Remaining Parts of Tokenization

Deadlines

- **https://people.cs.umass.edu/~hschang/cs685/schedule.html**

- **4/2:** Deadline of applying for the first round of API credit
  - https://piazza.com/class/m1kz66st9dn62i/post/146
- **4/7:** Midterm Review 1
  - Will show you the example questions and the slides that could be used to answer the question
- **4/9:** Midterm Review 2
- **4/11**: HW 2 due
- **4/18 (**Friday but Monday Schedule**)**: **Midterm**
  - Most questions are hard to answer by only watching Mohit's lecture
  - Most questions are multiple-choice problems.
    - One out of the four options is correct.
- **5/9: Final project report due**
  - I will make the novelty of the final project a bonus
  - If your scores are low, I will ask another TA or myself to provide a second opinion.
  - I understand the feelings of receiving negative feedback. Try to focus on how to address those concerns with the help of TAs.

Byte pair encoding

- Now, choose the most common pair (ug) and then merge the characters together into one symbol. Add this new symbol to the vocabulary. Then, retokenize the data

word	frequency
h+*ug*	10
p+*ug*	5
p+u+n	12
b+u+n	4
h+*ug*+s	5

character pair	frequency
un	16
h+ug	15
pu	12
p+ug	5
ug+s	5

...

Example from https://huggingface.co/transformers/tokenizer_summary.html

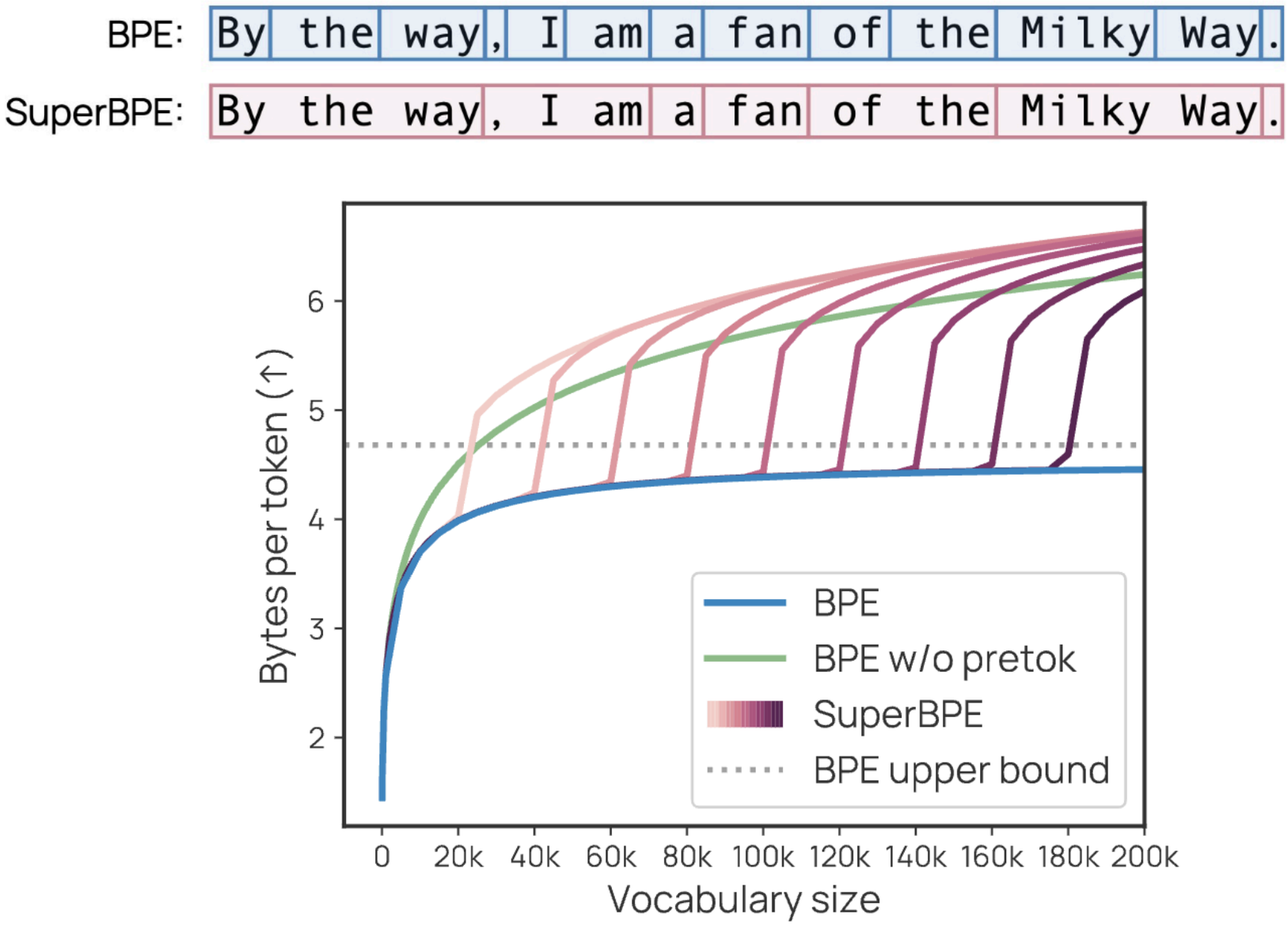
Weird LLM Behaviors from Tokenization

Tokenization is at the heart of much weirdness of LLMs. Do not brush it off.

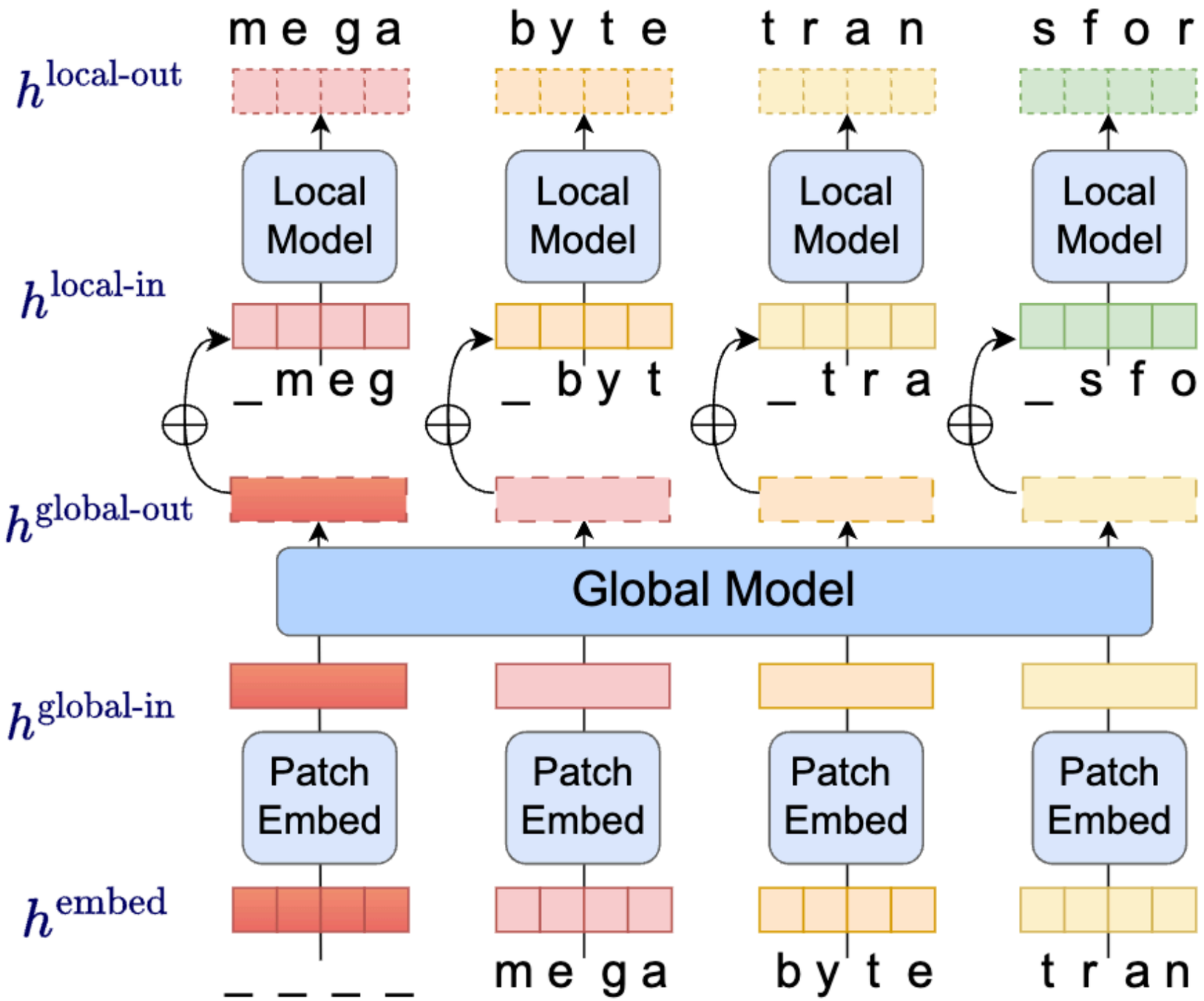
- Why can't LLM spell words? **Tokenization**.
- Why can't LLM do super simple string processing tasks like reversing a string? **Tokenization**.
- Why is LLM worse at non-English languages (e.g. Japanese)? **Tokenization**.
- Why is LLM bad at simple arithmetic? **Tokenization**.
- Why did GPT-2 have more than necessary trouble coding in Python? **Tokenization**.
- Why did my LLM abruptly halt when it sees the string "<|endoftext|>"? **Tokenization**.
- What is this weird warning I get about a "trailing whitespace"? **Tokenization**.
- Why the LLM break if I ask it about "SolidGoldMagikarp"? **Tokenization**.

Let's build the GPT Tokenizer (https://youtu.be/zduSFxRajkE?si=4CllS25JySOEQhtb)

SuperBPE



SuperBPE: Space Travel for Language Models (https://arxiv.org/pdf/2503.13423)



MEGABYTE: Predicting Million-byte Sequences with Multiscale Transformers (https://arxiv.org/pdf/2305.07185)

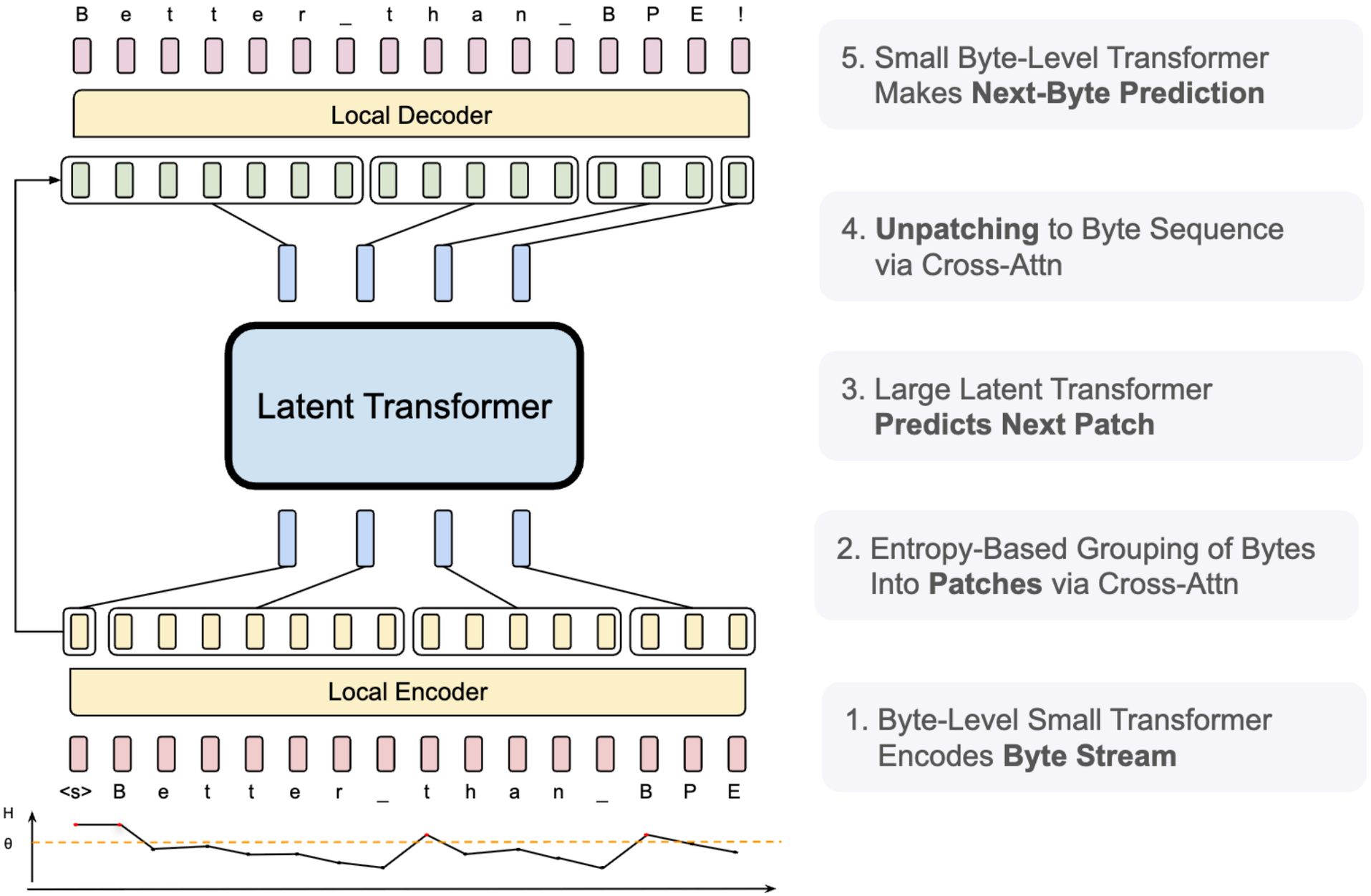
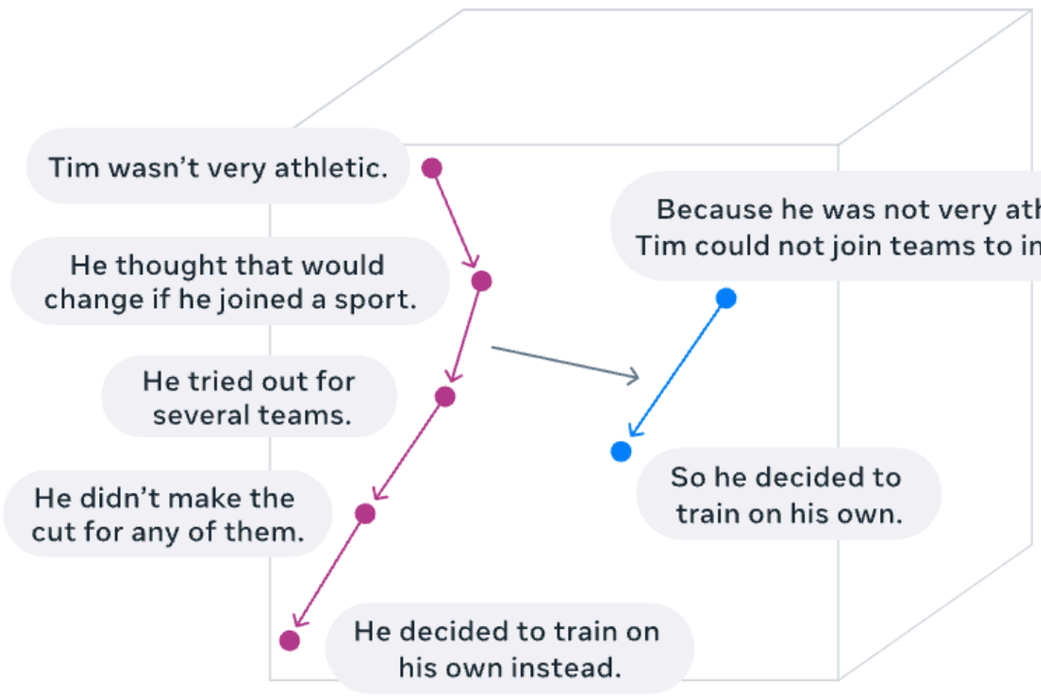


Figure 2 BLT comprises three modules, a lightweight *Local Encoder* that encodes input bytes into patch representations, a computationally expensive Latent Transformer over patch representations, and a lightweight *Local Decoder* to decode the next patch of bytes. BLT incorporates byte n-gram embeddings and a cross-attention mechanism to maximize information flow between the Latent Transformer and the byte-level modules (Figure 5). Unlike fixed-vocabulary tokenization, BLT dynamically groups bytes into patches preserving access to the byte-level information.

Byte Latent Transformer: Patches Scale Better Than Tokens (https://arxiv.org/pdf/2412.09871)

Large Concept Model



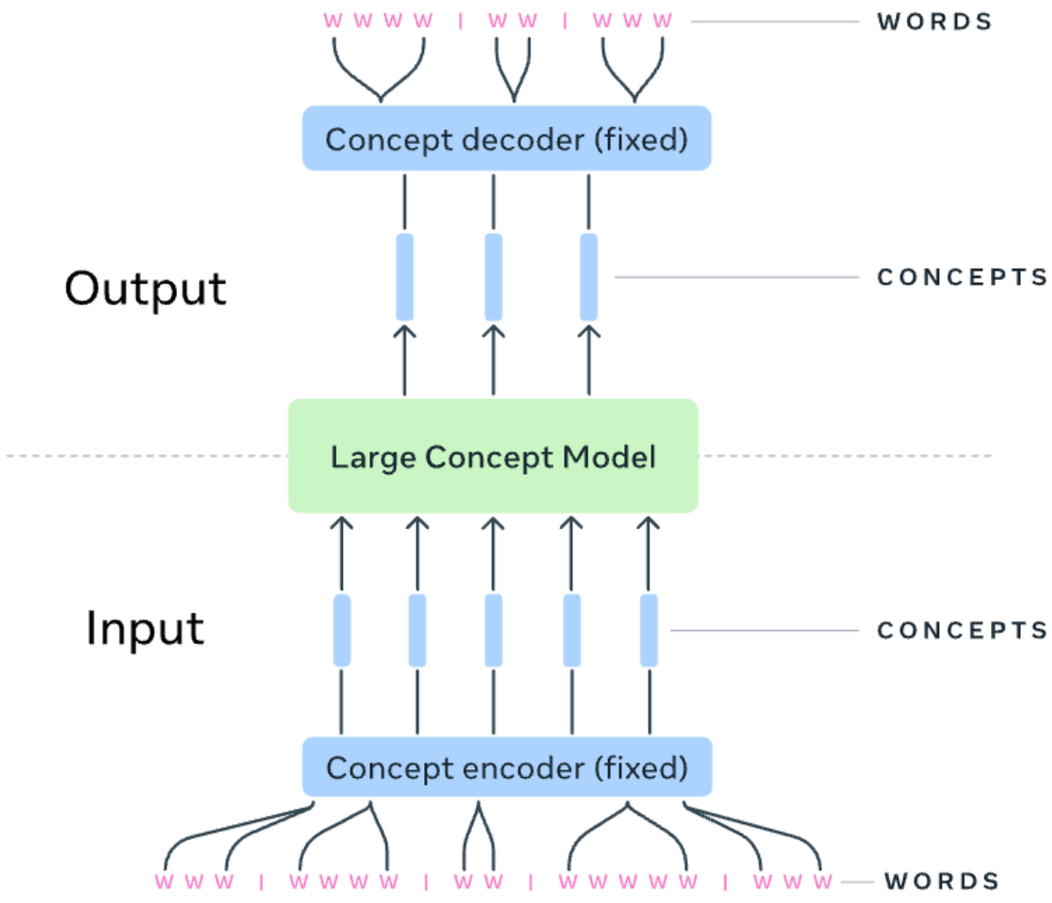


Figure 1 - Left: visualization of reasoning in an embedding space of concepts (task of summarization). Right: fundamental architecture of an LARGE CONCEPT MODEL (LCM). ⋆: concept encoder and decoder are frozen.

Large Concept Models: Language Modeling in a Sentence Representation Space (https://arxiv.org/abs/2412.08821)

Decoding from language models

CS685 Spring 2025

Advanced Natural Language Processing

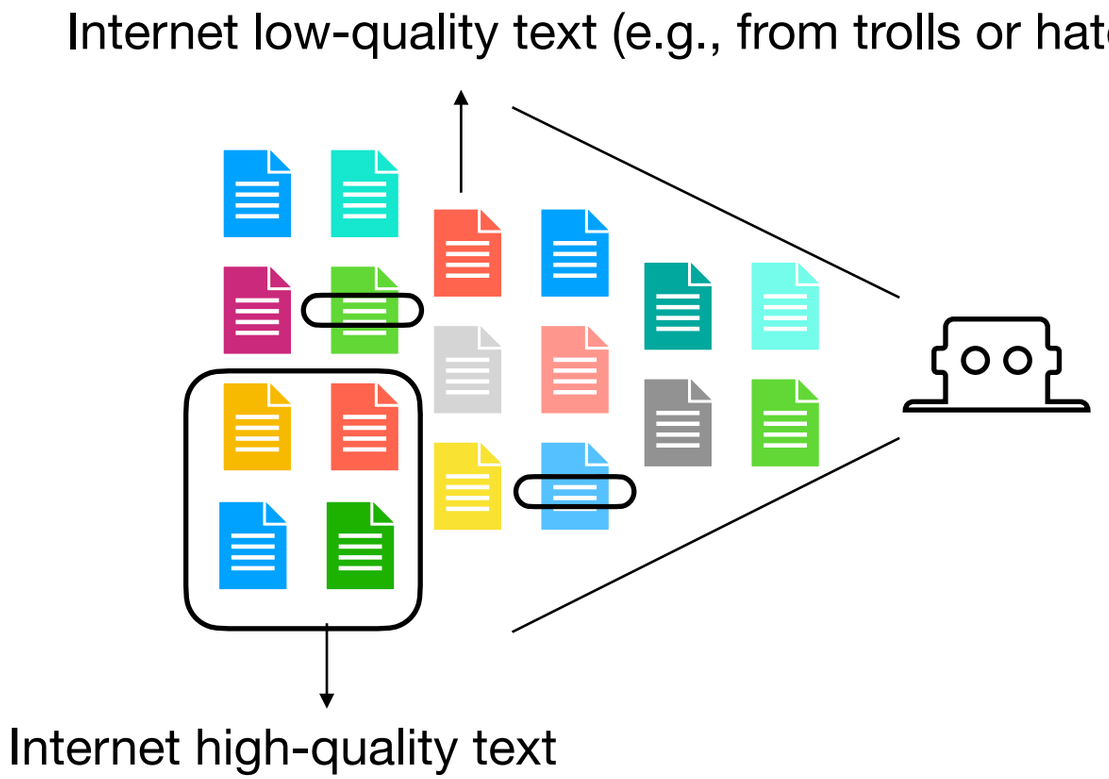
Haw-Shiuan Chang

College of Information and Computer Sciences

University of Massachusetts Amherst

Many slides from Mohit Iyyer

LLM Development



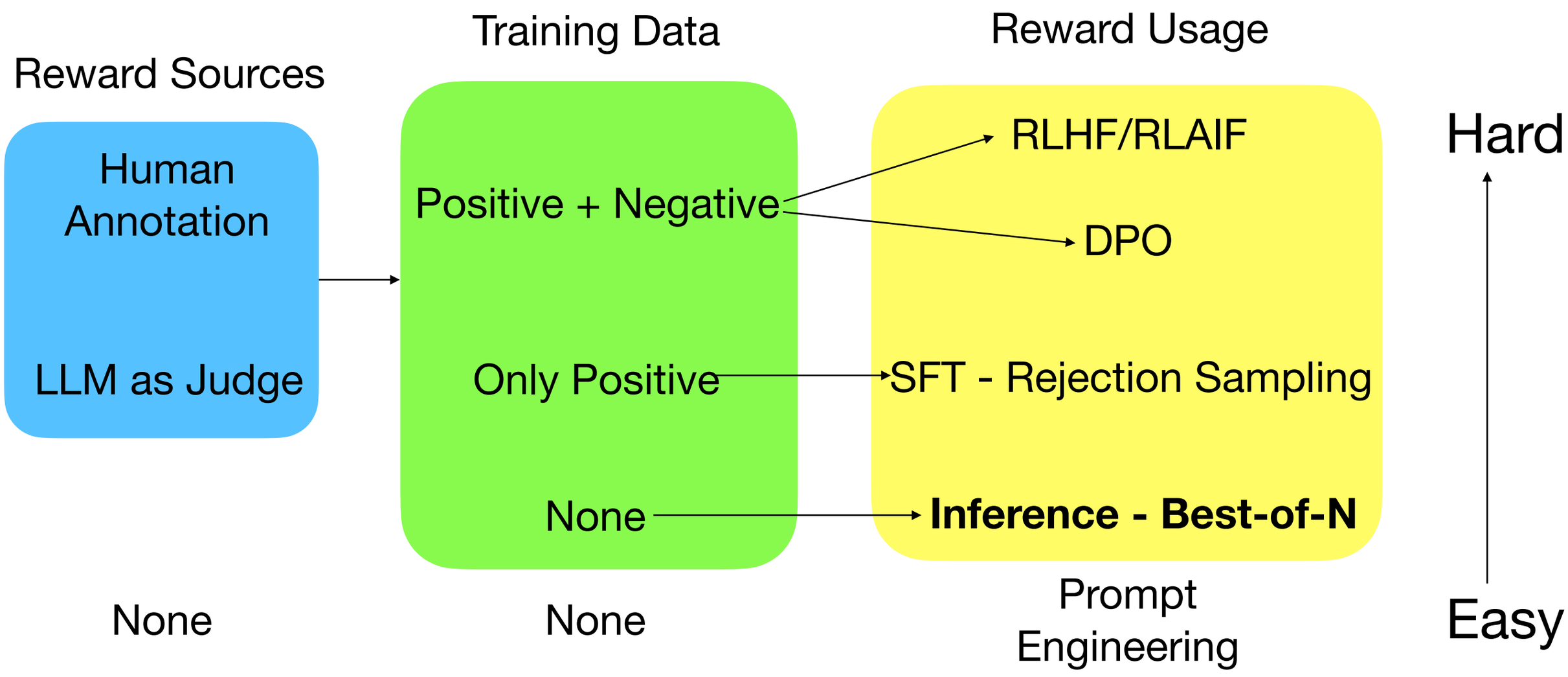
- Architectures
  - MLP
  - RNN
  - Transformer
- Training Stages
  - Pretraining
  - Supervised Fine-tuning (SFT)
  - Alignment
    - Learning from Human Feedback (LHF)
    - Reasoning

How do you improve LLM's performance without training?

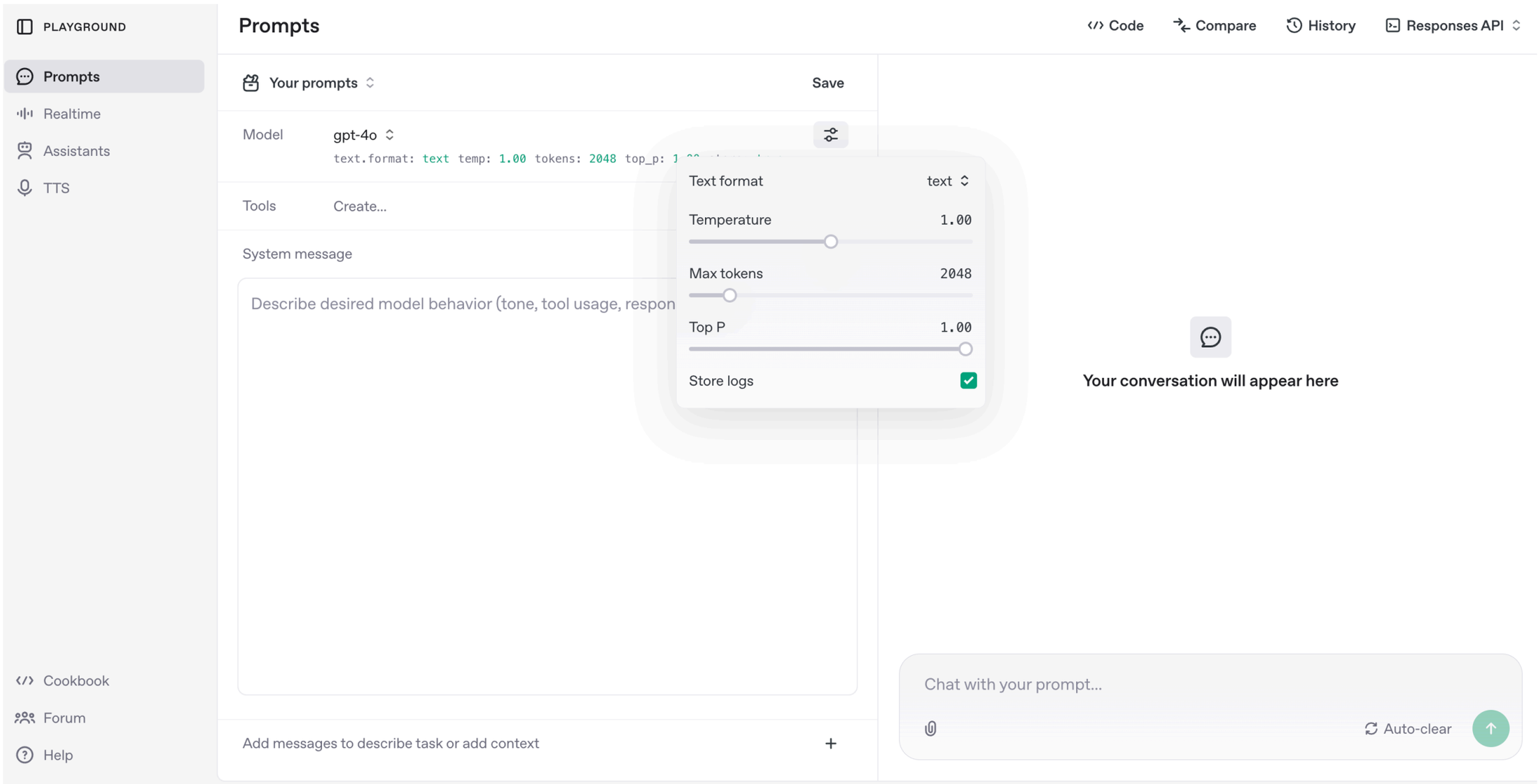
Inference-time Improvement

- Prompt engineering
  - In-context learning
- **Decoding**
- Agentic (won’t cover too much)
  - Tools
  - RAG
  - Multi-LLM collaboration

Inference Methods



The Meaning of Decoding Hyperparameters



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

Decoding

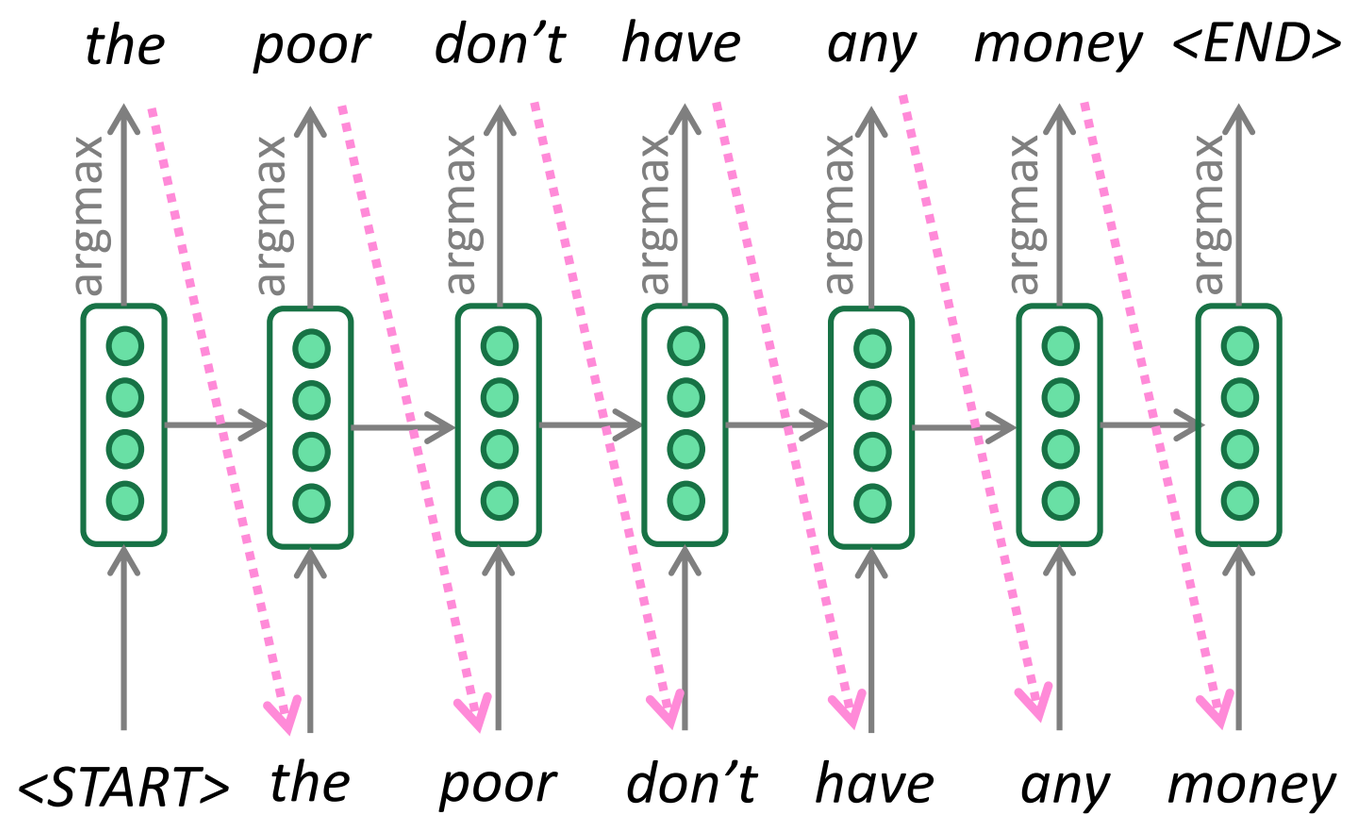
- Given a pretrained LM and a prefix, how do we generate the most probable continuation (or *any* probable continuation) of that prefix?

- More concretely, how do we find

$$\arg\max \prod_{i}^{L} p(w_i \mid w_1, w_2, \ldots, w_{i-1}, \text{prefix})$$

- Can we enumerate all possible generations given the prefix and then choose the one with the highest probability?

easiest option: **greedy decoding**



issues?

11

How many forward passes do you need for greedy decoding?

Beam search

- in greedy decoding, we cannot go back and revise previous decisions!
  - *les pauvres sont démunis (the poor don't have any money)*
  - → *the* ____
  - → *the poor* ____
  - → *the poor are* ____

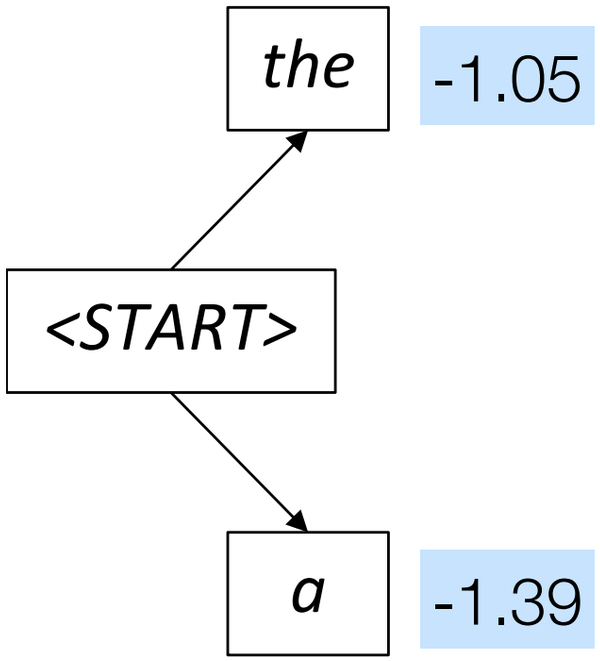
- fundamental idea of beam search: explore several different hypotheses instead of just a single one
  - keep track of *k* most probable partial translations at each decoder step instead of just one!

    the beam size *k* is usually 5-10

Beam search decoding: example

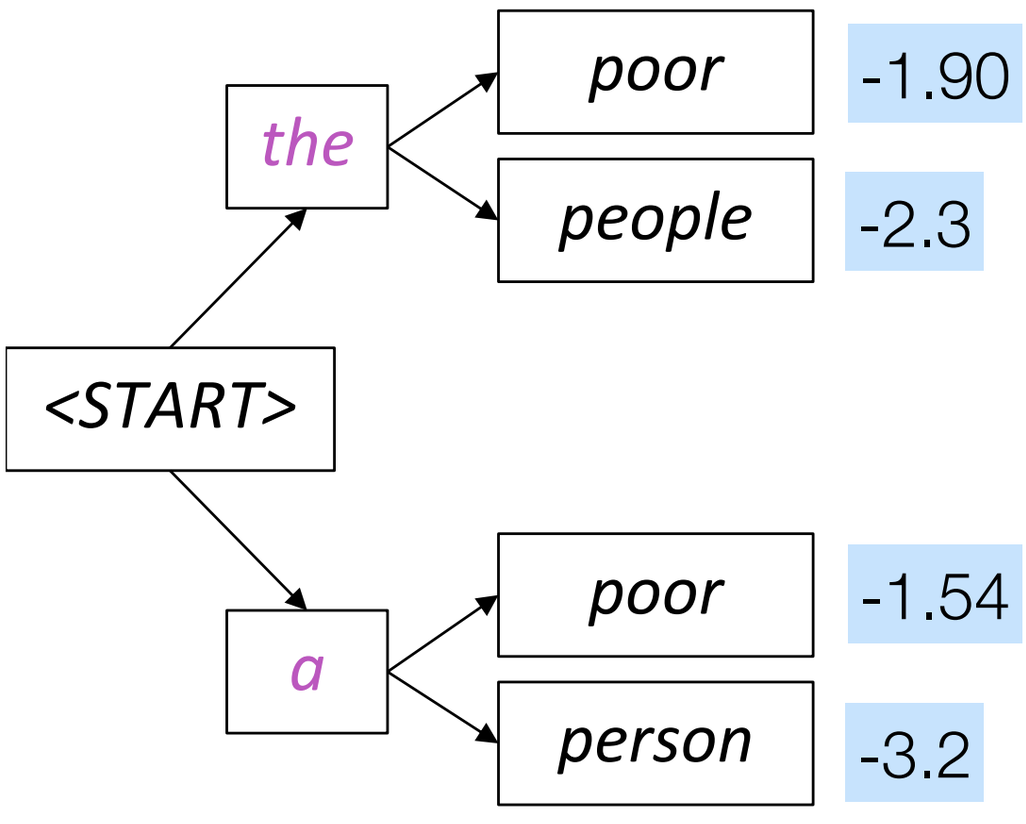
Beam size = 2

$$\log(\frac{exp(w^T x)}{\sum exp(Wx)})$$



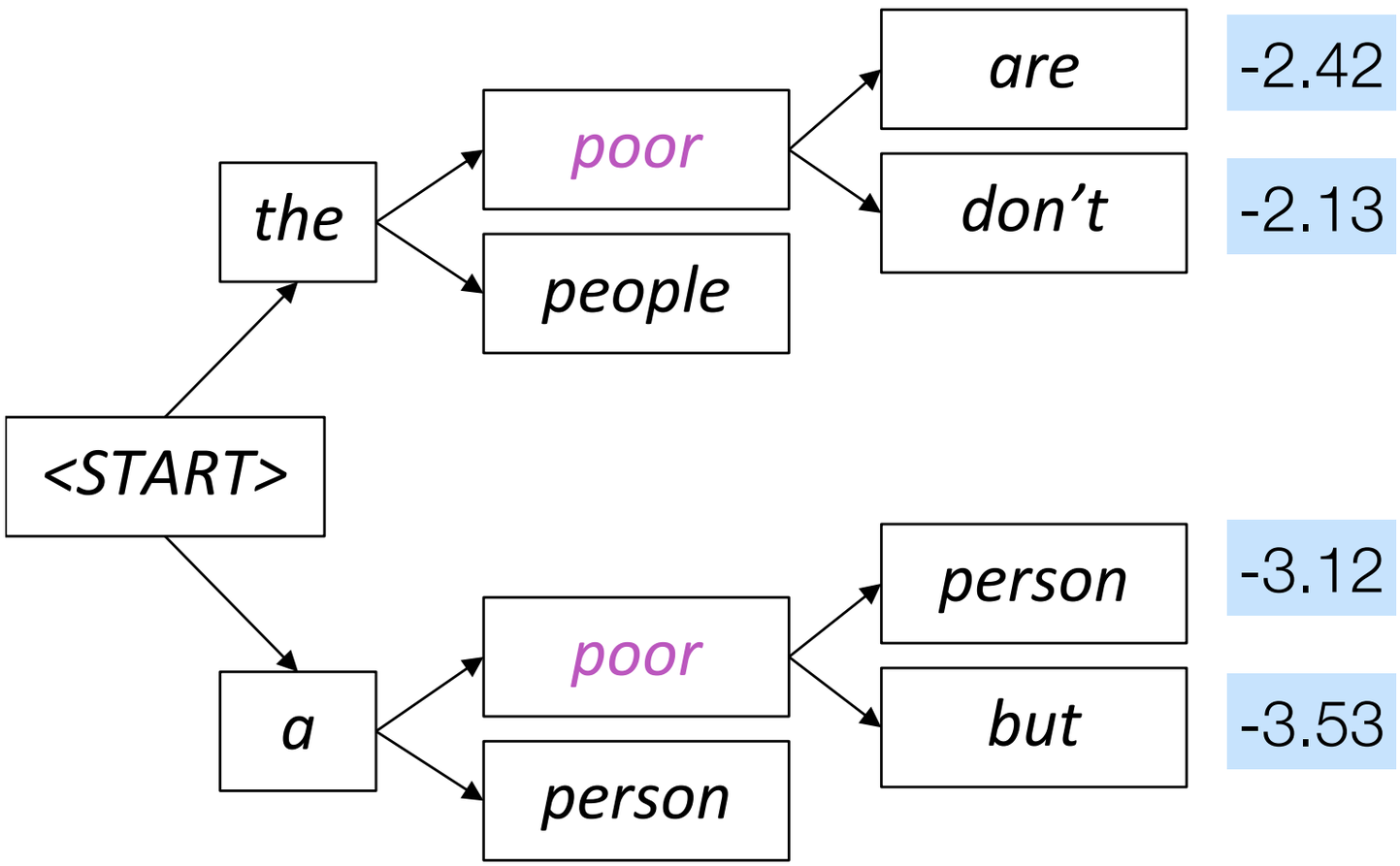
Beam search decoding: example

Beam size = 2



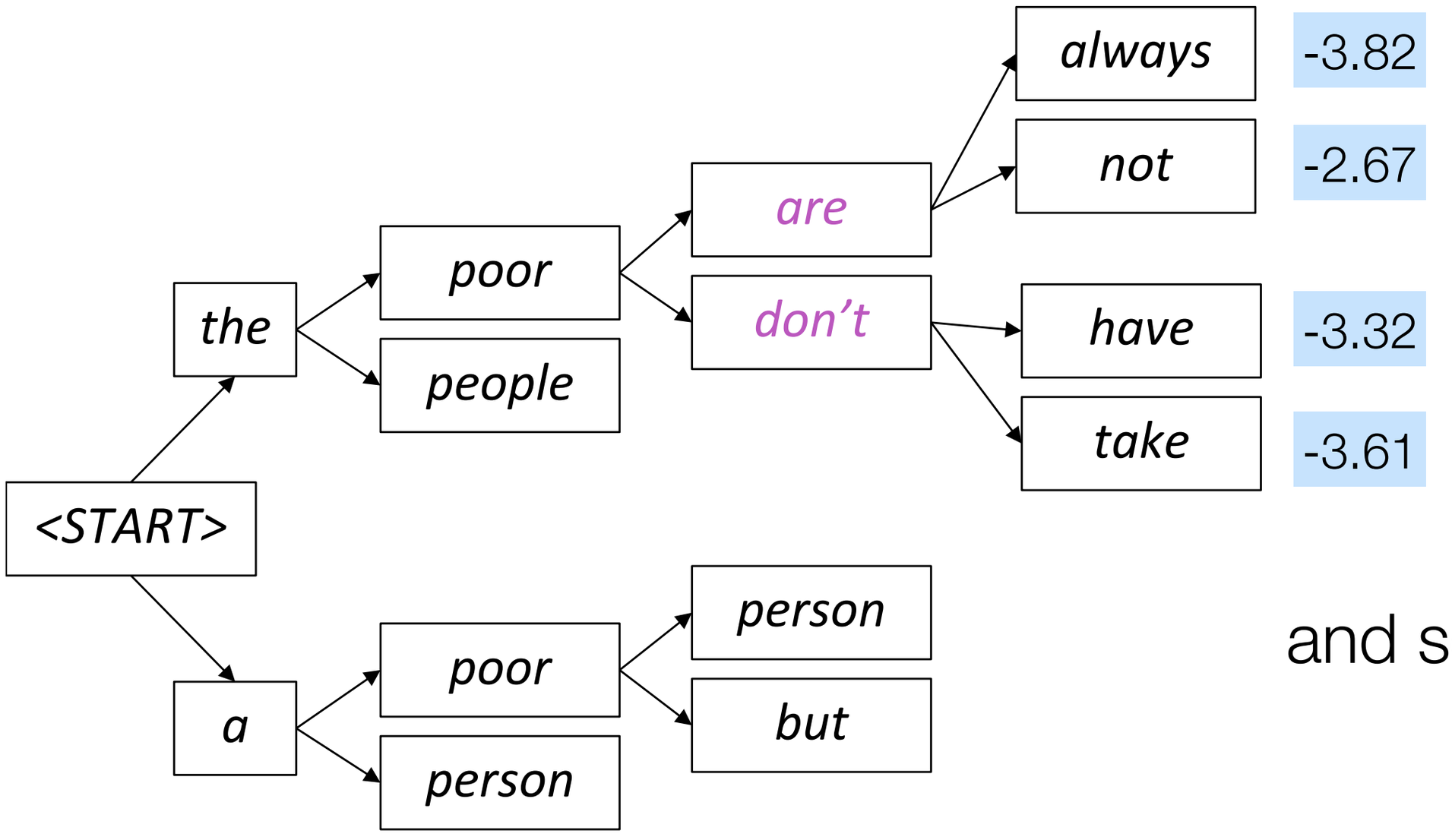
Beam search decoding: example

Beam size = 2



Beam search decoding: example

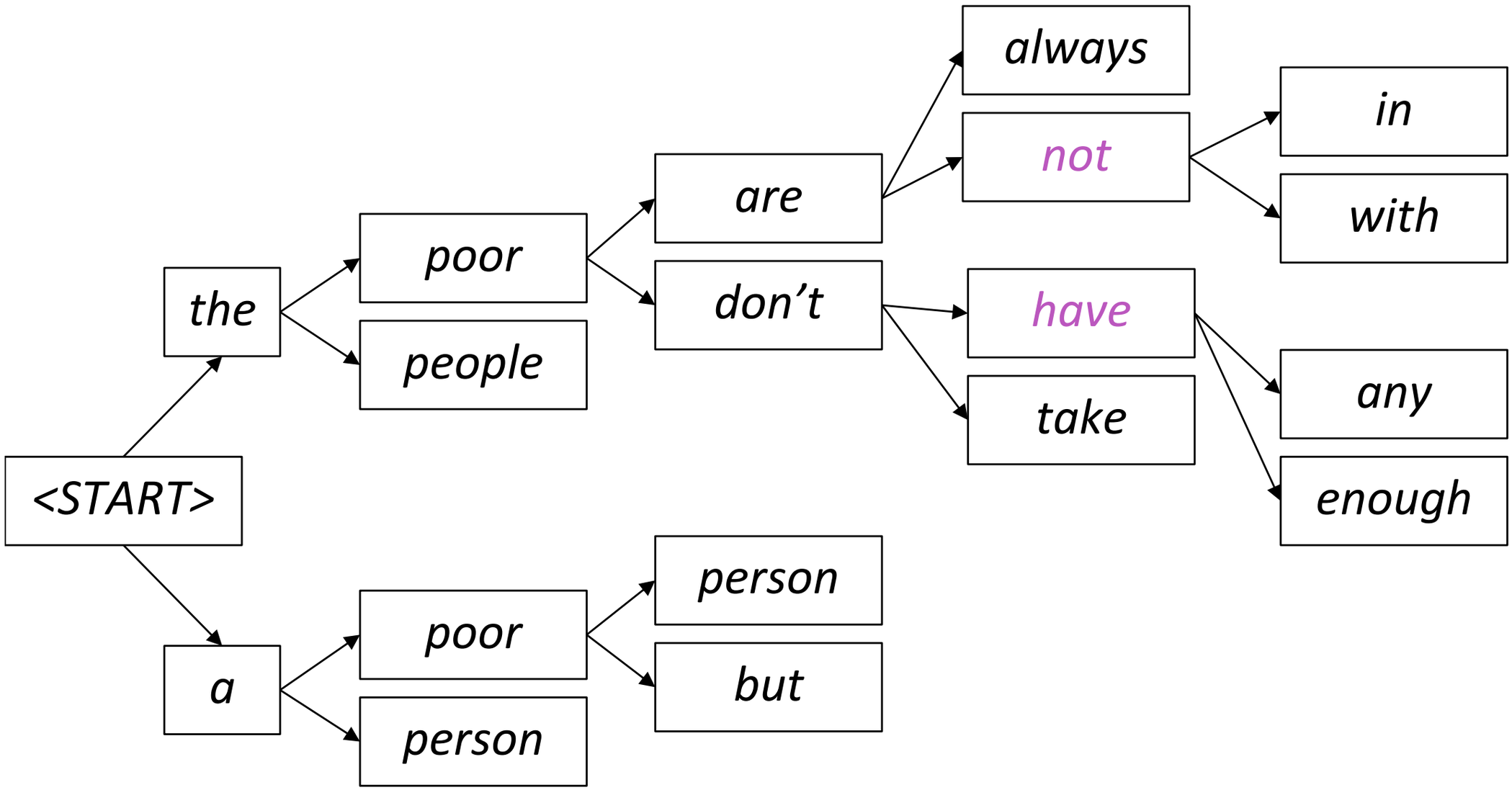
Beam size = 2



and so on…

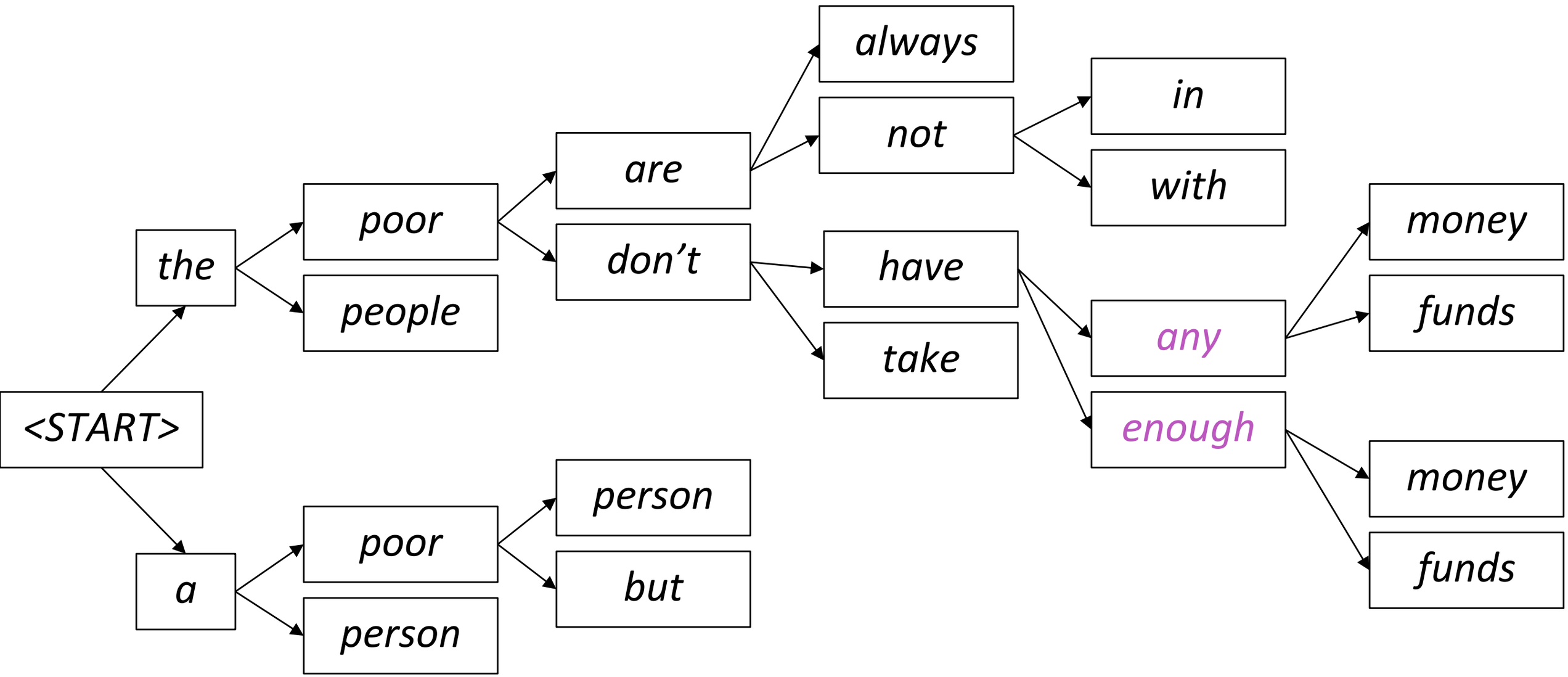
Beam search decoding: example

Beam size = 2



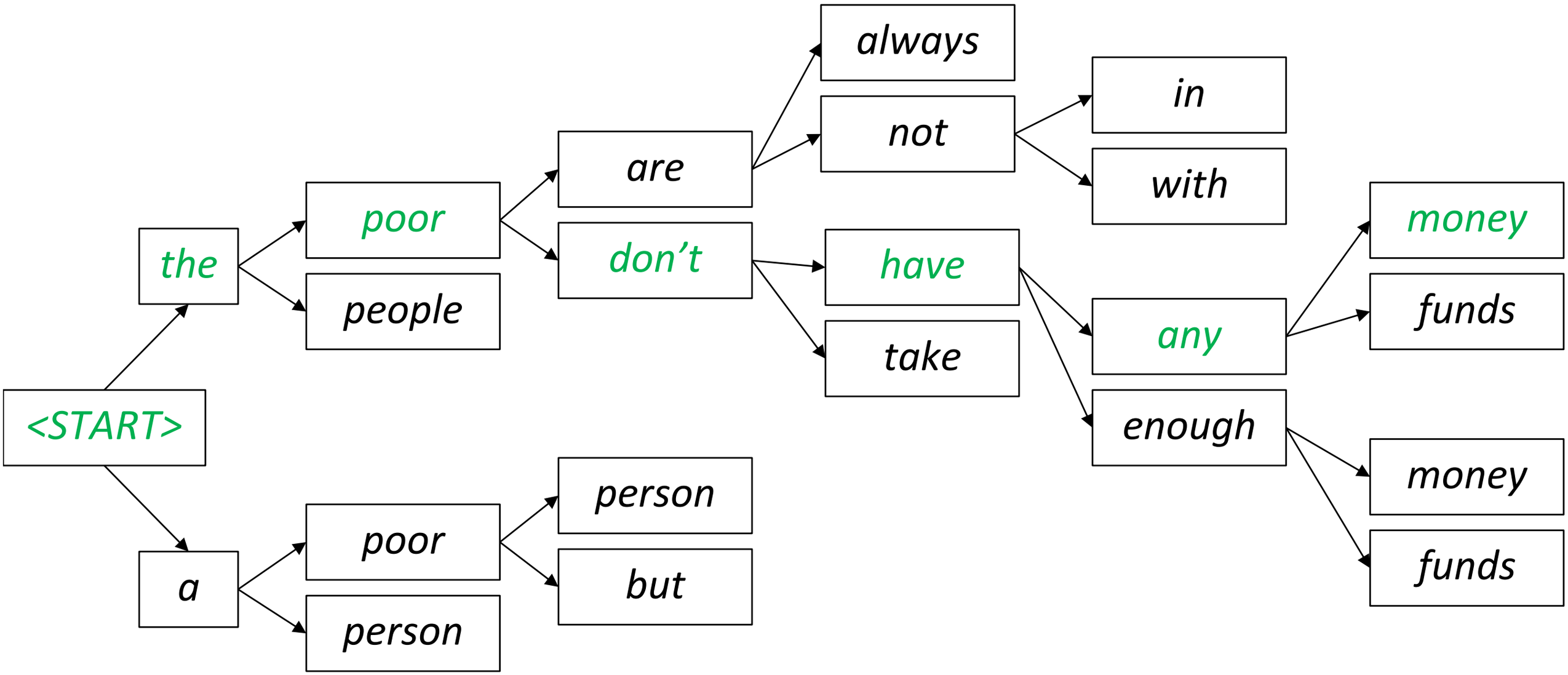
Beam search decoding: example

Beam size = 2



Beam search decoding: example

Beam size = 2



Does beam search always return the most probable continuation of the prefix?

What are the termination conditions of beam search?

How many forward passes do you need for beam search?

Is a More Probable Sequence a Better Sequence?

- Beam search can bring the most improvement in which of the following tasks?

- Translation
- Summarization
- Story generation
- Long-form question answering

What's the effect of changing beam size *k*?

- Small *k* has similar problems to greedy decoding (*k*=1)
  - Ungrammatical, unnatural, nonsensical, incorrect

- Larger *k* means you consider more hypotheses
  - Increasing *k* reduces some of the problems above
  - Larger *k* is more computationally expensive

- But increasing *k* can introduce other problems:
  - For NMT, increasing *k* too much decreases BLEU score (Tu et al, Koehn et al). This is primarily because large-k beam search produces too-short translations (even with score normalization!)
  - In open-ended tasks like chit-chat dialogue, large *k* can make output more generic (see next slide)

Neural Machine Translation with Reconstruction, Tu et al, 2017 https://arxiv.org/pdf/1611.01874.pdf
Six Challenges for Neural Machine Translation, Koehn et al, 2017 https://arxiv.org/pdf/1706.03872.pdf

Effect of beam size in chitchat dialogue

I mostly eat a fresh and raw diet, so I save on groceries

Human chit-chat partner

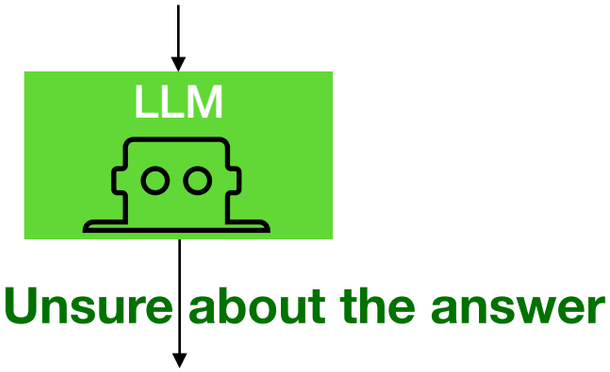
Beam size	Model response
1	*I love to eat healthy and eat healthy*
2	*That is a good thing to have*
3	*I am a nurse so I do not eat raw food*
4	*I am a nurse so I am a nurse*
5	*Do you have any hobbies?*
6	*What do you do for a living?*
7	*What do you do for a living?*
8	*What do you do for a living?*

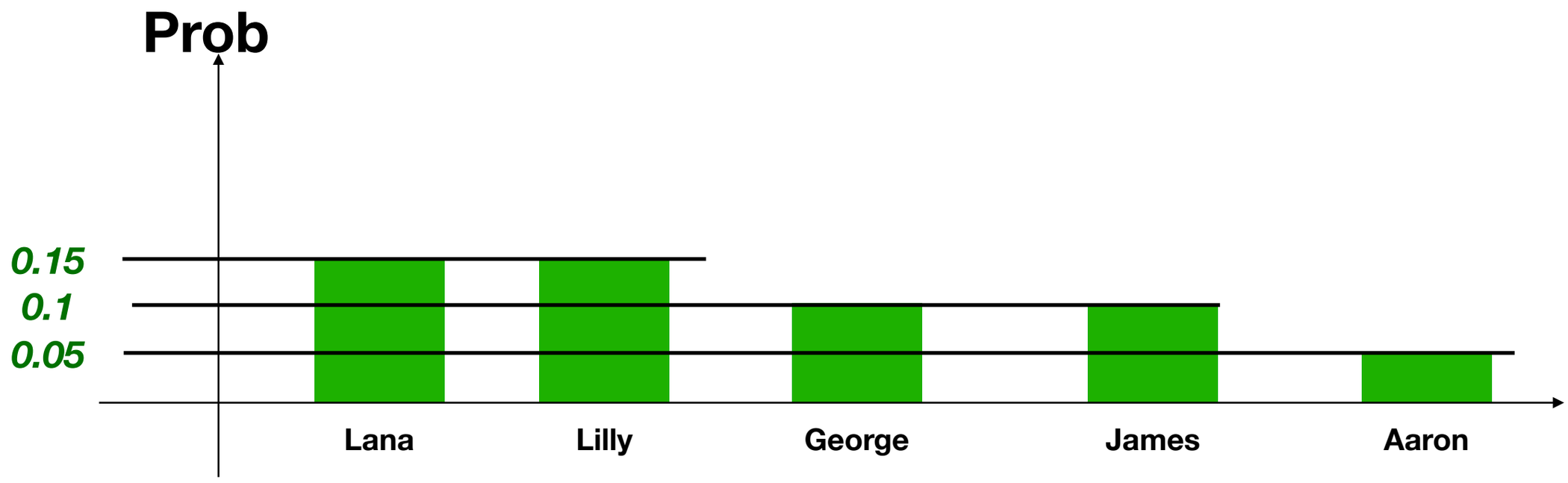
Low beam size:
More on-topic but nonsensical; bad English

High beam size:
Converges to safe, “correct” response, but it’s generic and less relevant

Pure Sampling (Ancestral Samling)

The screenwriter of The Matrix is ___





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

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

Sampling-based decoding

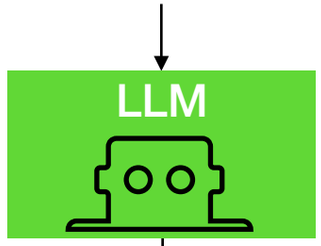
Both of these are **more efficient** than beam search – no multiple hypotheses

- Pure sampling (ancestral sampling)
  - On each step *t*, randomly sample from the probability distribution P_t to obtain your next word.
  - Like greedy decoding, but sample instead of argmax.
- Top-n sampling*
  - On each step *t*, randomly sample from P_t, restricted to just the top-n most probable words
  - Like pure sampling, but truncate the probability distribution
  - *n=1* is greedy search, *n=V* is pure sampling
  - Increase *n* to get more diverse/risky output
  - Decrease *n* to get more generic/safe output

*Usually called top-*k* sampling, but here we're avoiding confusion with beam size *k*

Top-p Sampling (Nucleus Samling)

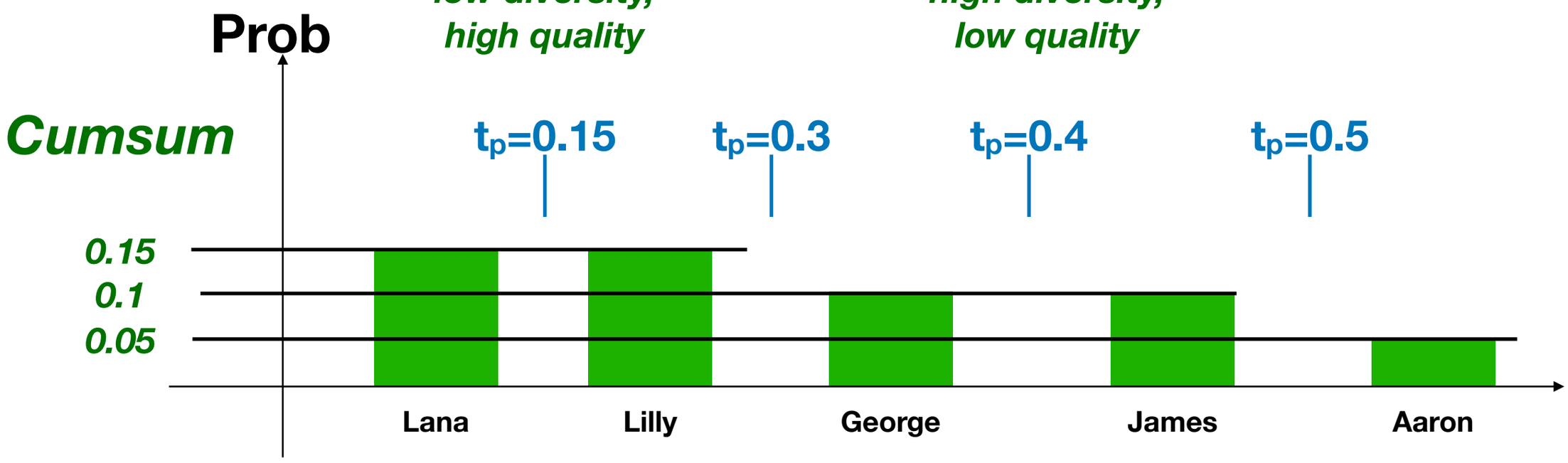
The screenwriter of The Matrix is ___



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

WebText

Beam Search, b=16

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Why does beam search encourage repetition?

The Curious Case of Neural Text Degeneration, Holtzman et al., 2020

WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

The Curious Case of Neural Text Degeneration, Holtzman et al., 2020

WebText

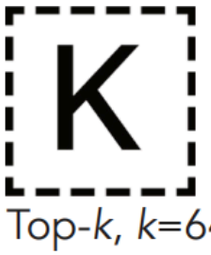
An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.



Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

The Curious Case of Neural Text Degeneration, Holtzman et al., 2020

WebText

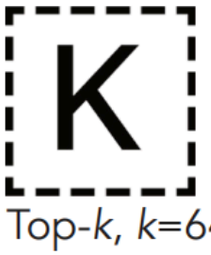
An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.

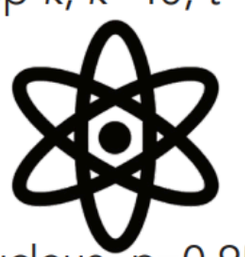


Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at:
www.abc.net.au/environment/species-worry/
in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.



Nucleus, *p*=0.95

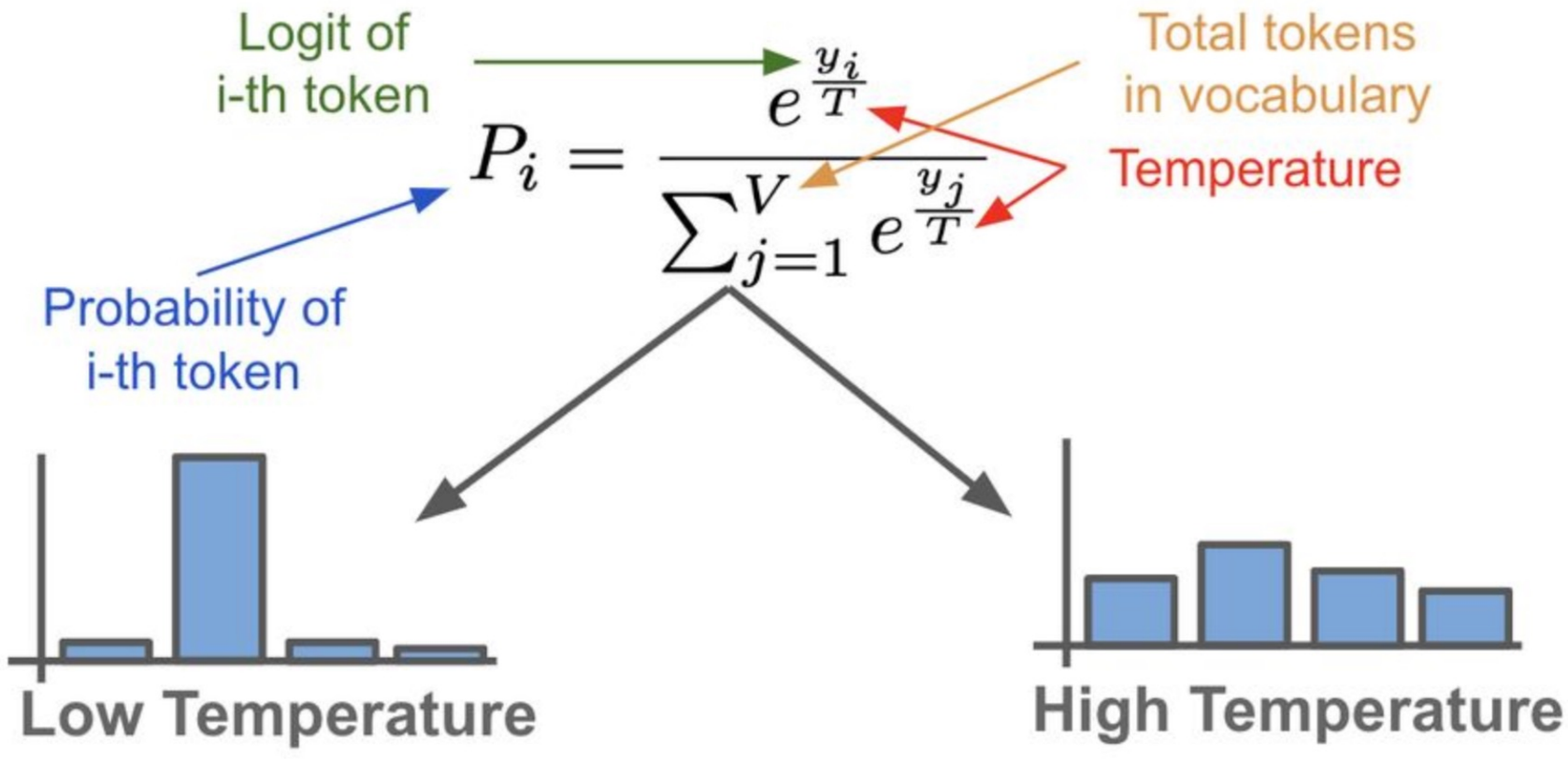
There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

Before RLHF: High Diversity -> Low Repetition

The Curious Case of Neural Text Degeneration, Holtzman et al., 2020

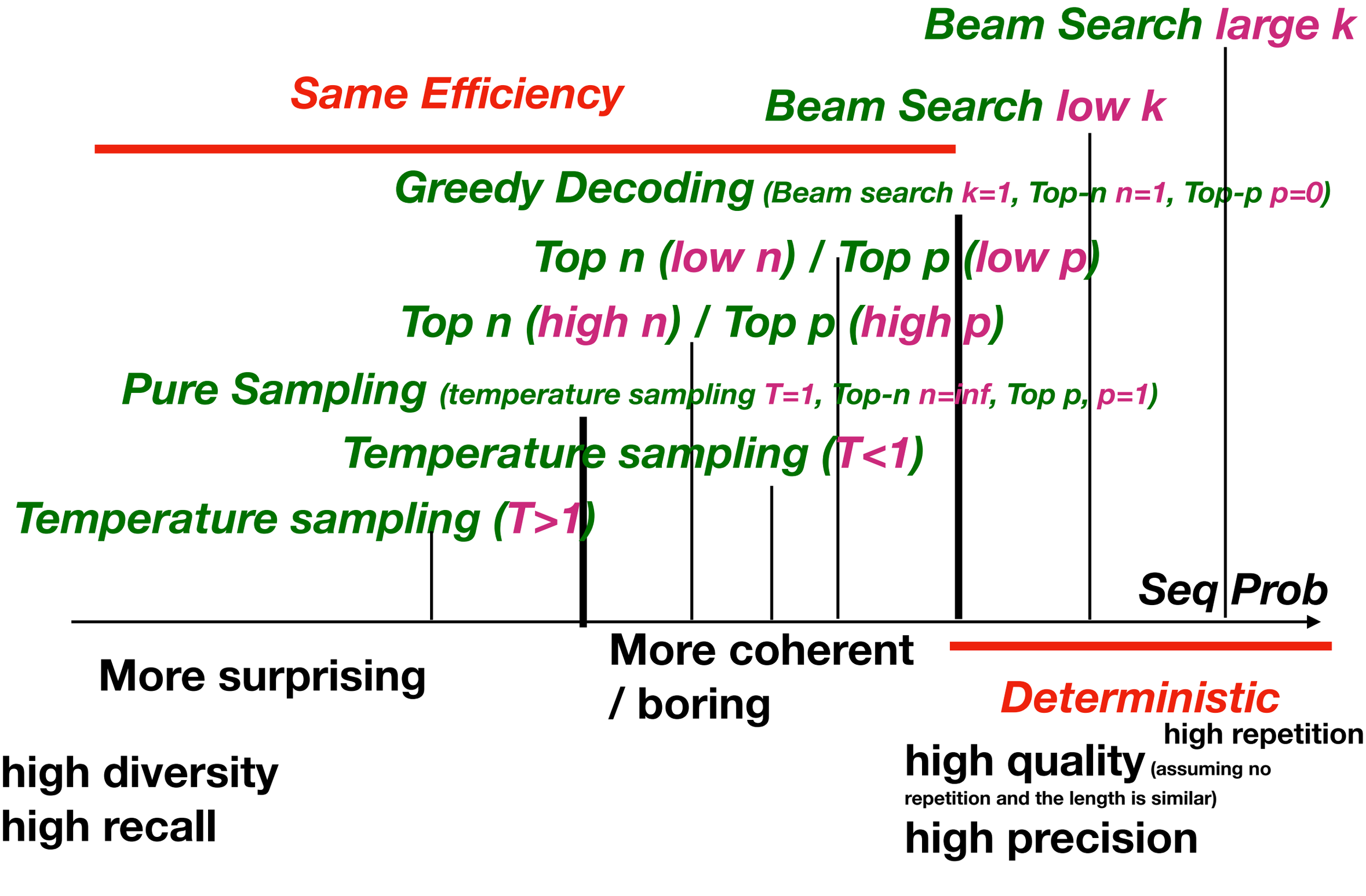
Temperature Sampling

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

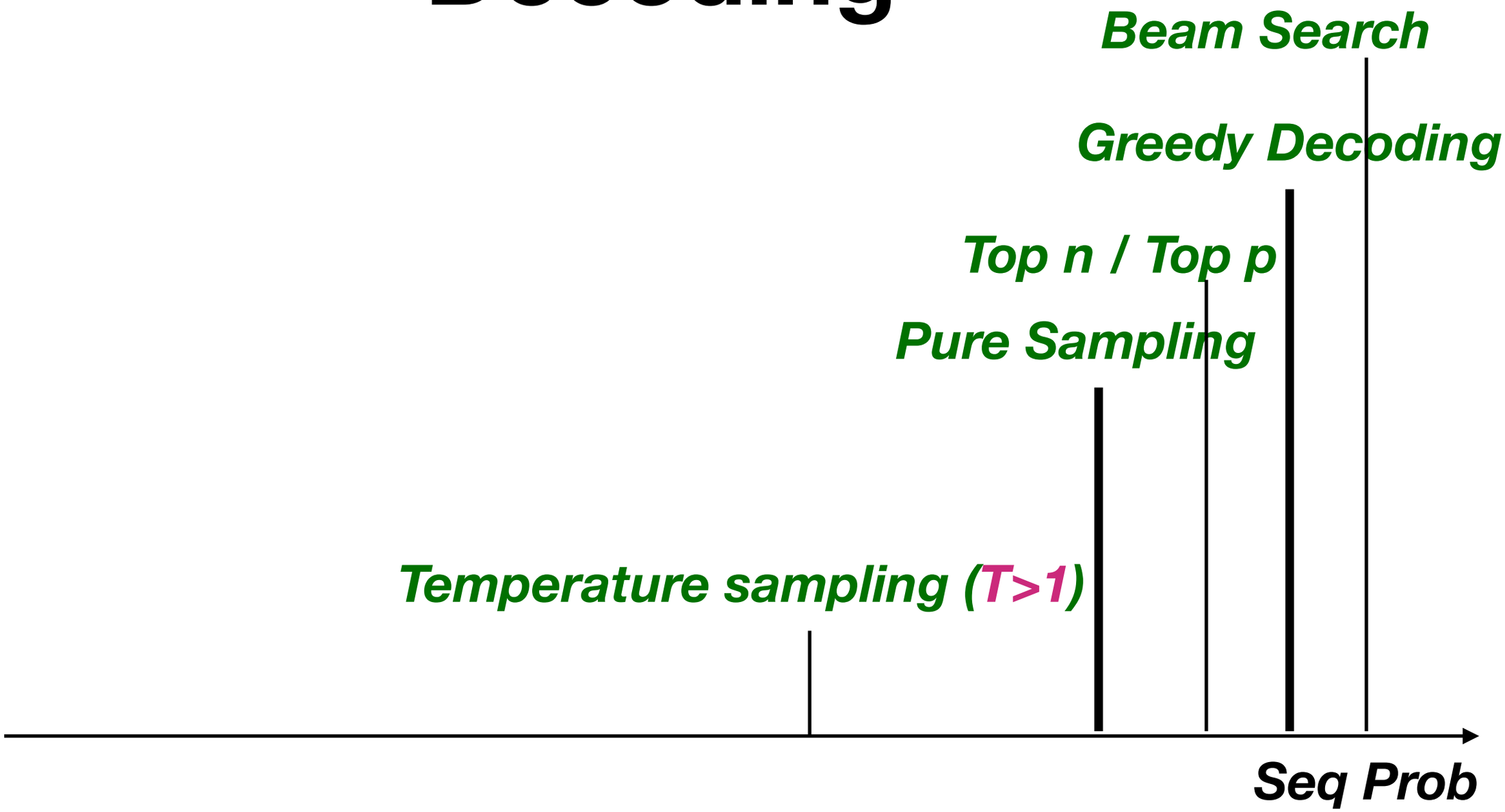


https://aman.ai/primers/ai/token-sampling/

Base LLM Decoding



Instruct LLM (after RLHF)
Decoding



high diversity
high recall

high quality
high precision