DECODING TECHNIQUES

SUBTITLE

Ramaseshan Ramachandran

PRE-TRAINING

- Process of learning linguistic patterns & world knowledge from massive text datasets
- Trains models to predict text sequences, building foundational language understanding

LLMS

- Output model from Pre-training - severe billion parameters
- Develop emergent capabilities through scale:
- Contextual word representations
- Cross-domain knowledge retention
- Pattern recognition across languages

ADVANTAGES OF LLMS

- State-of-the-art performance on NLP benchmarks:
  - Text generation (most transformative)
  - Semantic understanding
  - Few-shot learning
  - Particularly effective for generative tasks
    - Summarization
    - Machine Translation
    - Question Answering
    - Chatbot Interactions

DECODING TECHNIQUES

- Selecting next token from probability distribution
- Key components:
  - Context window (prior generated text)
  - Vocabulary probability scores
  - Decoding strategy algorithm
- Repeatedly choosing the next word conditioned on the previous choices - autoregressive/causal generation

RANDOM SAMPLING

- Generates sensible, high-probability words but also includes odd, low-probability words, resulting in weird sentences
- Will it effectively generate sentences with adequate and fluent structure?
- We look for quality and diversity in the generated text
- We want techniques that emphasize the most probable words

GREEDY APPROACH

- Model is computed using conditional probabilities. We want to generate a sentence $w_1, w_2, w_3, \cdots, w_n$ using
  - $\hat{w}_i = \arg\max_{w \in V} P(w_i | w_{<i})$
- The approach makes a locally optimal choices - Highest probability token is chosen
- Generate words that are likely in the context and less likely to generate equivalent words that are unlikely
- Generates sentences that are
  - More accurate
  - More coherent
  - More factual
  - Boring and more repetitive
- What happens if we choose the next word from the the middle of the distribution
  - May be more creative and diverse
  - Less factual and incoherent and no adequate

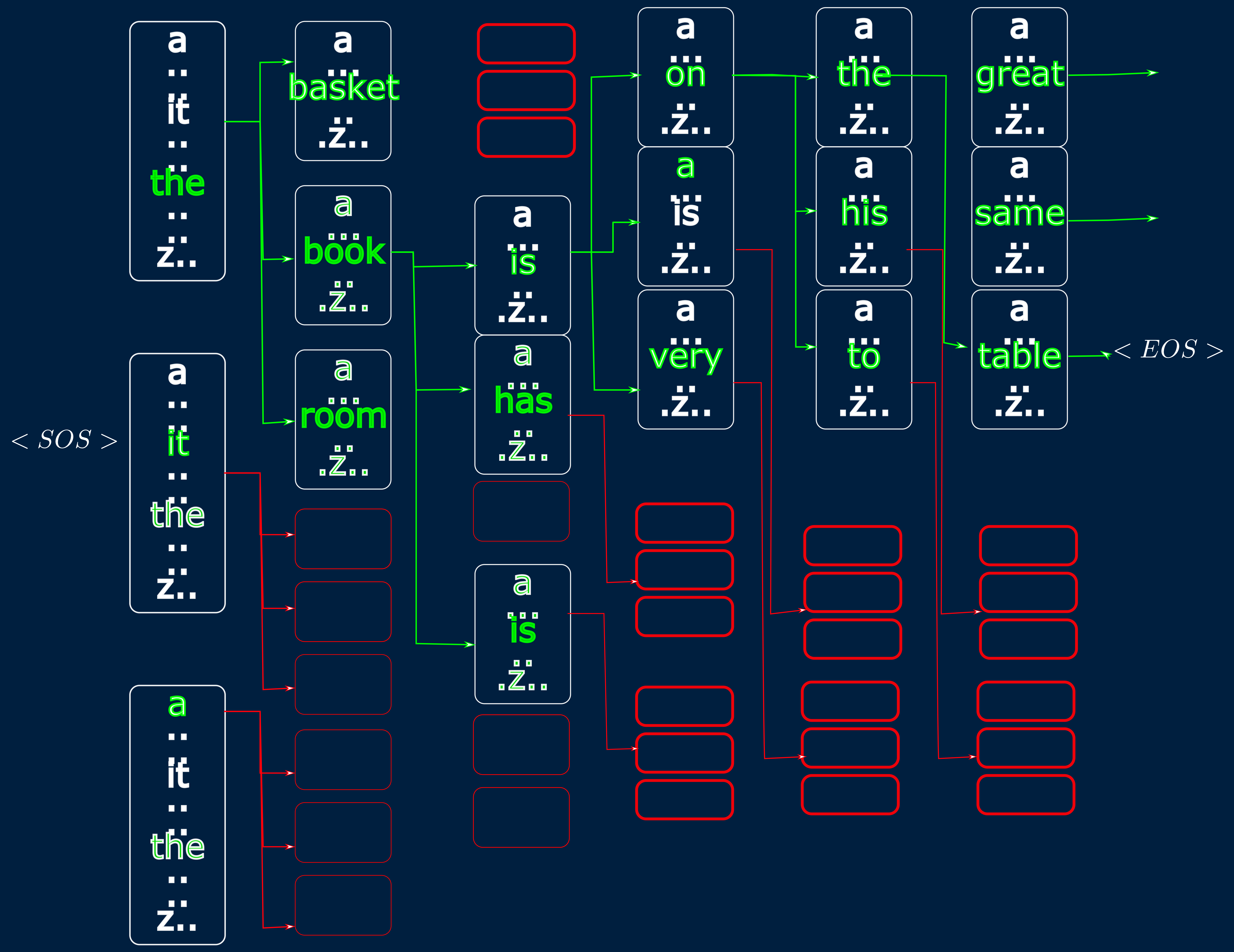
TEMPERATURE

- LLM Temperature Impact: Significantly affects text coherence
  - Controls token selection randomness
- **Quality of decoding:** Impacts the quality of the output generated by the LLM - Greedy→ Balanced choices
- $$P(w_t) = \frac{\exp(z_i/T)}{\sum \exp(z_j/T)}$$
  - $z = = \mathbf{h} \cdot W_{\text{vocab}} + \mathbf{b}$, represents raw $logit$ score before applying any activation function

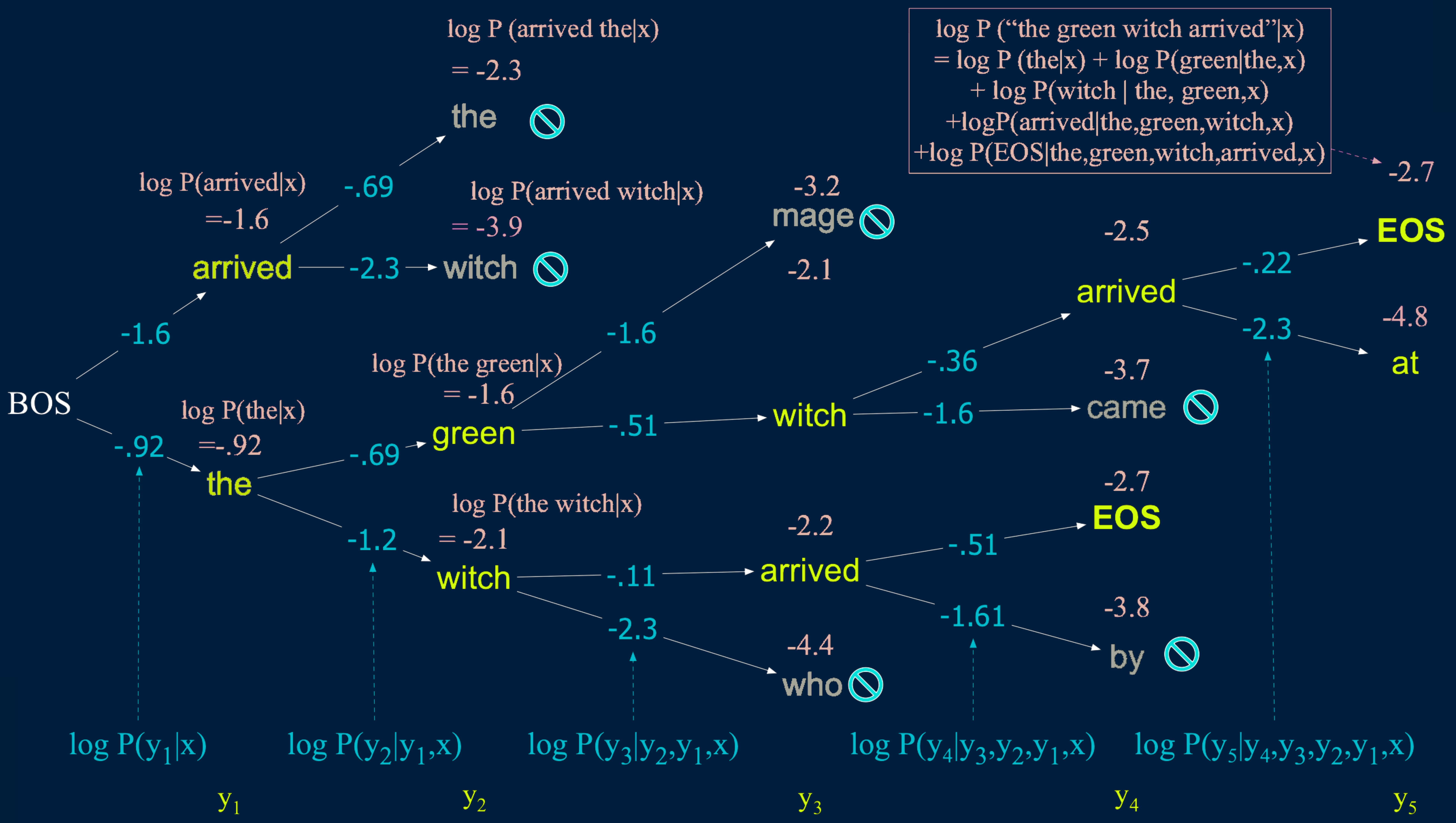
BEAM SEARCH I

- Selects a few candidate hypothesis from $|V|$. It reduces memory requirement by using only a $M < |V|$ candidates using a score.
  - Maintain M candidates/hypothesis at each time step
  - $C_t = (x_1^1, \ldots x_t^1) \ldots (x_1^M \ldots x_t^M)$
  - Compute C_{t+1} by expanding C_t and keeping the best M candidates
  - $$\tilde{C} = \bigcup_{i=1}^{M} C_{t-1}^{i}$$
- Typical Beam width = 5-10

BEAM SEARCH II



BEAM EXAMPLE ($M = 2$)



TOP-K SAMPLING

- Truncate the distribution to the *top-k* most likely words.
- Renormalized to produce a probability distribution
- A word is randomly sampled from within the *top-k* words according to their renormalized probabilities
- When $k = 1$, *top-k* sampling is identical to greedy decoding
- Setting *k* to a larger number
  - More diverse but still high-quality text
  - Impact on fluency - Low risk
- Selecting to the middle-probability words
  - More creative and more diverse
  - Impact on fluency - High risk

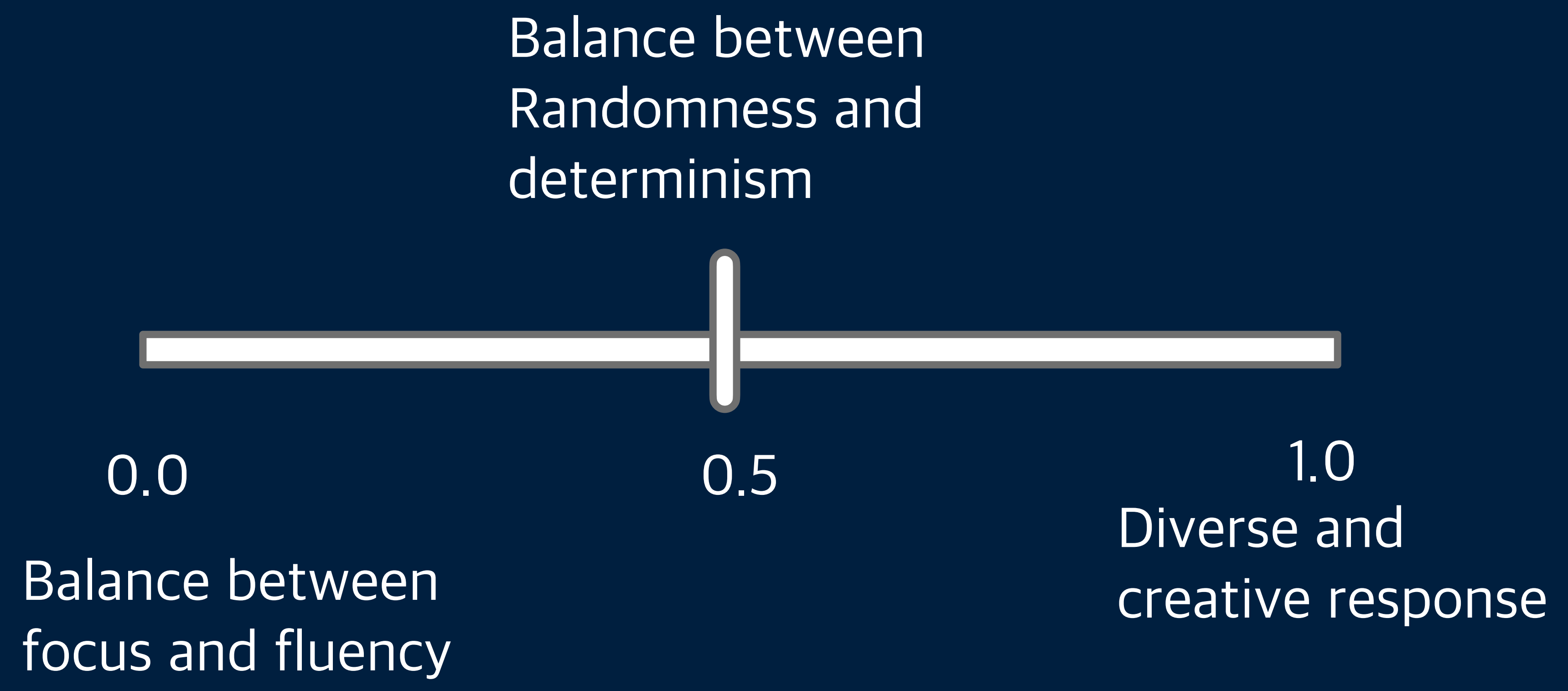
Use case	Top-k	Rationale
Technical writing	5-10	High-confidence tokens
Creative writing	20-50	Controlled generation for creativity
Conversational AI	10-30	Safety and engagement

N. N. Minh, A. Baker, C. Neo, A. G. Roush, A. Kirsch, and R. Shwartz-Ziv. Turning up the heat:Min-p sampling for creative and coherent LLM outputs. In The Thirteenth International Conference on Learning Representations, 2025.

TOP-P SAMPLING

- Keeps the *top-p* percent of the probability mass
  - Tokens selected $= \arg\min_{k} \sum_{i=1}^{k} P(w_i) \leq p$
  - Selects the smallest set of tokens whose cumulative probability exceeds threshold p
  - Removes very unlikely words
- Measures probability rather than the number of words
- Balances creativity and coherence by dynamically adjusting the candidate pool
- Rescales probability of the selected tokens so that $\sum_{i} p_i = 1$

ChatGpt uses top-p sampling



A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration, 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.

Sampling, t=0.9

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

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.

WebText

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

Figure 3: Example generations continuing an initial sentence. Maximization and top-k truncation methods lead to copious repetition (highlighted in blue), while sampling with and without temperature tends to lead to incoherence (highlighted in red). Nucleus Sampling largely avoids both issues.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration, 2020.

MIN-P SAMPLING

- Adjusts cutoff threshold based on model confidence in real-time
  - $\text{Threshold}_t = \max(P(x_t | x_{1:t-1})) \times \text{min\_p}$ Where:
    - $\max(P(x_t | x_{1:t-1}))$ = highest token probability at step
    - min_p = user-defined ratio (e.g., 0.05-0.2)

- At each generation step:
  - Compute token probabilities $P(x_t | x_{1:t-1})$ over vocabulary V
  - Identify top probability p_{max}
  - Calculate adaptive threshold:
    - $\text{Threshold} = p_{\text{max}} \times \text{min\_p3}$. Retain tokens where $$ p_i \geq \text{Threshold} $$
  - Sample from filtered distribution

TEMPERATURE BASED SAMPLING

- T=0.2: [▓▓▓▓▓░░░░░] Top token - *dominates*
- T=1.0: [▓▓▓▓░░░░░░] Balanced - *exploration*
- T=2.0: [▓▓▓▓▓▓▓░░░] High-risk - *diversity*
- Use Temperature When:
  - Creativity response is required (e.g., brainstorming, storytelling)
  - Balancing exploration/exploitation (e.g., chatbots)
- Avoid Temperature When:
  - Maximum determinism is required (e.g., legal contracts)
  - Using pure greedy/beam search

Decoding approach	Temperature used
Greedy	No
Beam search	No
Temperature Sampling	Yes
Top-K	Yes
Top-P	Yes
Min-P	Yes