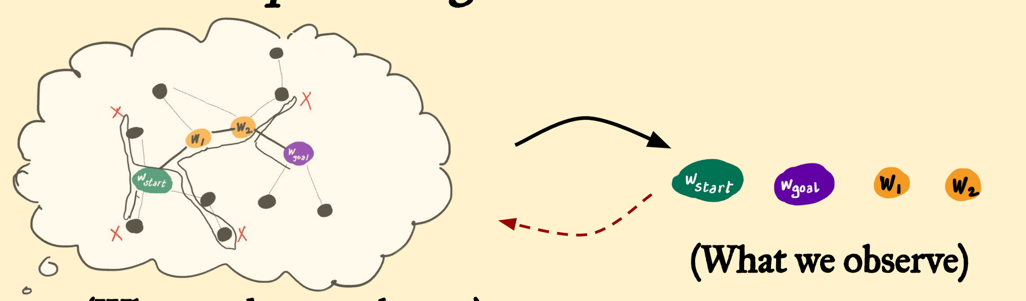
Code

In these tasks, we quantify limits of next-token learning & temperature sampling. We highlight alternatives, multi-token prediction & "seed-conditioning".

Google Research

Carnegie Mellon University School of Computer Science

Creative outputs are generated from an unobserved *leap of thought*



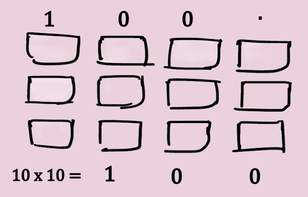
(What we do *not* observe)

(What we observe)

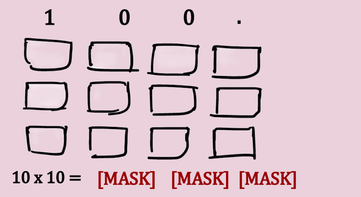
Can "*local*" next-token-learning (NTP) on the creative output infer the "*global*" end-to-end creative process?

We build on Bachmann and Nagarajan '2024's negative result.

Multi-token vs next-token learning

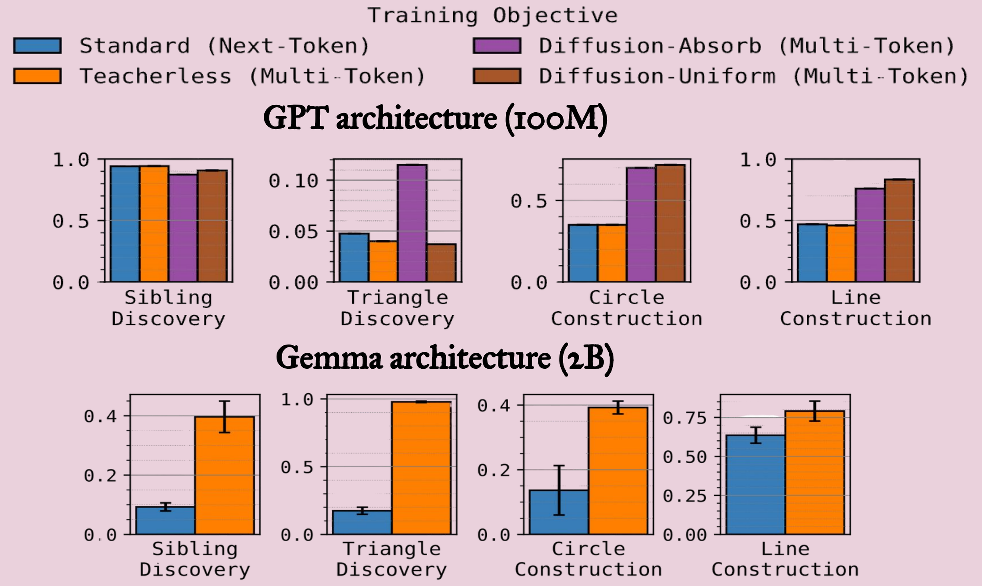


Standard next-token training (aka "teacher-forced")



Teacherless multi-token training (Bachmann and Nagarajan, 2024; Monea et al., 2023; Tschannen et al., 2023)

We also test diffusion models, another multi-token training method (Hoogeboom et al., 2021; Austin et al., 2021; Lou et al., 2023)



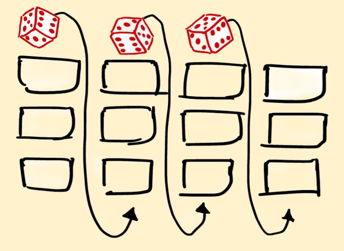
Multi-token tends to be most creative, especially for Gemma 1B, except for GPT sibling (where it is more creative under top-k).

Q: Why bother with alternatives to NTP, when NTP+RL+thinking can plan?

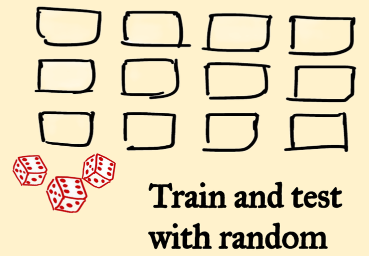
Ans: If RL only elicits existing skills, we must improve how base model picks up creativity from pretraining data!

Eliciting diversity with *seed-conditioning*

Temperature sampling



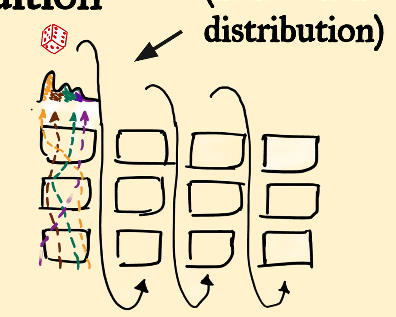
Seed-conditioning



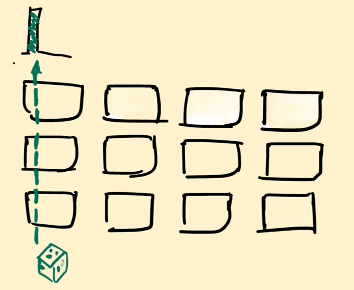
Train and test with random seed prefixes

Intuition

(next-token distribution)

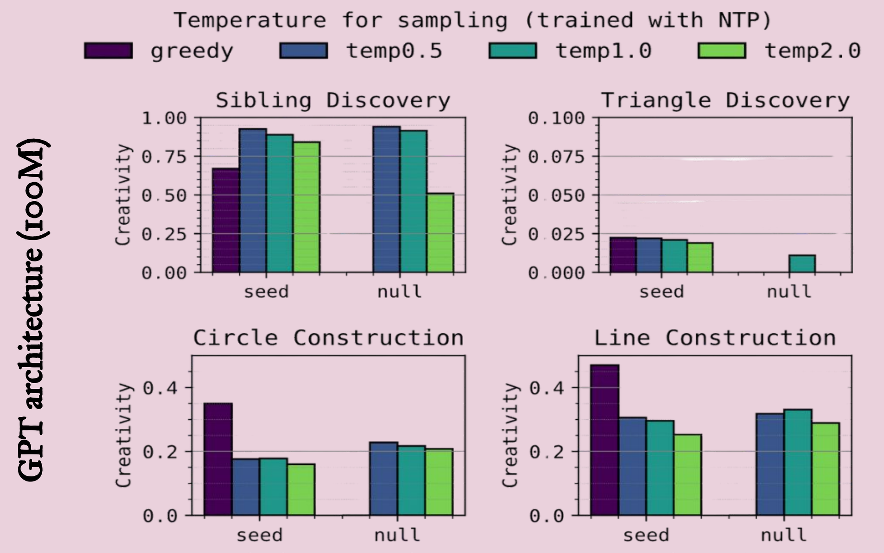


Must process *many thoughts in parallel* to produce diverse next-token distribution



Only needs to *focus on one thought* per seed

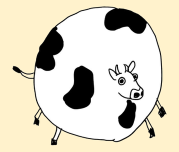
Seed-conditioning vs. temp sampling



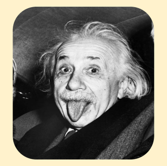
Despite the seeds being arbitrary, remarkably they offer non-trivial creativity even with greedy decoding; comparable or sometimes greater than temp sampling!

Limitations & future work

1. Don't 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



2. We haven't fully characterized or understood all effects in our tasks e.g., effect of model-size



3. How to capture "transformational creativity"?

Disclaimer: No AI was used in drawing the diagrams!