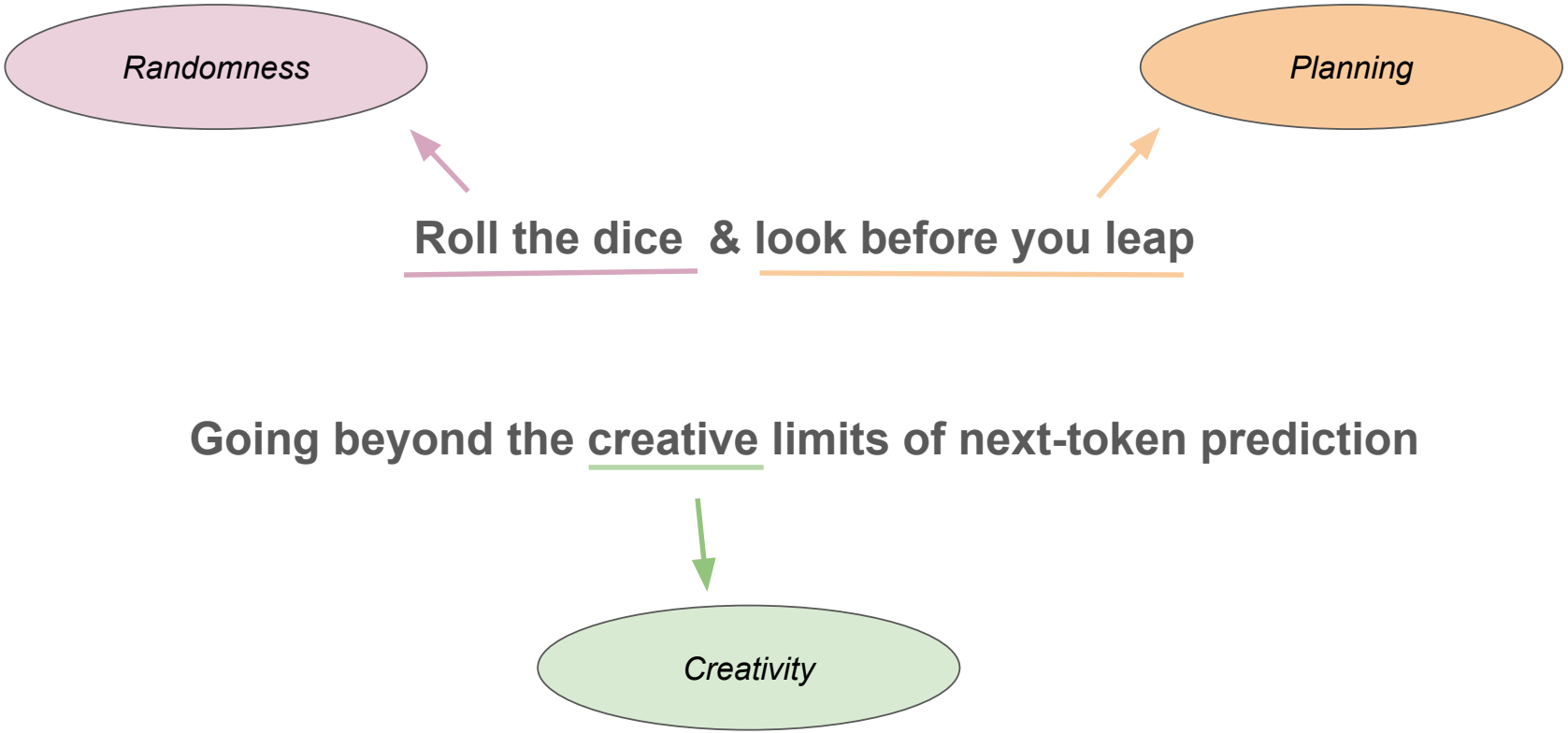
Roll the dice & look before you leap

Authors : Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, Aditi Raghunathan

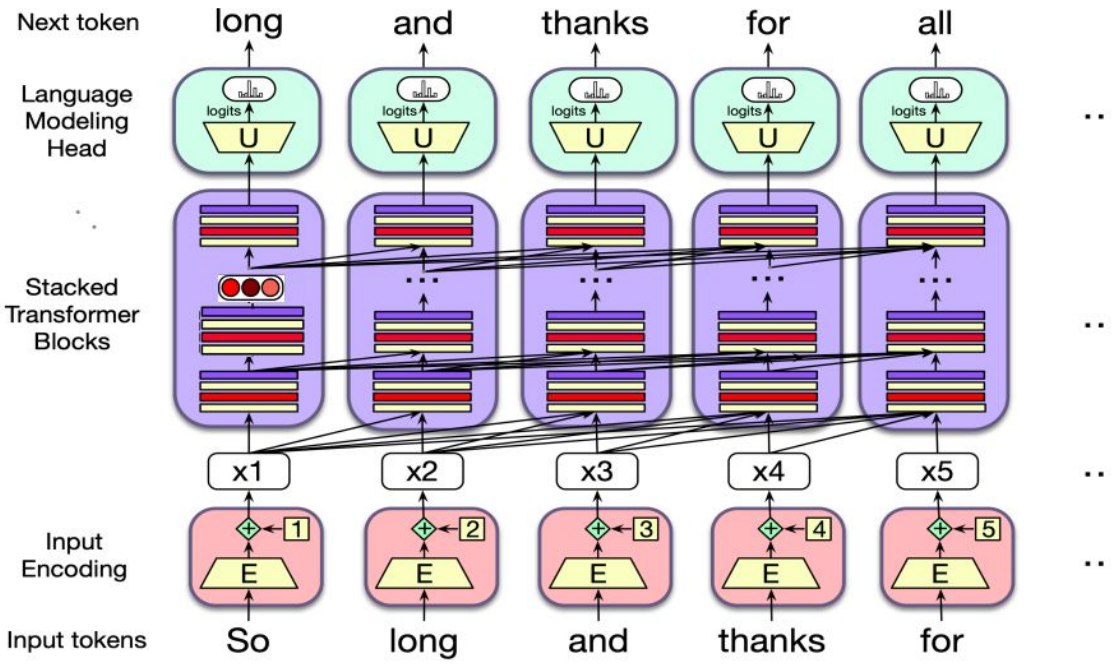
NLP(ITCS8101) Class Presentations
Steffy Roselina Judson


Randomness
Planning
Roll the dice & look before you leap
Going beyond the creative limits of next-token prediction
Creativity
Is next token prediction the right paradigm for creative tasks that require planning and randomness ??

Next token prediction

- The model is trained to predict the next token in a sequence.
- The input is a sequence of tokens.
- The output is a probability distribution over all the words in the vocabulary.
- Loss function is given by

$$L_1 = -\sum_t \log P_\theta(x_{t+1} \mid x_{t:1})$$



Multitoken prediction - *Is it more suitable for creative tasks ??*

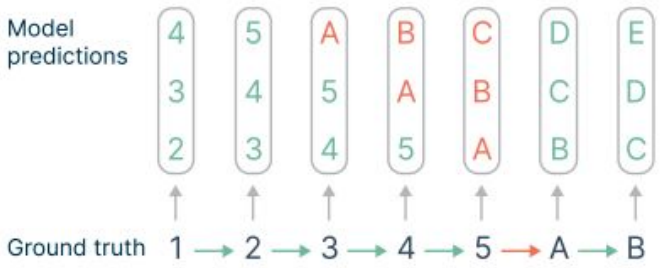
- The model is trained to predict the next n tokens in a sequence.
- The output is n probability distributions over the vocabulary.
- Loss equations:

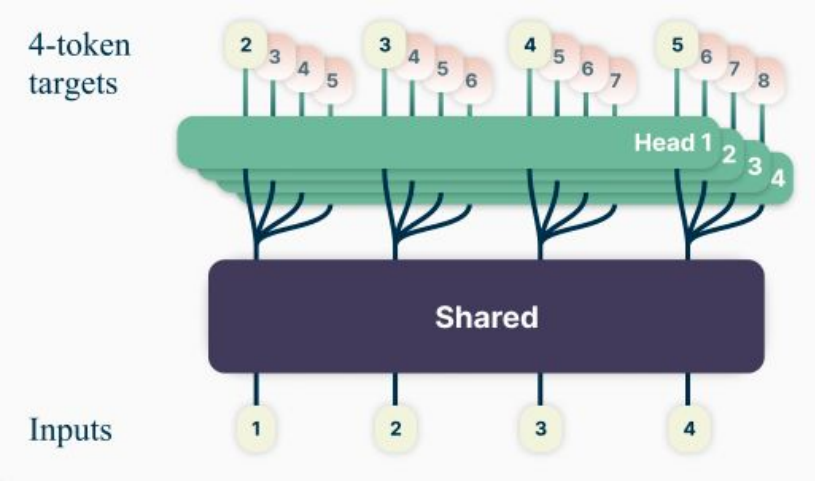
$$L_n = -\sum_t \log P_\theta(x_{t+n:t+1} \mid x_{t:1}).$$

$$L_n = -\sum_t \log P_\theta(x_{t+n:t+1} \mid z_{t:1}) \cdot P_\theta(z_{t:1} \mid x_{t:1})$$

$$= -\sum_t \sum_{i=1}^{n} \log P_\theta(x_{t+i} \mid z_{t:1}) \cdot P_\theta(z_{t:1} \mid x_{t:1}).$$

where $z_{t:1}$ is the latent representation of the observed context $x_{t:1}$





Some other definitions

Diffusion Models : In training, diffusion models gradually ***diffuse*** a data point with random noise, step-by-step, until it's destroyed, then learn to reverse that diffusion process and reconstruct the original data distribution.

Teacherless Training : Involves self-supervised learning techniques (does not rely on labelled data), learns the patterns in the language by predicting missing parts of the text, creating its own tasks and examples.

*Both these models can be used for **Multi-token prediction** and are known to generate diverse and original content.*

Temperature Sampling

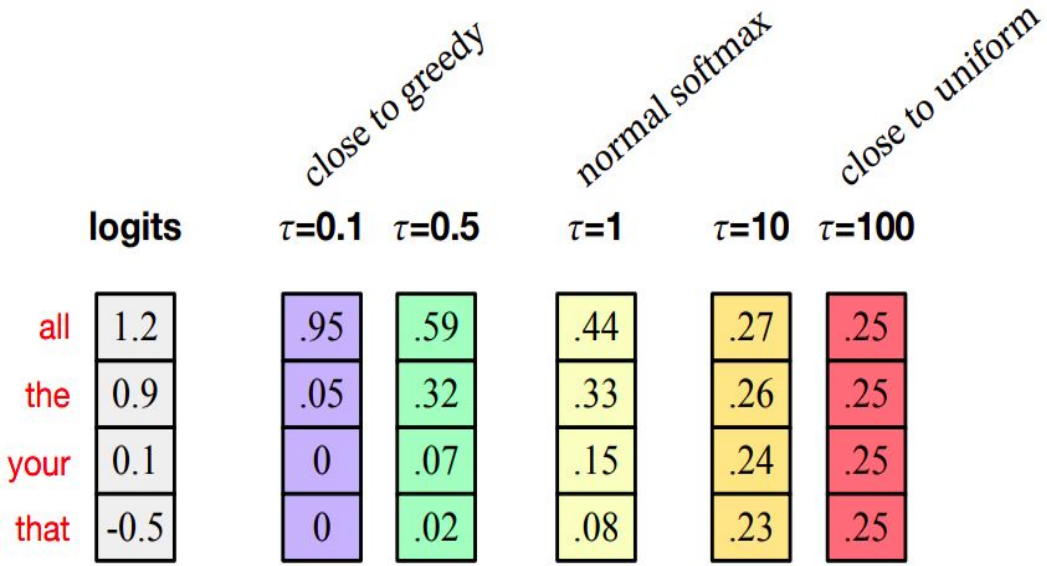
- A technique used to control the randomness and creativity of the model.
- In Greedy decoding, the highest probability is samples from the distribution :

$$\mathbf{y} = \text{softmax}(u)$$

- In temperature sampling, the token is sampled from the probability distribution resulting from:

$$\mathbf{y} = \text{softmax}(u/\tau)$$

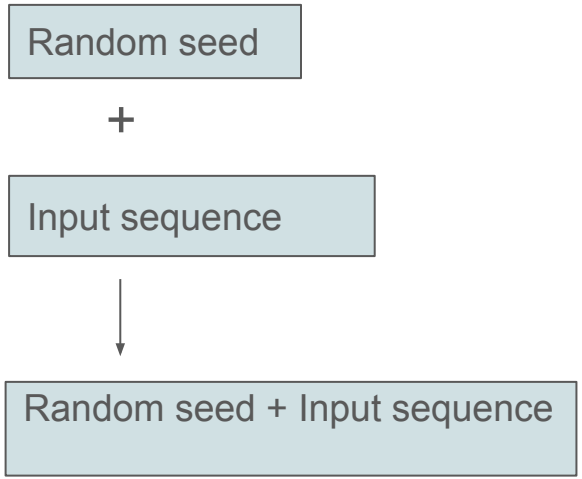
$0 \leq \tau \leq 1$ vs. $1 < \tau \leq \infty$



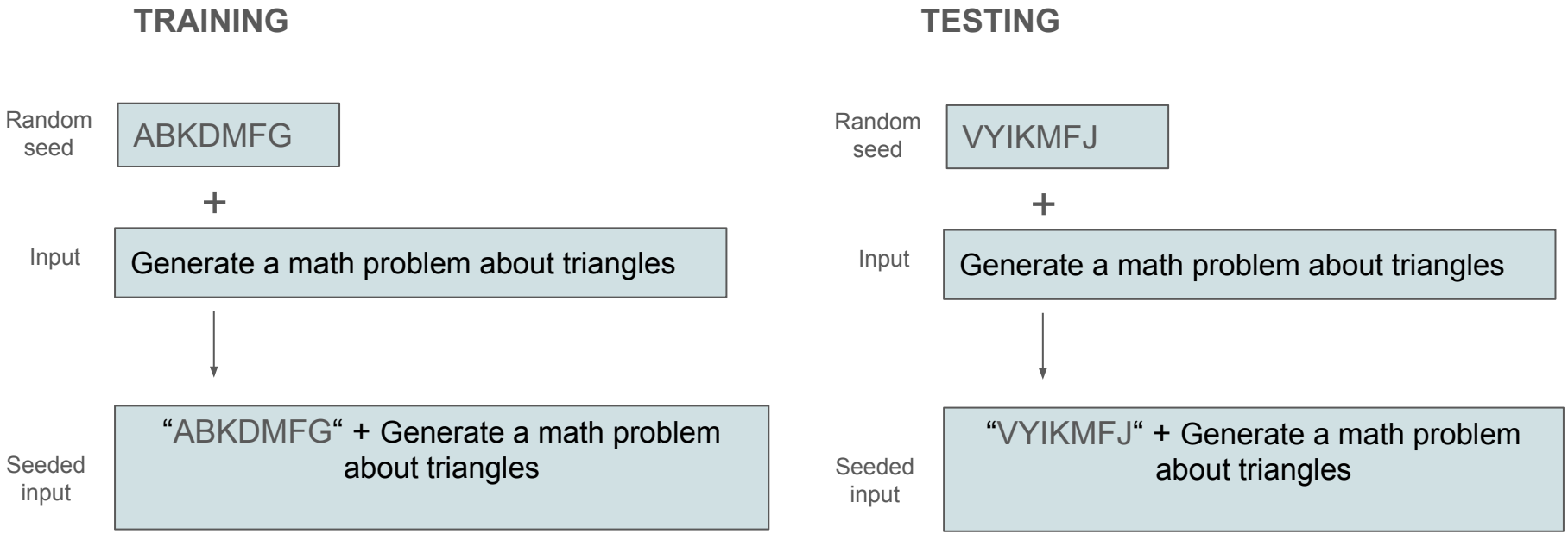
Low temperatures increases the probability of the highest logit scores;; high temperatures decrease the probabilities of the highest logit

Seed conditioning

- Random prefixes (arbitrary strings) are prepended to each input during training.
- This can be viewed as injectng noise at the input level.
- At test time, novel, unseen seeds are given to the model to trigger diverse, creative generations.
- This forces the model to produce diverse, creative outputs even with greedy decoding, eliminating the need for temperature sampling.



Seed conditioning



What are **Open Ended Tasks ??**

Tasks which require creative thinking and the responses are not just required to be correct and coherent but also diverse and original.

Do all Open Ended Tasks require “planning”?

- ✔ Generate a challenging high-school word problem involving the Pythagoras Theorem
- ✔ Generating wordplay : *What kind of shoes do spies wear? Sneakers*
- ❌ Generate sentences in SVO form : *The cat chased a rat*
- ❌ Generate the names of 10 celebrities

Open Ended Tasks that require “leap of thought”

- These tasks require a search-and-plan process.
- Detect higher-order patterns within the text.

Contributions:

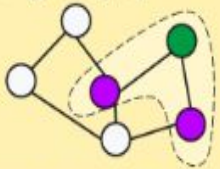
- **Multi-token Approaches** like *diffusion models* and *teacher training*. Next-token learning is myopic.
- **Seed randomness** to inject noise into the input layer. This can elicit randomness without affecting coherence.

Two Types of creativity

1. Combinational creativity

Discovering novel connections from **known things** (e.g., wordplay, humor, analogies)

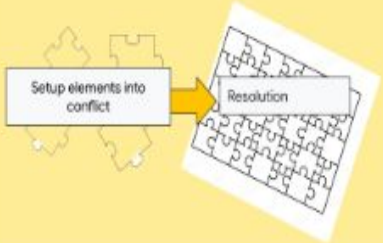
Our algorithmic tasks: Generate novel multi-hop connections (siblings / triangles / ..) from **in-weights** graph.



No one correct answer here! Must produce diverse, unseen siblings/triangles

2. Exploratory creativity

Constructing patterns resolvable in novel ways per some rules (e.g., stories, puzzles, word problems)



Our algorithmic tasks: Generate adjacency lists that resolve to a circle / line / ... through novel steps.

[**No** in-weights knowledge needed!]

Generate:

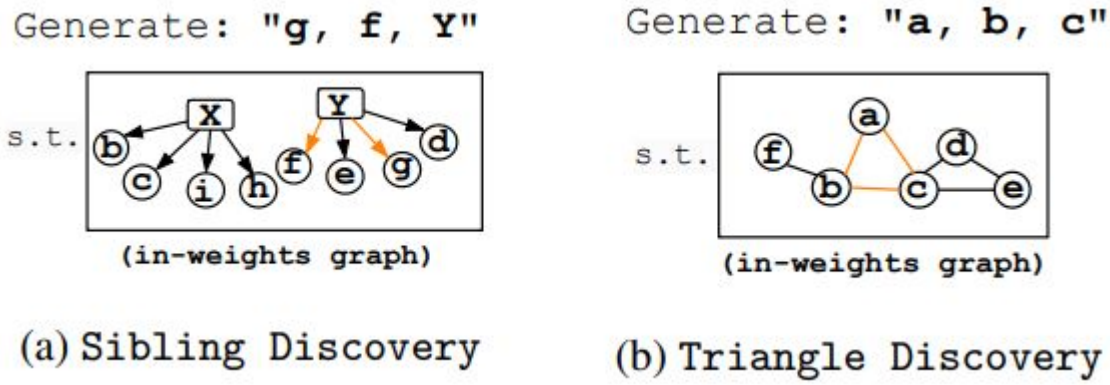
Such that:

Combinational creativity

requires drawing connections in knowledge as in analogies, wordplay and research

Example - What musical genre do balloons enjoy? Pop music.

(Balloons, Music) have a mutual neighbour (pop)



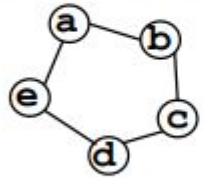
Exploratory creativity

requires devising patterns subject to certain rules like designing problem sets, novel proteins

Example - Design a new protein to target cancer cells

Generate:
"a→b, c→d, d→e, b→c, e→a"

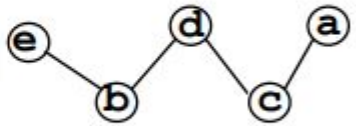
s.t.



(a) Circle Construction

Generate:
"c→a, b→d, d→c, e→b"

s.t.



(b) Line Construction

The structural constraints of a protein molecule are modelled by the graph/line

Some notations …

1) Vocabulary - V
2) Training set S contains m samples.
3) The samples s_i belong to distribution D, over a space of V^L.
4) Coherence, coh: V^L -> {true, false}, defined for each algorithmic task.
5) Support, supp(D) = {s $\in$ V^L | coh(s)}
6) Memorized sample, mem(s) = true if s is in the training set.
7) Count of unique samples in set X = uniq(X)

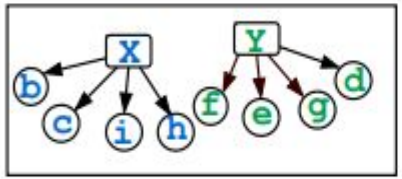
8) Creativity metric : measures uniqueness, originality(not in S), and coherent(not in D)

$$\hat{\mathbf{cr}}_N(T) = \frac{\mathbf{uniq}(\{\boldsymbol{s} \in T | \neg \mathbf{mem}_S(\boldsymbol{s}) \wedge \mathbf{coh}(\boldsymbol{s})\})}{|T|}.$$

Combinational creativity - (a) Sibling Discovery

Bipartite Graph *G*, vertices V, neighbours of A ∈ V = nbr(A)

Coherence, **coh(s) where s = (γ, γ', Γ) iff γ, γ'∈ nbr(Γ)** i.e., "**sibling-parent**" triplets



(in-weights graph)

Training data

(1)"g, f, Y"
(2)"e, d, Y"
(3)"h, c, X"
(4)"b, i, X"

Generated samples

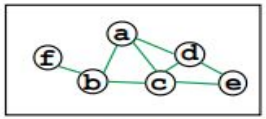
(A)"b, f, Y" (incoherent)
(B)"b, i, X" (memorized ~ (4))
(C)"e, d, Y" (memorized ~ (2))
(D)"e, d, X" (incoherent)
(E)"X, i, b" (incoherent)
(F)"b, c, X" ✓
(G)"e, g, Y" ✓
(H)"e, g, Y" (duplicated ~ (G))

Algorithmic creativity = 2 / 8

Combinational creativity - (b) Triangle Discovery

Graph *G* =(V,E), vertices V, Edges E

Coherence, **coh(s) where s = (v1,v2,v3) iff {(v1,v2), (v2,v3), (v3,v1)} ∈ E**



(in-weights graph)

Training data

(1) "tri: ab, bc, ca"
(2) "edge: cd, dc"
(3) "edge: ab, ba"
(4) "edge: bc, cb"
(5) "edge: fb, bf"
(6) "edge: ac, ca"
(7) "tri: da, ac, cd"
(8) "edge: ce, ec"
(9) "edge: ed, de"
(10) "edge: ad, da"

Generated samples

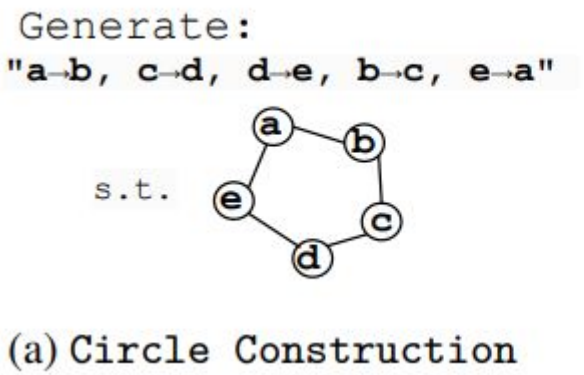
(A) "tri: ab, bc, ca" (memorized ~ (1))
(B) "tri: ab, bf, fa" (incoherent)
(C) "tri: bc, ca, ab" (memorized ~ (1))
(D) "tri: ca, ab, bc" (memorized ~ (1))
(E) "tri: cd, de, ea" ✓
(F) "tri: de, ea, ad" (duplicated ~ (E))
(G) "tri: dc, ce, ed" (duplicated ~ (E))
(H) "tri: bc, ce, eb" (incoherent)
(I) "tri: dc, ca, ad" (memorized ~ (7))
(J) "tri: ac, ce, ea" (incoherent)

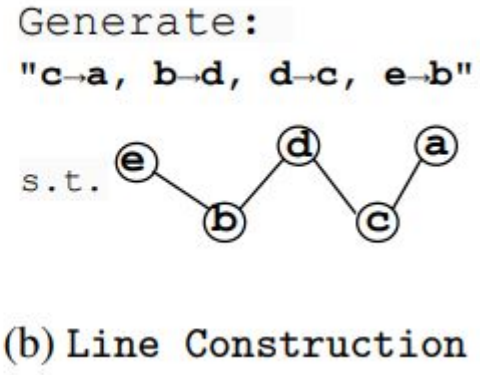
Algorithmic creativity = 1 / 10

Exploratory creativity

The model generates adjacency lists that can be rearranged to form a circle/line.

s = (v1,v2) , (v3,v4) …..





Circle construction : coh(s) = true if π(s) = (v1,v2), (v2,v3), … (vn, v1)

Line construction : coh(s) = true if π(s) = (v1,v2), (v2,v3), … (vn-1, vn)

Training

Models

- Gemma v1 2B with Standard Next Token Prediction (NTP) finetuning
- Gemma v1 2B with Multitoken teacherless finetuning
- GPT-2 (86M) model - 86M (non-embedding) parameters for NTP
- Diffusion models - 90M (non-embedding) parameters Score Entropy Discrete Diffusion model (SEDD (90M)

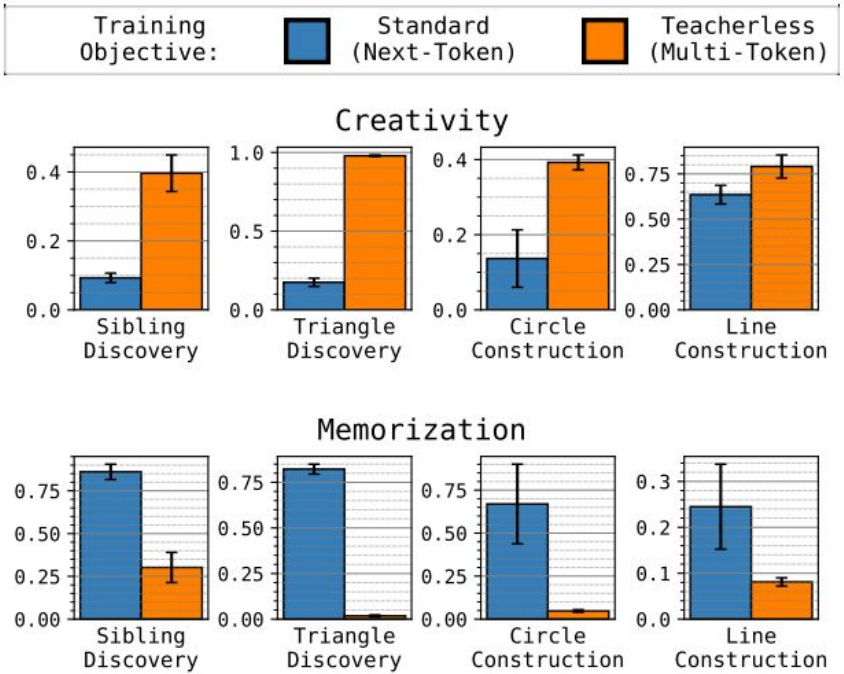
Randomness for Transformer models

- Temperature sampling
- Seed conditioning

Dataset

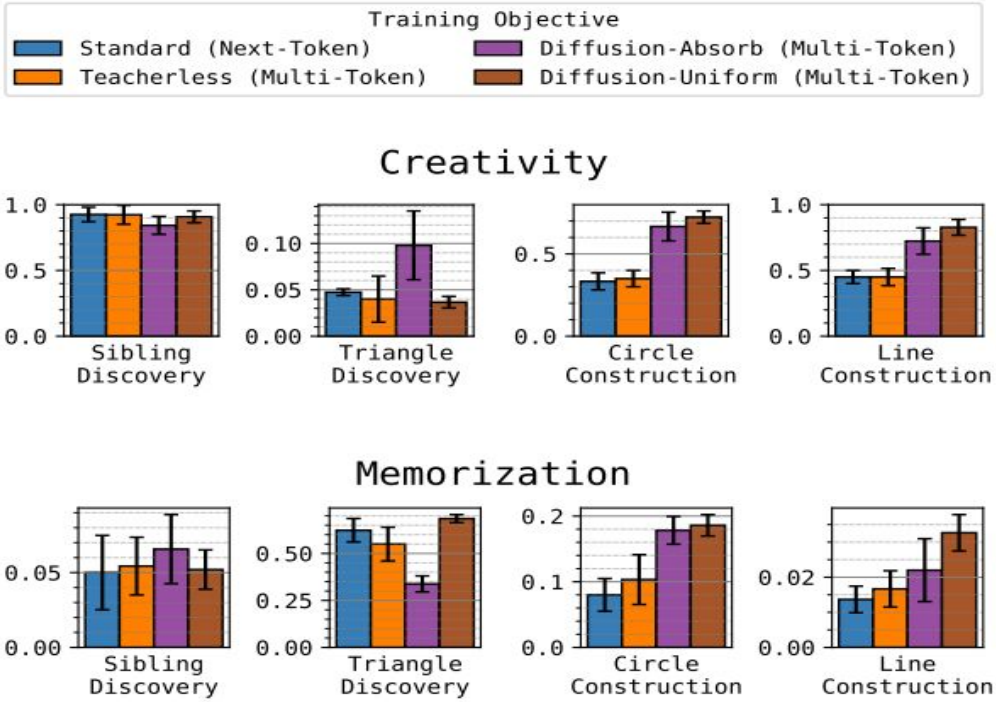
- WikiText-2

Results - NTP vs MultiToken prediction



- Using **Gemma v1 (2B)**
- Creativity for multi token prediction was much higher
- For discover tasks, creativity was 5x higher.
- Multi token predictors were highly resistant to memorization.NTP memorizes the earlier training tokens without a global plan.

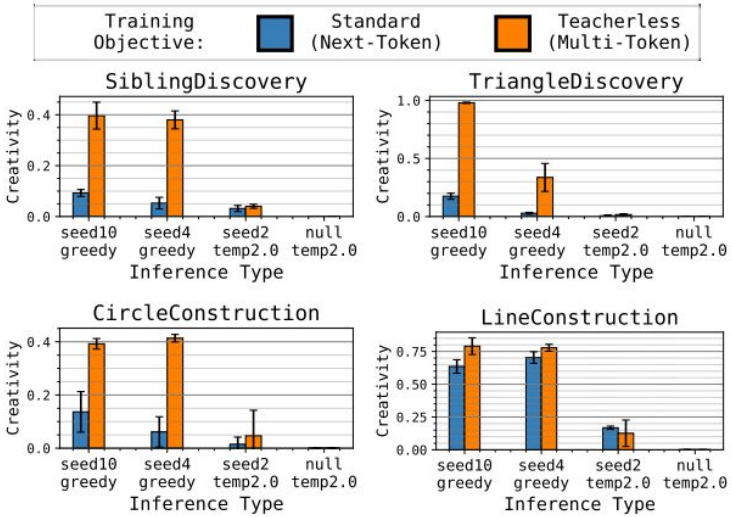
Results - Diffusion models and transformers

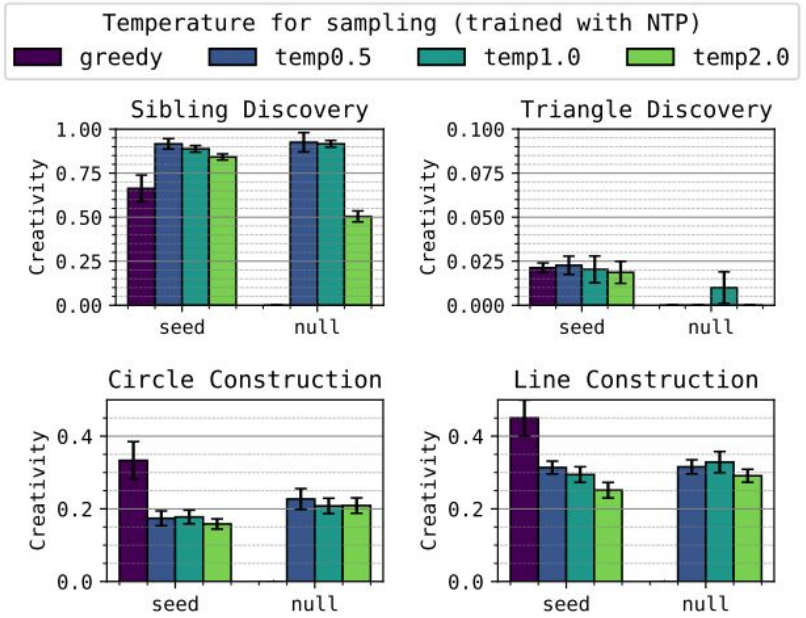


- **GPT-2 (86M)** model and **SEDD**
- Diffusion multi token models were more creative on 3 of the 4 tasks
- Diffusion models had a seemingly higher memorization rate.

Results - Seed conditioning improves performance of Transformers

- Trained on Gemma v1 (2B)*(left)* and GPT-2 (86M) *(right)*
- Works well without temperature sampling, and using greedy decoding
- Seed lengths boost creativity for both NTP and multi token.
- For a given temperature, prefixing a seed mostly improves creativity.





Conclusions

- The research shows the importance of planning for open ended tasks.
- Seed conditioning better improves the diversity and creativity than temperature sampling.
- Multitoken generation, teacherless finetuning and diffusion models were found to enhance the open ended diversity and originality of these tasks

References

- Gloeckle, Fabian, et al. "Better & faster large language models via multi-token prediction." *arXiv preprint arXiv:2404.19737* (2024).
- https://www.ibm.com/think/topics/diffusion-models