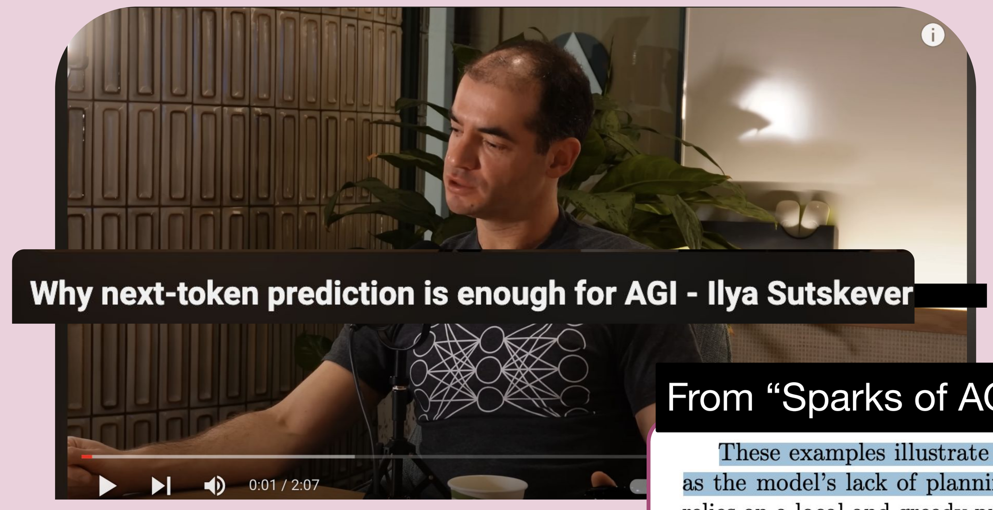
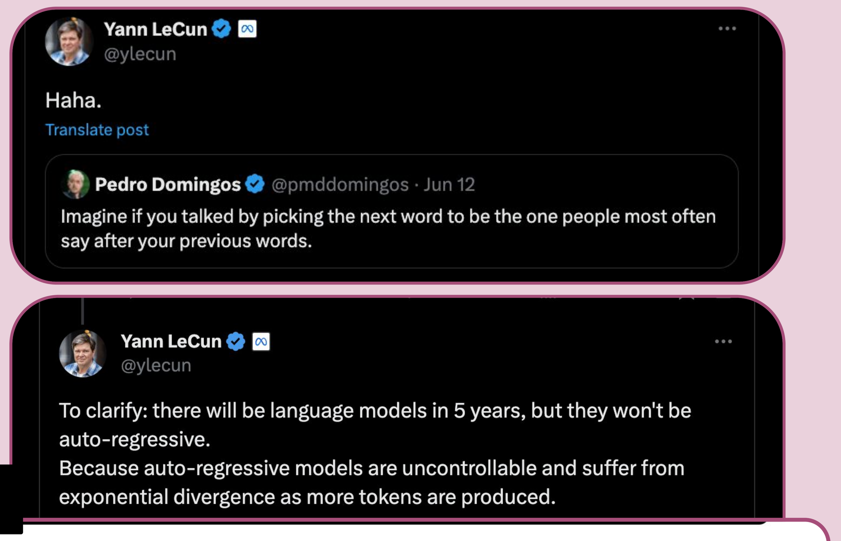
The Pitfalls of Next-Token Prediction (and how to fix them with multi-token prediction)

Gregor Bachmann* (ETH Zürich) & Vaishnavh Nagarajan* (Google Research NY)

The debate: Can NTP can model the way humans think?




Yann LeCun @ylecun

Haha.

Translate post

Pedro Domingos @pmddomingos · Jun 12

Imagine if you talked by picking the next word to be the one people most often say after your previous words.

Yann LeCun @ylecun

To clarify: there will be language models in 5 years, but they won't be auto-regressive.

Because auto-regressive models are uncontrollable and suffer from exponential divergence as more tokens are produced.

From "Sparks of AGI":

These examples illustrate some of the limitations of the next-word prediction paradigm, which manifest as the model's lack of planning, working memory, ability to backtrack, and reasoning abilities. The model relies on a local and greedy process of generating the next word, without any global or deep understanding of the task or the output. Thus, the model is good at producing fluent and coherent texts, but has limitations

Fallacy 1: "If humans just said the next token, they'd be speaking gibberish"

Wrong! Contradicts chain rule of probability!

Fact: Given full context, sampling from next-token probabilities = sampling from joint distribution.

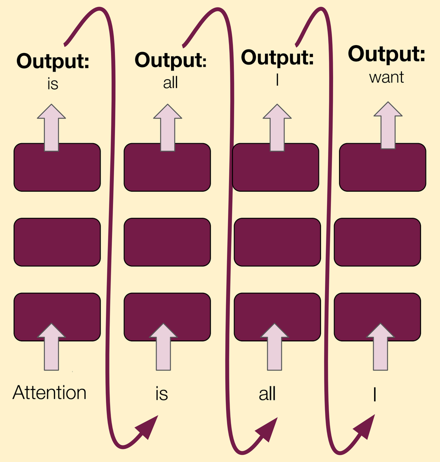
$$\Pr[\mathbf{Y} = (y_1,y_2,y_3,y_4,\dots)]$$
$$= \Pr[Y_1 = y_1] \times$$
$$\Pr[Y_2 = y_2 \mid \mathbf{Y}_{<1} = (y_1)] \times$$
$$\Pr[Y_3 = y_3 \mid \mathbf{Y}_{<2} = (y_1,y_2)] \times \dots$$

Fallacy 2: "The *core* issue is that next-token errors are compounding"

"Assume $\widehat{\Pr}$[**next token ✓** | **context**] $\approx 1-\epsilon$, then $\widehat{\Pr}$[**all tokens ✓**] $\approx (1-\epsilon)^{\#\text{tokens}}$"

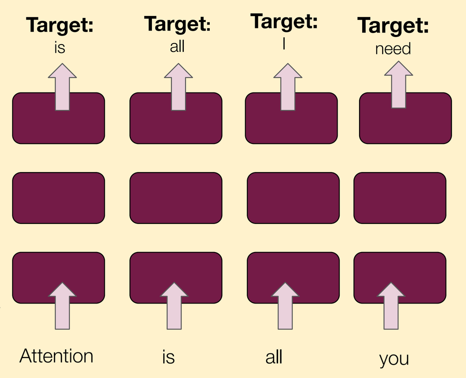
Only an inference-time issue, potentially solvable by post-hoc backtracking etc.,

Fallacy 3: Conflating the two phases of next-token prediction



Autoregression during inference = representing with NTP

≠ have orthogonal issues!



Teacher-forcing during training = learning with NTP

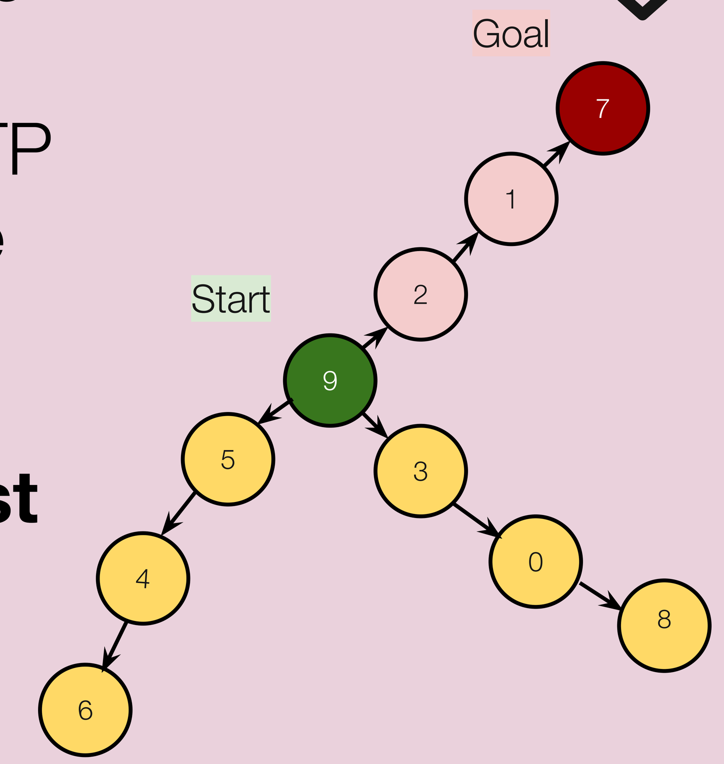
A next-token predictor can *represent* any sequence. But can it *learn* any sequence efficiently?

Is it really true that:

$\widehat{\Pr}$ [**next token ✓** | **context**] $\approx 1-\epsilon$?

Our Contributions:

- We point out popular fallacies in the next-token prediction (NTP) debate
- We argue that the core issue lies in **learning** with NTP — not **inference** with NTP.
- We provide **a first clear counter example** of NTP-learning failure.
- We propose an extremely simple **multi-token prediction** alternative.



Path-finding task which requires lookahead

ETH zürich

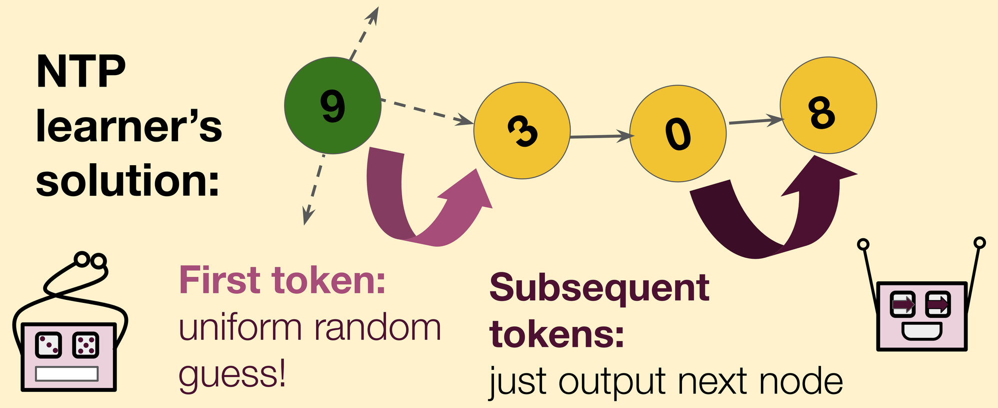
Google Research

A *straightforward* task where next-token learning fails (in-distribution!)

Path-finding on a path-star graph

PROBLEM PREFIX: Randomized adjacency list	2→1, 0→10, 9 → 3, 5 → 4, 4→6, 3→0, 1→7, 9→2,9→5 || find(9 → 7) ?
GROUND TRUTH SOLUTION: Start-Goal path	9 →2→1→7

NTP learner's solution:

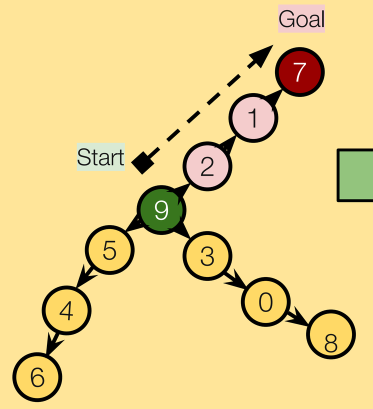


First token: uniform random guess!

Subsequent tokens: just output next node

Why does next-token learning fail?

Ideally, learn (problem ⇒ solution) mapping

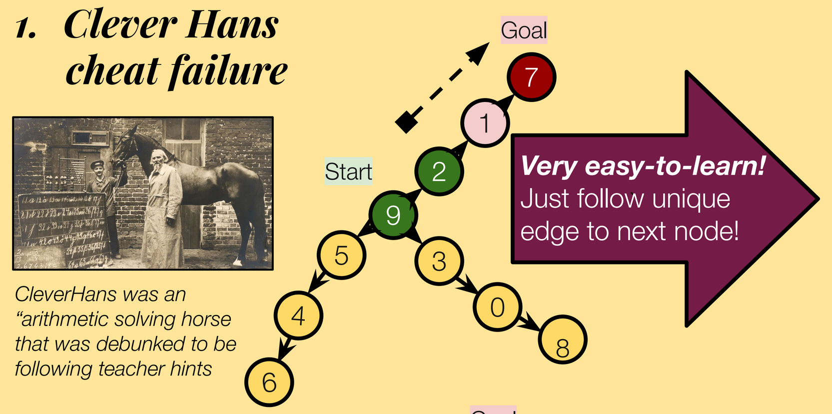


9 →2→1→7

Model needs to learn to plan ahead.

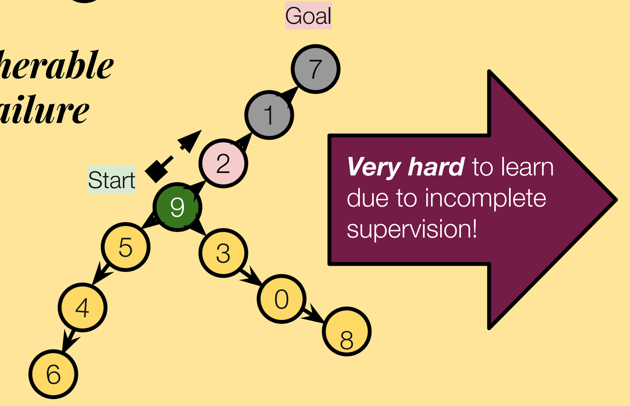
But next-token learning breaks this into *unequivalent* sub-problems!

1. ***Clever Hans cheat failure***



2→1→7

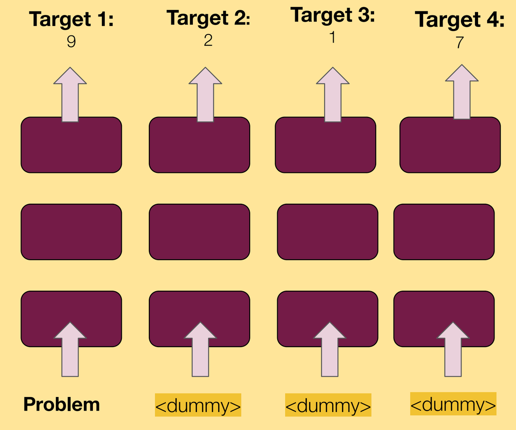
2. ***Indecipherable Token failure***



9→2

With supervision from later tokens lost to easy cheat, learning earlier tokens becomes inefficient in data/computation — **even impossible to learn.**

A simple multi-token objective



Teacherless training

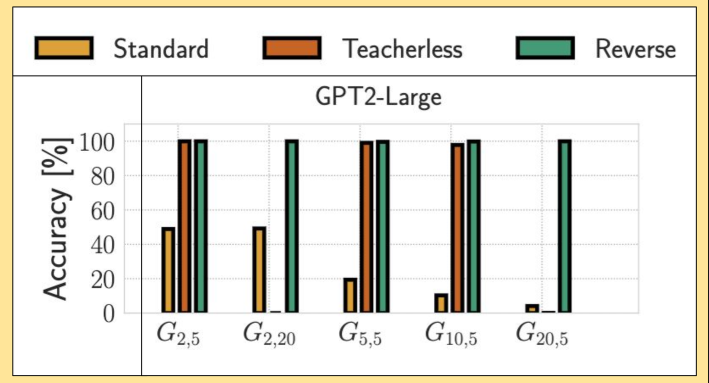
(Also see Monea et al., 23)

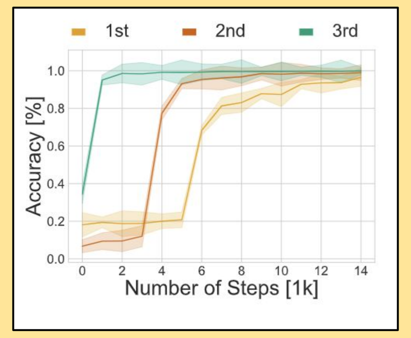
Simply replace the input-side answer tokens with dummy!

Target i can only see upto input i-k rather than upto input i-1!

Multi-token is able to learn where next-token cannot!

See paper for Mamba and GPT-mini.





Intuition: multi-token learner learns tokens in "correct" chronological order instead of left-to-right.

What's next?

We speculate this failure must occur in "lookahead" tasks e.g., poems or in story-writing where text is in "non-chronological order".

The debate needs more rigor and care. When we say "next-token prediction can't model human speech", our gut-feeling refers to limitations of next-token ***learning***.

Need more dedicated efforts going beyond next-token prediction!