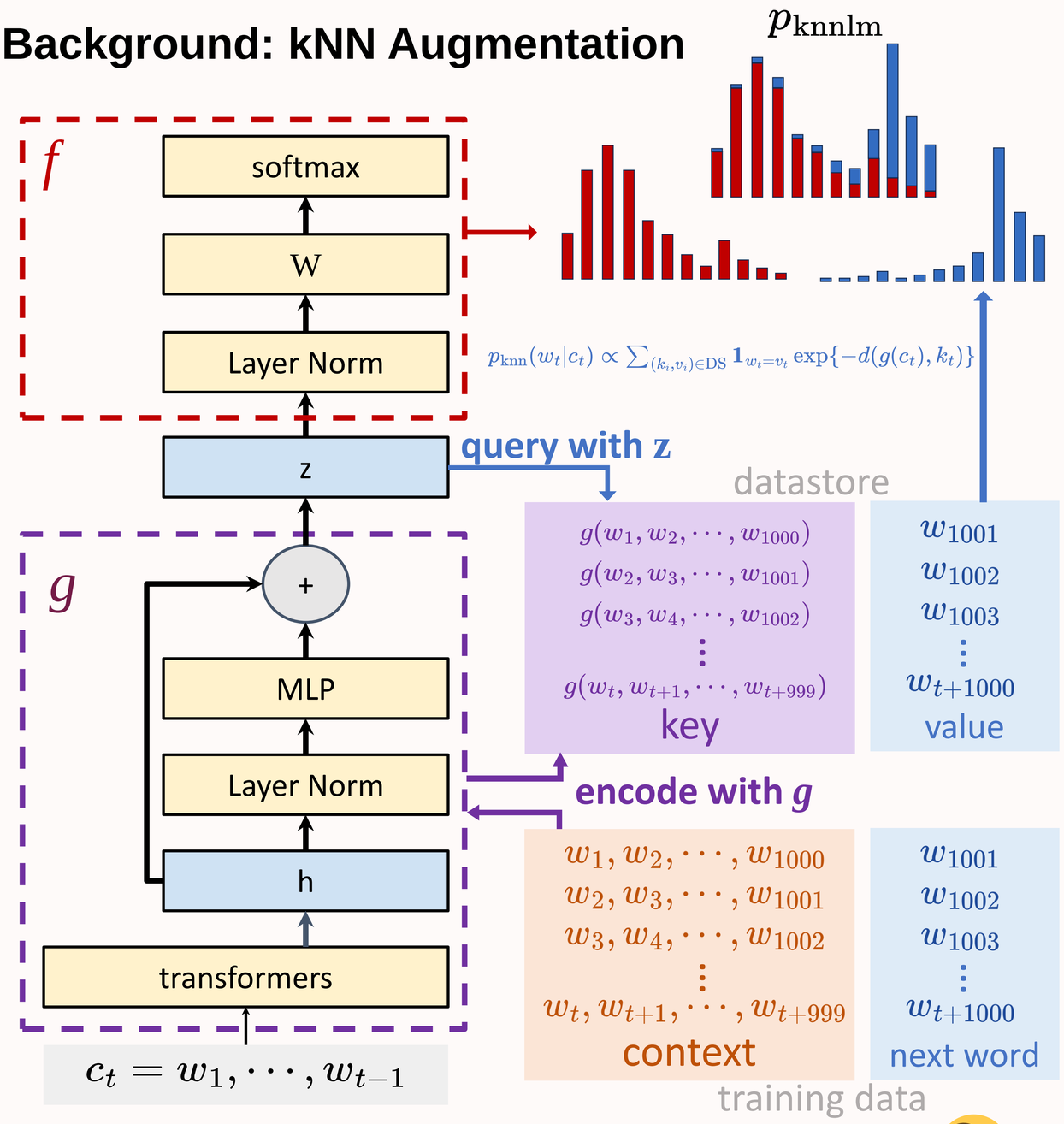
On Retrieval Augmentation and the Limitations of Language Model Training

Ting-Rui Chiang, Xinyan Velocity Yu, Joshua Robinson, Ollie Liu, Isabelle Lee, Dani Yogatama

tamagotchi lab @USC

Background: kNN Augmentation

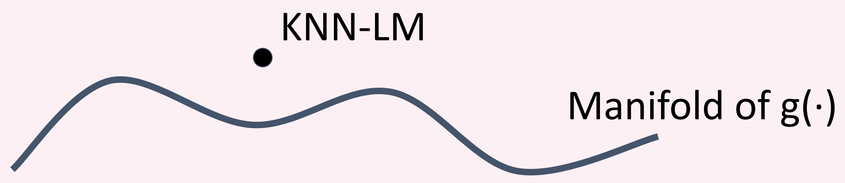


Without extra data, kNN augmentation can improve the LM

RQ1: Is the performance gap caused by the softmax bottleneck?

Softmax bottleneck (SB):
Given any W, no matter what z is, some distributions can't be generated.

Hypothesis (Xu et al., 2023)**:**
kNN-LMs can generate distributions that vanilla LMs can't generate



Inspecting the hypothesis:
(1) Find input z^* that makes the last layer approximate p_{knnlm}.

$$z^* \in \arg\min_{z\in\mathbb{R}^d} \text{KL}[f(z)\|p_{\text{knnlm}}] \qquad \text{Let } p_{\text{proj}} = f(z^*)$$

(2) Evaluate the approximation based on the perplexity of p_{proj}.

Original LMs		Projected
p_{lm}	p_{knnlm}	p_{project}
20.13	16.92	16.78

The projection can approximate p_{knnlm} well!

Thus, the softmax bottleneck is not the cause of the performance gap.

RQ2: Is the performance gap caused by underfitting?

Previous Hypothesis (Xu et al., 2023)**:**
kNN-LM performs better because it memorizes the training data better.

Our hypothesis:
The LM training objective has limitations and can't handle some cases. For example, we identify the over-specification scenario.

[Definition] Over-specification:
The prefix of a partial sentence contains information that is not causally relevant to its continuation.

[Proposed Dataset]: Macondo

Training set: *[villager], who [desc], is the parent of [child].*
Test set: *[villager], is the parent of [child].*

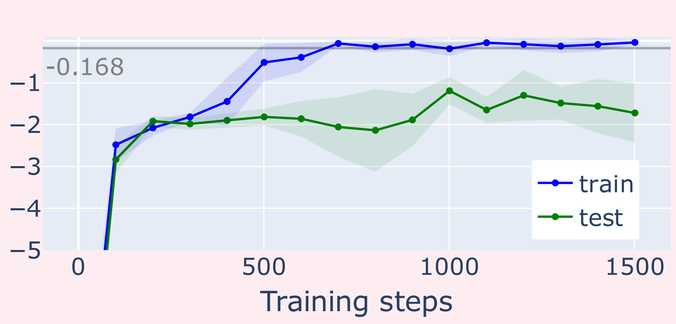
For example,
This part is not causally relevant to the continuation.
"Fifine Lottman, who used to work for Fox Broadcasting Company, is the parent of Hayward."

LMs trained with the above sentence may struggle to complete the following sentence

"Fifine Lottman is the parent of _____" (Hayward)

Fine-tuning with Macondo

GPT-2-XL: kNN-LM (dotted line) consistently outperforms the vanilla LM (solid lines).



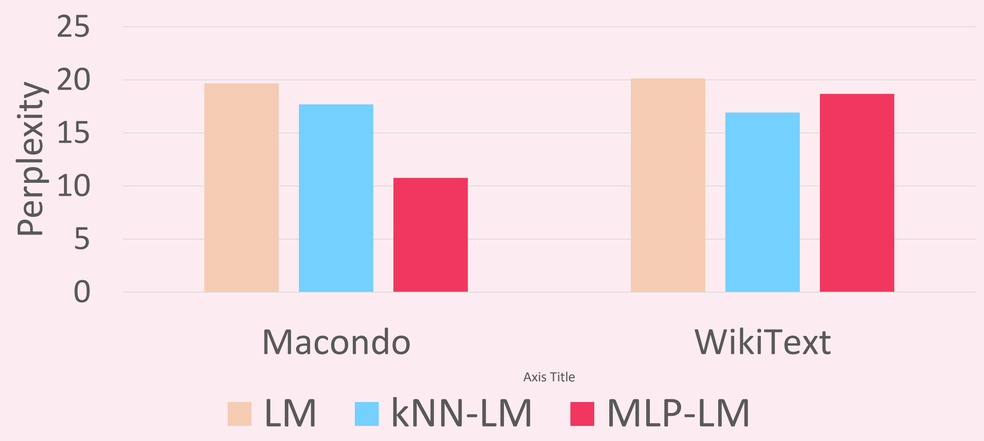
GPT-3-Turbo: It is still far from perfect.

The LM training objective has limitations and kNN augmentation can somewhat mitigate it.

RQ3: Any alternatives to kNN augmentation?

We propose *MLP-augmentation*:
Instead of using the key-value pairs in the datastore to build a kNN model, we use the pairs to train an MLP model.

Results:



Compared with a kNN datastore, an extra MLP layer requires only 4% of the storage space.

Augmenting with MLP may be a promising future direction.

Conclusion:

- **The LM training objective has limitations.**
- **Special training techniques could be a promising future direction.**

Reference
Frank F. Xu, Uri Alon, and Graham Neubig. 2023. Why do nearest neighbor language models work?

USC NLP

USC University of Southern California

paper on arxiv