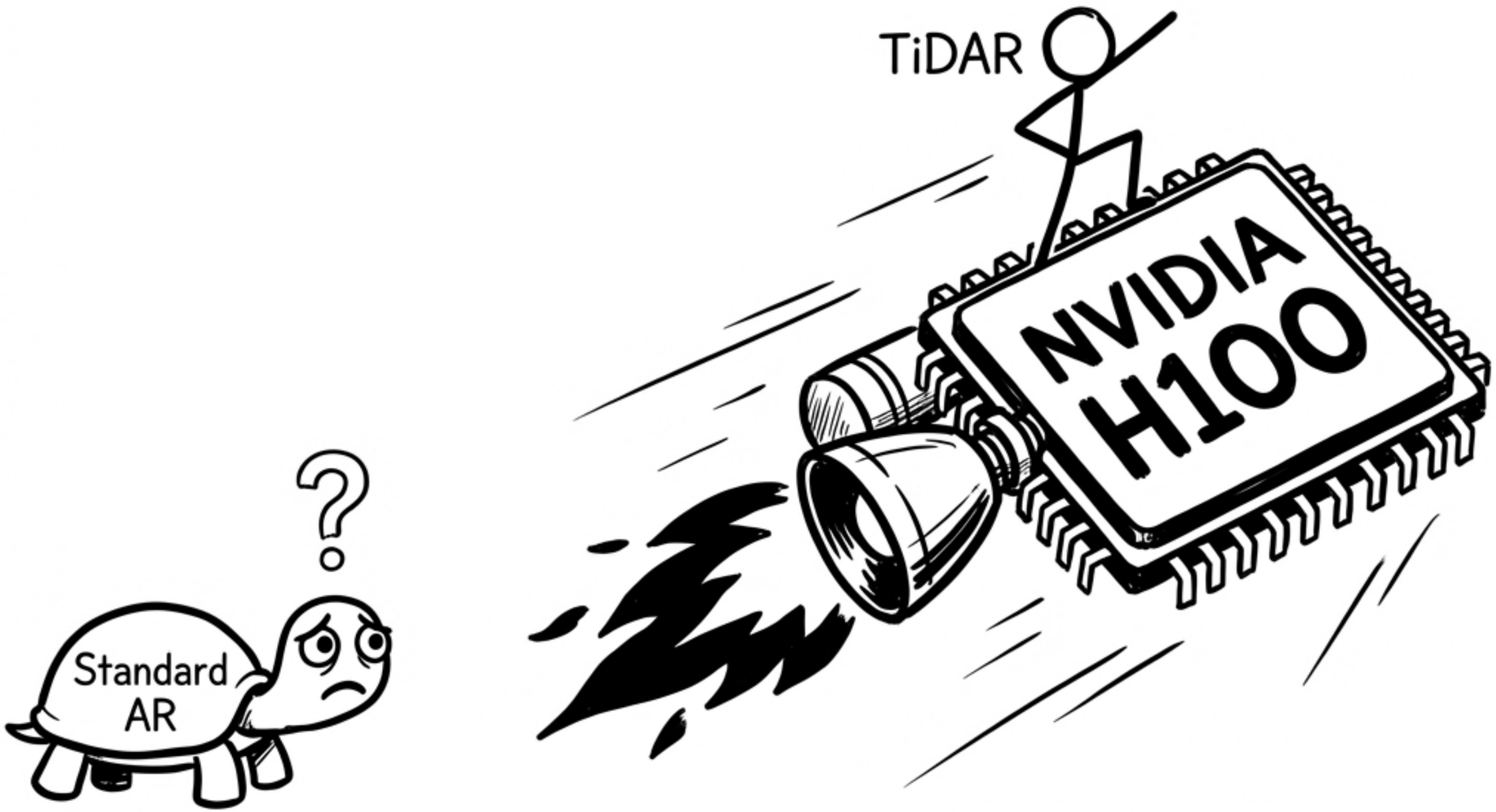
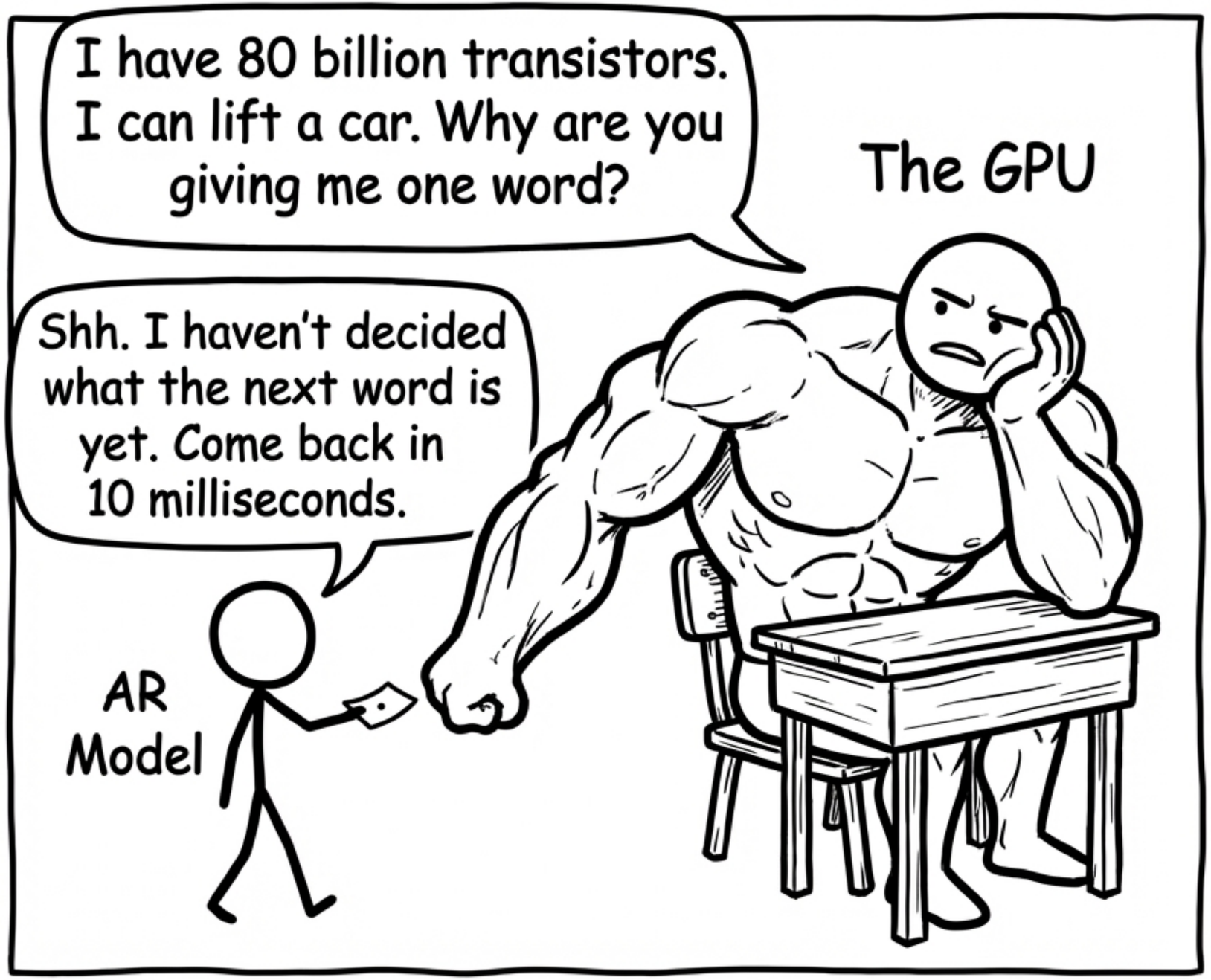
TiDAR: Think in Diffusion, Talk in Autoregression

Or: How I Learned to Stop Worrying and Love the Forward Pass.



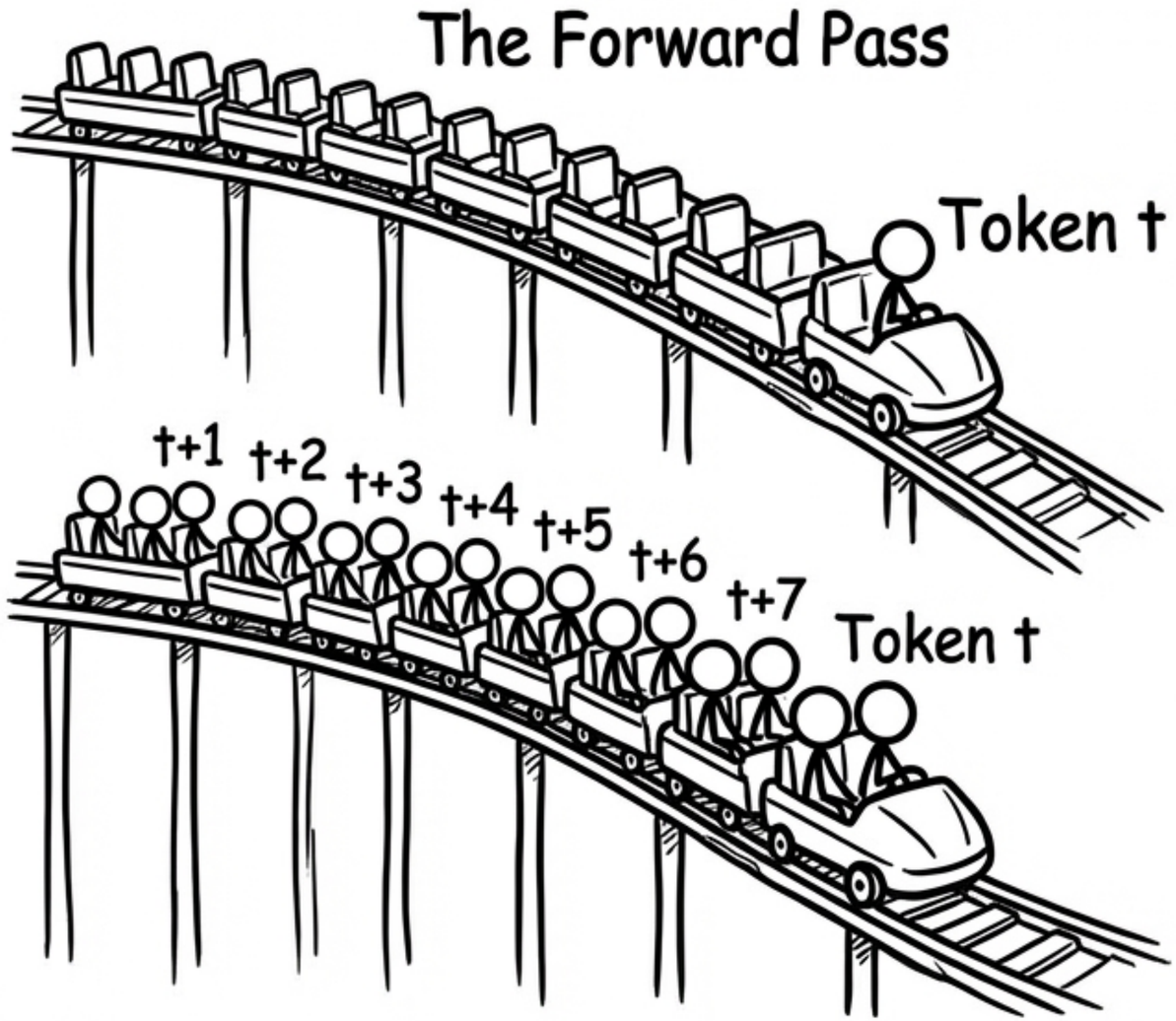
NVIDIA

Based on research by NVIDIA (Liu, Dong, et al.).

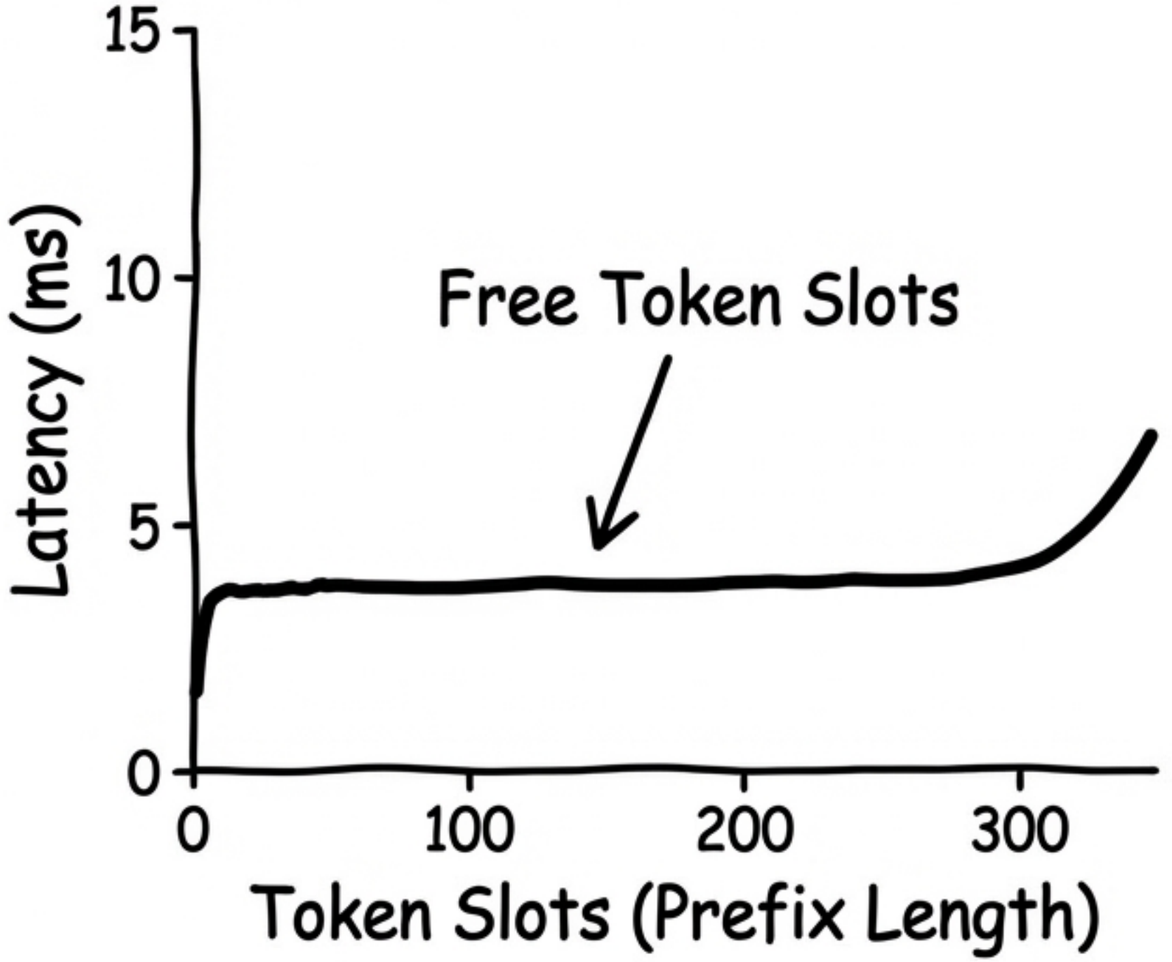


THE PROBLEM:

- LLM decoding is Memory Bound, not Compute Bound.
- Loading model weights takes longer than the actual math.
- We fetch the entire brain just to think of one syllable.

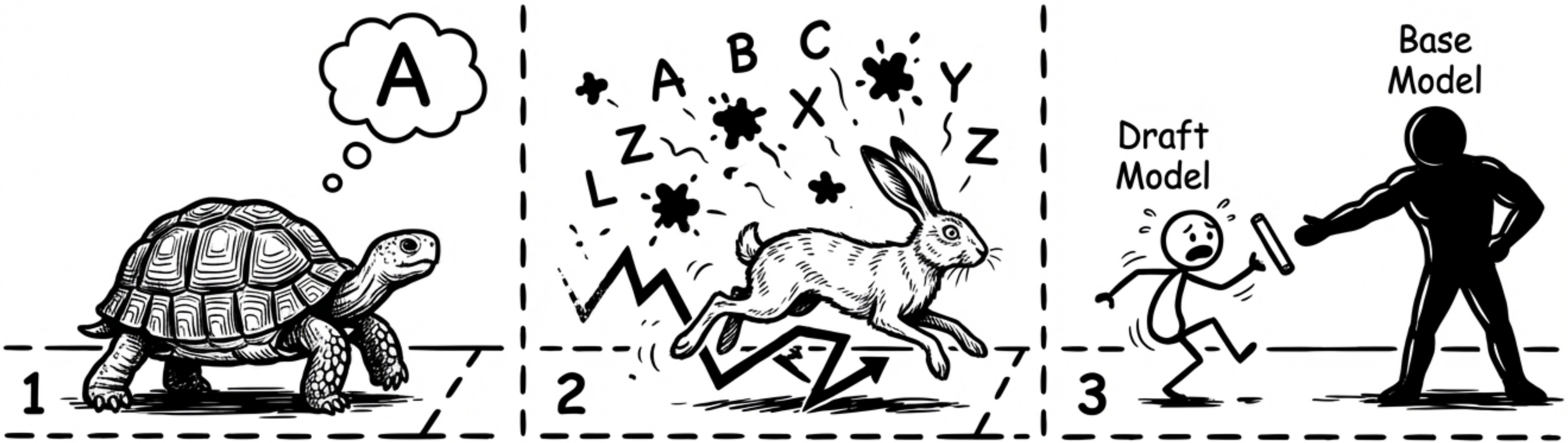


We are paying for the whole train, why not fill the seats?



Processing extra tokens incurs minimal to no latency increase.

The Failed Contenders



Standard AR

High Quality,
Low Speed.

Pure Diffusion

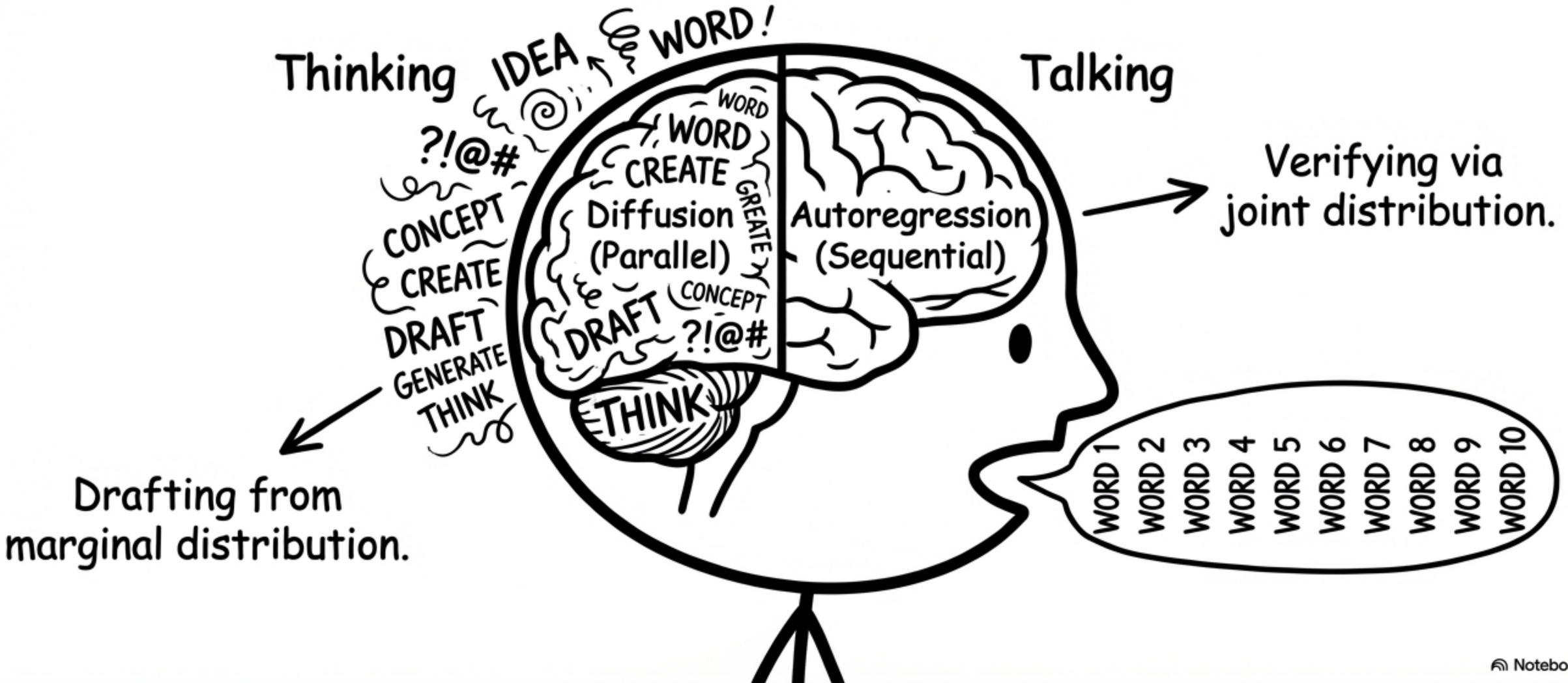
High Speed, Low Quality
(Independence Assumption).

Speculative Decoding

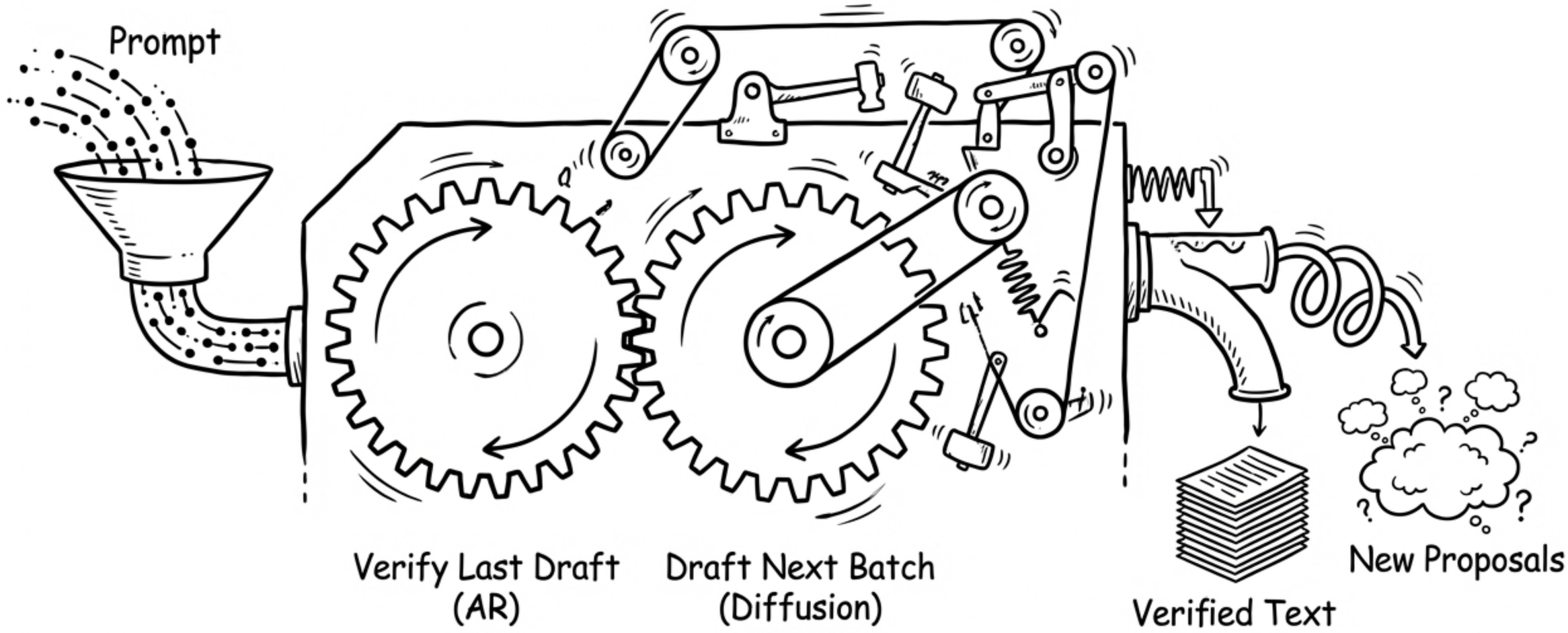
Bottlenecked by weaker
draft models.

Enter TiDAR: The Hybrid

Draft in Parallel. Verify Sequentially.



The Single Forward Pass Machine

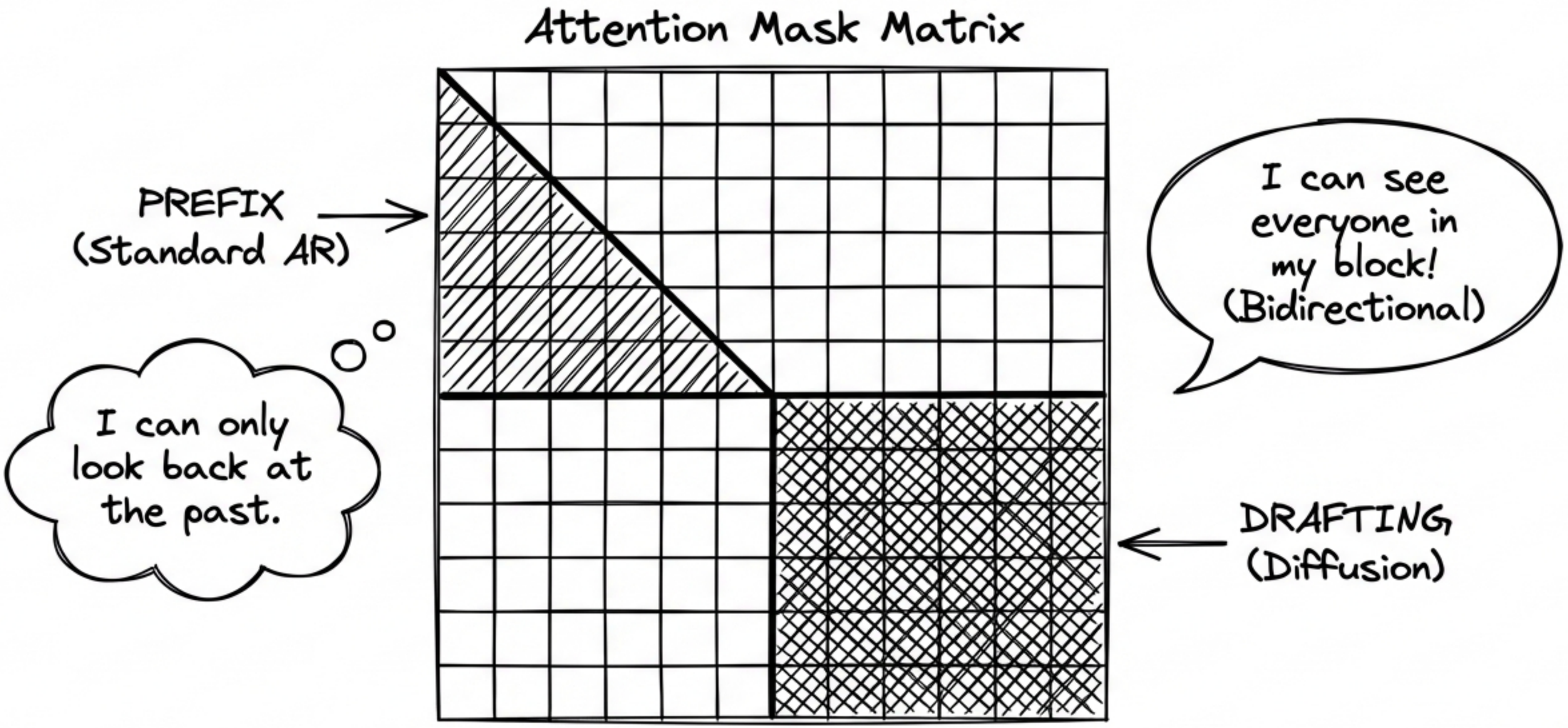


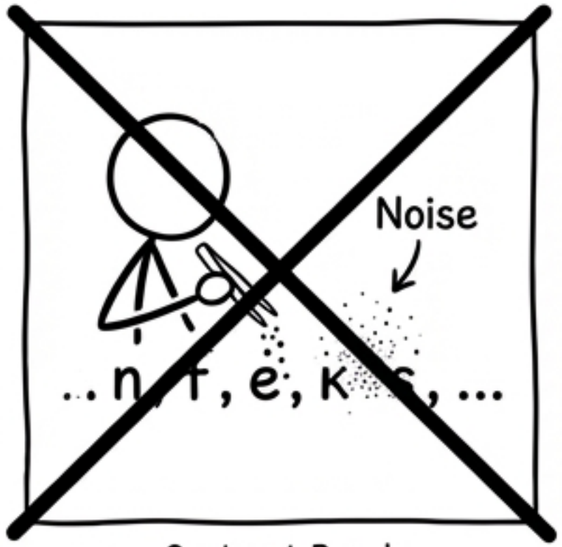
The Magic Trick: Simultaneity.

Unlike Speculative Decoding, TiDAR performs verification and drafting in the exact same GPU cycle.

Under the Hood: The Hybrid Attention Mask

Business in the front (Causal), Party in the back (Bidirectional)





Contrast Panel:
Tweezers and Noise

The Full Masking Strategy

1. Set all diffusion tokens to [MASK].

2. Predict them all at once (One-step diffusion).

3. Result: Denser loss signals and no complex noise schedules.

Self-Speculative Generation

DRAFT

Generate tokens via Diffusion.

VERIFY

Check against AR distribution.

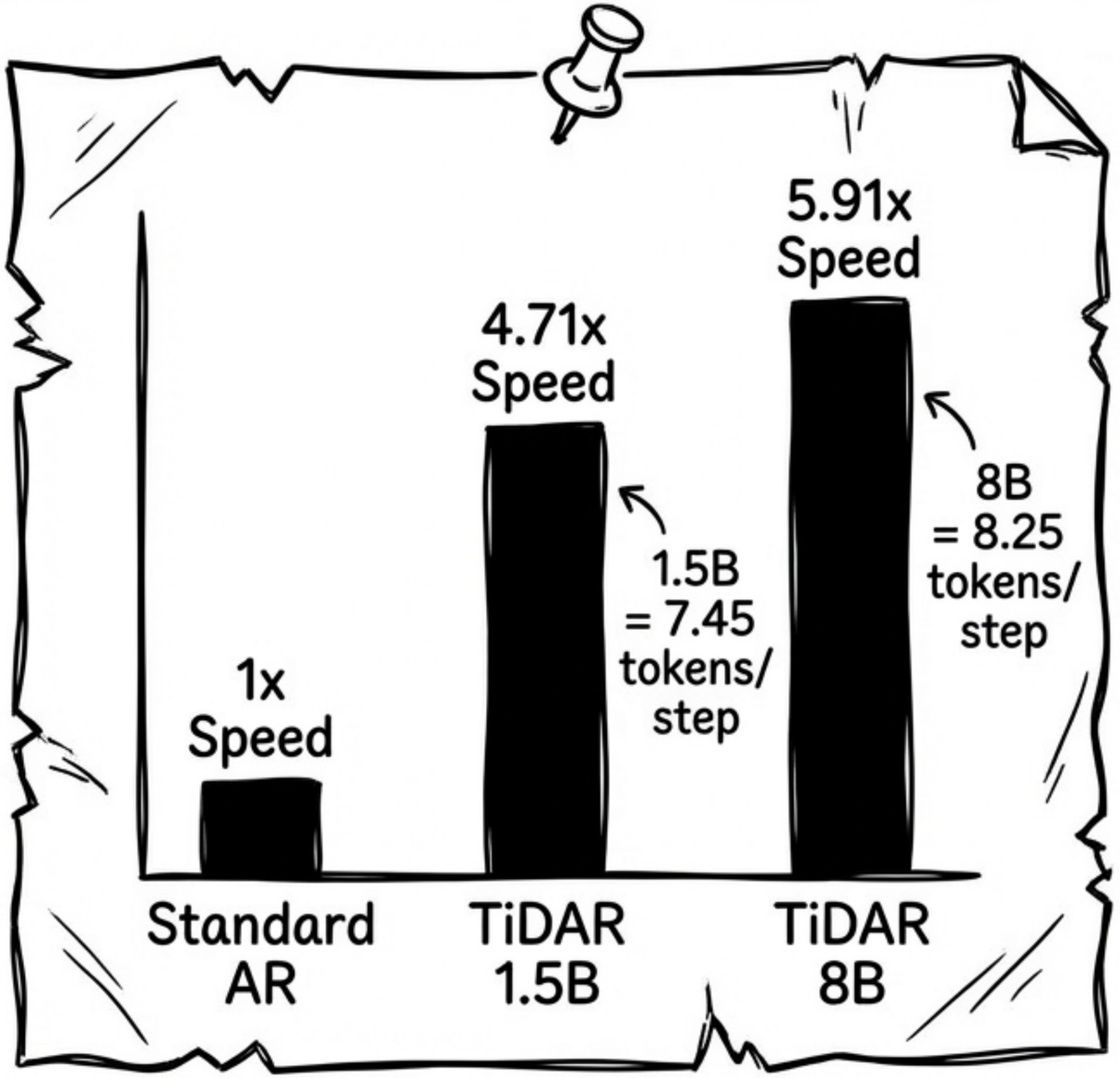
REJECTION SAMPLING

REPEAT

Use accepted tokens as new prefix.

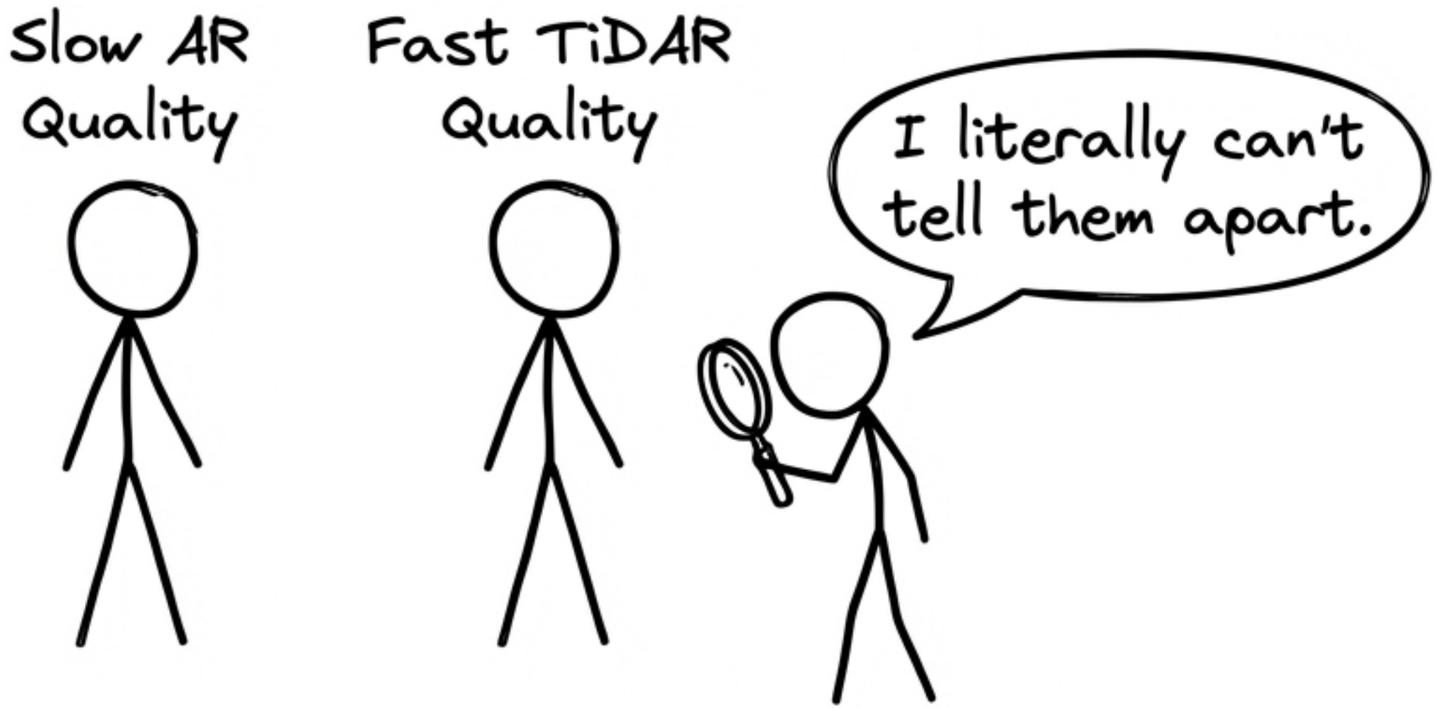
Efficiency: No Waste

No compute was harmed in the making of this sentence.

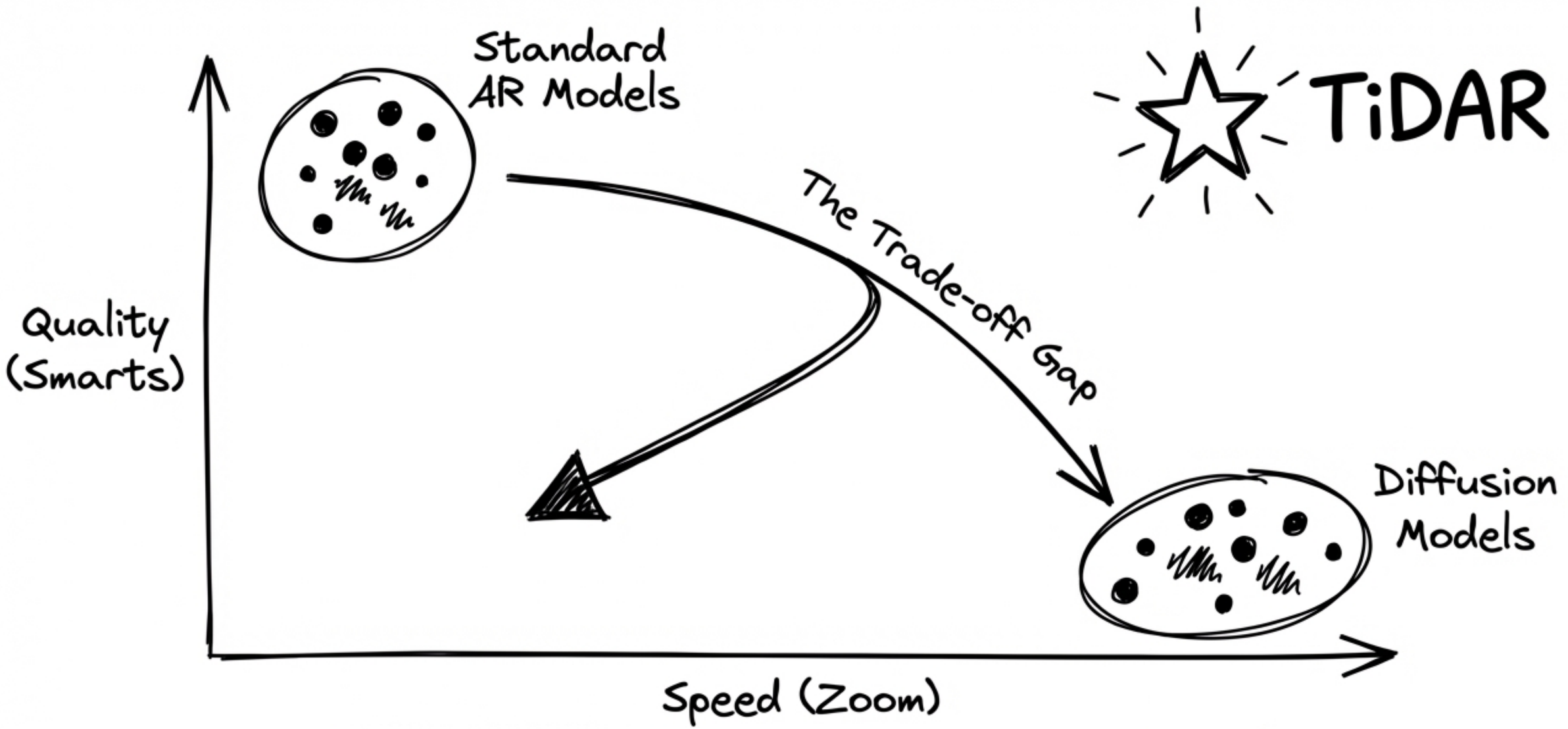


That is a LOT of free tokens.

The Results: Quality



HumanEval (Coding): Lossless Quality
GSM8K (Math): Competitive Performance
Beats 'Dream' and 'Llada' diffusion models.



The Pareto Frontier: We stopped trading off quality for speed. We just took both.

Old Assumptions

WHY THIS MATTERS:

- ☑ Deployment Ready:
  No extra draft models.
- ☑ Serving Friendly:
  Fits existing pipelines.
- ☑ Dense Compute:
  Maximizes GPU utilization.

Conventional wisdom says high quality costs high latency. Conventional wisdom is wrong.

TiDAR: Bridging Autoregressive Quality and Diffusion Efficiency.

Reference: TiDAR: Bridging Autoregressive Quality and Diffusion Efficiency (NVIDIA, 2025).