Nearest Neighbour LMs

NLP: Fall 2024

Anoop Sarkar

kNN LM

Improving Neural Language Models with a Continuous Cache

Edouard Grave, Armand Joulin, Nicolas Usunier
Facebook AI Research
{egrave,ajoulin,usunier}@fb.com

https://arxiv.org/abs/1612.04426

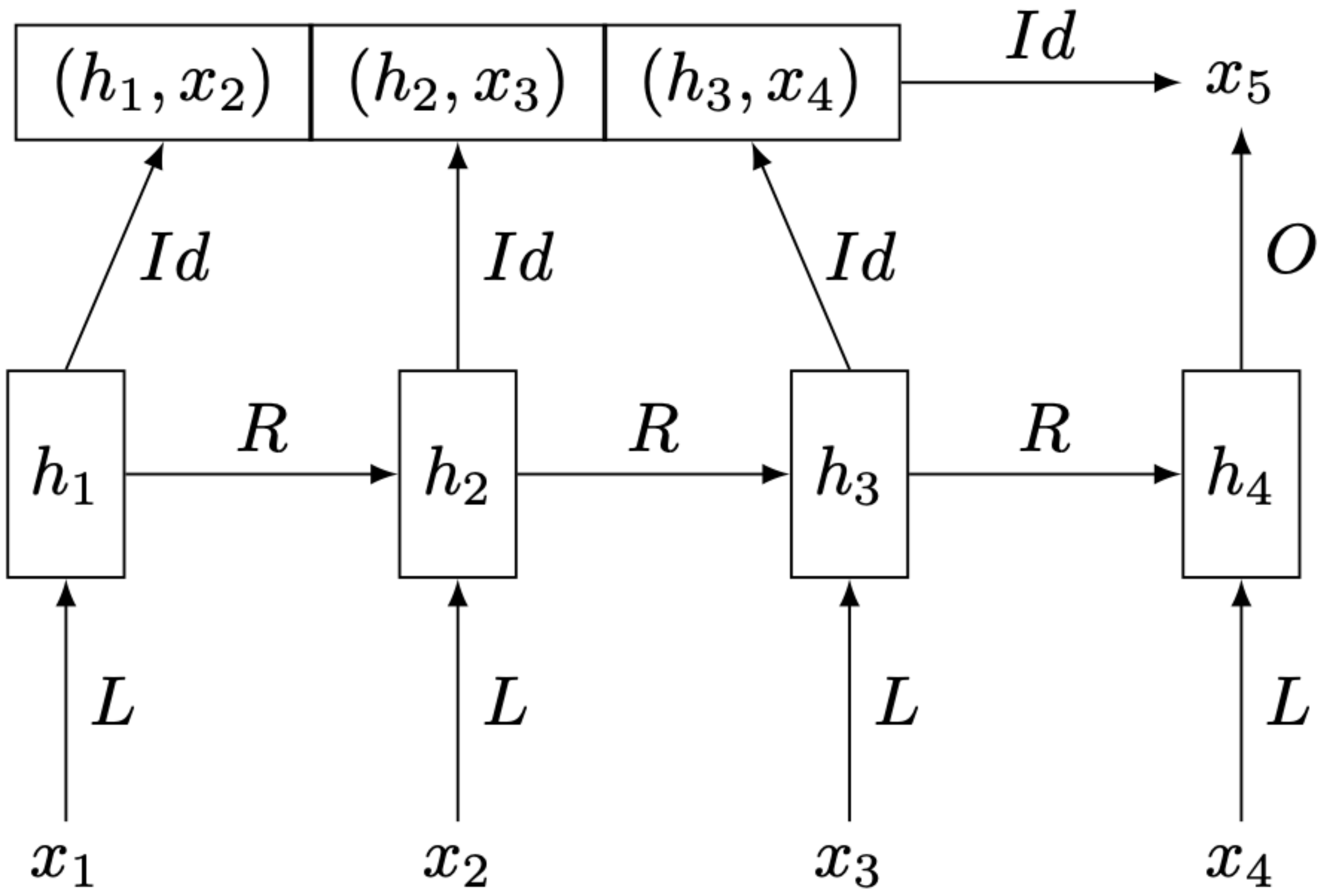


Figure 1: The neural cache stores the previous hidden states in memory cells. They are then used as keys to retrieve their corresponding word, that is the next word. There is no transformation applied to the storage during writing and reading.

$$p_{cache}(w \mid h_{1..t},\ x_{1..t}) \propto \sum_{i=1}^{t-1} \mathbb{1}_{\{w=x_{i+1}\}} \exp(\theta h_t^\top h_i)$$

Generalization through Memorization: Nearest Neighbor Language Models

Urvashi Khandelwal[†, *], **Omer Levy**[‡], **Dan Jurafsky**[†], **Luke Zettlemoyer**[‡] **& Mike Lewis**[‡]
[†]Stanford University
[‡]Facebook AI Research
{urvashik,jurafsky}@stanford.edu
{omerlevy,lsz,mikelewis}@fb.com

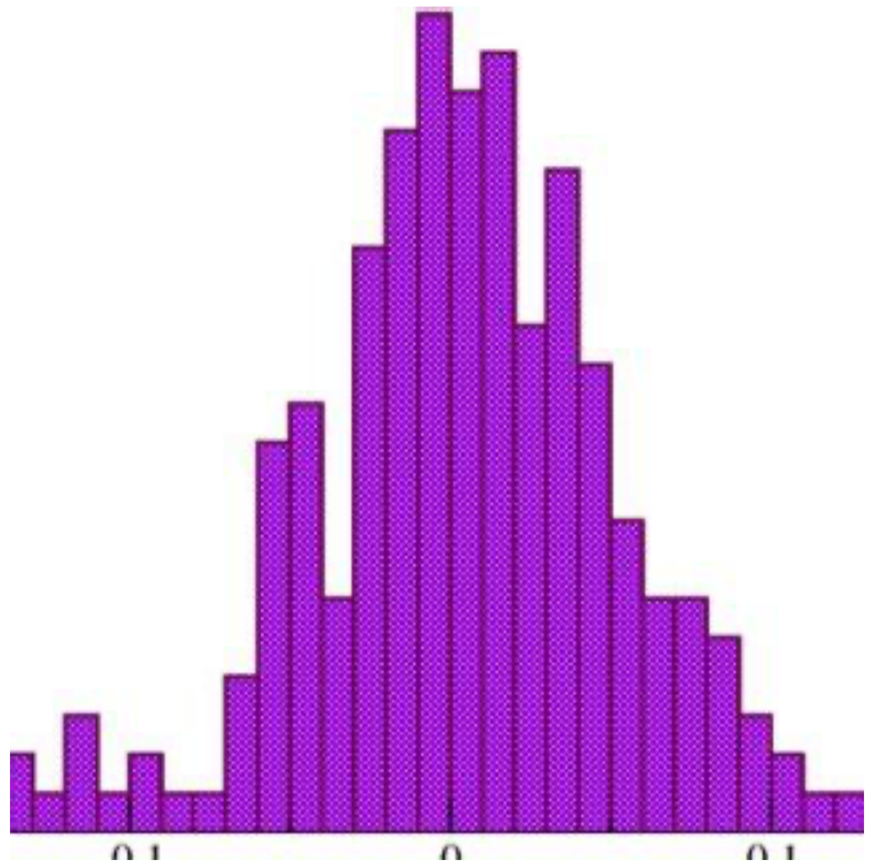
https://openreview.net/forum?id=HklBjCEKvH

Learning representations is easier than prediction

Dickens is the author of

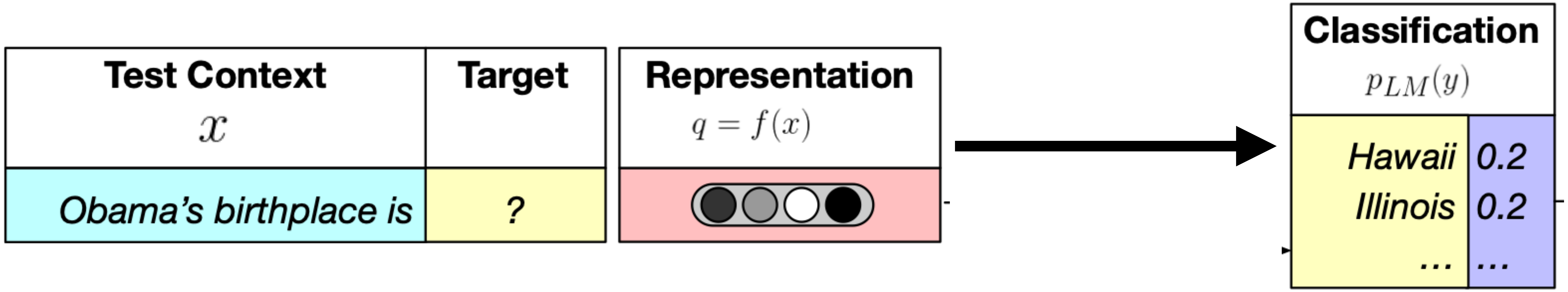
Dickens wrote

?



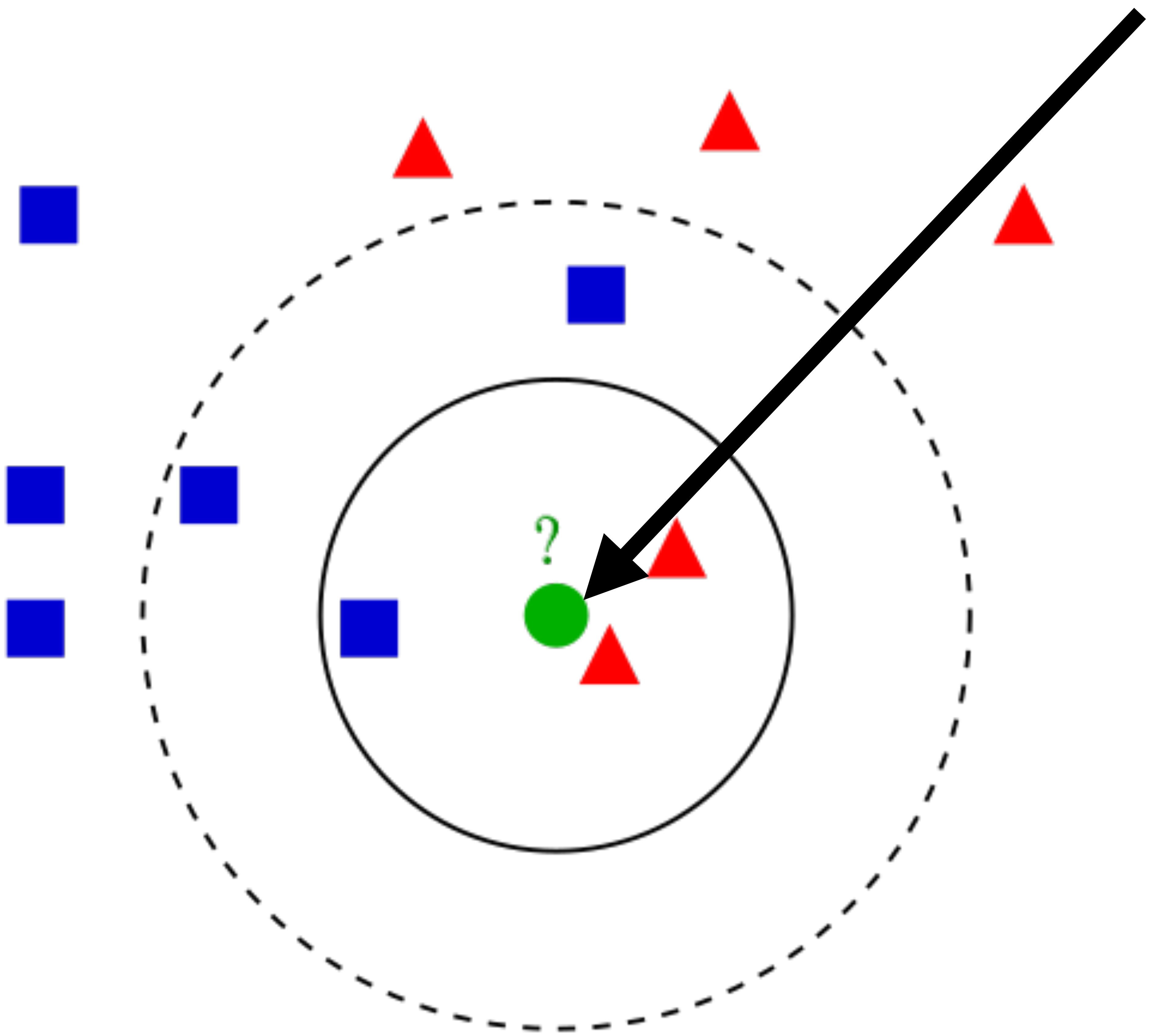
Even if you cannot predict the next token, you can predict that the distribution is identical over the vocabulary.

Standard LM prediction



Nearest Neighbour

k neighbours in vector space



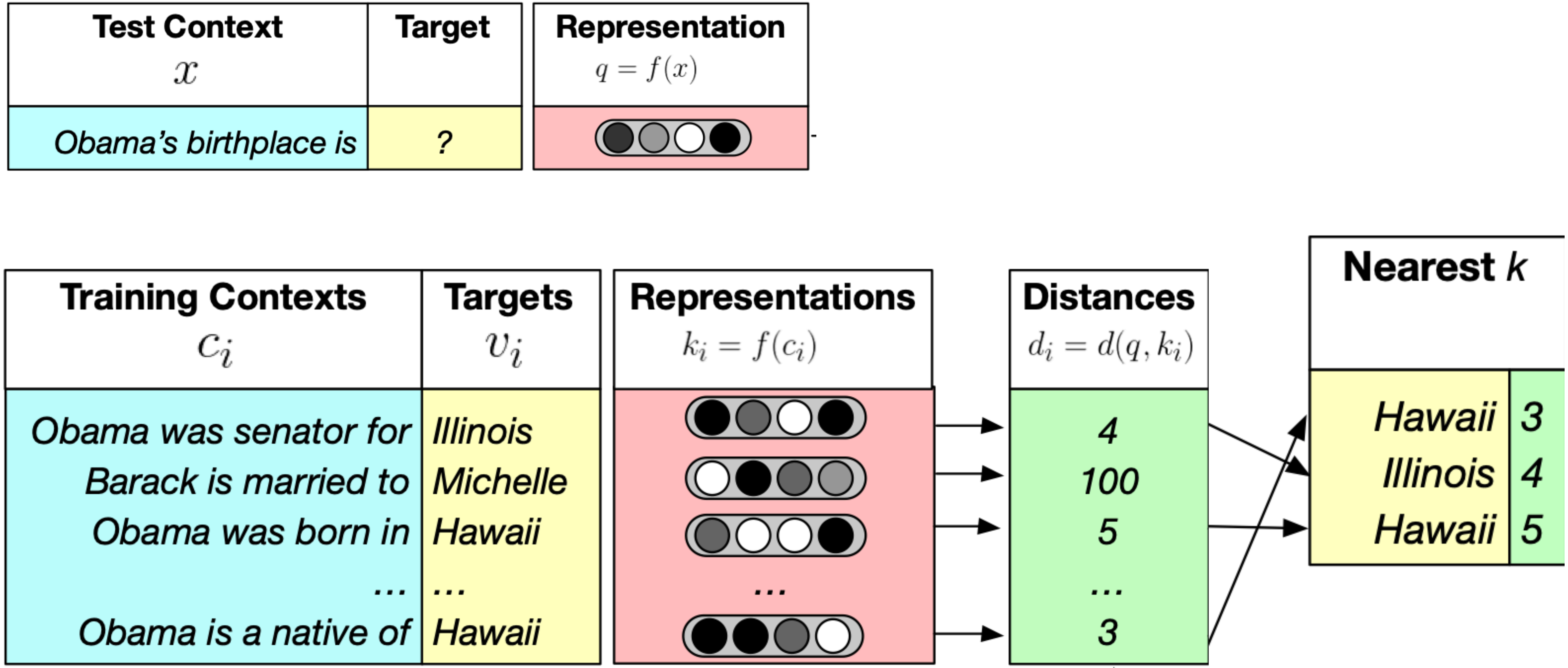
- The query vector is compared to other data vectors in the same vector space.
- Choose the class that has the most representatives inside the search perimeter.
- The search perimeter is determined by the cosine similarity of the query vector to the vectors stored in a kNN storage
- Efficient disk based kNN retrieval for very large sets of vectors is available: SCaNN, FAISS, annoy, etc.

kNN LM prediction (step 1)

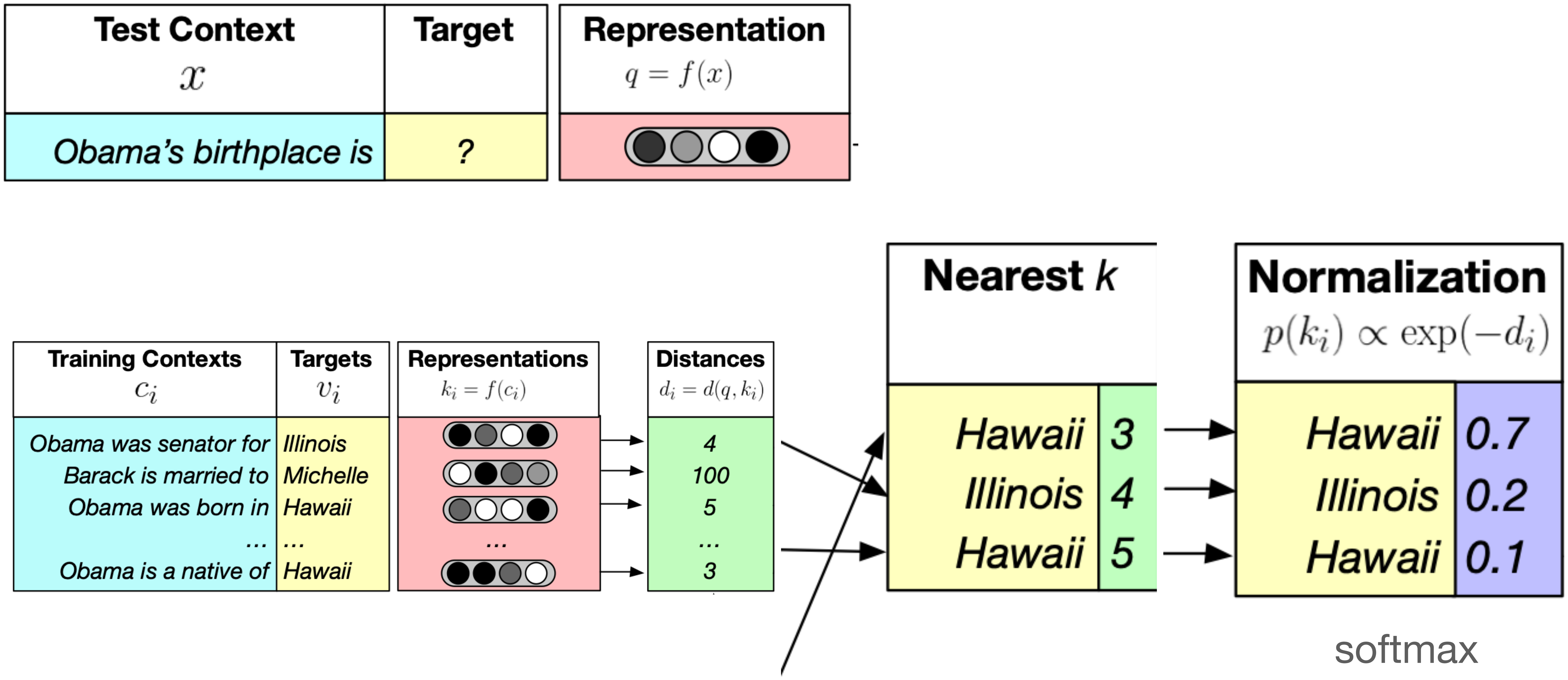
Test Context x	Target	Representation $q = f(x)$
Obama's birthplace is	*?*	

Training Contexts c_i	Targets v_i	Representations $k_i = f(c_i)$	Distances $d_i = d(q, k_i)$
Obama was senator for	*Illinois*		*4*
Barack is married to	*Michelle*		*100*
Obama was born in	*Hawaii*		*5*
...	...	...	...
Obama is a native of	*Hawaii*		*3*

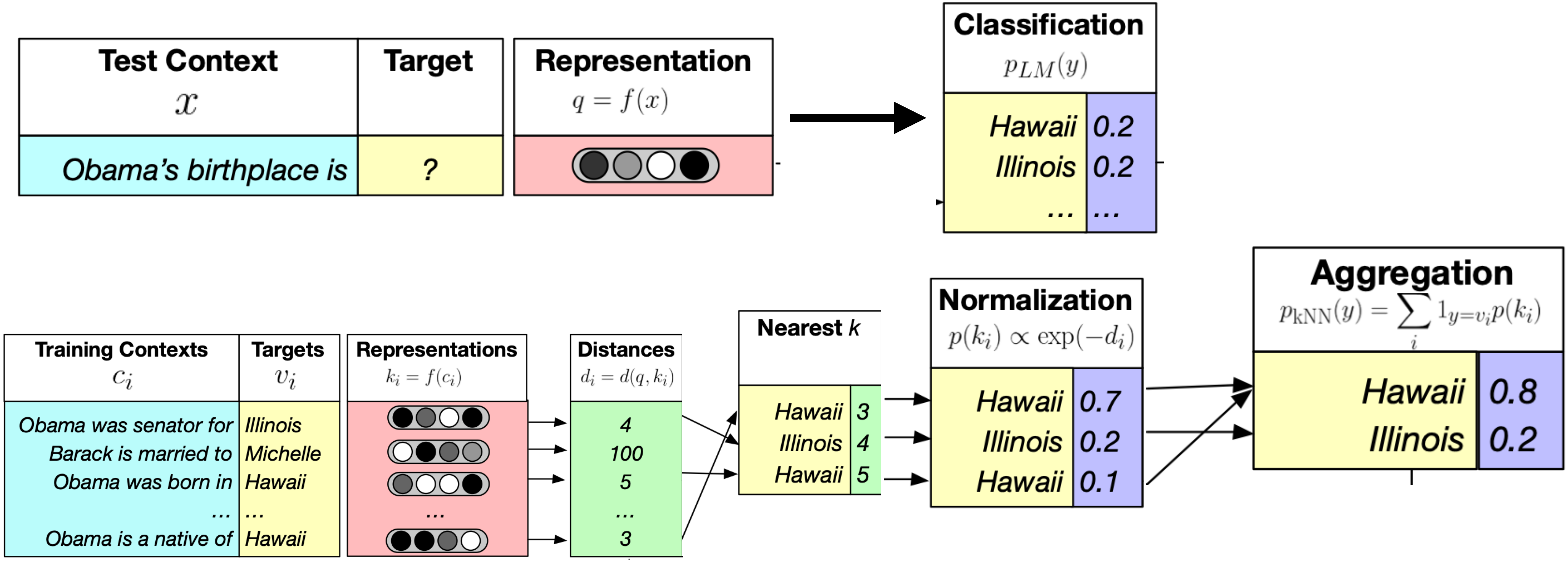
kNN LM prediction (step 2)



kNN LM prediction (step 3)



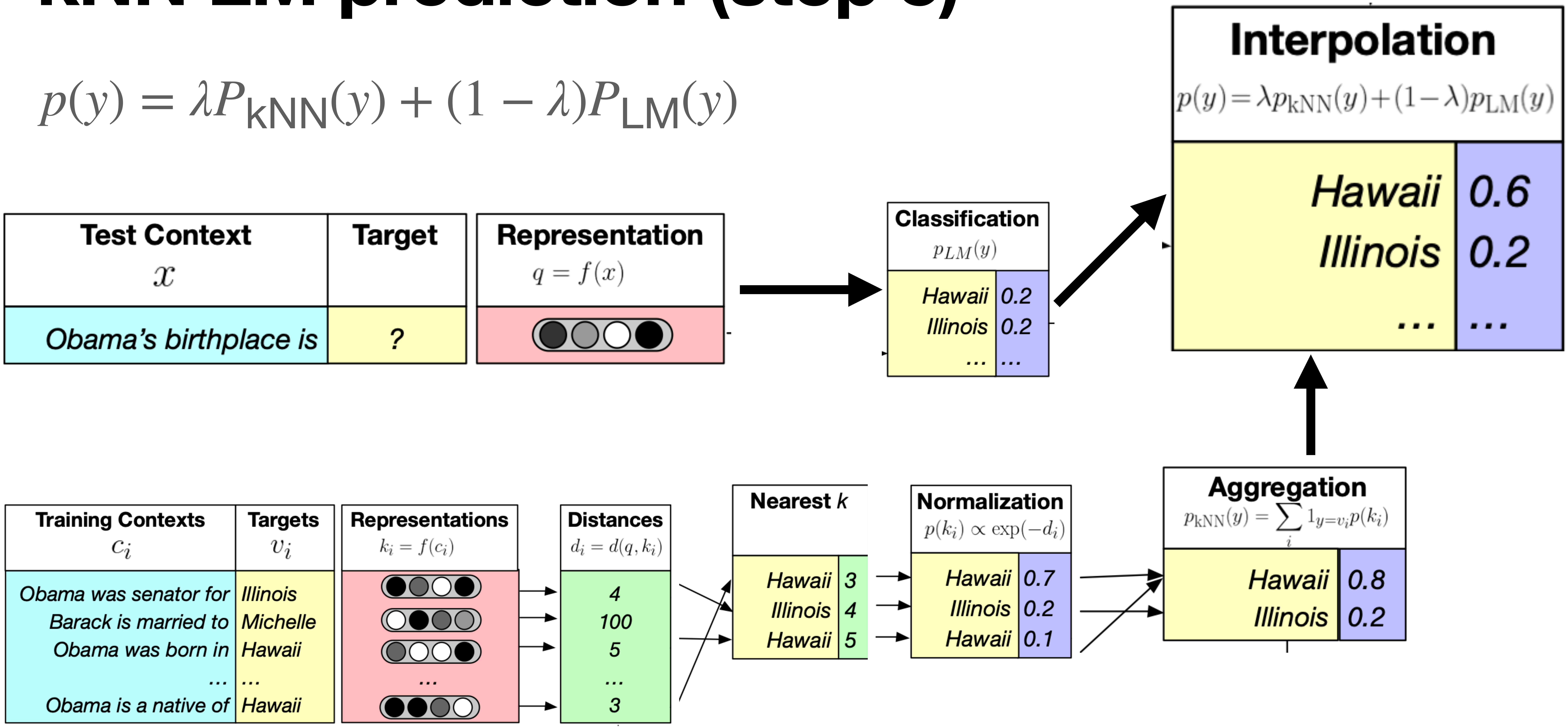
kNN LM prediction (step 4)



$$P_{\mathrm{kNN}}(y) \approx \sum_{k_i, v_i \in \mathcal{N}} 1_{y=v_i} \exp(-d(k_i, f(x)))$$

kNN LM prediction (step 5)

$$p(y) = \lambda P_{\mathrm{kNN}}(y) + (1 - \lambda) P_{\mathrm{LM}}(y)$$



Best representation for $f(c)$?

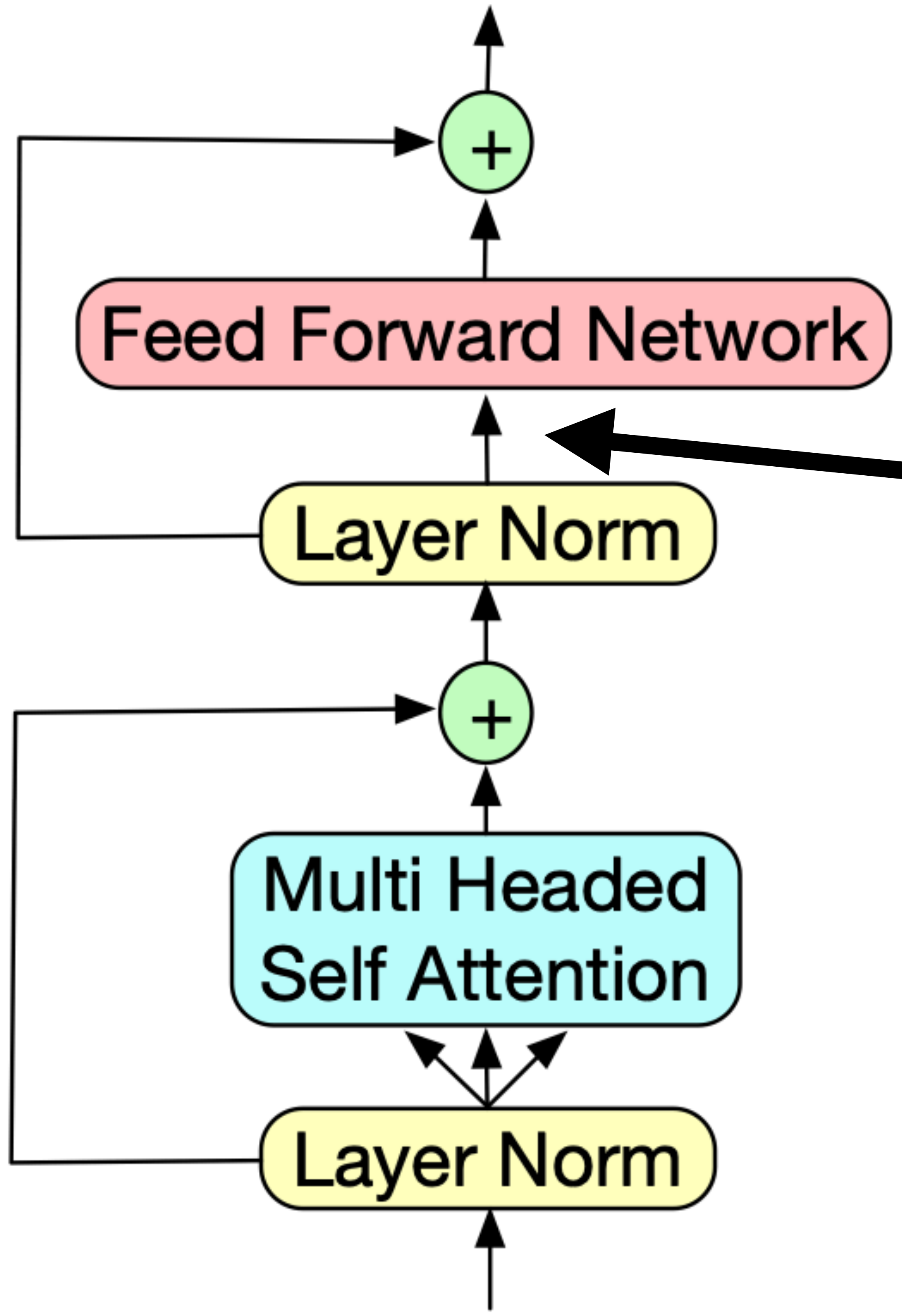


Figure 3: Transformer LM layer.

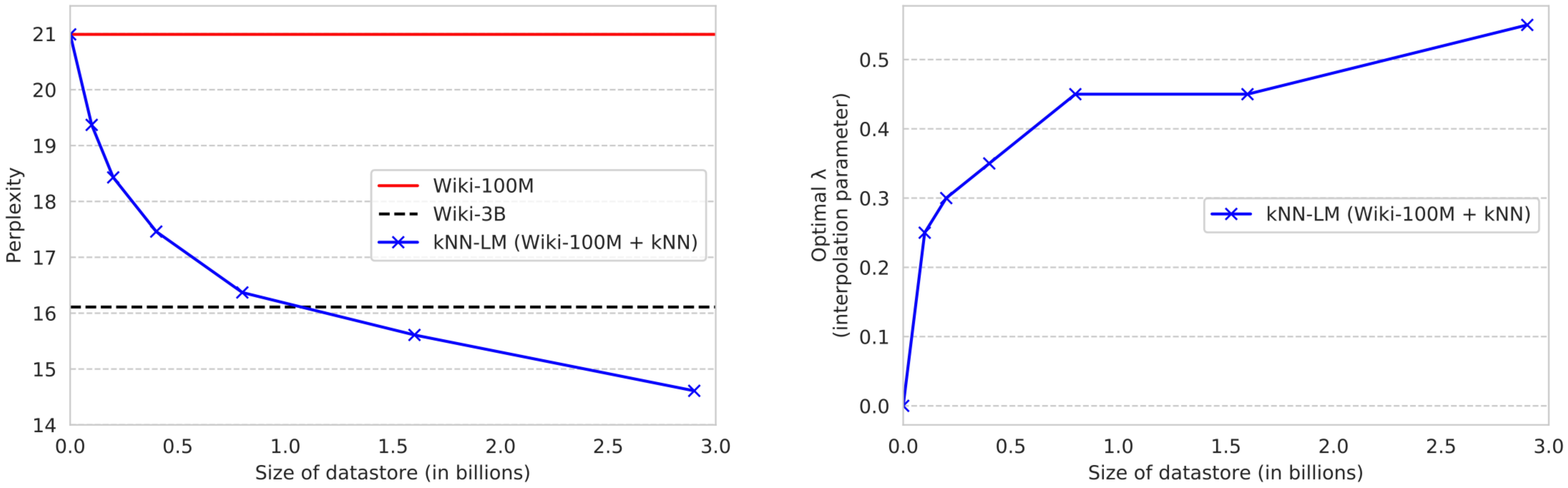
Key Type	Dev ppl. (↓)
No datastore	17.96
Model output	17.07
Model output layer normalized	17.01
FFN input after layer norm	**16.06**
FFN input before layer norm	17.06
MHSA input after layer norm	16.76
MHSA input before layer norm	17.14

Table 5: WIKITEXT-103 validation results using different states from the final layer of the LM as the representation function $f(\cdot)$ for keys and queries. We retrieve k=1024 neighbors and λ is tuned for each.

Output of FFN focuses on prediction; attention output focuses on representation

Model	Perplexity (↓) Dev	Test	# Trainable Params
Baevski & Auli (2019)	17.96	18.65	247M
+Transformer-XL (Dai et al., 2019)	-	18.30	257M
+Phrase Induction (Luo et al., 2019)	-	17.40	257M
Base LM (Baevski & Auli, 2019)	17.96	18.65	247M
+kNN-LM	**16.06**	**16.12**	247M
+Continuous Cache (Grave et al., 2017c)	17.67	18.27	247M
+kNN-LM + Continuous Cache	**15.81**	**15.79**	247M

Table 1: Performance on WIKITEXT-103. The kNN-LM substantially outperforms existing work. Gains are additive with the related but orthogonal continuous cache, allowing us to improve the base model by almost 3 perplexity points with no additional training. We report the median of three random seeds.



(a) Effect of datastore size on perplexities.

(b) Tuned values of λ for different datastore sizes.

Figure 2: Varying the size of the datastore. (a) Increasing the datastore size monotonically improves performance, and has not saturated even at about 3B tokens. A kNN-LM trained on 100M tokens with a datastore of 1.6B tokens already outperforms the LM trained on all 3B tokens. (b) The optimal value of λ increases with the size of the datastore.

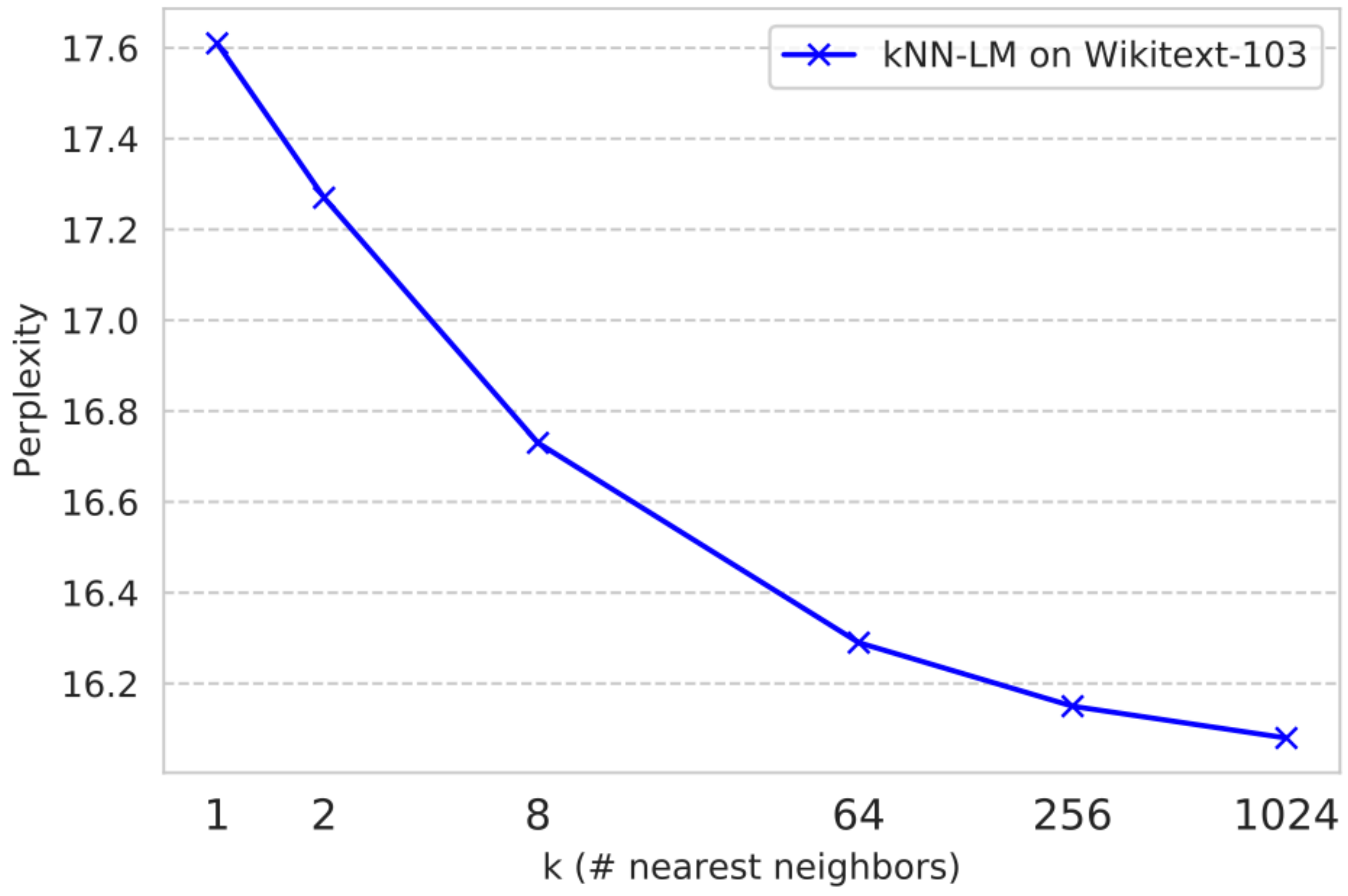


Figure 4: Effect of the number of nearest neighbors returned per word on WIKITEXT-103 (validation set). Returning more entries from the datastore monotonically improves performance.

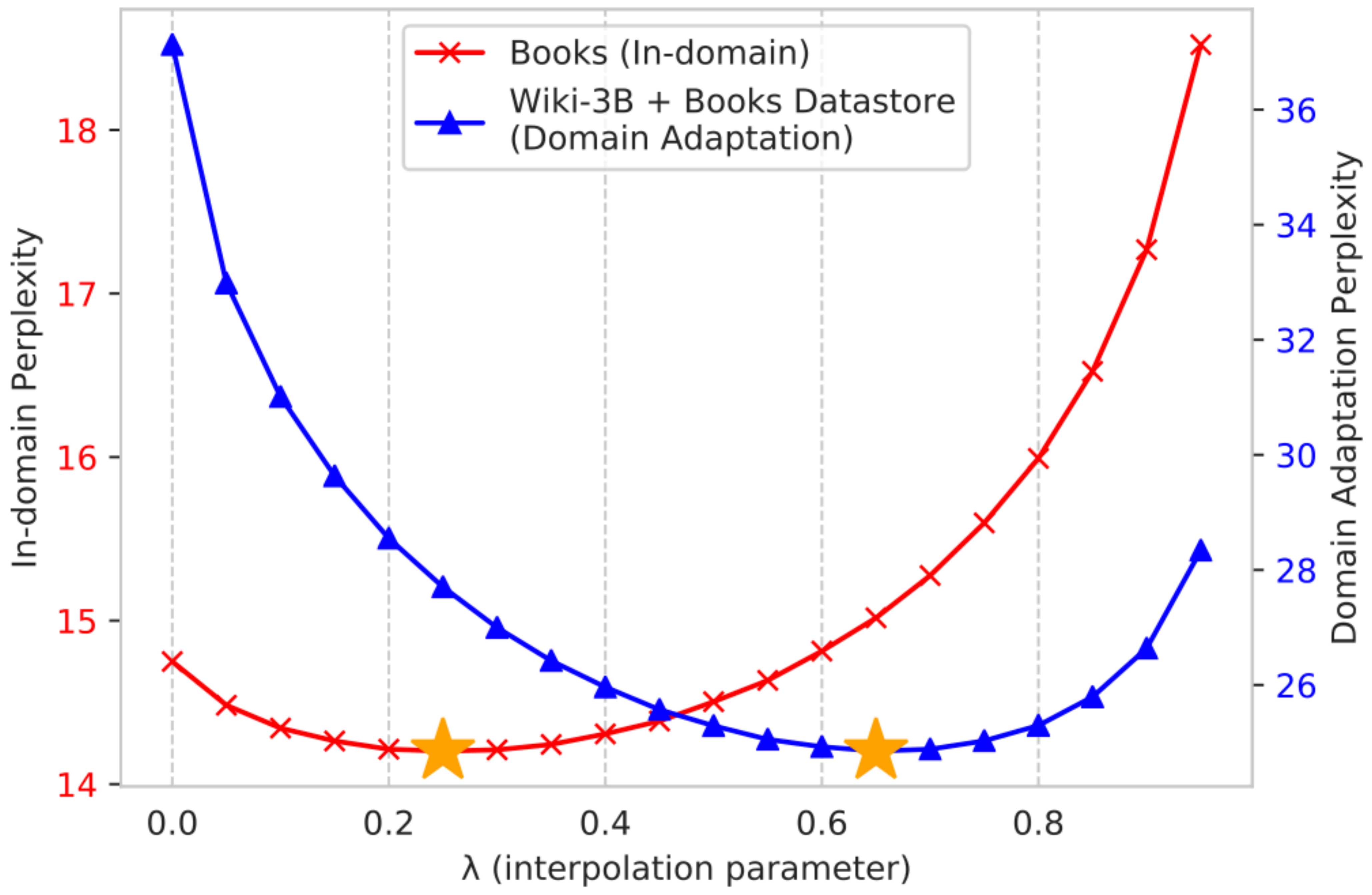


Figure 5: Effect of interpolation parameter λ on in-domain (left y-axis) and out-of-domain (right y-axis) validation set performances. More weight on p_{kNN} improves domain adaptation.

More data without training

Train LM on data; Store kNN on larger dataset

- Train LM on 100M token dataset, then run on 3B token dataset to store context vectors in kNN store
- Use kNN-LM to predict next token
- Surprisingly kNN-LM (100M + 3B) does better than LM trained on 3B token dataset
- "retrieving nearest neighbors from the corpus outperforms training on it"
- "rather than training language models on ever larger datasets, we can use smaller datasets to learn representations and augment them with kNN-LM over a large corpus"

Test Context $(p_{\mathrm{kNN}} = 0.998, p_{\mathrm{LM}} = 0.124)$	**Test Target**	
it was organised by New Zealand international player Joseph Warbrick, promoted by civil servant Thomas Eyton, and managed by James Scott, a publican. The Natives were the first New Zealand team to perform a haka, and also the first to wear all black. They played 107 rugby matches during the tour, as well as a small number of Victorian Rules football and association football matches in Australia. Having made a significant impact on the...	development	

Training Set Context	**Training Set Target**	**Context Probability**
As the captain and instigator of the 1888-89 Natives – the first New Zealand team to tour the British Isles – Warbrick had a lasting impact on the...	development	0.998
promoted to a new first grade competition which started in 1900. Glebe immediately made a big impact on the...	district	0.00012
centuries, few were as large as other players managed. However, others contend that his impact on the...	game	0.000034
Nearly every game in the main series has either an anime or manga adaptation, or both. The series has had a significant impact on the...	development	0.00000092

Test Context ($p_{\mathrm{kNN}} = 0.995$, $p_{\mathrm{LM}} = 0.025$)	**Test Target**	
For Australians and New Zealanders the Gallipoli campaign came to symbolise an important milestone in the emergence of both nations as independent actors on the world stage and the development of a sense of national identity. Today, the date of the initial landings, 25 April, is known as Anzac Day in Australia and New Zealand and every year thousands of people gather at memorials in both nations, as well as Turkey, to...	honour	
Training Set Context	**Training Set Target**	**Context Probability**
Despite this, for Australians and New Zealanders the Gallipoli campaign has come to symbolise an important milestone in the emergence of both nations as independent actors on the world stage and the development of a sense of national identity. Today, the date of the initial landings, 25 April, is a public holiday known as Anzac Day in Australia and New Zealand and every year thousands of people gather at memorials in both nations, and indeed in Turkey, to ...	honour	0.995
On the anniversary date of his death, every year since 1997, thousands of people gather at his home in Memphis to...	celebrate	0.0086
Twenty-five years after Marseille's death, fighter pilot veterans of World War II gathered to...	honour	0.0000041

Test Context $(p_{\text{kNN}} = 0.959, p_{\text{LM}} = 0.503)$	**Test Target**
U2 do what they're best at, slipping into epic rock mode, playing music made for the arena". In two other local newspaper reviews, critics praised the song's inclusion in a sequence of greatest hits. For the PopMart Tour of 1997–...	1998

Training Set Context	**Training Set Target**	**Context Probability**
Following their original intent, "Sunday Bloody Sunday" was not played during any of the forty-seven shows on the Lovetown Tour in 1989. The song reappeared for a brief period during the Zoo TV Tour, and late during the second half of PopMart Tour (1997–...	1998	0.936
They are 6 times Champions and they won the Challenge Cup in 1938, and have experienced two previous stretches in the Super League, 1997–...	2002	0.0071
About $40 million ($61.4 million in 2018 dollars) was spent on the property acquisition. After weather-related construction delays due to the El Nino season of the winter of 1997–...	1998	0.0015
This made it the highest-rated season of The X-Files to air as well as the highest rated Fox program for the 1997–...	98	0.00000048

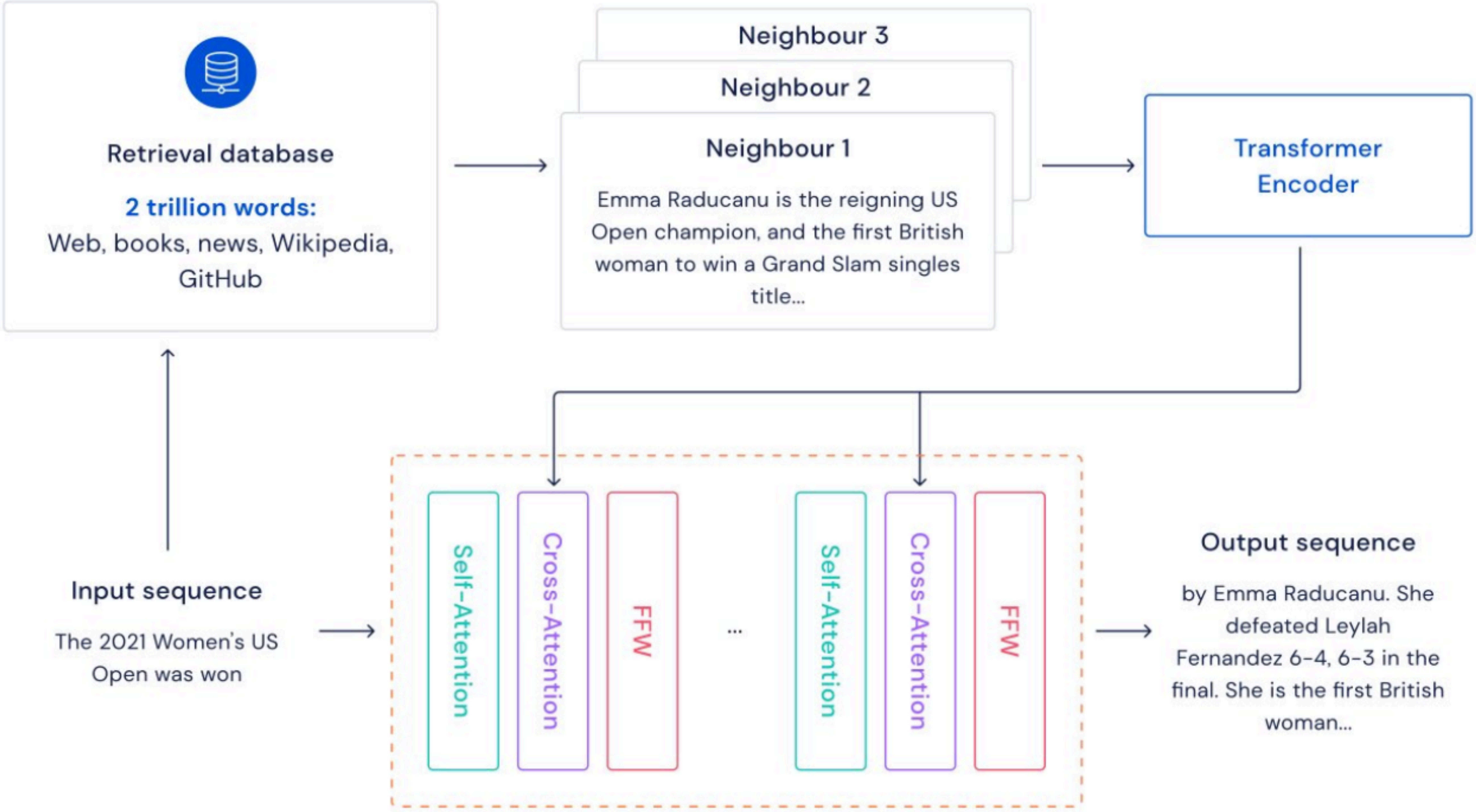
RETRO

Improving language models by retrieving from trillions of tokens

Sebastian Borgeaud[†], Arthur Mensch[†], Jordan Hoffmann[†], Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae[‡], Erich Elsen[‡] and Laurent Sifre[†,‡]

All authors from DeepMind, [†]Equal contributions, [‡]Equal senior authorship

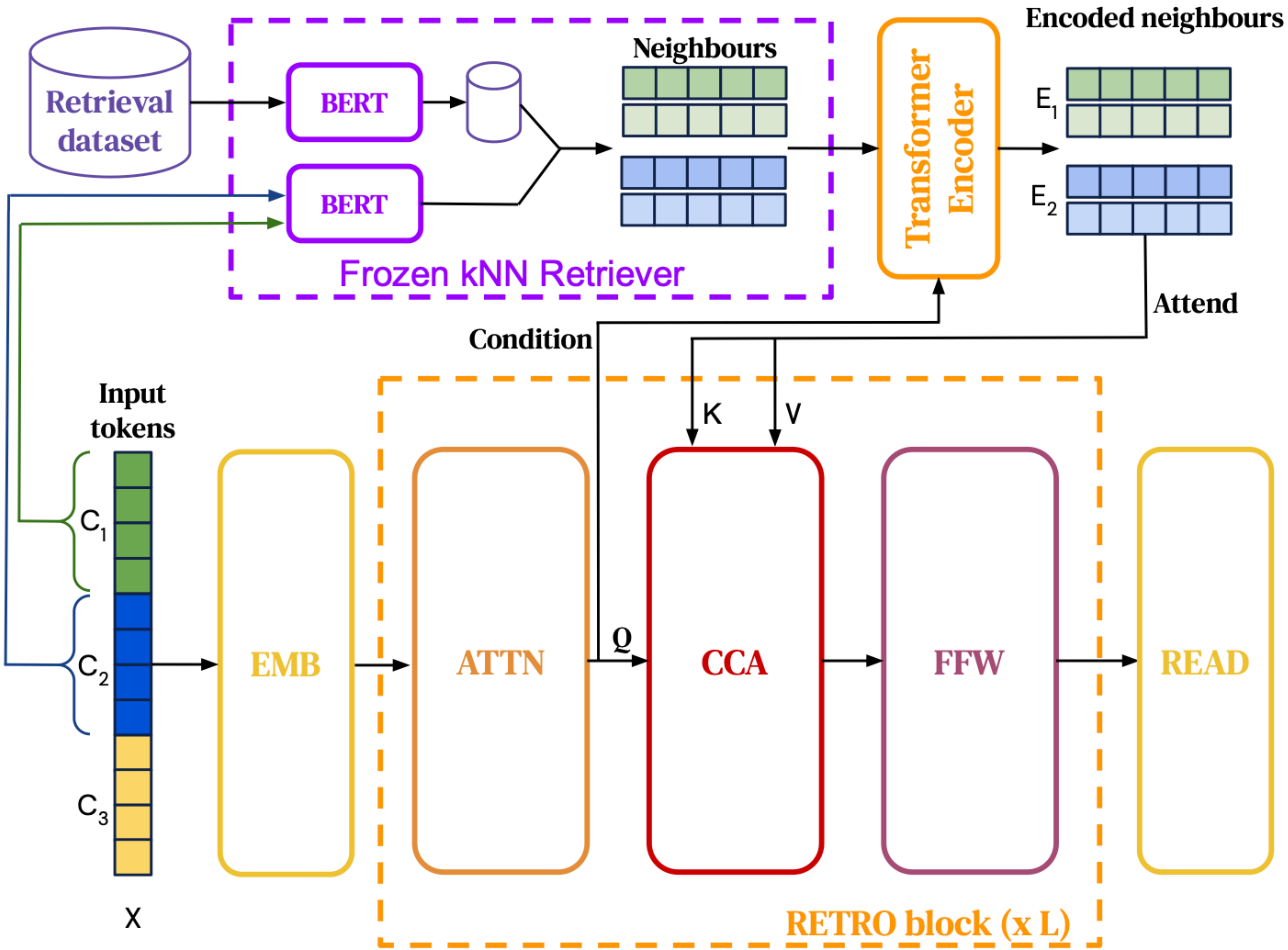
https://arxiv.org/abs/2112.04426



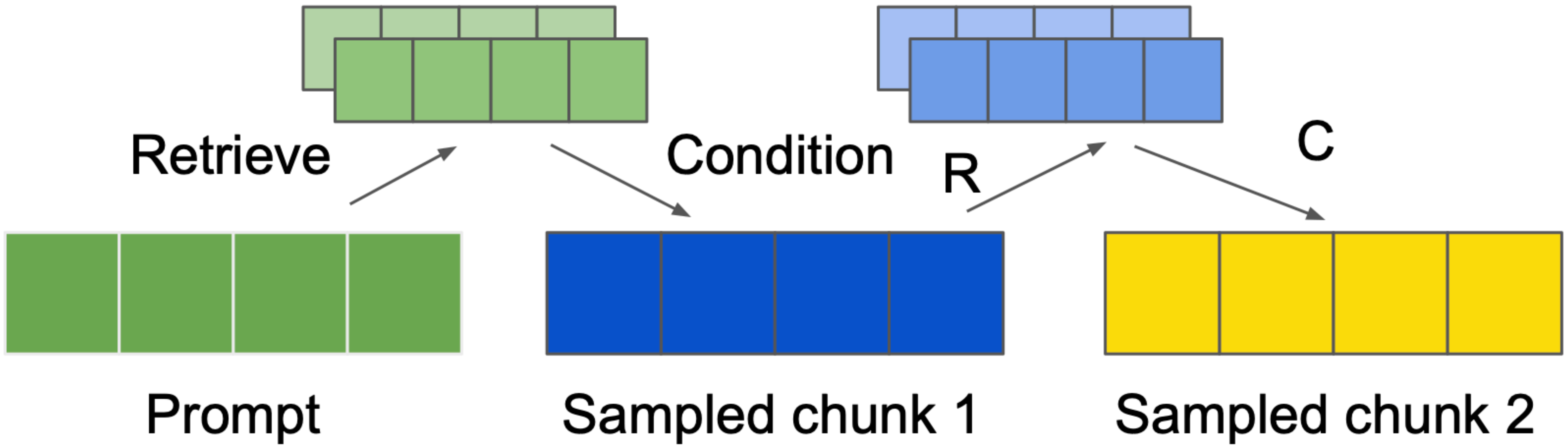
https://www.deepmind.com/blog/improving-language-models-by-retrieving-from-trillions-of-tokens

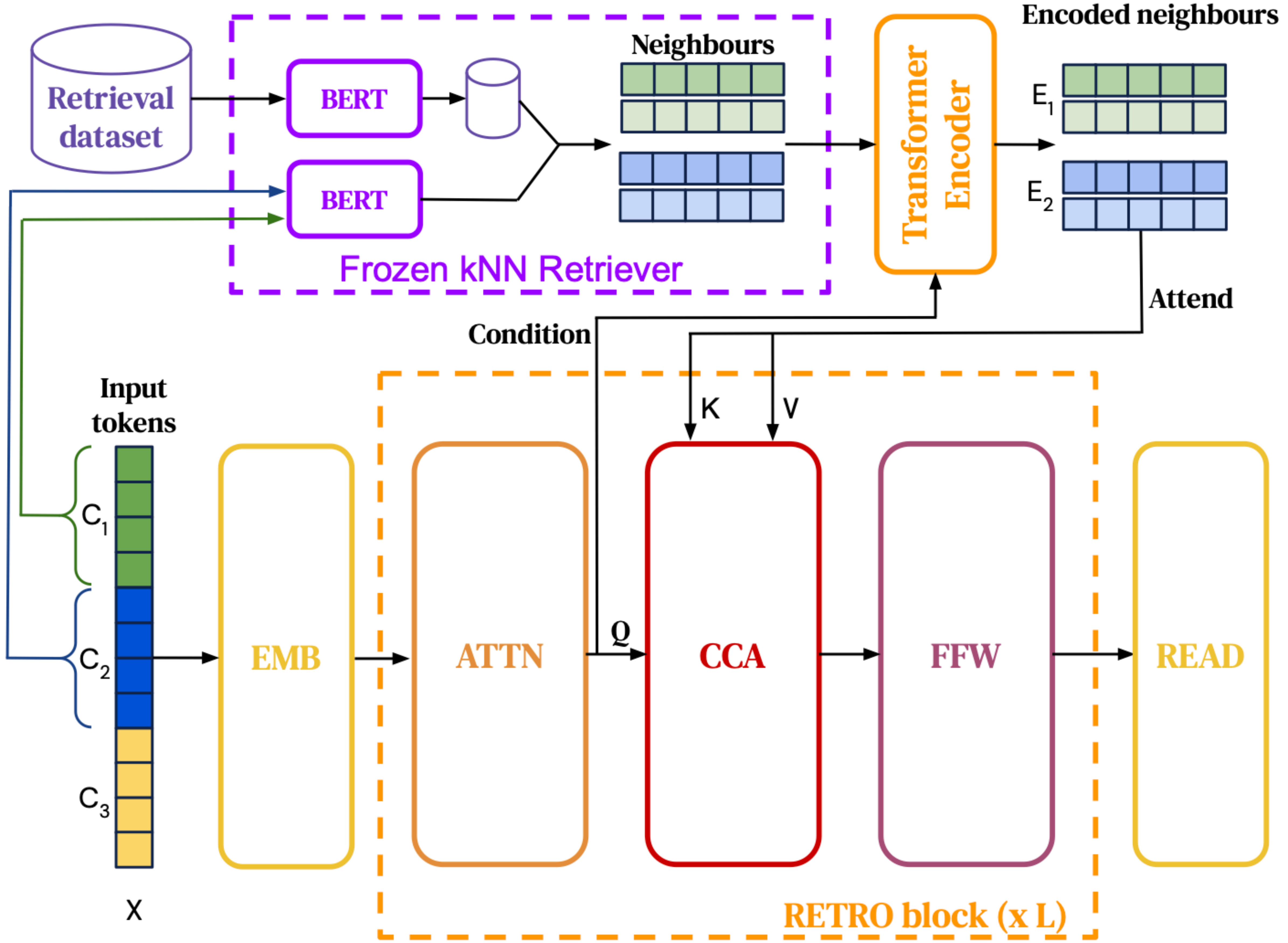
Table 3 | **Comparison of Retro with existing retrieval approaches.**

	# Retrieval tokens	Granularity	Retriever training	Retrieval integration
Continuous Cache	$\mathcal{O}(10^3)$	Token	Frozen (LSTM)	Add to probs
*k*NN-LM	$\mathcal{O}(10^9)$	Token	Frozen (Transformer)	Add to probs
Spalm	$\mathcal{O}(10^9)$	Token	Frozen (Transformer)	Gated logits
Dpr	$\mathcal{O}(10^9)$	Prompt	Contrastive proxy	Extractive QA
Realm	$\mathcal{O}(10^9)$	Prompt	End-to-End	Prepend to prompt
RAG	$\mathcal{O}(10^9)$	Prompt	Fine-tuned Dpr	Cross-attention
FiD	$\mathcal{O}(10^9)$	Prompt	Frozen Dpr	Cross-attention
Emdr^2	$\mathcal{O}(10^9)$	Prompt	End-to-End (EM)	Cross-attention
Retro (ours)	$\mathcal{O}(\mathbf{10^{12}})$	**Chunk**	**Frozen (Bert)**	**Chunked cross-attention**

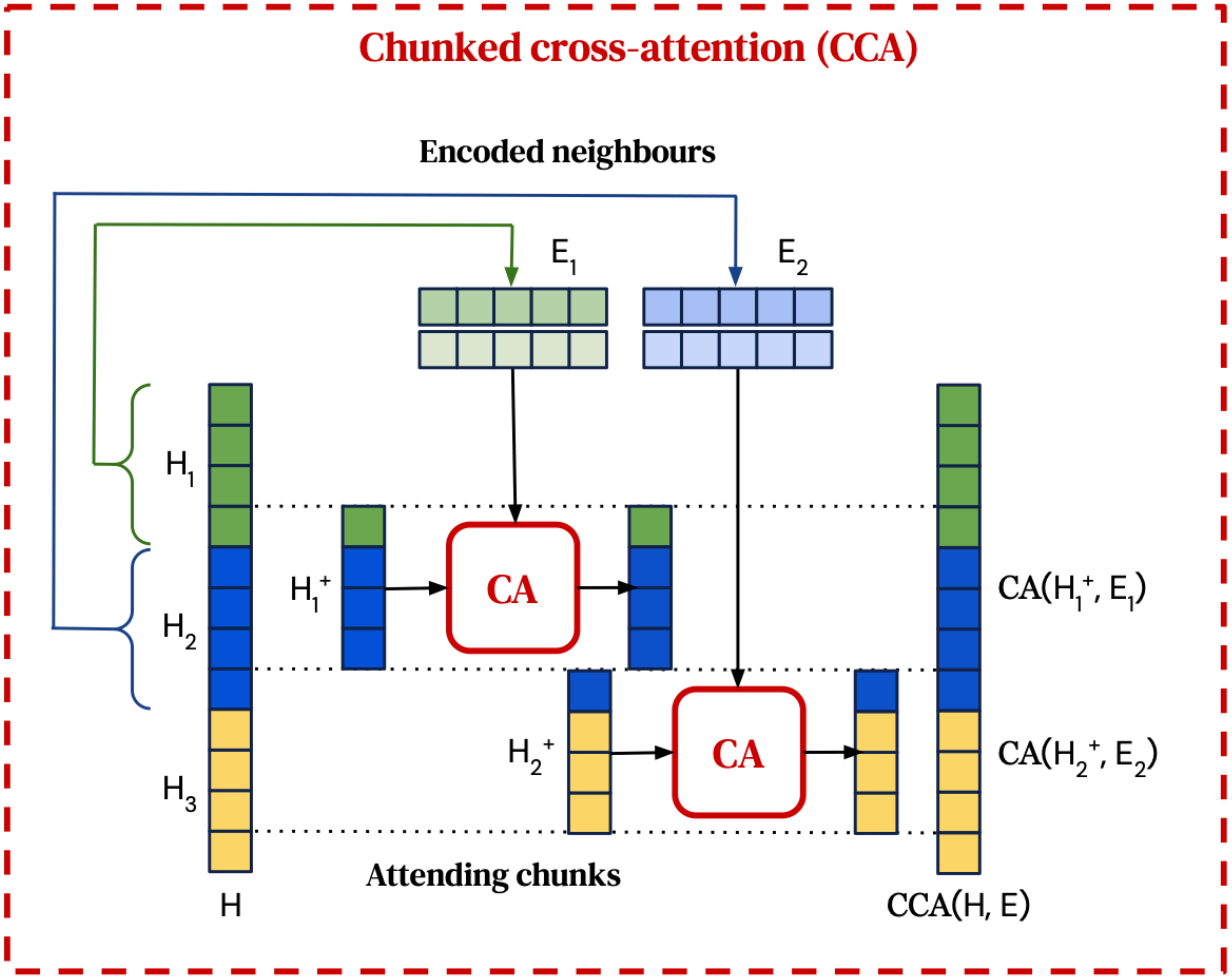


Retrieval dataset
BERT
BERT
Neighbours
Frozen kNN Retriever
Transformer Encoder
Encoded neighbours
E1
E2
Attend
Condition
Input tokens
C1
C2
C3
X
EMB
ATTN
Q
K
V
CCA
FFW
READ
RETRO block (x L)





https://icml.cc/media/icml-2022/Slides/17480_uuemO20.pdf



Chunked cross-attention (CCA)
Encoded neighbours
E_1
E_2
H_1
H_2
H_3
H_1^+
H_2^+
CA
CA
$CA(H_1^+, E_1)$
$CA(H_2^+, E_2)$
Attending chunks
H
CCA(H, E)

Language modelling: The Pile

https://icml.cc/media/icml-2022/Slides/17480_uuemO20.pdf

Gao et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling

RETRO Takeaways

- RETRO is a general architecture, that is fully autoregressive and enables large scale retrieval
- Adding a 2T token database yields a performance improvement that's constant with model size: Similar performance to models with 10x more parameters on the Pile
- Consistent performance across benchmarks
    - Retrieval does exploit train-test leakage more than standard language models
    - But performance also improves on held-out tokens
- Future work on few-shot evaluation

https://icml.cc/media/icml-2022/Slides/17480_uuemO20.pdf