Decoding Strategies

Wei Xu

(some slides from Greg Durrett)

Course Project

A lightweight Proposal

- Group project: forming a team of 2-4 students
- 1~2 pages, excluding references:
  - A project title
  - The names of all group members
  - A brief description of your project ideas and proposed plan

- Suggestions:
  - A brief review of relevant literature
  - Preliminary results
  - Potential novel contributions

Place to start?

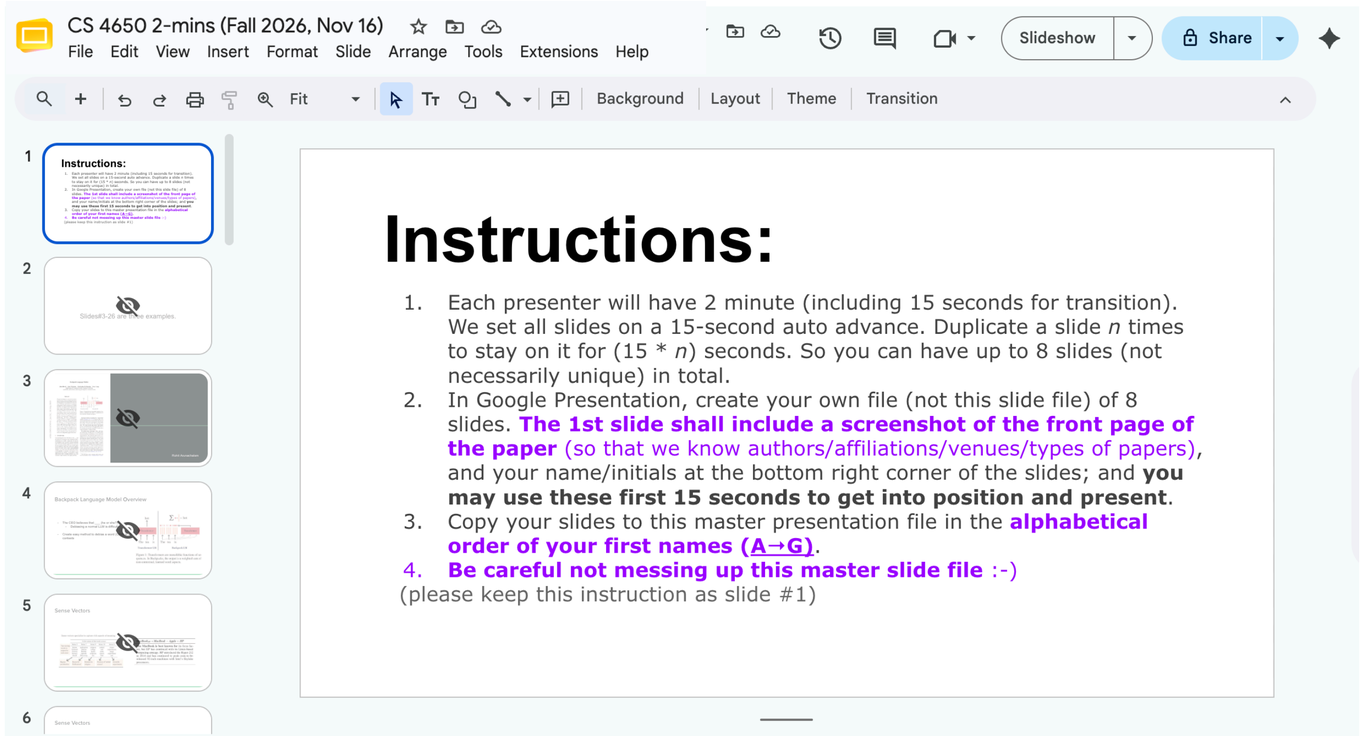
- Some people collect their own data for a project — **we like that!**
  - You can create and annotate a good amount of data
  - You can find some data in the wild, e.g., websites, social media, etc.
  - Look at research papers to see what data they use, how they get it
  - e.g., arXiv, NeurIPS Evaluations & Datasets (E&D) Track
- Others make use of existing datasets built by other researchers
  - Datasets used in other papers
  - You shall analyze the quality of the datasets you plan to use

Some example research directions

- Multilingual LLM evaluation
- Reasoning
- Domain-specific evaluation
- Memorization
- Dialects & stylistics
- Computational journalism
- etc.

2-min madness

- **In-class presentation of a recent research paper (2%)**
  - read recent publications (ACL/NAACL/EMNLP/ICLR/ICML/NeurIPS)
  - pick one and give a 2-minute oral presentation
    (detailed instructions on the submission will be released later)



Why 2-min madness?

- This will serve multiple purposes:
  - get started to think about what to do for the final course project
  - get to see a wide range coverage of topics on the latest NLP research
  - practice oral presentation skills
  - get to know other students in the class!

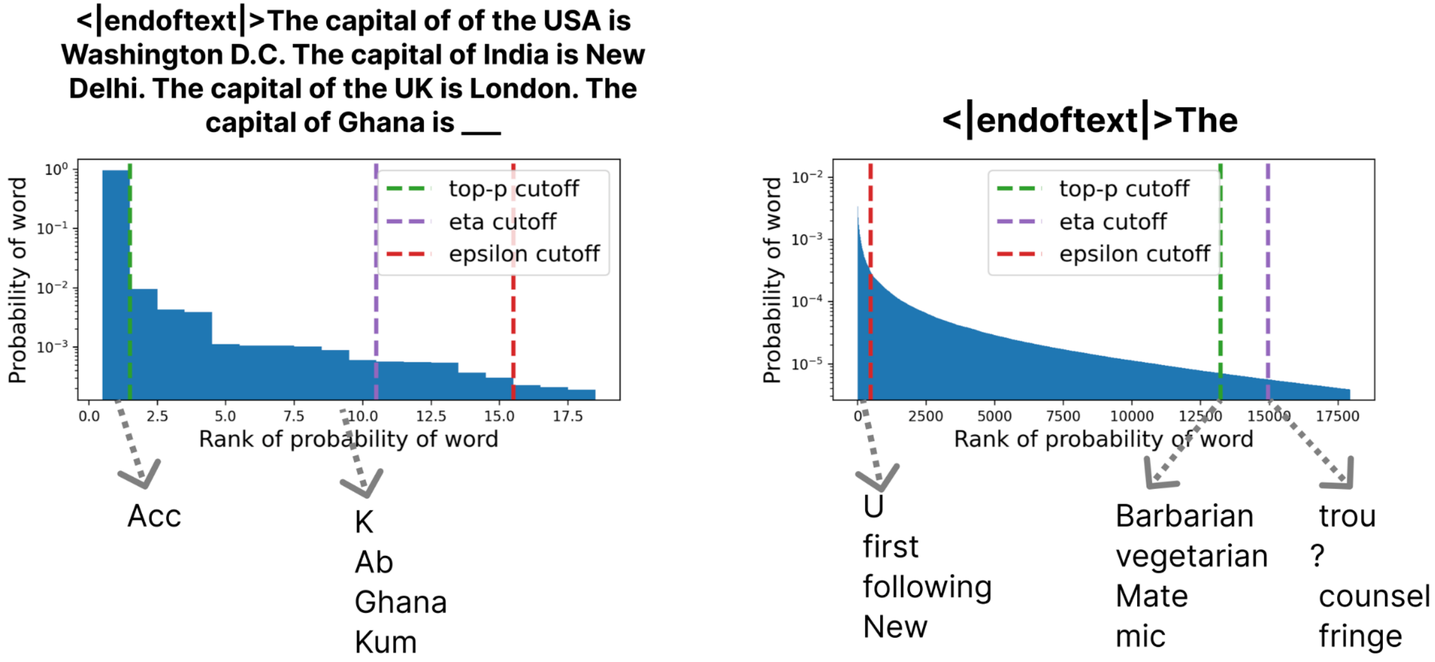
Decoding Strategies

Decoding Strategies

- LMs place a distribution $P(y_i | \ y_1, \ldots, y_{i-1})$
- seq2seq models place a distribution $P(y_i | \mathbf{x}, y_1, \ldots, y_{i-1})$
- Generation from both models looks similar; how do we do it?

Decoding

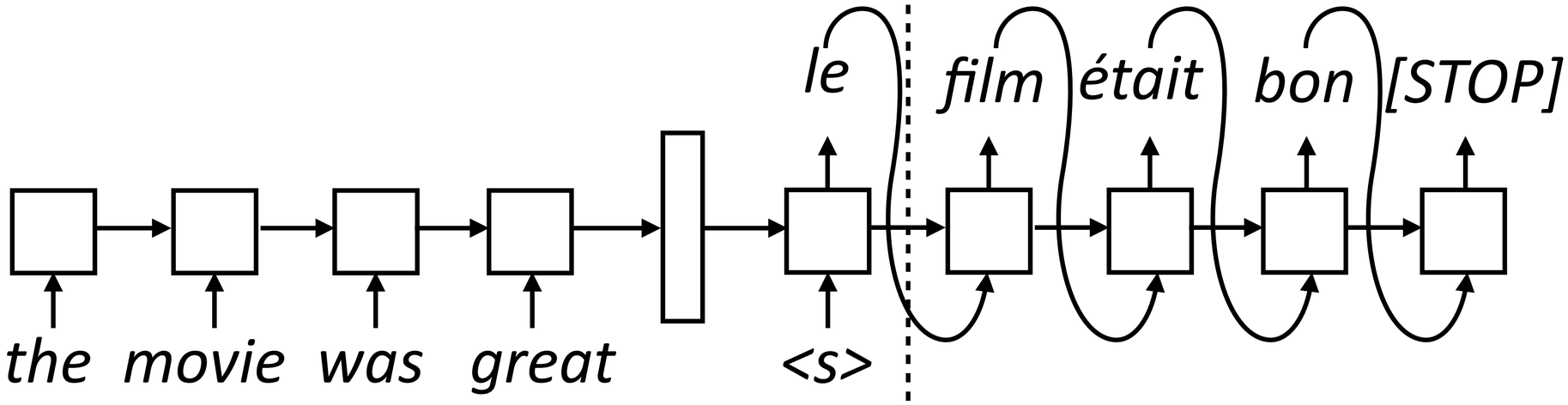
- Decoding is an inference-time solution to optimize LLM outputs
- Besides data and model size, inference-time algorithms can make a big impact.



Hewitt et al. (2022)

(Recap) Greedy Decoding

- Generate next word conditioned on previous word as well as hidden state



- During inference: need to compute the argmax over the word predictions and then feed that to the next RNN state. This is **greedy decoding**

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \text{softmax}(W\bar{h})$$

$$y_{\text{pred}} = \text{argmax}_y P(y|\mathbf{x}, y_1, \ldots, y_{i-1})$$

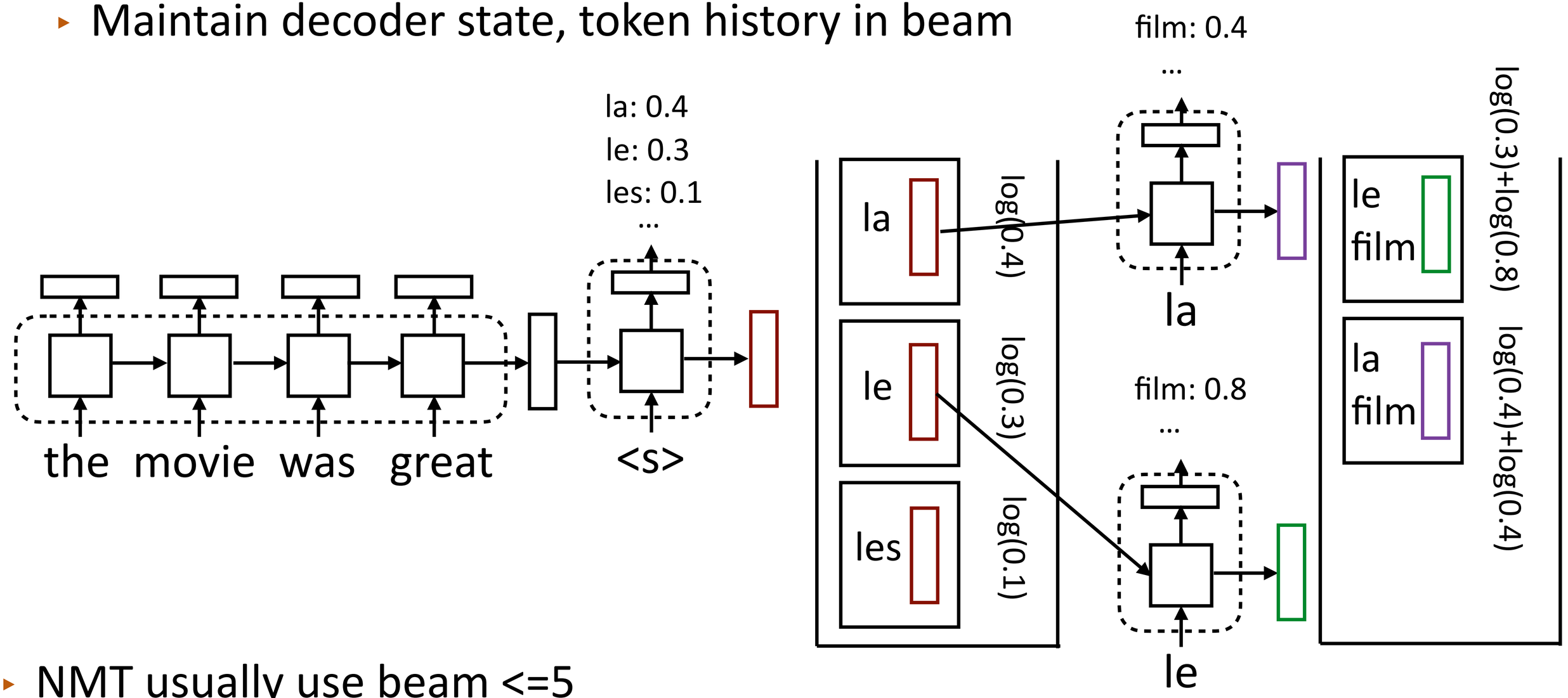
Problems with Greedy Decoding

- Only returns one solution, and it may not be optimal
- Can address this with **beam search**, which usually works better…but even beam search may not find the correct answer! (max probability sequence)

Stahlberg and Byrne (2019)

Beam Search

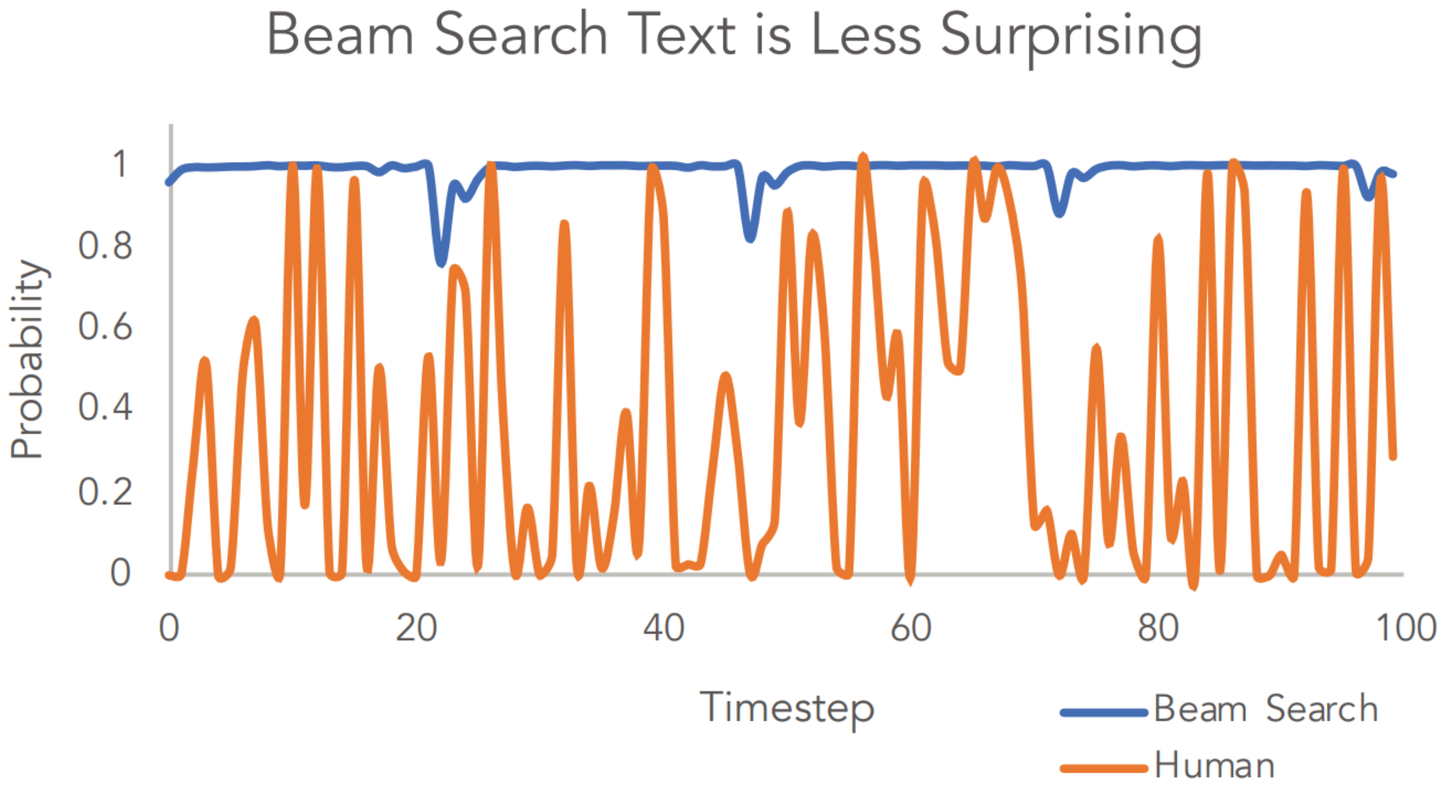
- Maintain decoder state, token history in beam



- NMT usually use beam <=5
- Keep **both** *film* states! Hidden state vectors are different

Beam Search vs. Human

- For less constrained generation tasks (e.g., story generation)



Holtzman et al. (2019)

Sampling

- Greedy solution can be uninteresting / vacuous for various reasons (so called text *degeneration)*.
- Beam search may give many similar sequences, and these actually may be *too close* to the optimal.
- Sampling can help - especially for some text generation tasks.

$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \text{softmax}(W\bar{h})$$

$$y_{\text{sampled}} \sim P(y|\mathbf{x}, y_1, \ldots, y_{i-1})$$

Beam Search vs. Sampling

- Sampling is better but sometimes draws too far from the tail of the distribution (relatively low prob. over thousands of candidate tokens).
- These are samples from an unconditioned language model GPT-2 (not seq2seq model)

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Pure Sampling:
They were cattle called Bolivian Cavalleros; they live in a remote desert uninterrupted by town, and they speak huge, beautiful, paradisiacal Bolivian linguistic thing. They say, 'Lunch, marge.' They don't tell what the lunch is," director Professor Chuperas Omwell told Sky News. "They've only been talking to scientists, like we're being interviewed by TV reporters. We don't even stick around to be interviewed by TV reporters. Maybe that's how they figured out that they're cosplaying as the Bolivian Cavalleros."

Holtzman et al. (2019)

Sampling

- Using methods more sophisticated than naive sampling
  - Top-k: take the top k most likely words (e.g. k=5), sample from those
  - Nucleus (top-p): take the top p% (e.g., 95%) of the distribution, sample from within that
  - Epsilon: simple truncation, allow any word with greater than ε probability

Generation Tasks



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.



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.

Holtzman et al. (2019)

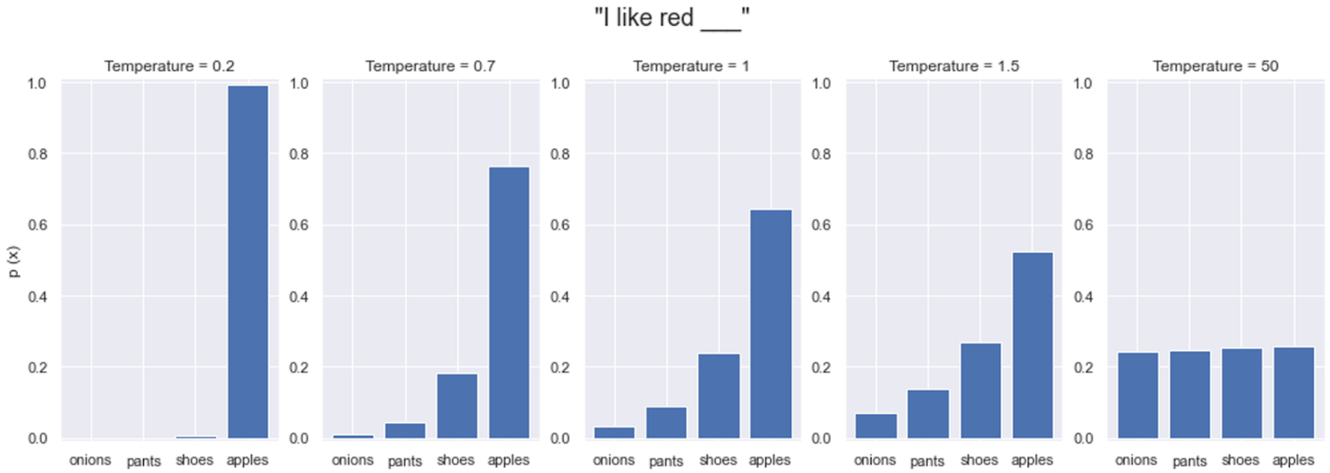
Softmax Temperature

- Formally, NNs typically use a softmax function to convert logits z_i into class probabilities:

temperature to smooth softmax distribution

$$P(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

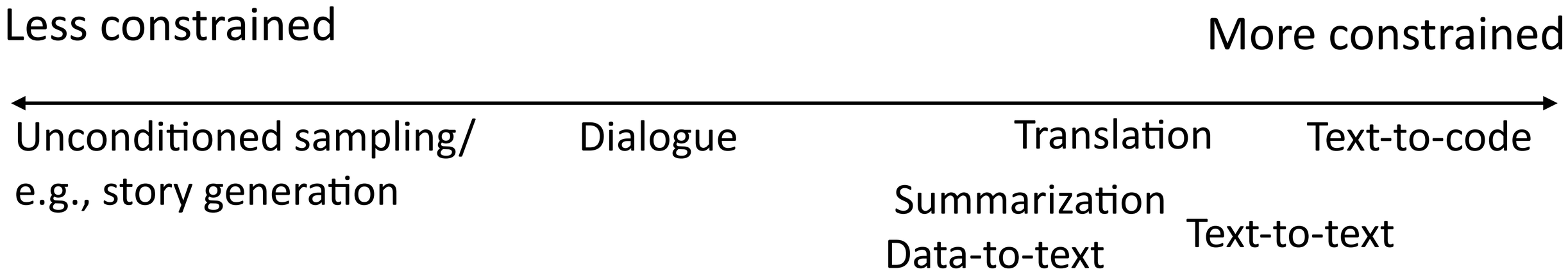
index of output space $\mathcal{Y}$



- T = 1: Standard Softmax

Generation Tasks

- For more constrained problems: greedy/beam decoding are usually best
- For less constrained problems: sampling introduces favorable variation in the output



Decoding Strategies

- LMs place a distribution $P(y_i | \ y_1, \ldots, y_{i-1})$
- seq2seq models place a distribution $P(y_i | \mathbf{x}, y_1, \ldots, y_{i-1})$
- Generation from both models looks similar: (1) Greedy, (2) Beam Search, (3) Sampling

- Q: Is the Viterbi algorithm suitable for LLM decoding?

More Advanced Decoding Strategies

(Recap) Decoding Strategies

- LMs place a distribution $P(y_i | y_1, \ldots, y_{i-1})$

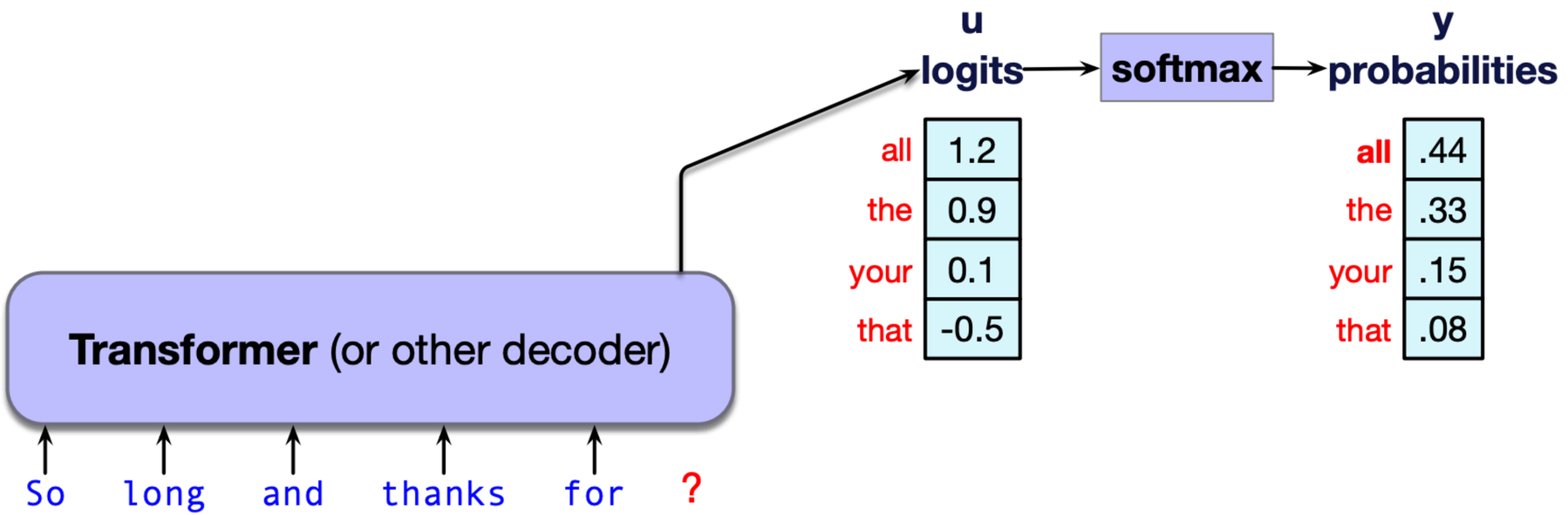


Image Credit: J&M Chapter 7.6

More Decoding Strategies

- Greedy
- Beam search
- Sampling, e.g.:

} predict the next token

  - Top-k: take the top k most likely words (e.g. k=5), sample from those
  - Nucleus: take the top p% (e.g., 95%) of the distribution, sample from within that
  - Epsilon: simple truncation, allow any word with greater than ε probability
- Reranking, e.g.:
  - Minimum Bayes Risk, Speculative decoding

} generate multiple sequences, then select

Minimal Bayes Risk (MBR) Decoding

- **Intuition**: the best output not only have high probability (same as maximum likelihood), but also is consistent or similar to the other candidate outputs.

Minimal Bayes Risk (MBR) Decoding

- **Step 1**: sample multiple sequences

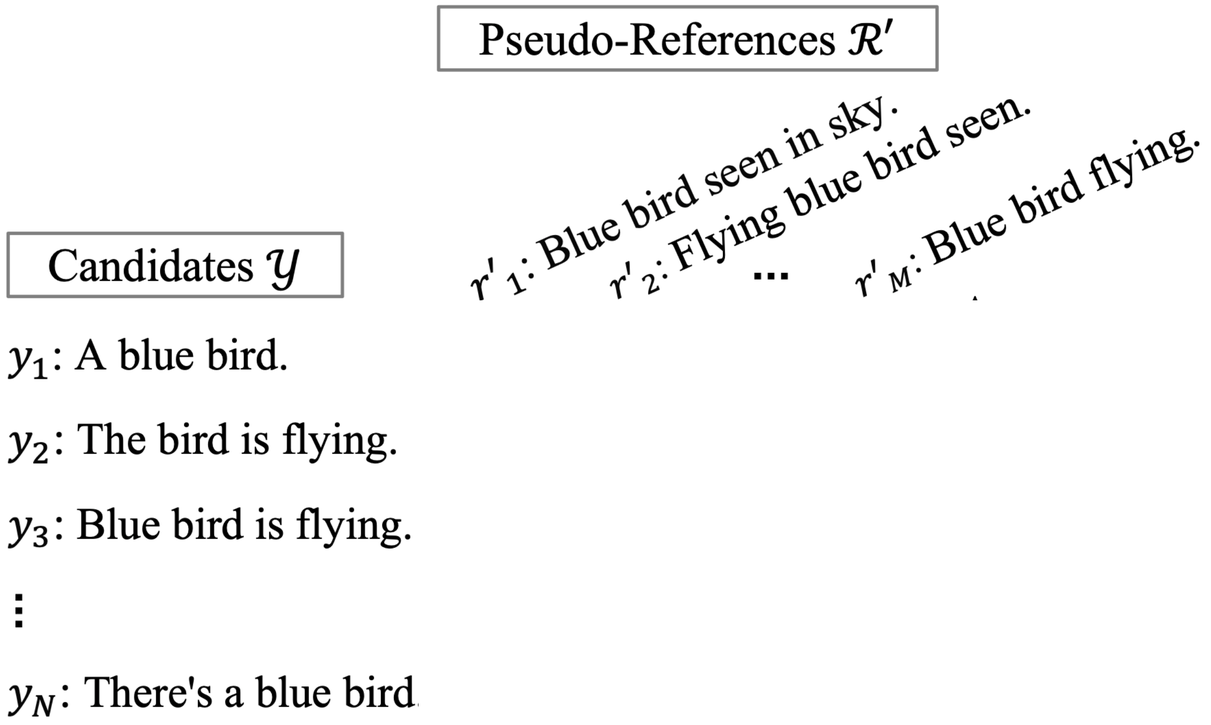


Image Credit: Ohashi et al. (2024)

Minimal Bayes Risk (MBR) Decoding

- **Step 1**: sample multiple sequences
- **Step 2**: Compare each seq. to the others by a utility function

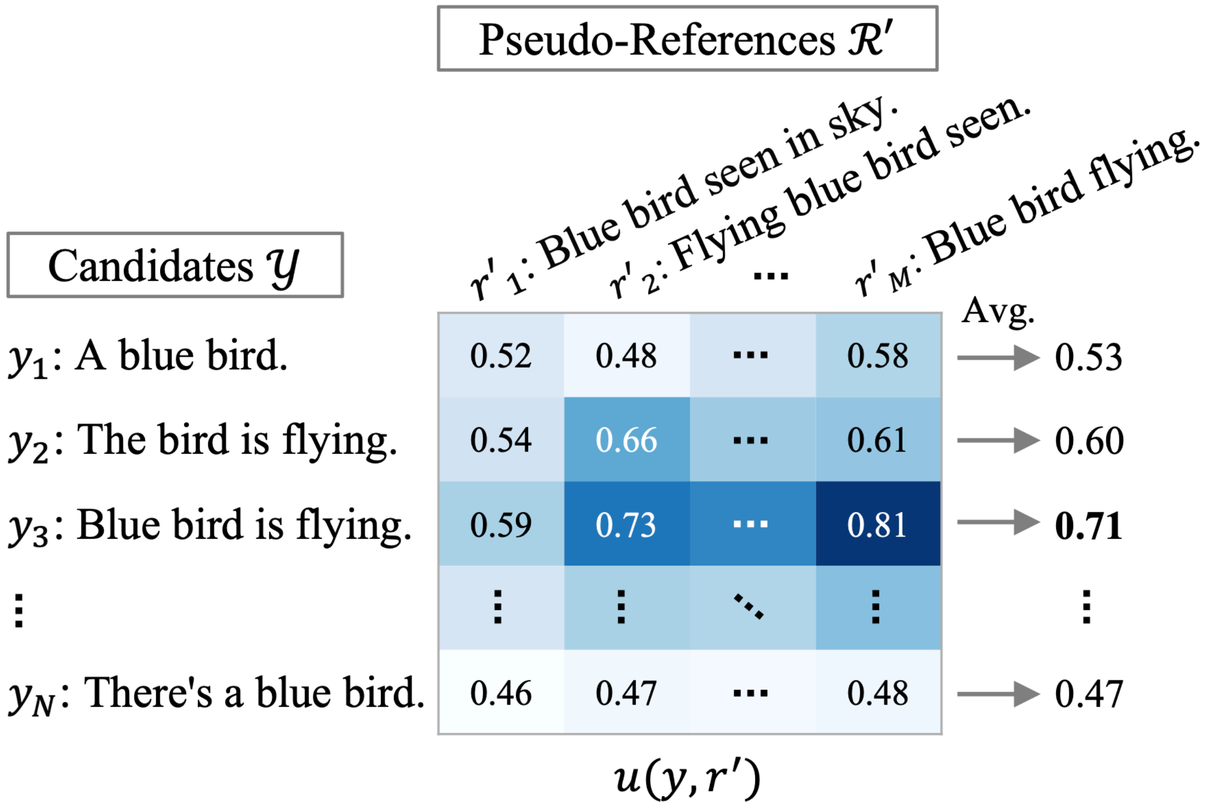


Image Credit: Ohashi et al. (2024)

Minimal Bayes Risk (MBR) Decoding

- **Step 1**: sample multiple sequences
- **Step 2**: Compare each seq. to the others by a utility function
- **Step 3**: Select the seq. that maximizes the expected utility

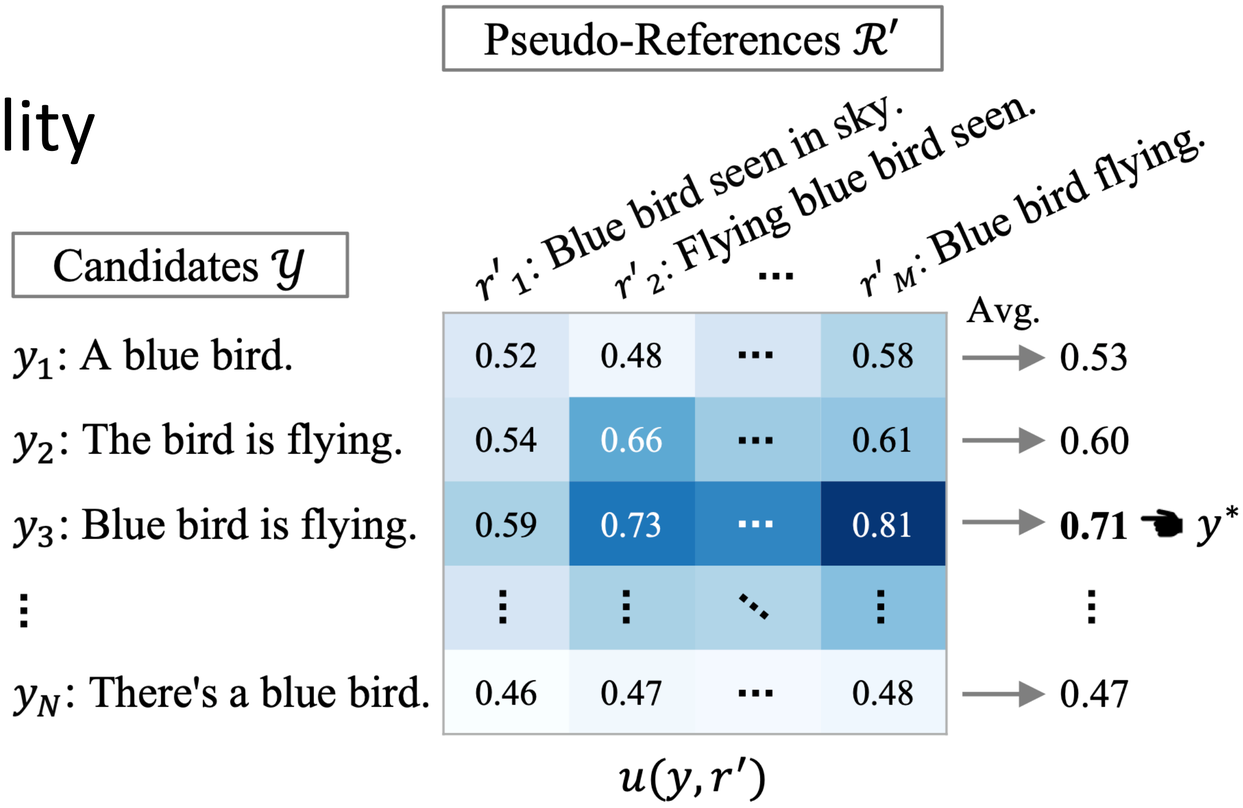
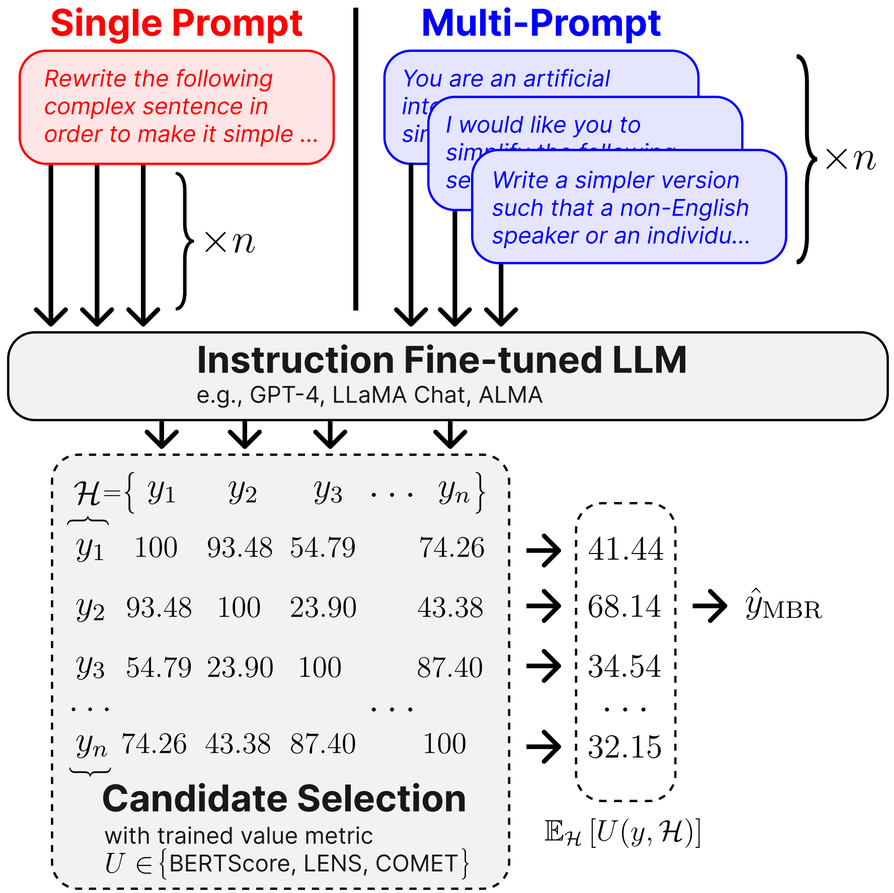
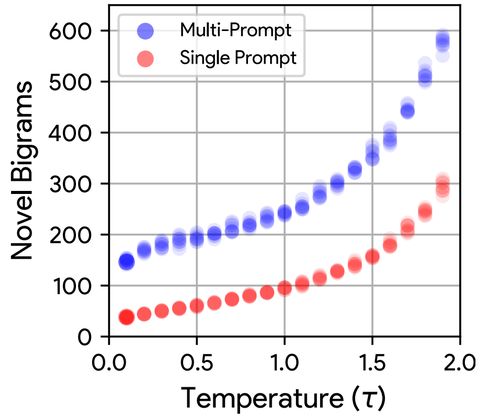


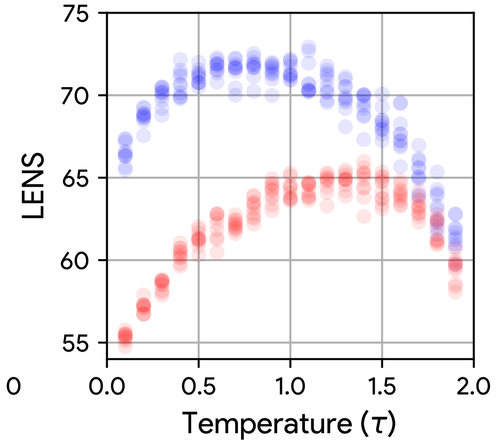
Image Credit: Ohashi et al. (2024)

Minimal Bayes Risk (MBR) Decoding

- MBR aims to find an output that maximizes the expected utility, i.e., metrics like COMET (for translation) or LENS (for simplification).







Heineman et al. (2024)

Minimal Bayes Risk (MBR) Decoding

- MBR aims to find an output that maximizes the expected utility, i.e., metrics like COMET (for translation) or LENS (for simplification).
- Diverse prompt + MBR decoding: consistent improvement across generation tasks, works for both open-source and black-box LLMs

Code Generation ($\|\mathcal{H}\| = 20$) – HUMANEVAL (pass@1)		
StarCoder 2 15B	44.51	49.39 (+4.88)
CodeLlama 7B	37.80	40.85 (+3.05)
CodeLlama 13B	43.29	48.17 (+4.88)
CodeLlama 34B	45.73	52.44 (+6.71)
CodeLlama 70B	61.59	68.90 (+7.31)
GPT-3.5	68.29	73.78 (+5.49)
GPT-4	81.71	82.93 (+1.22)

Text Simplification ($\|\mathcal{H}\| = 100$) – SIMPEVAL (LENS)		
Ctrl T5 3B	72.6	–
Ctrl T5 11B	74.4	–
Llama 2 7B Chat	75.71	80.38 (+4.67)
Llama 2 13B Chat	78.19	80.27 (+2.08)
Llama 2 70B Chat	82.21	83.28 (+1.07)
GPT-3.5	76.87	81.25 (+4.38)
GPT-4	76.47	81.56 (+5.09)

Translation ($\|\mathcal{H}\| = 100$) – WMT '22 EN-CS (COMET)		
WMT '22 Winners	91.9	–
MS Translate API	90.6	–
ALMA 7B R	89.17	89.94 (+0.77)
ALMA 13B R	89.41	90.45 (+1.04)
GPT-3.5	91.27	91.35 (+0.08)
GPT-4	92.24	92.47 (+0.23)

Heineman et al. (2024)

Minimal Bayes Risk (MBR) Decoding

- Often deliver several points of performance improvement, over the standard beam search or sampling methods.

	R-1	R-2	R-L	BLEU
Summarization	XSUM			
Sample-Once	37.9	16.1	30.6	11.4
Random	37.6	16.1	30.1	11.5
Majority Voting	37.8	16.2	30.6	11.4
MBRD-BLEURT	39.8	17.9	32.4	12.8
MBRD-BERTScore	**41.2**	**19.0**	**33.4**	**13.5**
Translation	DE → EN (German to English)			
Sample-Once	68.1	45.9	63.9	39.0
Random	68.5	46.1	64.0	39.5
Majority Voting	70.2	48.7	66.1	40.9
MBRD-BLEURT	71.9	50.7	68.2	43.7
MBRD-BERTScore	**73.3**	**52.6**	**69.6**	**45.8**

(Suzgun+, ’23)

	SARI	BScore	LENS	sBL↓	Human
Simplification	SIMPEVAL$_{2022}$				
T5-11B					
MLE$_{b=10}$	**46.4**	**93.8**	62.9	49.3	88.80
MBR-LENS$_{\|S\|=100}$	46.1	**93.8**	**74.4**	**44.6**	**90.13**
Close-source LLMs					
GPT-3.5 (0-shot)	41.4	93.4	60.7	31.8	90.77
GPT-3.5 (5-shot)	42.4	94.1	69.0	33.2	92.70
GPT-4 (0-shot)	43.7	**94.3**	**73.5**	**29.1**	**93.63**

(Maddela+, ’23)

Minimal Bayes Risk (MBR) Decoding

- **Intuition**: the best output not only have high probability (same as maximum likelihood), but also is consistent or similar to the other candidate outputs.
- **More formally**:

First sample a set of hypotheses $\mathcal{H}$ from the model, then select the output that maximizes the expected utility U (or minimize the expected risk) with respect to a set of references $\mathcal{R}$:

$$\hat{y}_{mbr} = \arg\max_{y \in \mathcal{H}} (\mathbb{E}_{\mathcal{H} \sim \text{model}}[U(y, \mathcal{R})])$$

Minimal Bayes Risk (MBR) Decoding

- **Intuition**: the best output not only have high probability (same as maximum likelihood), but also is consistent or similar to the other candidate outputs.
- **More formally**:

can be the same or different set, often about 10~1000 sequences generated by sampling or beam search

First sample a set of hypotheses $\mathcal{H}$ from the model, then select the output that maximizes the expected utility U (or minimize the expected risk) with respect to a set of references $\mathcal{R}$:

$$\hat{y}_{mbr} = \arg\max_{y \in \mathcal{H}} (\mathbb{E}_{\mathcal{H} \sim \text{model}}[U(y, \mathcal{R})])$$

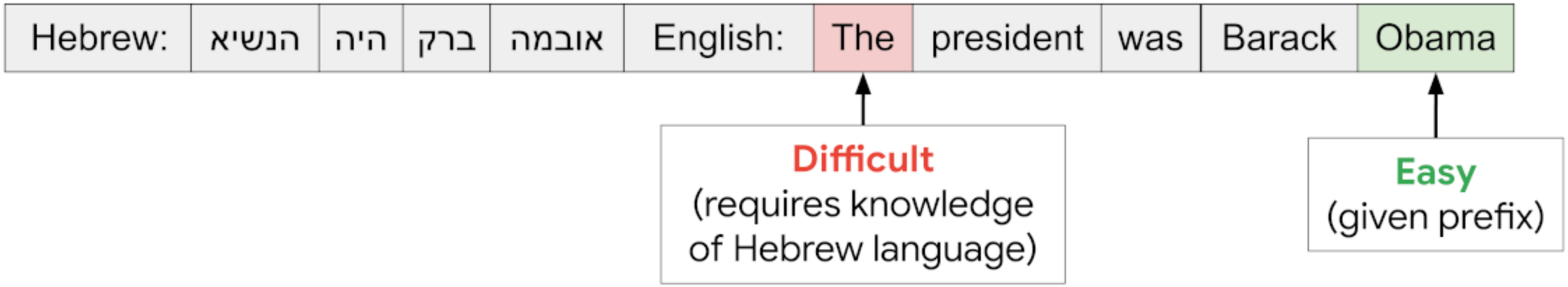
Minimal Bayes Risk (MBR) Decoding

- Main Limitations:
  - $O(|\mathcal{H}|^2)$ computation time for utility function
  - number of sample $|\mathcal{H}|$<< number of all possible hypotheses $|\mathcal{Y}|$

- Research directions:
  - choice of sampling algorithm to collect hypotheses (Ohashi+ ‘24)
  - model-based estimation (Jinnai+ ‘24)
  - promoting diversity (Heineman+ ’24)
  - Reducing computation time for utility function (Tomani+ ’24)

Speculative Decoding

- **Intuition**: Some tokens in the sequence are easier (can use a small LLM) to generate than others (ideally, use a larger LLM).



- How to combine small and large LLMs to do this more efficiently?

Leviathan et al. (2023)

Speculative Decoding

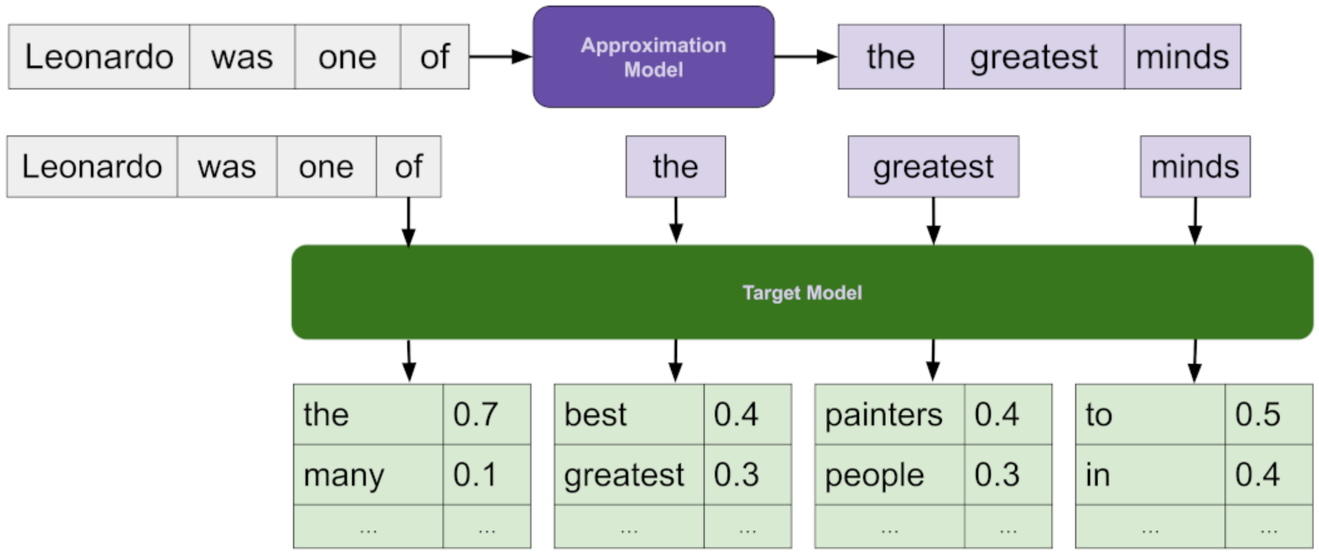
- **Step 1**: generate k tokens by a small (approximation) model, autoregressively



Leviathan et al. (2023)

Speculative Decoding

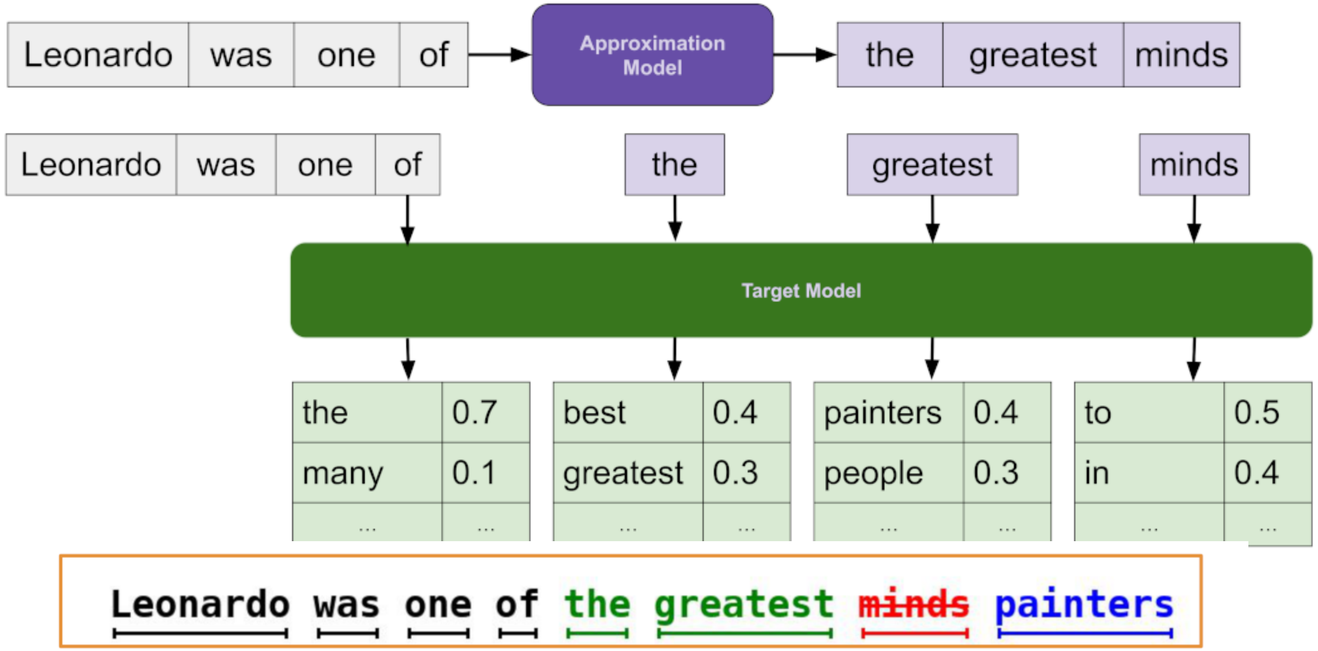
- **Step 1**: generate k tokens by a small (approximation) model, autoregressively
- **Step 2**: use a large (target) model to generate next-token distributions for all k+1 prefixes — i.e., evaluate small model's generations **in parallel!**



Leviathan et al. (2023)

Speculative Decoding

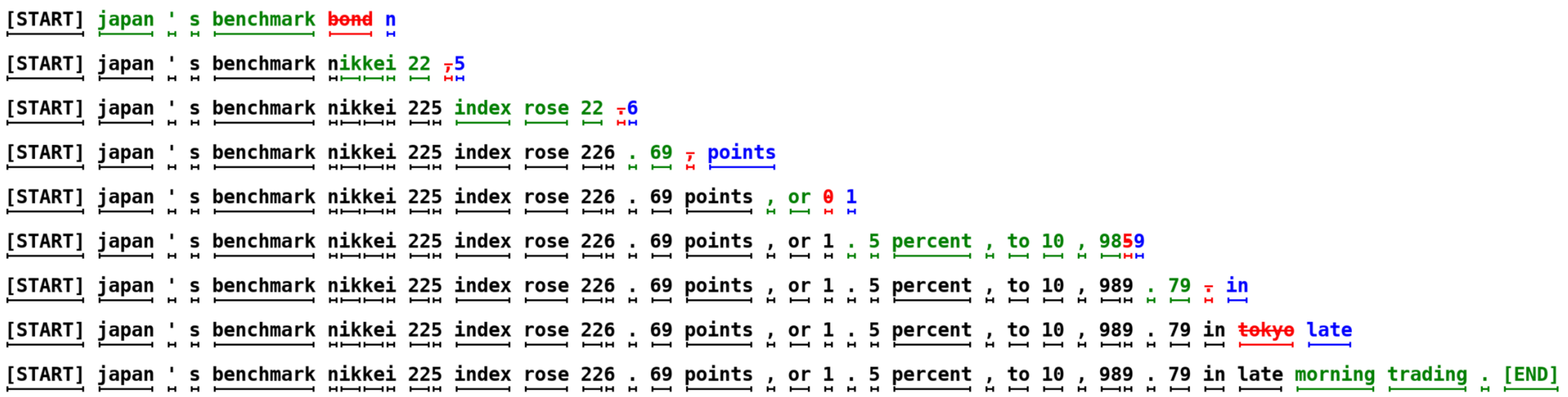
- **Step 1**: generate k tokens by a small (approximation) model, autoregressively
- **Step 2**: use a large (target) model to generate next-token distributions for all k+1 prefixes — i.e., evaluate small model's generations **in parallel!**
- **Step 3**: Decide which tokens to accept or reject (with a probability) based on the large model, and sample one more token from the large model



Leviathan et al. (2023)

Speculative Decoding

- **Step 3**: Decide which tokens from a small LLM to accept or reject (with a probability) based on the large model, and sample one more token from the large model



Each line is one iteration

Leviathan et al. (2023)

Collaborative Decoding

- **Intuition**: The decision of deferral from a small model to a large model is learnt as a latent variable $Z_t \in \{0, 1, \ldots, M\}$ by optimizing the marginal likelihood:

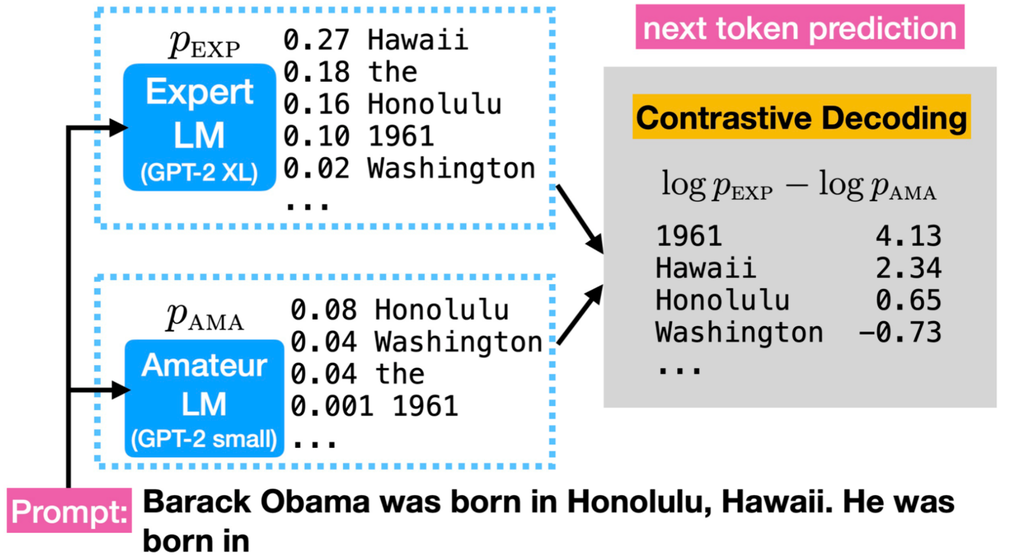
$$P(X) = \prod_{t=1}^{T} \Big(\sum_{Z_t=0}^{M} P_\theta(Z_t | X_{<t}) P_{Z_t}(X_t | X_{<t}) \Big)$$



Shen et al. (2024)

Contrastive Decoding

- **Intuition**: The failures of larger LLMs (“expert”) are even more prevalent in smaller LMs (“amateur”), thus the difference between the two can be a useful signal.

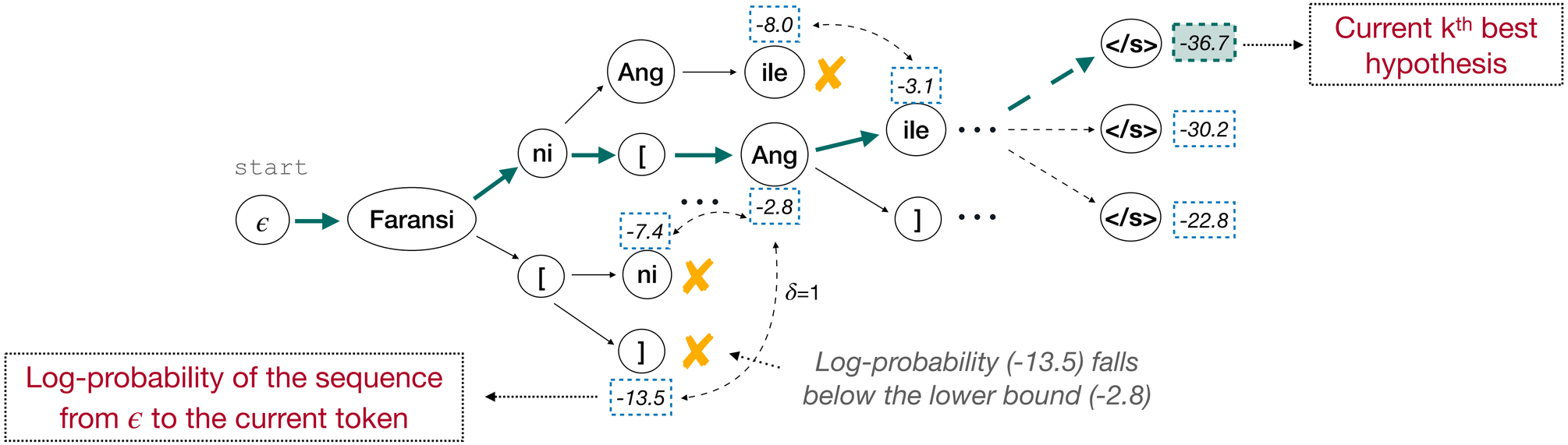


Li et al. (2023)

Constrained Decoding

- CODEC: A branch-and-bound search algorithm with a heuristic lower bound

Input: x = "Only France and Britain backed Fischler 's proposal ." x^{mark} = "Only France and [Britain] backed Fischler 's proposal ." y^{tmpl} = "Faransi ni Angiletɛri dɔrɔn de ye Fischler ka laɲini dɛmɛ ."



✘ *Prune branches based on a heuristic lower-bound*

Le et al. (2024)