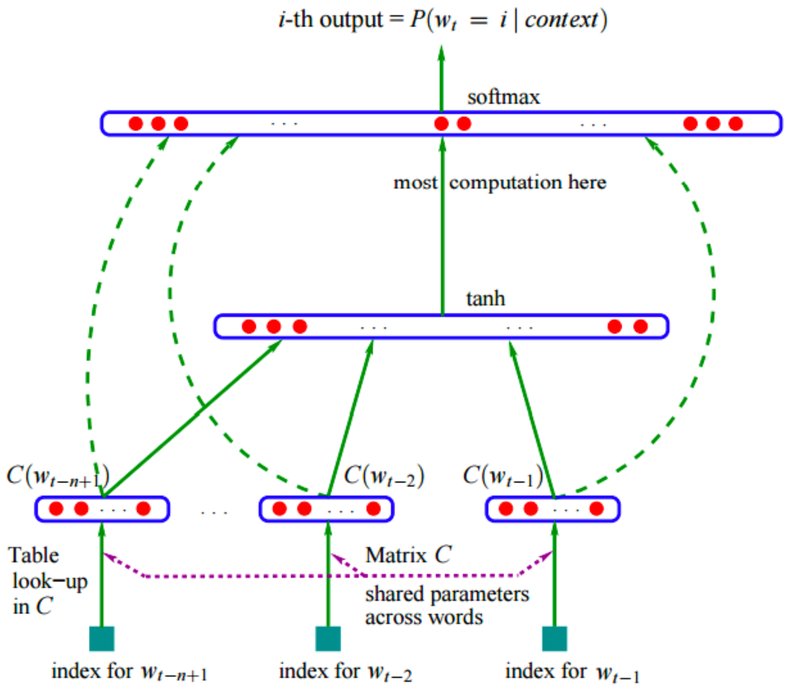
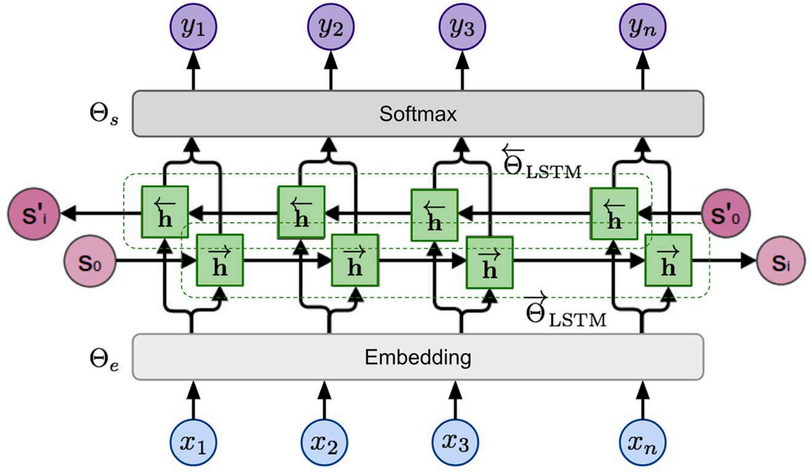
Building Semi-Parametric Language Models to Integrate World Knowledge

Speaker: Wenhu Chen
Acknolwedgement: Michel de Jong, Pat Verga, William Cohen

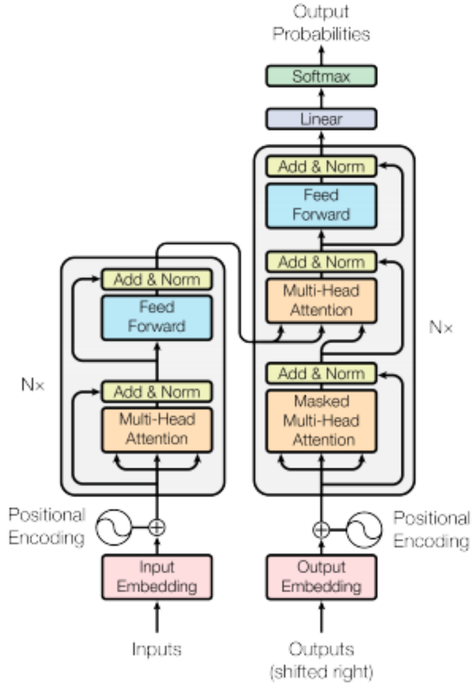
Language Models



FeedForward Language Model (Bengio et al. 2003)



LSTM Language Model (Sutskever et al. 2014)



Attention Language Model (Vaswani et al. 2017)

Why is Language Model so Hard?

Micheal Jordan is an American businessman and former professional basketball player.
… winning six championships with ???

Chicago Bulls

Syntactic Knowledge

```
(ROOT
 (S
  (NP (NNP Micheal) (NNP Jordan))
  (VP (VBZ is)
   (NP (DT an)
    (NML
     (NML (JJ American) (NN businessman))
     (CC and)
     (NML (JJ former) (JJ professional) (NN basketball)))
    (NN player)))))
```

Language Model is beyond Language

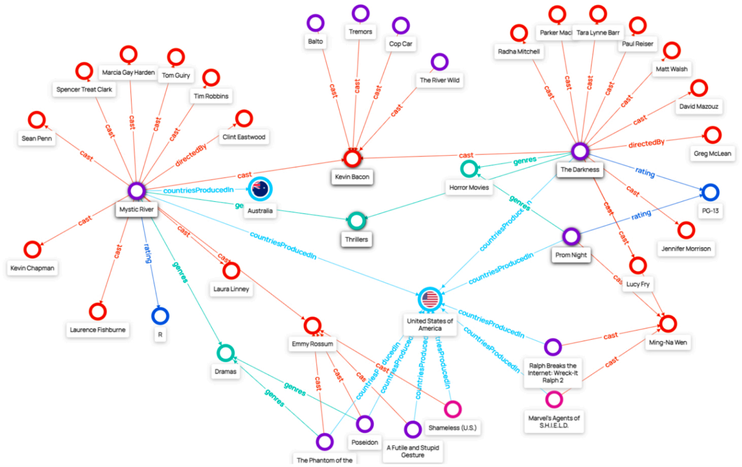
Micheal Jordan is an American businessman and former professional basketball player.
… winning six championships with ???

Chicago Bulls

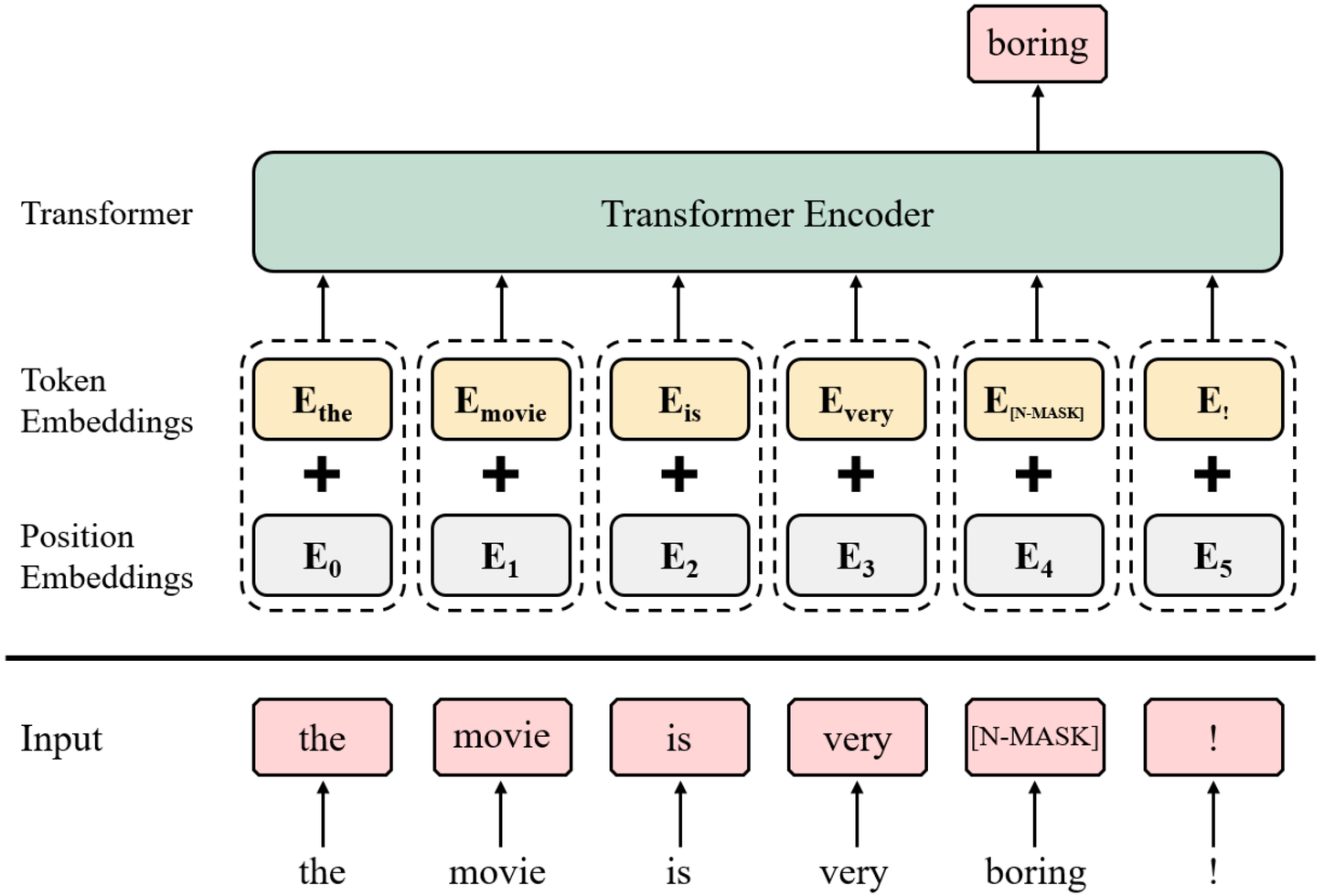
Syntactic Knowledge

```
(ROOT
 (S
  (NP (NNP Micheal) (NNP Jordan))
  (VP (VBZ is)
   (NP (DT an)
    (NML
     (NML (JJ American) (NN businessman))
     (CC and)
     (NML (JJ former) (JJ professional) (NN basketball)))
    (NN player)))))
```

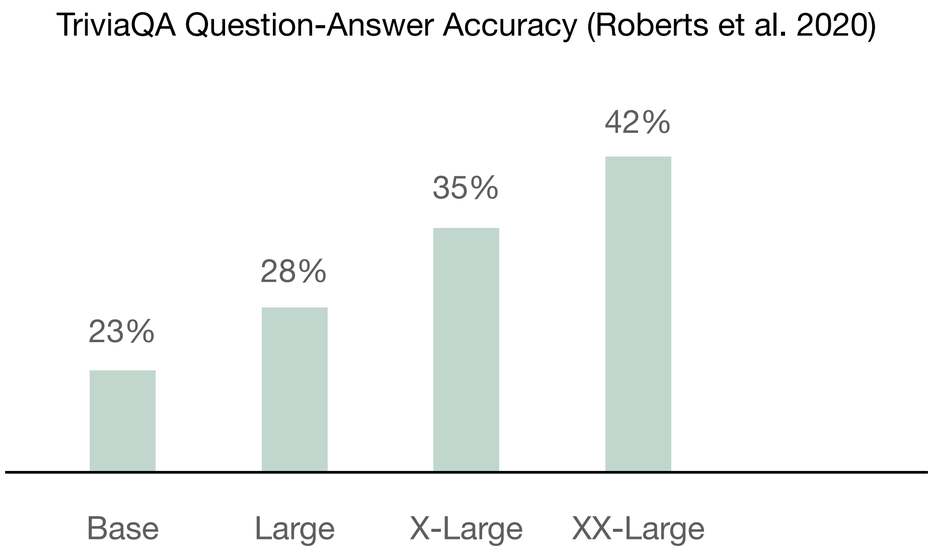
Entity Knowledge



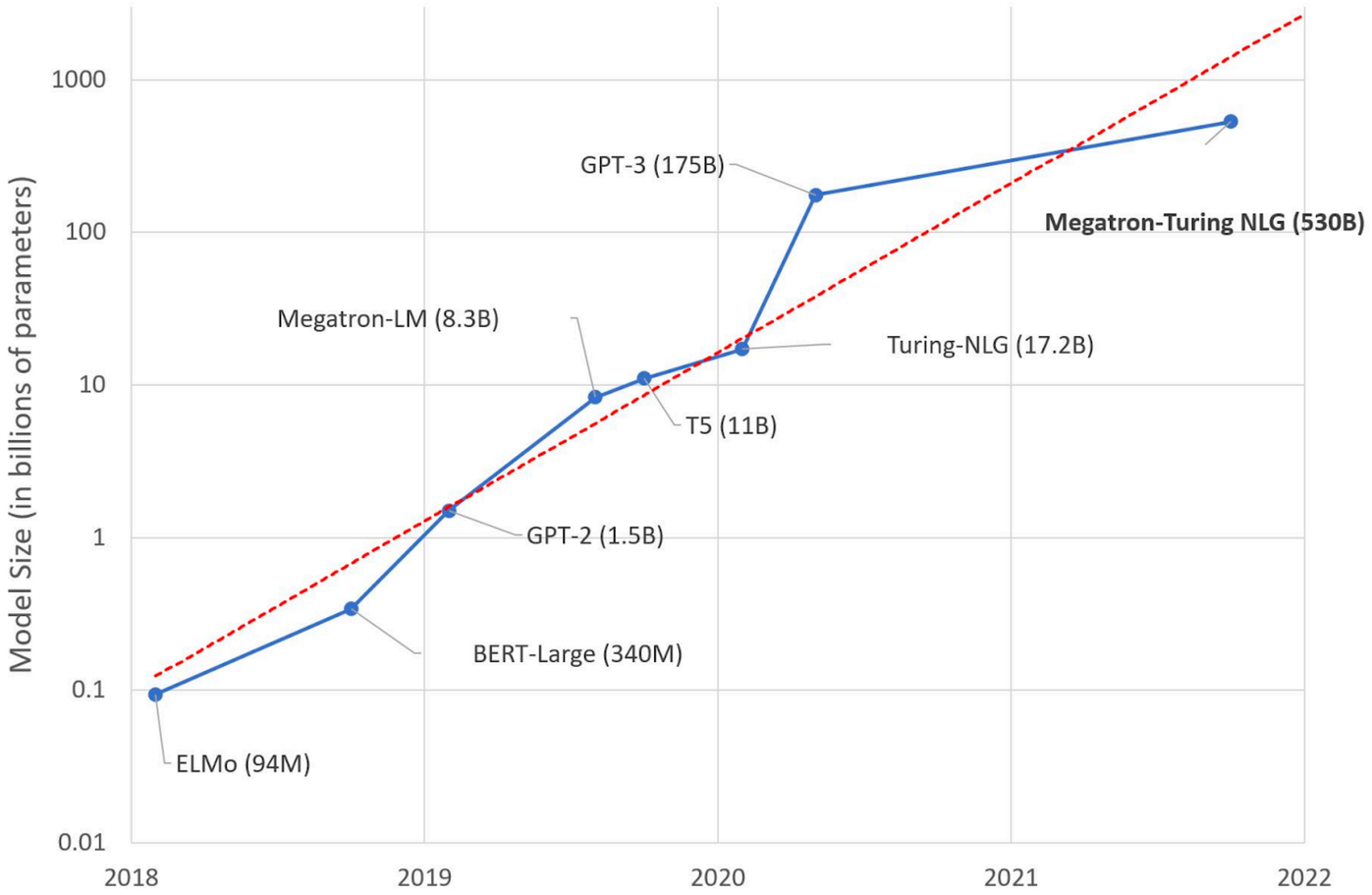
Pre-training Language Model



Larger Model, Larger Capacity



New Moore's Law

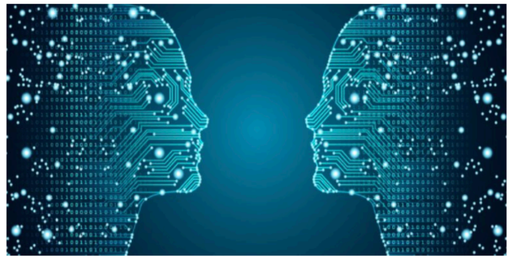


Drawbacks of Large Language Models

March 15, 2021

Experts Disagree on the Utility of Large Language Models

Alex Woodie



(Ryzhi/Shutterstock)

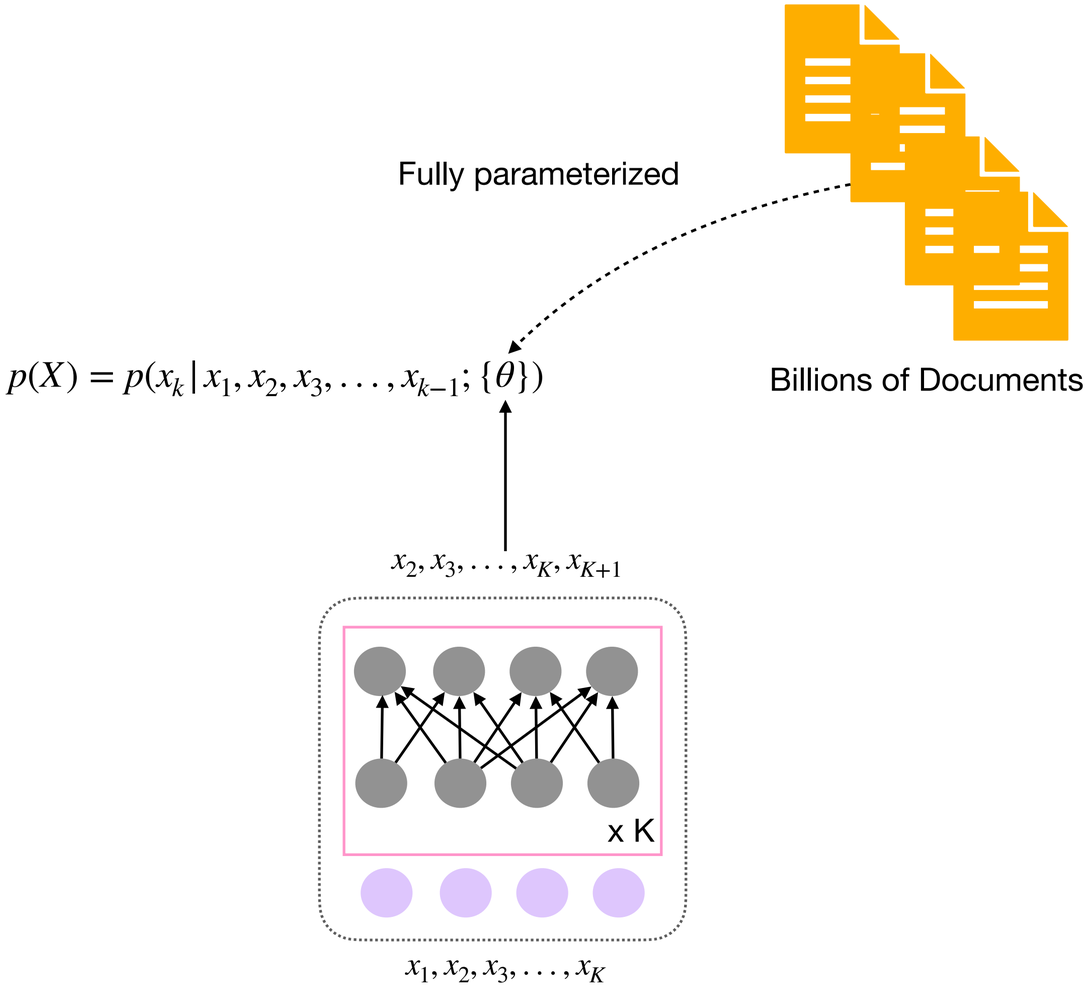
Large language models like OpenAI's GPT-3 and Google Brain's Switch Transformer have caught the eye of AI experts, who have expressed surprise at the rapid pace of improvement. However, not everybody is jumping onto the bandwagon, and others see significant limitations in the new technology, as well as ethical implications.

There's a veritable arms race occurring in the world of natural language processing (NLP), as large neural networks continually raise the bar on computer understanding of the written word. Over a matter of months, as OpenAI released GPT-3 last summer to Google Brain's launch of Switch Transformer earlier this year, the number of parameters in these specialized transformer networks has gone from the hundreds of millions into the trillions.

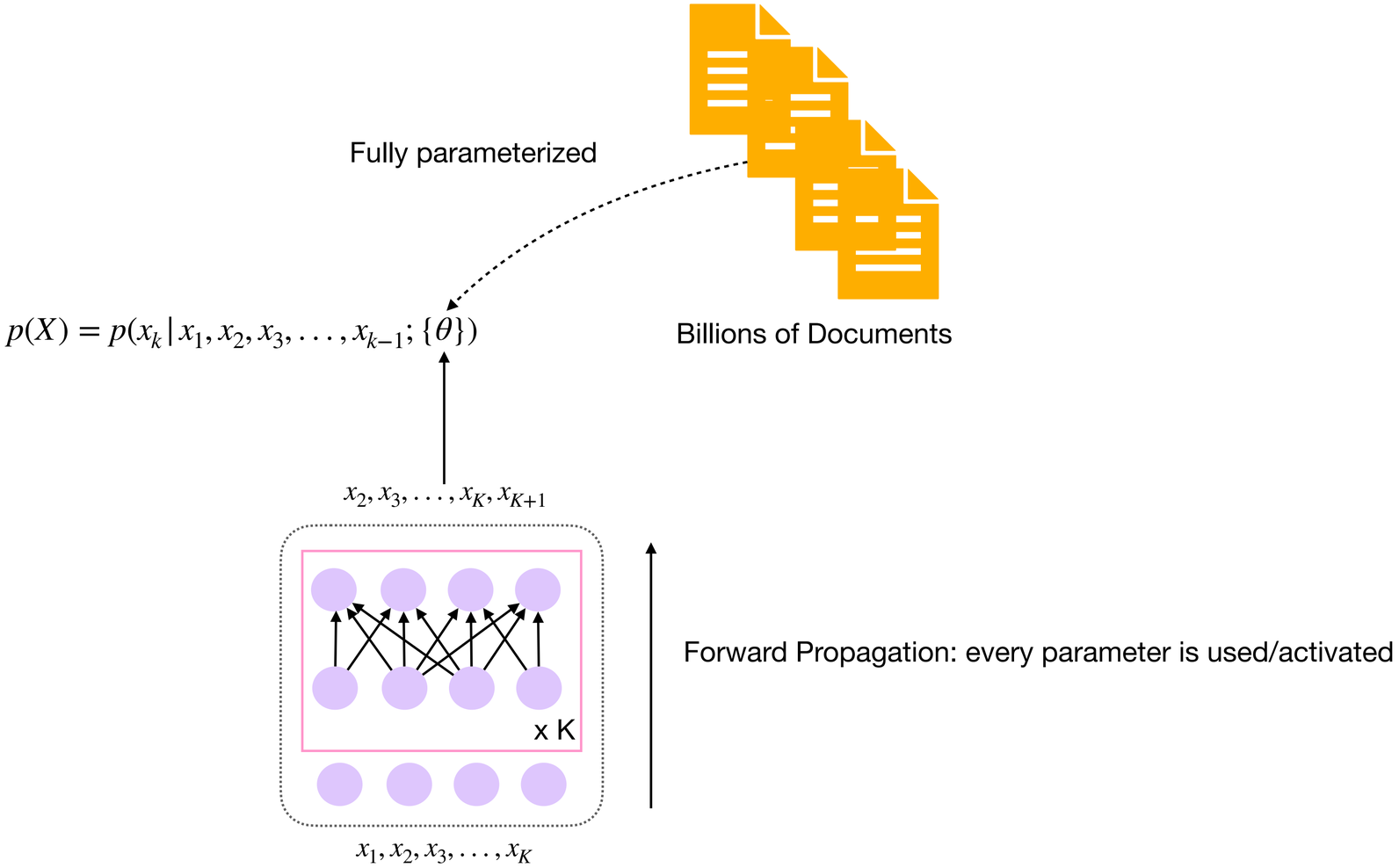
Trained upon huge corpuses of words, these language models have displayed marked improvement in the capability to understand how words go together to express higher-level ideas. Simultaneously, the generative models have also shown greater skill in outputting text, to the point where it can be difficult to tell whether a chunk of prose (or a length of verse) was written by man or machine.

Computation Cost ↑ Memory Cost ↑ Carbon Footprint ↑ Interpretability ↓ Portability ↓

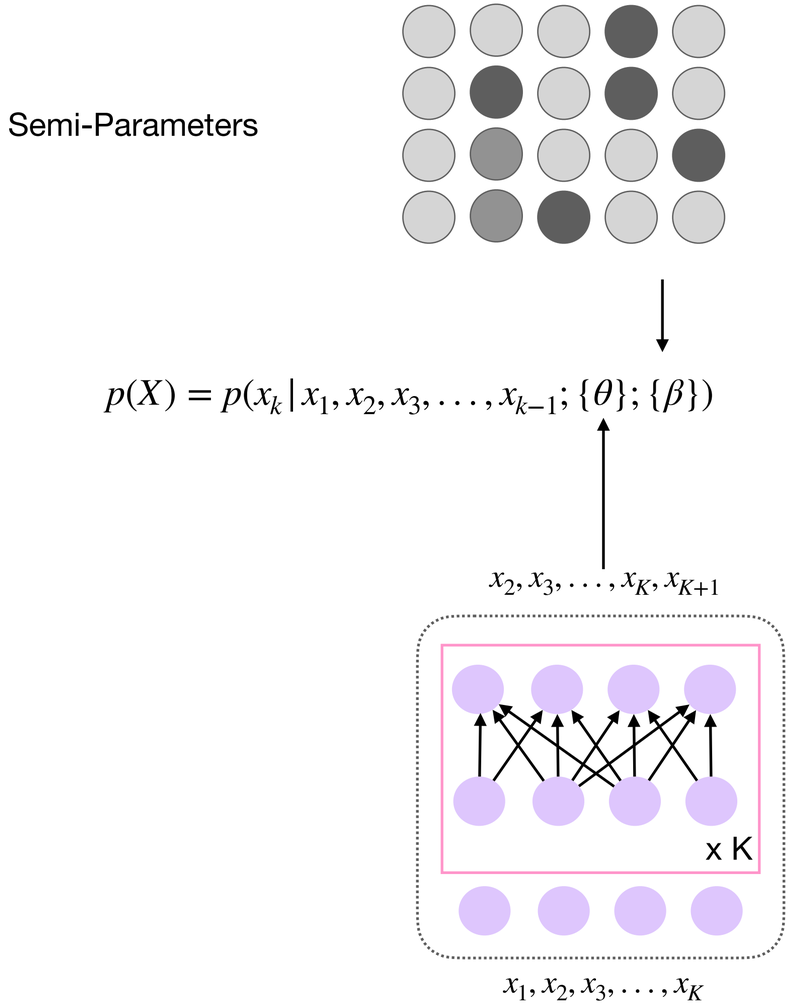
Fully-Parametric Language Model



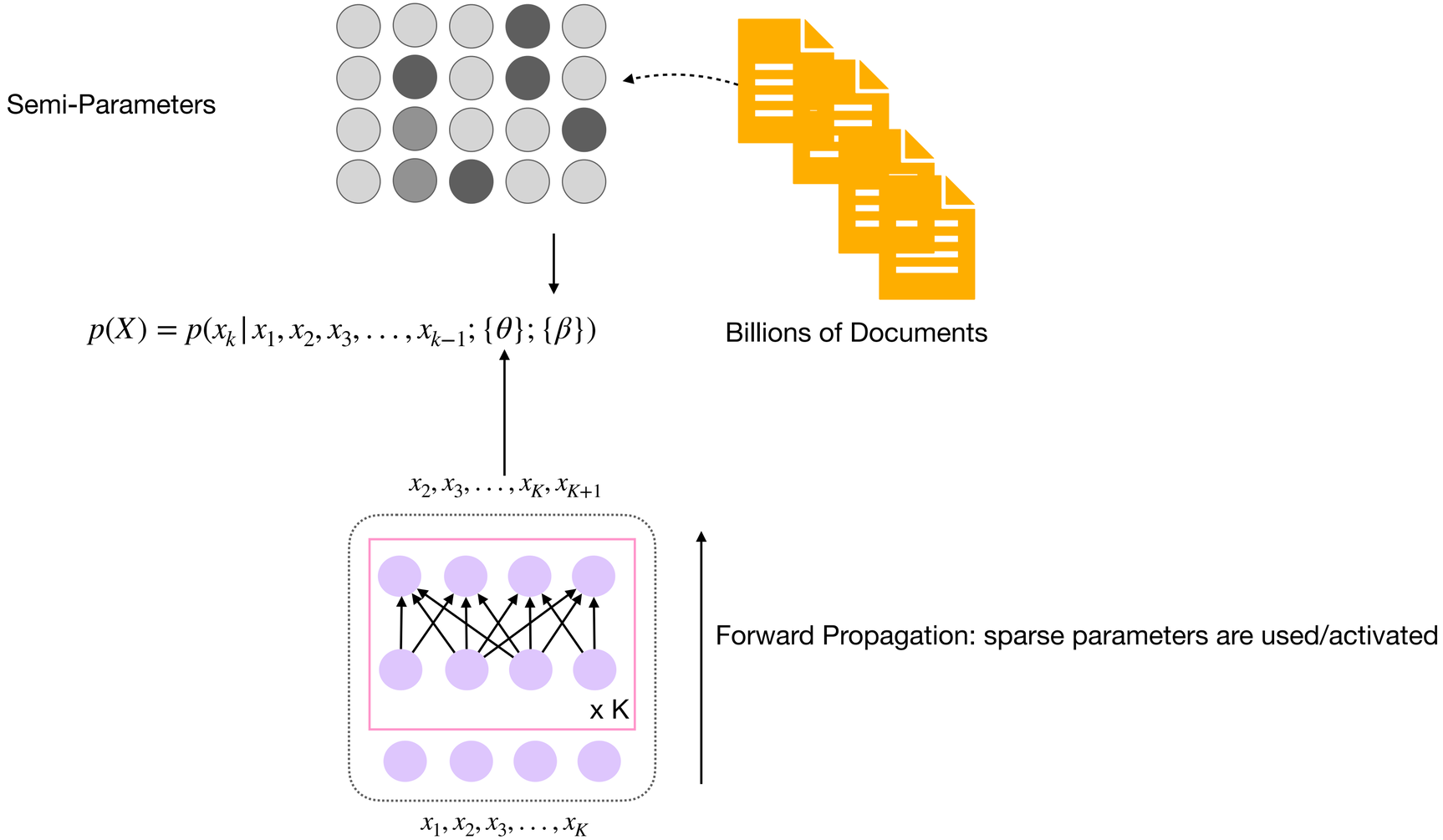
Fully-Parametric Language Model



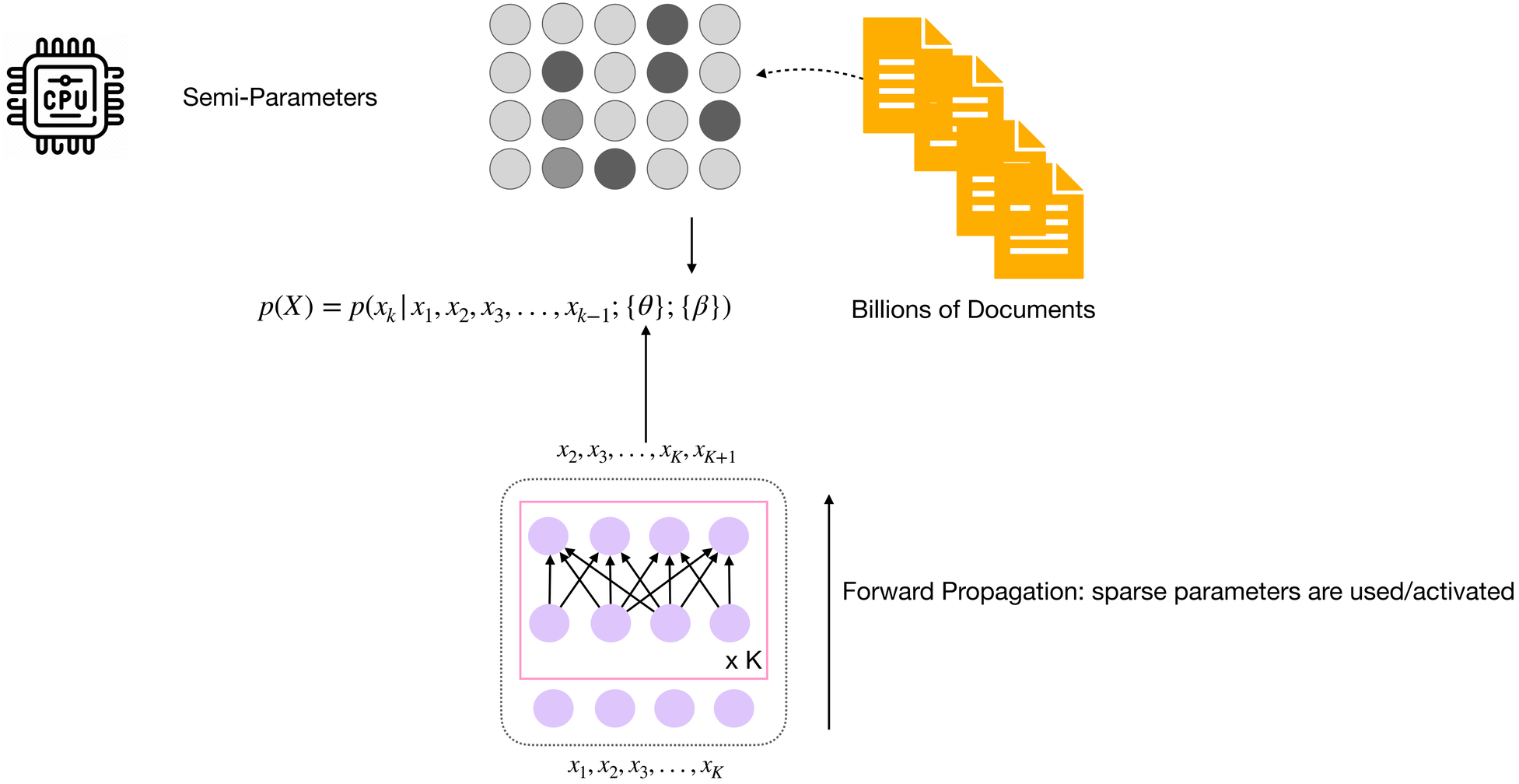
Semi-Parametric Language Model



Semi-Parametric Language Model

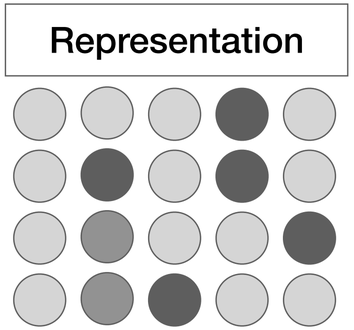


Semi-Parametric Language Model

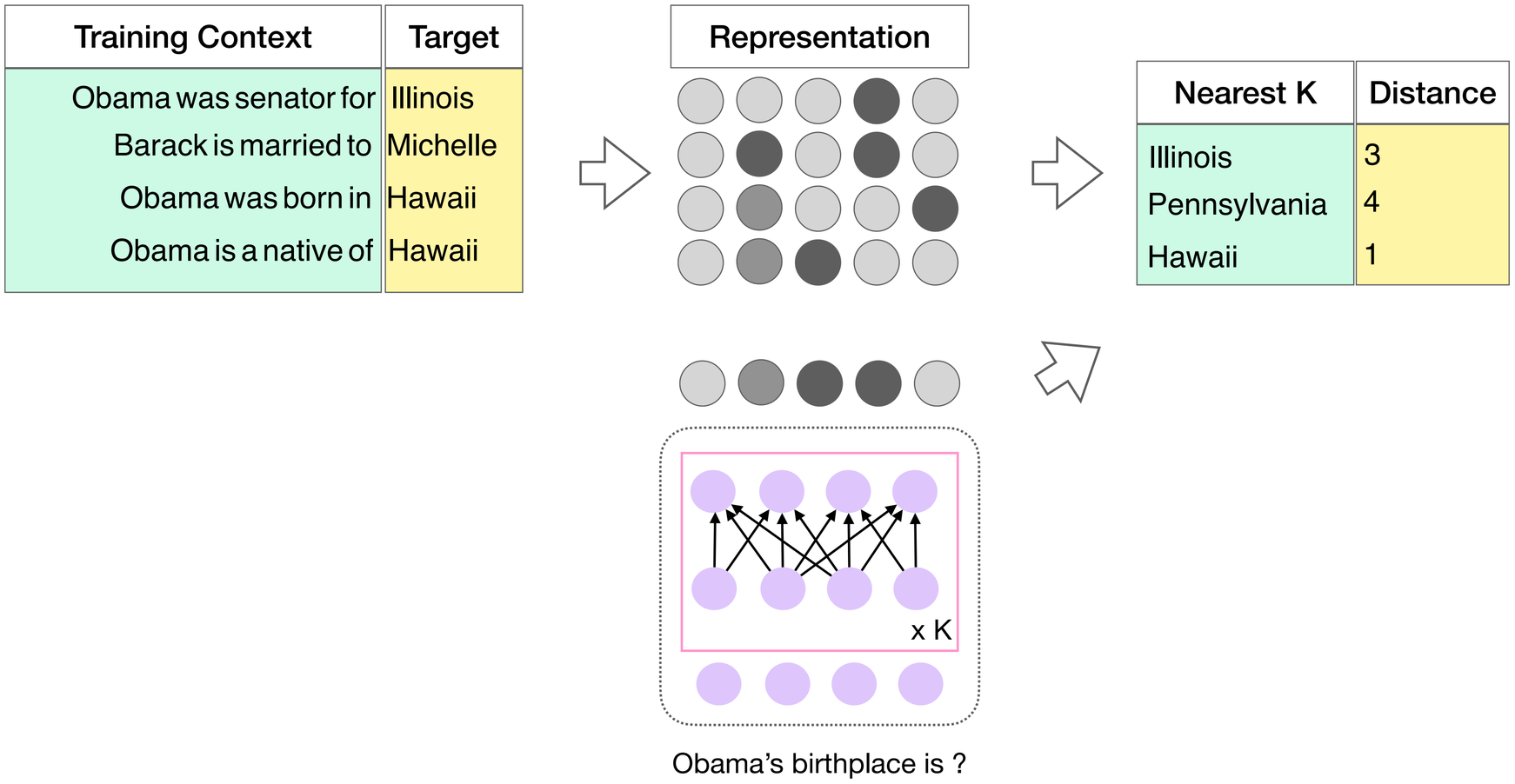


K-Nearest Neighbor LM (Khandelwal et al. 2020)

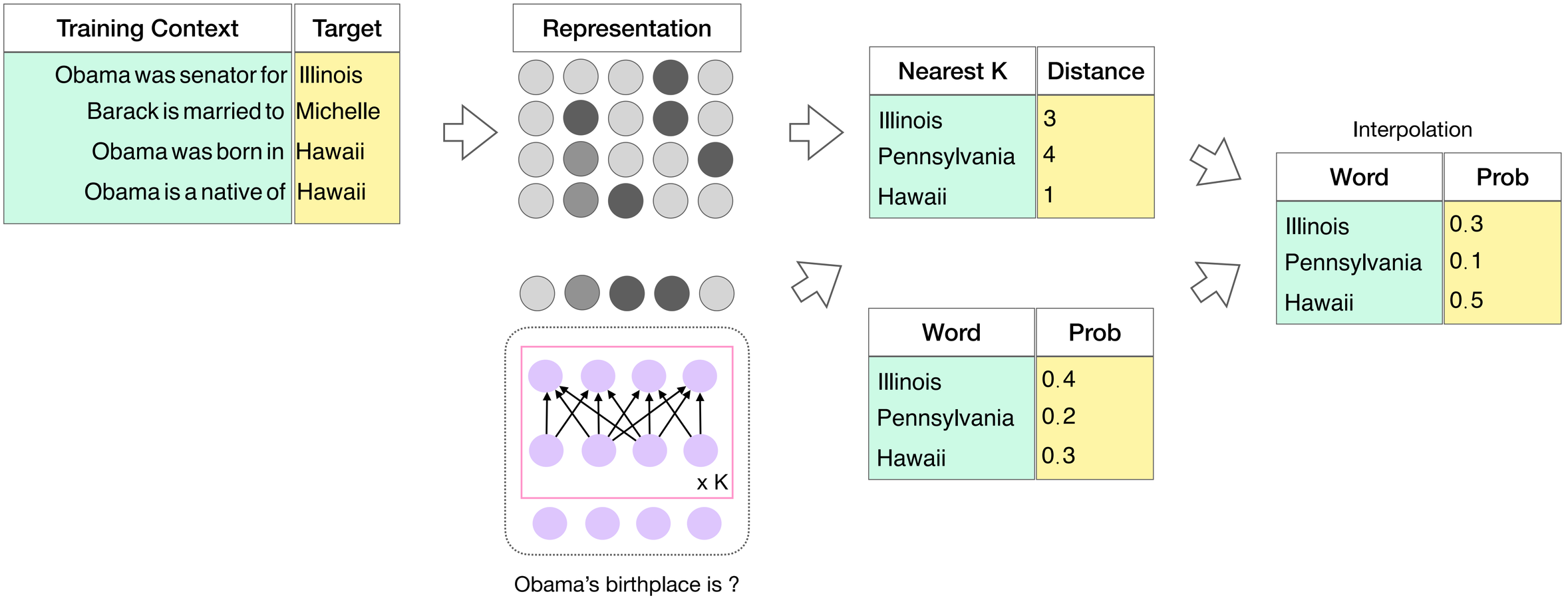
Training Context	Target
Obama was senator for	Illinois
Barack is married to	Michelle
Obama was born in	Hawaii
Obama is a native of	Hawaii



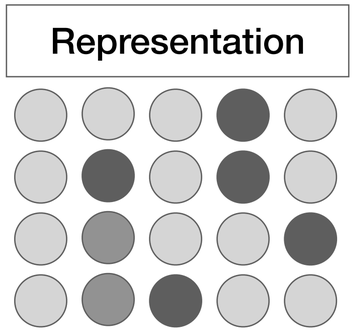
K-Nearest Neighbor LM (Khandelwal et al. 2020)



K-Nearest Neighbor LM (Khandelwal et al. 2020)

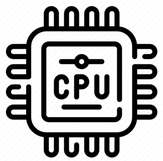


KNN-LM: Advantages



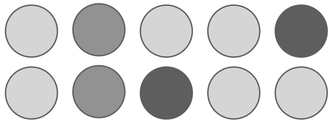
100M Vectors

Pre-computed, saved in CPU RAM, no added GPU memory cost for model



Sparse activated parameters, very few computation overhead.

Top-K Vector



Nearest K	Distance
Illinois	3
Pennsylvania	4
Hawaii	1

No need to back propagate gradient, no added training complexity

KNN-LM: Results

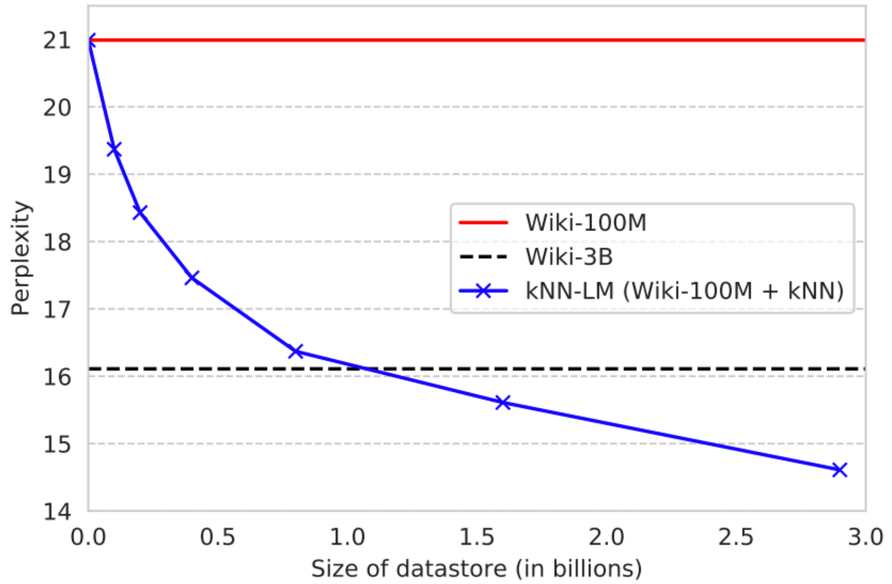
State-of-the-art LM performance on Wiki-100M Corpus

Language Model Metrics: Perplexity, the lower the better

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

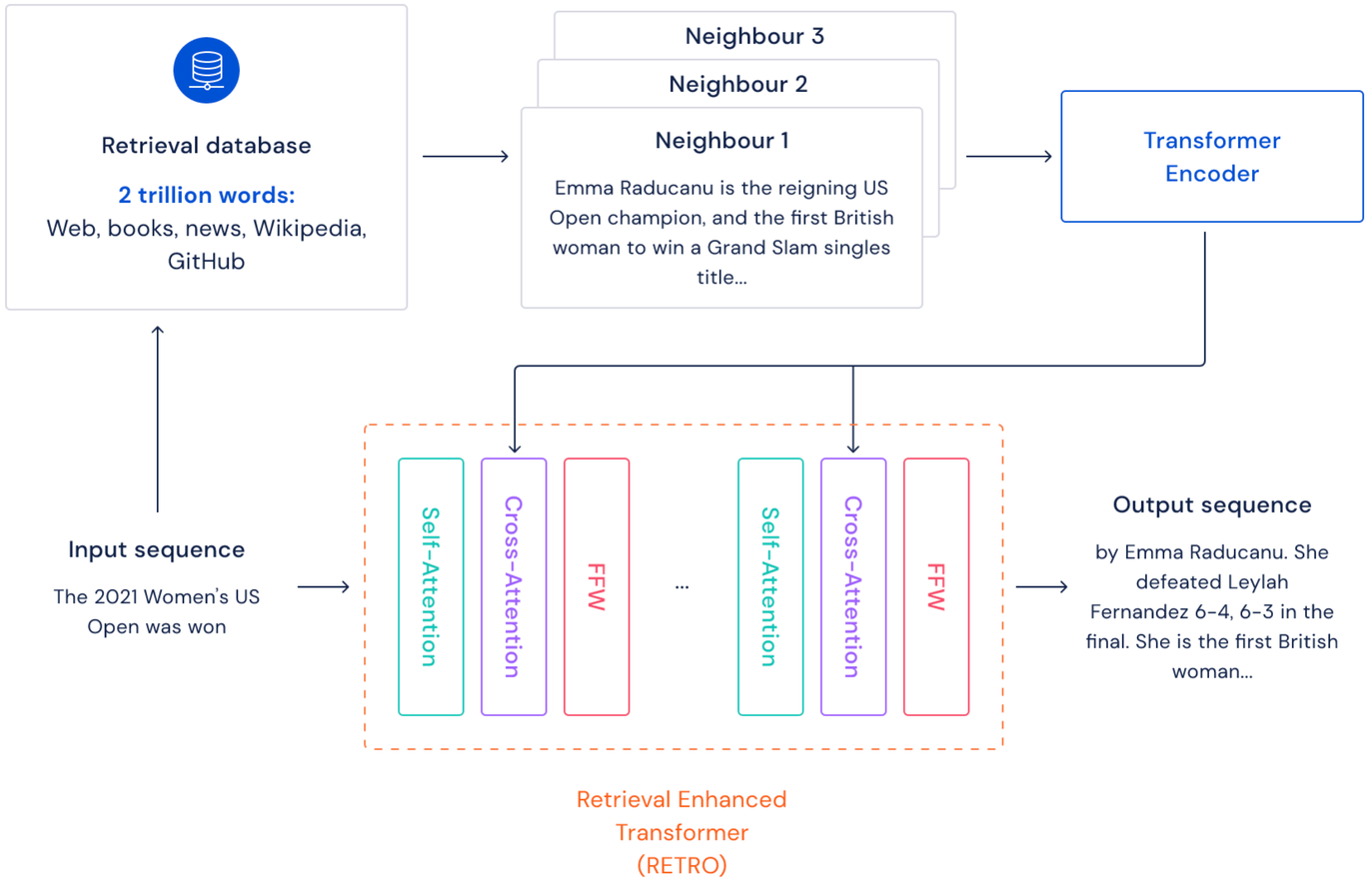
KNN-LM: Results

- Train on 100M token corpus
- Train on 3B token corpus
- Train on 100 token corpus but uses 3B token corpus as memory



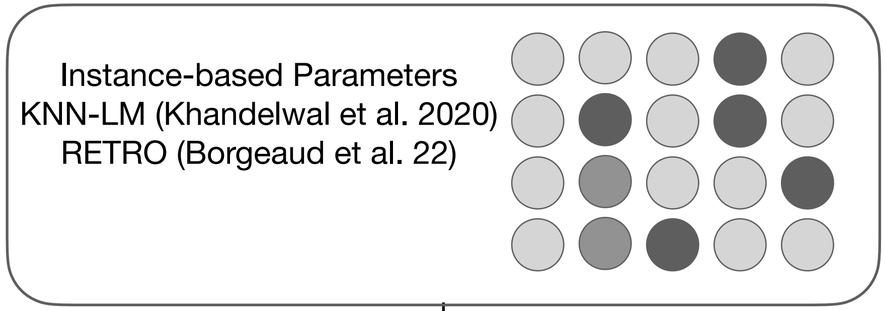
Deepmind RETRO (Borgeaud et al. 2022)

- Attend over 2-trillion words.
- Attend once in the intermediate step.
- RETRO obtains comparable performance as GPT-3 on LM corpus though having 25x less parameters.

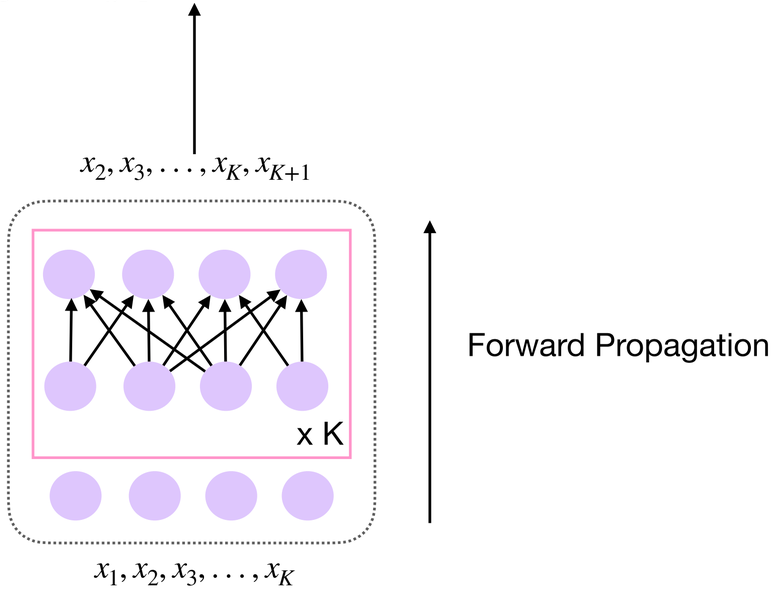


Instance-based Semi-parametric Model

Not Interpretable.
Not modularized.

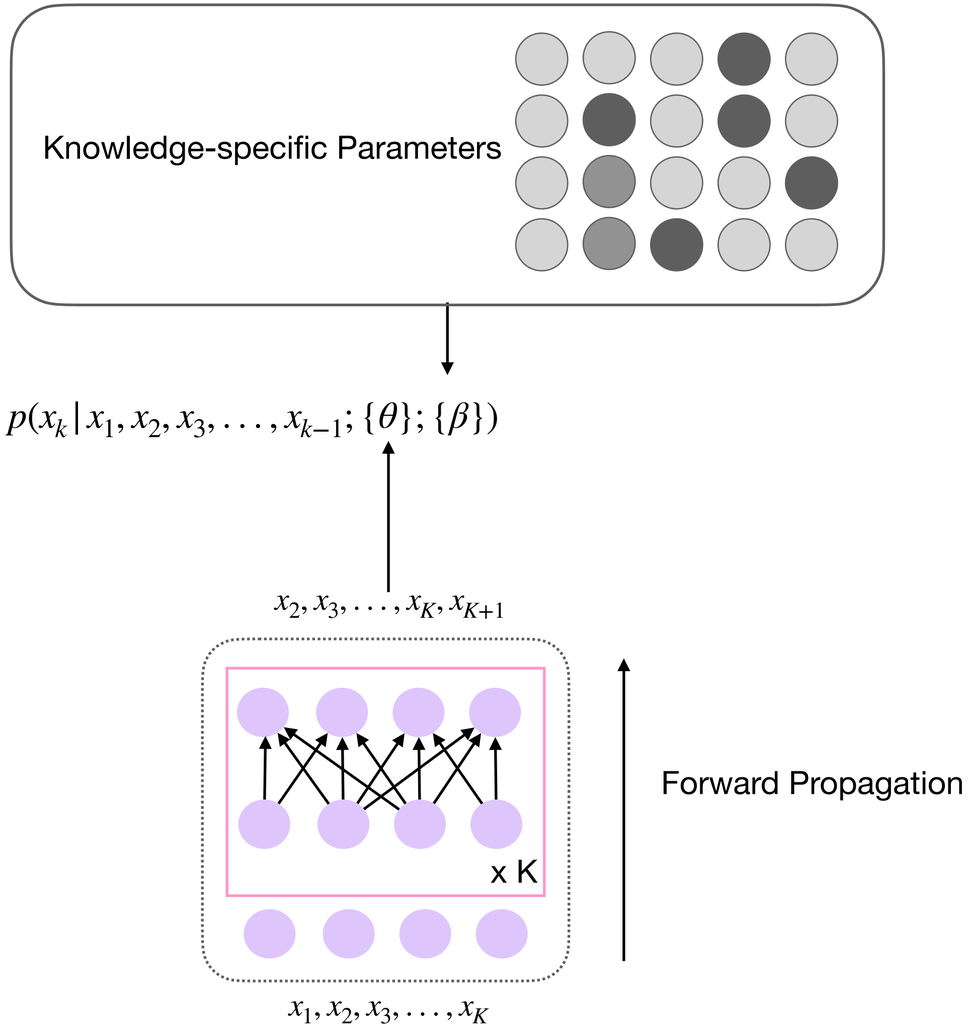


$$p(X) = p(x_k \mid x_1, x_2, x_3, \ldots, x_{k-1}; \{\theta\}; \{\beta\})$$

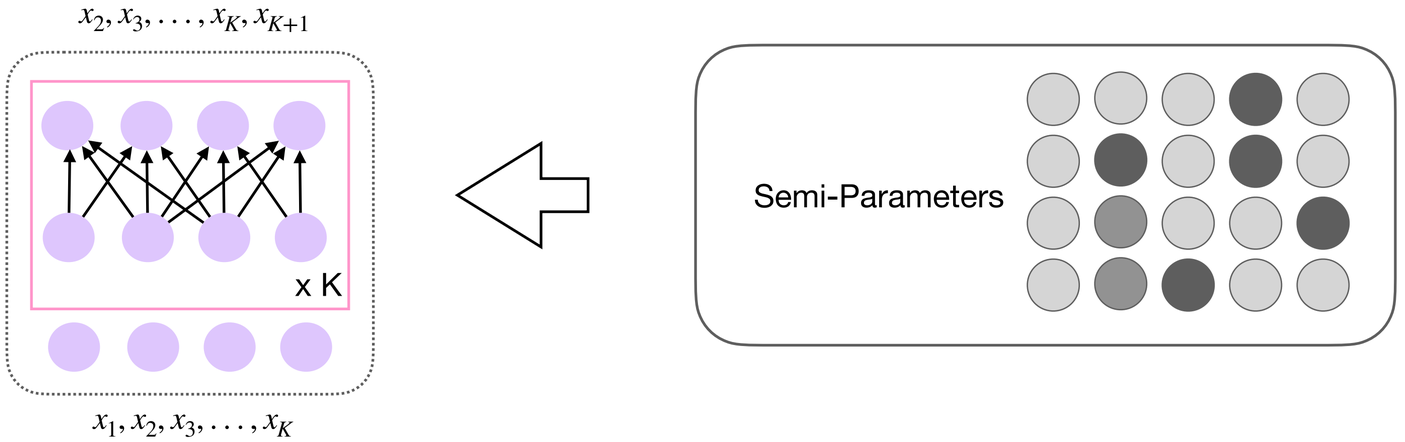


Knowledge-based Semi-parametric Model

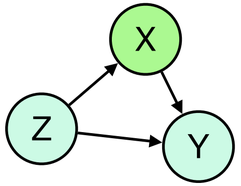
Every parameter has physical meaning.
Disentangle Language/Knowledge



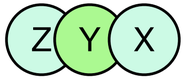
Memory-Augmented Language Models



Entities Memory
Fevry et al. 2020



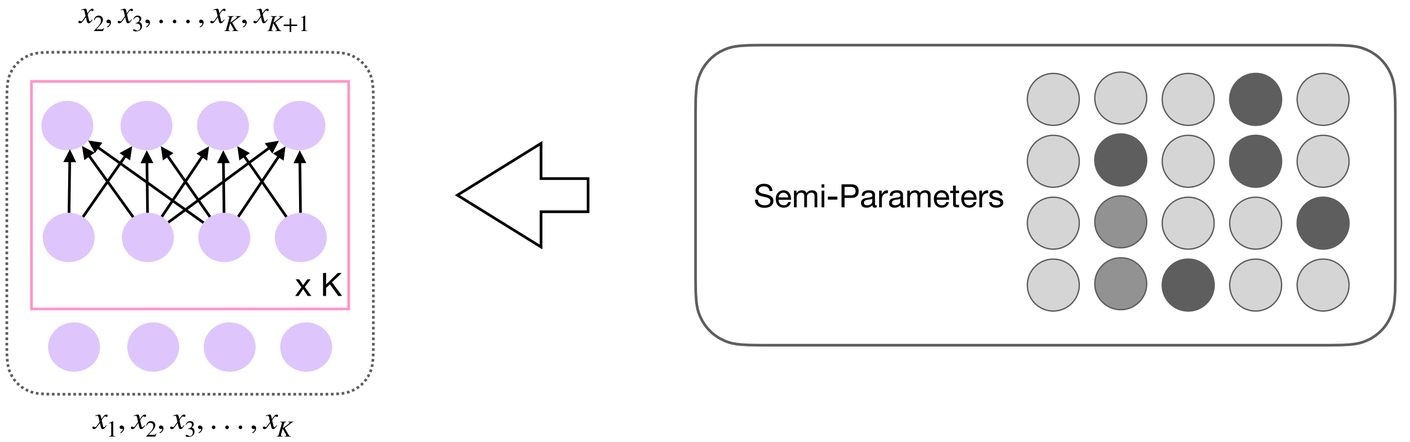
KB Memory
Sun & Verga et al. 2021

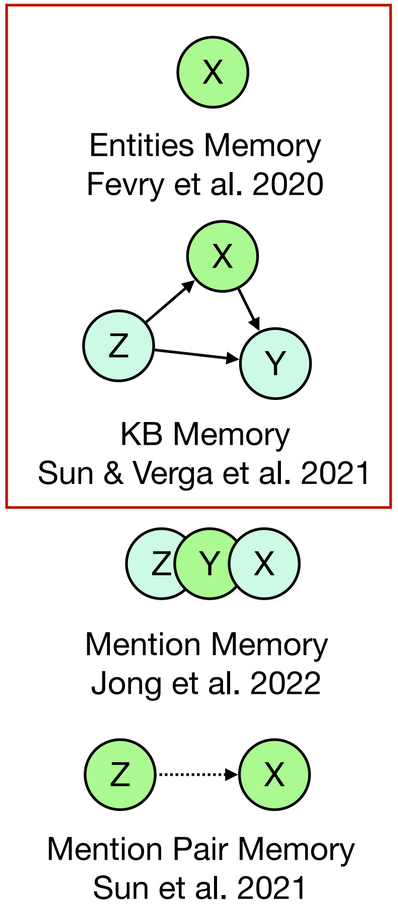


Mention Memory
Jong et al. 2022

Mention Pair Memory
Sun et al. 2021

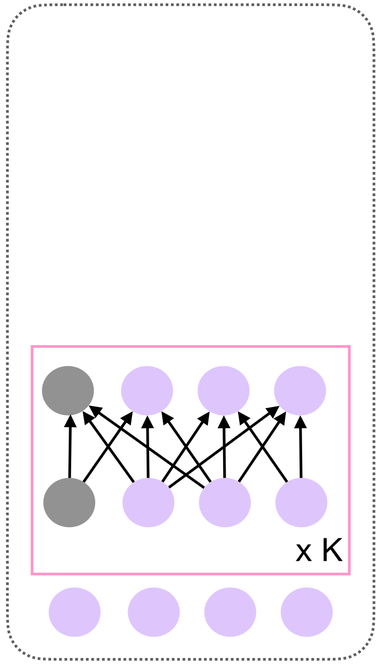
Memory-Augmented Language Models





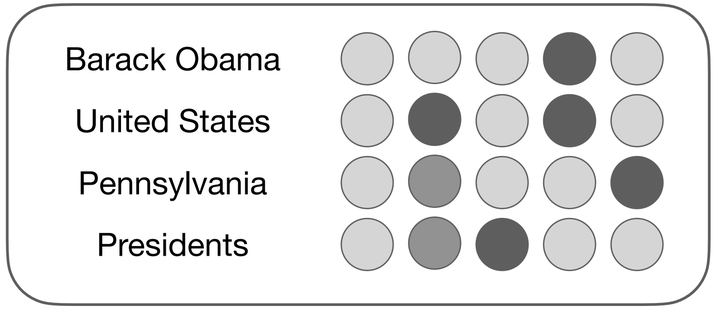
Semi-parametric Model with Entity Memory

Entities as Experts (Févry et al. 2020)



[MASK] was born in Hawaii and served as the President of USA …

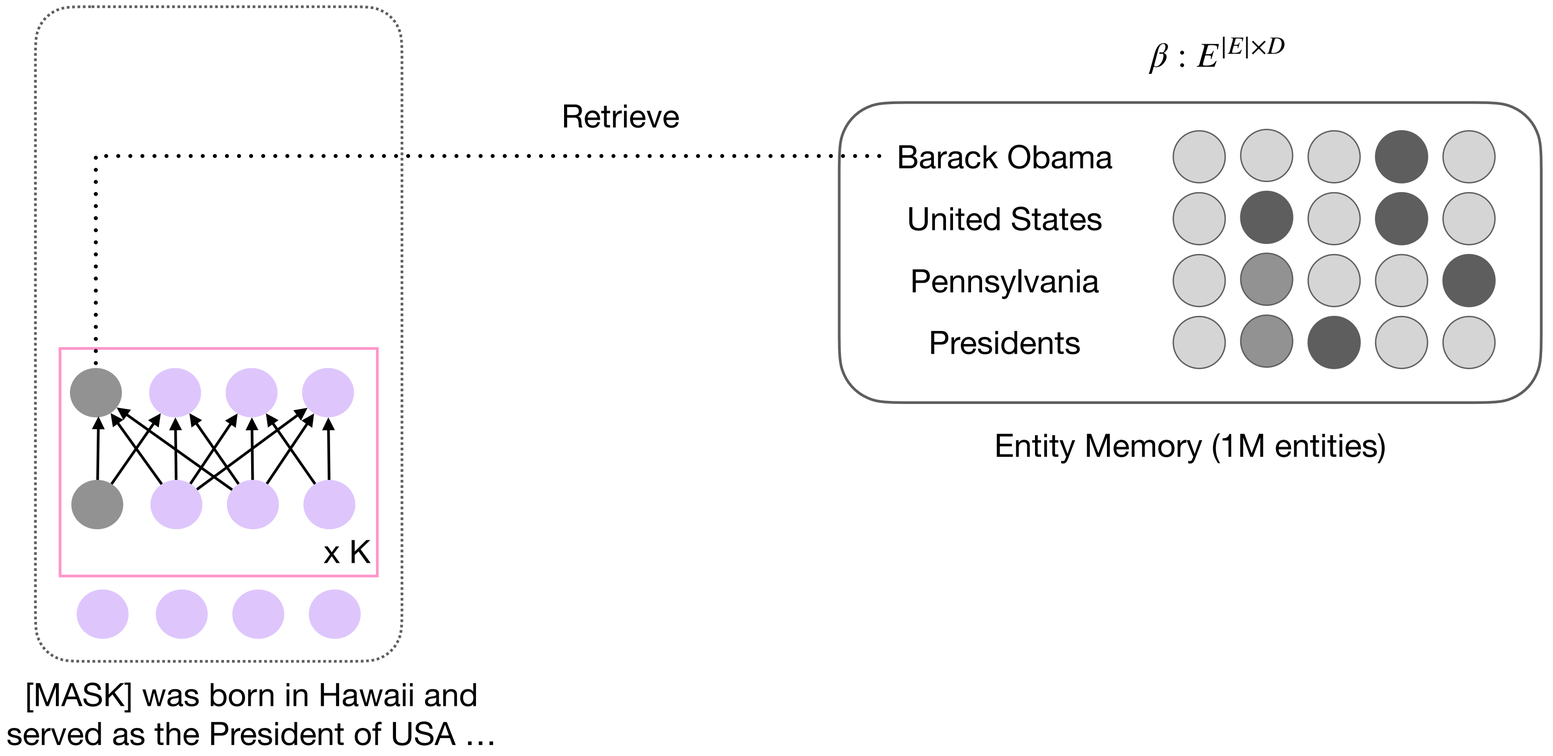
$\beta : E^{|E| \times D}$



Entity Memory (1M entities)

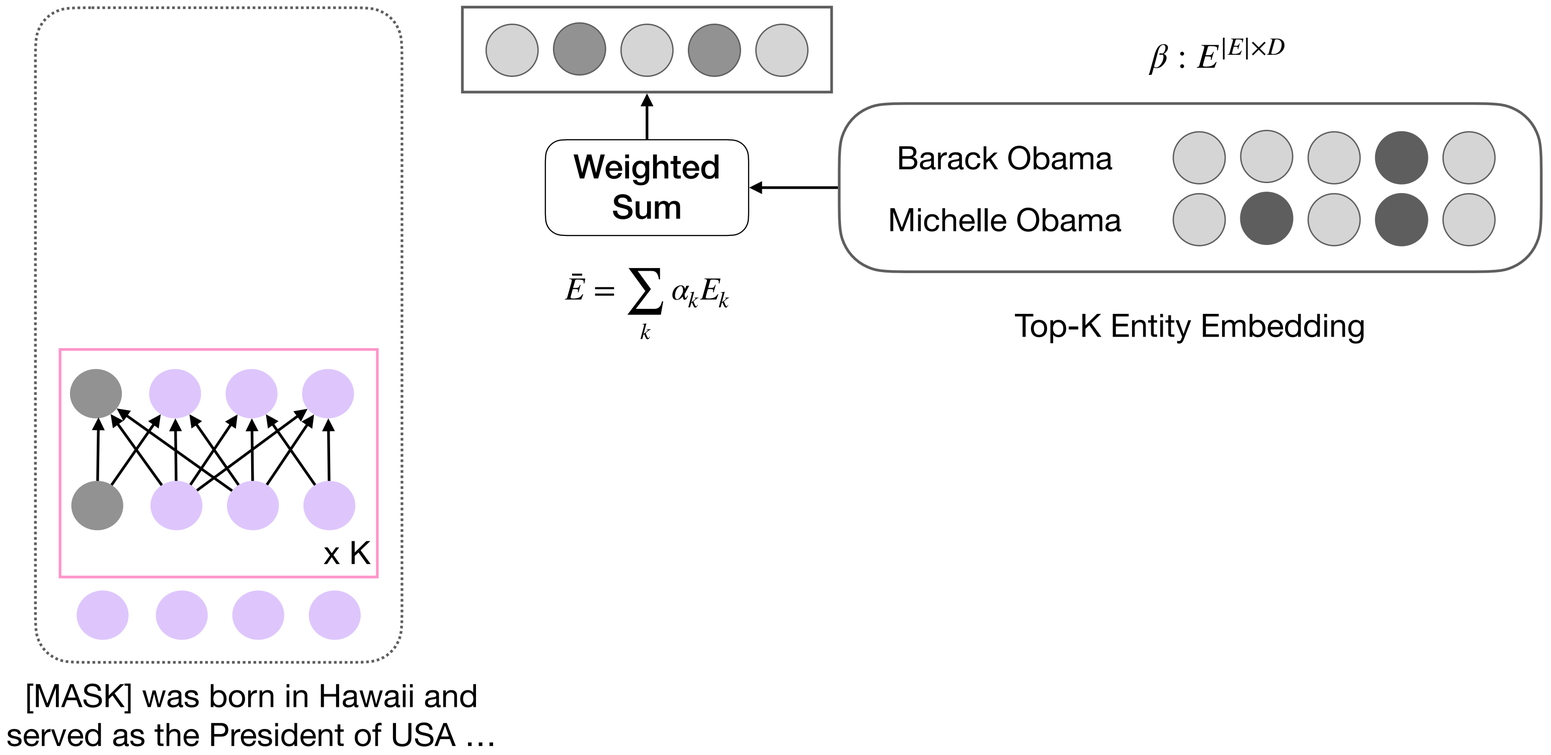
Semi-parametric Model with Entity Memory

Entities as Experts (Févry et al. 2020)



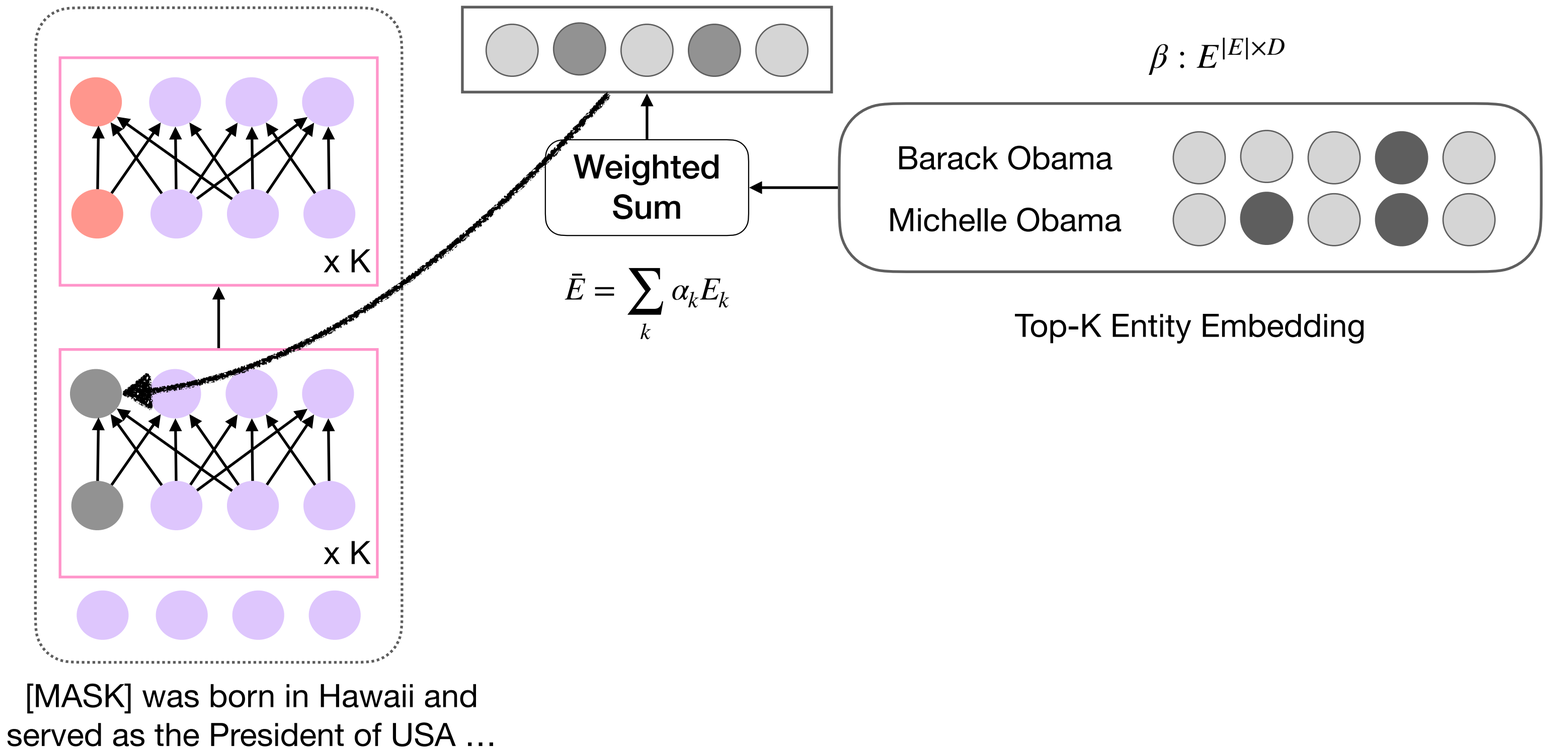
Semi-parametric Model with Entity Memory

Entities as Experts (Févry et al. 2020)



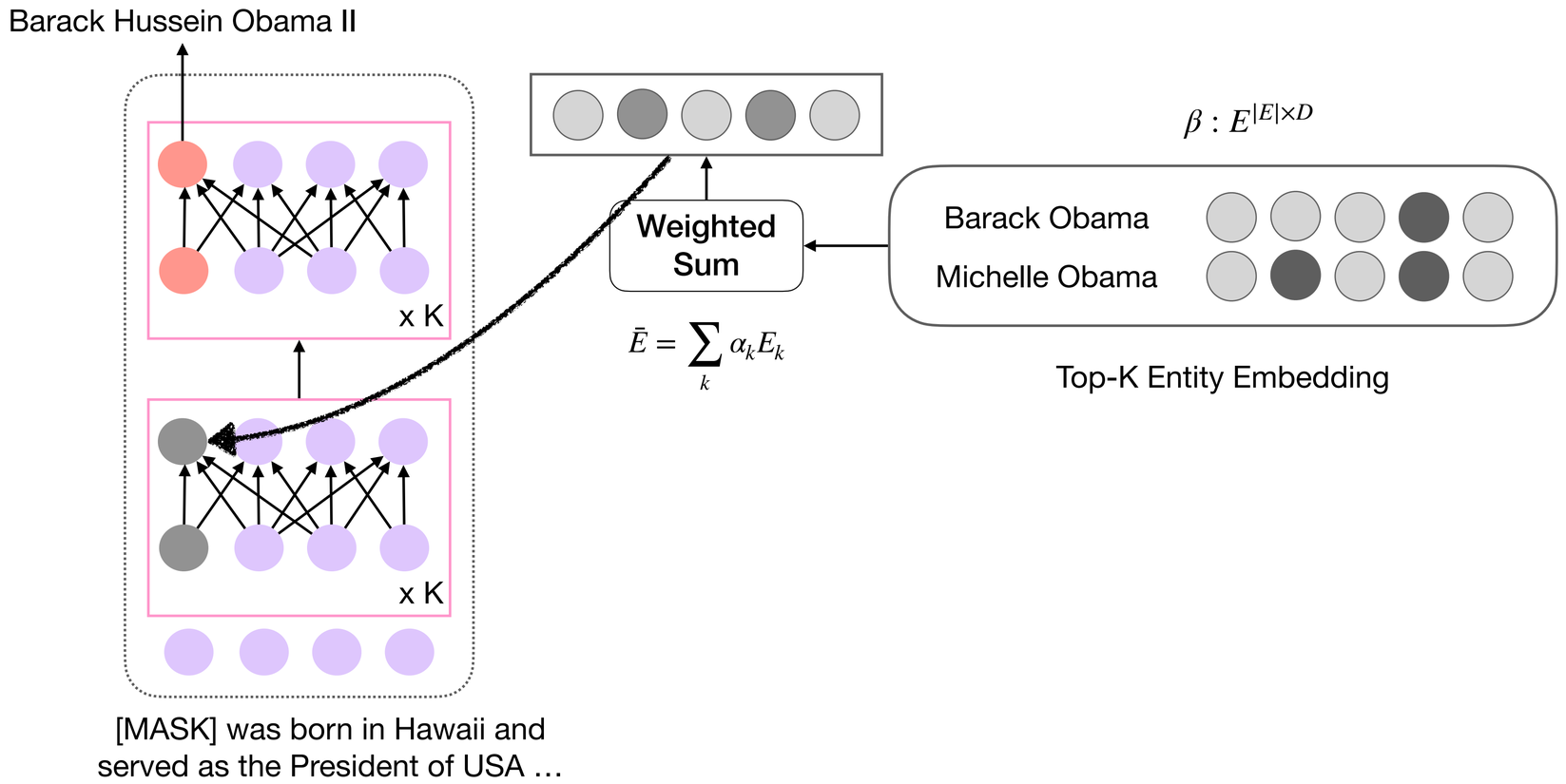
Semi-parametric Model with Entity Memory

Entities as Experts (Févry et al. 2020)



Semi-parametric Model with Entity Memory

Entities as Experts (Févry et al. 2020)



Pre-Training & Fine-Tuning

Pre-Training: Entity Linking from Wikipedia to WikiData

Florence Bell (scientist)

From Wikipedia, the free encyclopedia

Florence Ogilvy Bell (1 May 1913 – 23 November 2000), later **Florence Sawyer**, was a British scientist who contributed to the discovery of DNA. She was an X-ray crystallographer in the lab of William Astbury. In 1938 they published a paper in *Nature* that described the structure of DNA as a "Pile of Pennies".

WIKIPEDIA
The Free Encyclopedia

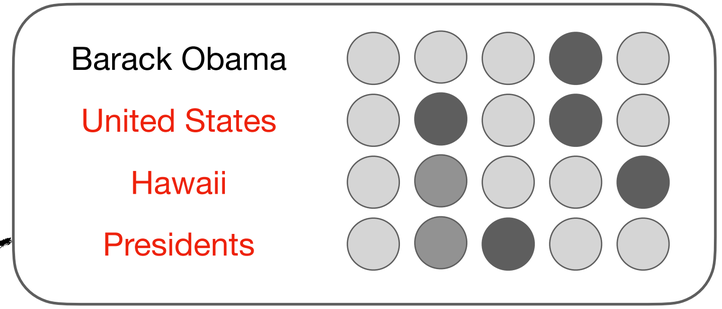
Florence Bell (Q52581420)

Place of birth (P19): London (Q84)
Date of birth (P569): 1 May 1913
Date of death (P570): 23 November 2000
Occupation (P106): Scientist (Q901)
Employer (P108): University of Leeds (Q503424)
Doctoral thesis: (P1026): X-ray and related studies of the structure of the proteins and nucleic acids (Q59314810)
Doctoral advisor (P184): William Astbury (Q562321)

WIKIDATA

WIKIMEDIA COMMONS

Fine-tuning: Question-Answering Task

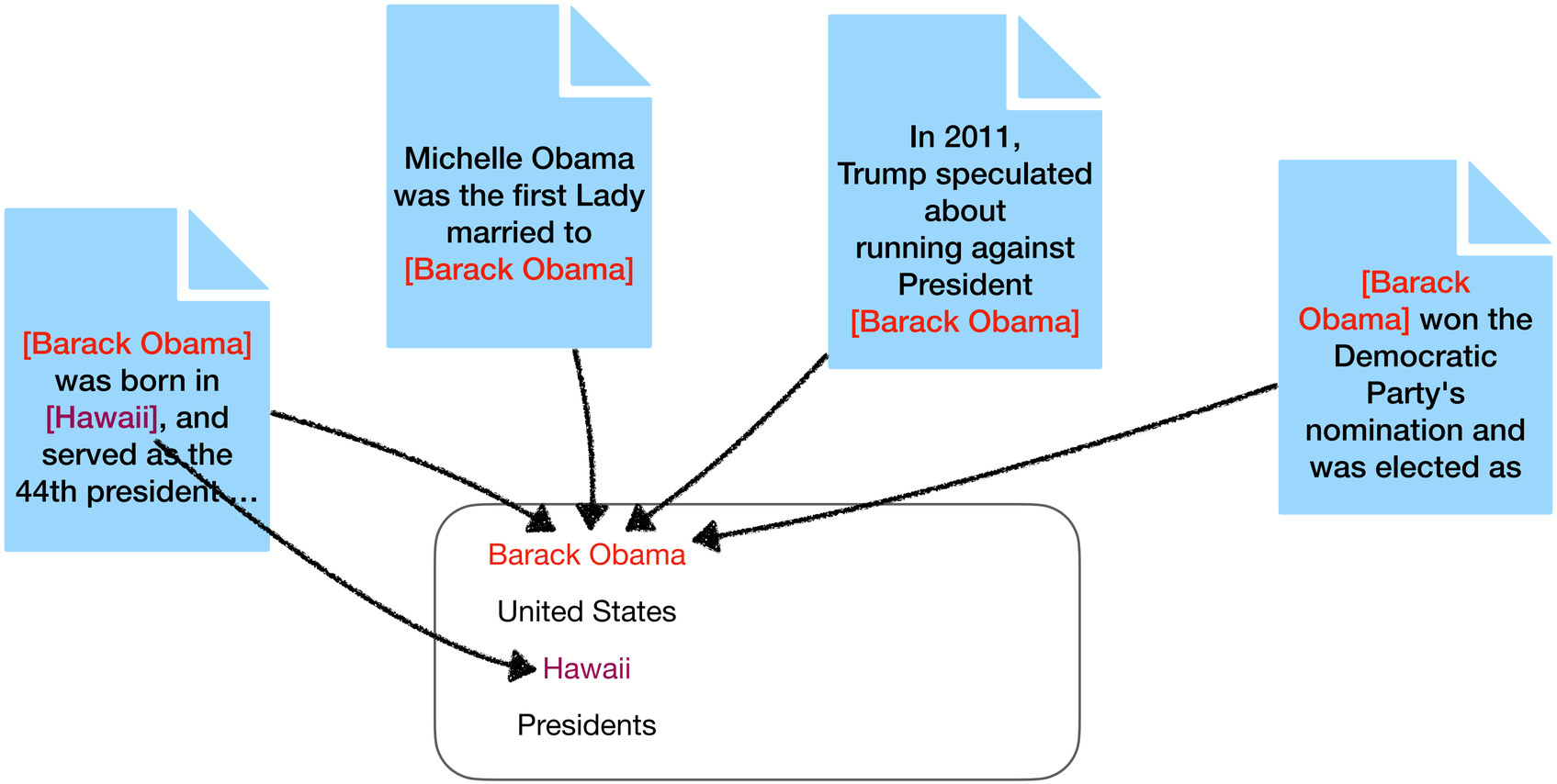


Entity Memory (1M entities)

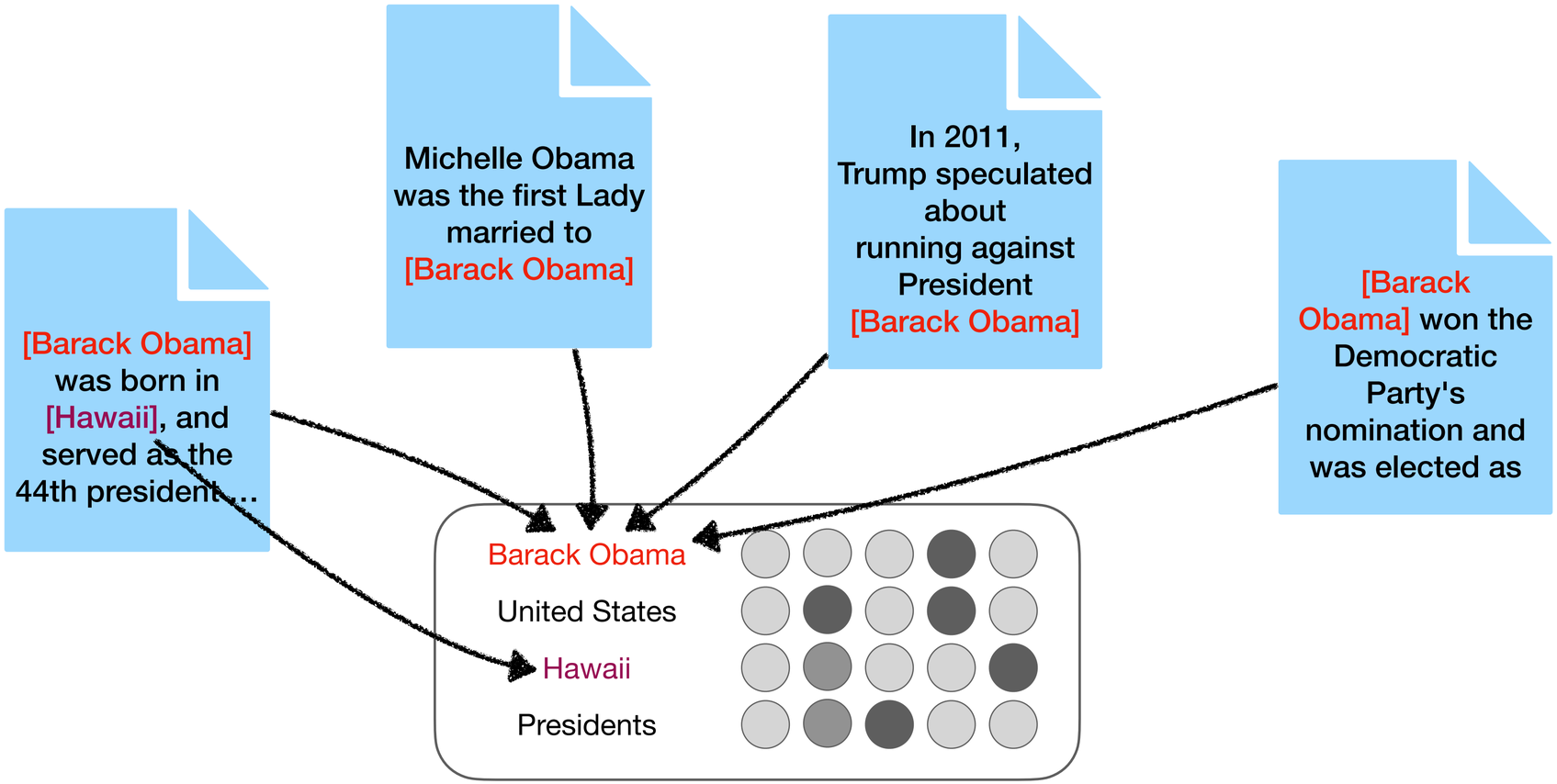
Who was born in Hawaii and served as the President of USA? **[ENTITY]**

Entity List	Prob
John Kennedy	0.1
Barack Obama	0.4
Donald Trump	0.2

Core Idea

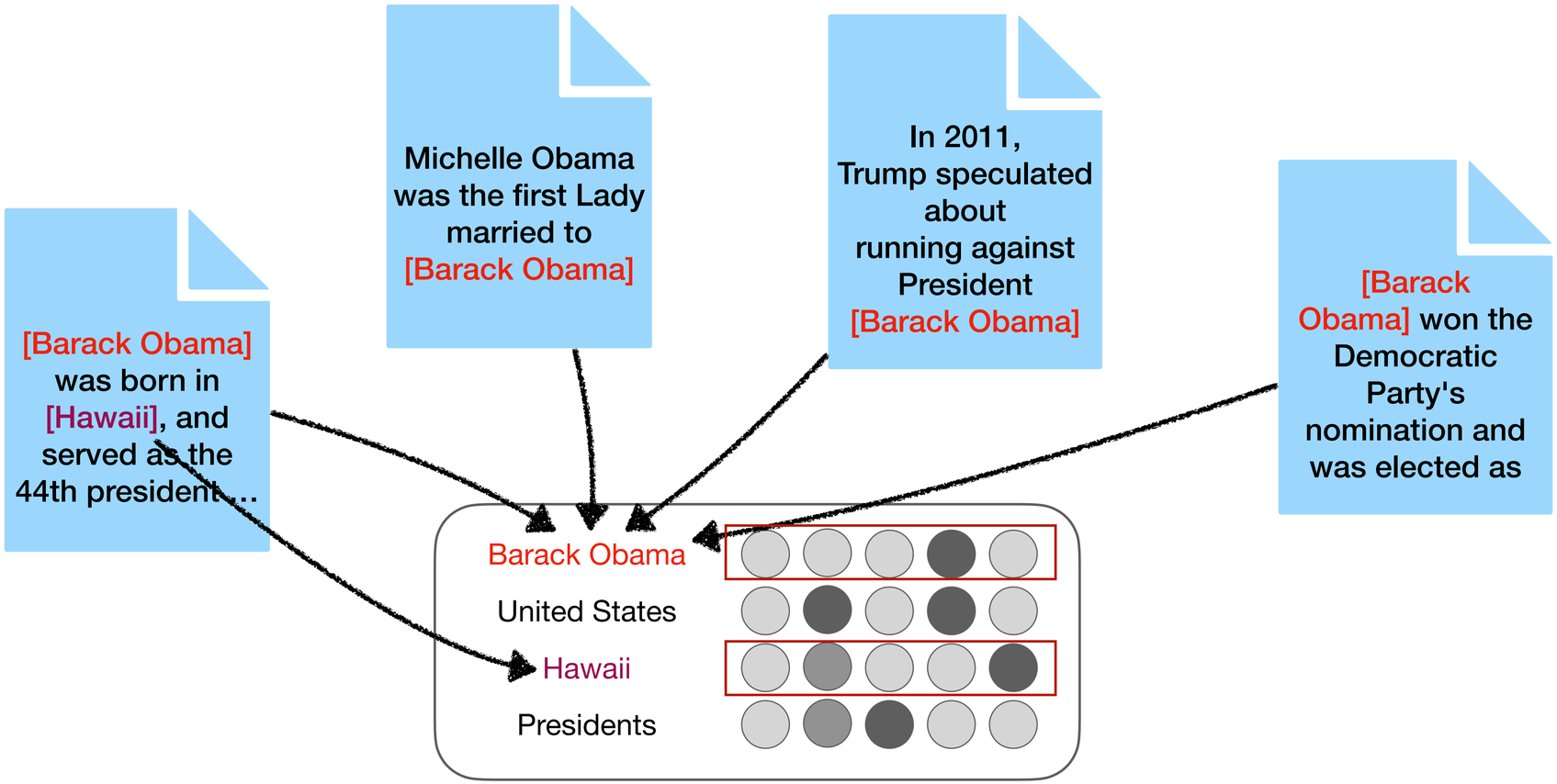

[Barack Obama] was born in [Hawaii], and served as the 44th president ...
Michelle Obama was the first Lady married to [Barack Obama]
In 2011, Trump speculated about running against President [Barack Obama]
[Barack Obama] won the Democratic Party's nomination and was elected as
Barack Obama
United States
Hawaii
Presidents

Core Idea



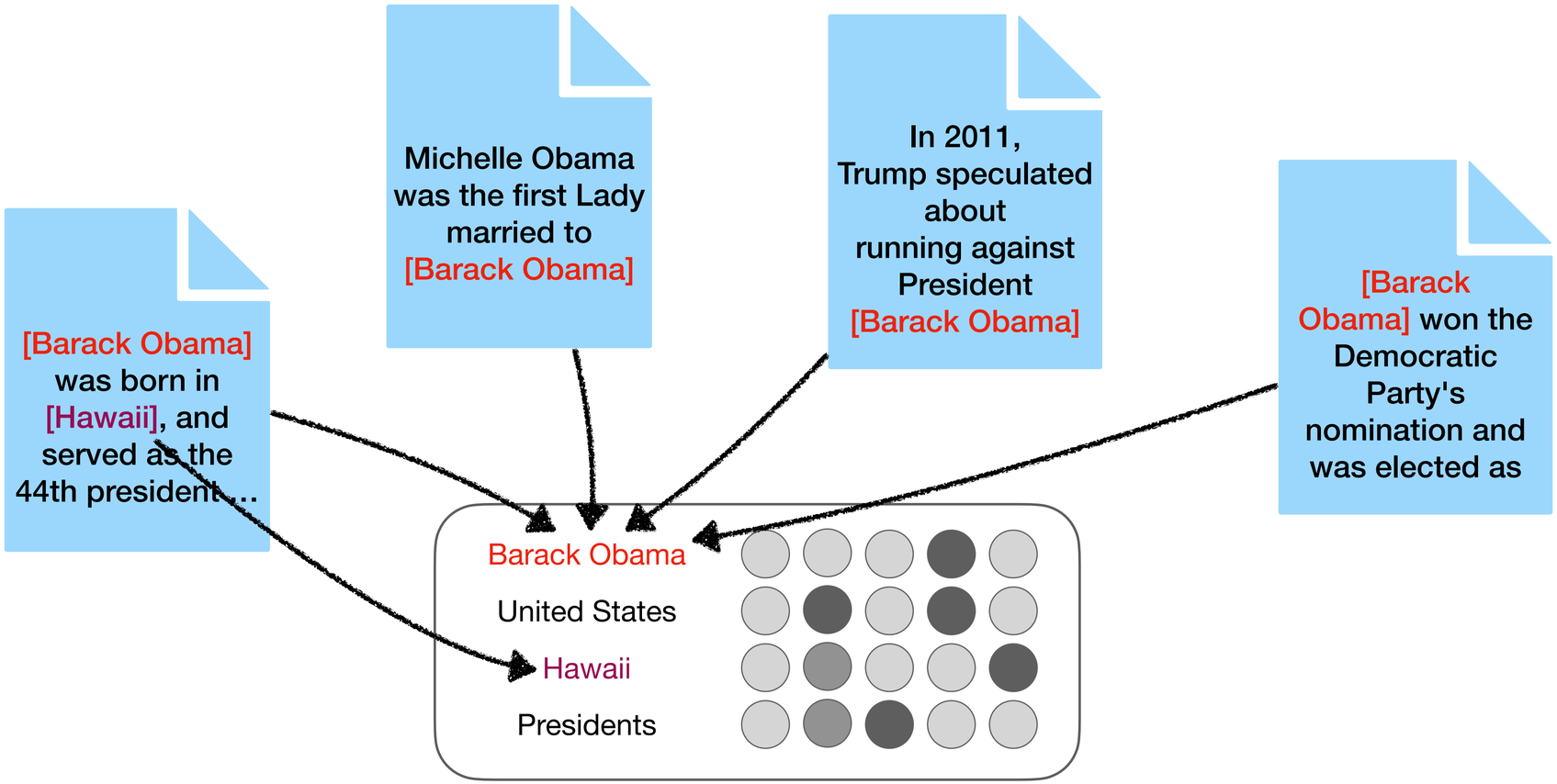
Aggregating to entity-context information into the embedding

Core Idea



Learn the entity-to-entity correlation based on their embedding

Limitation

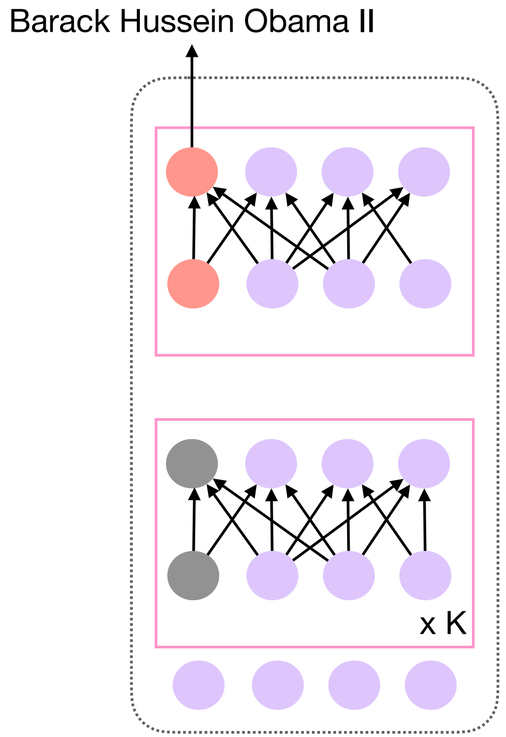


Learning Co-occurrences of entities based on a large collections of text

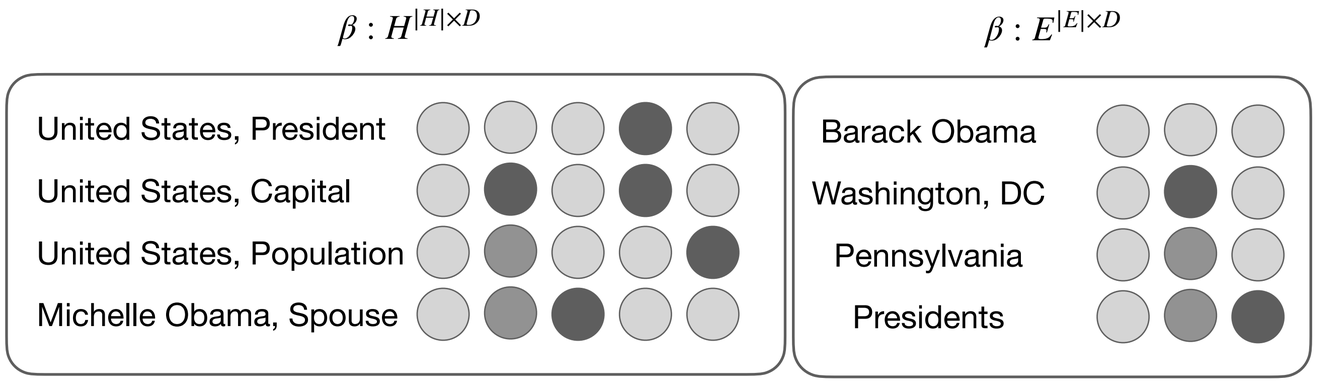
- No “knowledge” in the memory, not interpretable.
- Learning statistical correlations between entities rather than reasoning.

Semi-parametric Model with KG Memory

Fact-Injected Language Model (Sun & Verga et al. 2021)



[MASK] was born in Hawaii and served as the President of USA …

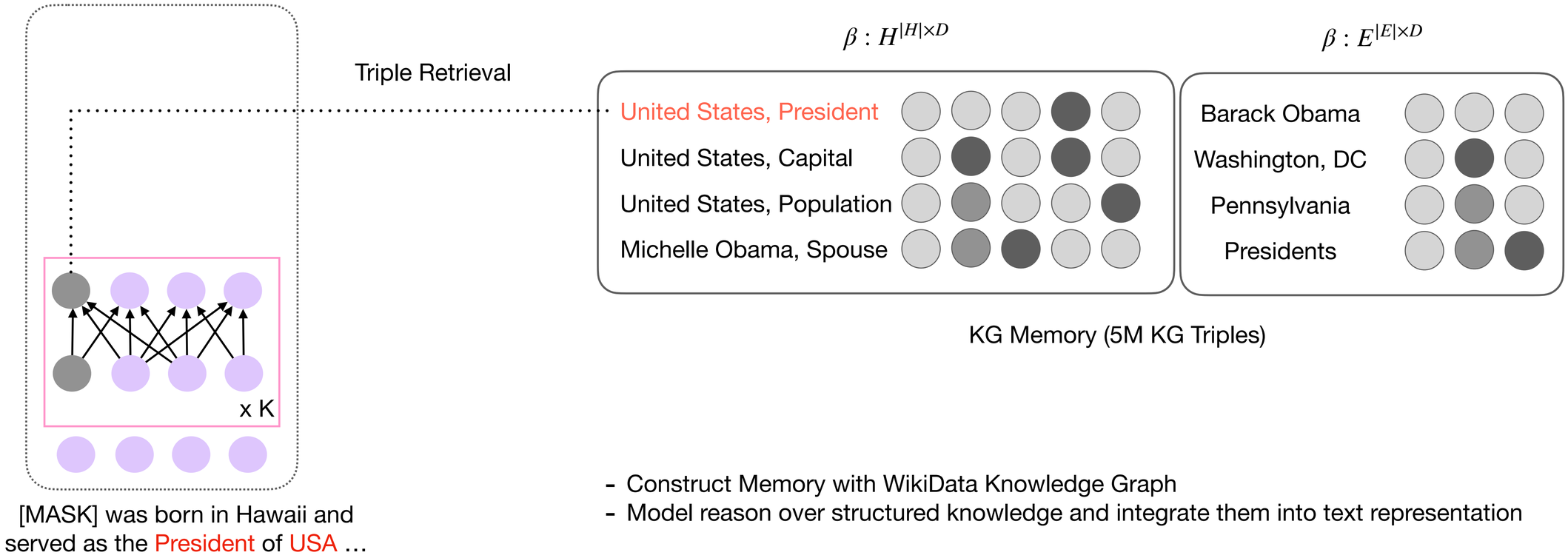


KG Memory (5M KG Triples)

- Construct Memory with WikiData Knowledge Graph
- Model reason over structured knowledge and integrate them into text representation

Semi-parametric Model with KG Memory

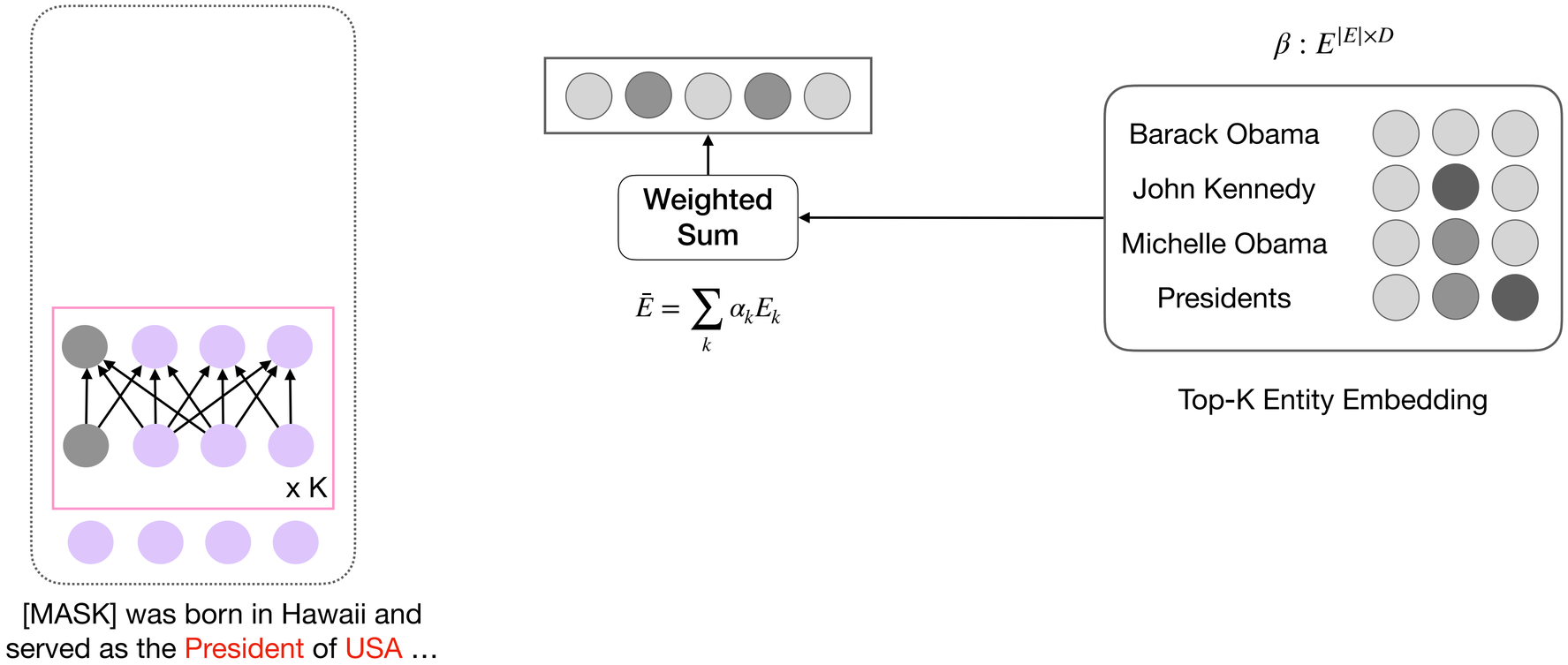
Fact-Injected Language Model (Sun & Verga et al. 2021)



- Construct Memory with WikiData Knowledge Graph
- Model reason over structured knowledge and integrate them into text representation

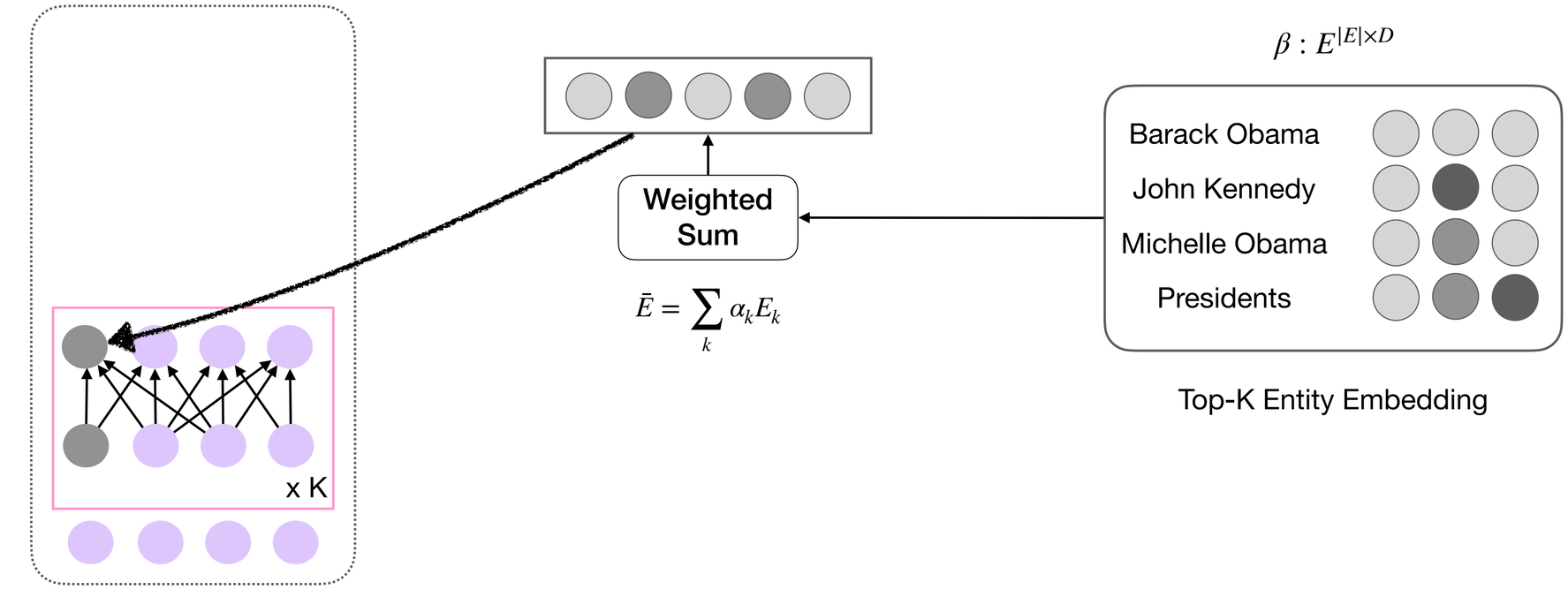
Semi-parametric Model with KG Memory

Fact-Injected Language Model (Sun & Verga et al. 2021)



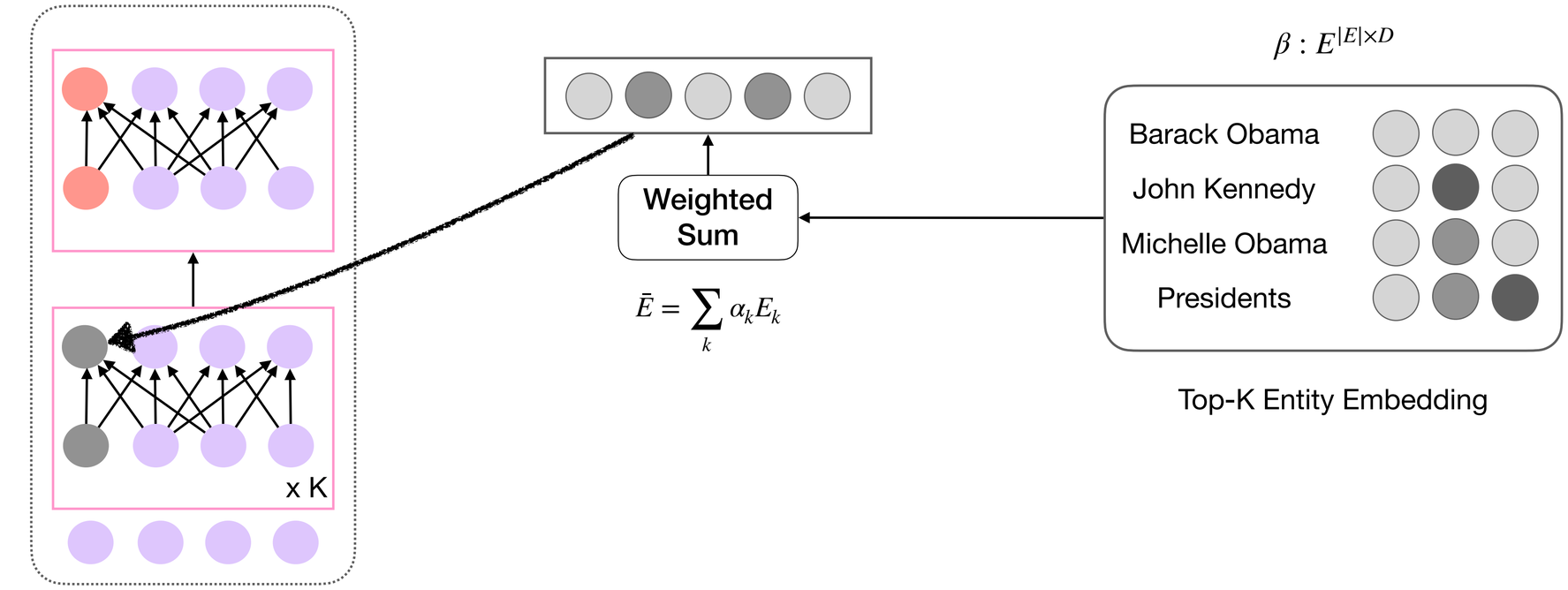
Semi-parametric Model with KG Memory

Fact-Injected Language Model (Sun & Verga et al. 2021)



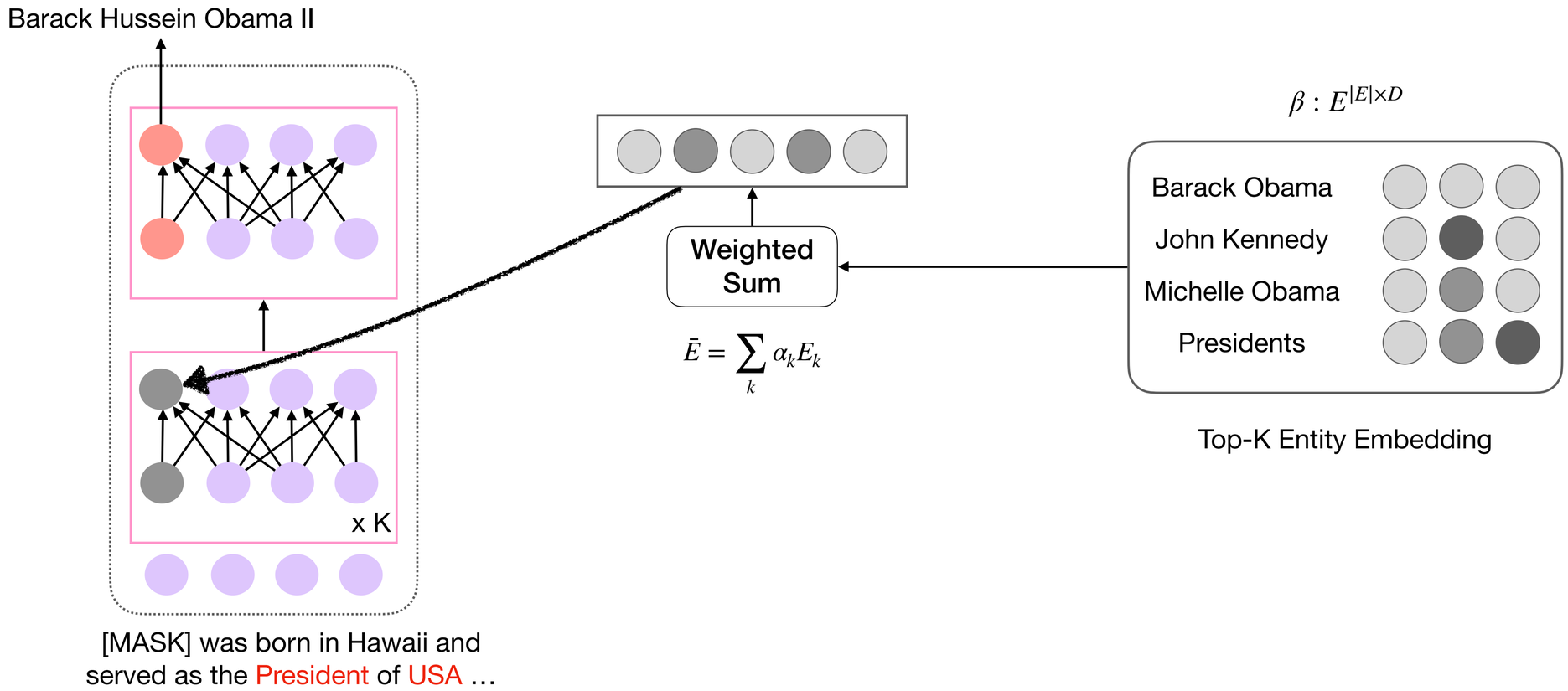
Semi-parametric Model with KG Memory

Fact-Injected Language Model (Sun & Verga et al. 2021)



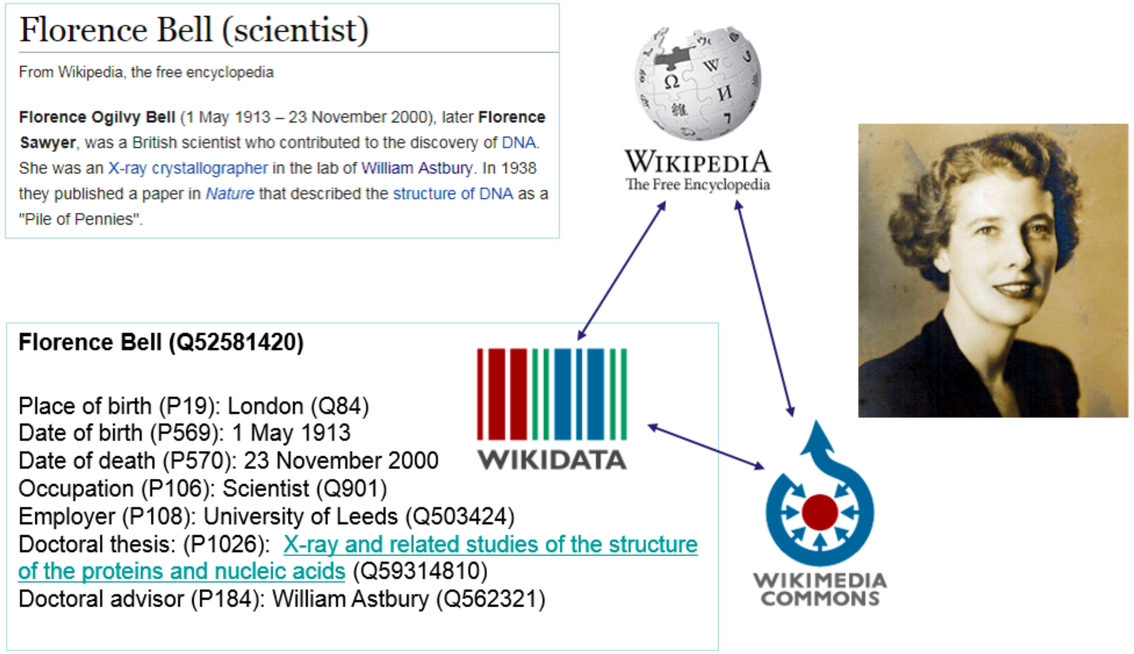
Semi-parametric Model with KG Memory

Fact-Injected Language Model (Sun & Verga et al. 2021)

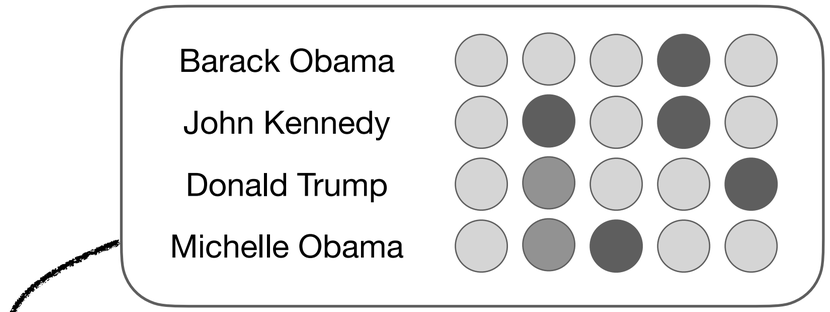


Implementation

Pre-Training: Entity Linking from Wikipedia to WikiData



Fine-tuning: Question-Answering Task



KG Memory (5M KG triples)

Who was born in Hawaii and served as the President of USA? **[ENTITY]**

Entity List	Prob
John Kennedy	0.1
Barack Obama	0.4
Donald Trump	0.2

Experimental Results

Using Memory vs. No-Memory: Parameter Efficiency Increases by 30x.

Model	Params	TriviaQA	TriviaQA No Overlap	WebQuestion	WebQuestion SP
T5-3B	3B	42.3	-	37.4	49.7
Entity Memory	110M	43.2	9.1	39.0	47.4
KG Memory	110M	-	15.6	-	54.7

Results taken from:

Verga, Pat, et al. "Adaptable and interpretable neural memoryover symbolic knowledge." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

Févry, Thibault, et al. "Entities as Experts: Sparse Memory Access with Entity Supervision." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

Experimental Results

Using KG-Memory vs. Entity-Memory: models learns to perform Multihop reasoning on WebQuestionSP

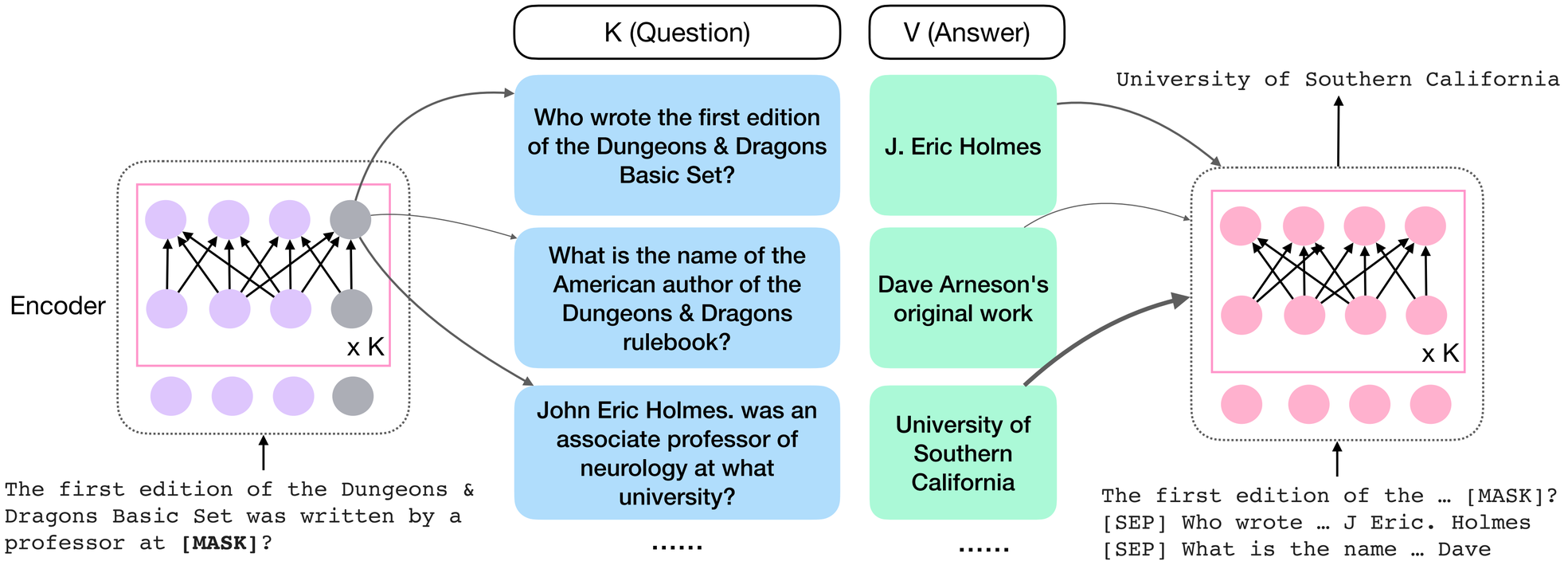
Model	Params	TriviaQA	TriviaQA No Overlap	WebQuestion	WebQuestion SP
T5-3B	3B	42.3	-	37.4	49.7
Entity Memory	110M	43.2	9.1	39.0	47.4
KG Memory	110M	-	15.6	-	54.7

Results taken from:

Verga, Pat, et al. "Adaptable and interpretable neural memoryover symbolic knowledge." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

Févry, Thibault, et al. "Entities as Experts: Sparse Memory Access with Entity Supervision." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.

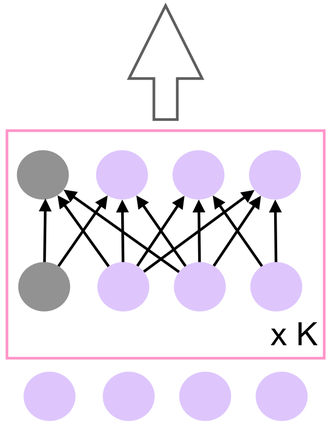
My work: Semi-parametric Model with QA Memory



Limitation of Entity/KG Memory

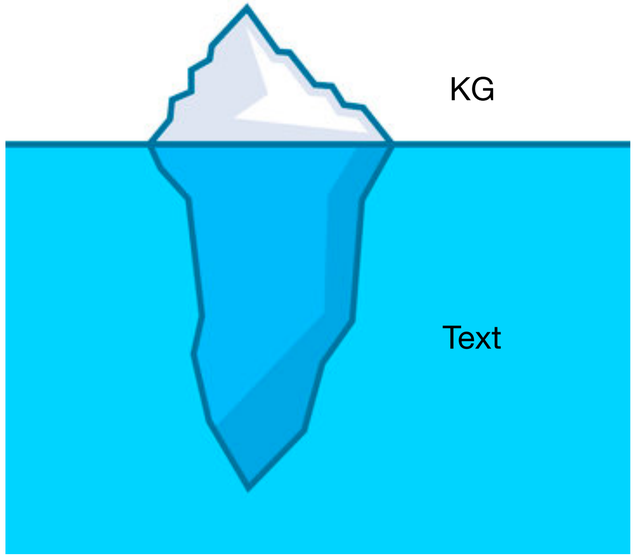
Answer Coverage:
The prediction is restricted only to the pre-defined 1M entities

Entity List	Prob
John Kennedy	0.1
Barack Obama	0.4
Donald Trump	0.2



What if the answer is a number or date or noun phrase? [ENTITY]

KG Coverage:
The memory can only host knowledge triple Pre-trained in WikiData



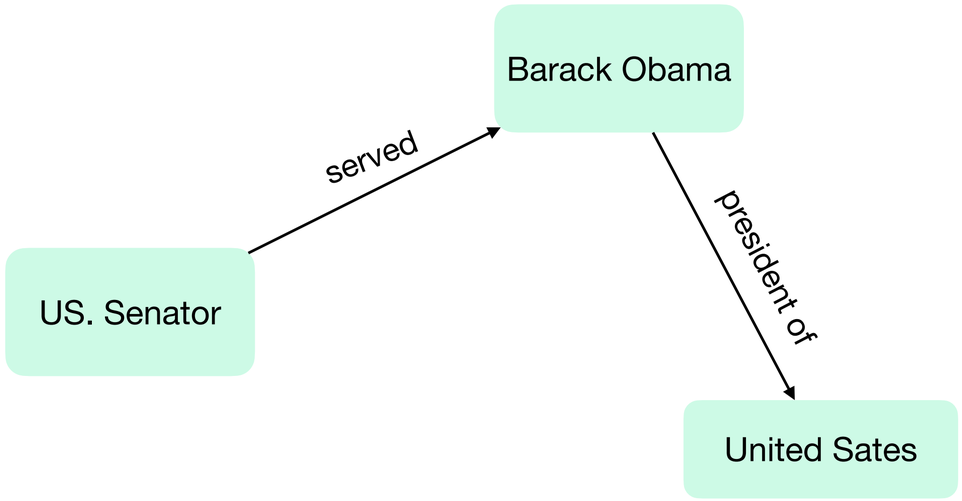
Google Freebase
>95% incomplete Person-Birth-place
>80% incomplete Person-Nationality

Performance:
The entity/KG-augmented Models are still lagging behind text-augmented models

Model	Params	TriviaQA
T5-3B	3B	42.3
Entity Memory	110M	43.2
KG Memory	110M	-
RAG (Text)	620M	55.8
FiD (Text)	440M	65.5

KG vs. Text

Knowledge Graph



1. Highly Interpretable
2. Highly Atomic
3. Coarse-Grained Relation
4. Low Coverage

Textual Passage

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, Obama was the first African-American president of the United States.[3] He previously served as a U.S. senator from Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.

Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, he enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004.

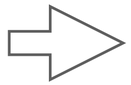
1. Low Interpretability
2. Non-Atomic
3. Fine-Grained Description
4. High Coverage

Q-A as Virtual Knowledge Graph

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, Obama was the first African-American president of the United States.[3] He previously served as a U.S. senator from Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.

Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, he enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004.

Question Generation
(Lewis et al. 2021)



Q: What number president of the United States?

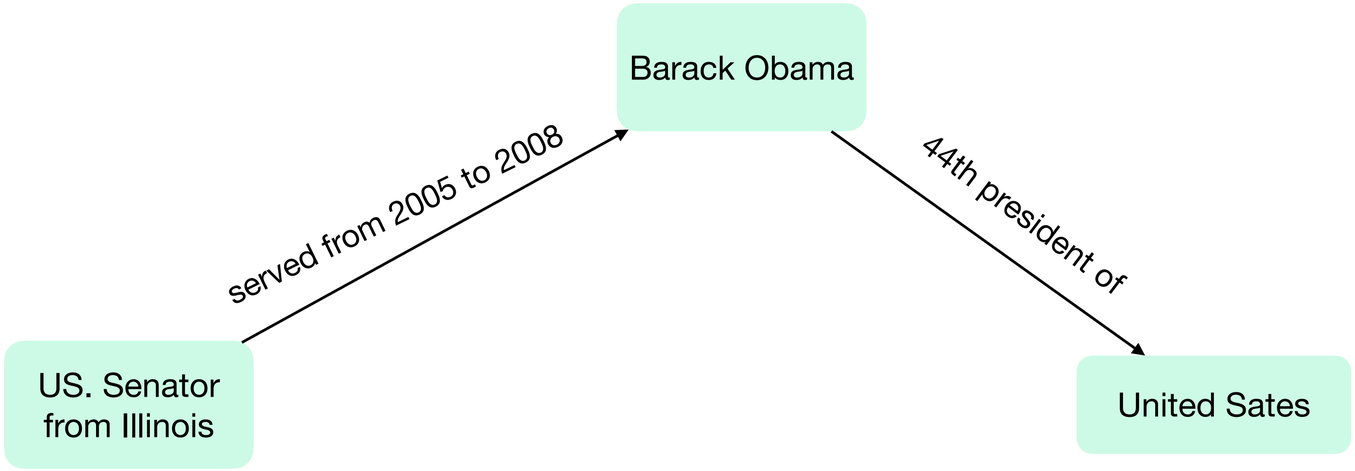
A: 44th

Q: Which organization did Barack Obama serve from 2005-08?

A: U.S. Senator from Illinois

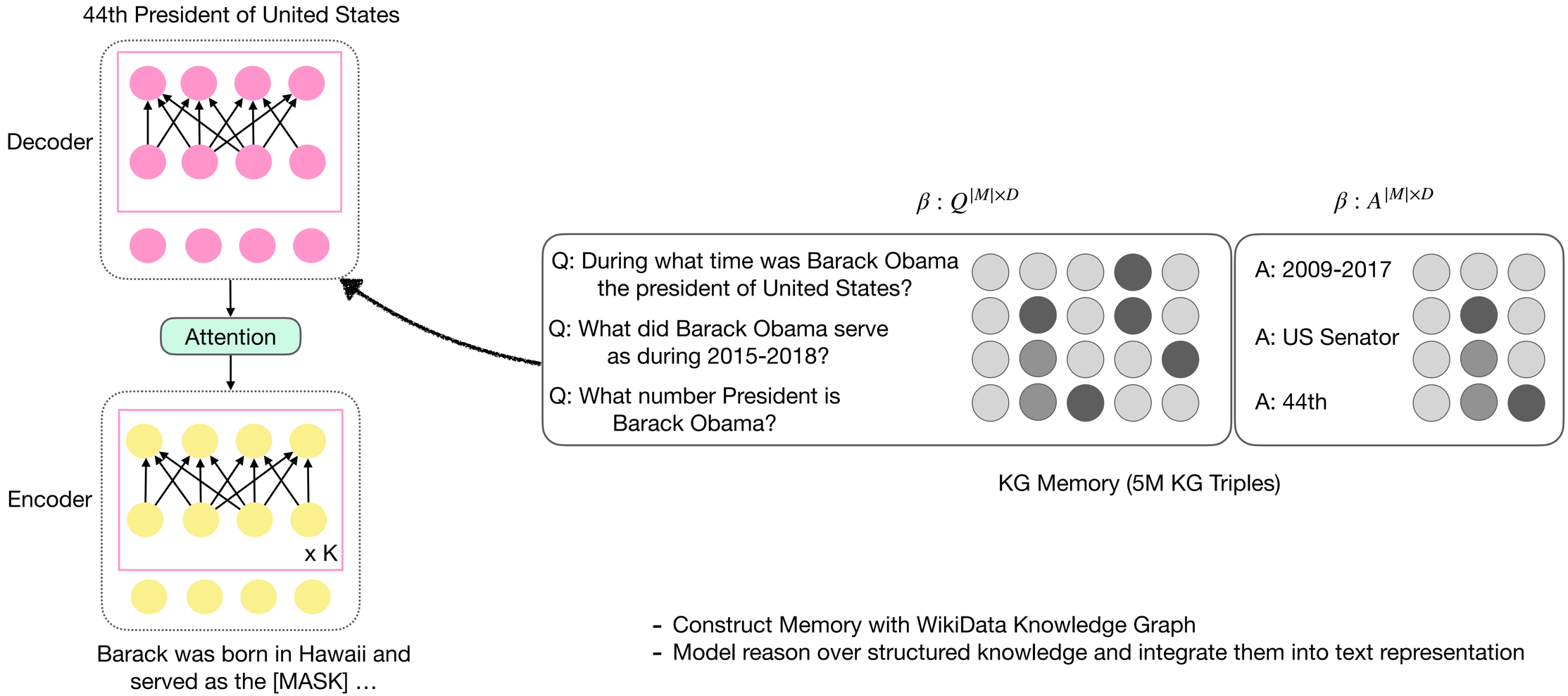
Q: Where was Barack Obama born?

A: Honolulu, Hawaii



1. High Interpretability
2. Atomic
3. Fine-Grained Description
4. Medium Coverage

Big Picture

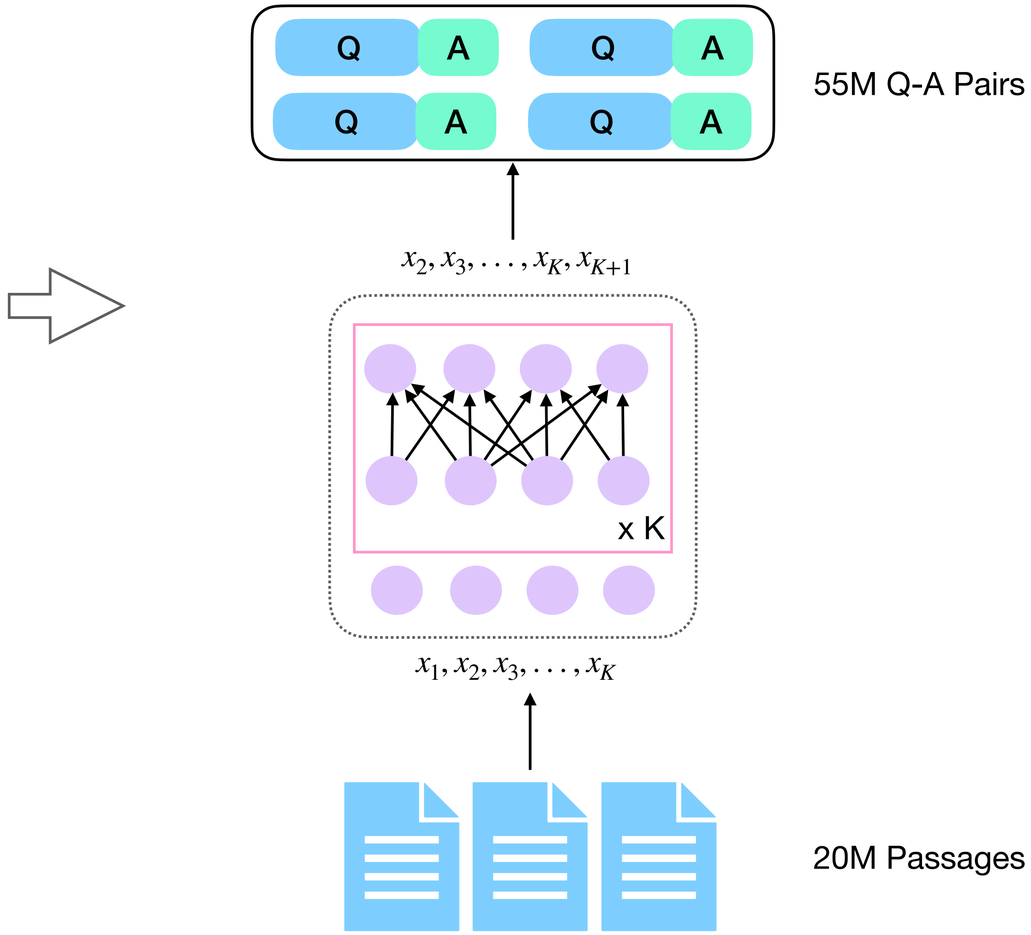


- Construct Memory with WikiData Knowledge Graph
- Model reason over structured knowledge and integrate them into text representation

Construct Dataset

Supervised Question Generation

1. Natural Questions (Kwiatkowski et al. 2019)
2. SQuAD (Rajpurkar et al. 2017)
3. TriviaQA (Joshi et al. 2017)
4. WebQuestions (Berant et al. 2013)
5. HotpotQA (Yang et al. 2019)

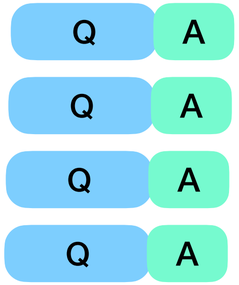


Notation and Objective

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States [MASK1]. A member of the Democratic Party, Obama was the first African-American president of the United States. He previously served as a [MASK2] from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.

$X : Corrupted - Passage$

Memory: 55 Million



$m_i \in M : R^{|M| \times D}$

[MASK1] from 2009-2017 [MASK2] US Senator from Illinois

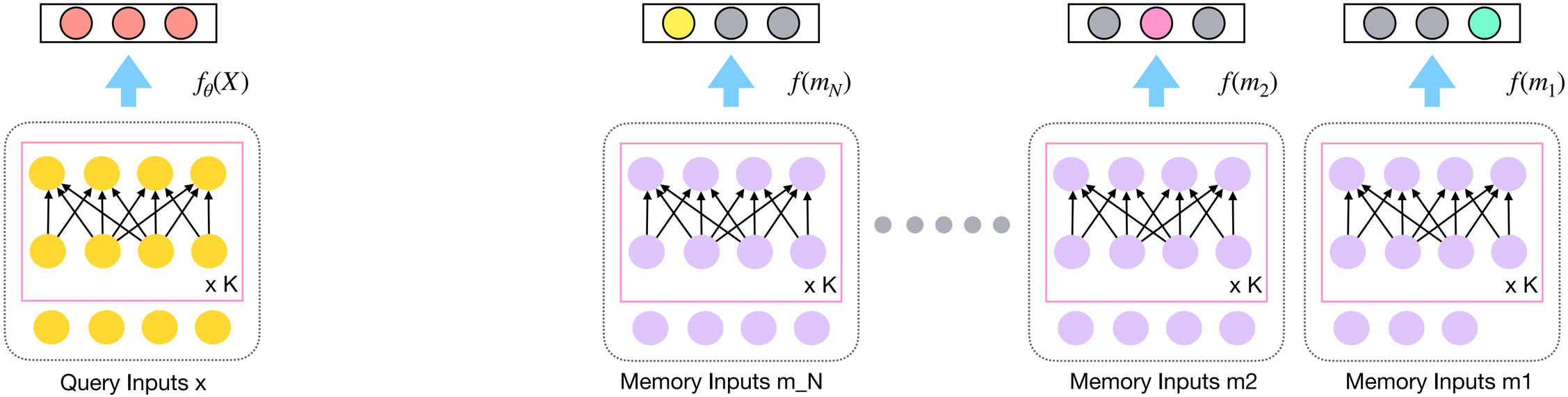
$Y : CorruptedSpan$

LM Training Objective:

$$P(Y|X) = \sum_{m_i \in M} P(Y|X, m_i) \approx \sum_{m_i \in \bar{M}} P(Y|X, m_i)$$

$\bar{M}$ is the top-K subset of the whole memory

Basics: Encoder



Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States [MASK1]. A member of the Democratic Party, Obama was the first African-American president of the United States. He previously served as a [MASK2] from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.

Q: During what time was Barack Obama the president of US?

A: 2009-2017

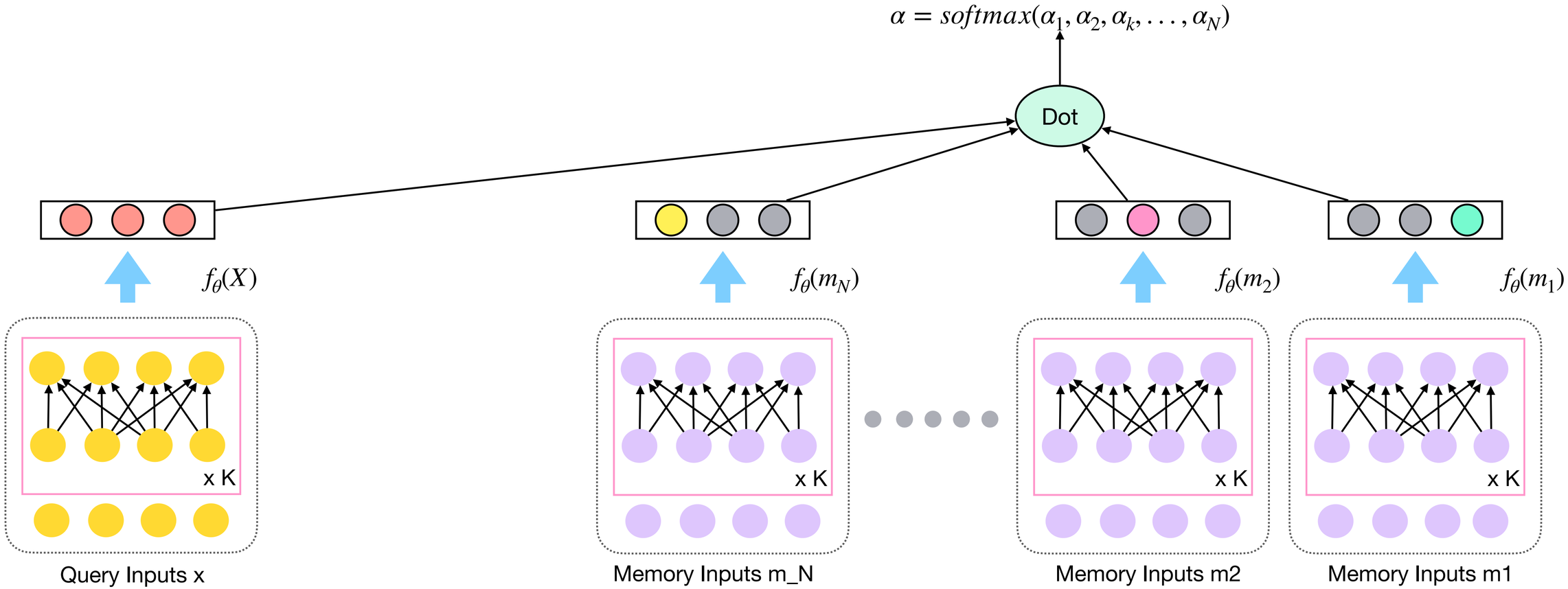
Q: Which organization did Barack Obama serve from 2005-08?

A: U.S. Senator from Illinois

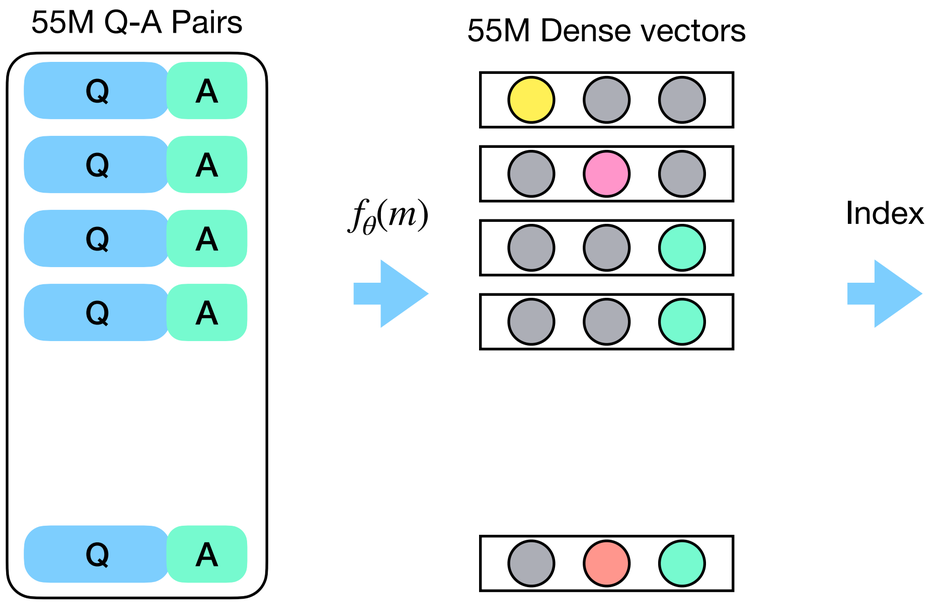
Q: Which district or state did Barack Obama serve from 2005-08?

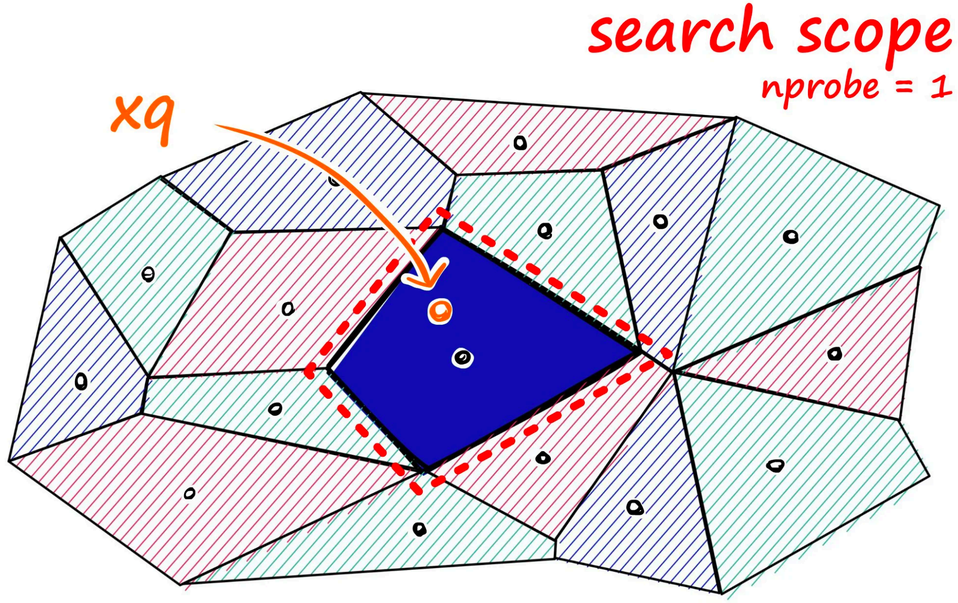
A: Illinois

Basics: Retriever



Basics: Maximum Inner Product Search



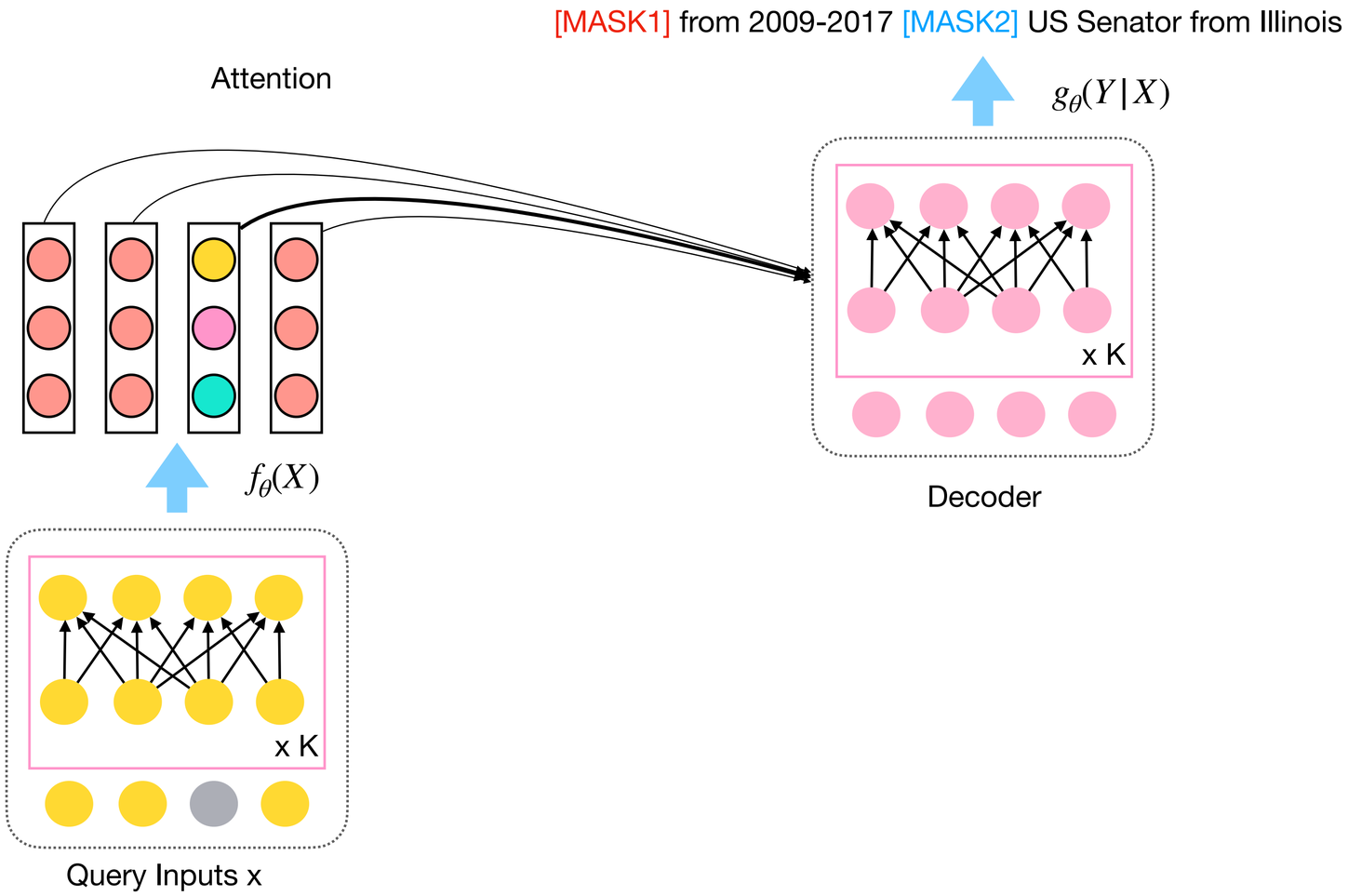


Clustering + Hierarchical Search + Quantization

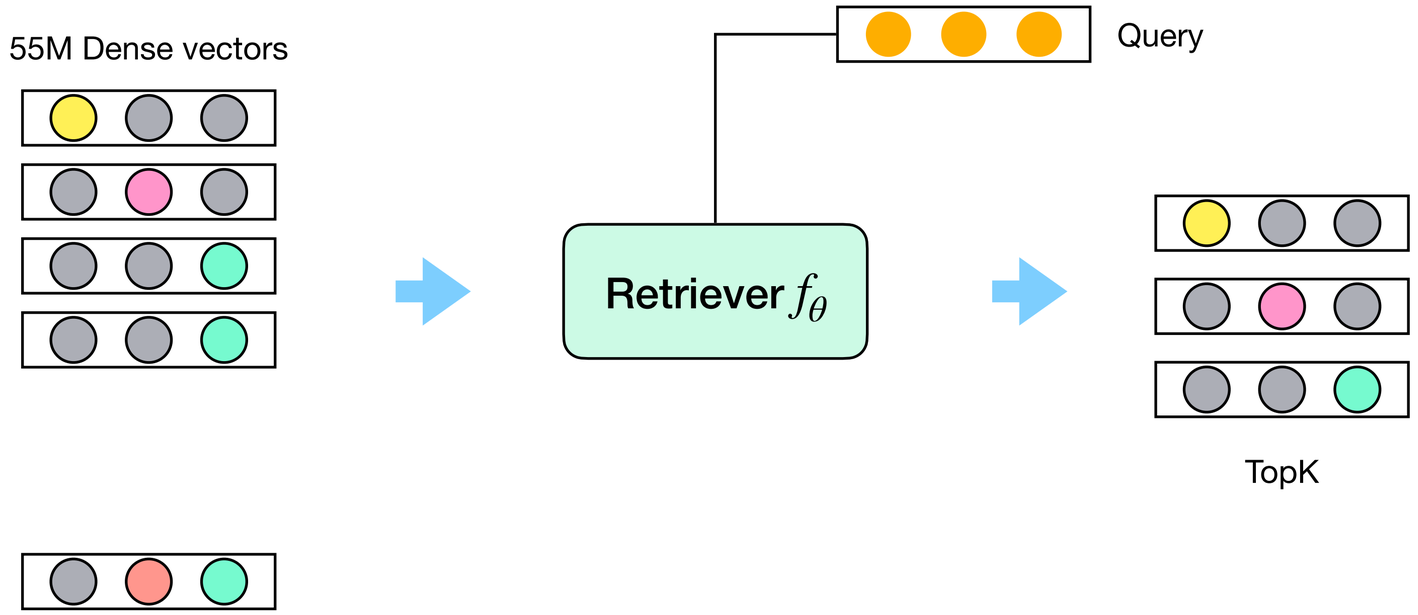
Time Complexity: Approximately O(log(|M|))

https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/

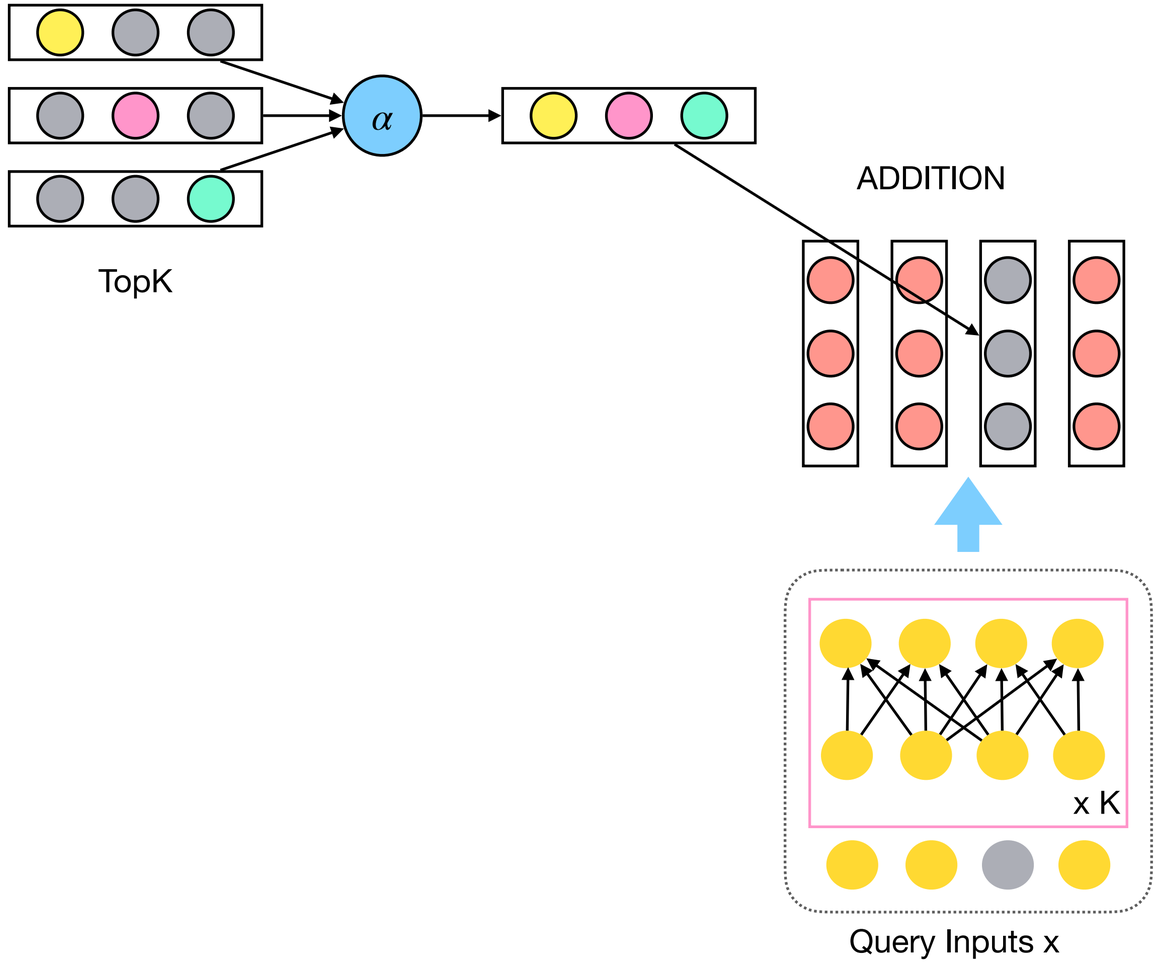
Basics: Decoder



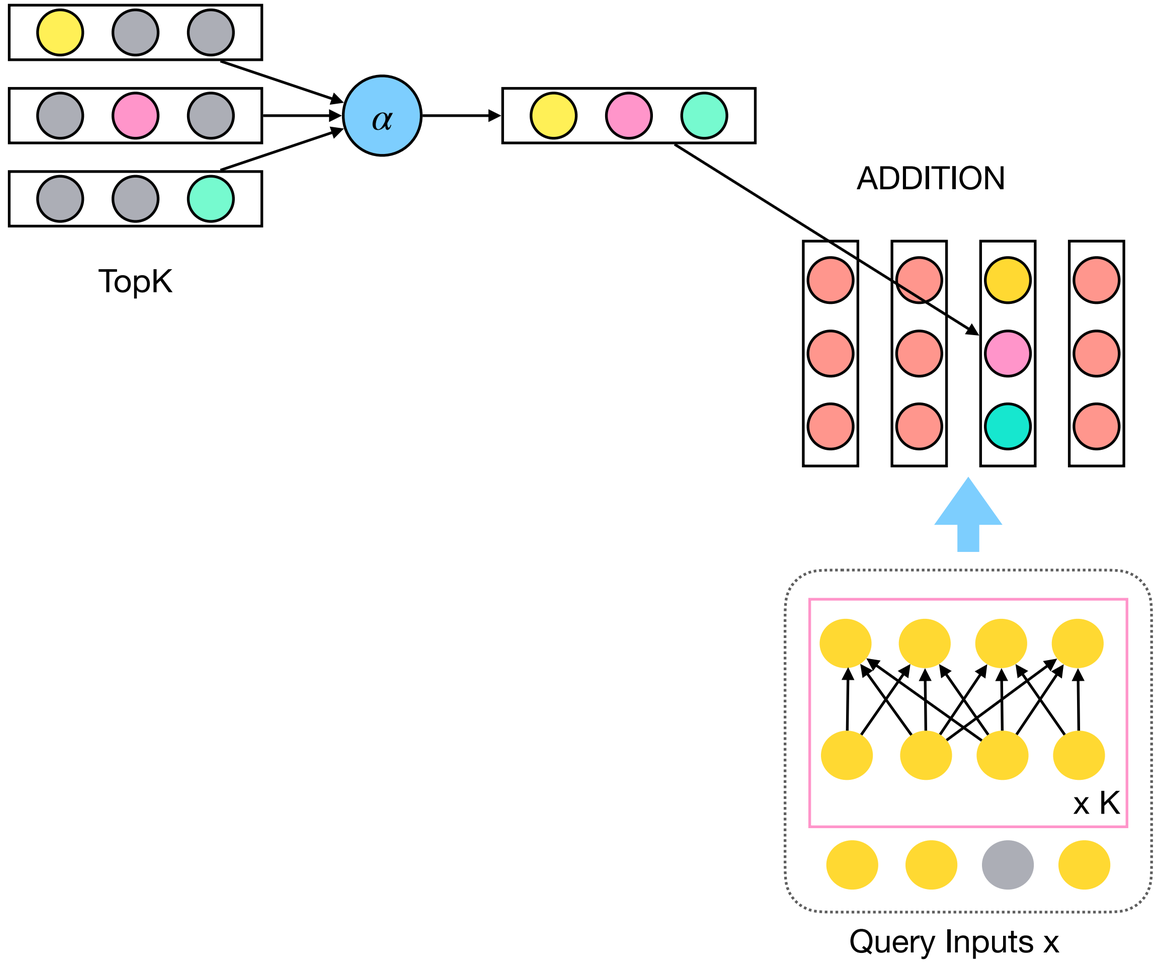
Workflow: Retrieval Top-K



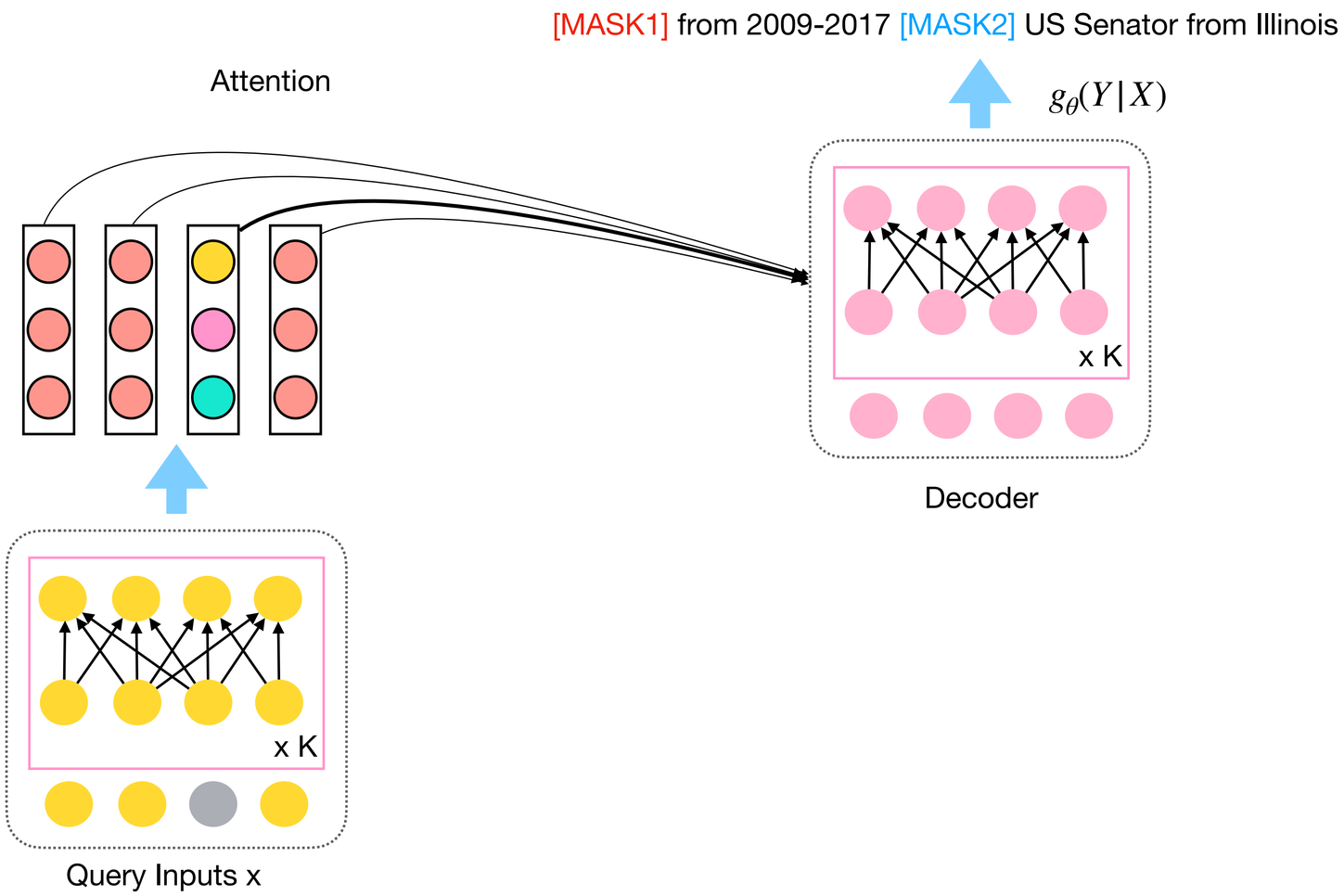
Workflow: Weighted-Sum Aggregation



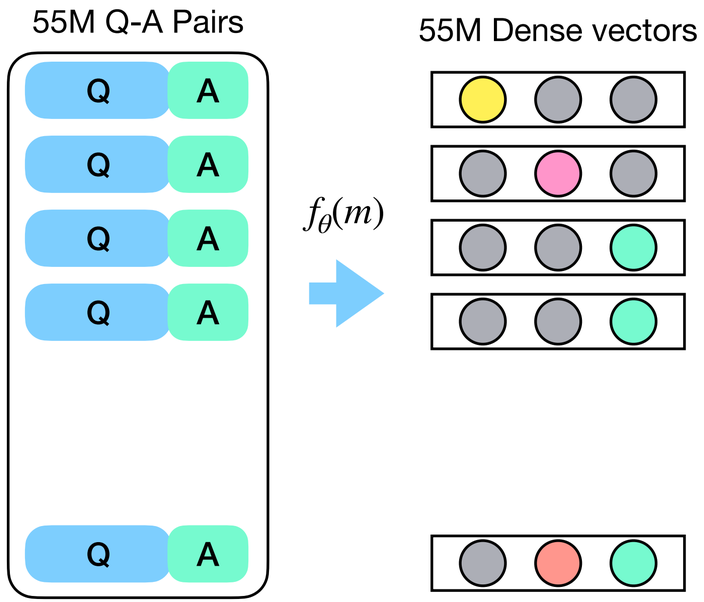
Workflow: Addition



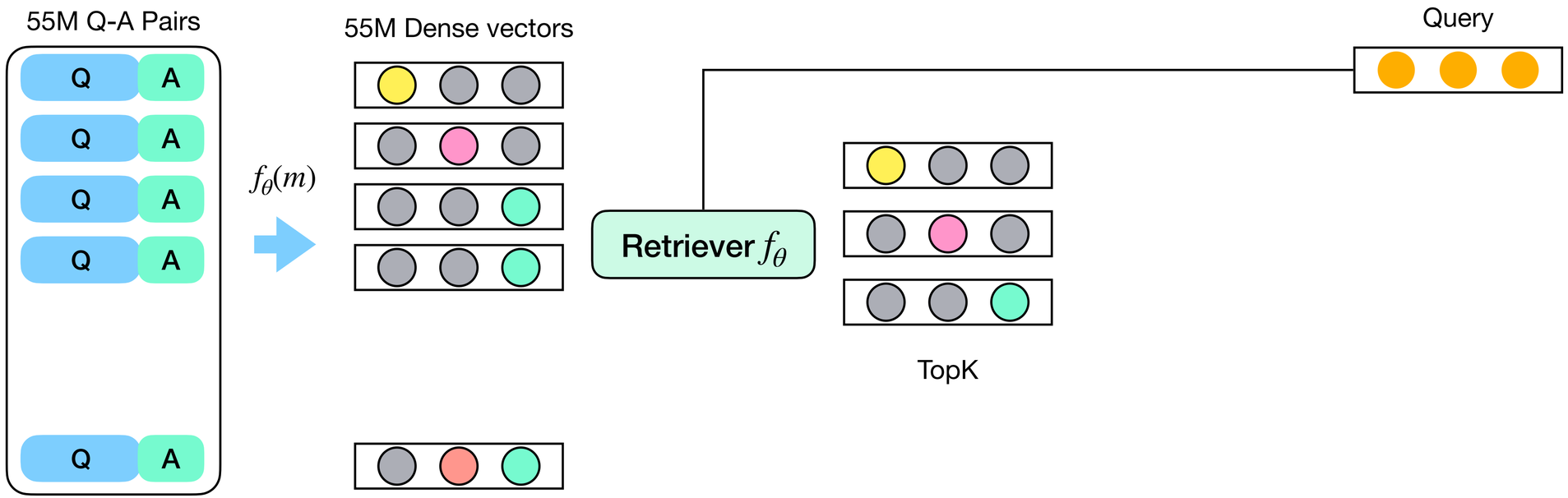
Workflow: Decoding



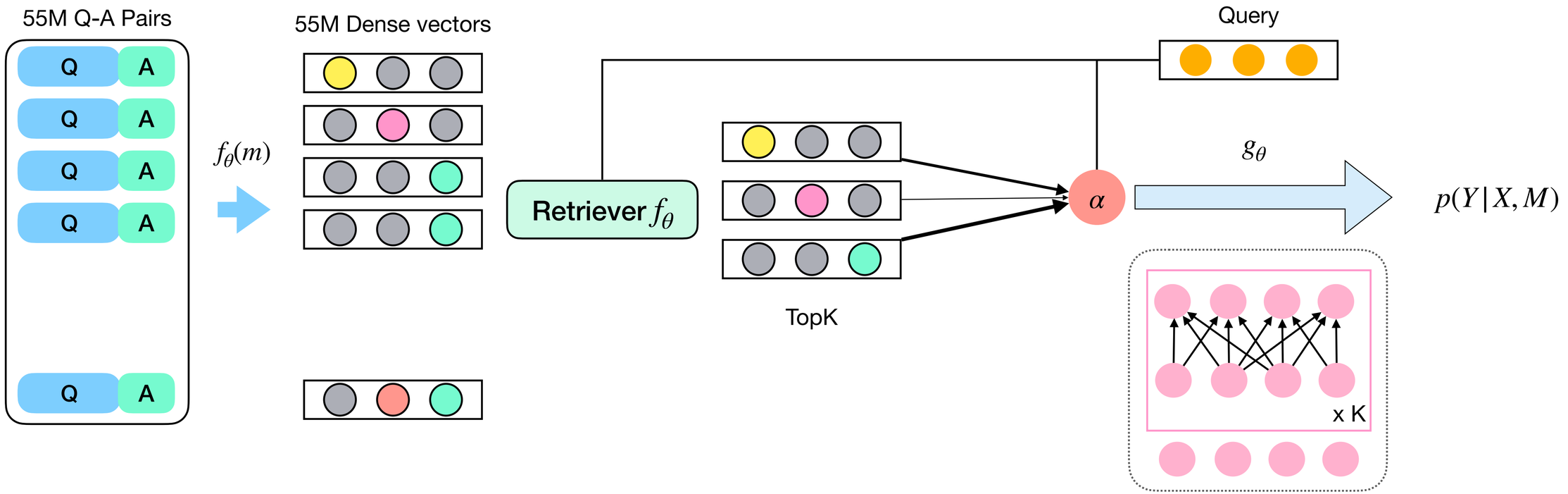
Pre-training: Computation Challenges



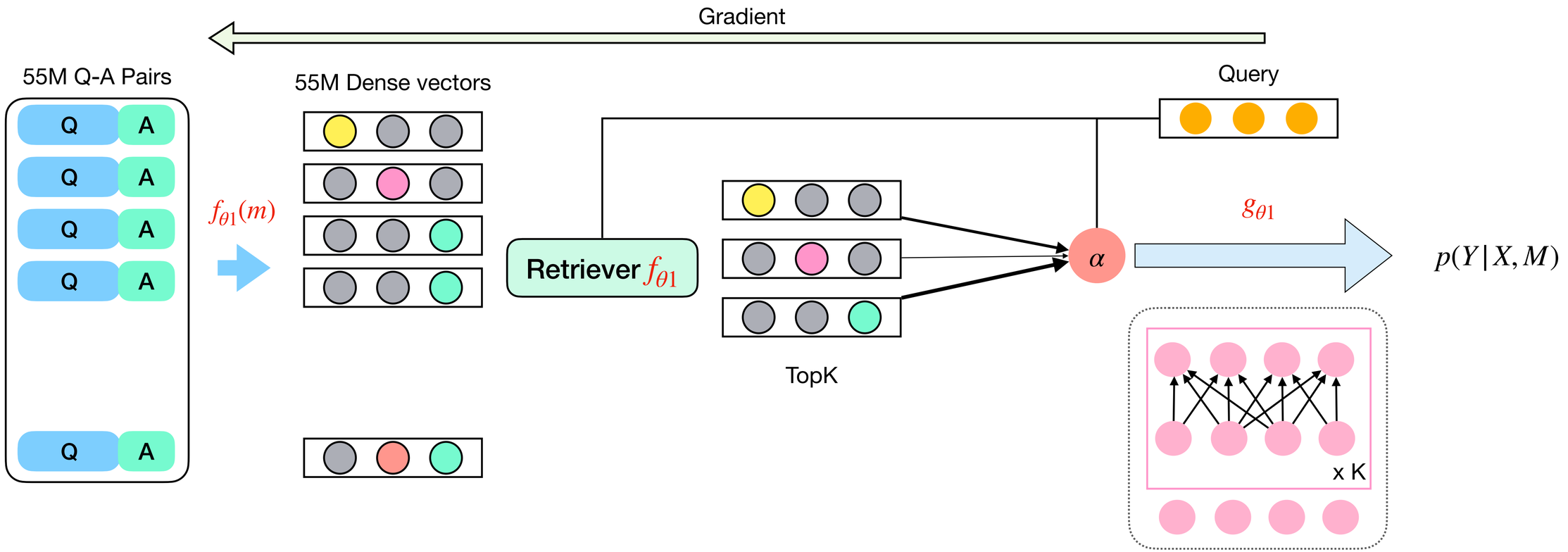
Pre-training: Computation Challenges



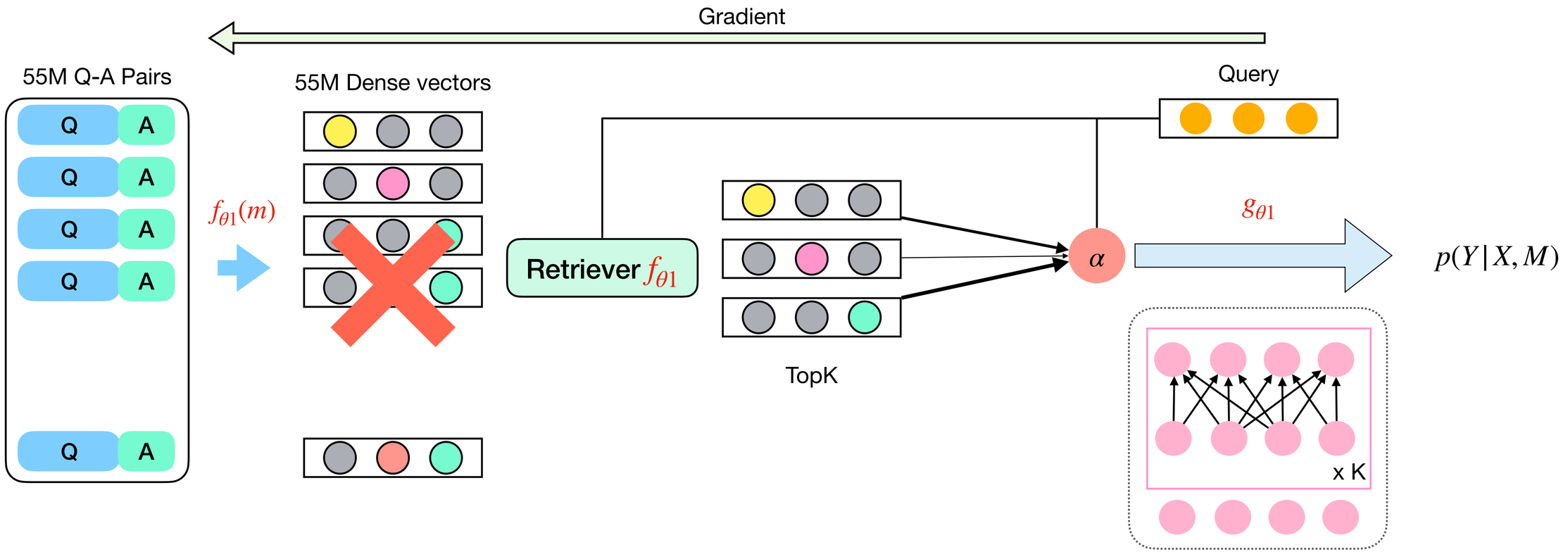
Pre-training: Computation Challenges



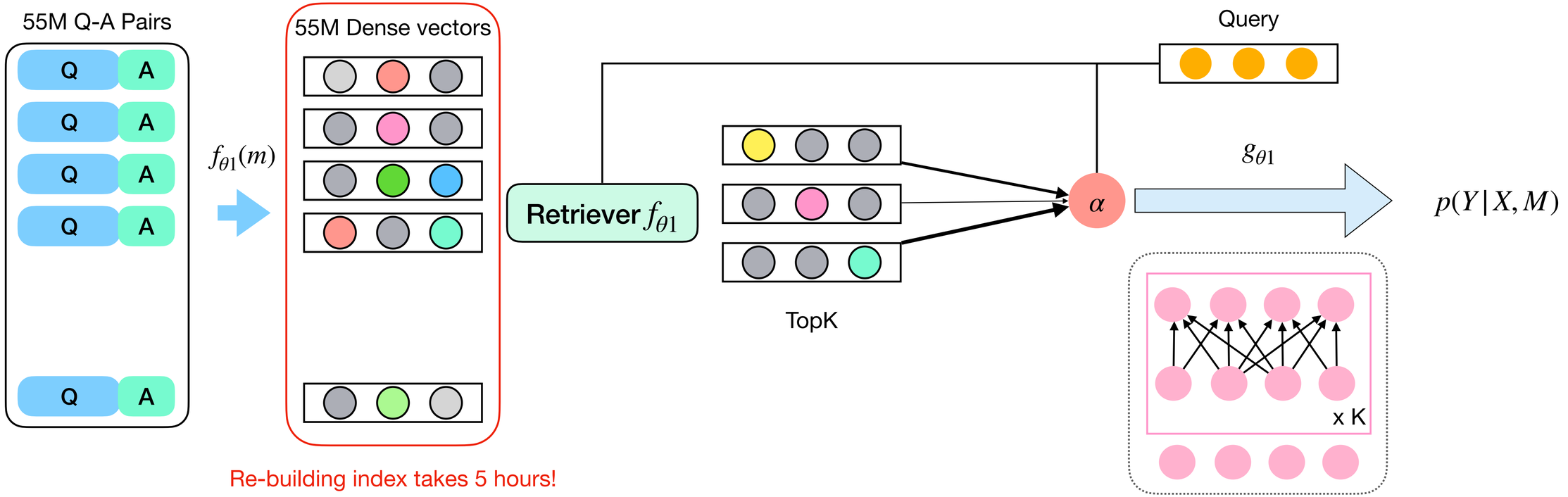
Pre-training: Computation Challenges



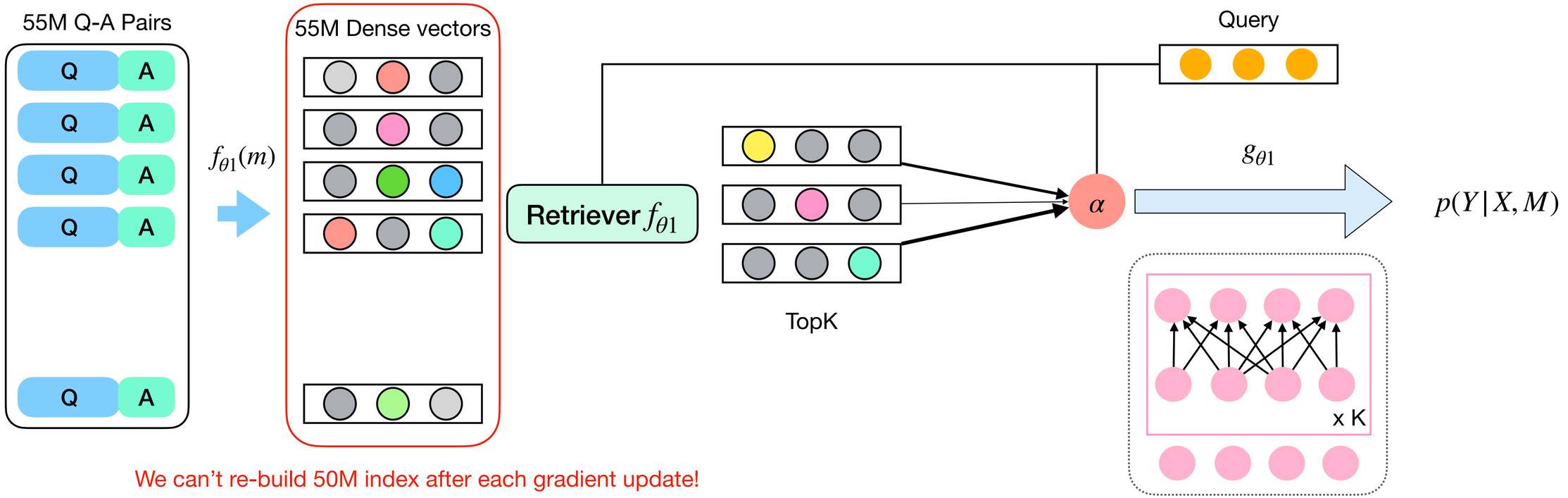
Pre-training: Computation Challenges



Pre-training: Computation Challenges

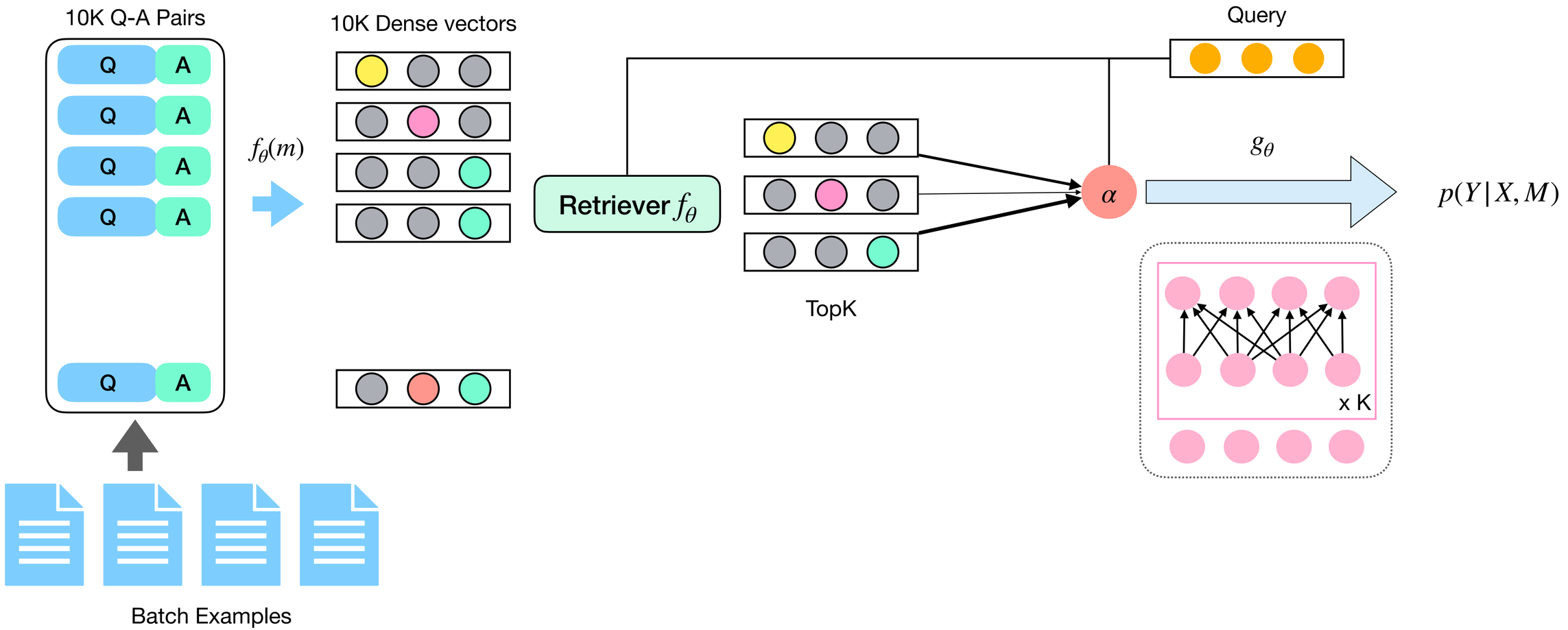


Pre-training: Computation Challenges

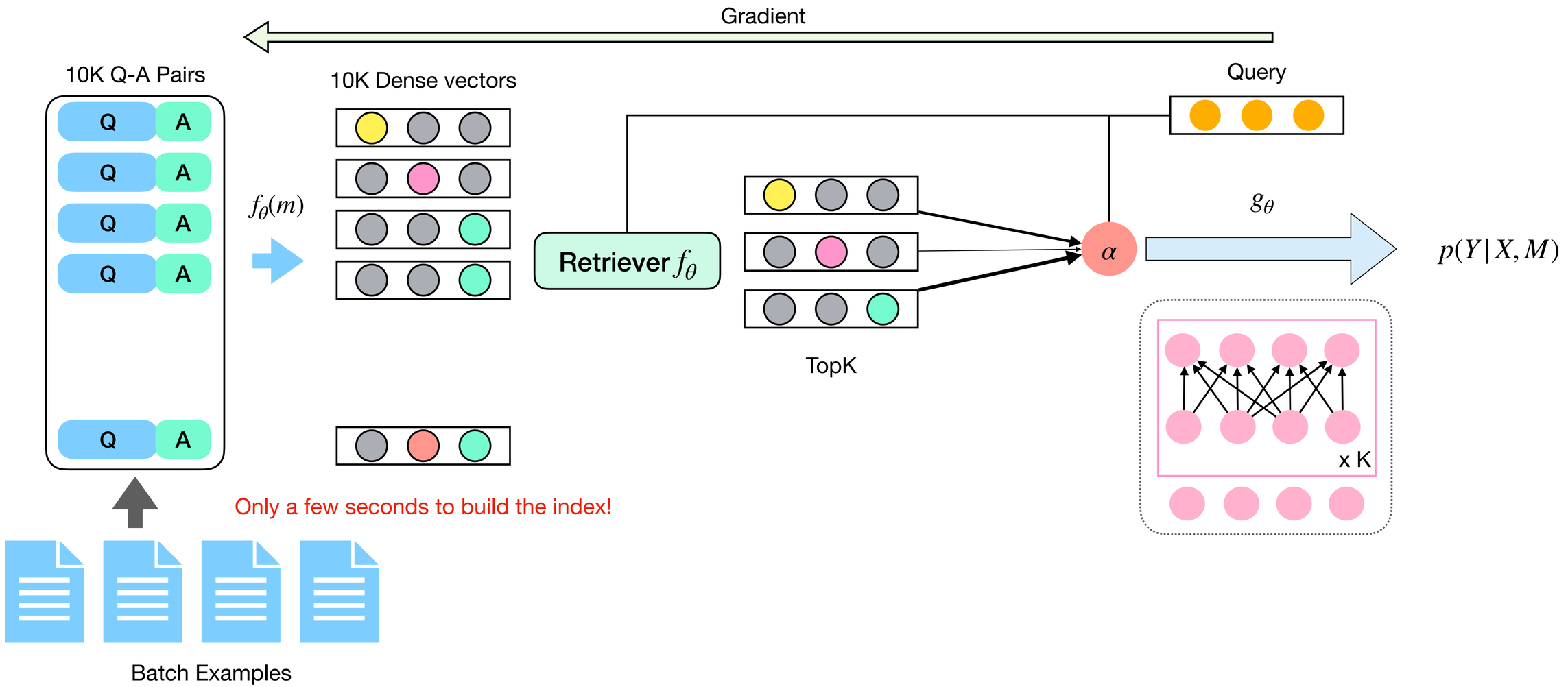


We can’t re-build 50M index after each gradient update!

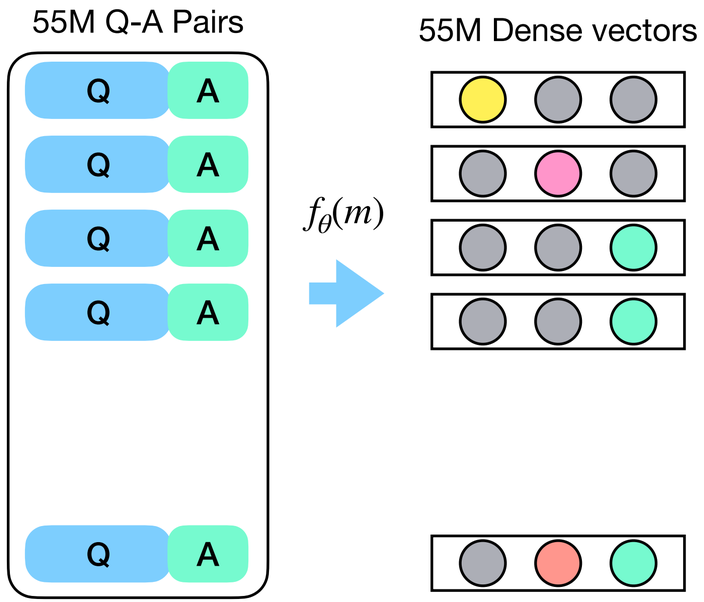
Pre-training: In-Batch Training



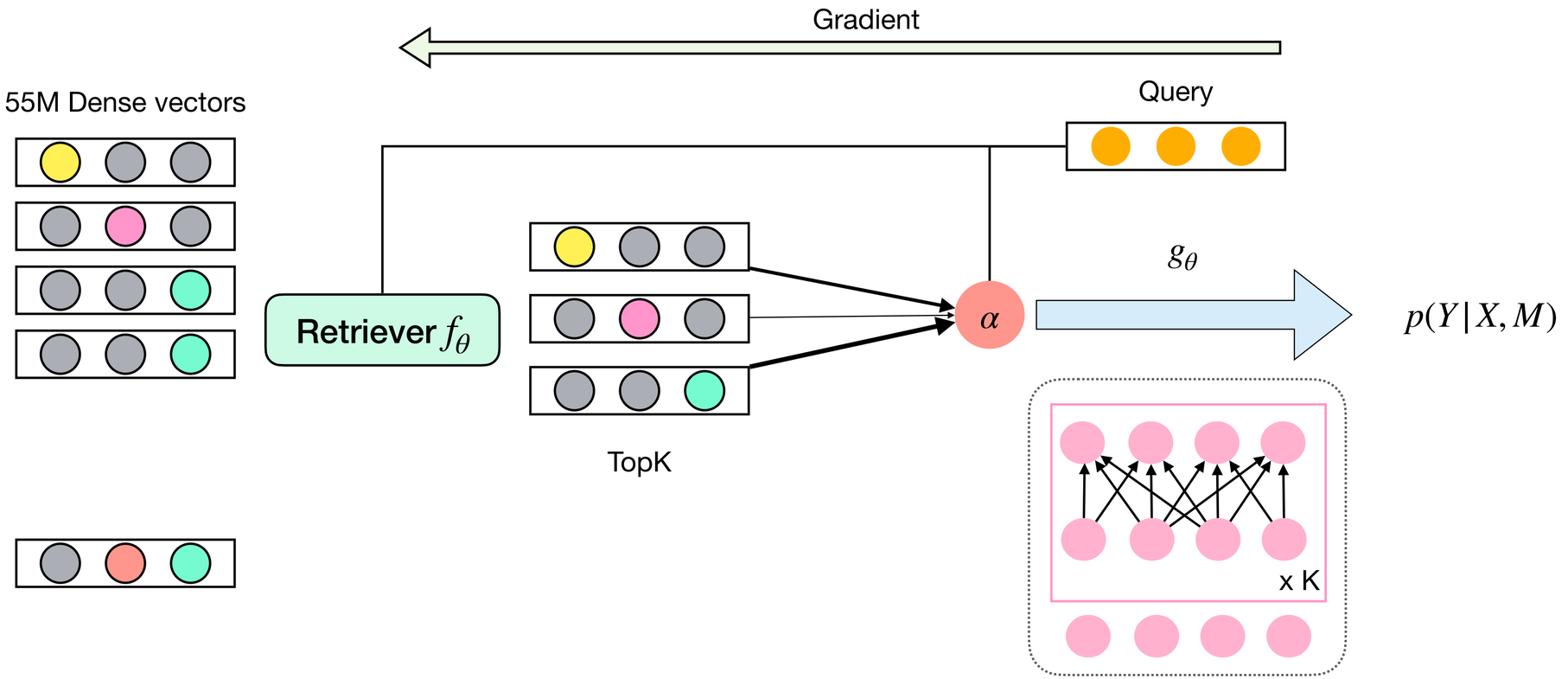
Pre-training: In-Batch Training



Pre-training: Global Training

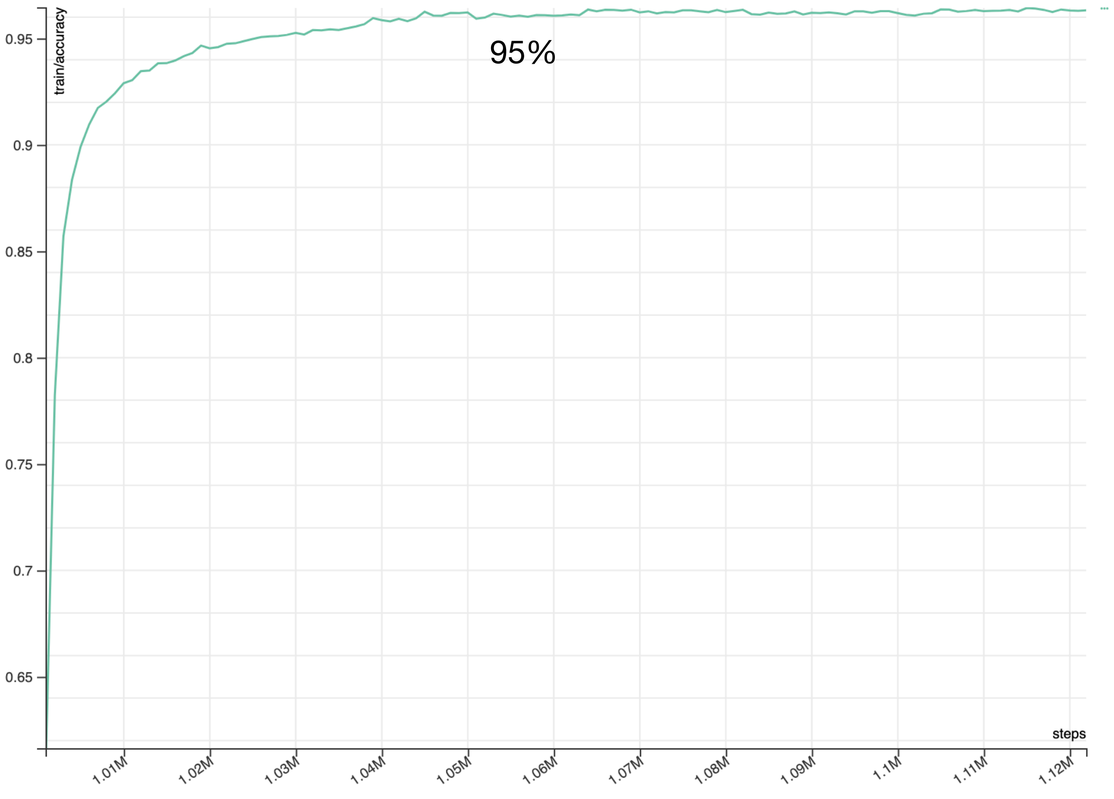


Pre-training: Global Training

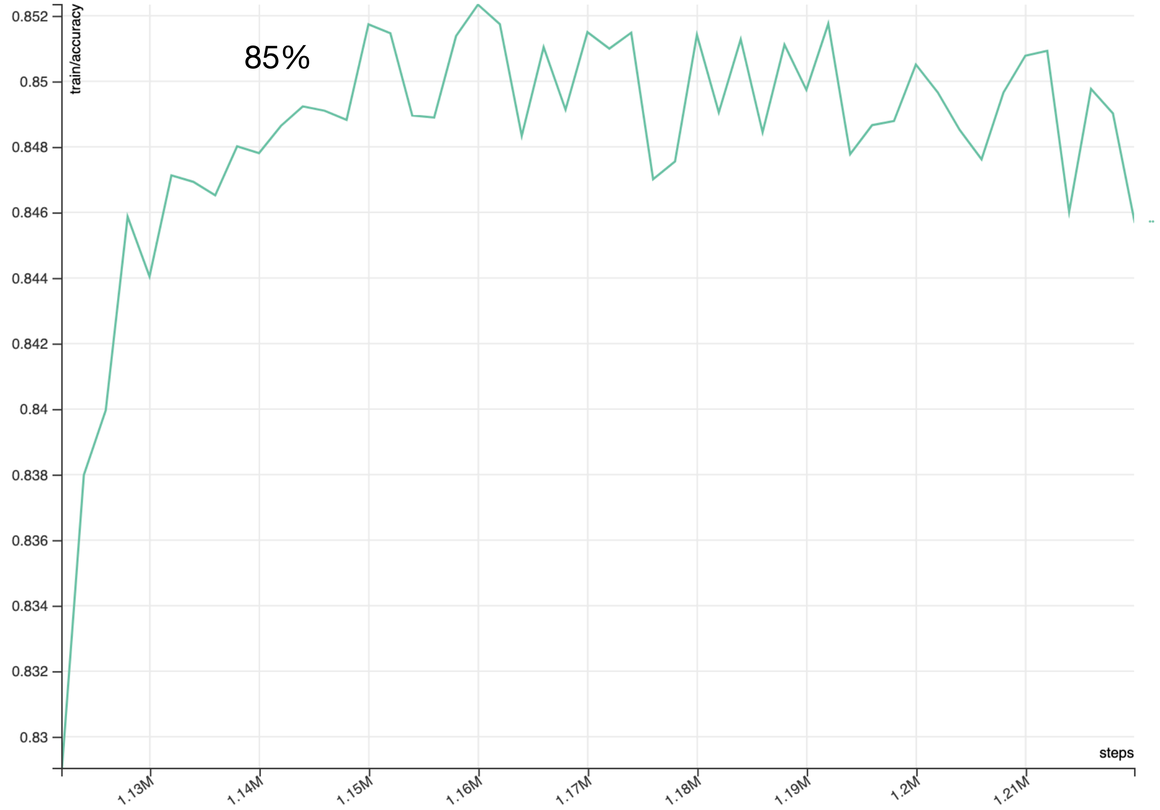


Pre-training Accuracy

Token-Level Accuracy: In-Batch Initialization



Token-Level Accuracy: Global Tuning



Downstream Fine-tuning

Model	Params	TriviaQA	WebQuestion	NaturalQA
T5-3B	3B	42.3	37.4	30.4
Entity Memory	110M	43.2	39.0	-
KG Memory	110M	-	-	-
QA Memory	220M	53.2	43.2	44.5
QA Memory (Large)	770M	54.8	43.8	45.5
RePAQ	770M	50.7	37.6	47.7
RAG (Text)	620M	56.8	45	44.5

Compositional Reasoning

Q: Who founded the party which Barack Obama served in?

Barack Hussein Obama II (born August 4, 1961) is an American politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, Obama was the first African-American president of the United States.[3] He previously served as a U.S. senator from Illinois from 2005 to 2008 and as an Illinois state senator from 1997 to 2004.

The Democratic Party is one of the two major contemporary political parties in the United States. It was founded in 1828 by supporters of Andrew Jackson, making it the world's oldest active political party. Since the 1860s, its main political rival has been the Republican Party.

Question Generation (Lewis et al. 2021)

Q: Which party did Barack Obama serve?

A: Democratic Party

Q: Who was the first African-American President?

A: Barack Obama

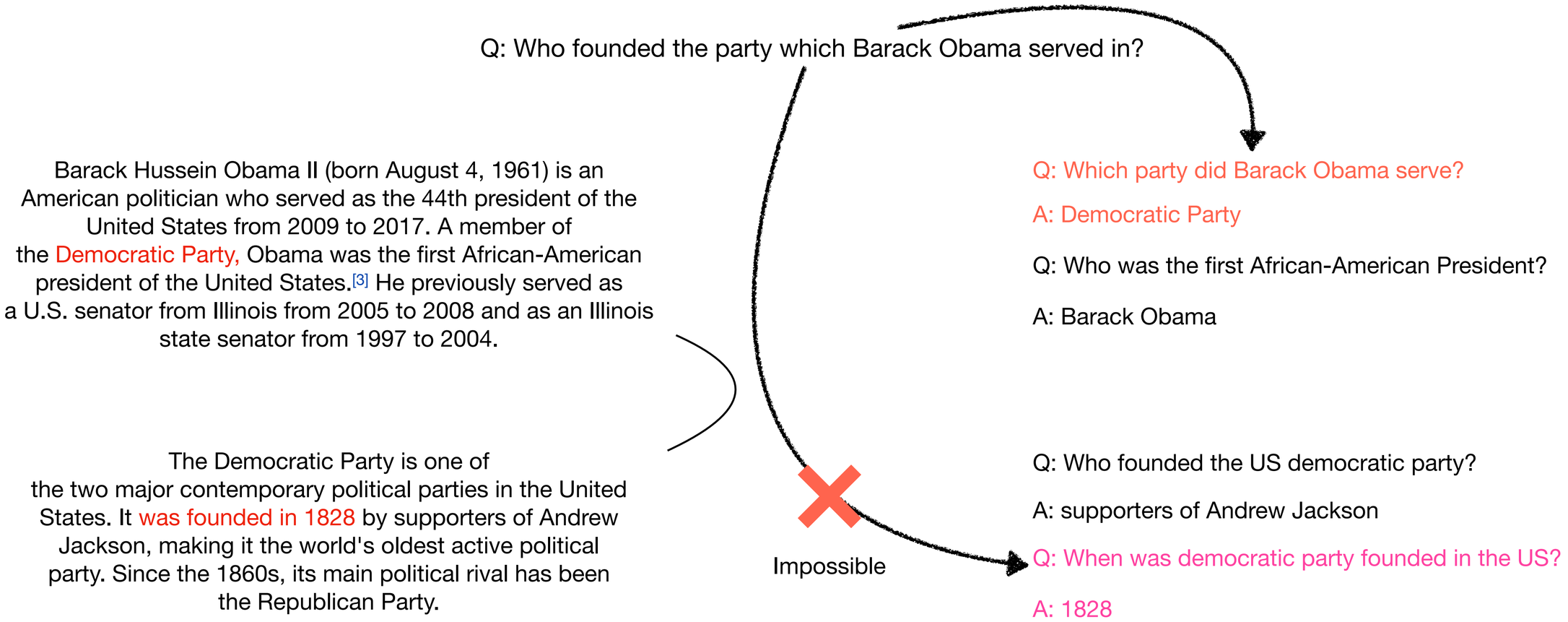
Q: Who founded the US democratic party?

A: supporters of Andrew Jackson

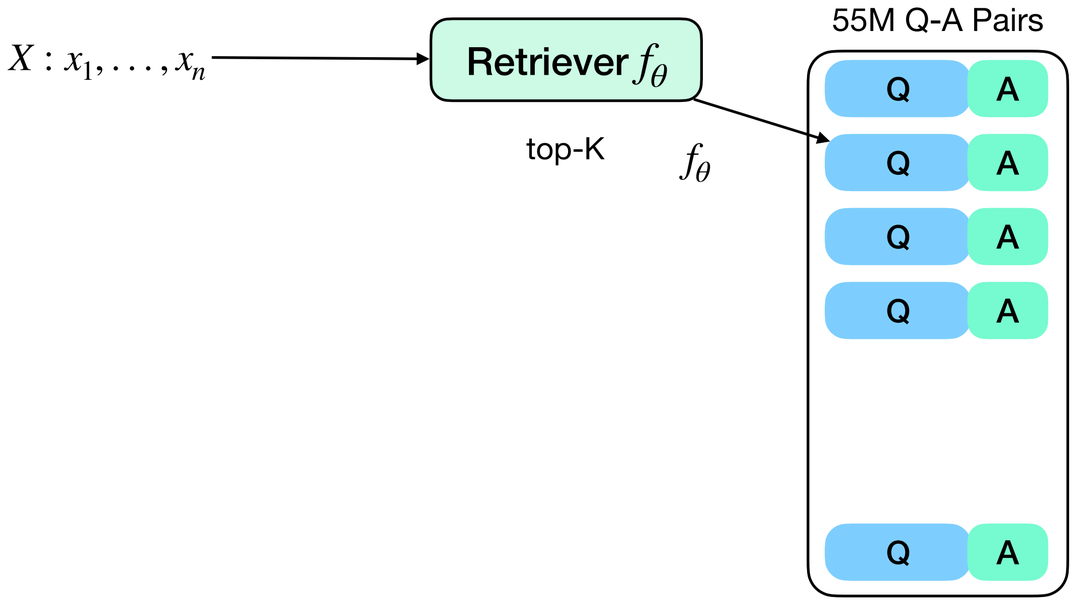
Q: When was democratic party founded in the US?

A: 1828

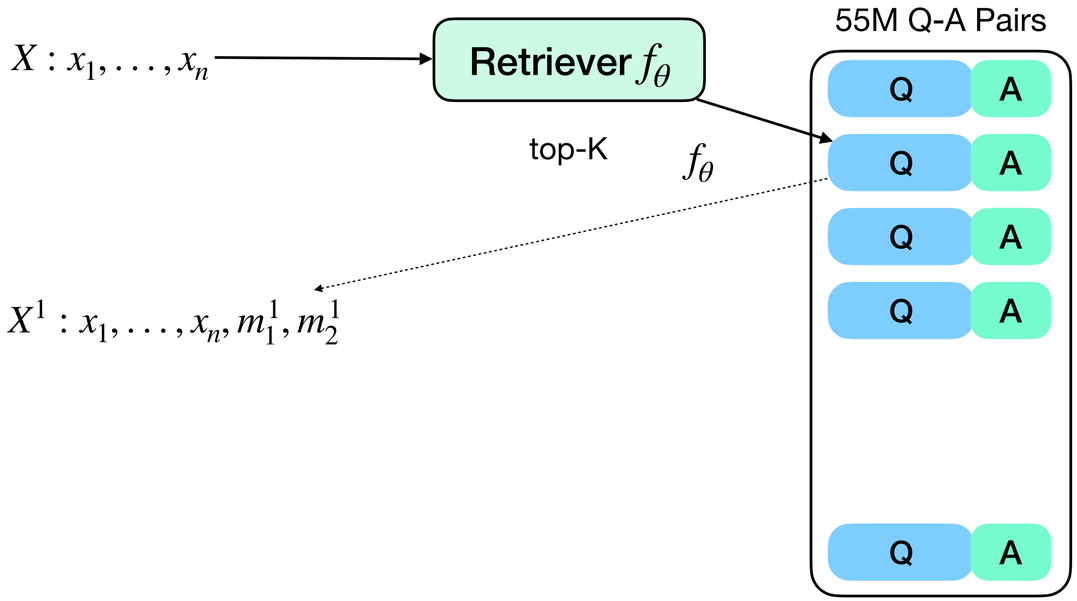
Compositional Reasoning



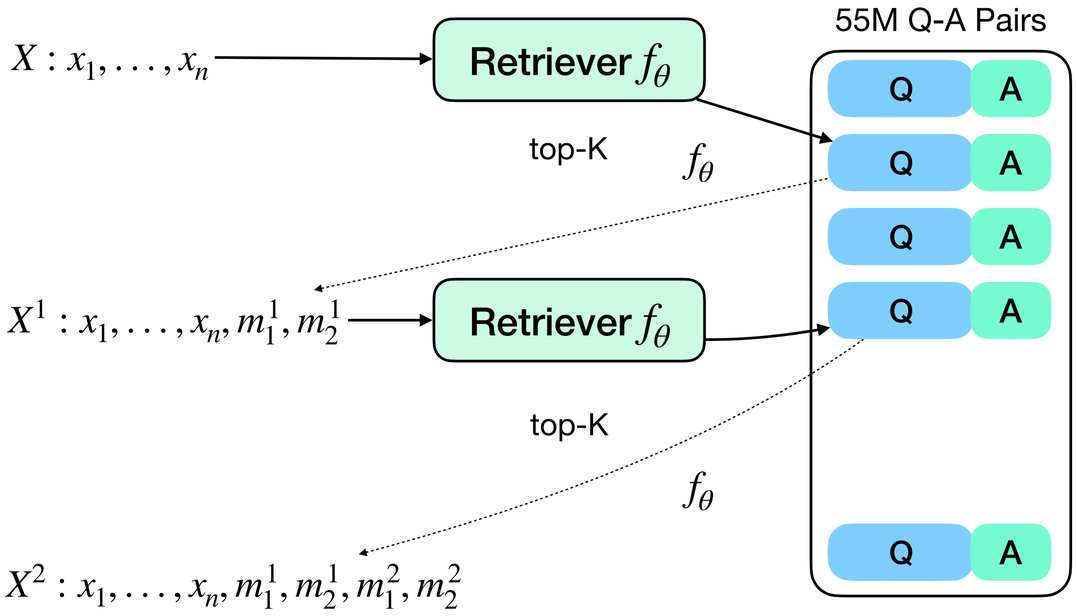
Stacked Version



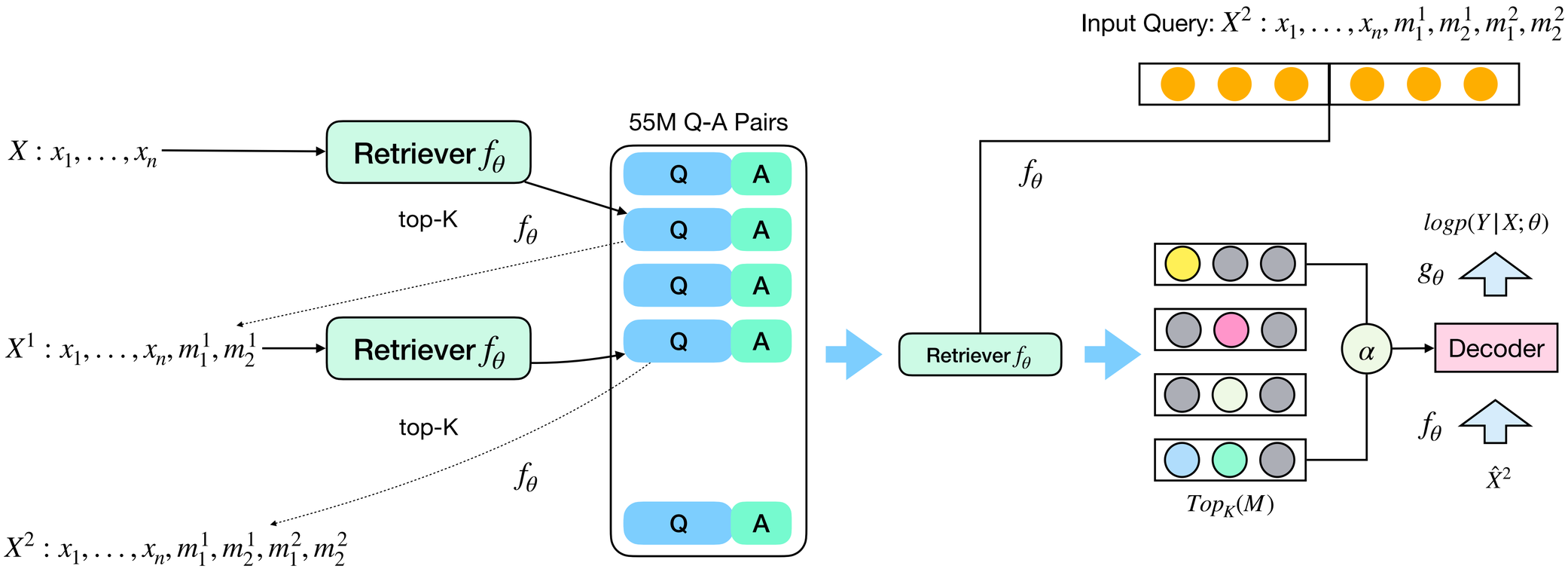
Stacked Version



Stacked Version



Stacked Version



Experimental Results

Model	HotpotQA-EM	HotpotQA-F1
Stacked QA Memory	45%	57.1%
Cognitive Graph (Ding et al. 2019)	37.6%	49.4%
DecompRC (Min et al. 2019)	-	43.3%
Semantic Retriever (Nie et al. 2019)	46.5%	58.8%
Multi-Step Retriever - RAG 16 PSG (Xiong et al. 2021)	51.2%	63.9%

Conclusion

Semi-parametric Language Model

- We are able to prove that using semi-parameters to represent knowledge can make model more efficient and decrease the computation/memory cost.

- We are still exploring what's the best unit to represent world knowledge, the best representation is still passage (non-atomic and barely interpretable).

- The current framework only prove effective on several knowledge-intensive tasks like QA or Fact-Verification, not able to see gains on more broader spectrum of NLP tasks.


Q&A