Cascaded Text Generation with Markov Transformers

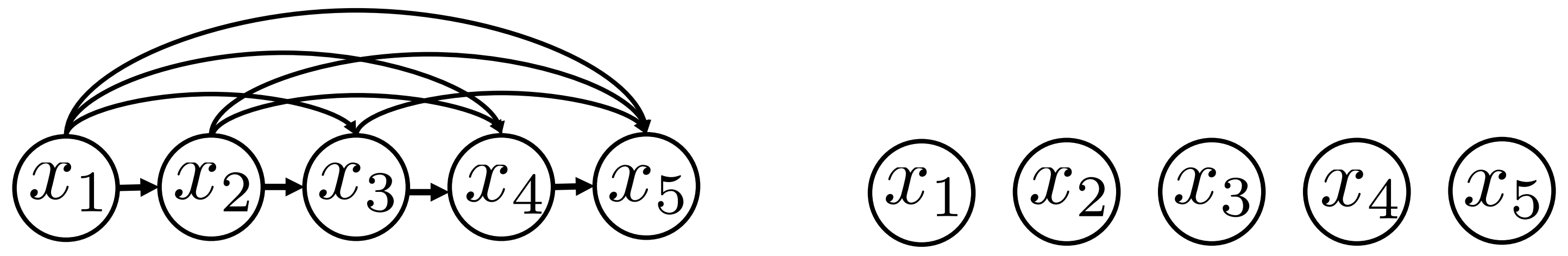
Yuntian DENG, Alexander M. RUSH
Harvard University, Cornell University

github.com/harvardnlp/cascaded-generation

We present a **fast and fluent** text generation method using **bounded-order** Markov models. To decode from high-order models efficiently, we propose a cascaded decoding approach that **prunes the search space** using lower-order models, and we introduce a Markov transformer that can **parameterize the entire cascade**.

1/ Motivation

Fully Autoregressive (AR) | **Nonautoregressive (NAR)**



- Decoding: beam search
- Fluent but serial

- Decoding: argmax
- Parallel but disfluent

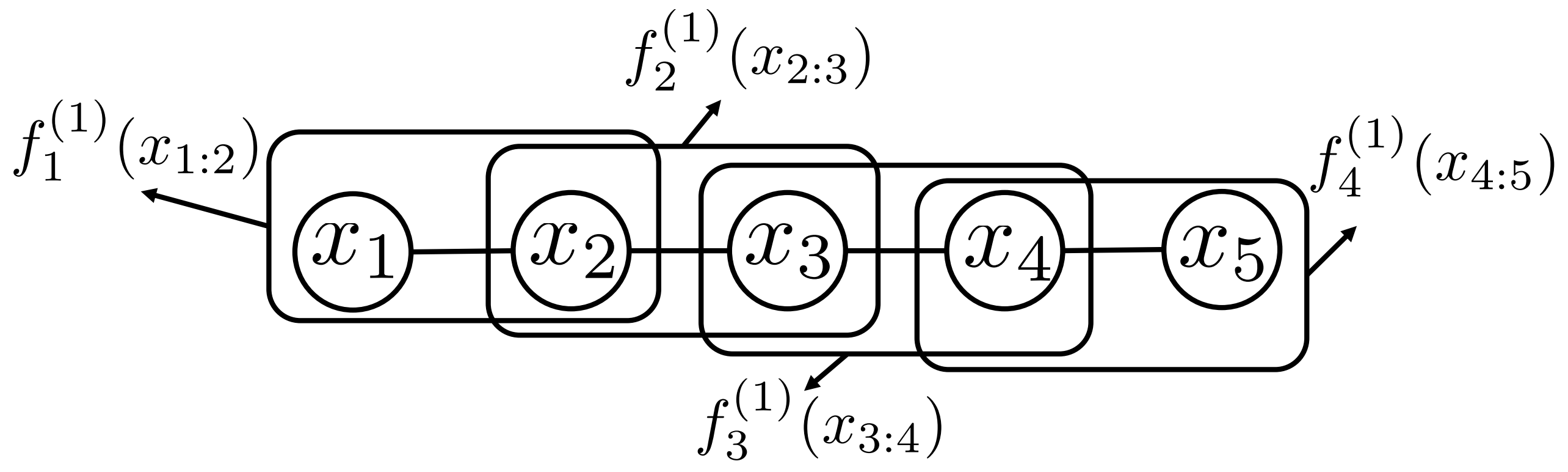
- We propose a method that's both fast and fluent

2/ Model: Bounded-Order MRF

- An m-th order Markov Random Field (MRF)

$$P^{(m)}(x_{1:L};\theta) \propto \exp \sum_{l=1}^{L-m} f_l^{(m)}(x_{l:l+m};\theta)$$

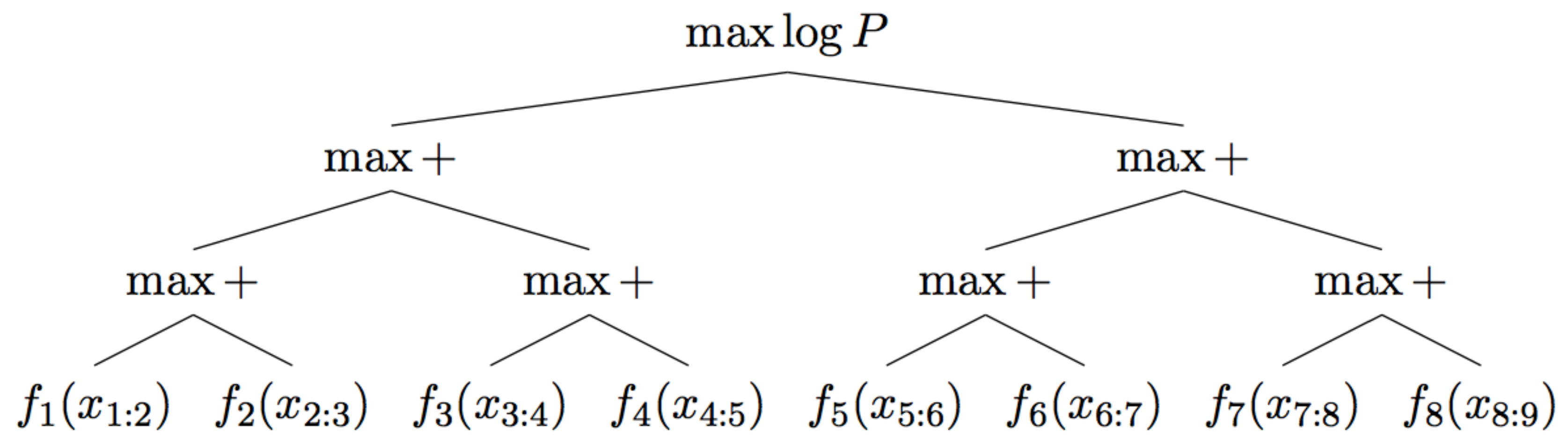
- Each $f_l^{(m)}$ models dependencies among adjacent $(m+1)$ words
- Example: $m=1$, $L=5$



- Special Cases:
  - ○ $m=0$: nonautoregressive
  - ○ $m = \mathrm{L}-1$: fully autoregressive
  - ○ $0 < m < \mathrm{L}-1$: **bounded-order models** (this work)

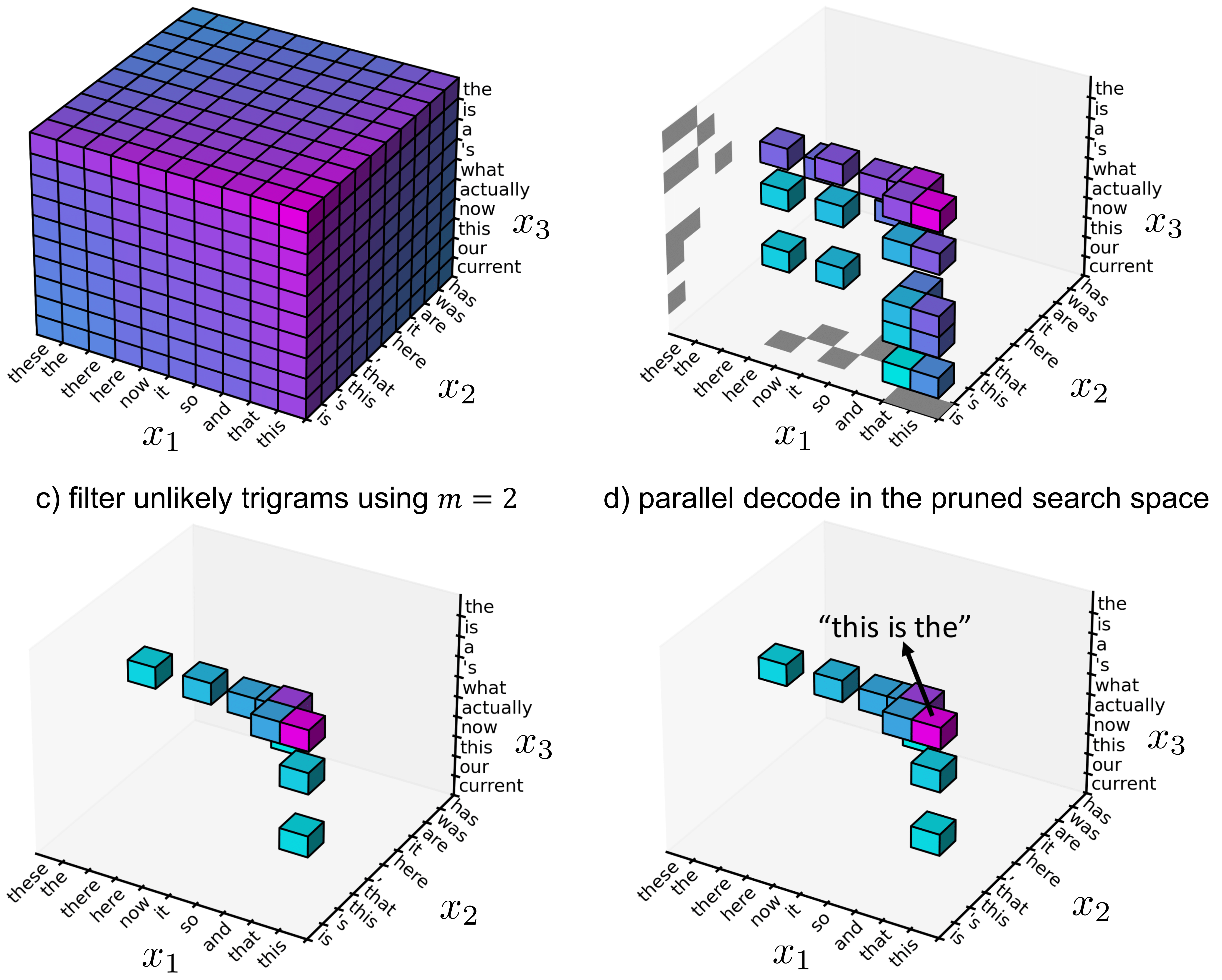
3/ Parallel Decoding

- Bounded-order models can be decoded in parallel
- In practice, takes <1% total time



4/ Cascaded Decoding

- Parallel decoding for an M-th order model takes $O(V^{M+1} \log L)$ time
- Impractical even when M=1
- Cascaded decoding filters unlikely n-grams using lower-order models
- Example: M= 3


a) filter unlikely unigrams using $m=0$

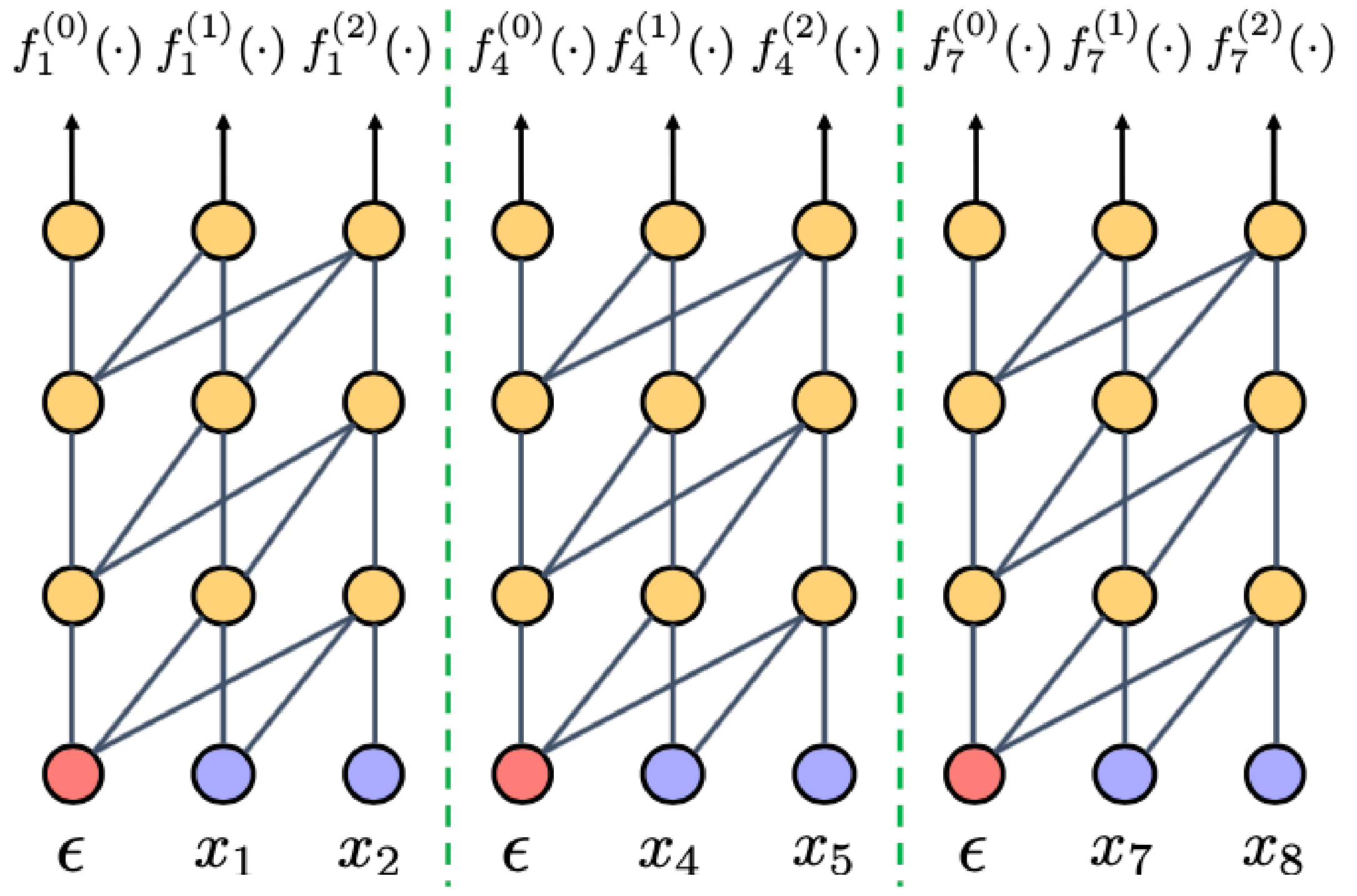
b) filter unlikely bigrams using $m=1$

c) filter unlikely trigrams using $m=2$

d) parallel decode in the pruned search space

5/ Parameterization: Markov Transformer

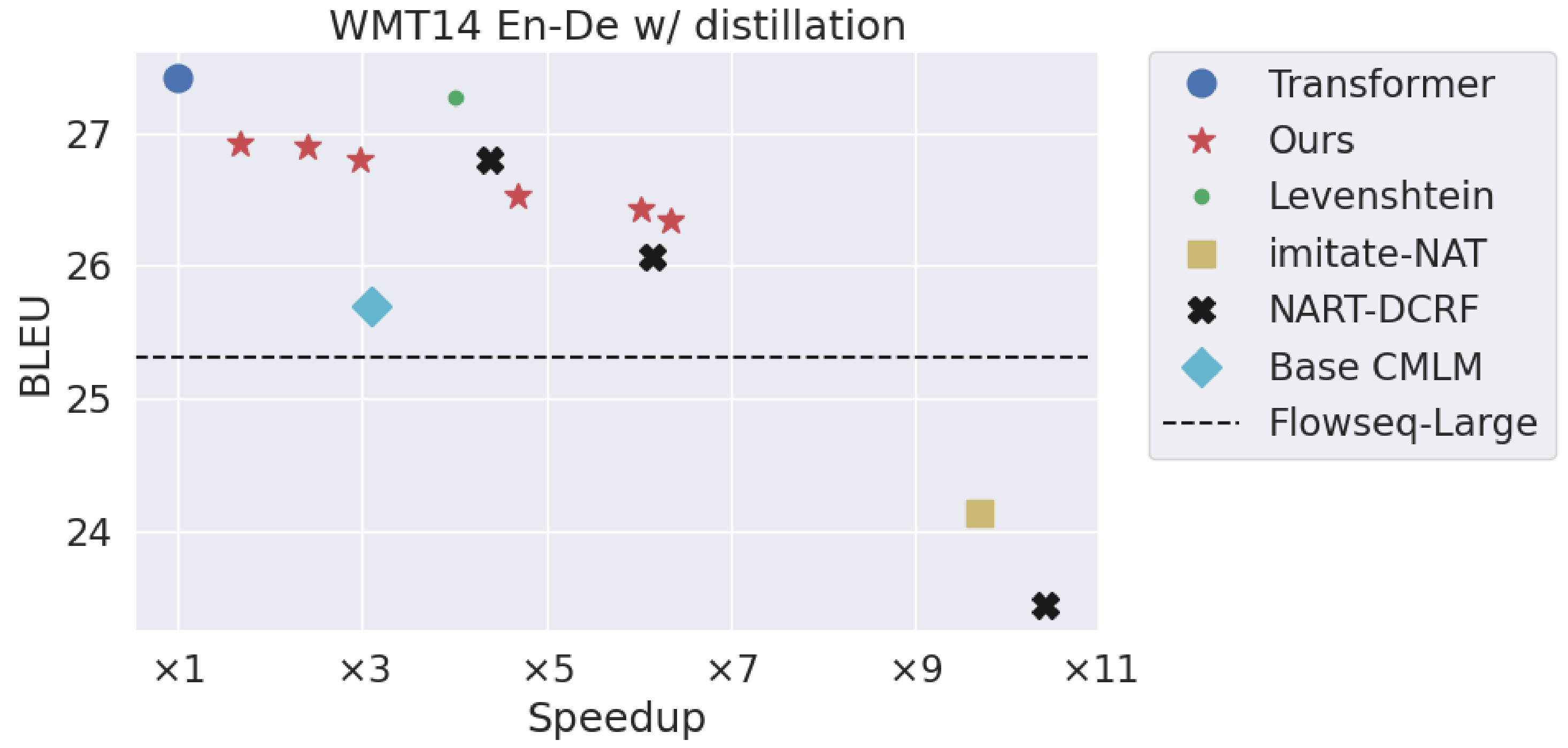
- Parameterize the entire cascade with a single transformer
- Training: insert M-spaced barriers with random offset



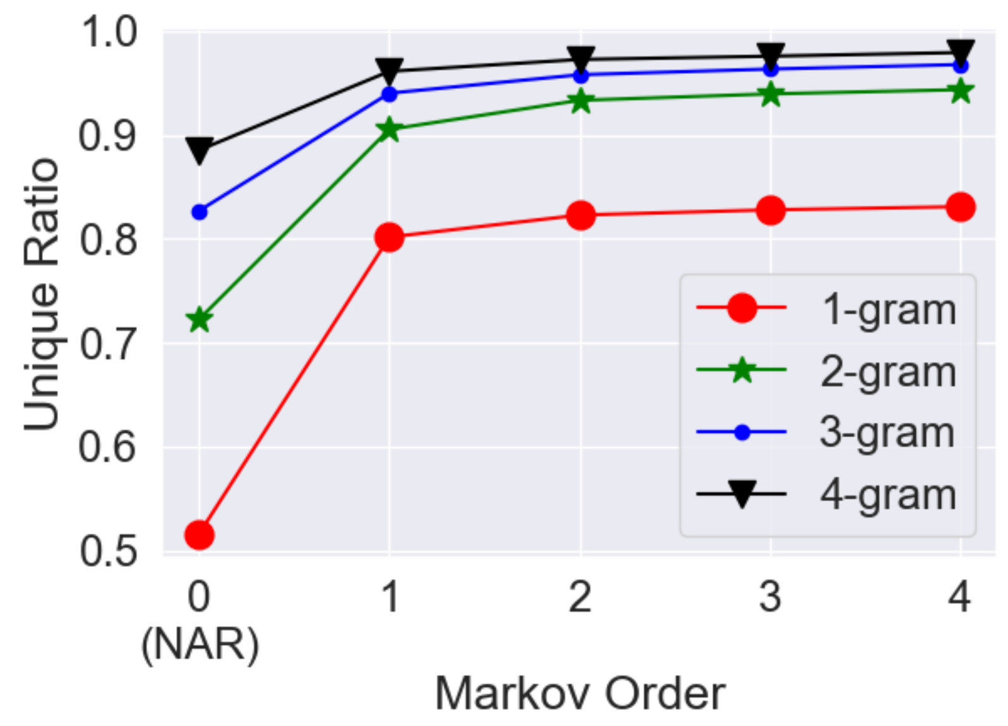
- Test: can be applied as any Markov model with $m < M$

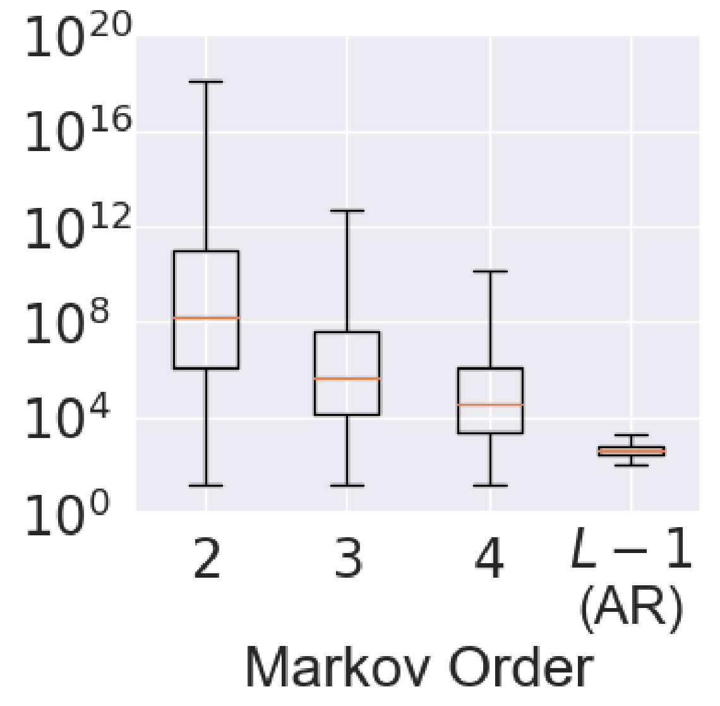
6/ Results & Analysis

- Speed/accuracy tradeoff



- Fewer repetitions
- More hypotheses scored


Unique Ratio
(NAR)
Markov Order
1-gram
2-gram
3-gram
4-gram


10^{20}
10^{16}
10^{12}
10^{8}
10^{4}
10^{0}
$L-1$
(AR)
Markov Order