FiLM: Fill-in Language Models for Any-Order Generation

Tianxiao Shen Hao Peng Ruoqi Shen Yao Fu Zaid Harchaoui Yejin Choi

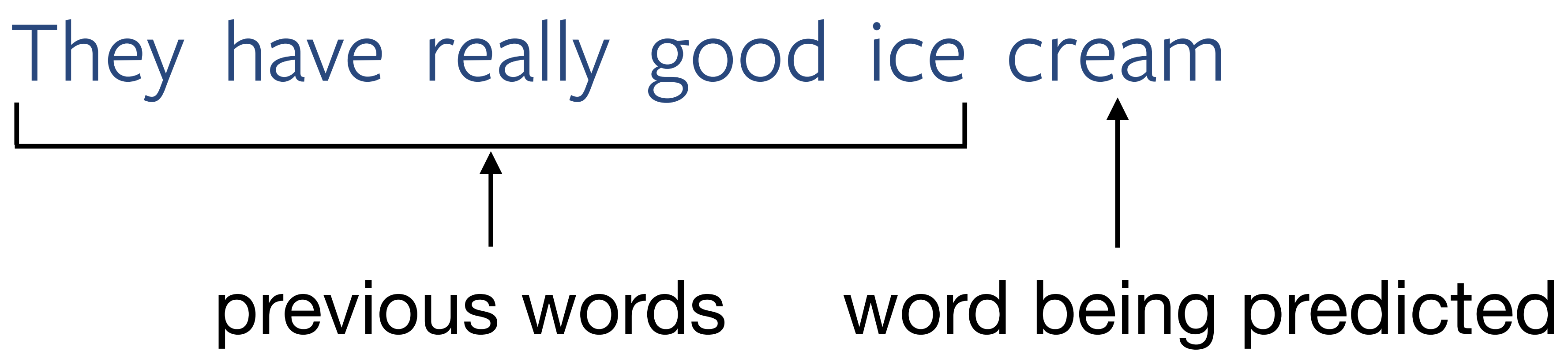
UWNLP

Causal Language Model (CLM)

✓ Generate from left to right

✗ Start with partially specified text

- text editing
- template filling
- code completion
- …


They have really good ice cream

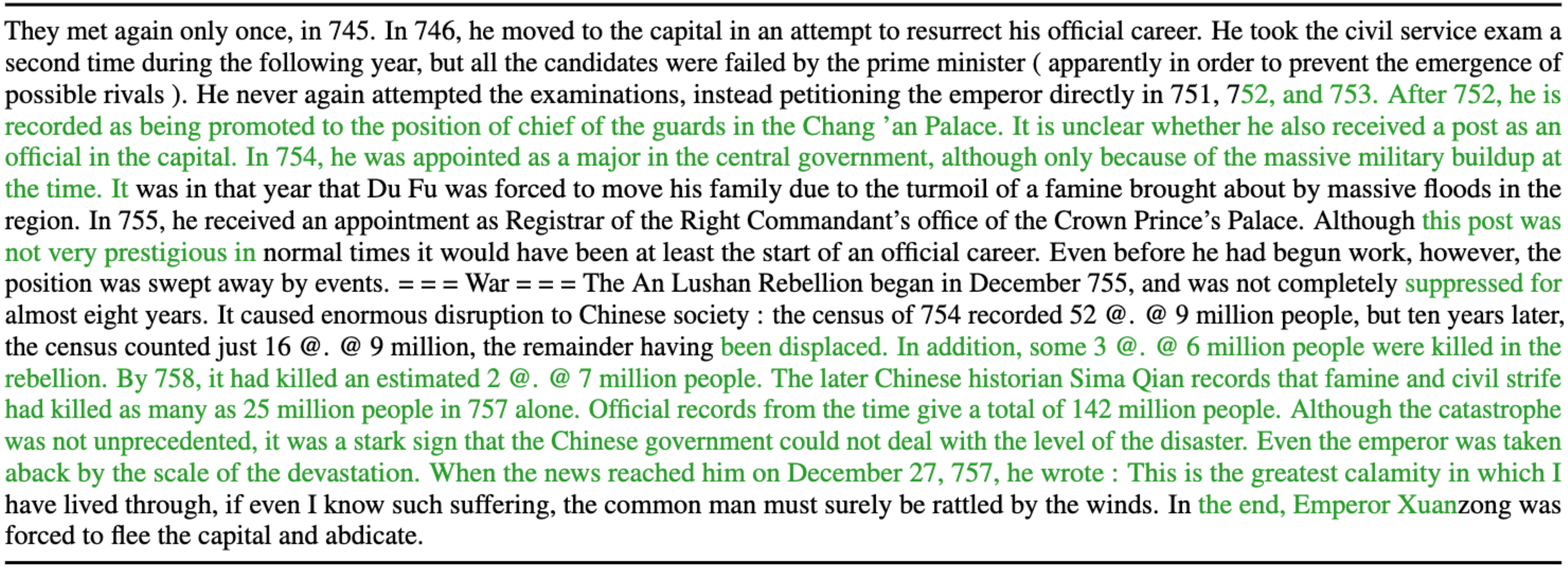
previous words

word being predicted

Fill-in Language Model (FiLM)

- ✓ Generate text from scratch in any order
- ✓ Start from partial text and fill in the missing parts
- ✓ Respect preceding and following context

Fill-in Language Model (FiLM)

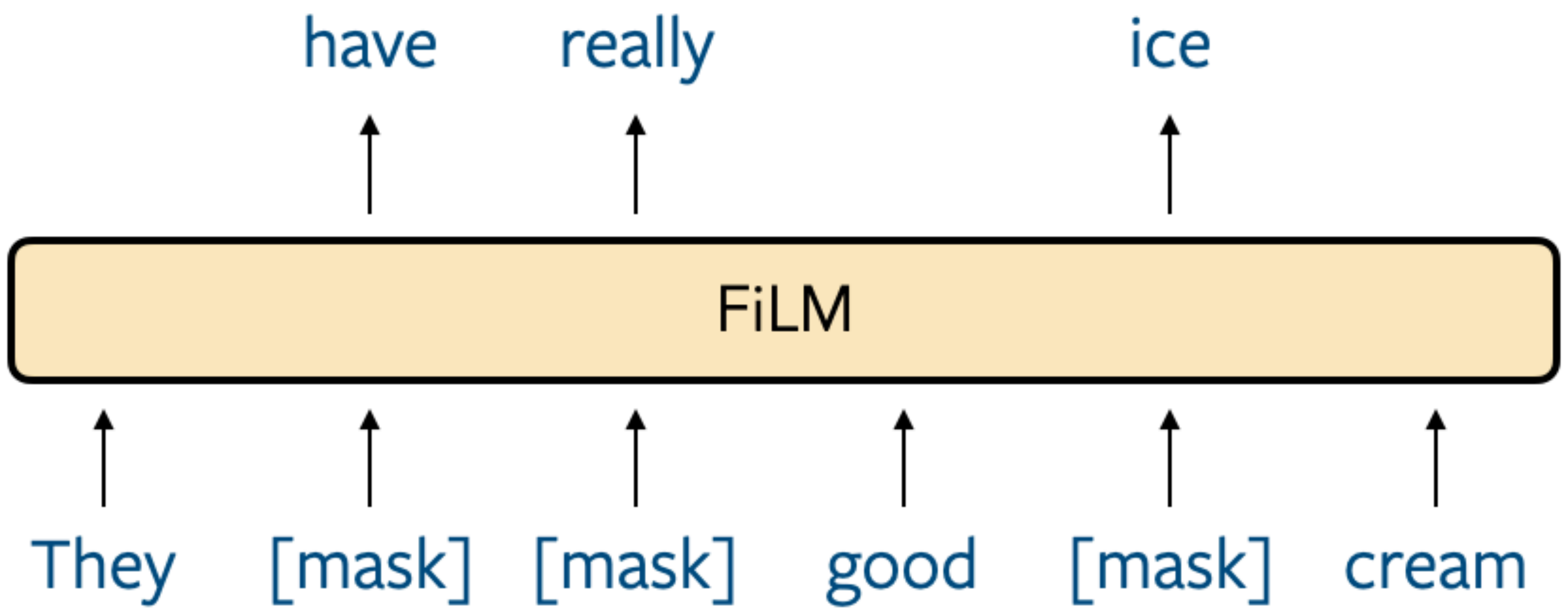


They met again only once, in 745. In 746, he moved to the capital in an attempt to resurrect his official career. He took the civil service exam a second time during the following year, but all the candidates were failed by the prime minister (apparently in order to prevent the emergence of possible rivals). He never again attempted the examinations, instead petitioning the emperor directly in 751, 752, and 753. After 752, he is recorded as being promoted to the position of chief of the guards in the Chang 'an Palace. It is unclear whether he also received a post as an official in the capital. In 754, he was appointed as a major in the central government, although only because of the massive military buildup at the time. It was in that year that Du Fu was forced to move his family due to the turmoil of a famine brought about by massive floods in the region. In 755, he received an appointment as Registrar of the Right Commandant's office of the Crown Prince's Palace. Although this post was not very prestigious in normal times it would have been at least the start of an official career. Even before he had begun work, however, the position was swept away by events. = = = War = = = The An Lushan Rebellion began in December 755, and was not completely suppressed for almost eight years. It caused enormous disruption to Chinese society : the census of 754 recorded 52 @. @ 9 million people, but ten years later, the census counted just 16 @. @ 9 million, the remainder having been displaced. In addition, some 3 @. @ 6 million people were killed in the rebellion. By 758, it had killed an estimated 2 @. @ 7 million people. The later Chinese historian Sima Qian records that famine and civil strife had killed as many as 25 million people in 757 alone. Official records from the time give a total of 142 million people. Although the catastrophe was not unprecedented, it was a stark sign that the Chinese government could not deal with the level of the disaster. Even the emperor was taken aback by the scale of the devastation. When the news reached him on December 27, 757, he wrote : This is the greatest calamity in which I have lived through, if even I know such suffering, the common man must surely be rattled by the winds. In the end, Emperor Xuanzong was forced to flee the capital and abdicate.

Flexible sequence infilling by FiLM-1.6B. Given context is in black, generated text is in color.

FiLM — Training

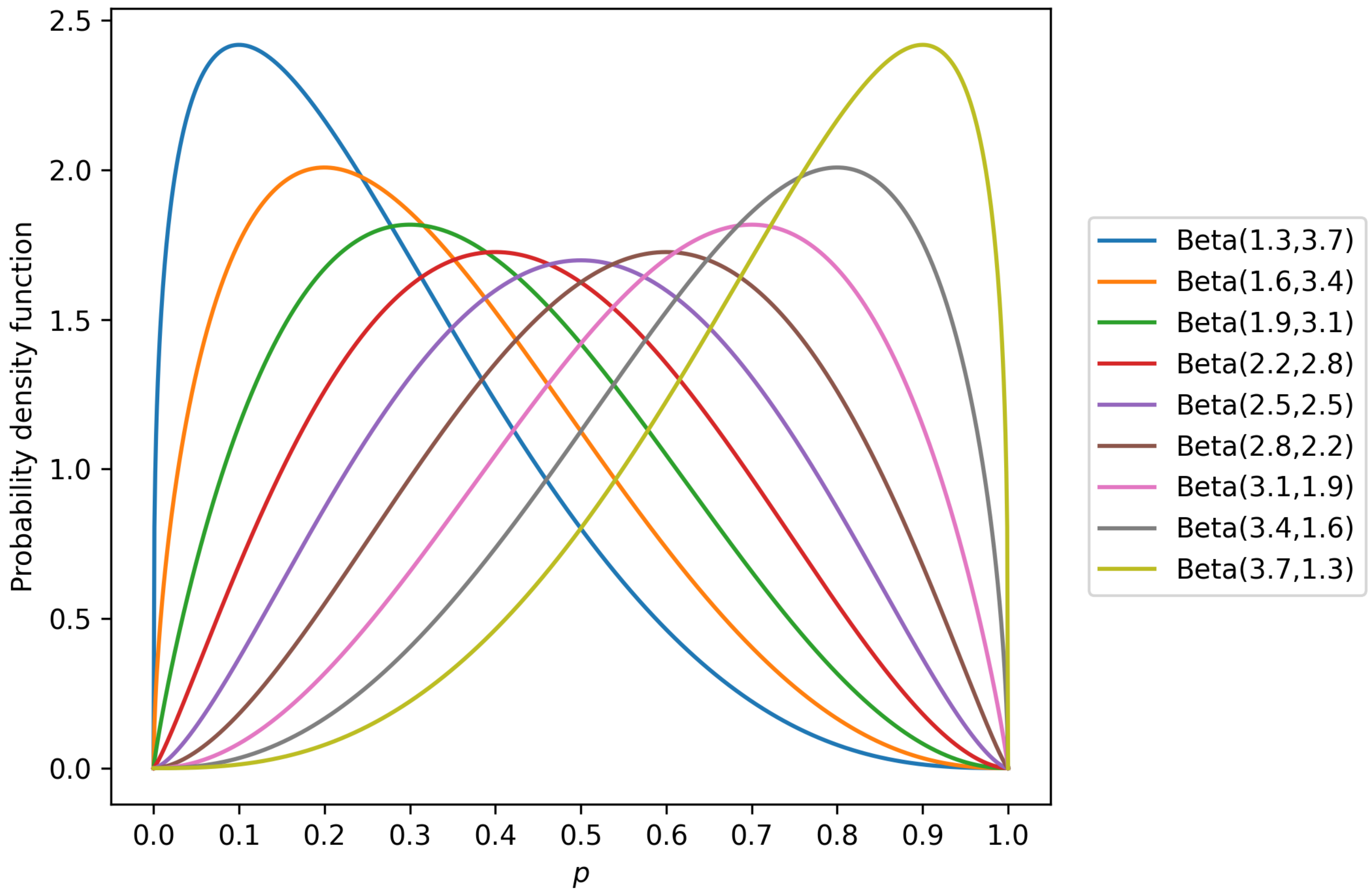
- Choose a mask prob p
- Independently mask each token with prob p
- Predict the original tokens from the masked sequence



FiLM — Training

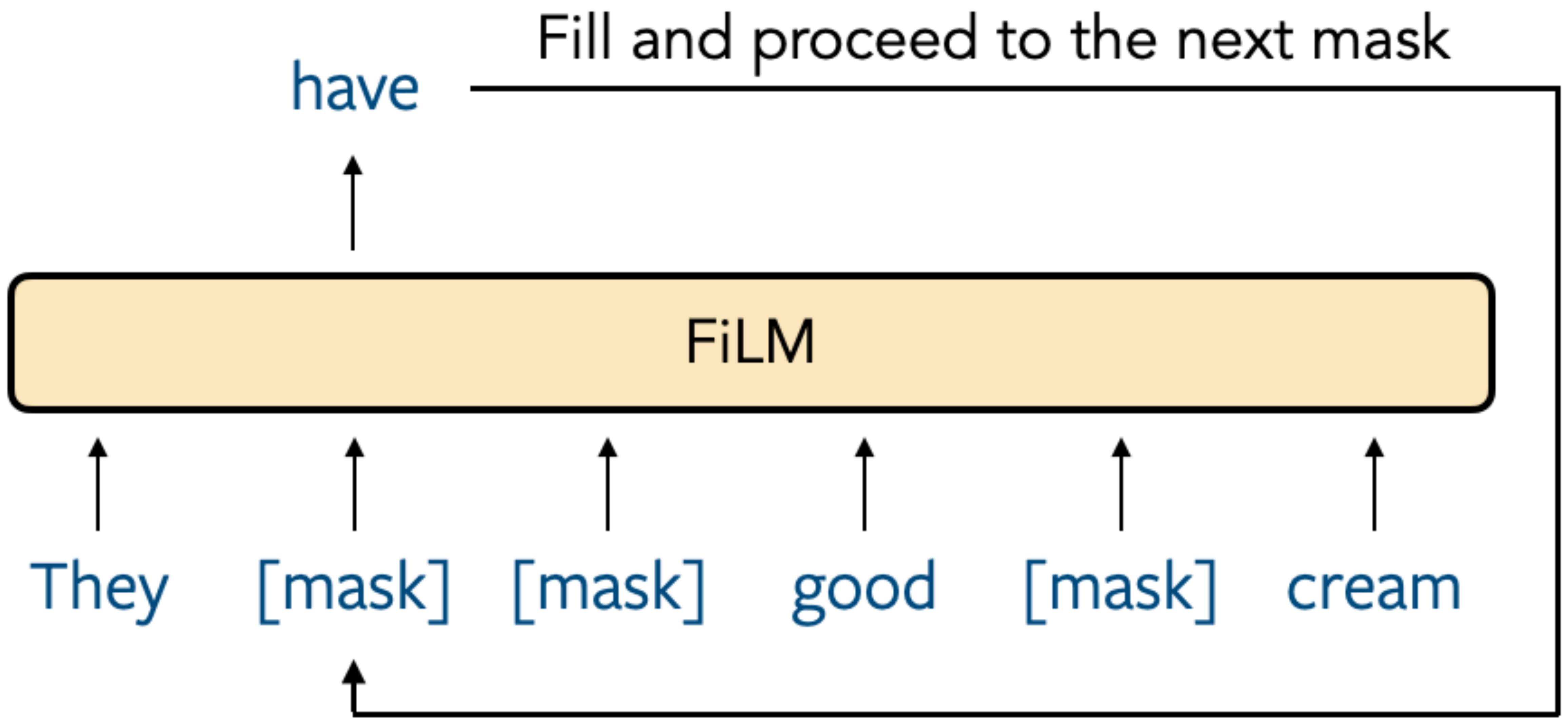
How to choose p?

- fixed → masked language model (MLM), e.g., BERT uses 0.15
- sample from $U[0,1]$
- sample from Beta distribution



FiLM — Decoding

Fill in one mask at each step, conditioned on the given context and previous predictions



FiLM — Decoding

Which mask to fill in first?

- random
- left-to-right
- right-to-left
- min-entropy (“easy-first”)
- max-entropy (“hard-first”)

Once the decoding order is determined, conventional decoding algorithms of CLM (sampling, greedy decoding, beam search, etc.) are also applicable to FiLM

FiLM — Perplexity

Given a sequence $x = (x_1, \ldots, x_n)$,

- Computing $p_\theta(x)$ requires marginalizing over $n!$ decoding orders
- Computing $p_\theta(x; \sigma)$ for a specific order σ is tractable

$$\log p_\theta (x; \sigma) = \log p_{len} (n) + \sum_{t=1}^{n} \log p_\theta \left(x_{\sigma_t} \,|\, x_{\sigma_1}, \ldots, x_{\sigma_{t-1}}, n \right)$$

e.g., $\log p_\theta \left(x_1, x_2, x_3; \sigma = (3,1,2) \right) = \log p_{len} (3)$

$$+\log p_\theta \left(x_3 \,|\, [mask], [mask], [mask] \right)$$

$$+\log p_\theta \left(x_1 \,|\, [mask], [mask], x_3 \right)$$

$$+\log p_\theta \left(x_2 \,|\, x_1, [mask], x_3 \right)$$

FiLM — Perplexity

Given a sequence $x = (x_1, \ldots, x_n)$,

- Computing $p_\theta(x)$ requires marginalizing over $n!$ decoding orders
- Computing $p_\theta(x; \sigma)$ for a specific order σ is tractable

$$\log p_\theta(x; \sigma) = \log p_{len}(n) + \sum_{t=1}^{n} \log p_\theta\left(x_{\sigma_t} \mid x_{\sigma_1}, \ldots, x_{\sigma_{t-1}}, n\right)$$

$$Perplexity = \exp\left(-\frac{1}{n+1} \log p_\theta(x; \sigma)\right)$$

Dividing by $n + 1$ to ensure comparability with CLM, which appends an extra [eos] to $(x_1, \ldots, x_n)$

Experiments

- Analysis of FiLM
  evaluate various training and decoding strategies to find the optimal configuration

- Language modeling
  compare perplexity with CLM

- Text infilling
- Story completion
  compare with SOTA infilling methods under automatic and human evaluations

Analysis of FiLM

Datasets:

- WikiText-103
  document-level, chunked into 512 tokens, 103M words in total
- One Billion Word
  sentence-level, average length 28.5 tokens, 1B words in total

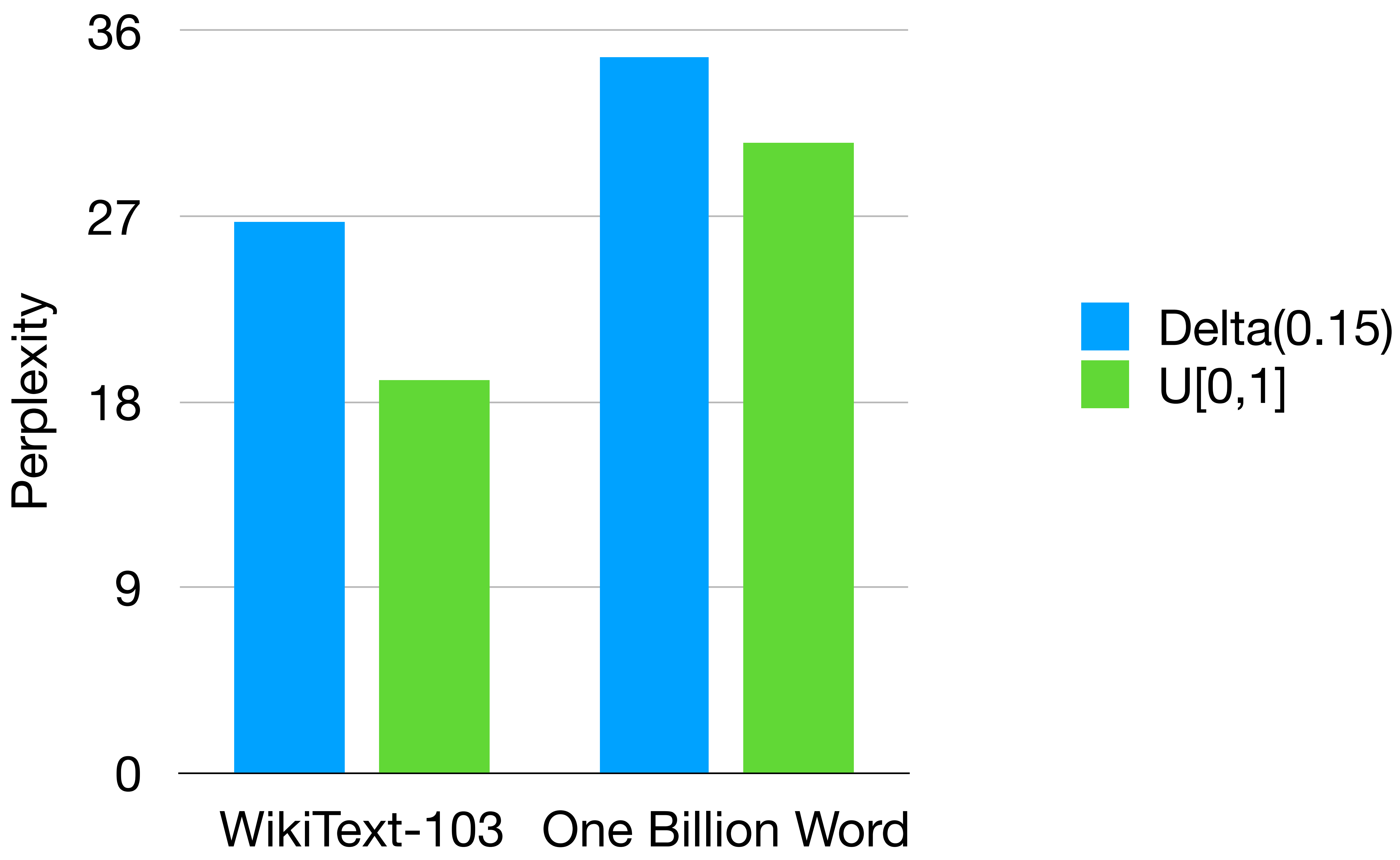
Analysis of FiLM

Pretrained-models:

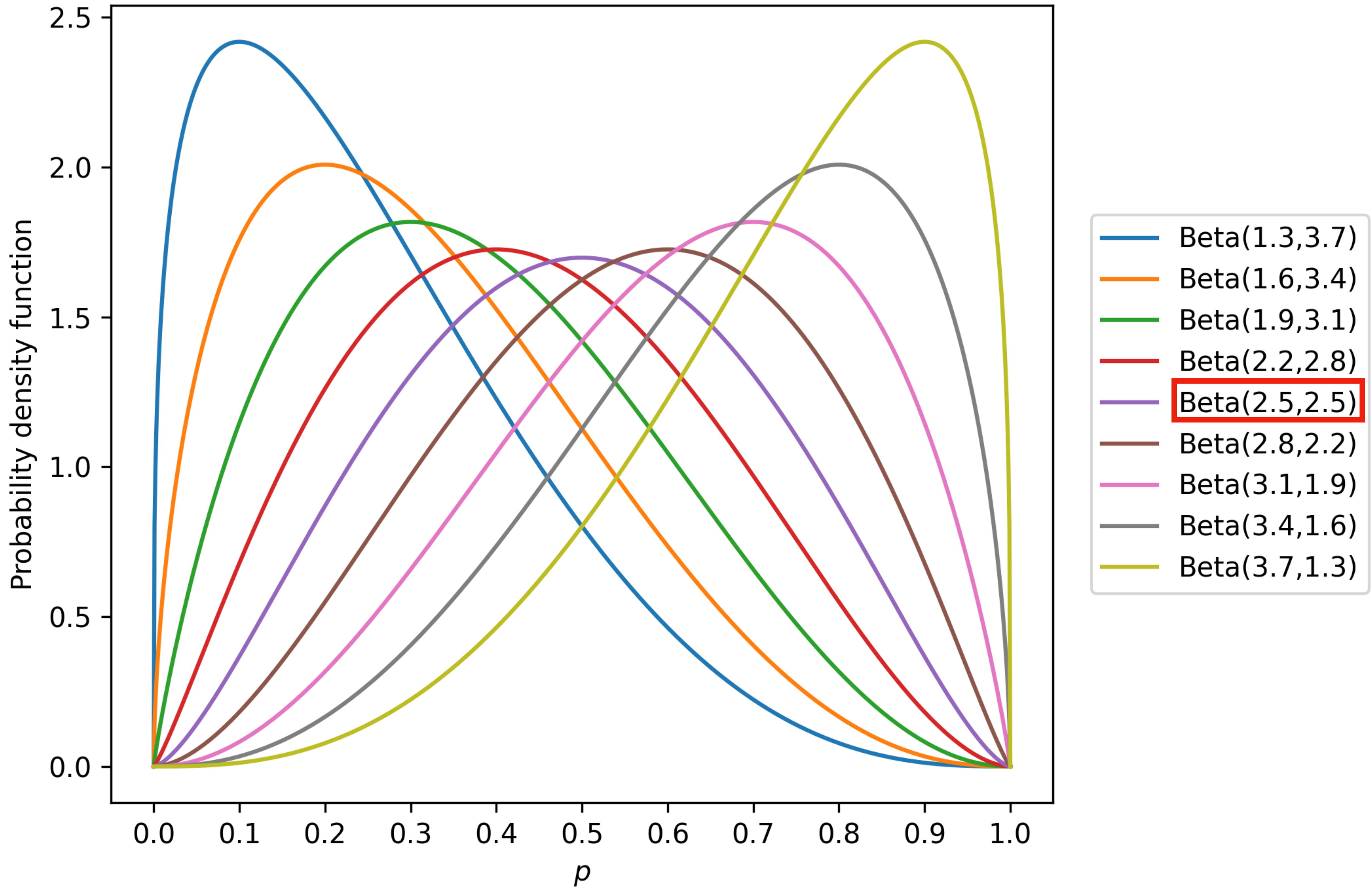
- RoBERTa (MLM)
  base (124M), large (355M)
- GPT2 (CLM)
  small (124M), medium (355M), large (774M), xl (1558M)
  disable causal mask, unshift logits

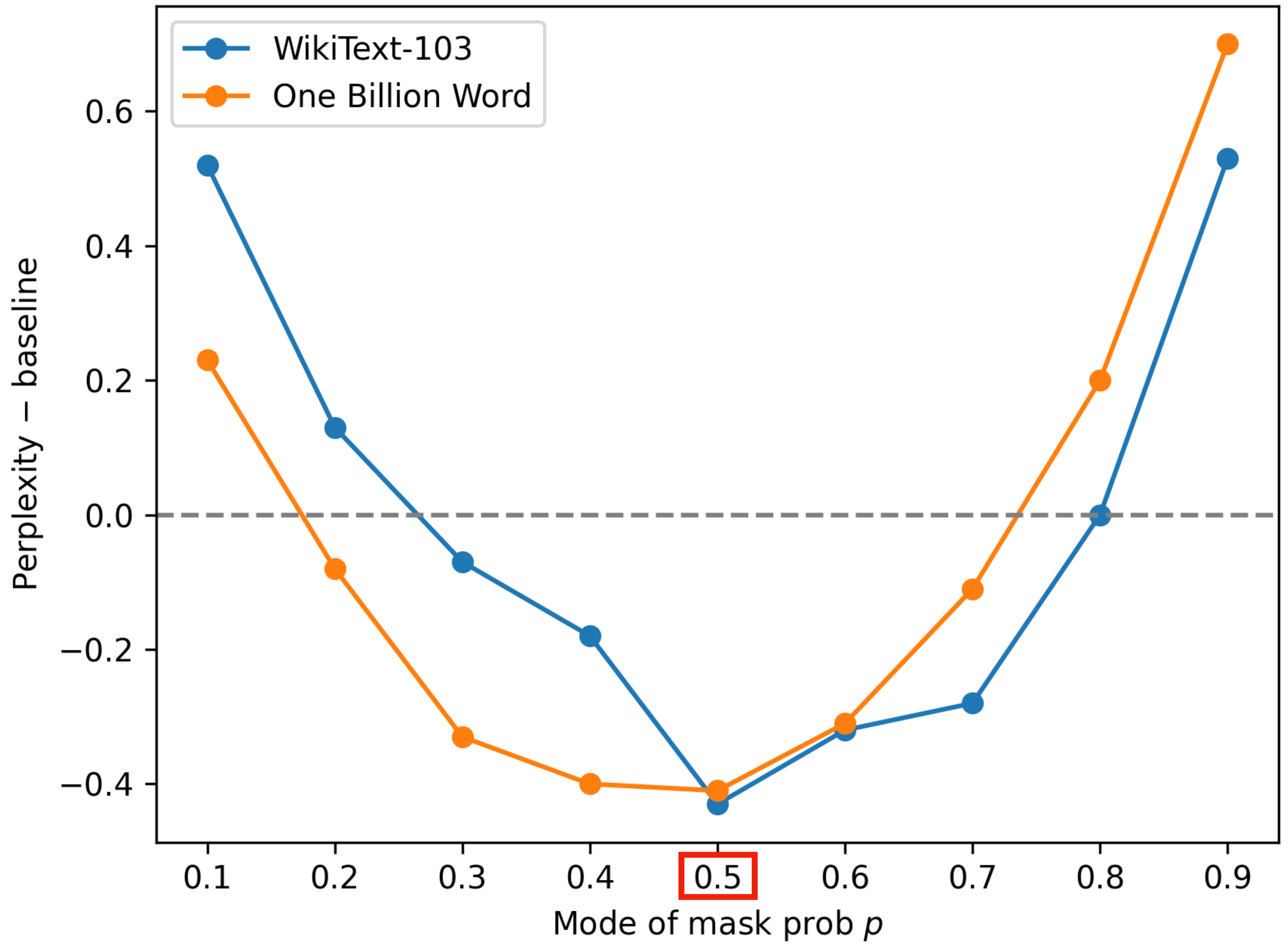
Analysis of FiLM — Training

- Pretrained model: RoBERTa-base
- Decoding order: left-to-right



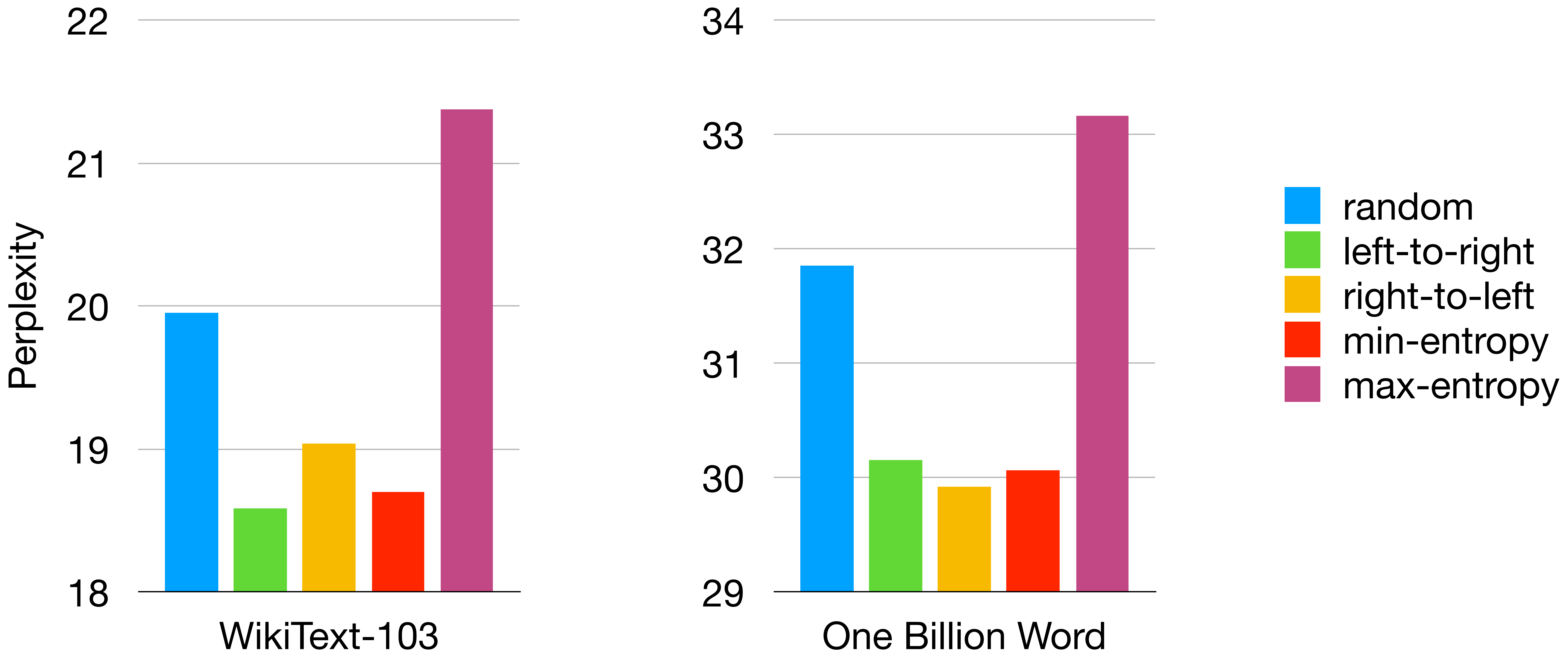
Analysis of FiLM — Training





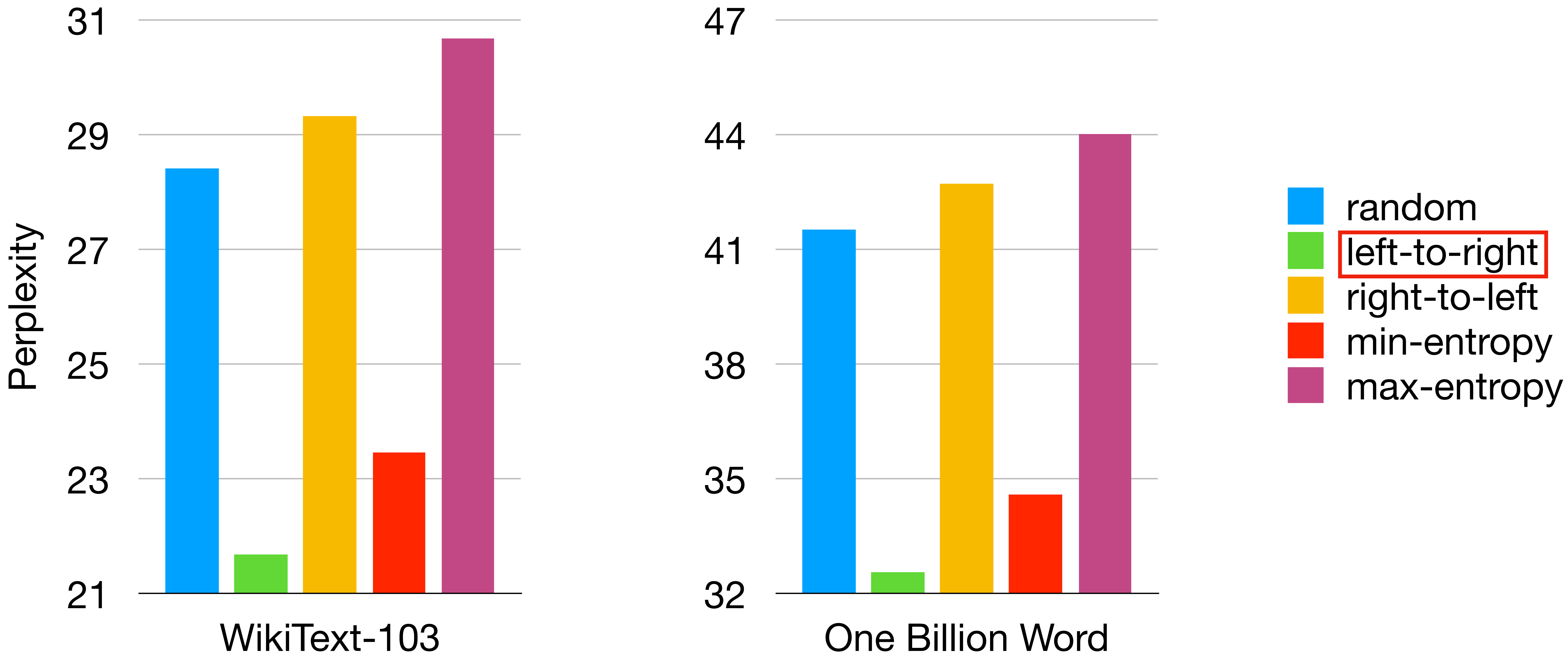
Analysis of FiLM — Decoding

- Pretrained model: RoBERTa-base

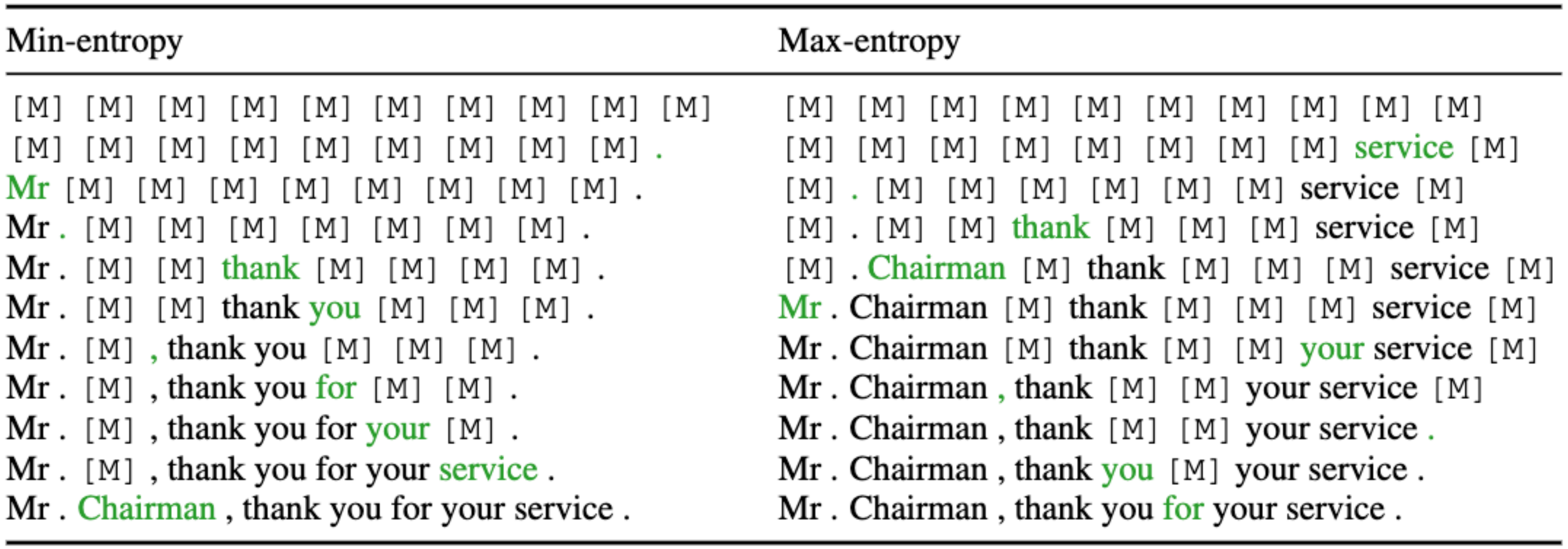


Analysis of FiLM — Decoding

- Pretrained model: GPT2-small



Analysis of FiLM — Decoding

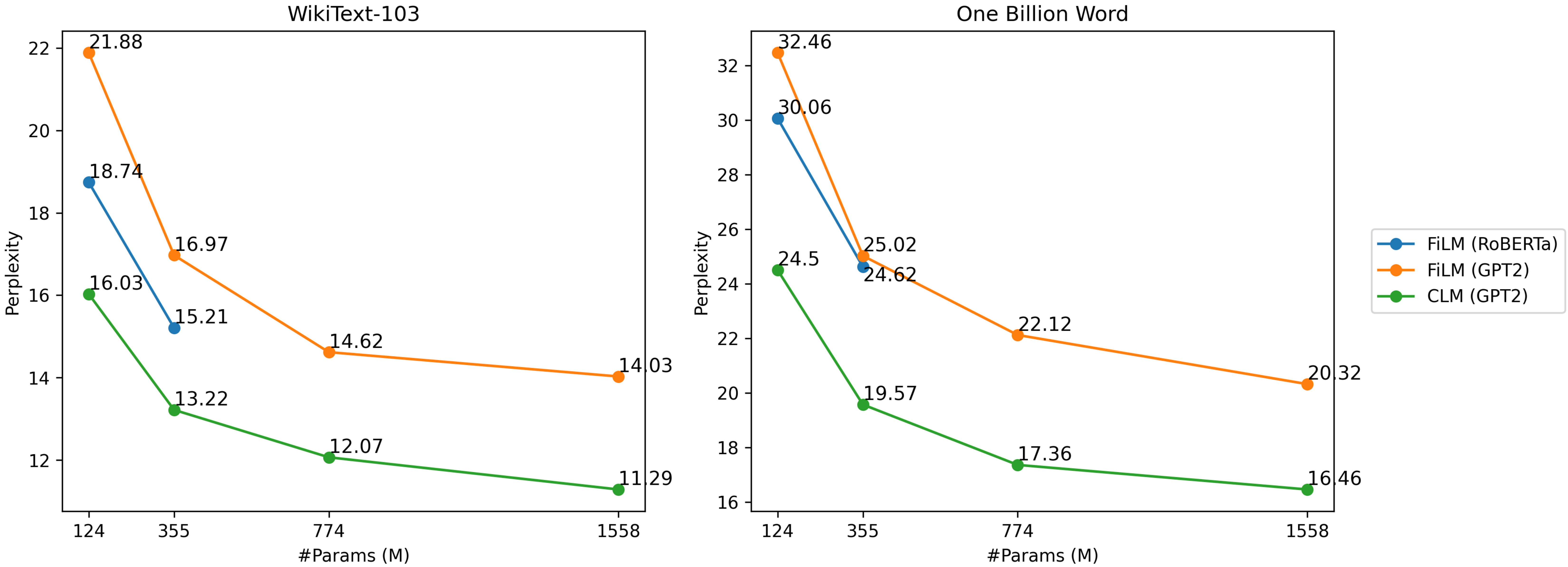


Min-entropy	Max-entropy
[M] [M] [M] [M] [M] [M] [M] [M] [M] [M]	[M] [M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M] .	[M] [M] [M] [M] [M] [M] [M] [M] service [M]
Mr [M] [M] [M] [M] [M] [M] [M] [M] .	[M] . [M] [M] [M] [M] [M] [M] service [M]
Mr . [M] [M] [M] [M] [M] [M] [M] .	[M] . [M] [M] thank [M] [M] [M] service [M]
Mr . [M] [M] thank [M] [M] [M] [M] .	[M] . Chairman [M] thank [M] [M] [M] service [M]
Mr . [M] [M] thank you [M] [M] [M] .	Mr . Chairman [M] thank [M] [M] [M] service [M]
Mr . [M] , thank you [M] [M] [M] .	Mr . Chairman [M] thank [M] [M] your service [M]
Mr . [M] , thank you for [M] [M] .	Mr . Chairman , thank [M] [M] your service [M]
Mr . [M] , thank you for your [M] .	Mr . Chairman , thank [M] [M] your service .
Mr . [M] , thank you for your service .	Mr . Chairman , thank you [M] your service .
Mr . Chairman , thank you for your service .	Mr . Chairman , thank you for your service .

Decoding process with adaptive orders. Selected position at each step is highlighted in color.

- Min-entropy generates text in a segmented order (“thank you”, “for your service”), deferring the uncertain name after “Mr.” to the end
- Max-entropy selects distant positions at each step

Language Modeling



- FiLM fine-tuned from bidirectional RoBERTa outperforms unidirectional GPT2
- The perplexity gap between FiLM and CLM decreases as model size increases

Text Infilling

- Datasets: for a sequence x of length n, sample the number of spans m from 1 to 5, then sample $2m$ endpoints from 1 to n, and mask tokens between a_{2i-1} and a_{2i} in x

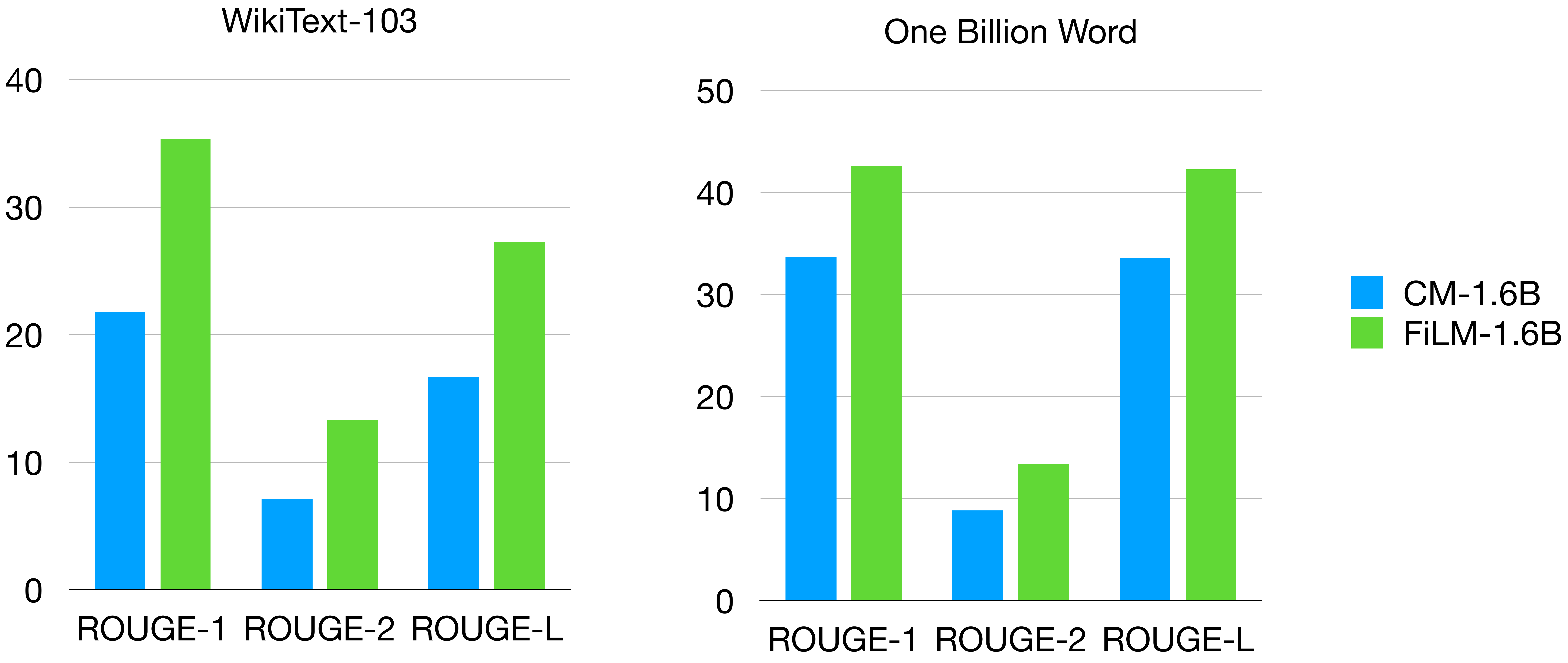
- Baseline: train CLM on rearranged data (Donahue et al., 2020; Aghajanyan et al., 2022; Fried et al., 2022; Bavarian et al., 2022), named causal masking (CM)

They have really good ice cream

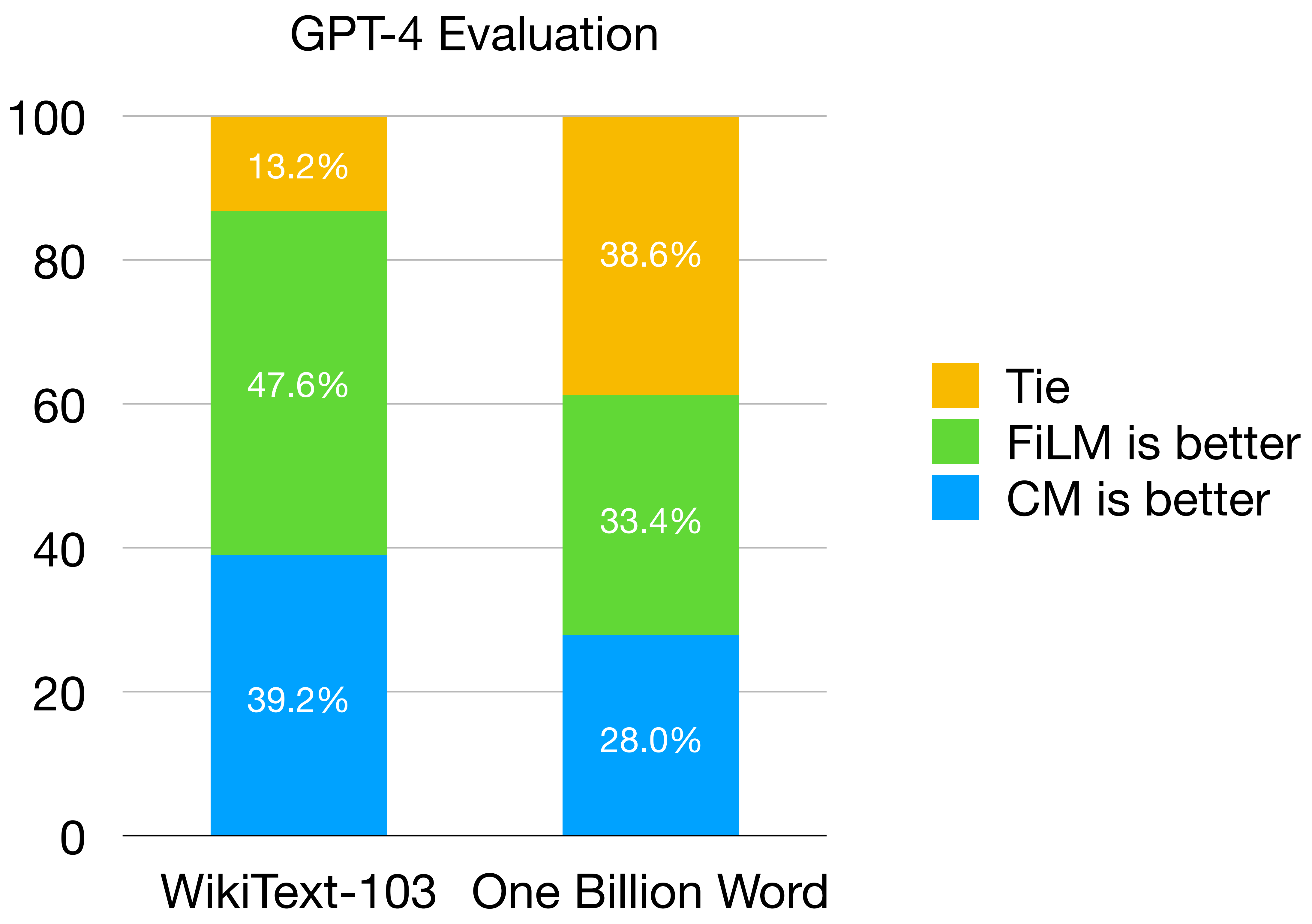
They [MASK:0] good [MASK:1] cream [FILL:0] have really [FILL:1] ice

- Evaluations:
  - ROUGE scores between generated and original text
  - GPT-4 to evaluate which output is more grammatically fluent and coherent with surrounding context

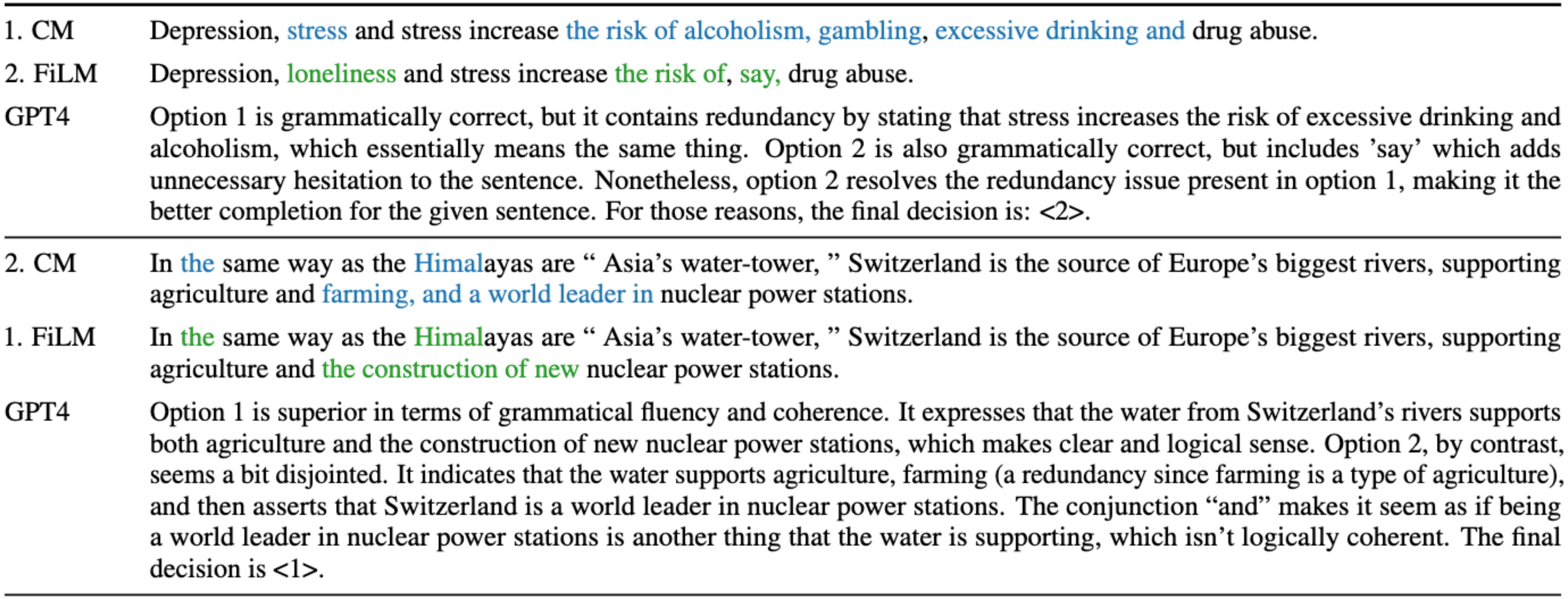
Text Infilling



Text Infilling



Text Infilling

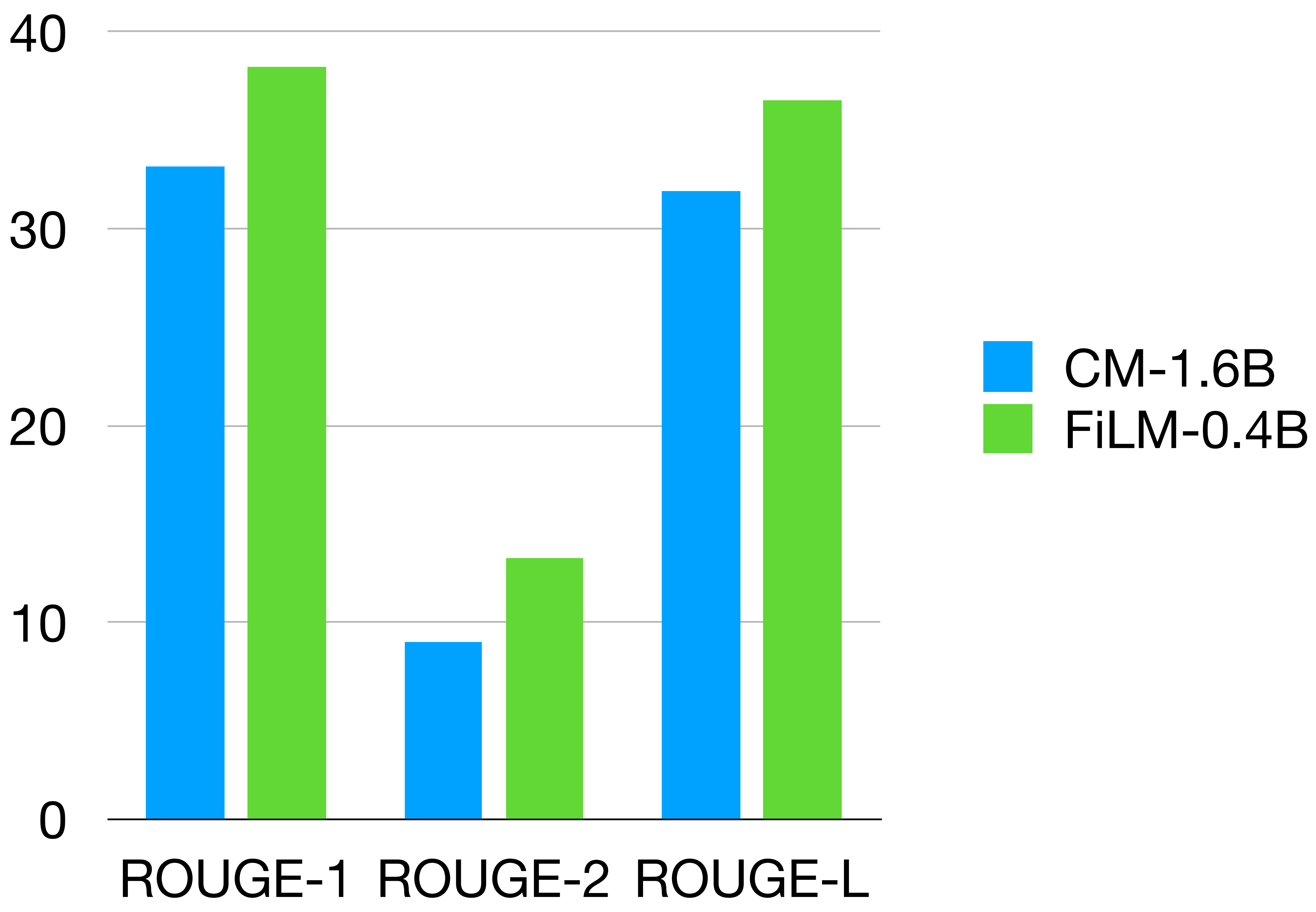


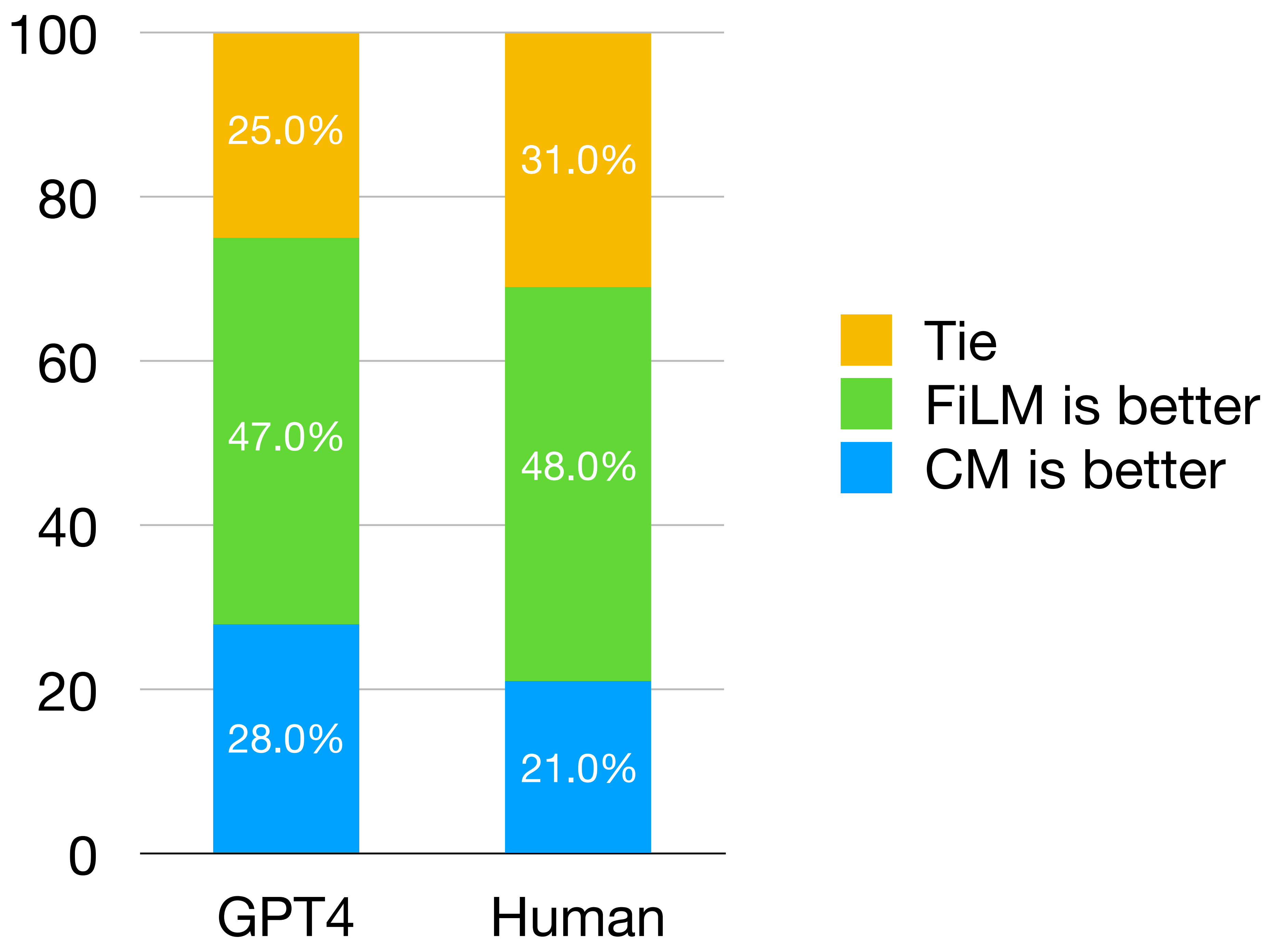
1. CM	Depression, stress and stress increase the risk of alcoholism, gambling, excessive drinking and drug abuse.
2. FiLM	Depression, loneliness and stress increase the risk of, say, drug abuse.
GPT4	Option 1 is grammatically correct, but it contains redundancy by stating that stress increases the risk of excessive drinking and alcoholism, which essentially means the same thing. Option 2 is also grammatically correct, but includes ’say’ which adds unnecessary hesitation to the sentence. Nonetheless, option 2 resolves the redundancy issue present in option 1, making it the better completion for the given sentence. For those reasons, the final decision is: <2>.
2. CM	In the same way as the Himalayas are “ Asia’s water-tower, ” Switzerland is the source of Europe’s biggest rivers, supporting agriculture and farming, and a world leader in nuclear power stations.
1. FiLM	In the same way as the Himalayas are “ Asia’s water-tower, ” Switzerland is the source of Europe’s biggest rivers, supporting agriculture and the construction of new nuclear power stations.
GPT4	Option 1 is superior in terms of grammatical fluency and coherence. It expresses that the water from Switzerland’s rivers supports both agriculture and the construction of new nuclear power stations, which makes clear and logical sense. Option 2, by contrast, seems a bit disjointed. It indicates that the water supports agriculture, farming (a redundancy since farming is a type of agriculture), and then asserts that Switzerland is a world leader in nuclear power stations. The conjunction “and” makes it seem as if being a world leader in nuclear power stations is another thing that the water is supporting, which isn’t logically coherent. The final decision is <1>.

Story Completion

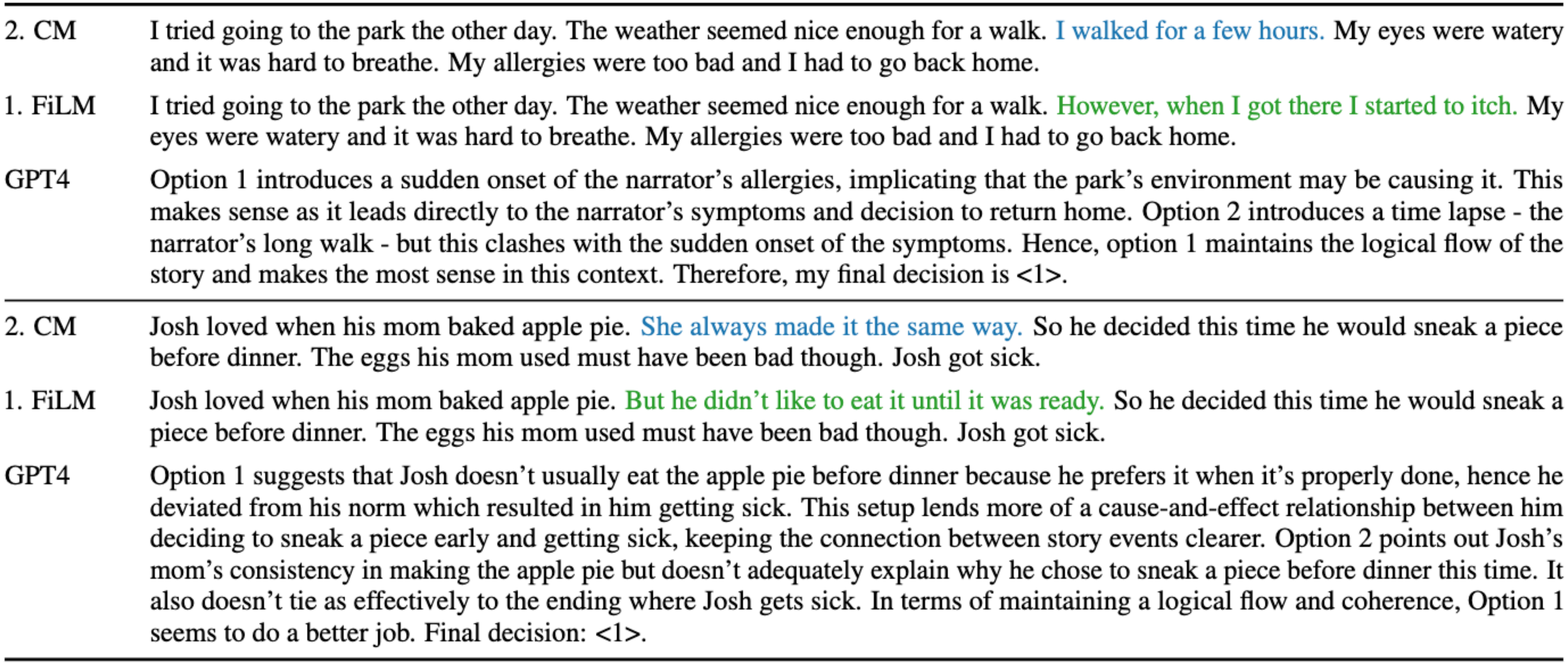
- ROCStories dataset:
  five-sentence commonsense stories, average length 51.4 tokens, 5M words in total
  randomly remove a sentence
- Evaluations:
  - ROUGE scores
  - GPT-4 to evaluate which output best preserves story's logical flow and coherence
  - Human evaluation

Story Completion





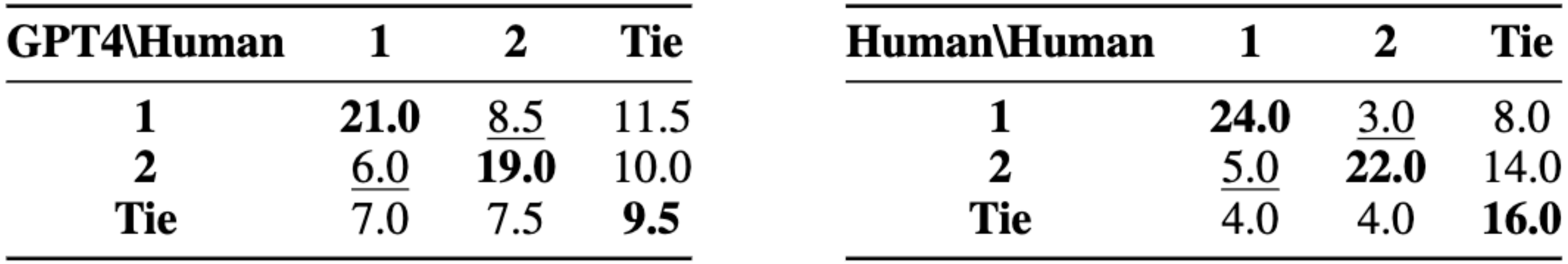
Story Completion



2. CM	I tried going to the park the other day. The weather seemed nice enough for a walk. I walked for a few hours. My eyes were watery and it was hard to breathe. My allergies were too bad and I had to go back home.
1. FiLM	I tried going to the park the other day. The weather seemed nice enough for a walk. However, when I got there I started to itch. My eyes were watery and it was hard to breathe. My allergies were too bad and I had to go back home.
GPT4	Option 1 introduces a sudden onset of the narrator's allergies, implicating that the park's environment may be causing it. This makes sense as it leads directly to the narrator's symptoms and decision to return home. Option 2 introduces a time lapse - the narrator's long walk - but this clashes with the sudden onset of the symptoms. Hence, option 1 maintains the logical flow of the story and makes the most sense in this context. Therefore, my final decision is <1>.
2. CM	Josh loved when his mom baked apple pie. She always made it the same way. So he decided this time he would sneak a piece before dinner. The eggs his mom used must have been bad though. Josh got sick.
1. FiLM	Josh loved when his mom baked apple pie. But he didn't like to eat it until it was ready. So he decided this time he would sneak a piece before dinner. The eggs his mom used must have been bad though. Josh got sick.
GPT4	Option 1 suggests that Josh doesn't usually eat the apple pie before dinner because he prefers it when it's properly done, hence he deviated from his norm which resulted in him getting sick. This setup lends more of a cause-and-effect relationship between him deciding to sneak a piece early and getting sick, keeping the connection between story events clearer. Option 2 points out Josh's mom's consistency in making the apple pie but doesn't adequately explain why he chose to sneak a piece before dinner this time. It also doesn't tie as effectively to the ending where Josh gets sick. In terms of maintaining a logical flow and coherence, Option 1 seems to do a better job. Final decision: <1>.

Story Completion

Agreement percentages

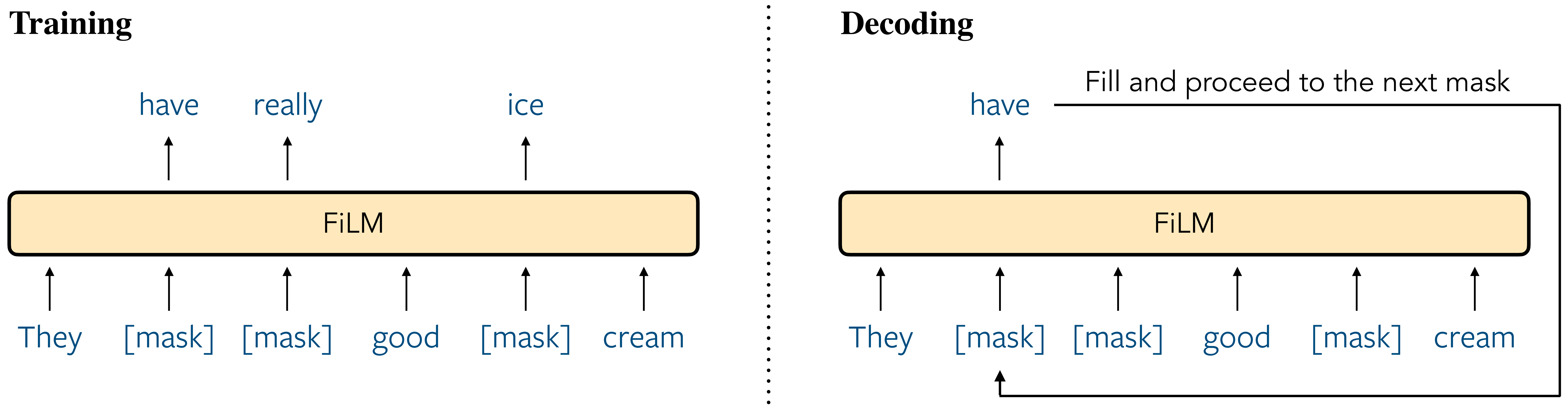


GPT4\Human	1	2	Tie
1	**21.0**	8.5	11.5
2	6.0	**19.0**	10.0
Tie	7.0	7.5	**9.5**

Human\Human	1	2	Tie
1	**24.0**	3.0	8.0
2	5.0	**22.0**	14.0
Tie	4.0	4.0	**16.0**

- GPT4-Human consensus: 49.5%
- Human-human consensus: 62%

Summary



- Training: sample the mask probability from Beta distribution
- Decoding: from left to right or select the position with min entropy
- FiLM's perplexity approaches CLM as model size increases → potential as an alternative LLM
- FiLM excels in text infilling and story completion and outperforms strong baselines

https://github.com/shentianxiao/FiLM