

Decoding 2 and Rotational Embeddings

CS685 Spring 2025

Advanced Natural Language Processing

Haw-Shiuan Chang

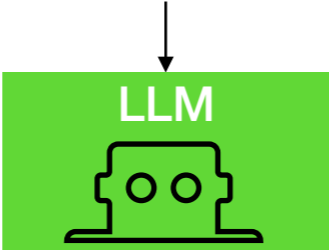
College of Information and Computer Sciences
University of Massachusetts Amherst

Logistics

- <https://people.cs.umass.edu/~hschang/cs685/schedule.html>
- **I need to leave very soon after today's class**
 - If you have some complex questions, come to my office hour
- **4/2: Deadline of applying for the first round of API credit**
 - <https://piazza.com/class/m1kz66st9dn62i/post/146>
 - The credits are for API calls or very cheap fine-tuning (e.g., fine-tuning GPT)
 - We will have two rounds of credit allocations.
- **4/7:** Midterm Review 1
- **4/9:** Midterm Review 2
- **4/11:** HW 2 due
- **4/18 (Friday but Monday Schedule): Midterm**
- **5/9: Final project report due**

Top-p Sampling (Nucleus Sampling)

The screenwriter of The Matrix is _____

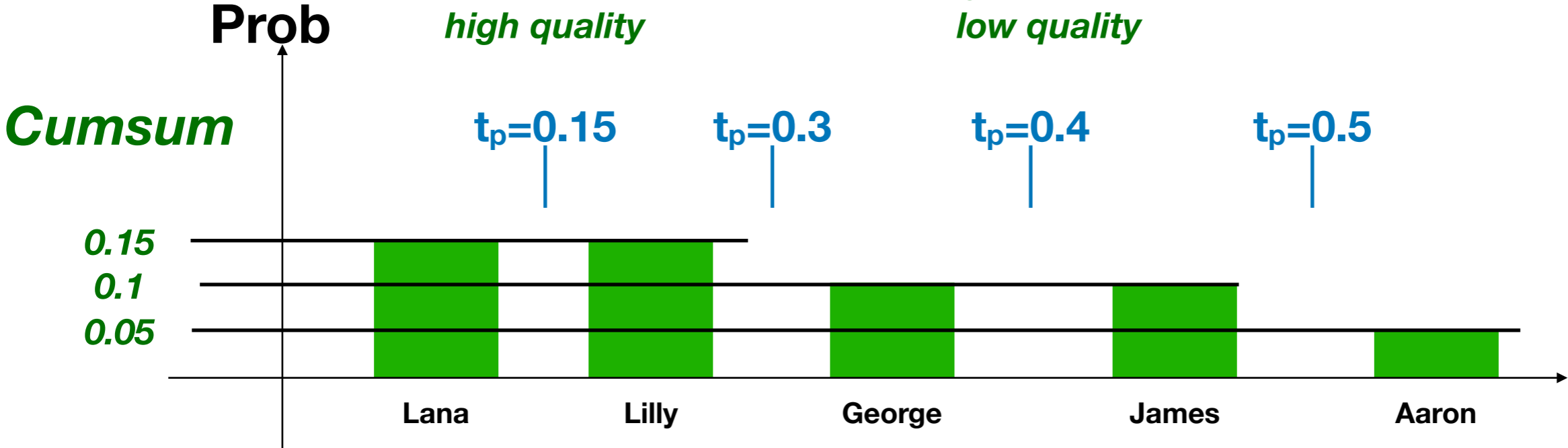


Unsure about the answer



*Low p ->
low diversity,
high quality*

*High p ->
high diversity,
low quality*

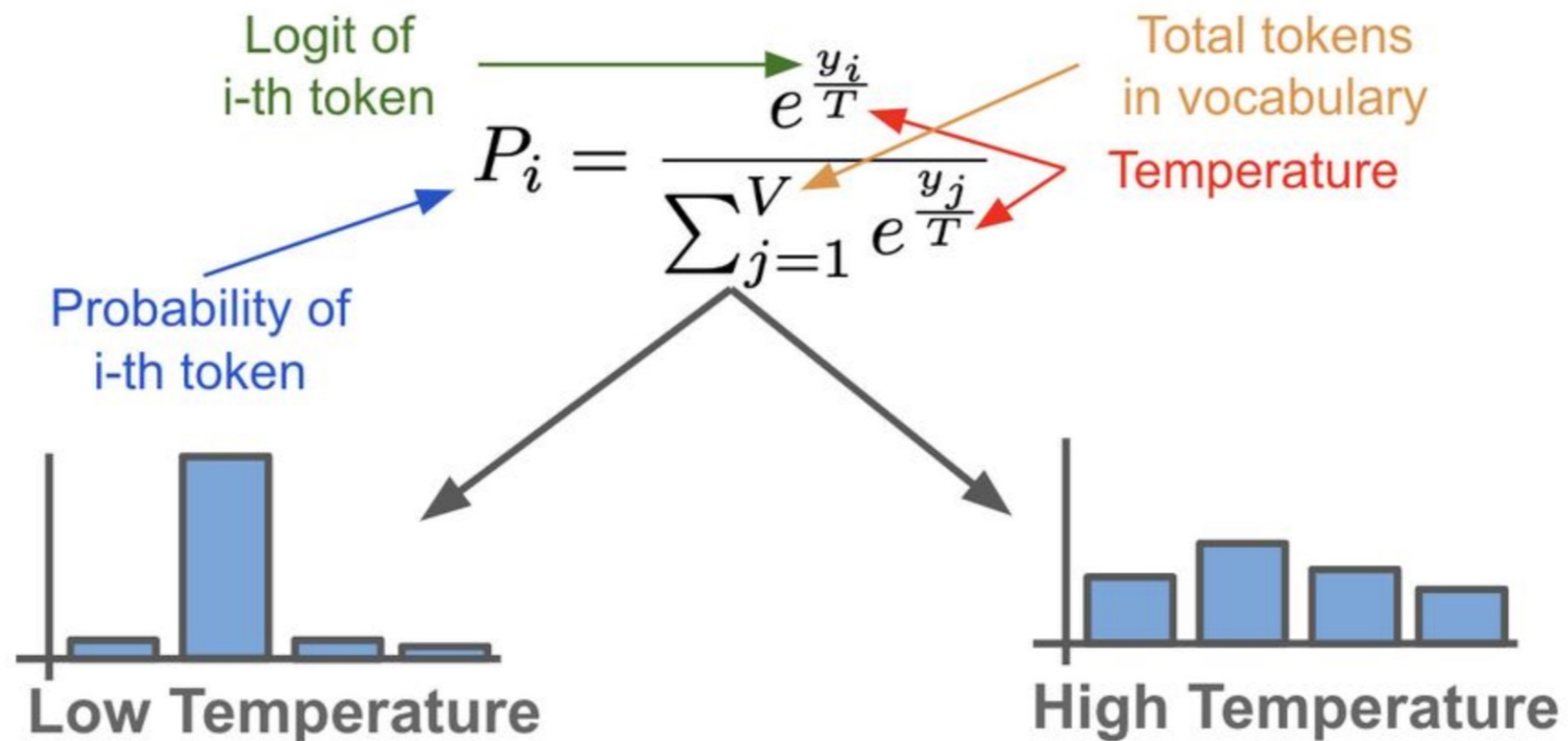


https://commons.wikimedia.org/wiki/File:Andy_and_Lana_Wachowski_%282012%29.JPG

<https://www.flickr.com/photos/nunoluciano/5396200604>

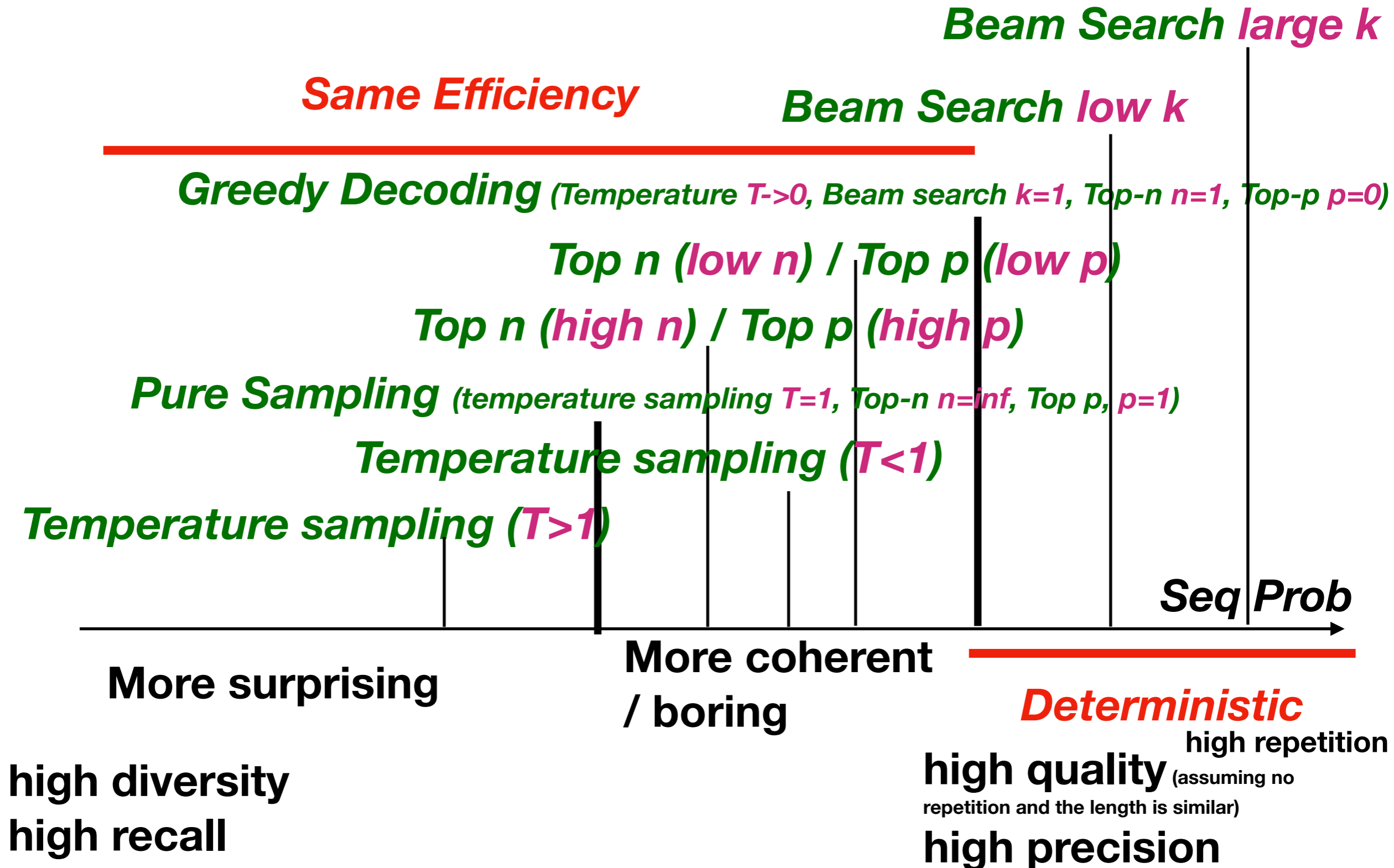
Temperature Sampling

We can adjust the temperature to modulate the uniformity of the token distribution produced by the softmax transformation



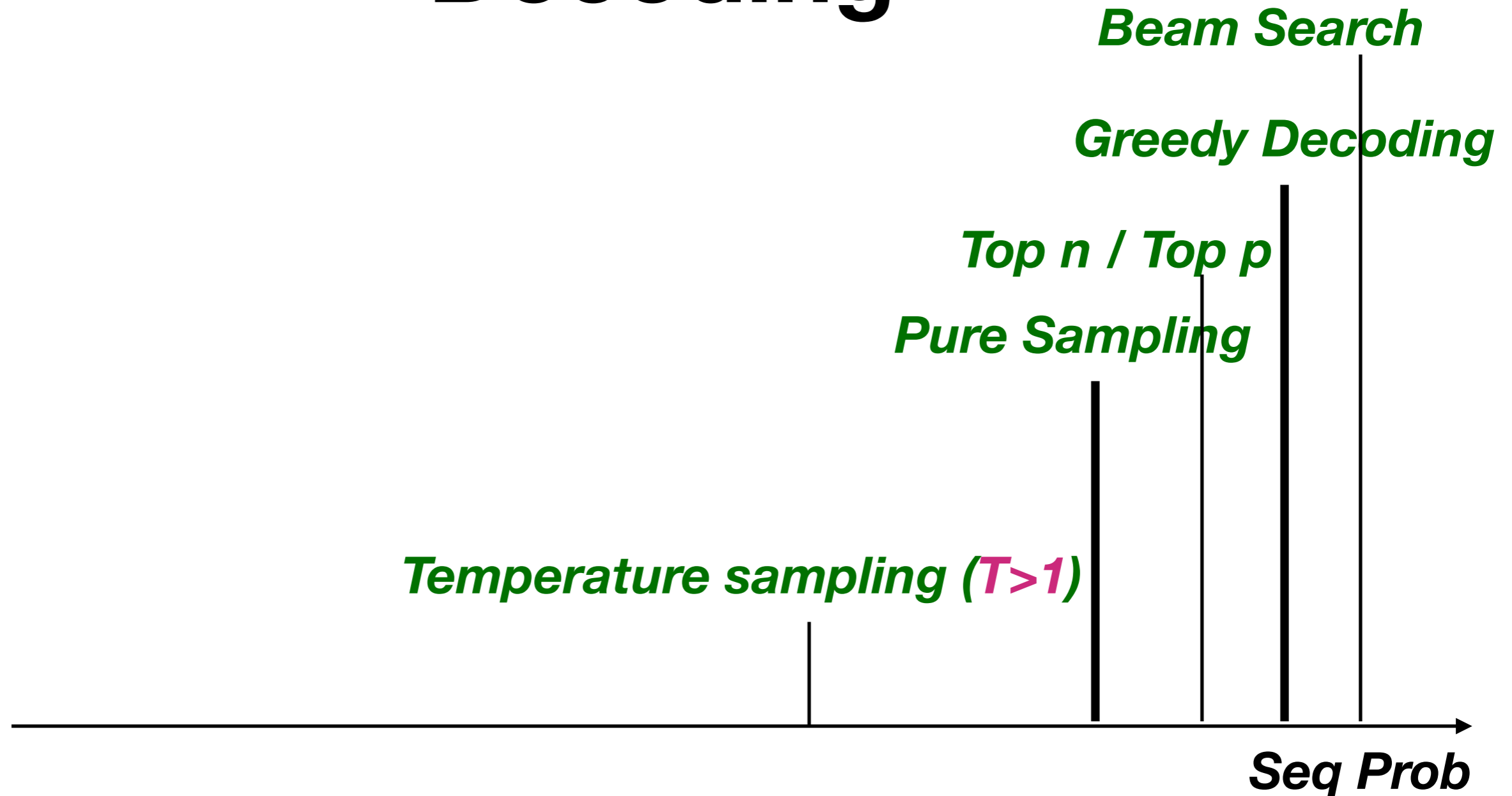
<https://aman.ai/primers/ai/token-sampling/>

Base LLM Decoding



Instruct LLM (after SFT/RLHF)

Decoding

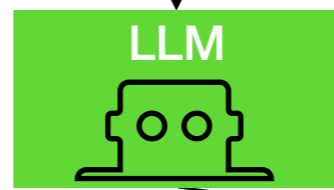


high diversity
high recall

high quality
high precision

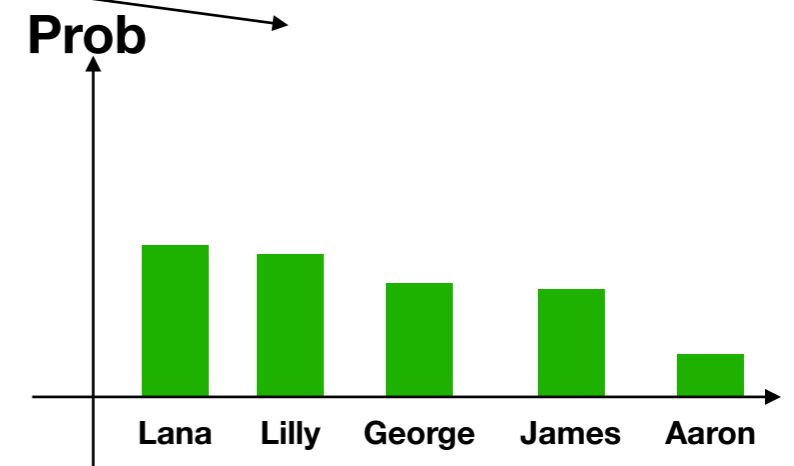
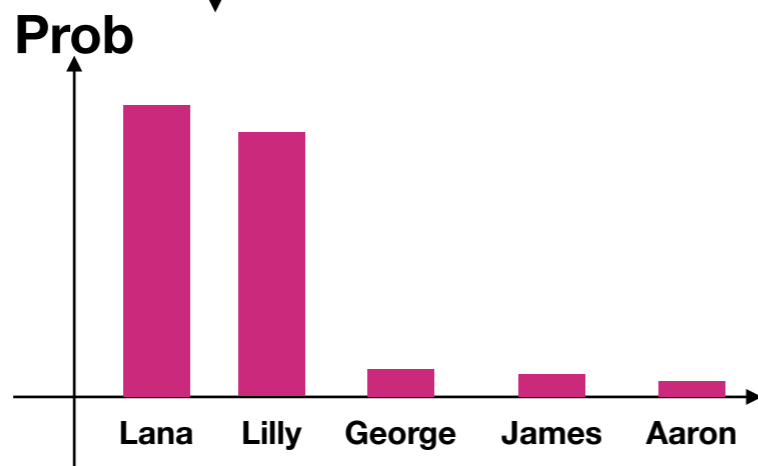
Hallucination from High-Entropy Distribution

The screenwriter of The Matrix is ____



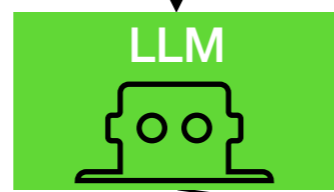
Pretty sure about the answer

Unsure about the answer



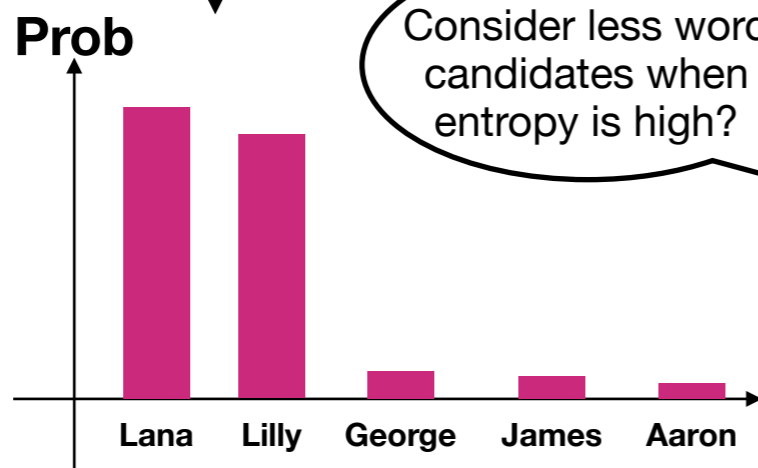
Hallucination from High-Entropy Distribution

The screenwriter of The Matrix is ____

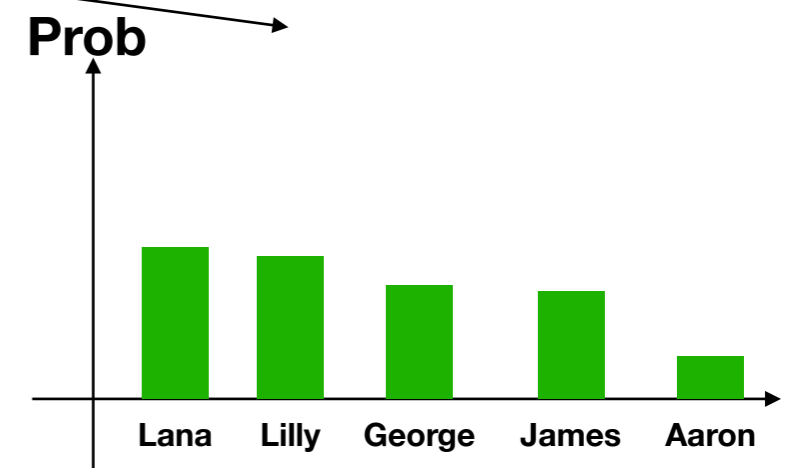
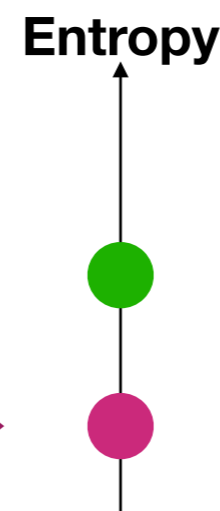


Pretty sure about the answer

Unsure about the answer



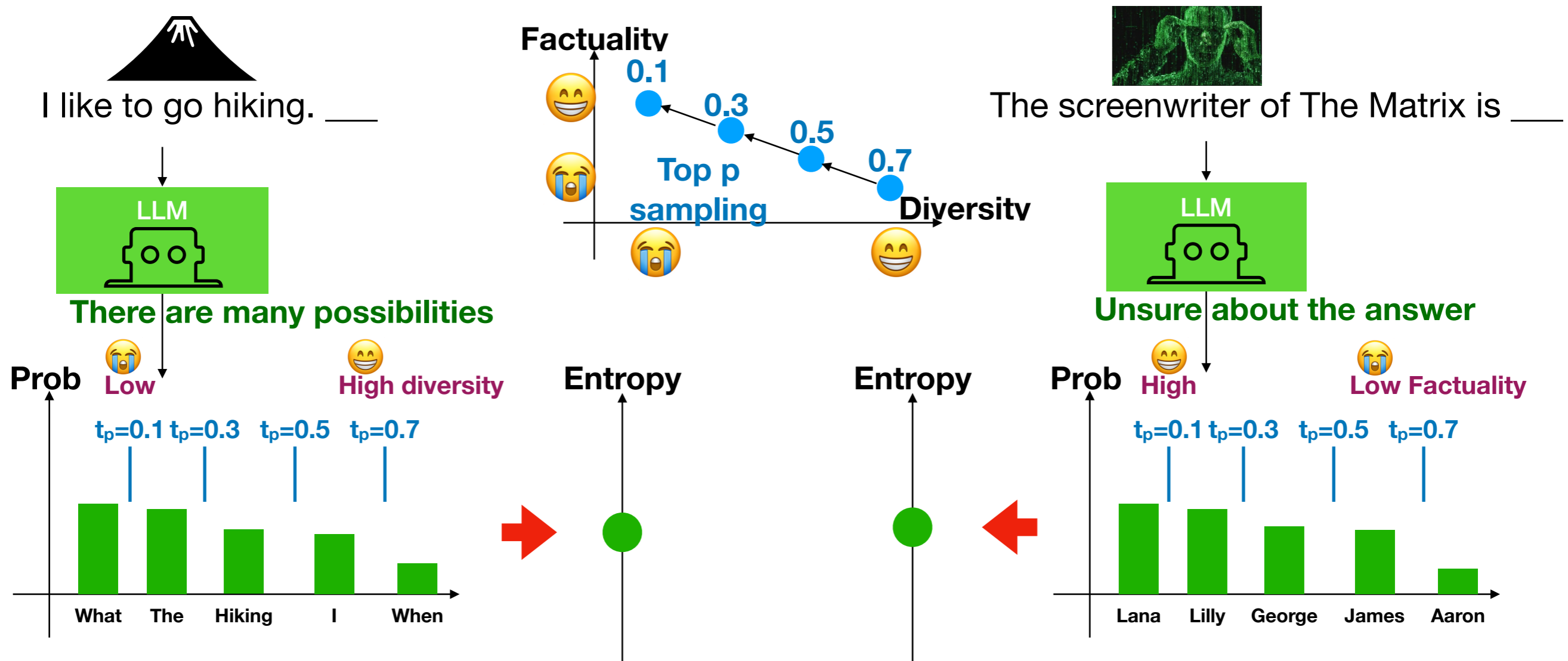
Consider less word candidates when entropy is high?



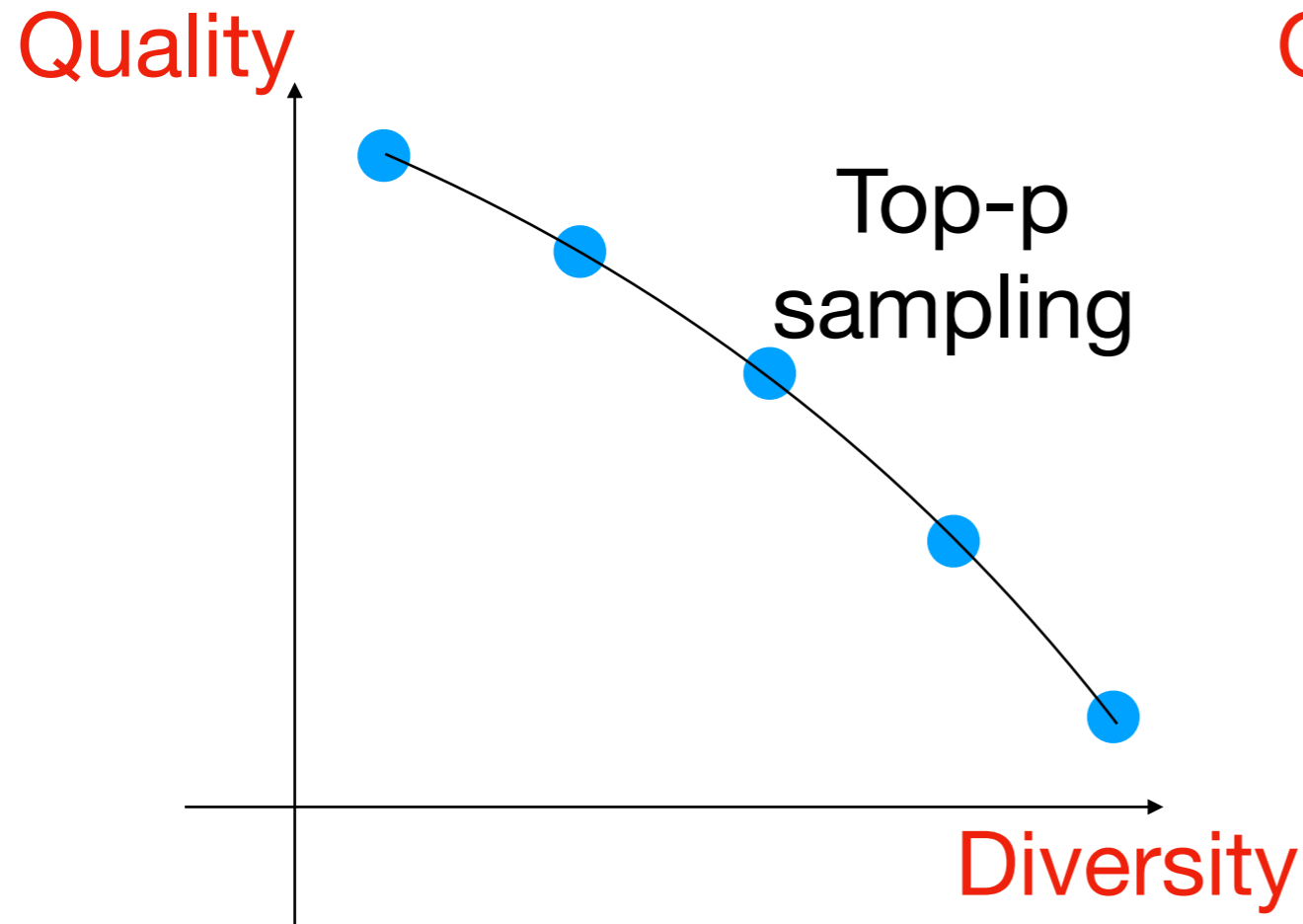
<https://www.flickr.com/photos/nunoluciano/5396200604>
https://commons.wikimedia.org/wiki/File:Andy_and_Lana_Wachowski_%282012%29.JPG

$$H(q) = - \sum_x q(w_t = x | w_{<t}) \log q(w_t = x | w_{<t})$$

Tradeoff between Factuality and Diversity



Trade-off between Quality and Diversity



Lee, Nayeon, et al. "Factuality enhanced language models for open-ended text generation." *Advances in Neural Information Processing Systems* 35 (2022): 34586-34599.

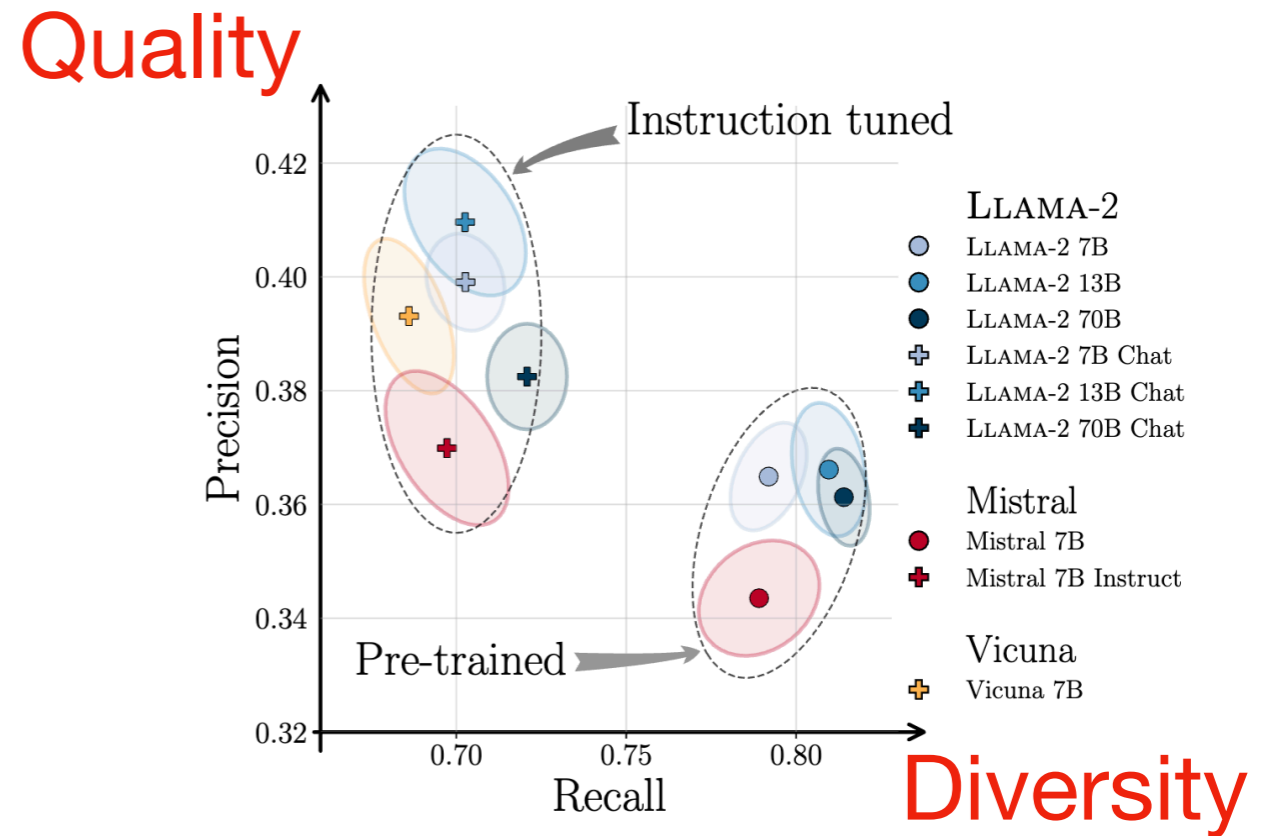


Figure 1: Precision and Recall of various models on generating the WebText dataset, with the 2 standard deviation error ellipsis. Chat and pre-trained models different behaviors are clearly captured by evidenced.

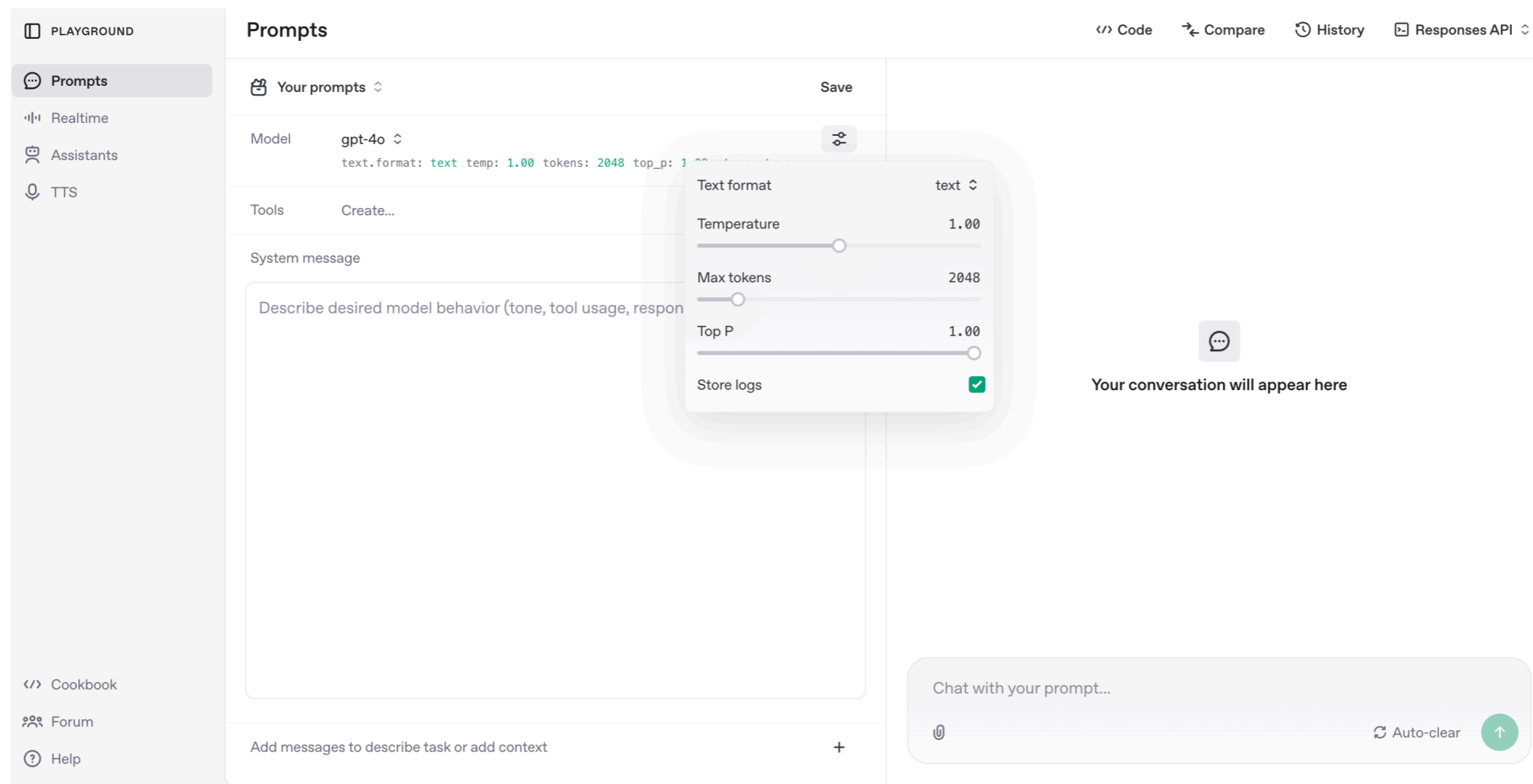
Le Bronnec, Florian, et al. "Exploring Precision and Recall to assess the quality and diversity of LLMs." *62nd Annual Meeting of the Association for Computational Linguistics*. 2024.

Midterm Example Question

- Changing the temperature in the sampling and RLHF could both reduce the entropy/diversity of generation. What are their differences?

Midterm Example Question

- How do you remove the randomness in the generation?

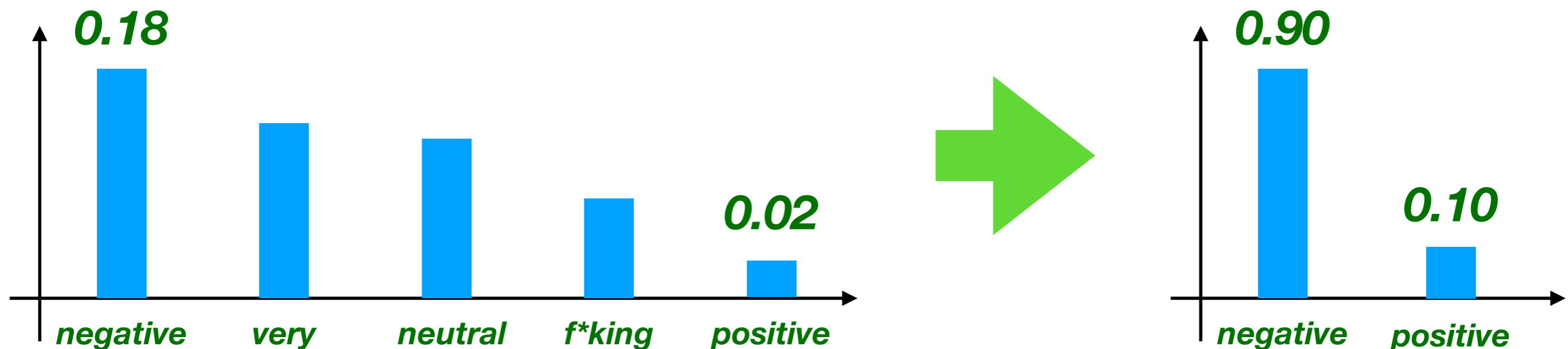


The screenshot shows the OpenAI Playground interface. On the left is a sidebar with navigation options: PLAYGROUND, Prompts (selected), Realtime, Assistants, TTS, Cookbook, Forum, and Help. The main area is titled 'Prompts' and includes a 'Save' button. Below this, the 'Model' is set to 'gpt-4o' with a dropdown arrow. A status bar shows 'text.format: text temp: 1.00 tokens: 2048 top_p: 1.00'. The 'Tools' section has a 'Create...' button. The 'System message' section contains a text area with the placeholder 'Describe desired model behavior (tone, tool usage, respon...'. A settings overlay is centered on the screen, showing sliders for 'Temperature' (set to 1.00), 'Max tokens' (set to 2048), and 'Top P' (set to 1.00). The 'Text format' is set to 'text' and 'Store logs' is checked. On the right, there is a placeholder for the conversation: 'Your conversation will appear here'. At the bottom, there is a text input field 'Chat with your prompt...' with an 'Auto-clear' button and a send button.

<https://platform.openai.com/playground/prompts?mode=chat&models=gpt-4o>

Constrained Decoding

- Context
 - {input sentence}
 - e.g., This movie is far from being great
- The sentiment of the movie is ____

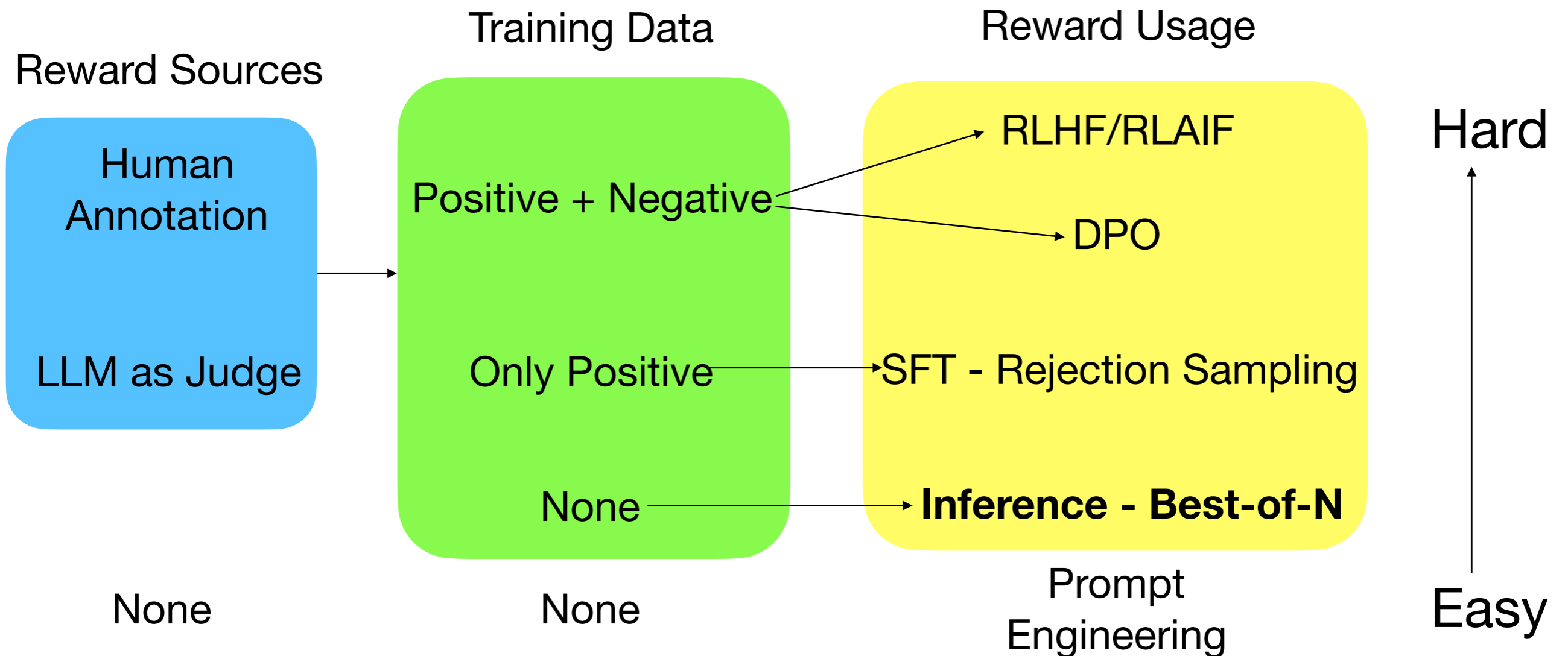


Constrained Decoding

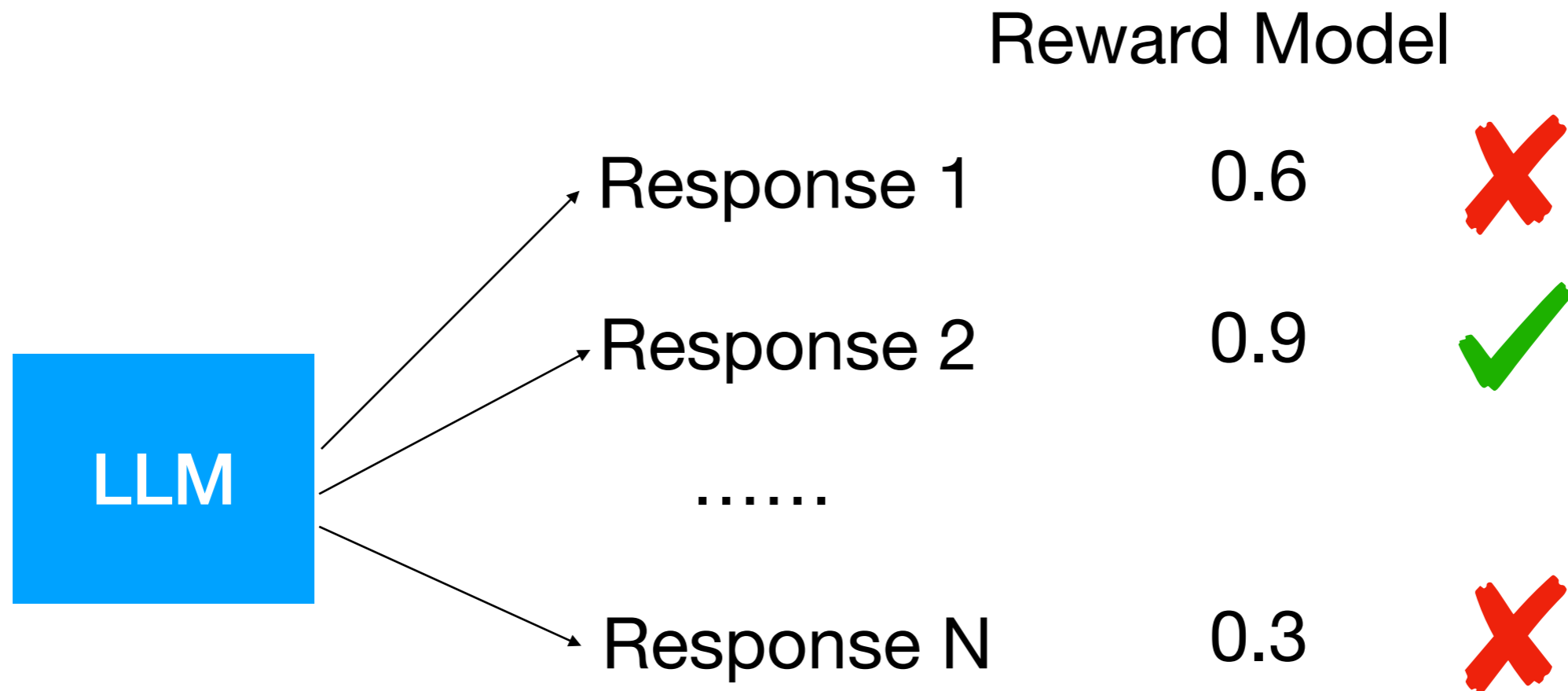
- Sometimes, LLMs (especially worse ones) are not very good at following the negative constraints. For example,
 - you ask a model to continue a story, but you don't want the model to mention the main character's name
 - you ask a model to generate a tweet, but you don't want the model to generate the hashtags
- You don't want the model to generate short responses

Test-time Scaling

Inference Methods



Best of N



The reward model could be anything. For example, LM probability (beam search), answer quality scorer, profanity/toxicity filter, sentiment classifier, PRM

Best-of-N vs Beam Search

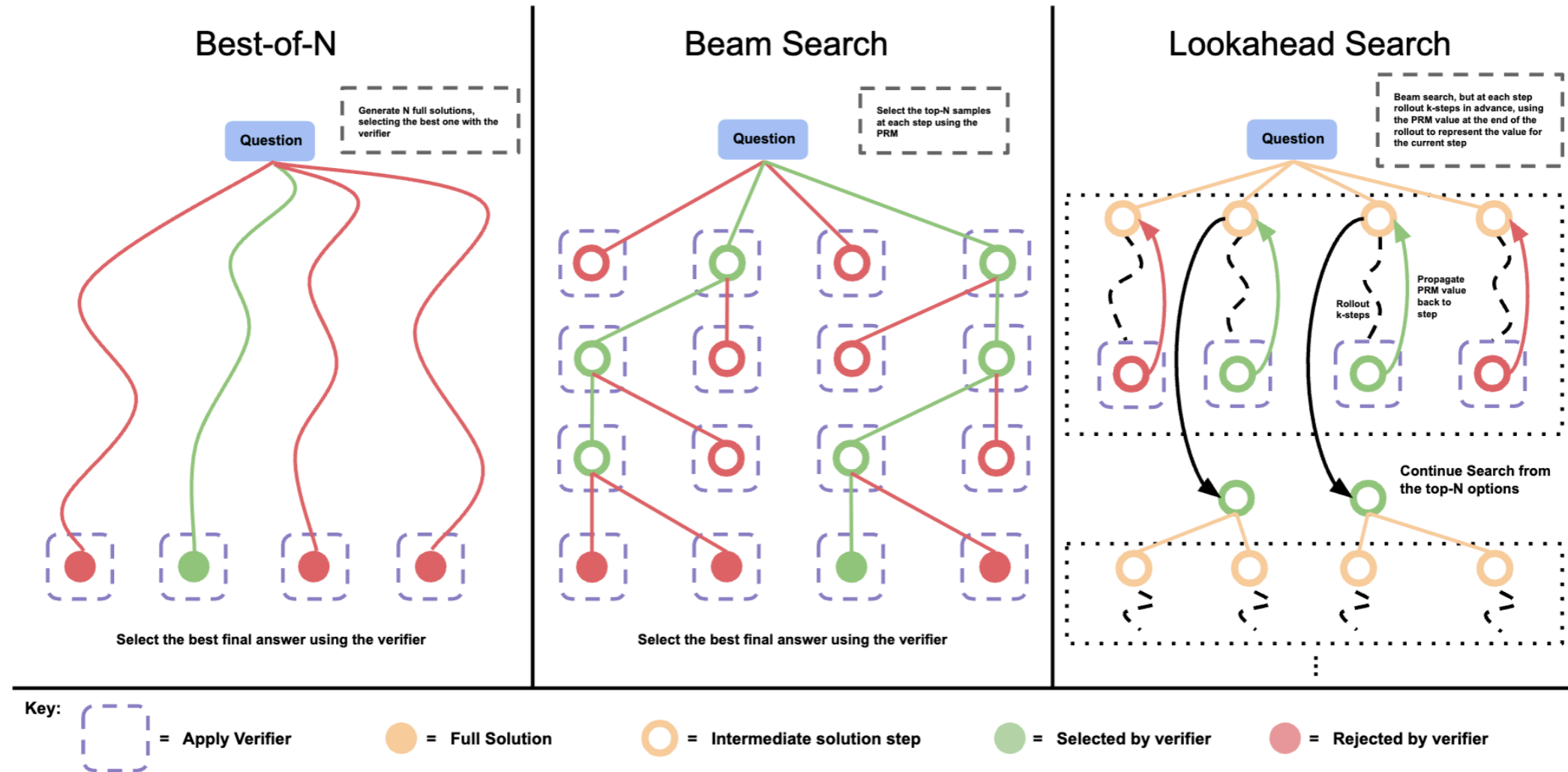
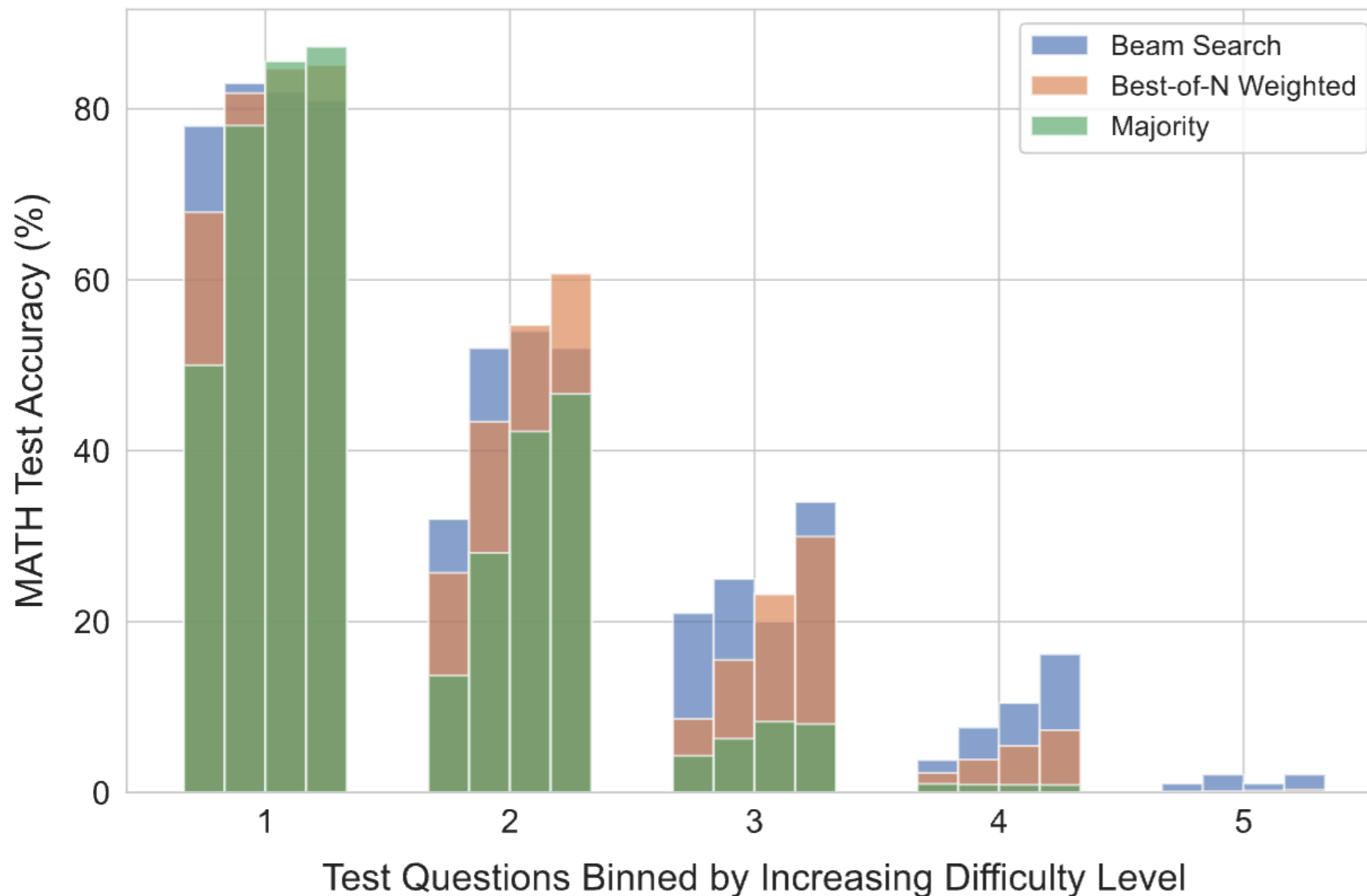


Figure 2 | *Comparing different PRM search methods.* **Left:** Best-of-N samples N full answers and then selects the best answer according to the PRM final score. **Center:** Beam search samples N candidates at each step, and selects the top M according to the PRM to continue the search from. **Right:** lookahead-search extends each step in beam-search to utilize a k-step lookahead while assessing which steps to retain and continue the search from. Thus lookahead-search needs more compute.

**Scaling LLM Test-Time Compute Optimally can
be More Effective than Scaling Model Parameters (<https://arxiv.org/pdf/2408.03314>)**

Usefulness of Guidance Depends on the Difficulty

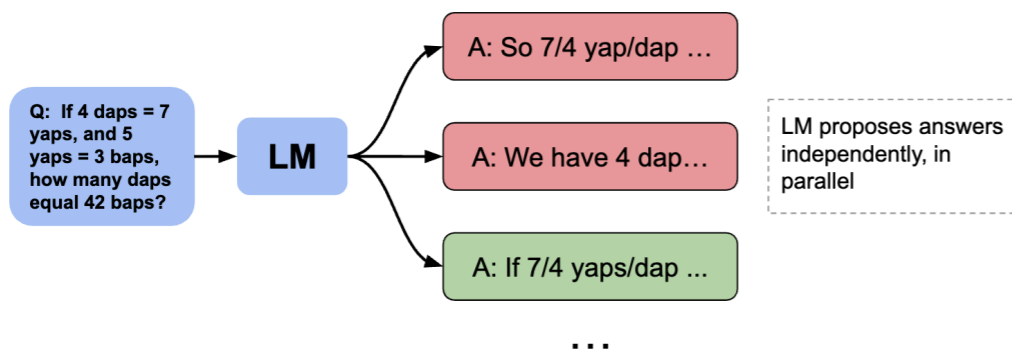
Comparing Beam Search and Best-of-N by Difficulty Level



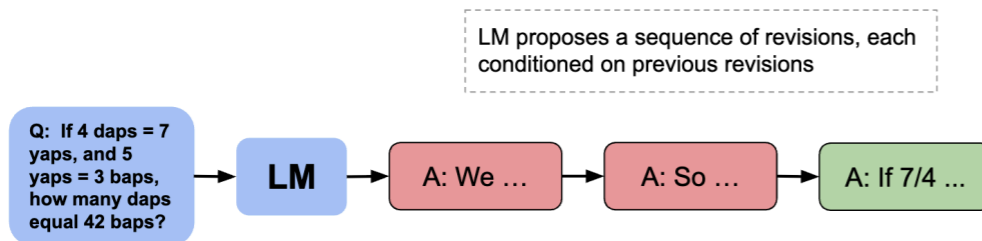
- For LLM rather than LRM

Parallel vs Sequential

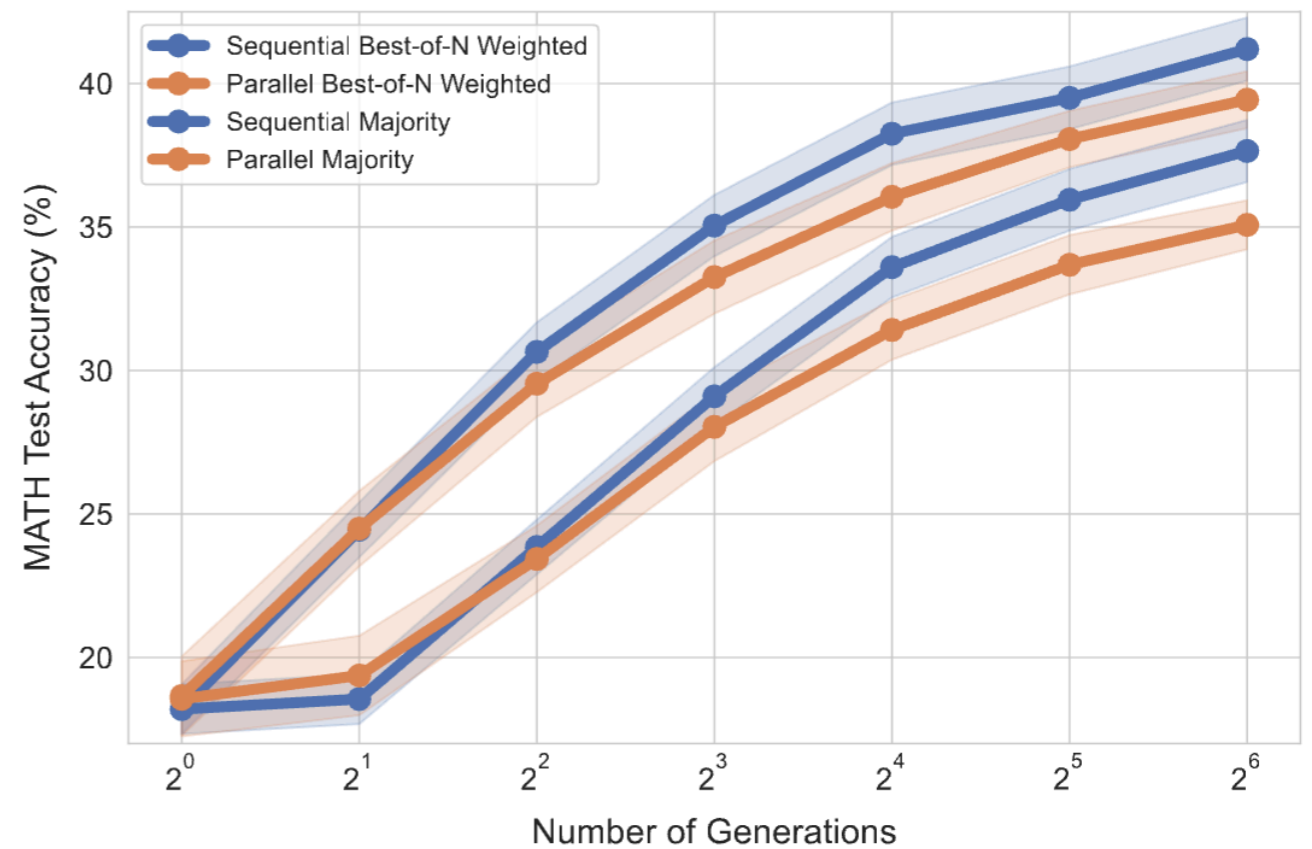
Parallel Sampling



Sequential Revisions



Revision Model Parallel Verses Sequential



Sequential with Distillation from LRM is Much Better

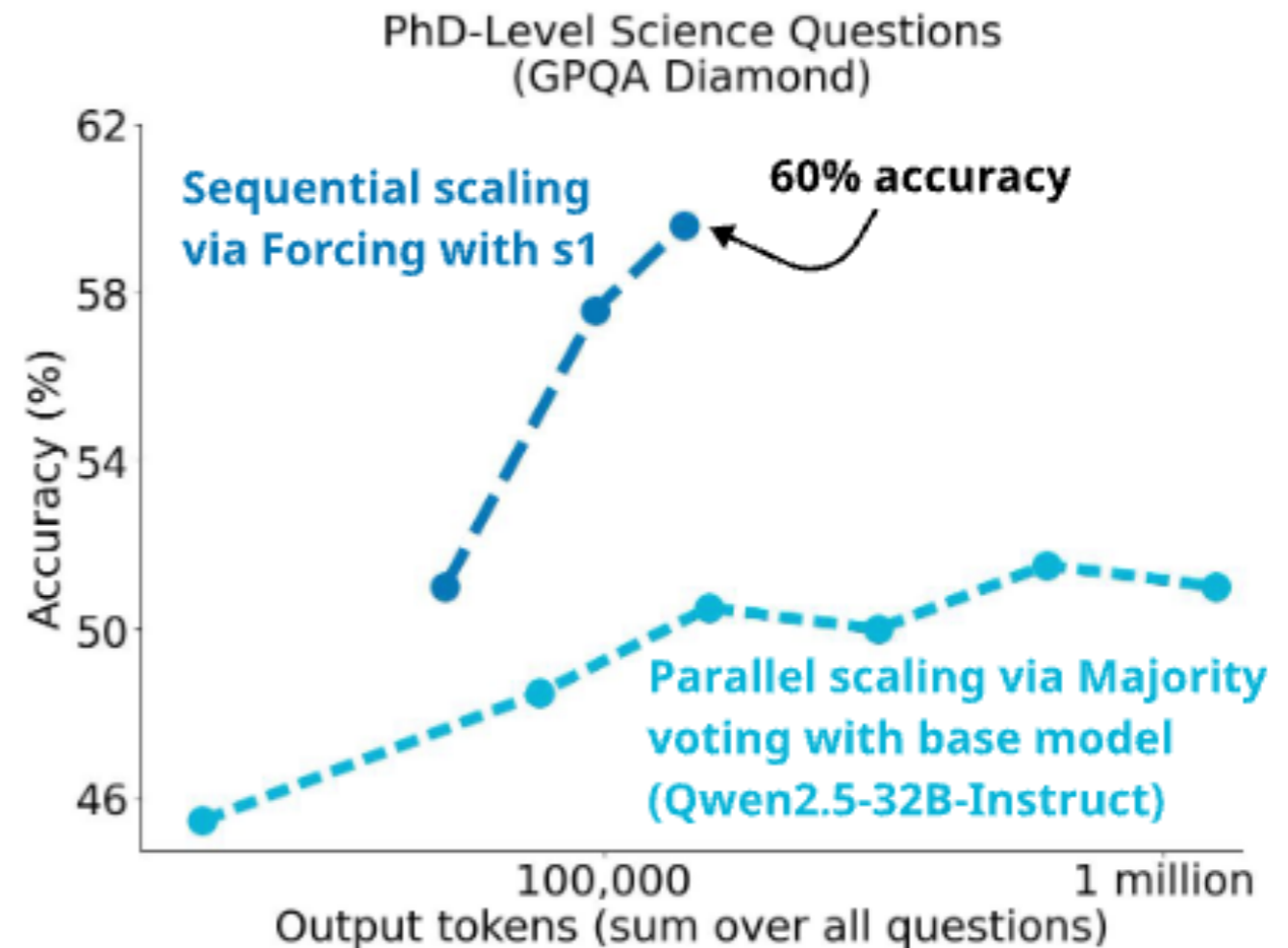
How many r in raspberry? **Question**

Let's break down the process of counting the letter 'r' in the word "raspberry" ...
* First letter: 'r' - This is an 'r', count = 1.
* Second letter: 'a' - Not an 'r', count remains 1 ...
* Sixth letter: 'e' - Not an 'r', count remains 1.
* Seventh letter: 'r' - This is an 'r', count = 2.
* Eighth letter: 'y' - Not an 'r', count remains 2 ...
The number of 'r's in "raspberry" is 2.
Wait, let's re-read the question carefully. It asks "How many r in raspberry?" ... * r - a - s - p - b - e - r - r - y ... * First 'r' ... * Second 'r' ... * Third 'r' ... Count = 3 ... **Reasoning trace**

My initial answer of 2 was incorrect due to a quick reading of the word. **Final Answer:** The final answer is **3** **Response**

Figure 3. **Budget forcing with s1-32B.** The model tries to stop after "...is 2.", but we suppress the end-of-thinking token delimiter instead appending "Wait" leading s1-32B to self-correct its answer.

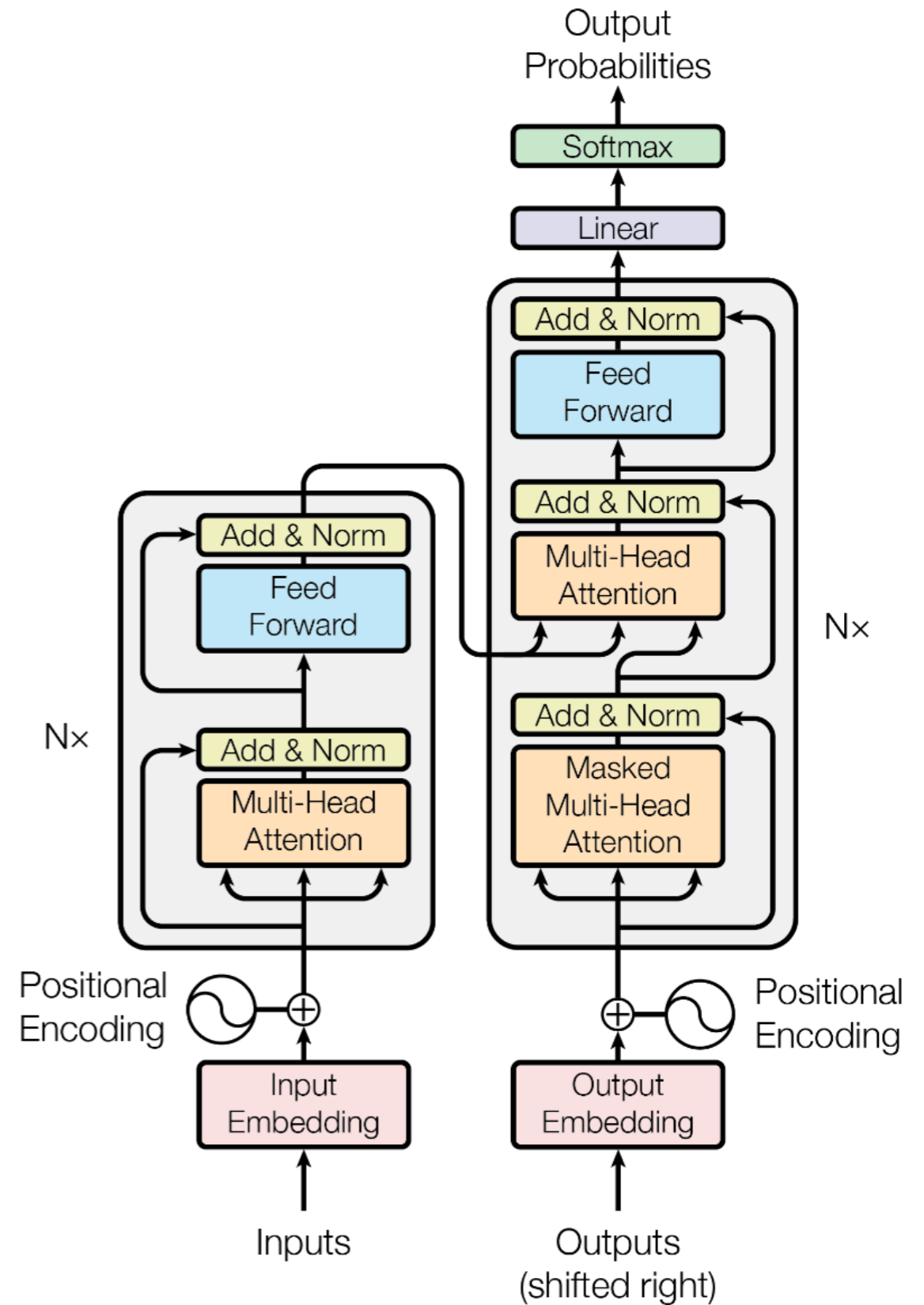
s1 also uses constrained decoding



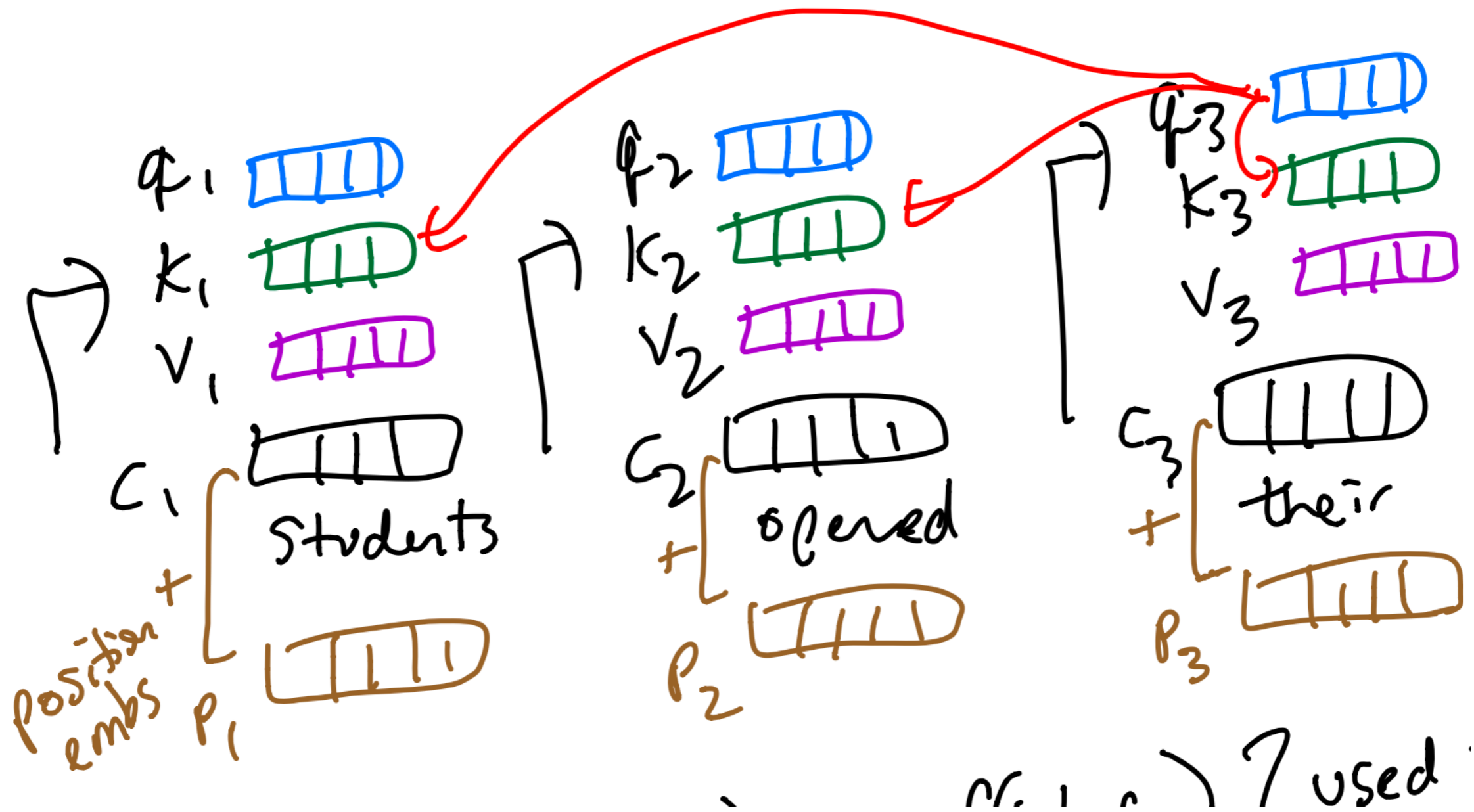
(b) Parallel scaling via majority voting

s1: Simple test-time scaling (<https://arxiv.org/pdf/2501.19393>)

Position Embedding



Without Positional Embeddings, Transformer can only Handle Set rather than Sequence



Desired Properties of Positional Embeddings

- Shift Invariant
- Context-Dependent
 - Control the attention range
 - Long distant dependency
- Not overfitting



...Green apple.....An apple is ___
p₃ p₄ p₃₁ p₃₂ p₃₃



An apple is usually red.....
p₀ p₁ p₂ p₃ p₄



.....An apple is usually red.....
p₃₁ p₃₂ p₃₃ p₃₄ p₃₅

Last Year Note

RoPE

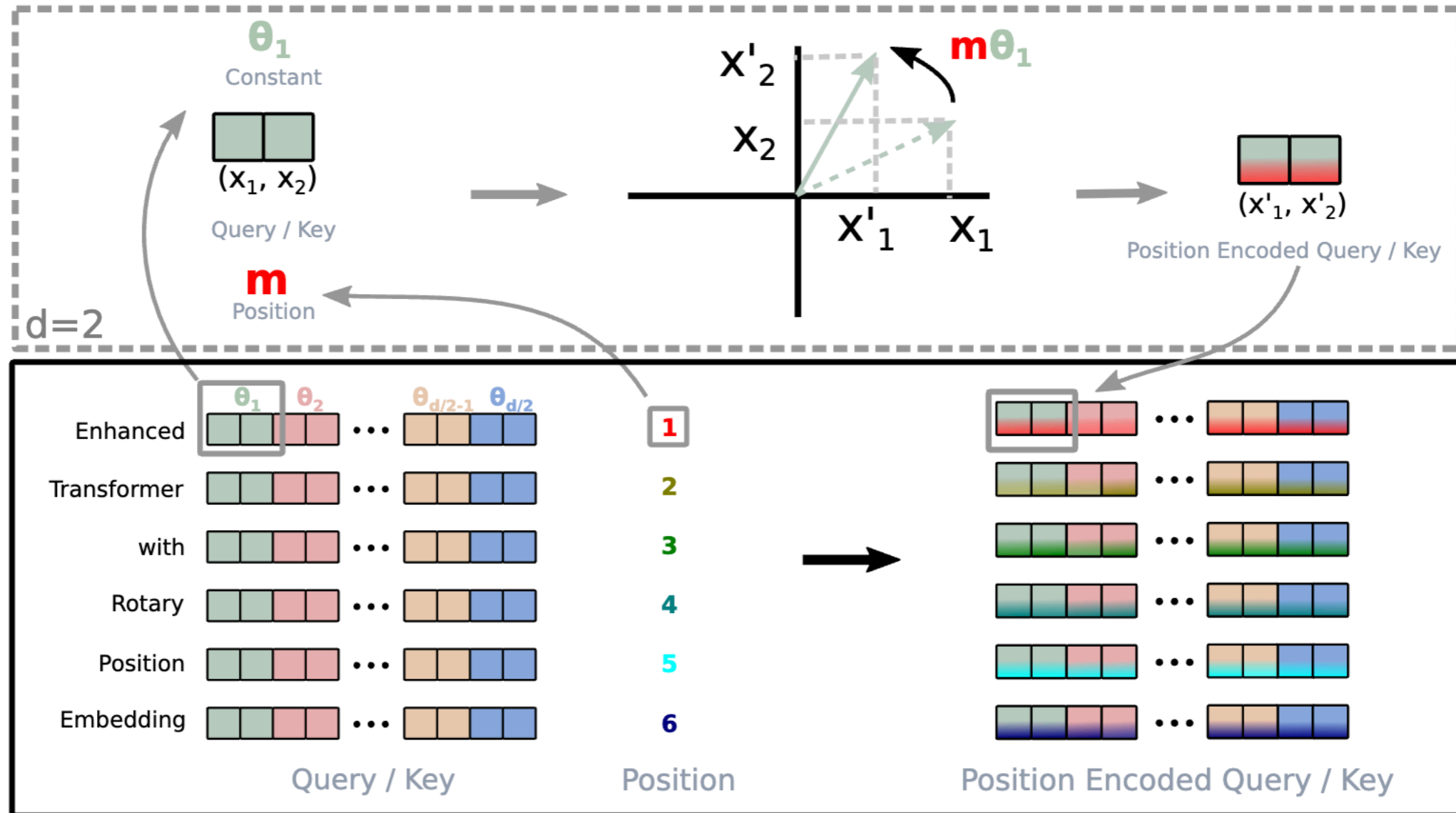


Figure 1: Implementation of Rotary Position Embedding(RoPE).

RoPE

**Long distance
attention**

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \mathbf{R}_{\Theta, m}^d \mathbf{W}_{\{q,k\}} \mathbf{x}_m$$

$$\mathbf{R}_{\Theta, m}^d = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \dots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \dots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \dots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

**Short distance
attention**

$$\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, \dots, d/2]\}.$$

Longer Context Matters Less

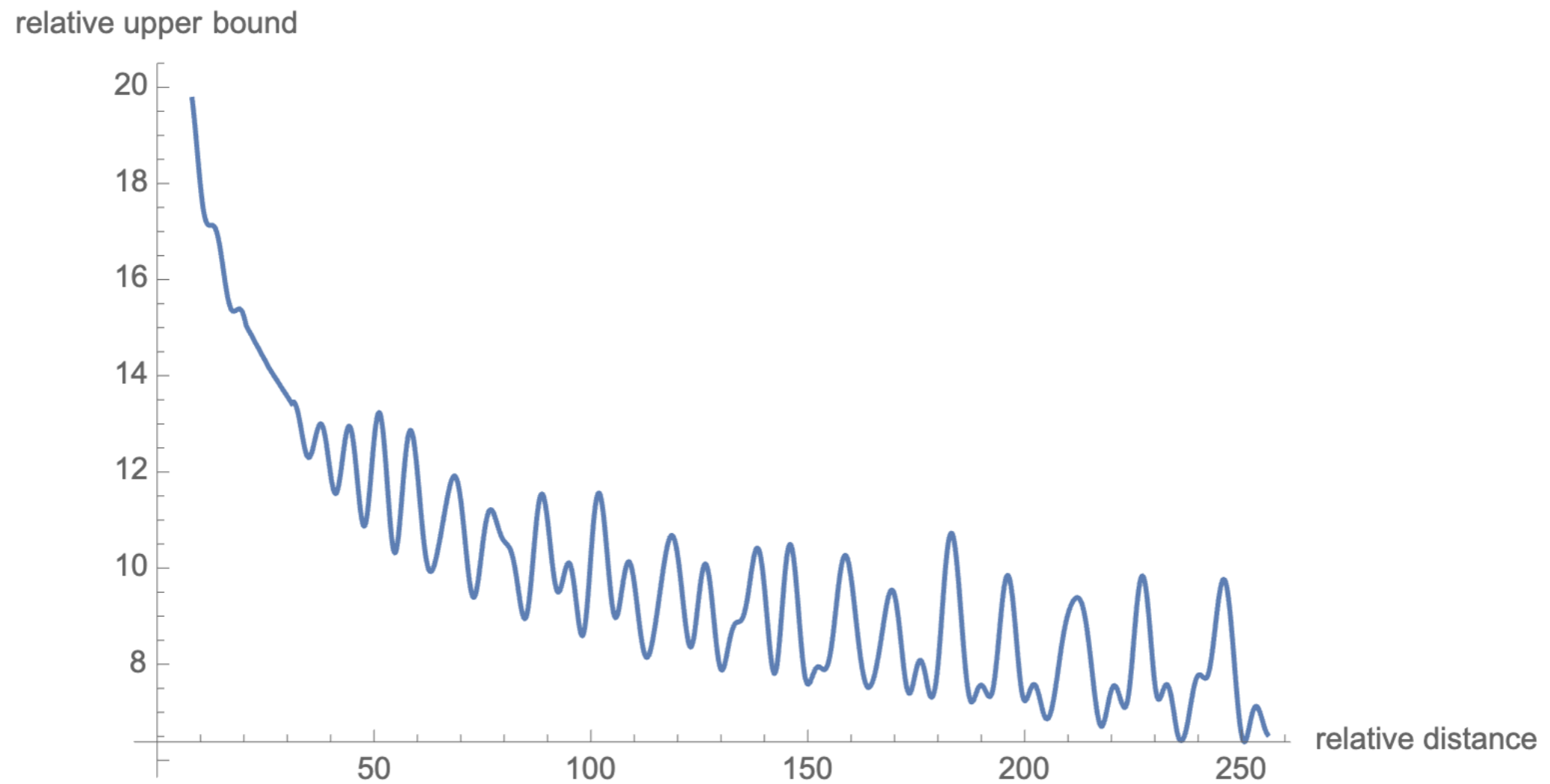


Figure 2: Long-term decay of RoPE.

Desired Properties of Positional Embeddings

- Shift Invariant
- Context-Dependent
 - Control the attention range
 - Long distant dependency
- Not overfitting



...Green apple.....An apple is __



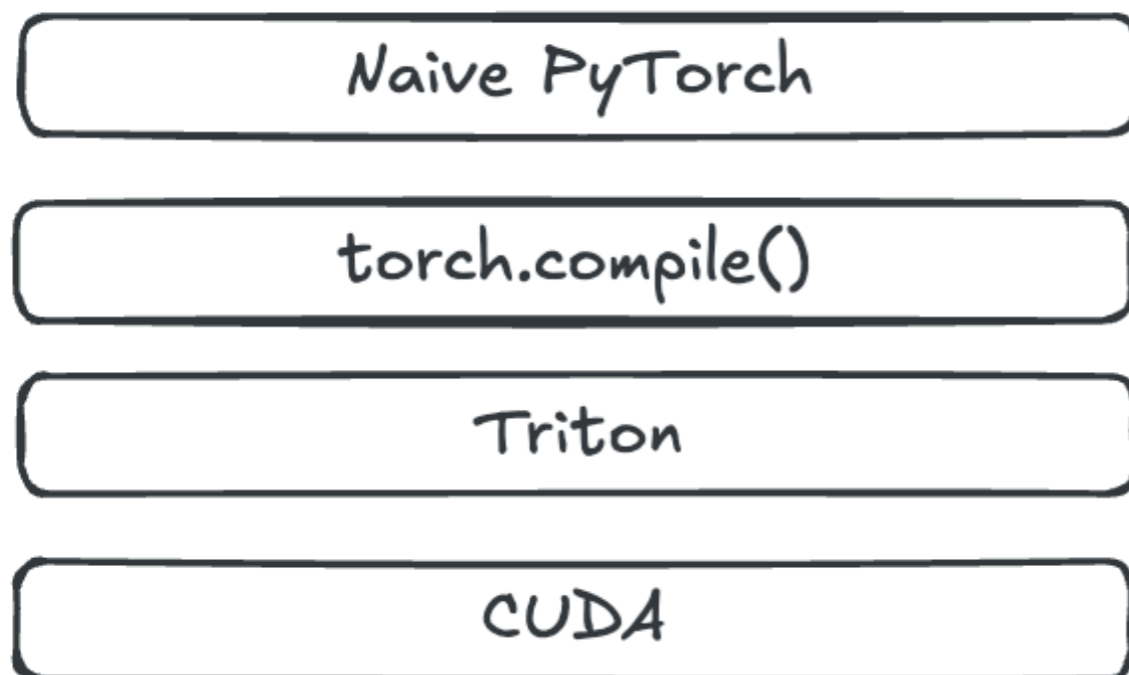
An apple is usually red.....



.....An apple is usually red.....

GPU Usage Optimization

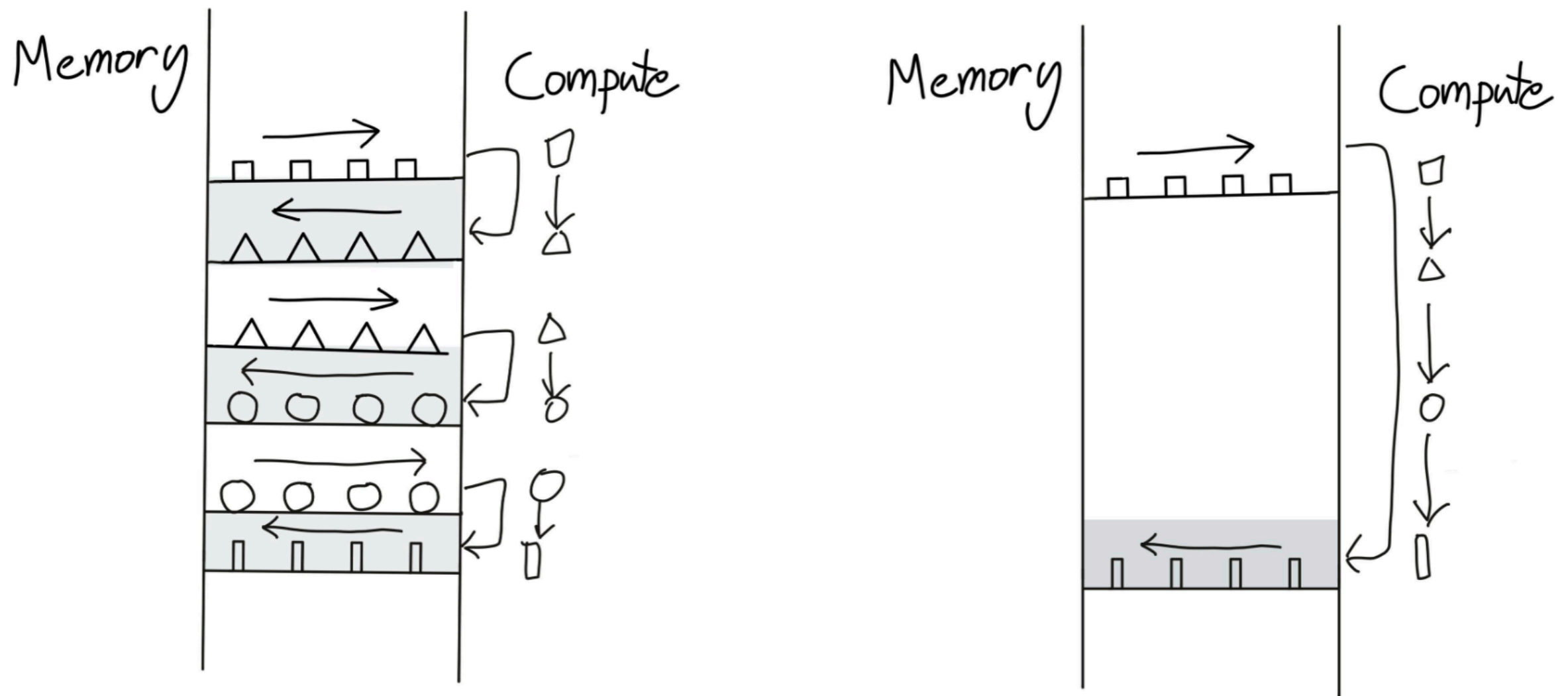
1. kernel



high-level language
slower, less control

low-level language
faster, full control

Memory Optimization

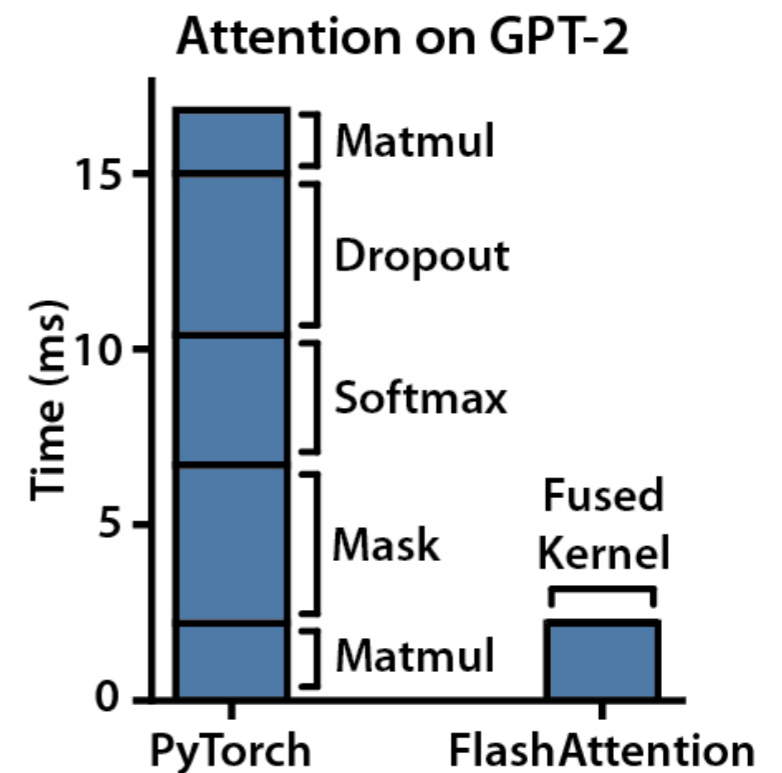
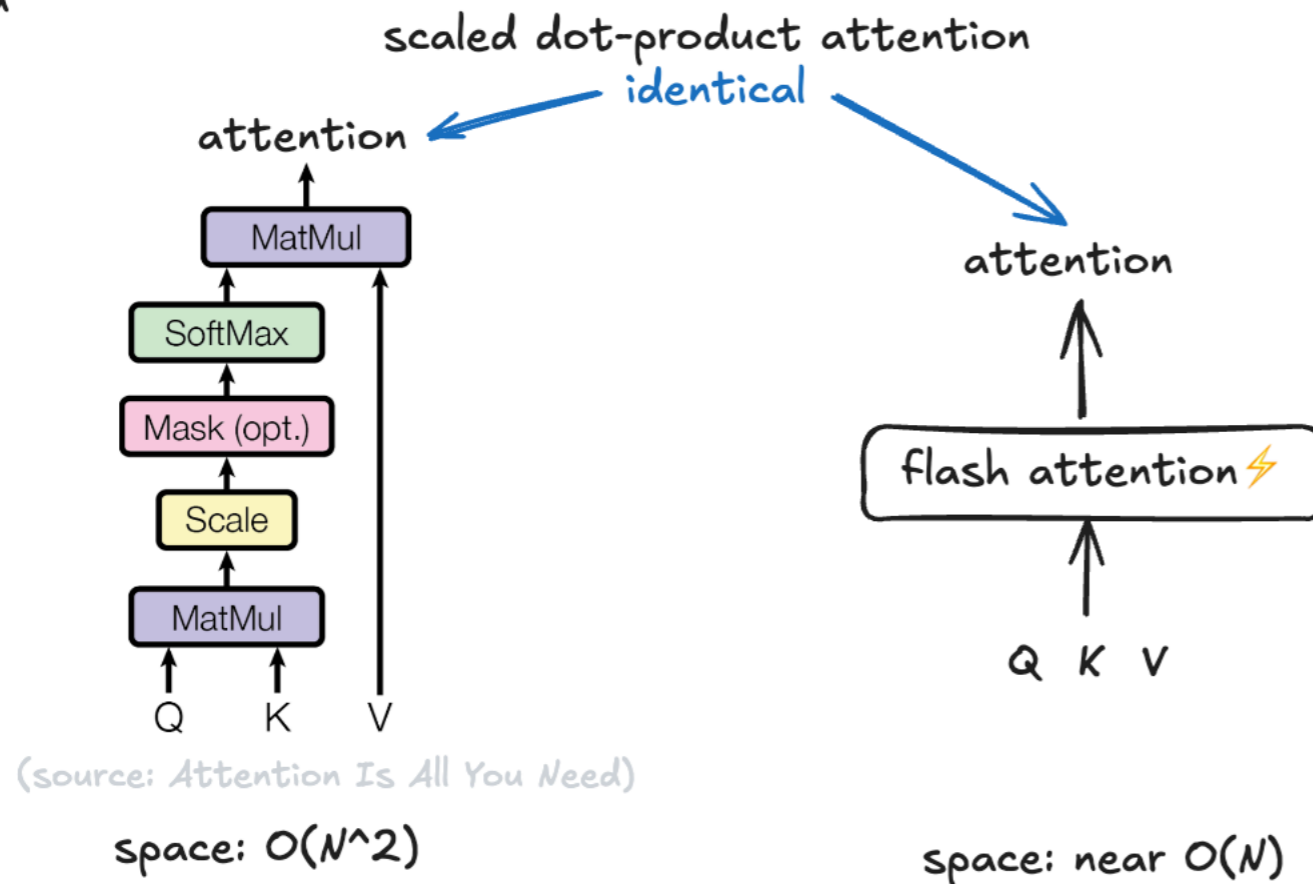


Here's how a sequence of pointwise operators might look like.

https://horace.io/brrr_intro.html

Efficient Attention

2.1 Speed



<https://youtu.be/mpuRca2UZtl?si=RierGptMLhO1p4mA>

<https://github.com/Dao-AILab/flash-attention>