Revisiting the Architectures like Pointer Networks to Efficiently Improve the Next Word Distribution, Summarization Factuality, and Beyond

Haw-Shiuan Chang*[1,2], Zonghai Yao*[1], Alolika Gon[1], Hong Yu[1], Andrew McCallum[1]

[1]CICS, University of Massachusetts, Amherst, [2]Amazon Alexa AI

Introduction

- Background & Motivation

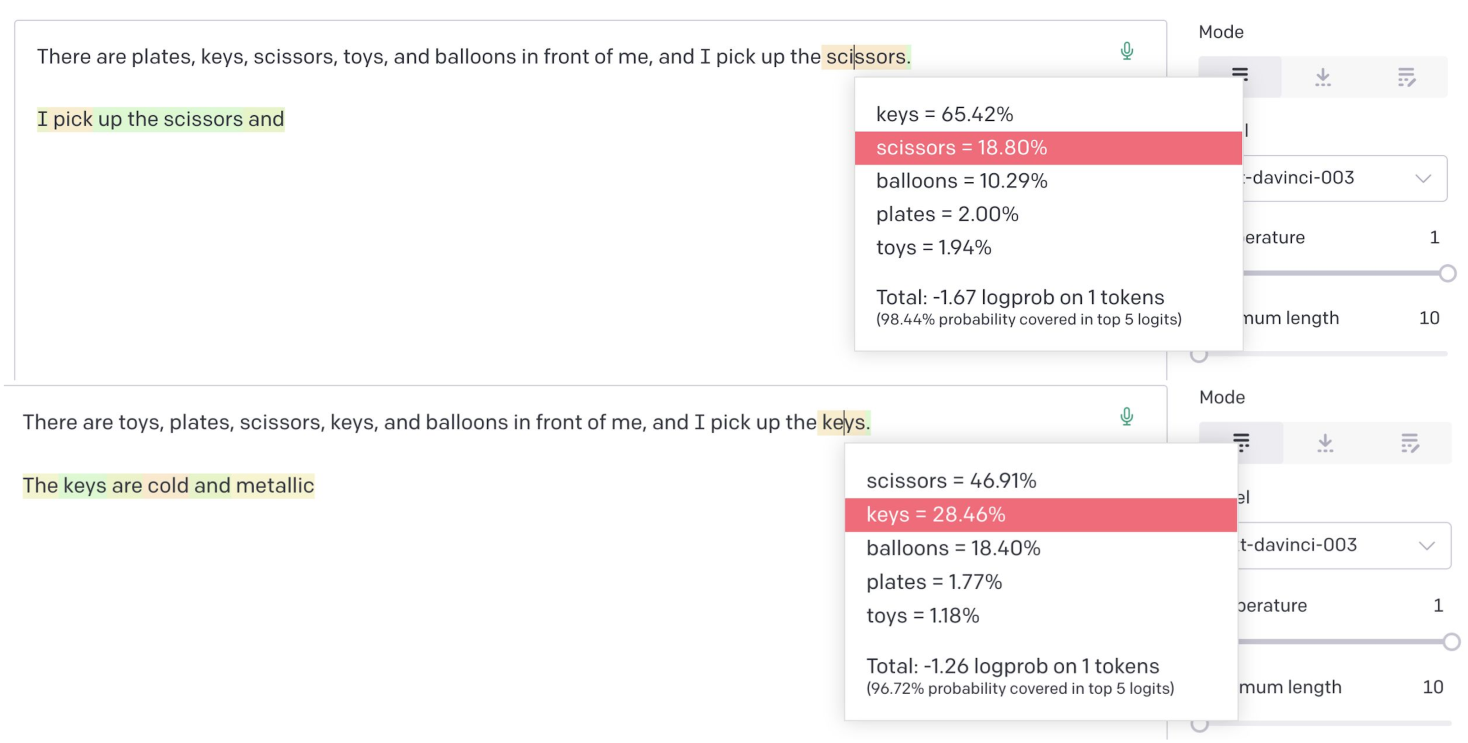
1. Can Large LM Learn to Output Arbitrary Next Word Distribution? NO

- There are plates, keys, scissors, toys, and balloons in front of me, and I pick up the ...

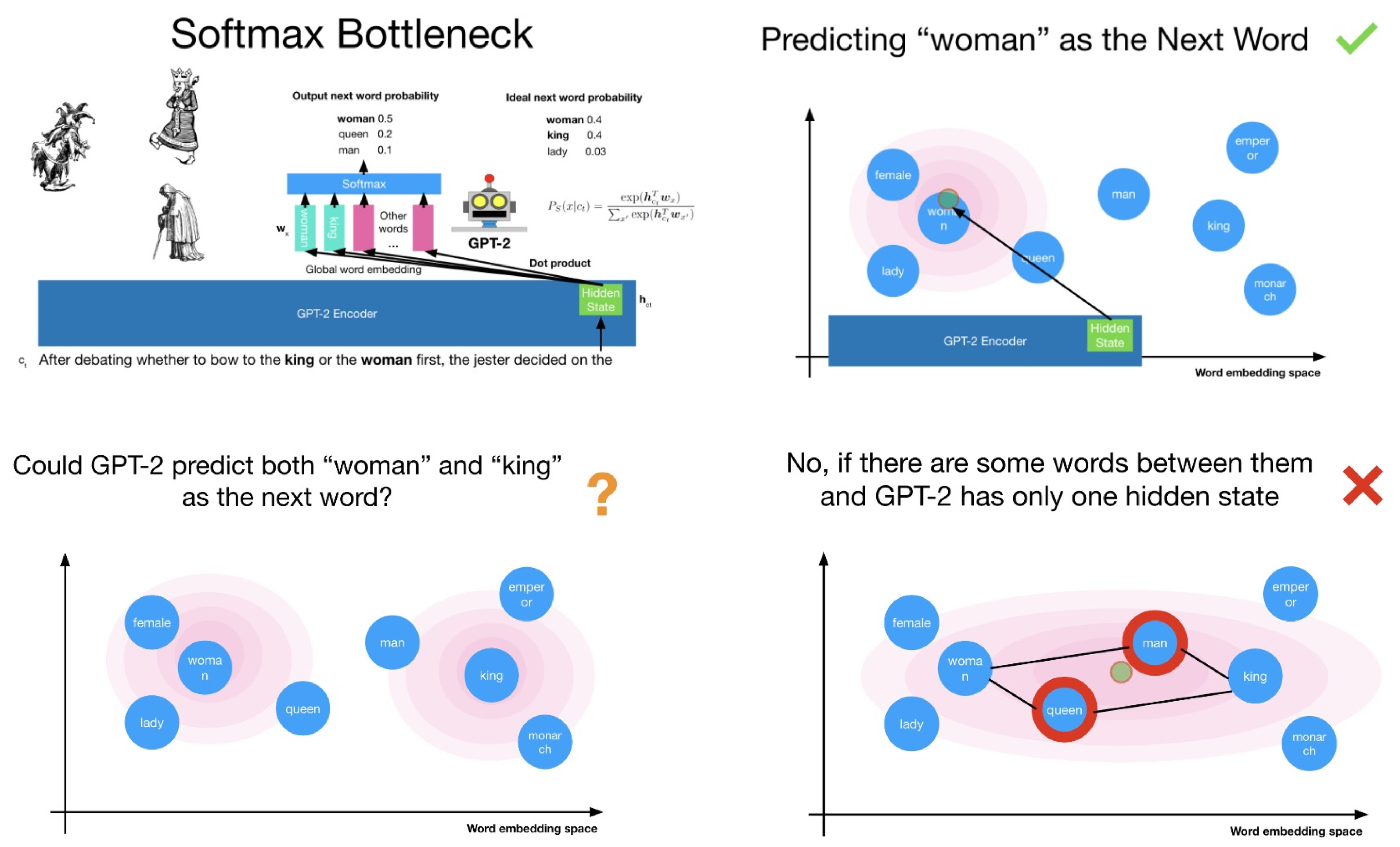
Ideal distribution
- plates ~0.2
- keys ~0.2
- scissors ~0.2
- toys ~0.2
- balloons ~0.2

- There are plates, keys, scissors, toys, and balloons in front of me, and I pick up the ...
  - phone (from GPT-2)?
  - Hallucination
    - Should copy but not copy
- I like tennis, baseball, golf, basketball, and ...
  - tennis (from GPT-2)?
  - Repetition
    - Should not copy but copy

GPT3.5's output



2. Why is Softmax Unable to Learn to Copy Properly (Chang and McCallum, 2022)?



- Contributions:

1. We propose a series of efficient softmax alternatives that unify the ideas of pointer network, reranker, multiple embeddings, and vocabulary partitioning.
2. We evaluate the proposed softmax alternatives in text completion tasks and summarization tasks using various metrics to identify where our methods improve the most.
3. Our experiments indicate pointer networks and our proposed alternatives can still improve the modern transformer-based LMs. By breaking the softmax bottleneck, our methods learn sometimes to copy the context words to reduce generation hallucination and sometimes exclude the context words to reduce the repetition.

Methods: Softmax-CPR

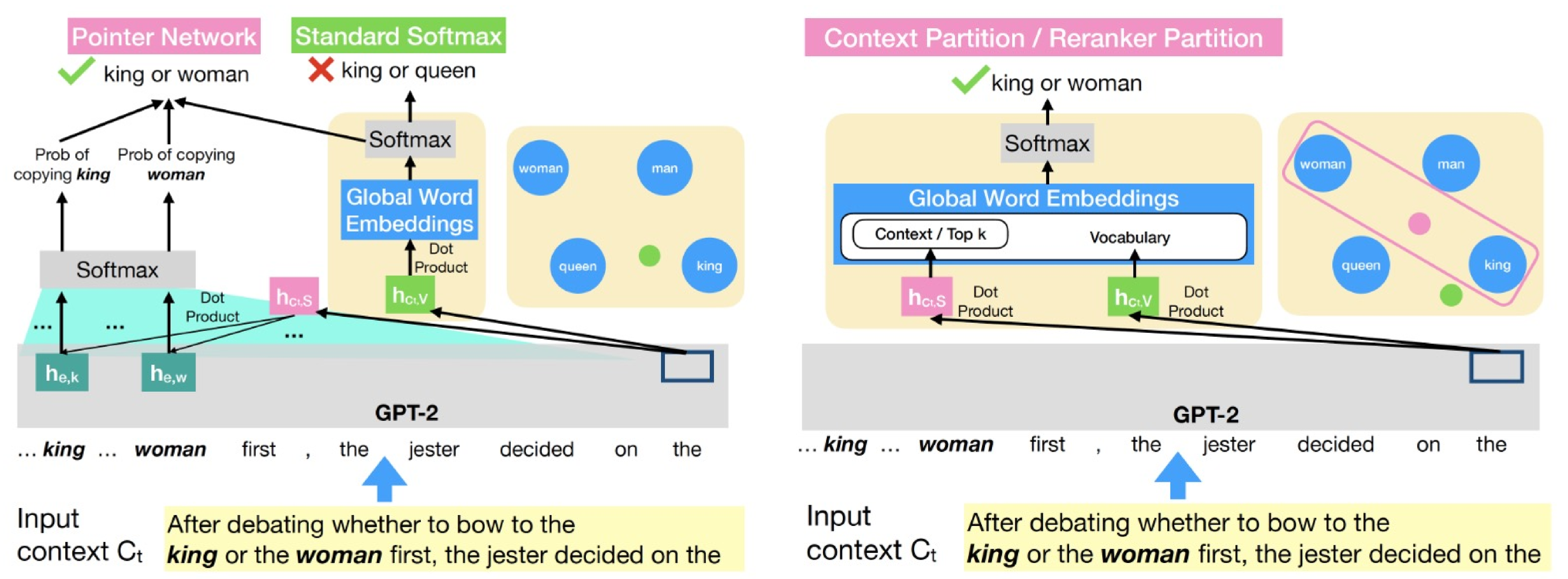


Figure 1: Left: Illustration of the softmax bottleneck and pointer network. Right: We simplify the pointer network / reranker by using another embedding $\mathbf{h}_{c_t,S}$ for the words in the context / the top-k likely words.

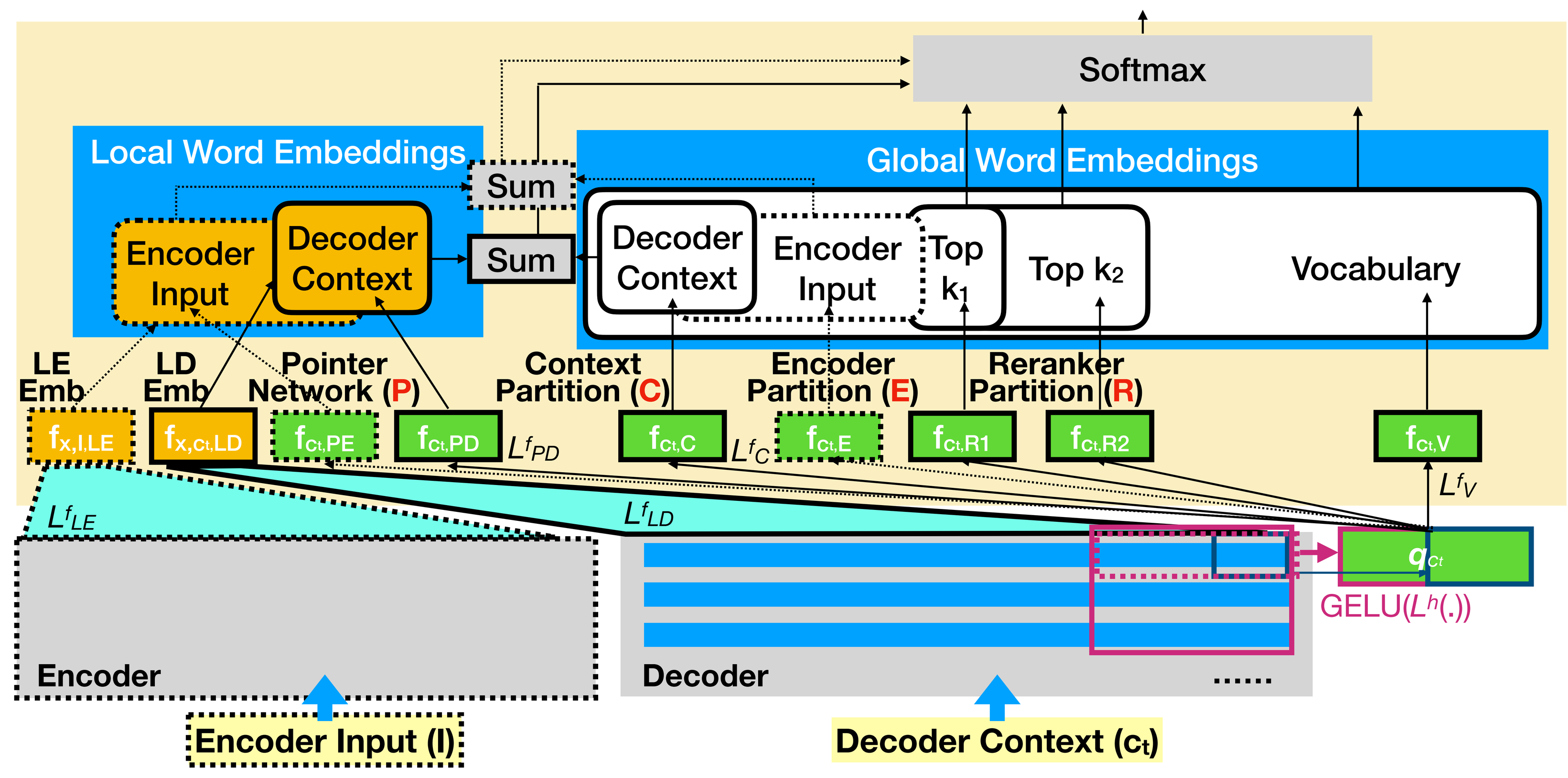


Figure 2: Architectures of our method for T5/BART that computes Logit_{CEPR}. In GPT-2, we use same architecture except that we take the 3x3 input hidden state block rather than the 1x3 block and there are no encoder-related components, which are marked by dotted lines.

Experimental Results

GPT-2 Perplexity Comparison

Model Name	Size (GPT-2 Small)	Time (ms)	OWT (↓)	Wiki (↓)
Softmax (GPT-2)	125.0M	82.9	18.96	24.28
Softmax + Mi	130.9M	85.6	18.74	24.08
Mixture of Softmax (MoS) (Yang et al., 2018)	126.2M	130.2	18.97	24.10
MoS + Mi (Chang and McCallum, 2022)	133.3M	133.2	18.68	23.82
Pointer Generator (PG) (See et al., 2017)	126.2M	106.0	18.67	23.70
Pointer Sentinel (PS) (Merity et al., 2017)	126.2M	94.1	18.70	23.79
Softmax + R:20 + Mi	132.1M	90.4	18.67	24.03
Softmax + R:20,100 + Mi	133.3M	101.1	18.69	23.93
Softmax + C + Mi	132.1M	94.8	18.48	23.56
Softmax + P + Mi	133.3M	99.1	18.58	23.66
PG + Mi	133.3M	111.2	18.43	23.43
PS + Mi	133.3M	98.0	18.48	23.53
Softmax + CR:20,100 + Mi	134.5M	113.3	18.46	23.48
Softmax + CPR:20,100 + Mi	136.8M	119.9	18.43	23.42
MoS + CPR:20,100 + Mi	139.2M	165.1	**18.39**	**23.29**

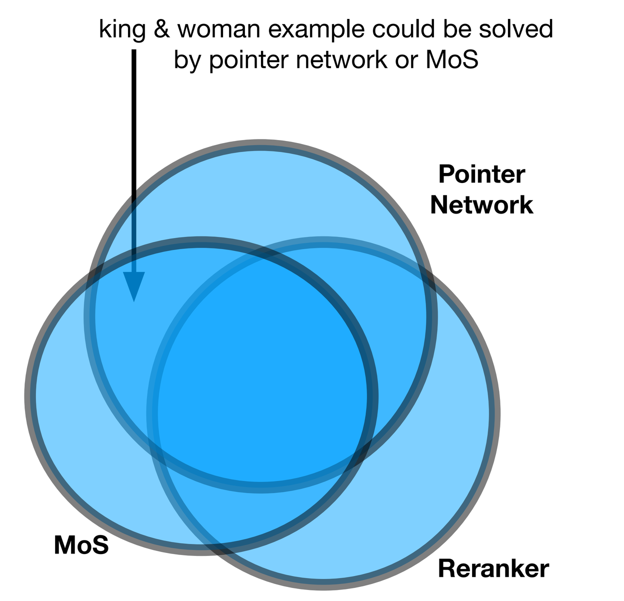


Figure 3: This table shows that dynamic partitioning are very helpful in terms of perplexity. Lower perplexity is better

Summarization Experiments

- Improve BookSum more
  - Probably because the names in narrative text are usually locally defined

Model Name	CNN/DM R1	CNN/DM CIDEr	CNN/DM factCC	CNN/DM MAUVE	XSUM R1	XSUM CIDEr	XSUM factCC	XSUM MAUVE	BookSum Paragraph R1	BookSum Paragraph CIDEr	BookSum Paragraph factCC	BookSum Paragraph MAUVE	SAMSUM R1	SAMSUM CIDEr	SAMSUM factCC	SAMSUM MAUVE
T5-Small																
Softmax (S)	38.255	0.442	0.462	0.861	28.713	0.446	0.254	0.939	16.313	0.083	0.424	0.328	39.472	0.817	0.577	0.898
CopyNet (Gu et al., 2016)	37.990	0.438	0.482	0.865	28.573	0.442	0.274	0.940	16.666	0.092	0.439	0.402	39.525	0.853	0.579	0.924
PG (See et al., 2017)	37.913	0.442	0.467	0.874	28.777	0.450	0.257	0.931	16.432	0.088	0.429	0.376	32.451	0.585	0.552	0.153
PS (Merity et al., 2017)	38.058	0.444	[illegible]	[illegible]	[illegible]	[illegible]	0.267	0.932	16.408	0.090	0.436	[illegible]	38.731	0.817	0.578	0.865
S + R:20	37.881	0.433	[illegible]	[illegible]	[illegible]	[illegible]	0.256	0.931	16.336	0.086	0.431	[illegible]	39.073	0.752	0.579	0.847
S + E	38.137	0.441	[illegible]	[illegible]	[illegible]	[illegible]	0.272	0.942	16.542	0.090	0.435	0.390	39.056	0.784	0.579	0.904
S + CE	38.461	0.460	0.475	0.874	29.155	0.464	0.270	**0.948**	16.628	0.093	0.436	0.403	40.055	0.835	**0.583**	0.943
S + CER:20	38.346	0.450	**0.482**	**0.890**	29.067	0.459	**0.276**	0.942	16.638	0.093	0.436	0.400	**40.505**	0.846	0.580	0.915
S + CEPR:20	**38.807**	**0.456**	0.481	0.877	**29.395**	**0.474**	0.273	0.942	**16.894**	**0.098**	**0.440**	0.418	40.127	**0.891**	0.582	**0.946**
S + CEPR:20 + Mi	38.675	0.451	0.475	0.878	29.348	0.470	0.275	0.946	16.738	0.096	0.438	**0.426**	40.328	0.874	0.582	0.932
T5-Base																
Softmax (S)	40.198	0.504	0.478	0.907	33.571	0.667	0.249	0.979	16.761	0.096	0.424	0.467	44.348	1.046	0.574	**0.986**
CopyNet (Gu et al., 2016)	39.940	0.507	0.484	0.903	33.557	0.666	0.253	0.979	16.918	0.101	0.430	0.531	44.141	1.052	0.570	0.973
PG (See et al., 2017)	39.982	0.489	0.485	0.911	33.605	0.663	0.255	0.982	16.611	0.095	0.423	0.463	37.597	0.784	0.548	0.140
PS (Merity et al., 2017)	40.018	0.495	0.483	0.914	33.638	0.672	0.249	**0.983**	16.905	0.100	0.428	0.504	43.098	1.008	0.575	0.946
S + CEPR:20	40.354	**0.511**	**0.487**	**0.919**	33.700	0.675	0.260	0.980	**16.997**	0.100	**0.432**	**0.549**	**44.860**	**1.064**	0.573	0.963
S + CEPR:20 + Mi	**40.510**	0.506	0.481	0.918	**33.853**	**0.683**	**0.263**	**0.983**	16.975	**0.101**	0.431	0.546	44.488	1.055	**0.576**	0.980

Comparable to some reranker methods

+ 30%

Figure 4: The performance on test sets of four summarization datasets.

Conclusion

1. We propose softmax-CPR and softmax-CEPR, which unify the ideas of the pointer network, reranker, and mixture of softmax (MoS)
   (a) Alleviate hallucination and repetition problem
   (b) mostly by learning to copy the words from context properly
2. Pointer networks significantly boost summarization factuality
   (a) their improvements mainly come from breaking the softmax bottleneck rather than its attention mechanism
   (b) Softmax-CPR could bring even more improvements

Reference

Chang, Haw-Shiuan, and Andrew McCallum. "Softmax bottleneck makes language models unable to represent multi-mode word distributions." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.