COMP 4446 / 5046 Lecture 9: Training – Unsupervised

Jonathan K. Kummerfeld

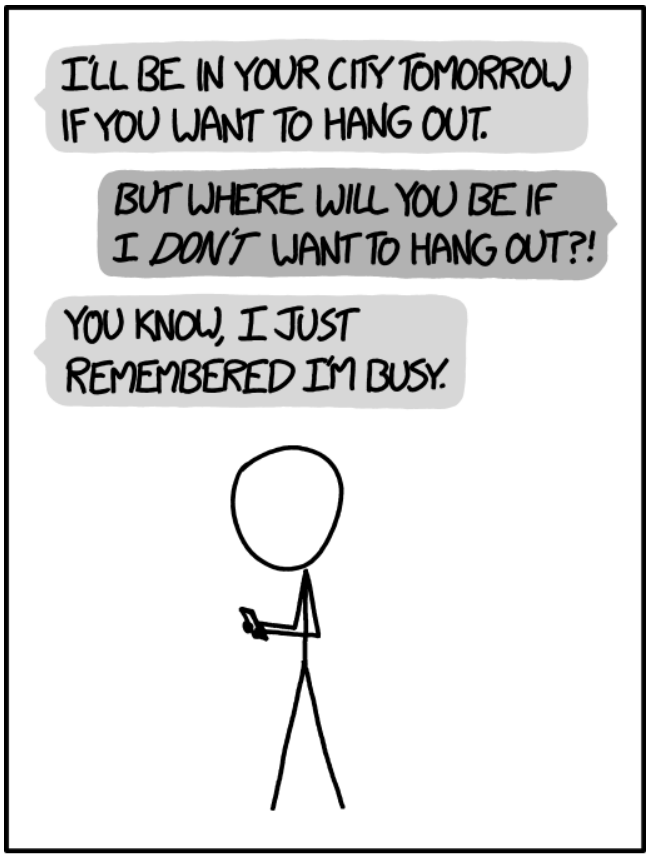
Semester 1, 2025

THE UNIVERSITY OF SYDNEY

menti.com 3821 2301

["If you're done being pedantic, we should get dinner."
"You did it again!"
"No, I didn't."]

Conditionals



WHY I TRY NOT TO BE PEDANTIC ABOUT CONDITIONALS.

Source: https://xkcd.com/1652/

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Efficiency

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

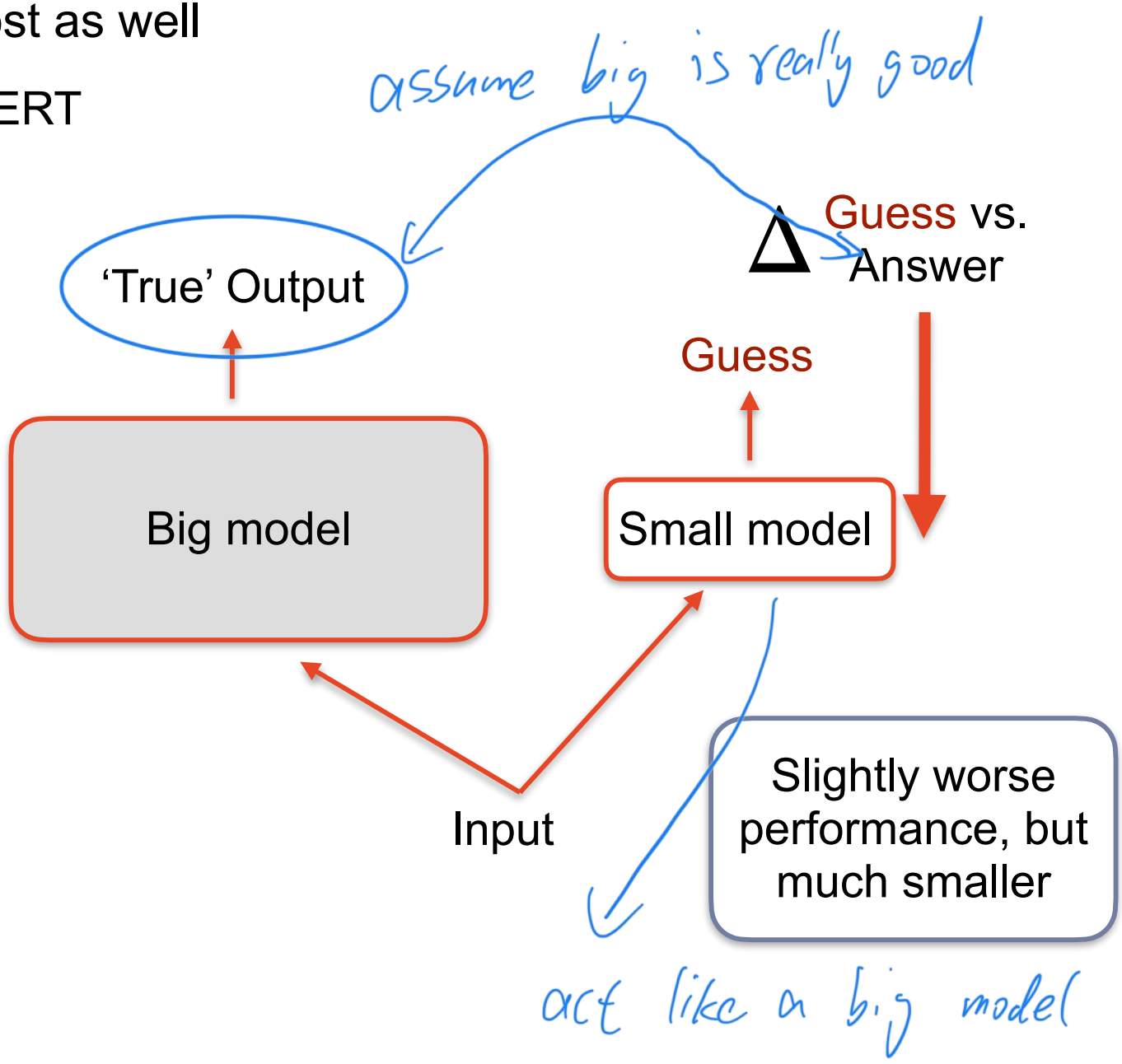
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can use a big model to train a smaller model to do almost as well

DistilBERT



COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

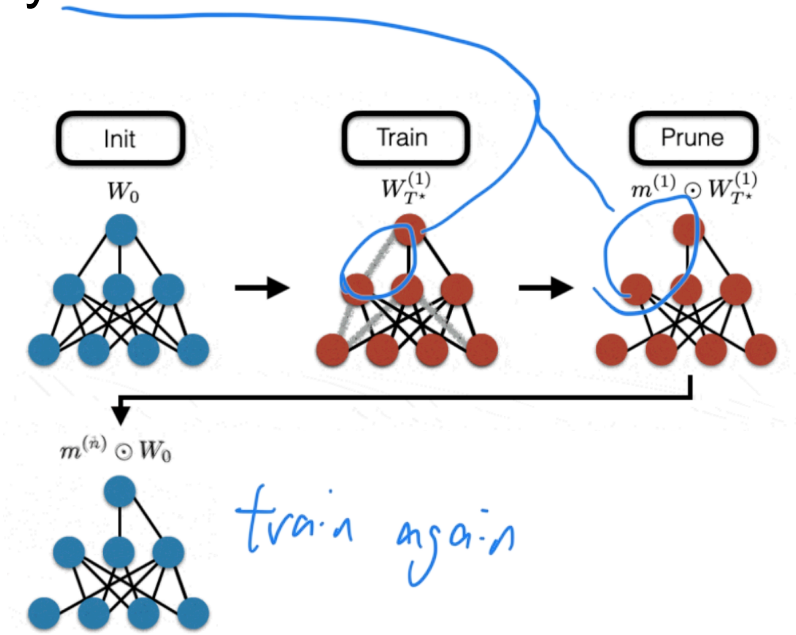
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can prune a big model and still get the same accuracy

Adding sparsity



Berkeley NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

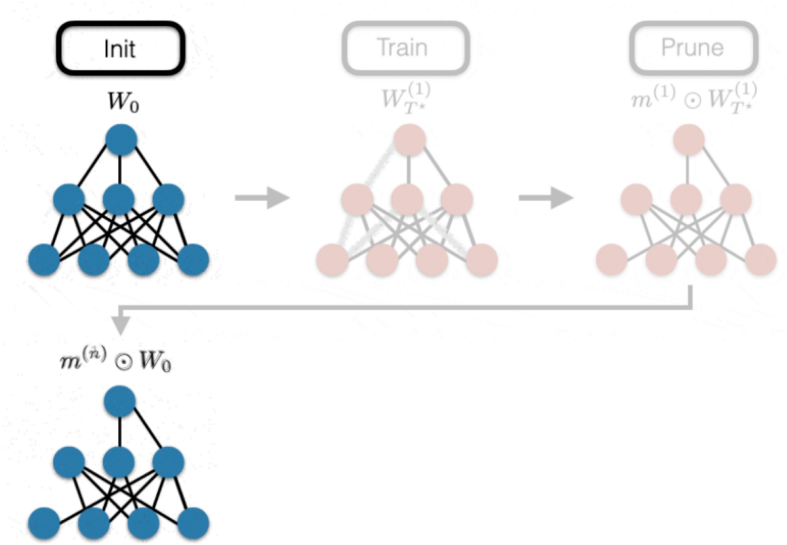
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can prune a big model and still get the same accuracy

Adding sparsity



Berkeley NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

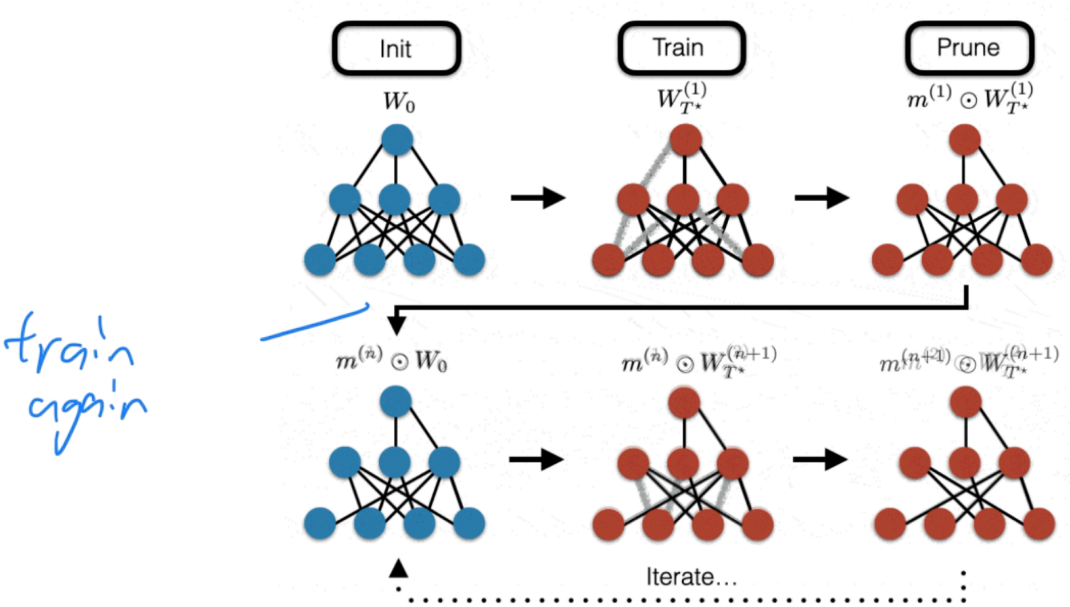
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can prune a big model and still get the same accuracy

Adding sparsity



Not often used in practise because the pruning is sparse, which doesn't help on GPUs

Berkeley NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

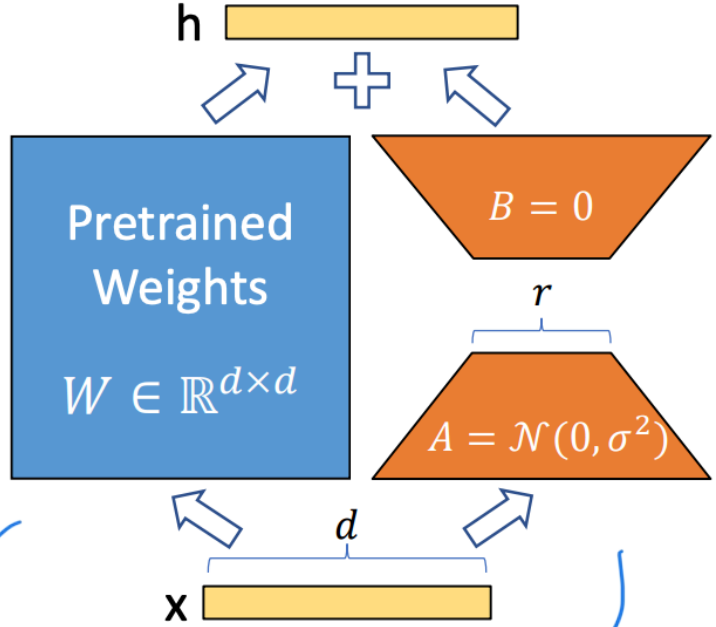
menti.com 2376 2478

We can fine-tune a big model more efficiently

approximation

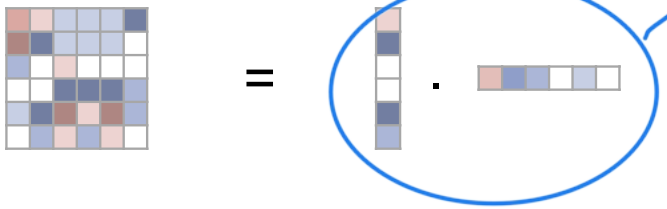
LoRA

reduce weight matrix size, use less memory



then project back to original space

lower dimensional space



This is a form of Parameter-Efficient Fine Tuning (PEFT)

Hu et al. (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

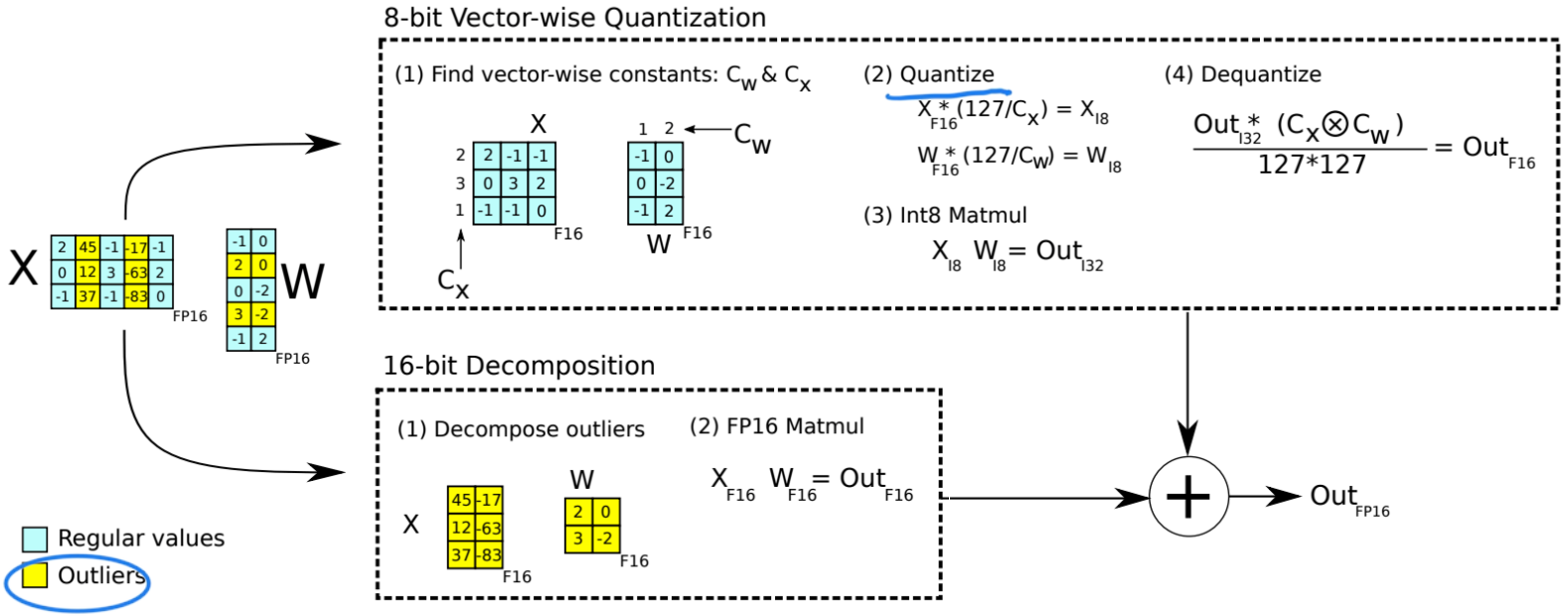
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can use a reduced numerical precision version of the model

LLM.int8



What about if the outliers aren't in one dimension? Lucky for us, they are in practise!

Dettmers et al. (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can use a reduced numerical precision version of the model

LLM.int8

			Largest Model that can be run	
Class	Hardware	GPU Memory	8-bit	16-bit
Enterprise	8x A100	80 GB	**OPT-175B / BLOOM**	**OPT-175B / BLOOM**
Enterprise	8x A100	40 GB	**OPT-175B / BLOOM**	OPT-66B
Academic server	8x RTX 3090	24 GB	**OPT-175B / BLOOM**	OPT-66B
Academic desktop	4x RTX 3090	24 GB	**OPT-66B**	OPT-30B
Paid Cloud	Colab Pro	15 GB	**OPT-13B**	GPT-J-6B
Free Cloud	Colab	12 GB	**T0/T5-11B**	GPT-2 1.3B

Dettmers et al. (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

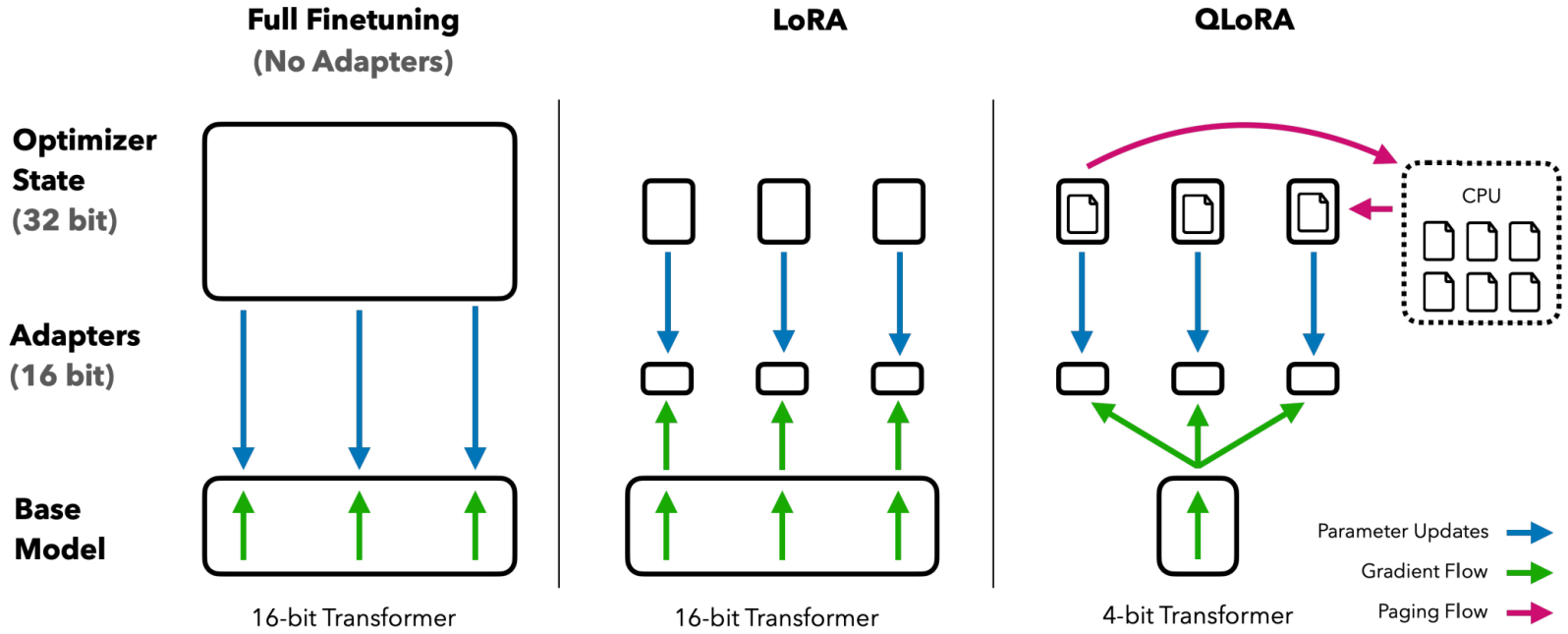
Workshop Preview

menti.com 2376 2478

We can combine these ideas

QLoRA

Quantization



Dettmers et al. (2023)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

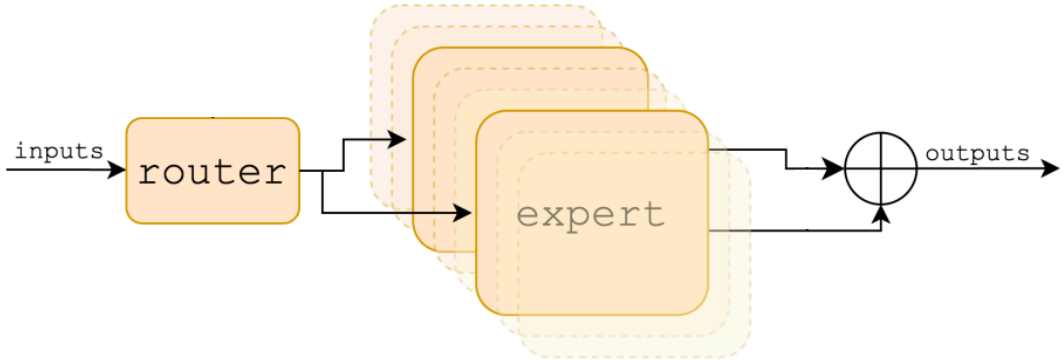
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can run a set of smaller models together

Sparse Mixture of Experts (SMoE)



Jiang et al. (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

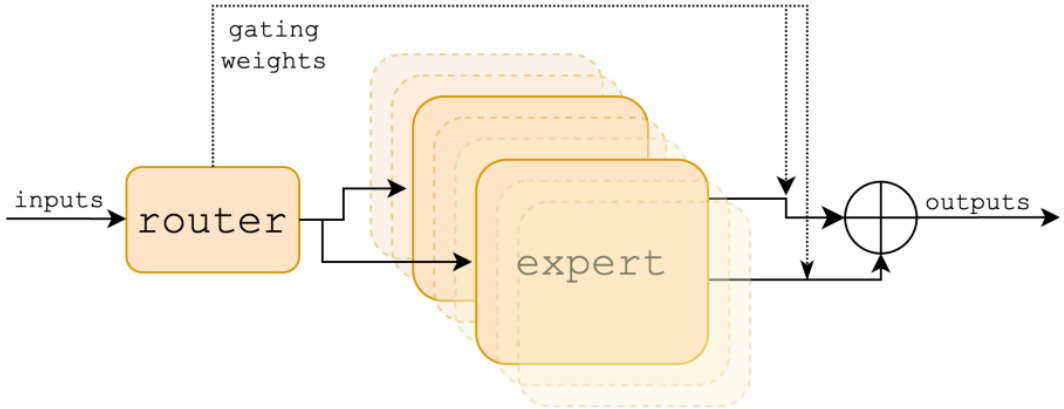
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can run a set of smaller models together

Sparse Mixture of Experts (SMoE)



Jiang et al. (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

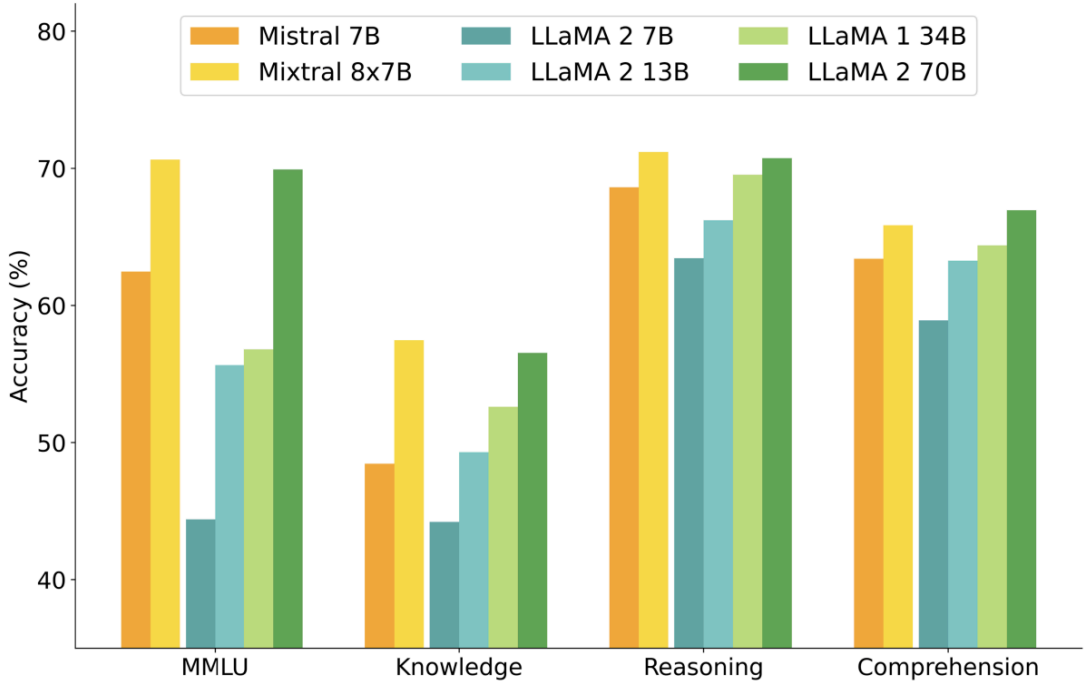
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

We can run a set of smaller models together

Sparse Mixture of Experts (SMoE)



Jiang et al. (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Recap: Efficiency

Reducing model size: To make models smaller we can use a large model to train a small model (distillation) or we can prune parts of a large model. Pruning is hard to do in a way that is computationally useful, so distillation is much more common

Increasing training memory efficiency: During training, fitting all the updates to the weights in memory can be expensive. Methods like LoRA allow us to approximate the changes, enabling training on lower memory GPUs.

Numerical approximation: We do not need full precision for our models. Reducing numerical precision can save GPU memory.

LLM.int8

Mixtures of models: To improve performance we can make models that are composed of a set of smaller models. Only one (or a few) models are active at a time.

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

In-Context Learning

COMP 4446 / 5046
Lecture 9, 2025

menti.com 2376 2478

What is “N-shot”?

i.e. Give 0 example

0-shot: Ask a question, no examples for training

1-shot: Give one example, then ask a question

N-shot: Give N examples, then ask a question

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

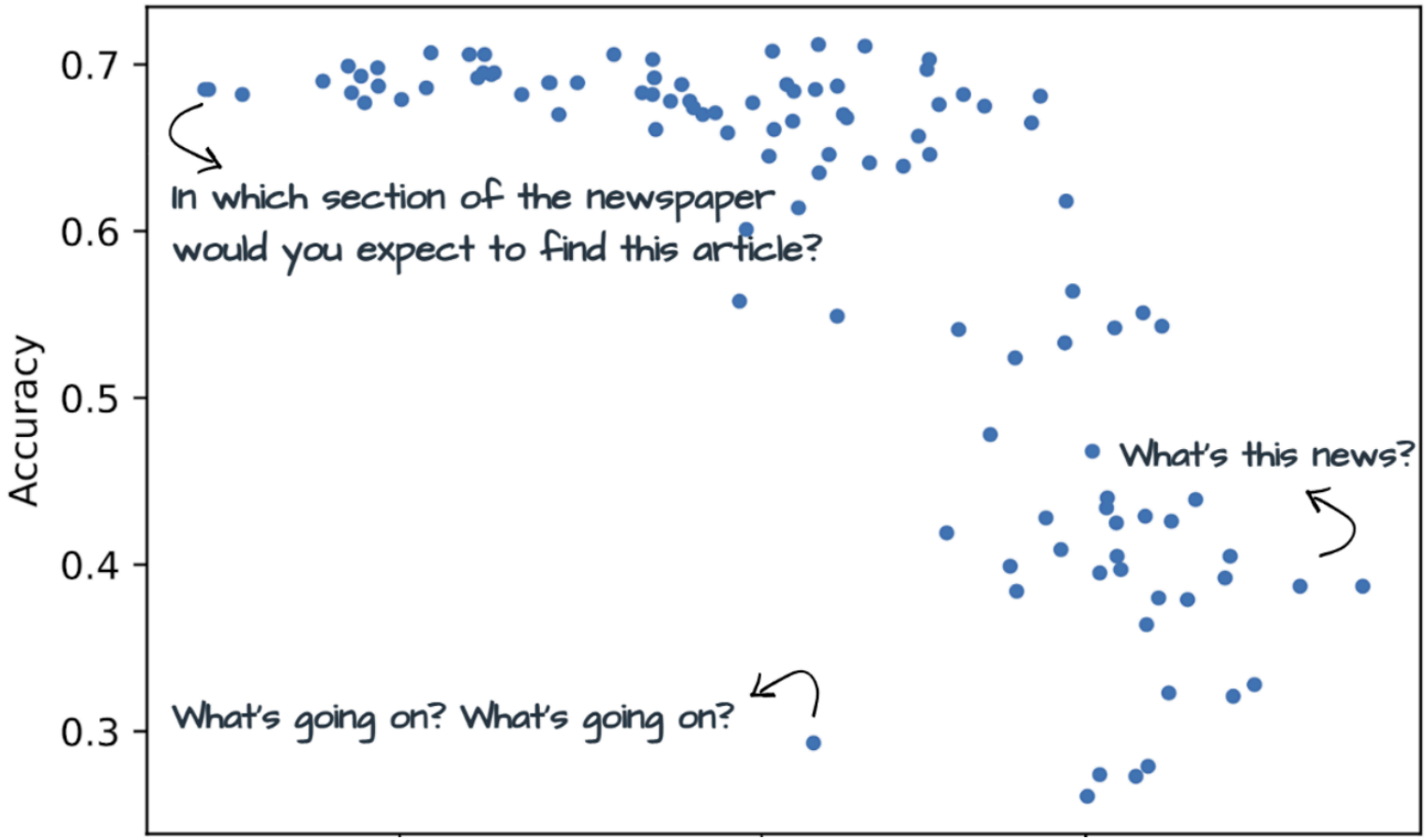
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Prompts vary in quality a lot



Gonen et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Zero-short prompting

Review:
The movie's acting could've been better, but the visuals and directing were top-notch.

Out of positive, negative, or neutral, this review is

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Zero-short prompting

Verb

Template 'verbalizer'

Review:
The movie's acting could've been better, but the visuals and directing were top-notch.

input text

Out of positive, negative, or neutral, this review is

predict from these labels

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Zero-short prompting

Template ‘verbalizer’

Review:
The movie’s acting could’ve been better, but the visuals and directing were top-notch.

On a 1 to 4 star scale, the reviewer would probably give this movie

UT Austin NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Measuring LM scores rather than reading 1 best

positive

negative

neutral

UT Austin NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Can also reverse it:

4 stars - The movie's acting could've been better, but the visuals and directing were top-notch.

$P(y \mid x)$ Vs. $P(x \mid y)P(y) \propto P(x \mid y)$

Context is different base on their order

UT Austin NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

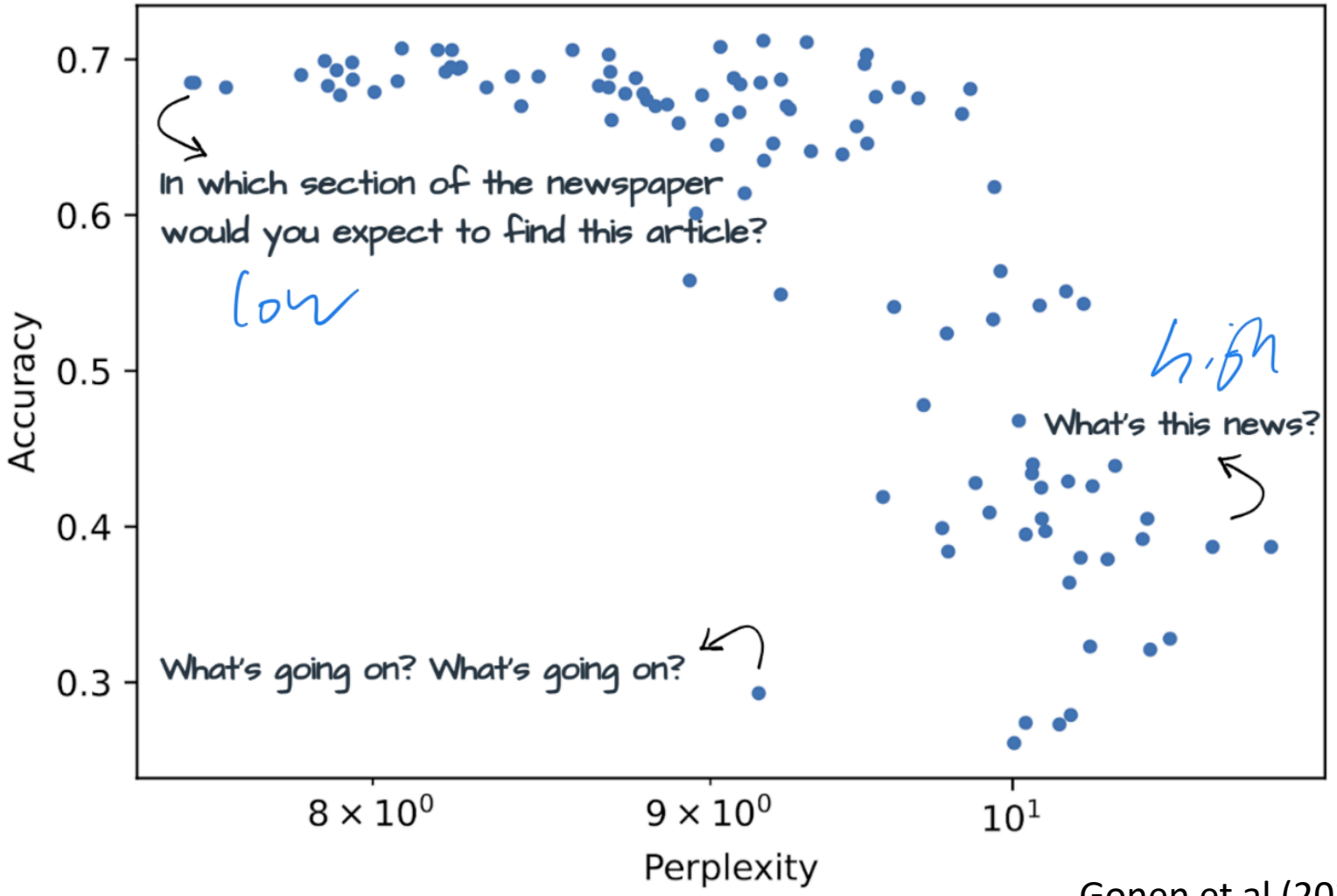
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Low perplexity prompts are better



Gonen et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

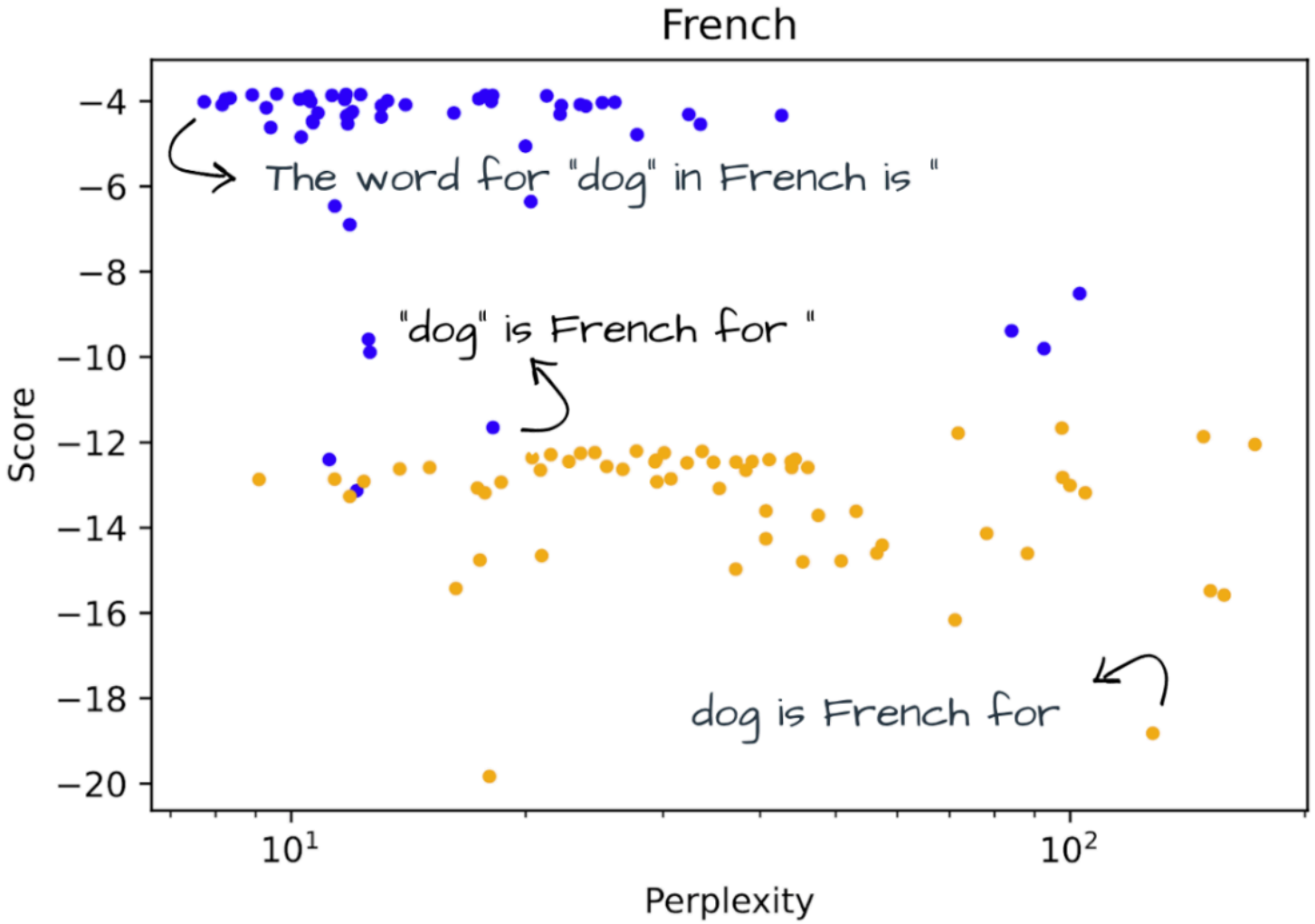
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Low perplexity prompts are better



Gonen et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Low perplexity prompts are better (mostly)

	OPT			Bloom		
Task	low-ppl	manual	Δ	low-ppl	manual	Δ
GLUE Cola	51.7	48.5	3.1	64.5	60.9	3.6
Newspop	80.6	70.4	10.2	90.0	80.0	10.0
AG News	68.4	61.9	6.5	51.0	63.5	-12.5
IMDB	90.4	88.9	1.4	91.3	88.8	2.5
DBpedia	46.0	51.7	-5.7	31.2	30.2	1.0
Emotion	21.6	22.6	-1.1	35.8	32.1	3.6
Tweet Offensive	48.4	50.6	-2.3	48.6	40.8	7.8

Gonen et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Few-shot prompting

given few examples

Review: The cinematography was stellar; great movie!
Sentiment (positive or negative): positive

Review: The plot was boring and the visuals were subpar.
Sentiment (positive or negative): negative

Review: The movie's acting could've been better, but the visuals and directing were top-notch.
Sentiment (positive or negative):

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

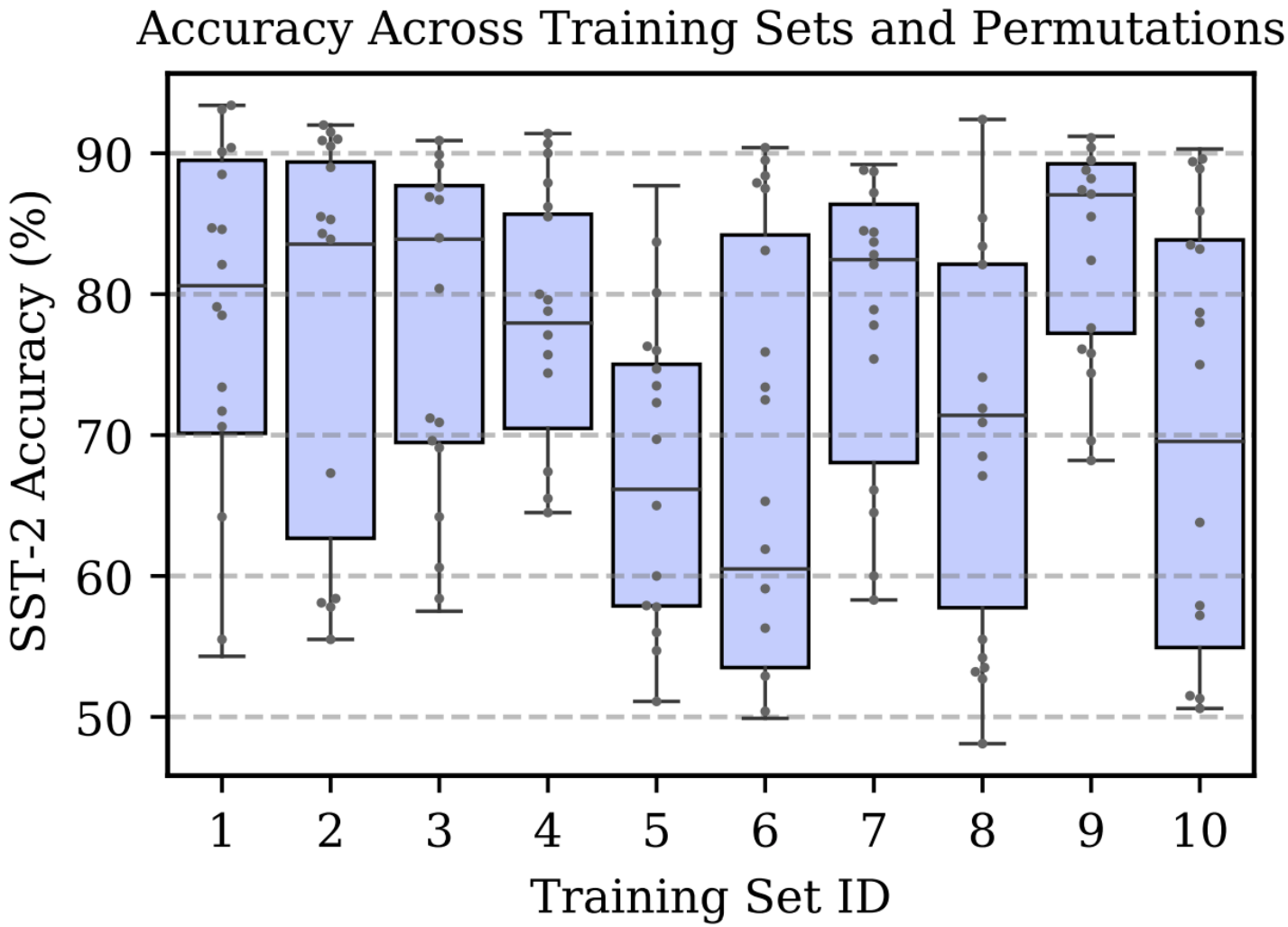
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Variability - different examples

Order of example matters



Zhao et al (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

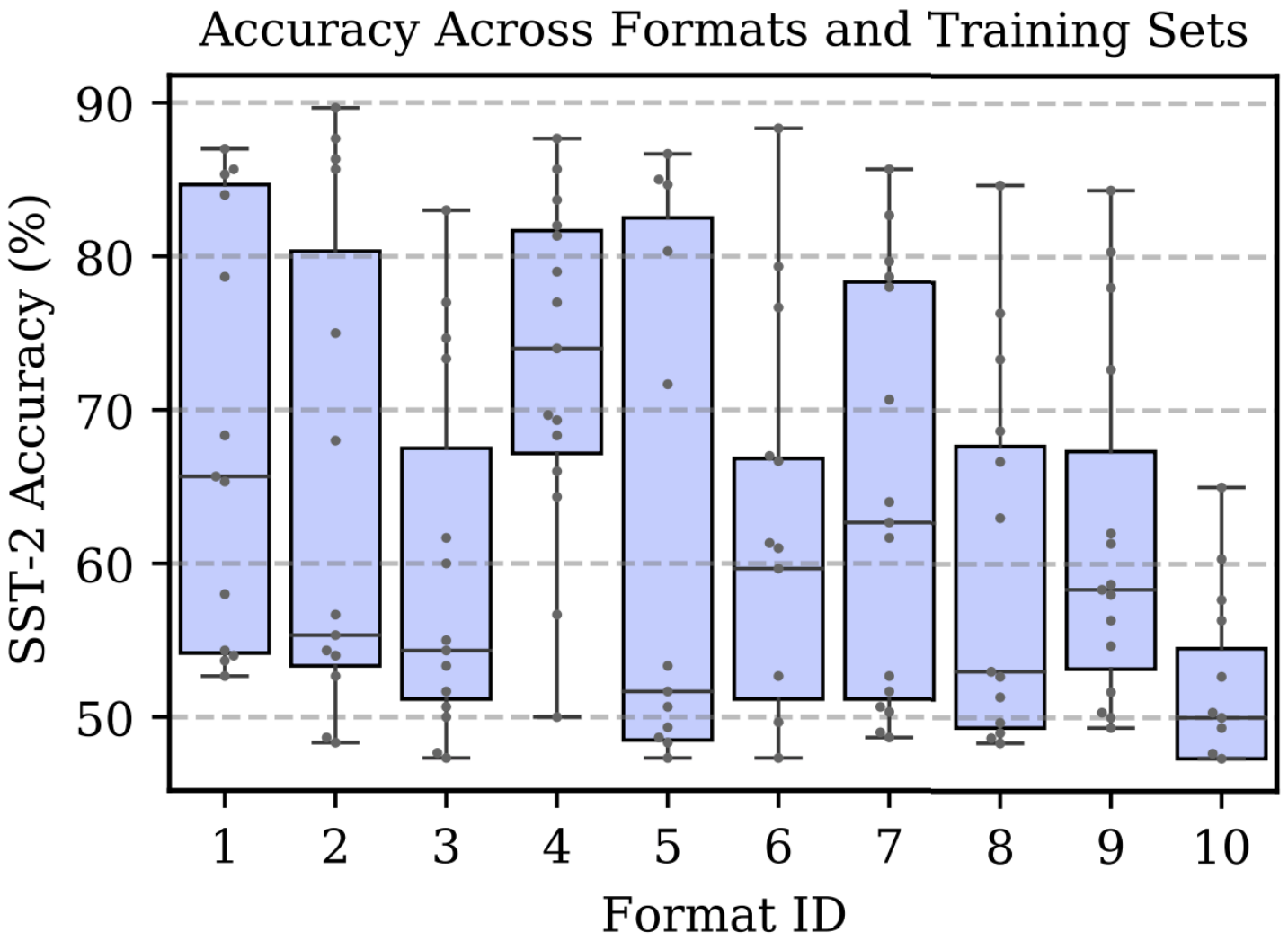
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Variability - different formats



Zhao et al (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

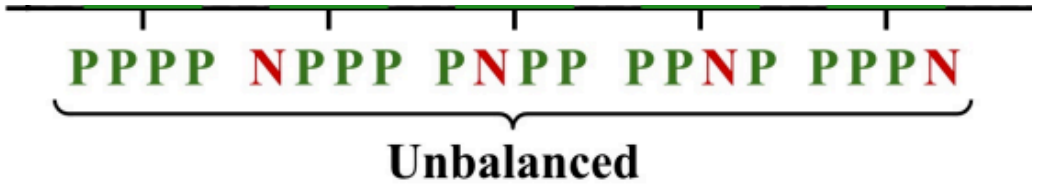
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Variability - ordering



Zhao et al (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

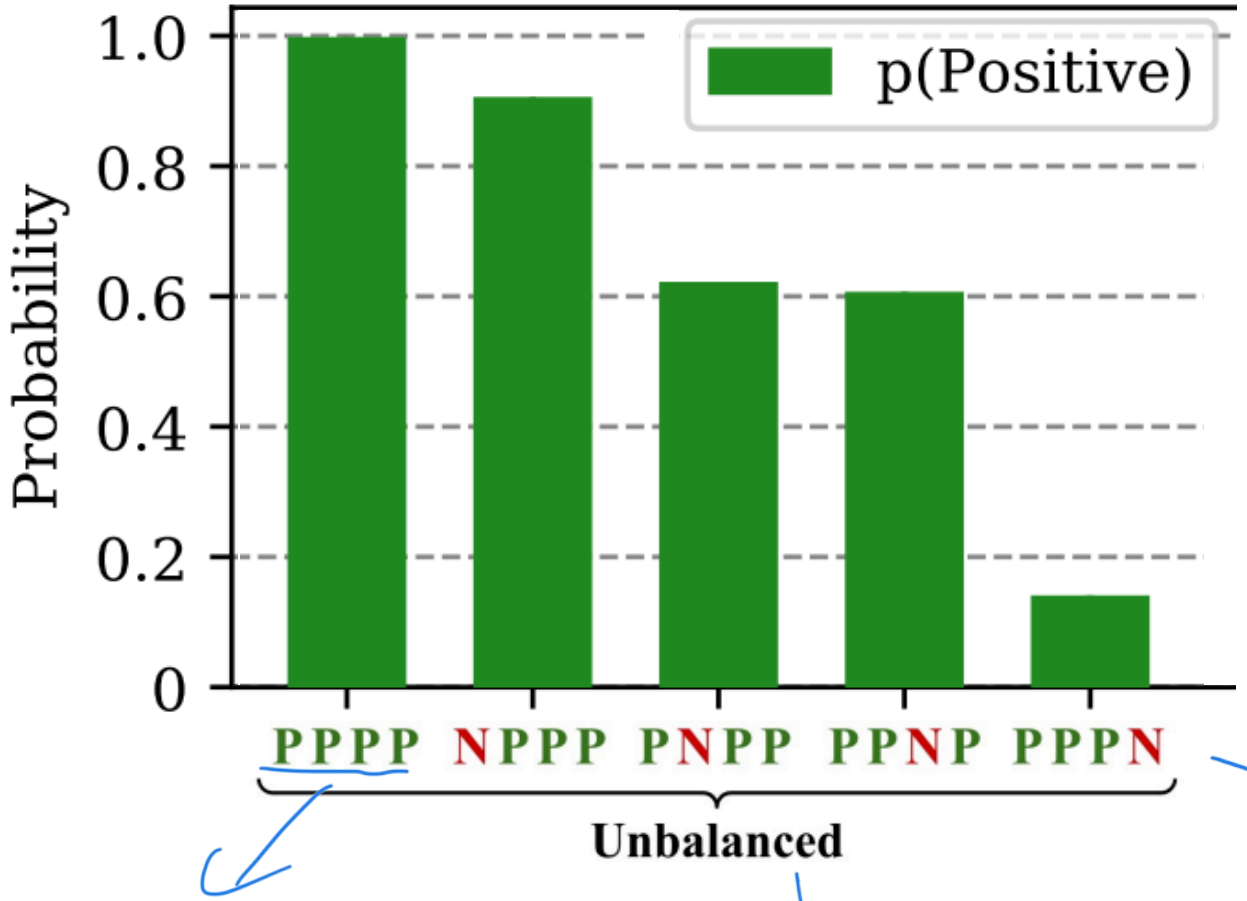
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Variability - ordering



predict with high p be positive

example

bias p predict positive

Zhao et al (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

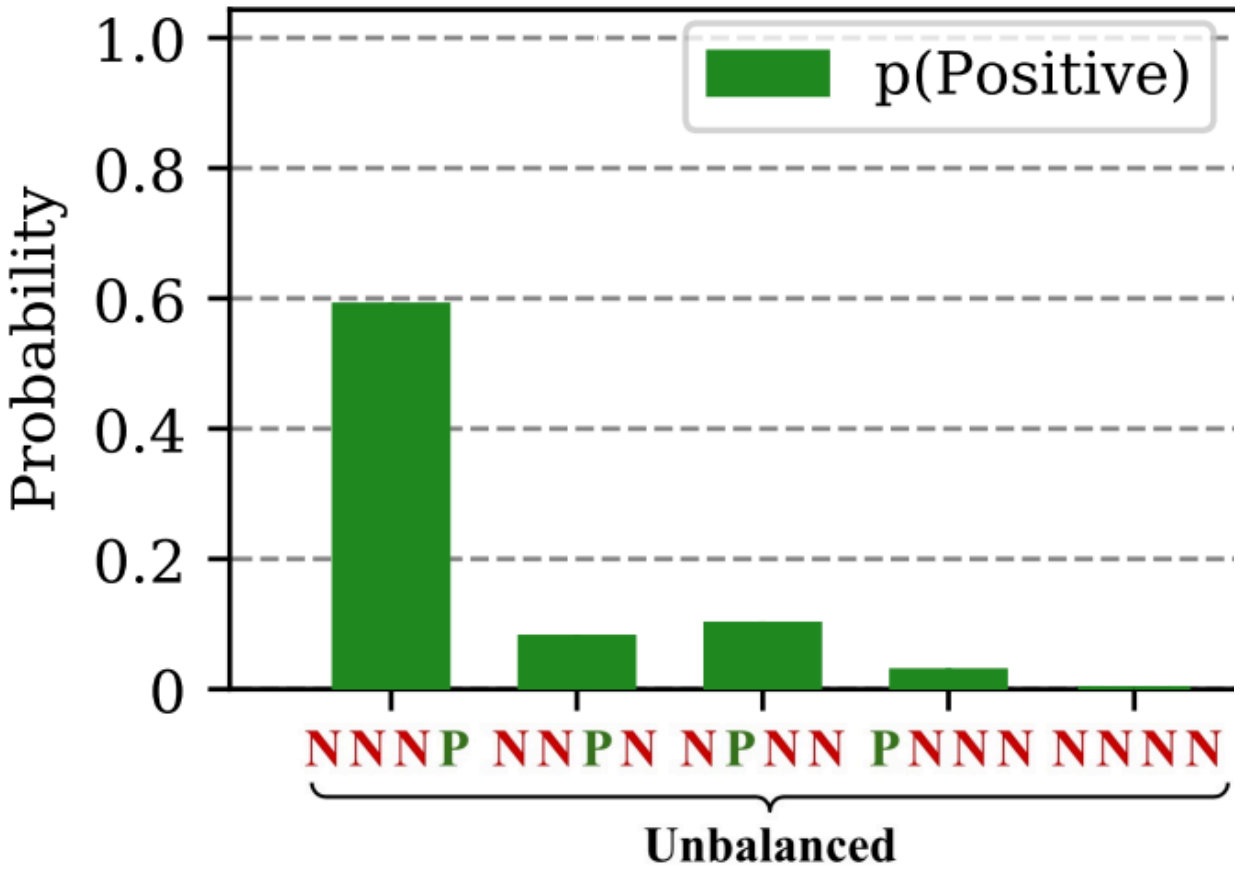
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Variability - ordering



Zhao et al (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

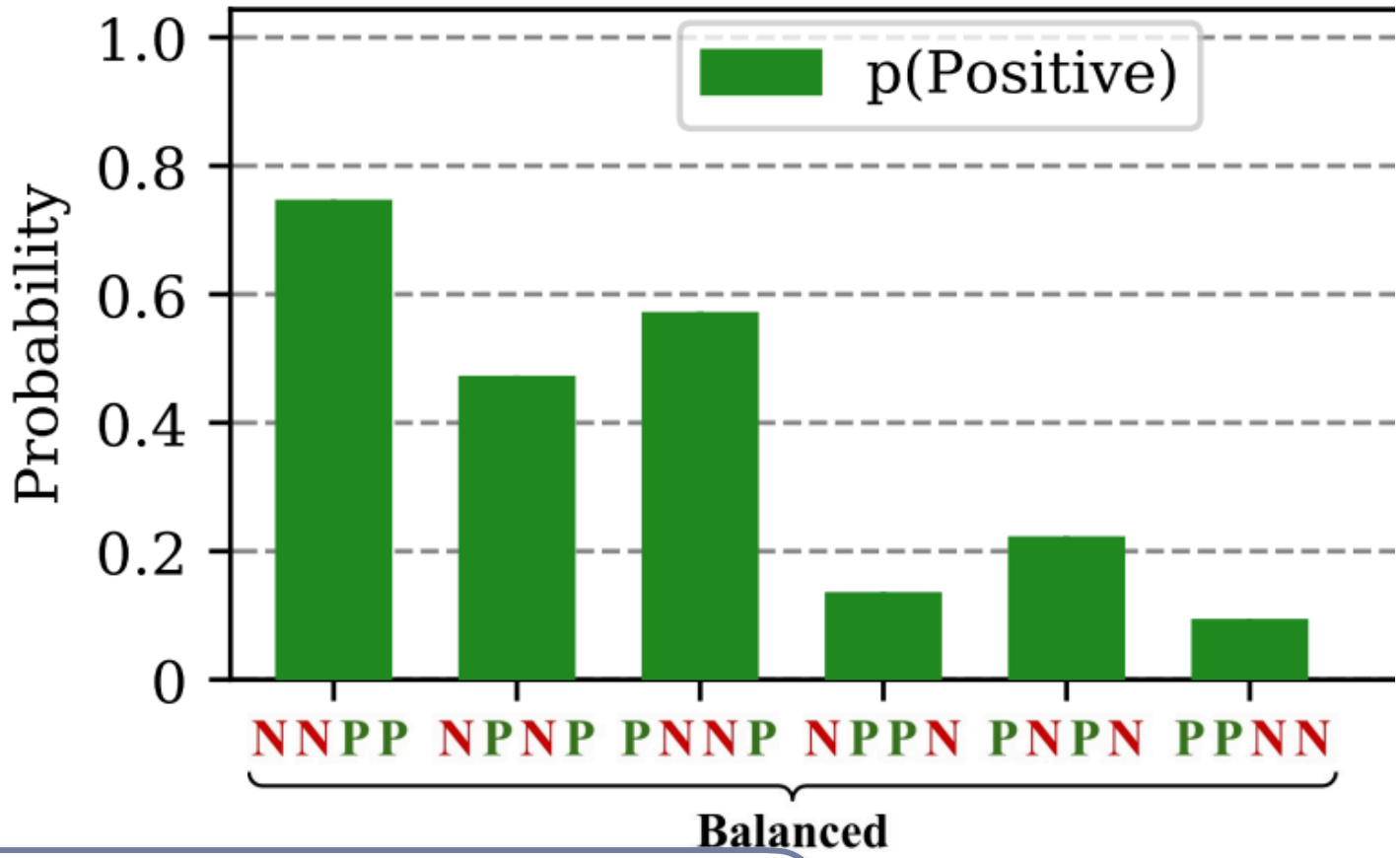
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Variability - ordering



I expect this effect is much smaller today than it was in 2021

Zhao et al (2021)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

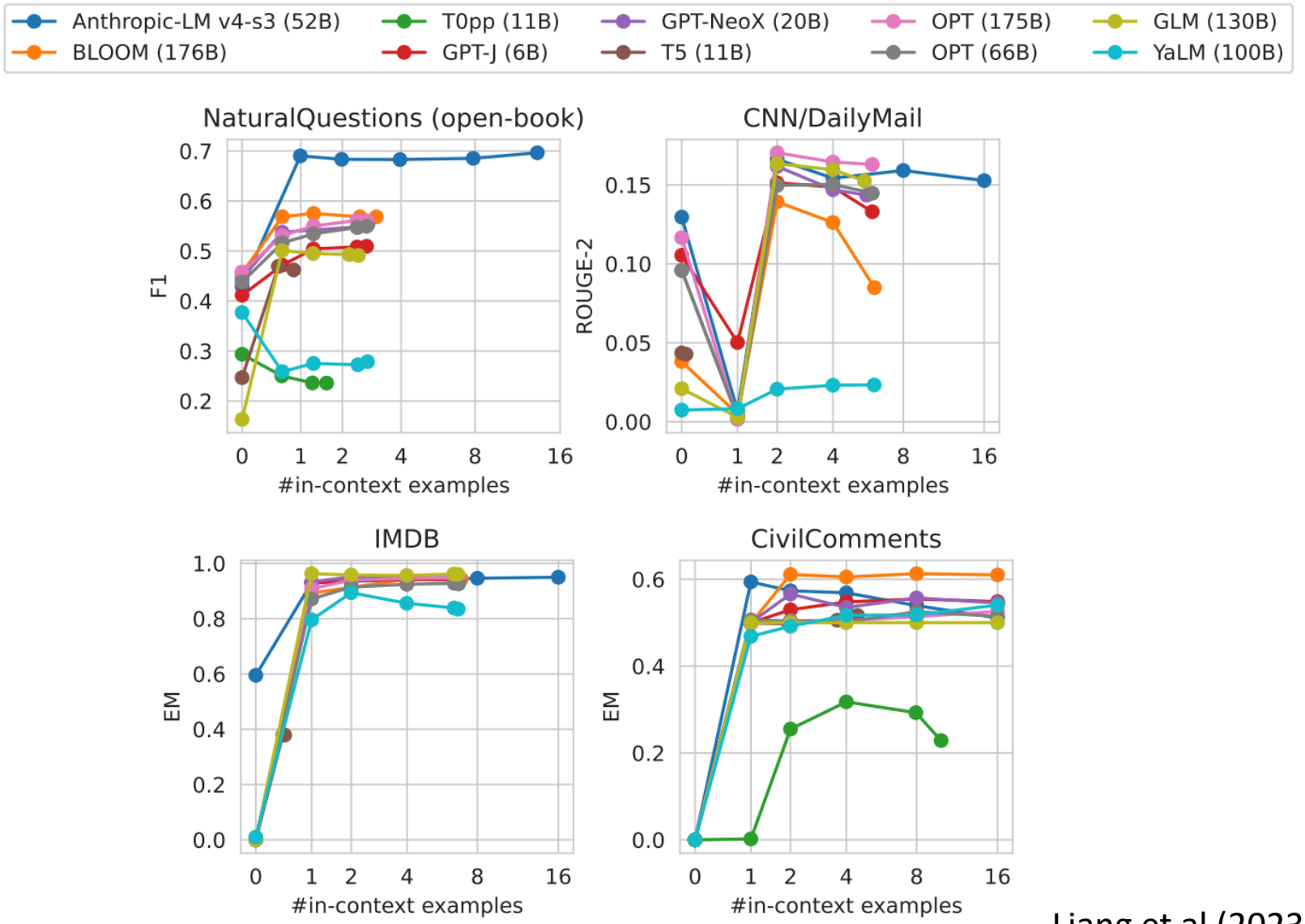
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

How many examples are needed?



Liang et al (2023)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

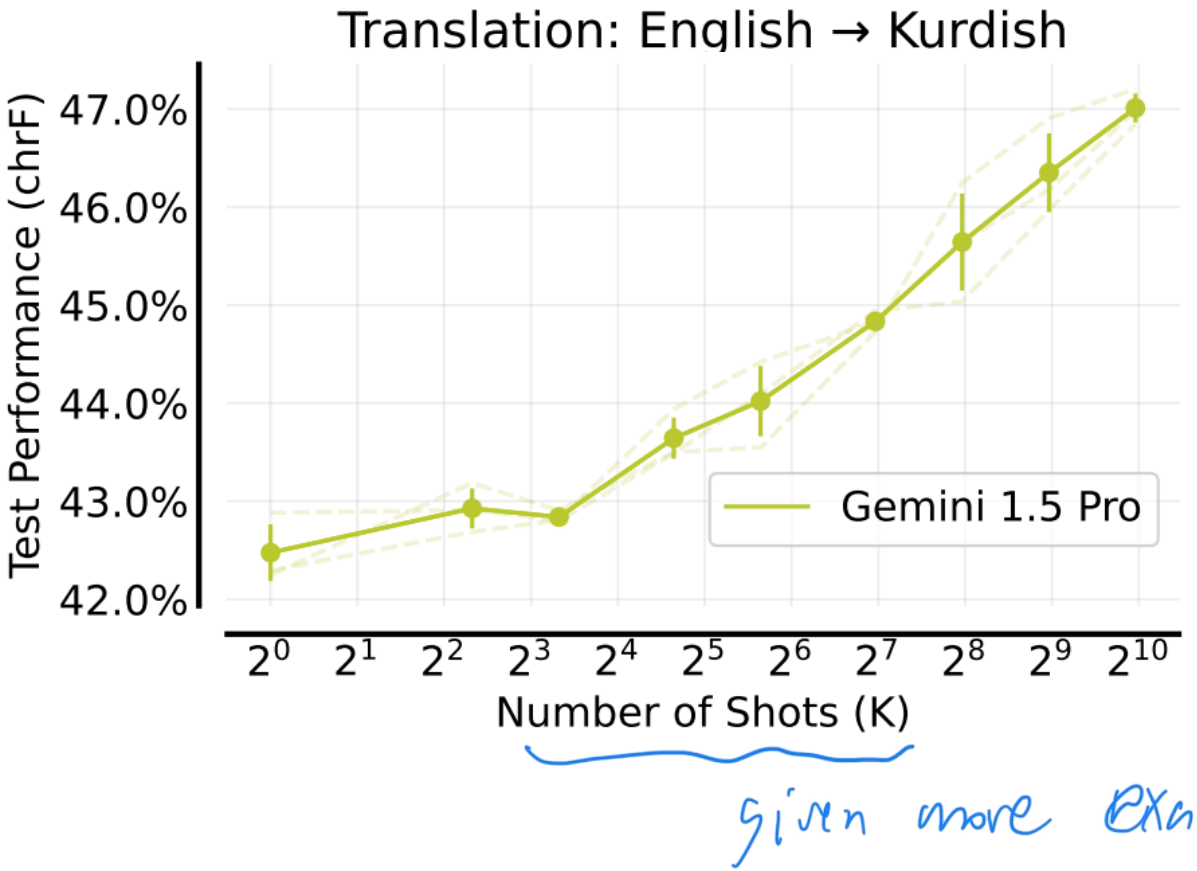
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

What if we scale up the number of examples?



Agarwal et al (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

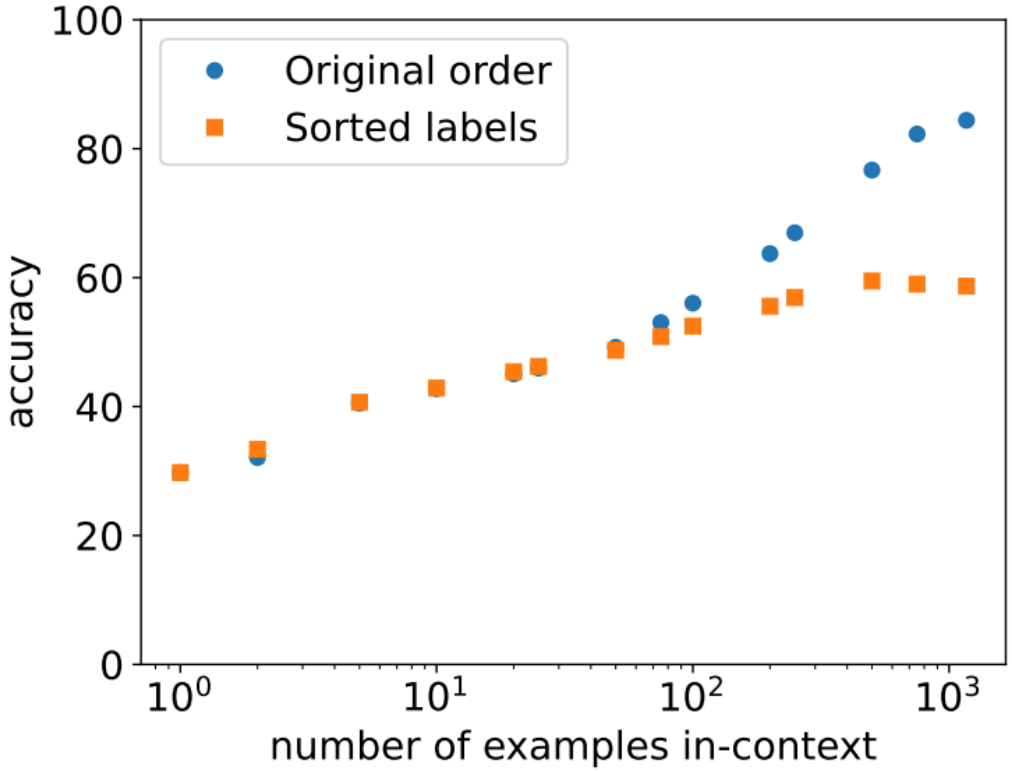
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

What if we scale up the number of examples?

performance for original and sorted are different



Bertsch et al (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

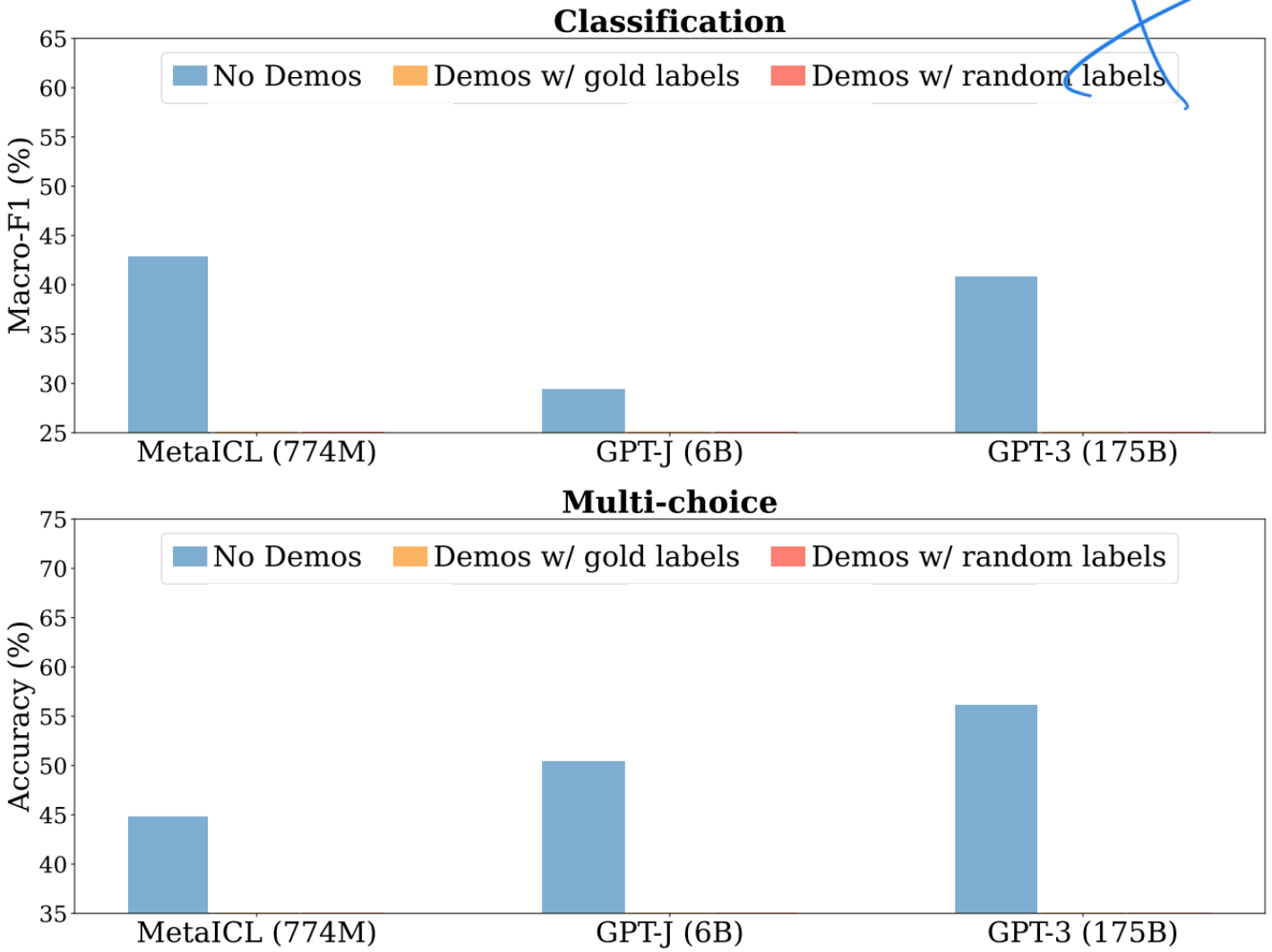
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

What labels help?



Min et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

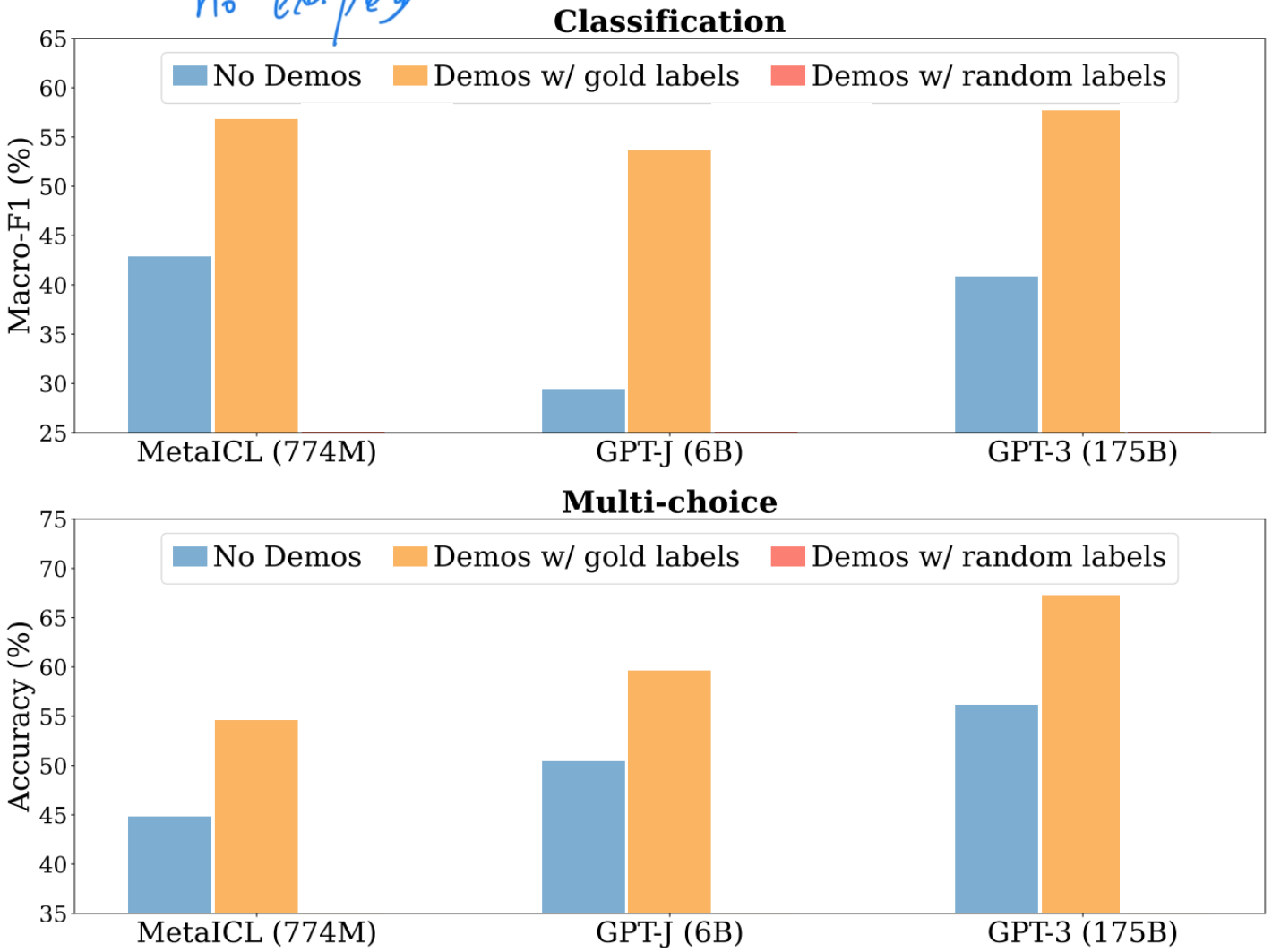
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

What labels help?

no examples



Min et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

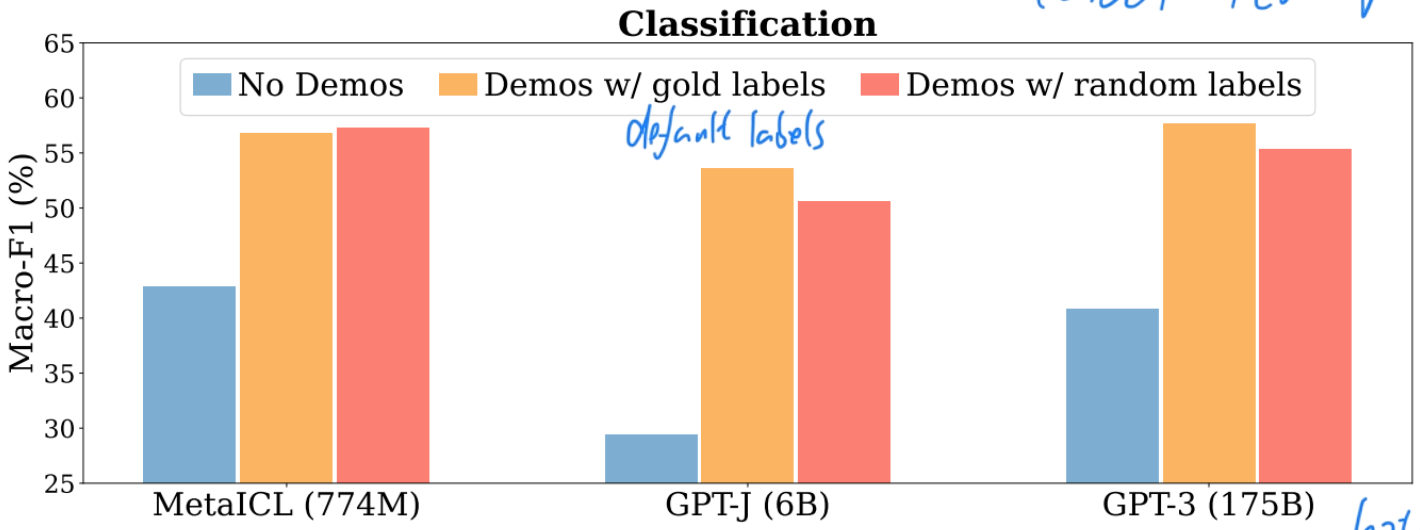
Retrieval Augmentation

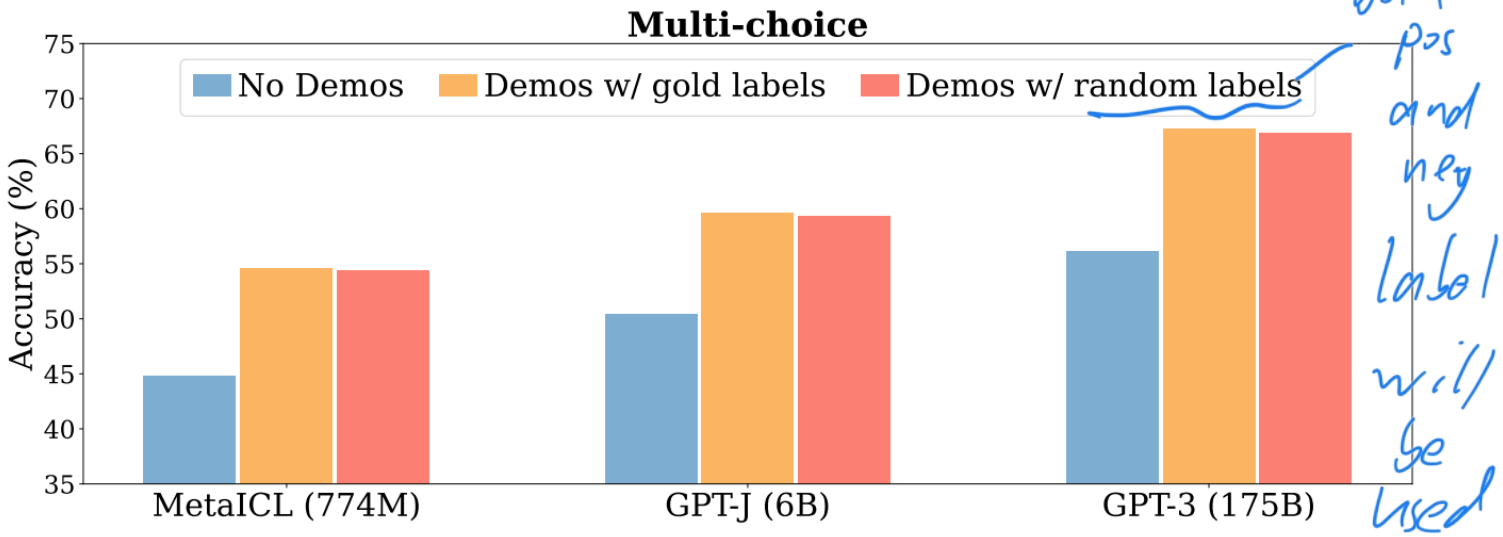
Workshop Preview

menti.com 2376 2478

What labels help?

structure of example is much more important that label itself





Min et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

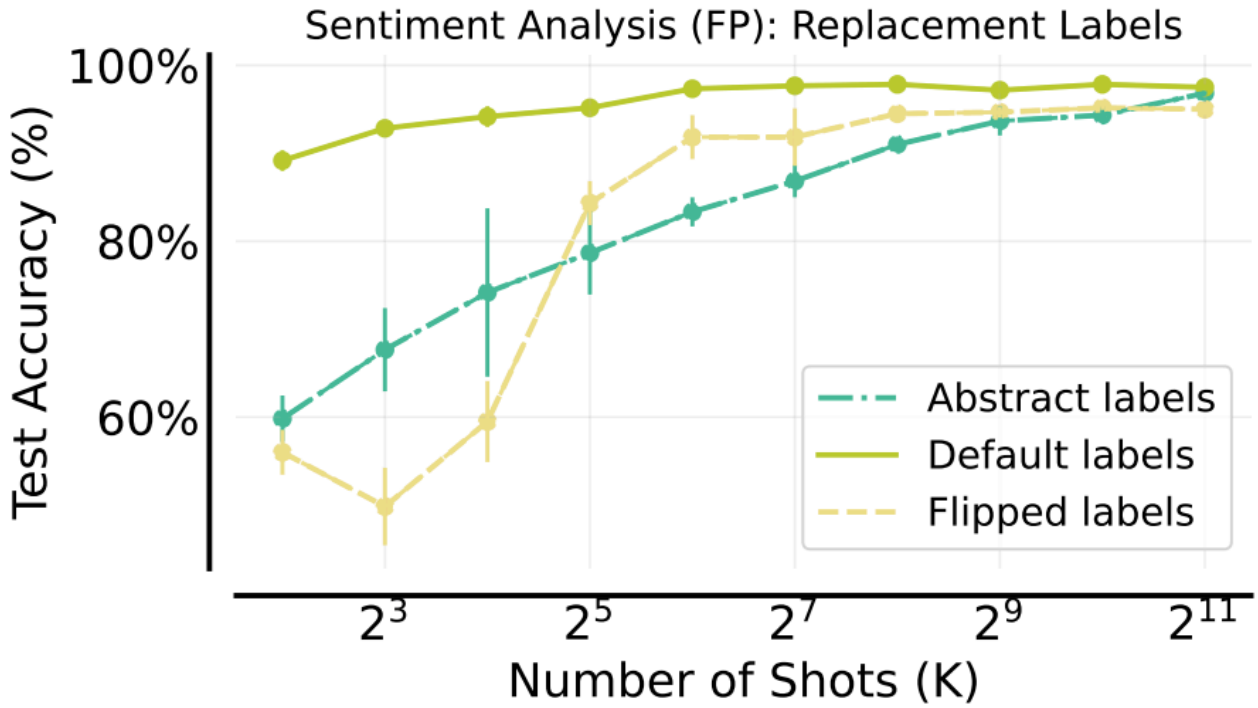
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

What labels are useful at scale?



Flipped - Intentionally change the label names
(e.g., swap 'positive' and 'negative')
Abstract - Give meaningless names
(e.g., 'A' instead of 'positive')

Agarwal et al (2024)

means we need make sure label relates to sentence

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Chain of Thought

Context: *Christopher agrees with Kevin.* […] **Q**: *Who hangs out with a student?*

Mary

Standard few-shot learning, no explanation , no explanation

Context: *Christopher agrees with Kevin.* […] **Q**: *Who hangs out with a student?*

Mary, because Mary hangs out with Danielle and Danielle is a student.

Predict-explain: answer **is not** conditioned on output explanation (original E-SNLI LSTM)

Context: *Christopher agrees with Kevin.* […] **Q**: *Who hangs out with a student?*

Because Mary hangs out with Danielle and Danielle is a student**, the answer is **Mary.

Explain-predict: answer is conditioned on output explanation (Chain of Thought)

explain the logic and make prediction

From the UT Austin NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

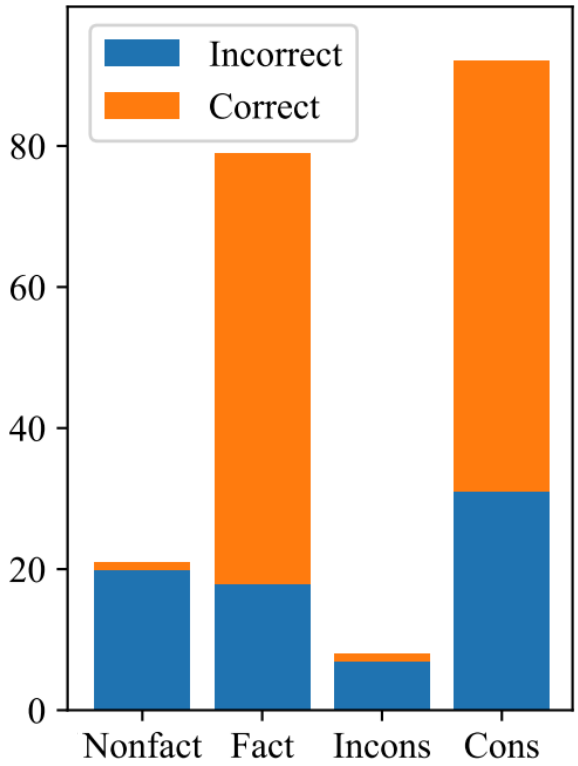
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Chain of Thought

What is in these explanations?
Manually label them for facts and consistency, then see



Bad explanations usually lead to the wrong answer!

Ye and Durrett (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

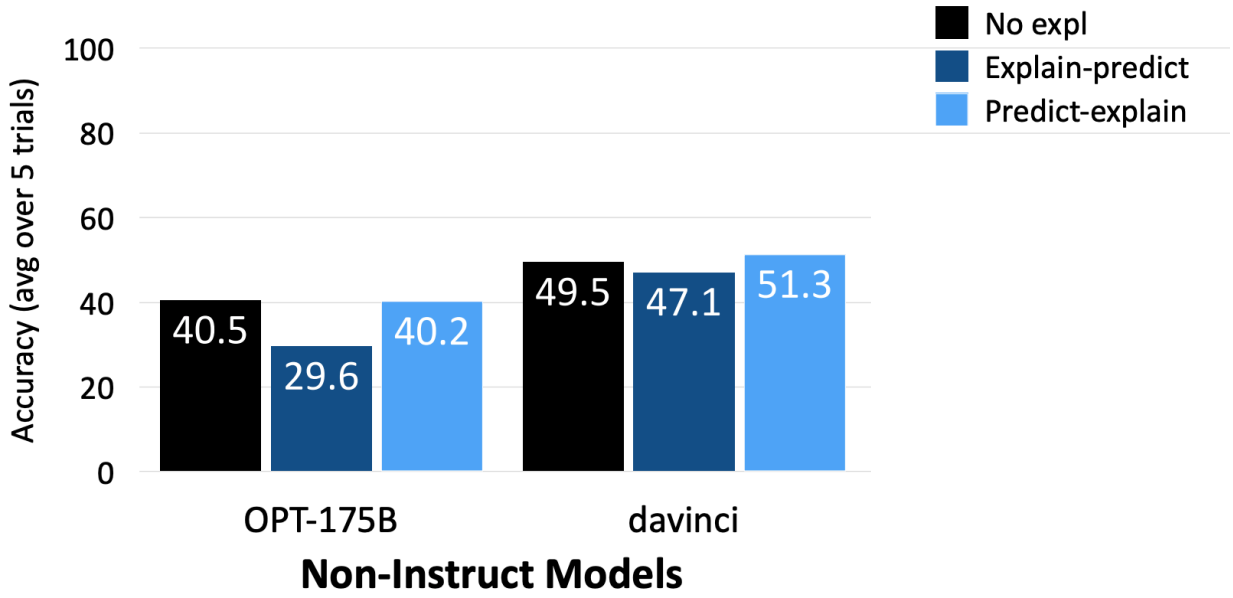
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Chain of Thought



From the UT Austin NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

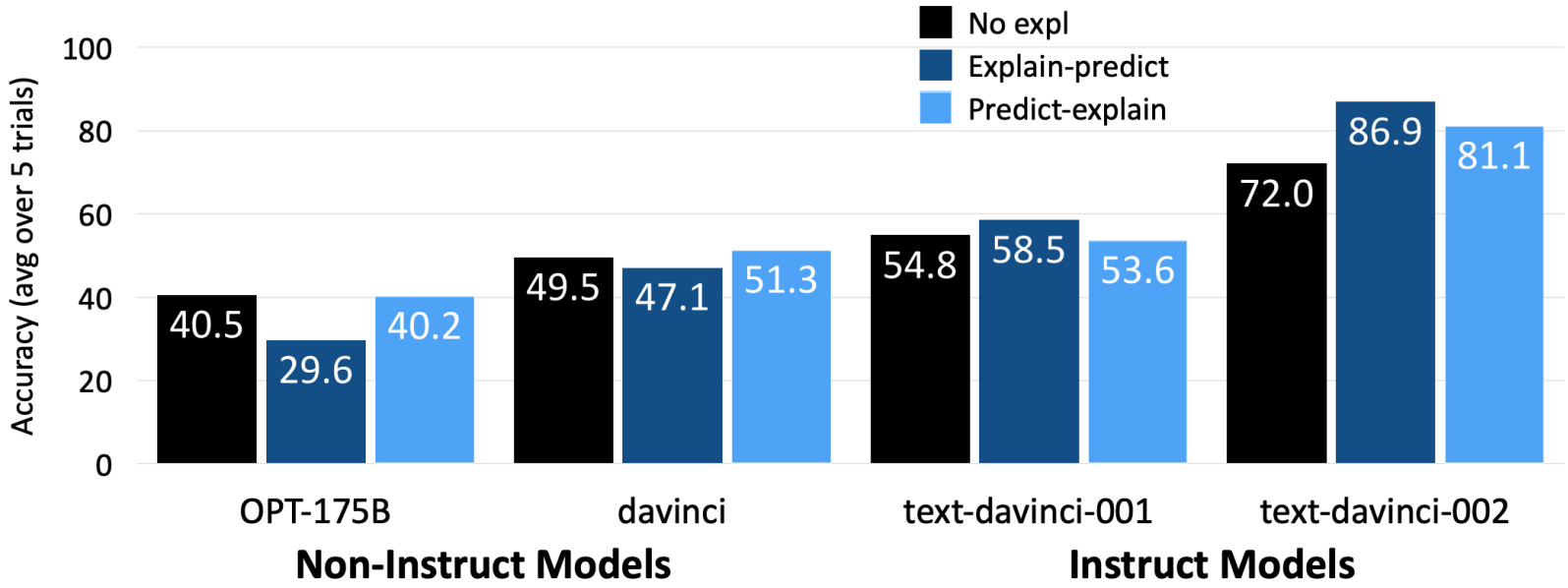
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Chain of Thought



From the UT Austin NLP Course

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Zero-Shot version: “Step-by-Step”

No.	Category	Template	Accuracy
1	instructive	Let’s think step by step.	**78.7**
2		First, (*1)	77.3
3		Let’s think about this logically.	74.5
4		Let’s solve this problem by splitting it into steps. (*2)	72.2
5		Let’s be realistic and think step by step.	70.8
6		Let’s think like a detective step by step.	70.3
7		Let’s think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don’t think. Just feel.	18.8
11		Let’s think step by step but reach an incorrect answer.	18.7
12		Let’s count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It’s a beautiful day.	13.1
-		(Zero-shot)	17.7

we need to ask relevant info

Kojima et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

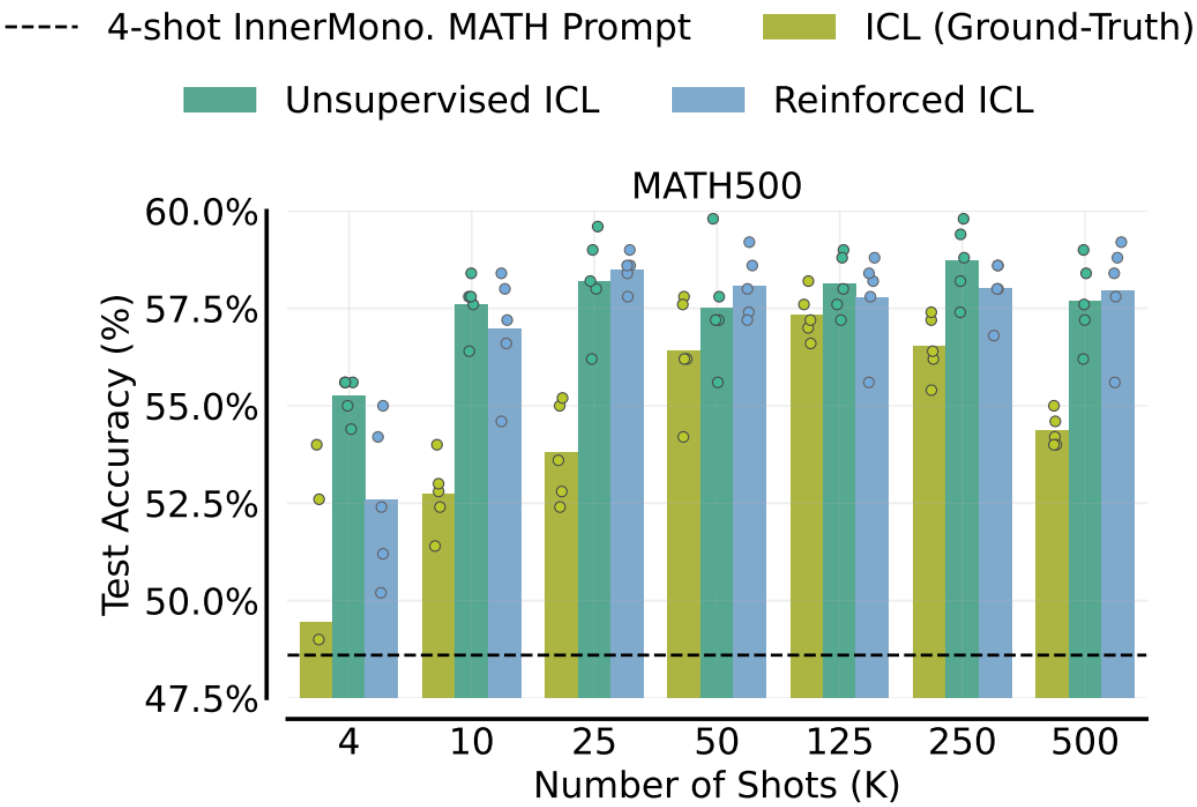
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

How do rationales interact with demonstrations?



Agarwal et al (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

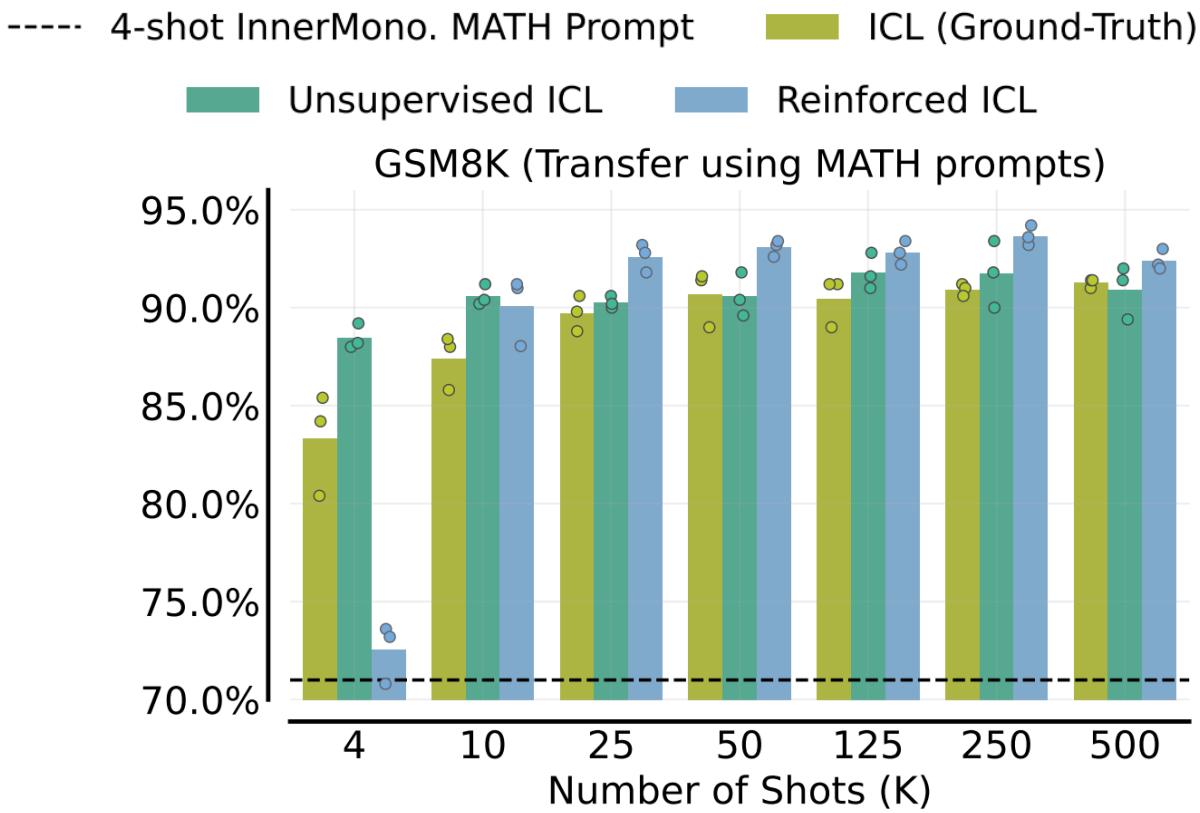
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

How do rationales interact with demonstrations?



Agarwal et al (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Recap: In-Context Learning (ICL)

ICL: Rather than modifying the weights of the model, provide instructions and examples in the input and hope that it will somehow use those to model the task.

LLMs appear to benefit from:

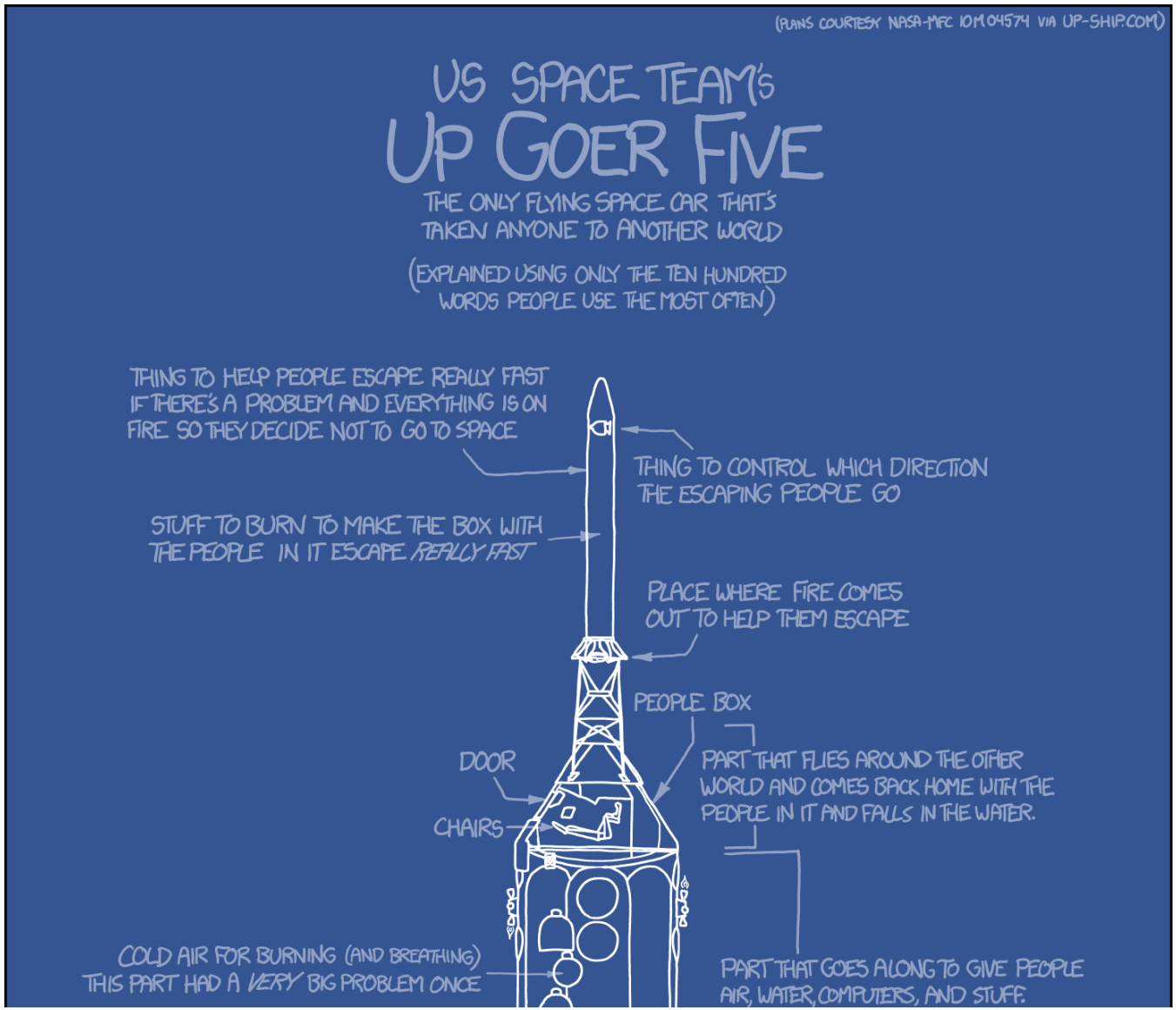
- Examples providing the label space
- Seeing the distribution of inputs
- Seeing the format of the intended output

You should:

- Use as many demonstrations as you can
- Avoid patterns in the order of examples

3 minute Break - stretch and visit Menti

menti.com
2376 2478



Up Goer Five

[Another thing that is a bad problem is if you're flying toward space and the parts start to fall off your space car in the wrong order. If that happens, it means you won't go to space today, or maybe ever.]

Source: https://xkcd.com/1133/

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Retrieval Augmentation

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Why do we want to access external data?

Hallucination	Jonathan teaches NLP, Basket Weaving, and AI
New Events	The current Prime Minister is Scott Morrison
Rare Info	The Old Darlington School is a Gothic Revival style building
Errors	Chocolate is not particularly delicious

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

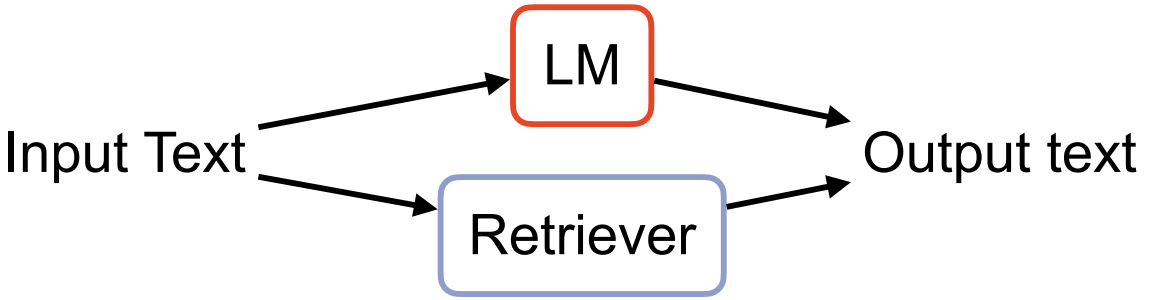
Retrieval Augmentation

Workshop Preview

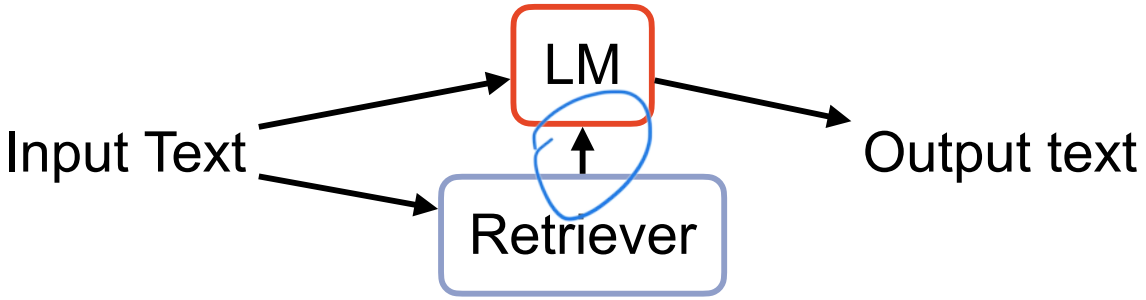
menti.com 2376 2478

Three general forms

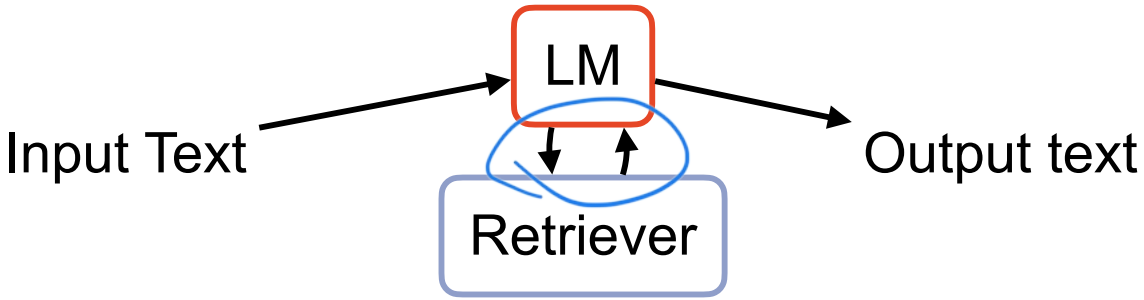
Parallel Interaction



Sequential Single Interaction



Sequential Multiple Interaction



COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

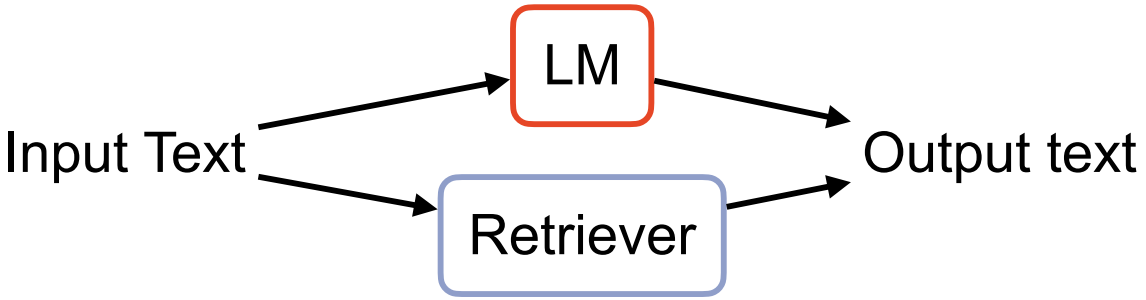
Retrieval Augmentation

Workshop Preview

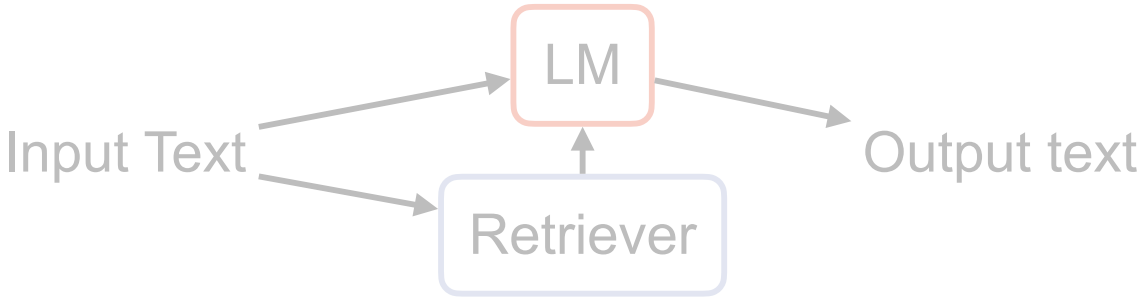
menti.com 2376 2478

Three general forms

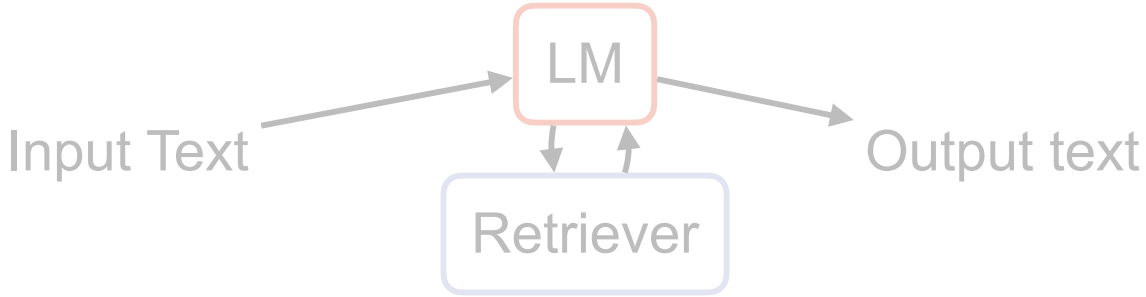
Parallel Interaction



Sequential Single Interaction



Sequential Multiple Interaction



COMP 4446 / 5046
Lecture 9, 2025

Efficiency

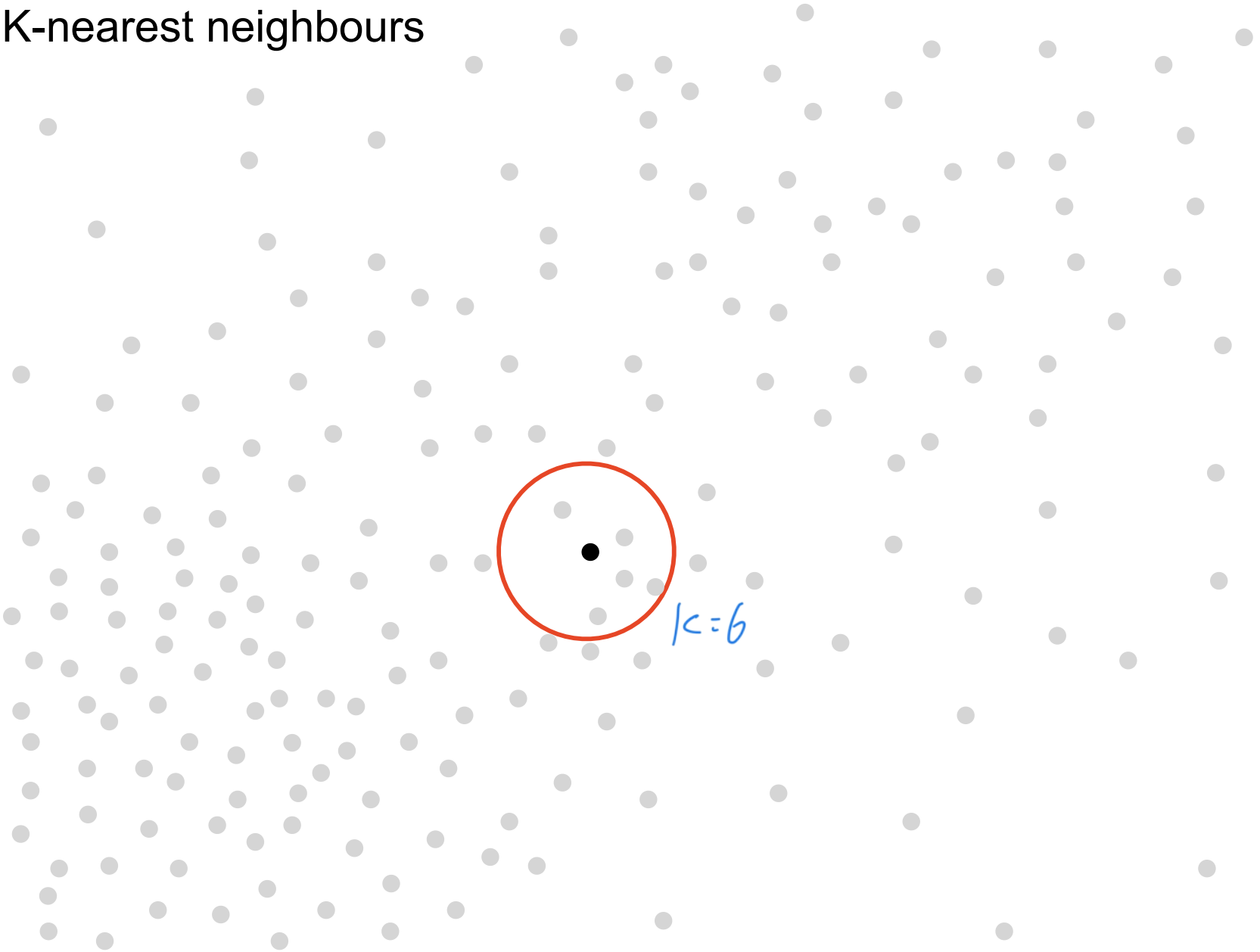
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

K-nearest neighbours



COMP 4446 / 5046
Lecture 9, 2025

Efficiency

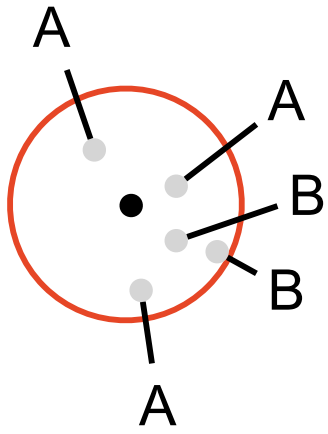
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

K-nearest neighbours



Return: A

Majority is A

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM

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

Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

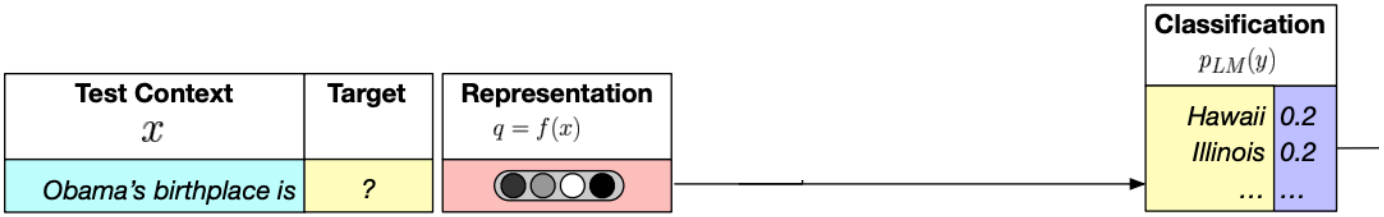
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

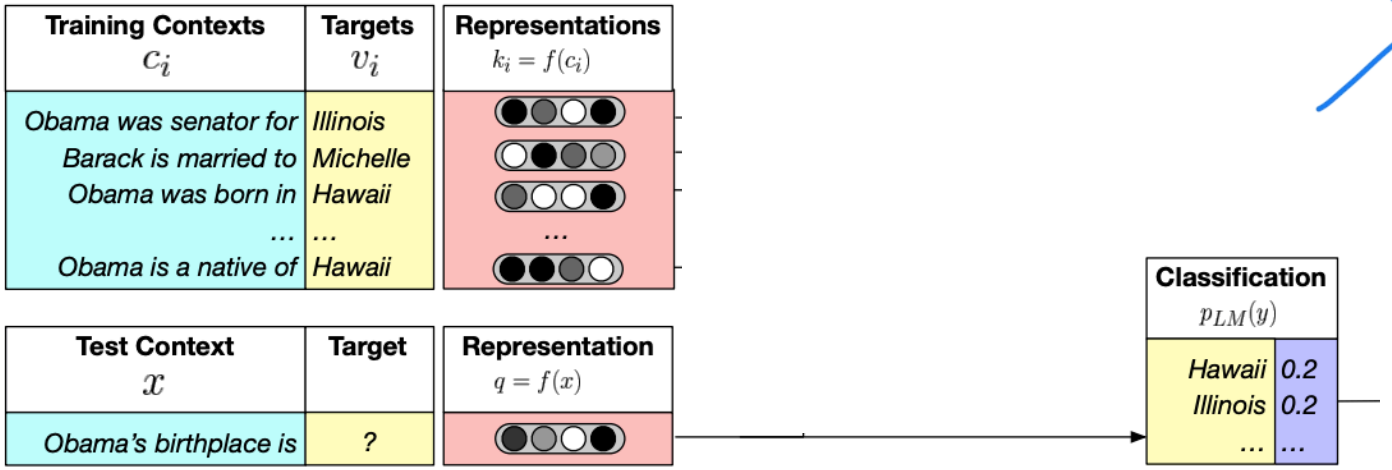
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM

Train data

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

far away because irrelevant

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

input

Classification $p_{LM}(y)$	
Hawaii	0.2
Illinois	0.2
...	...

Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

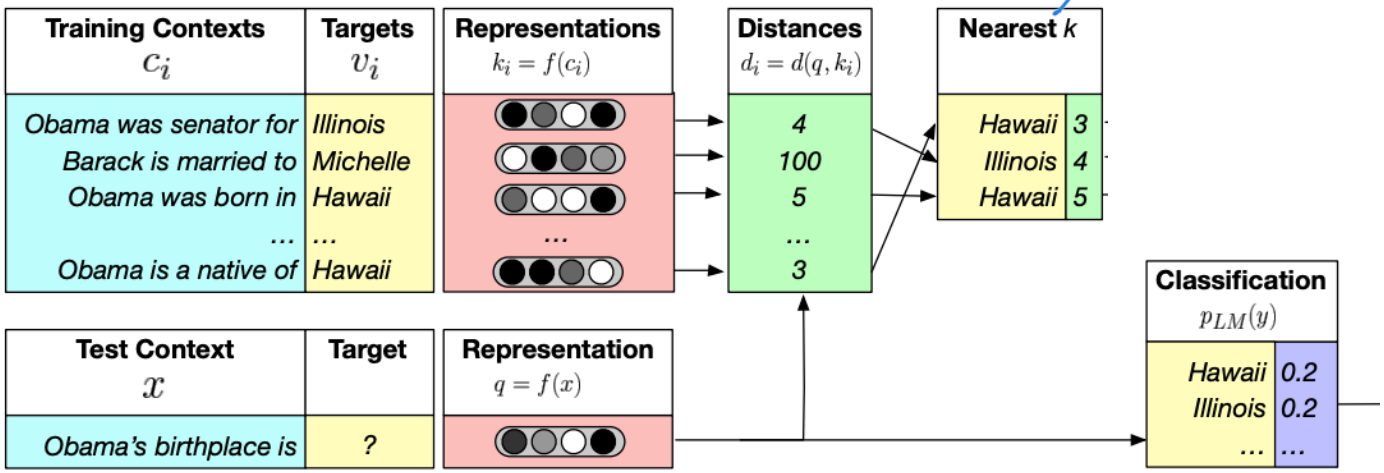
Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM

normalize to get distribution



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

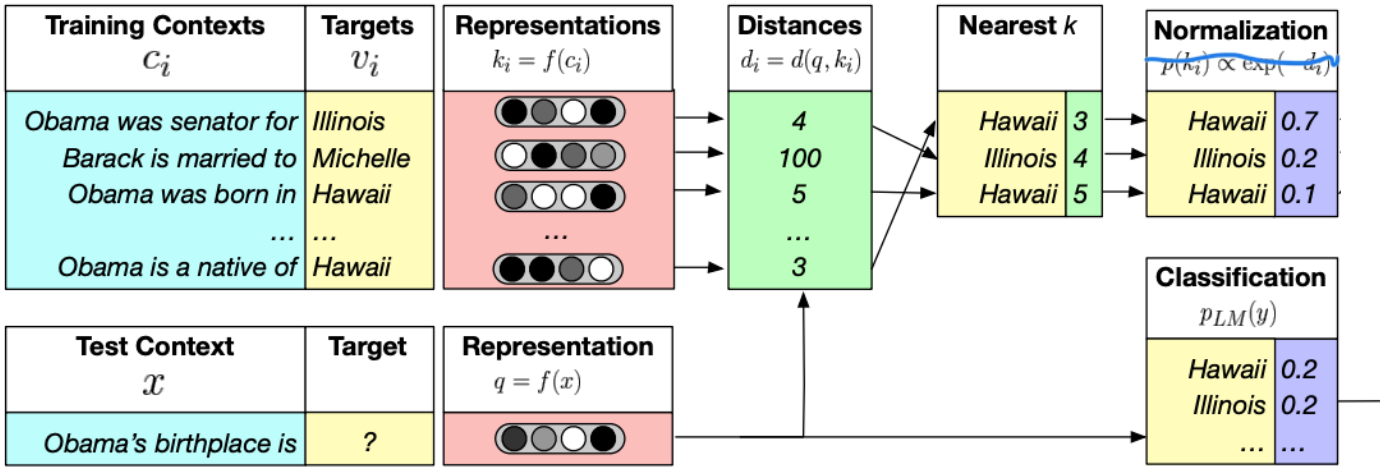
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

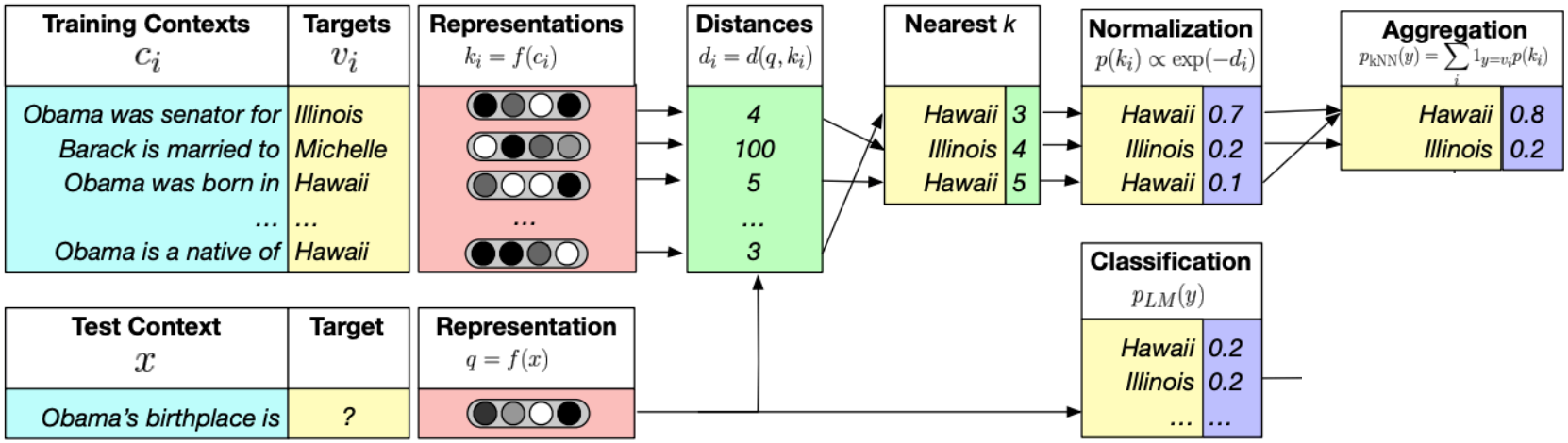
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

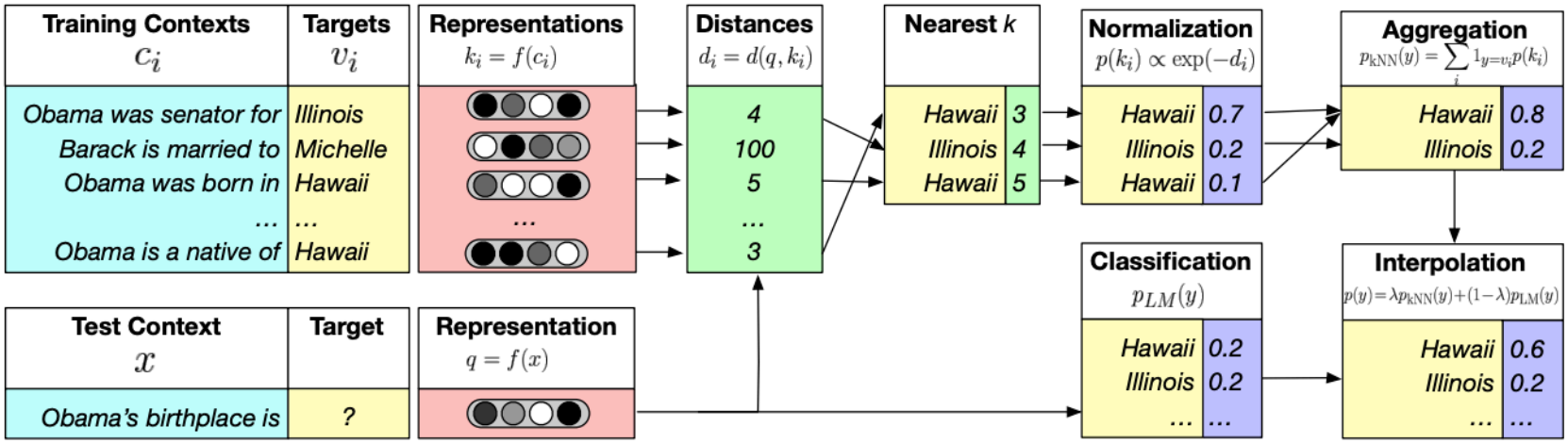
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

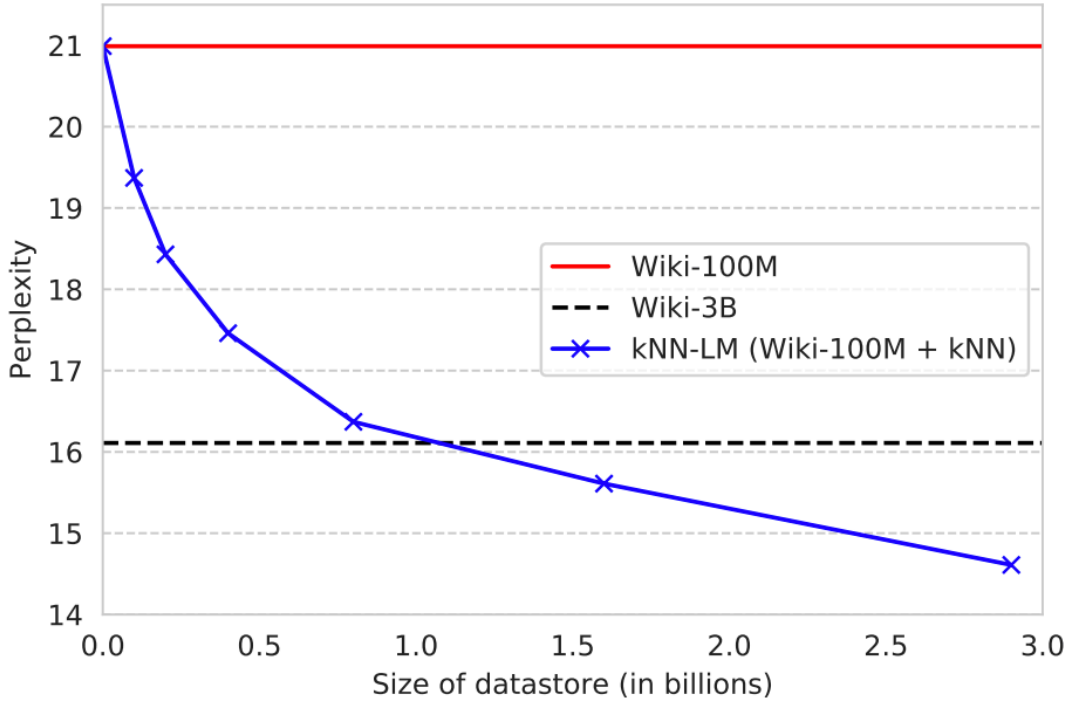
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

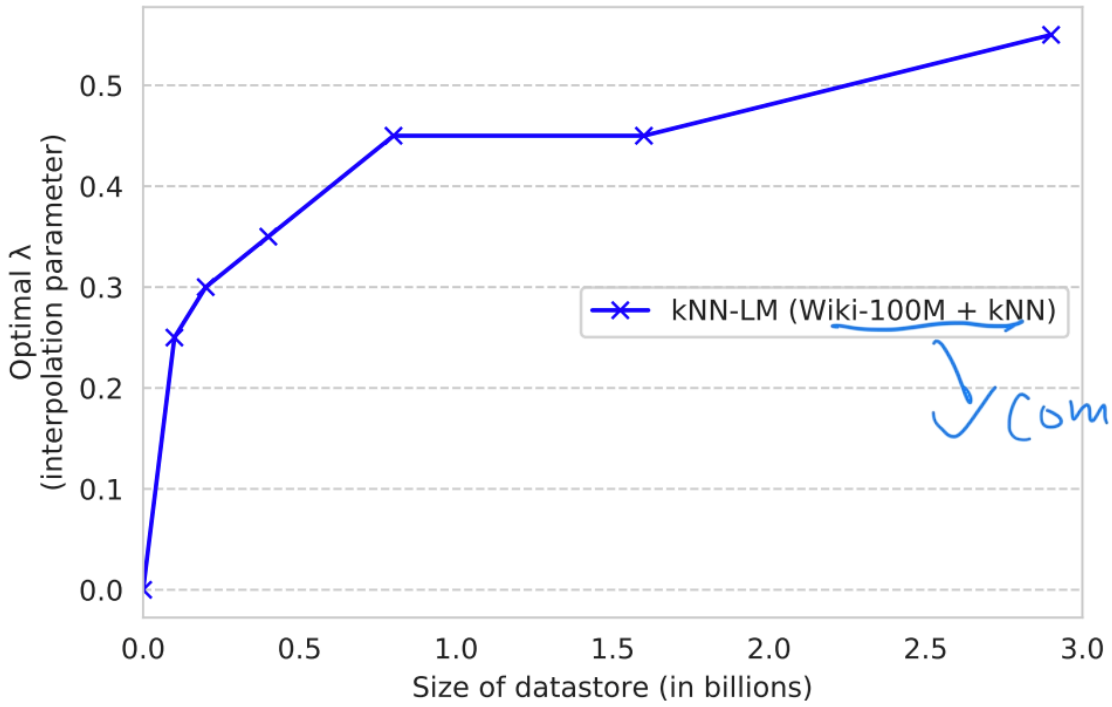
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



combine classifica and knn

Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

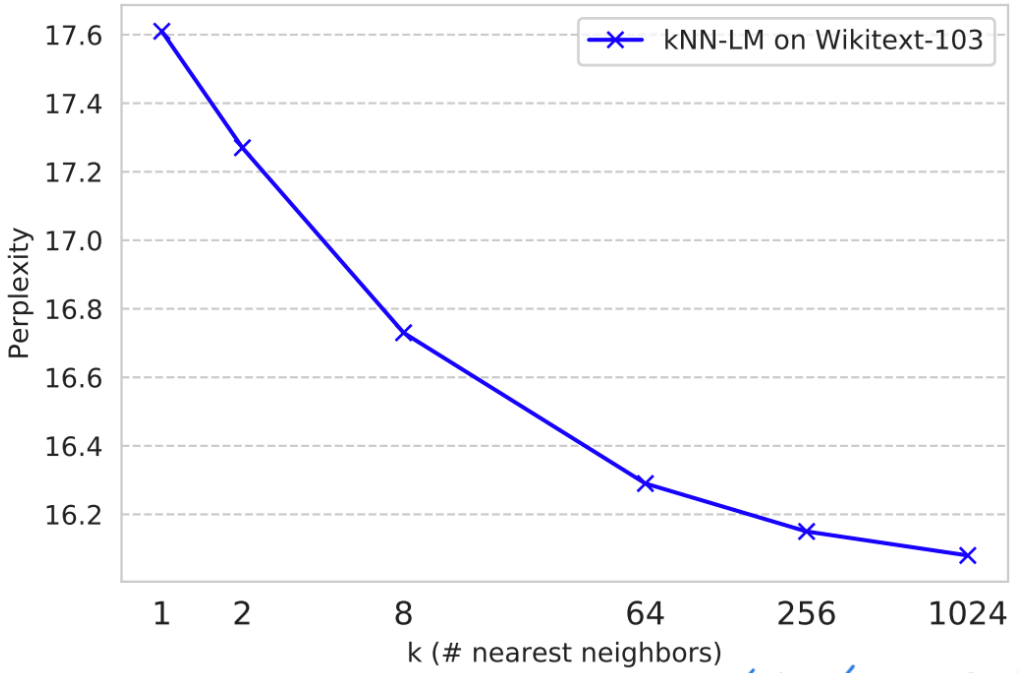
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



K 越大 perform 越好

Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

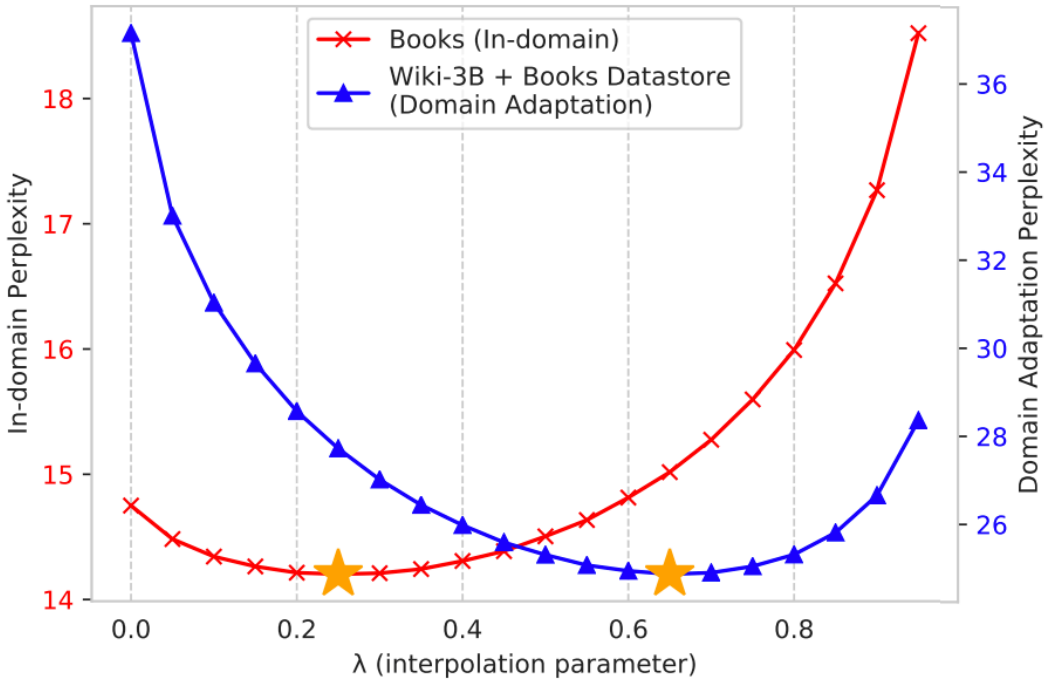
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

kNN-LM



Khandelwal et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

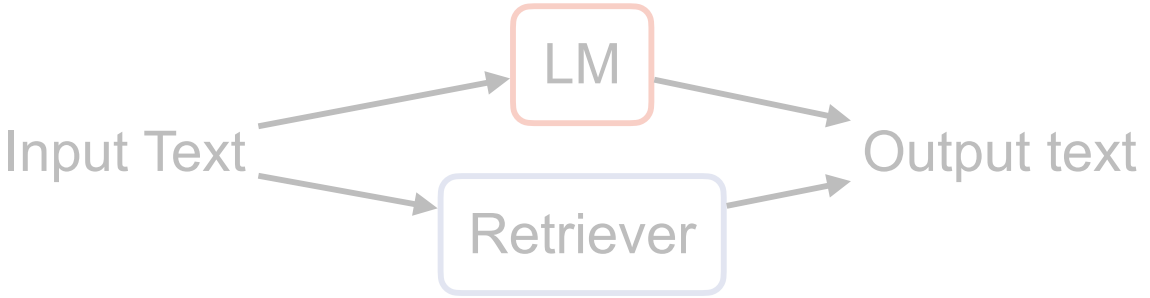
Retrieval Augmentation

Workshop Preview

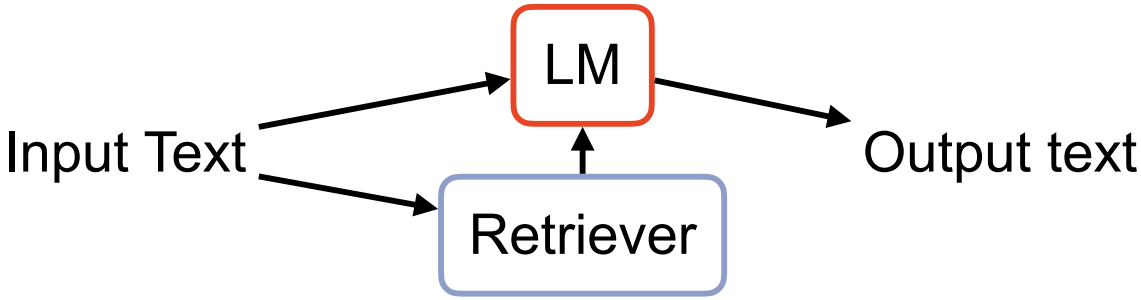
menti.com 2376 2478

Three general forms

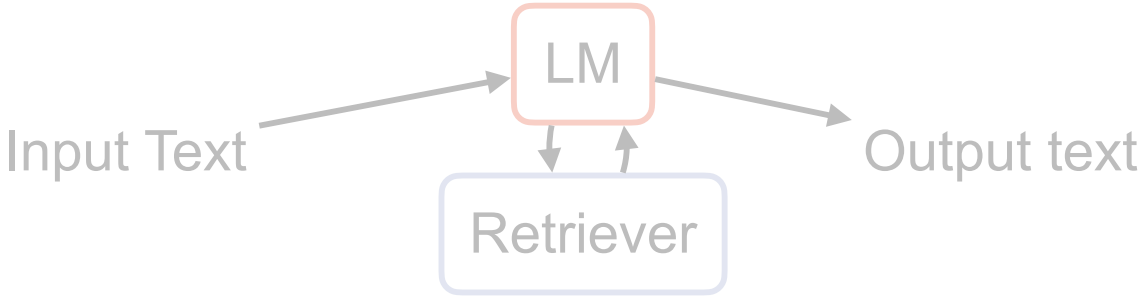
Parallel Interaction



Sequential Single Interaction



Sequential Multiple Interaction



COMP 4446 / 5046
Lecture 9, 2025

Efficiency

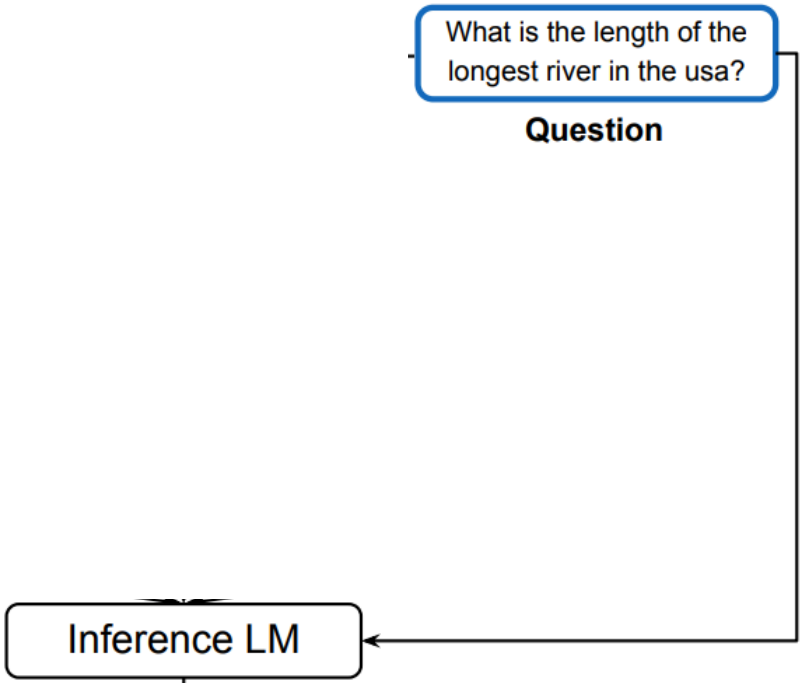
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Providing examples based on retrieval



Rubin et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

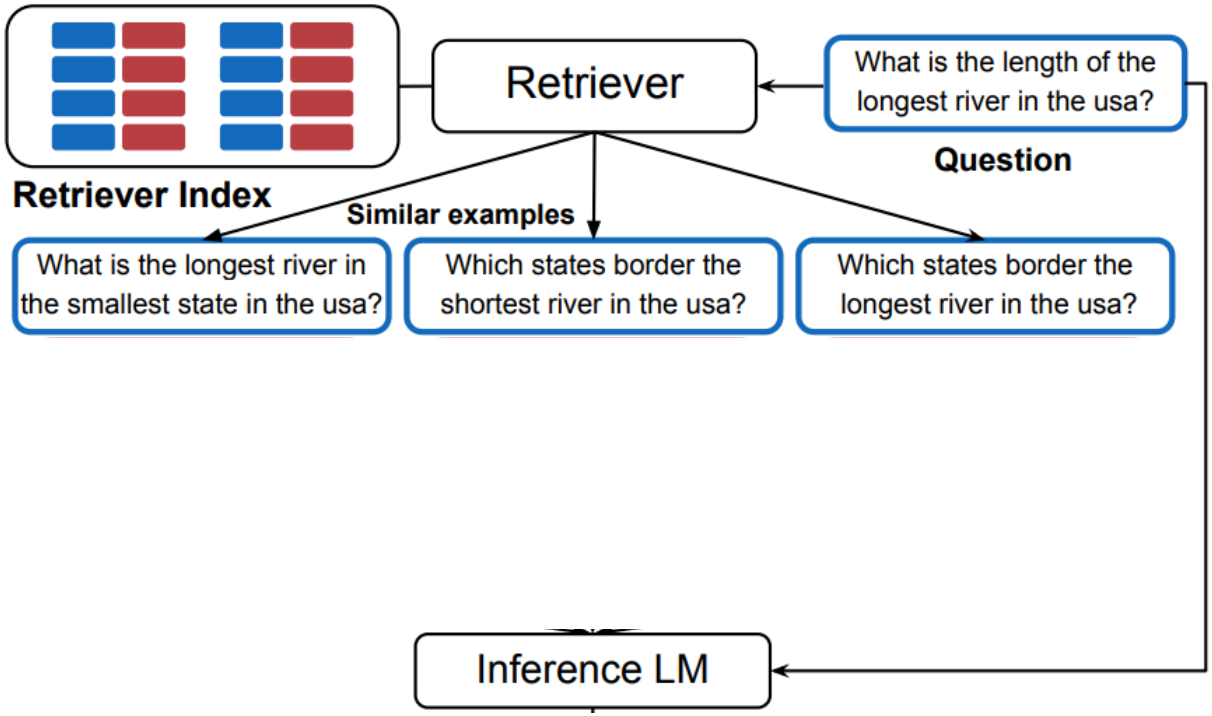
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Providing examples based on retrieval



Rubin et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

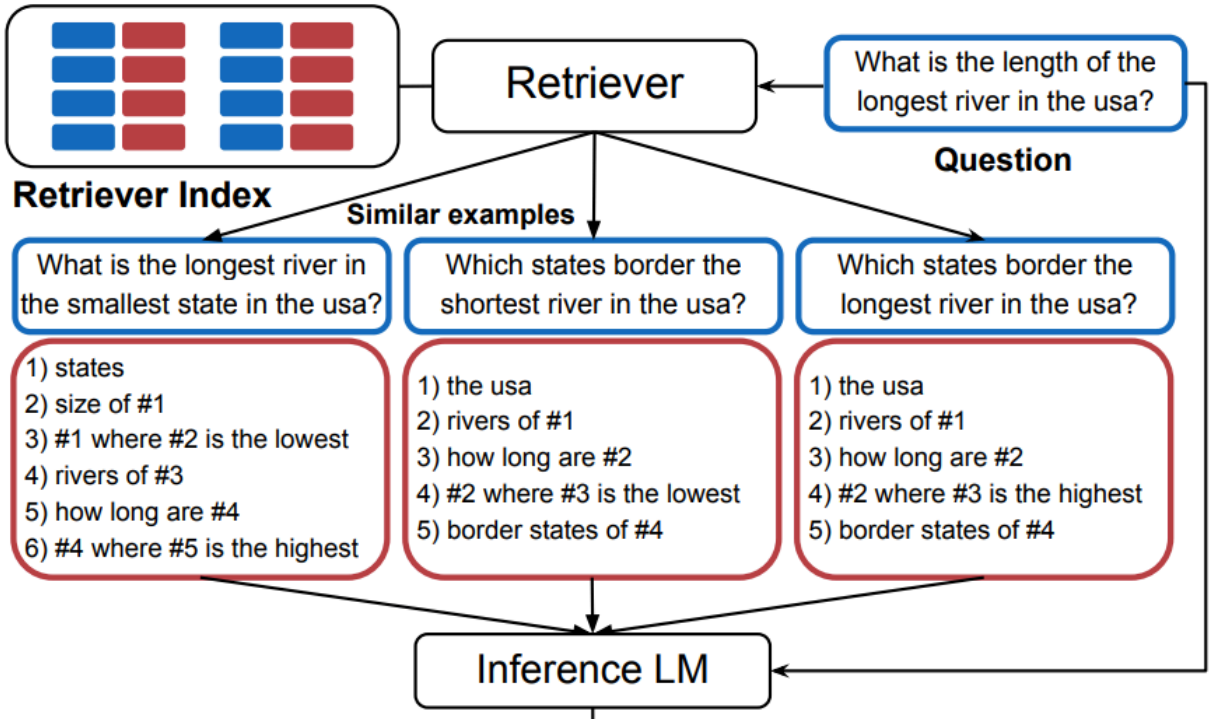
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Providing examples based on retrieval



Rubin et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

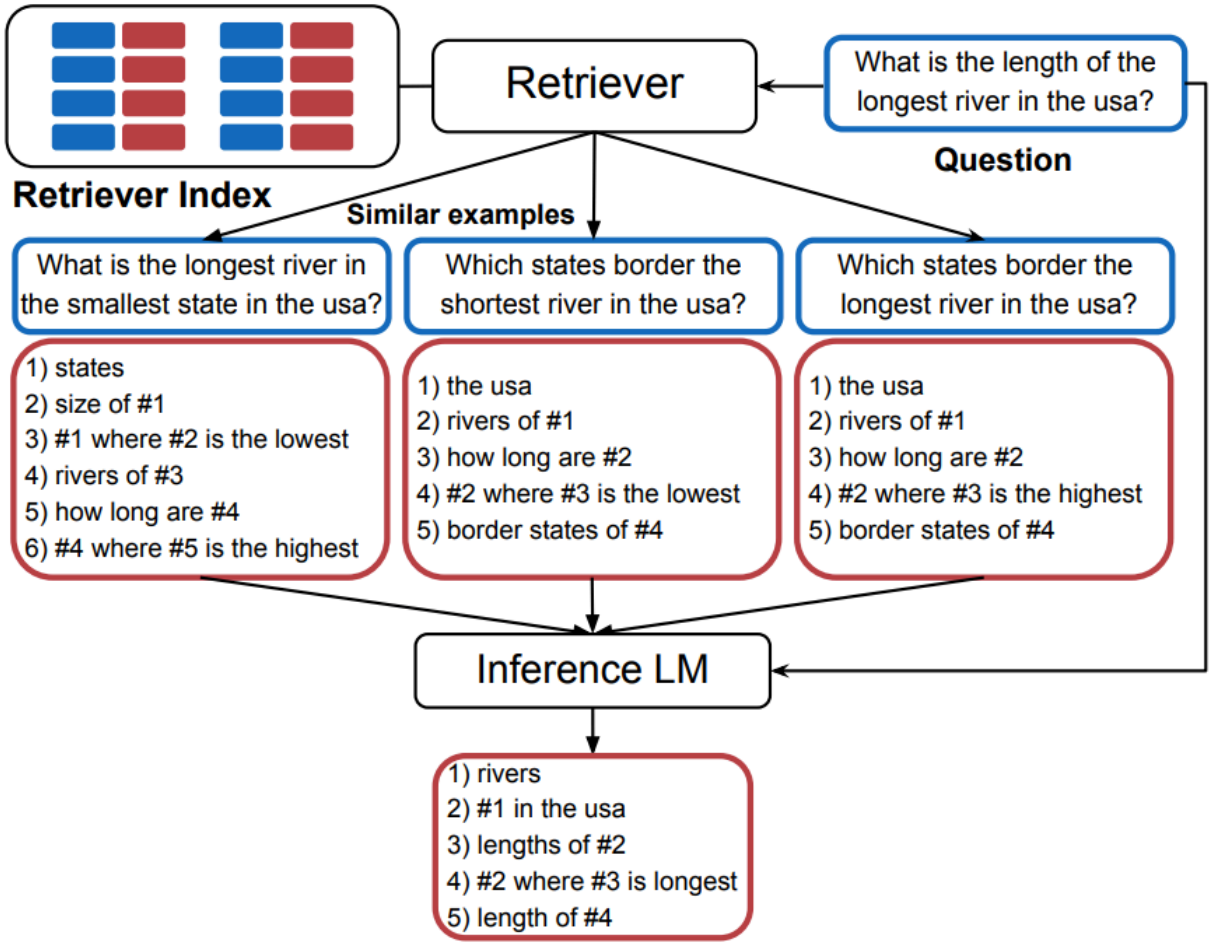
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Providing examples based on retrieval



Rubin et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

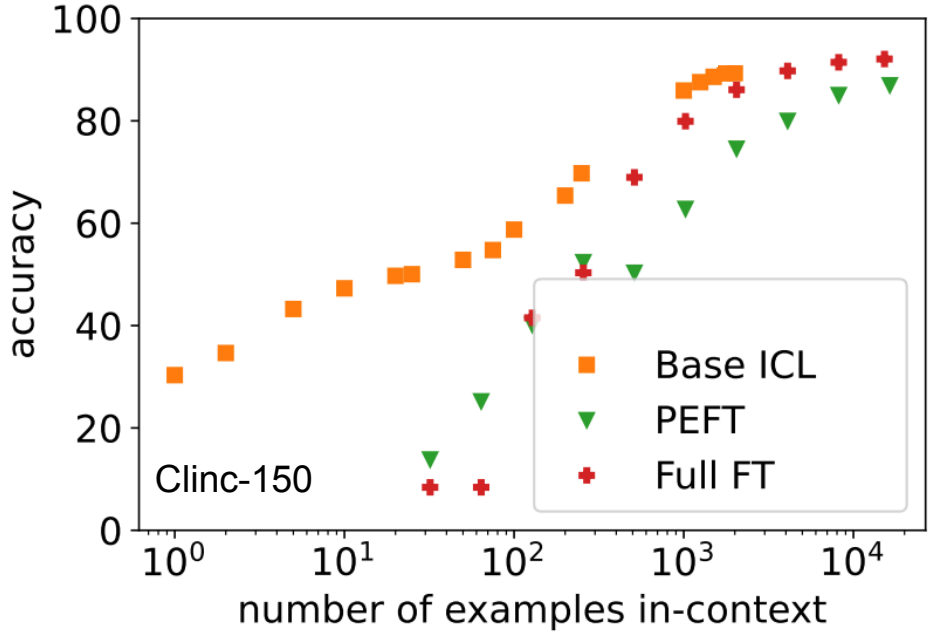
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

How does this interact with the number of examples?



Retrieval ICL:
Find the most relevant examples in your training data and use them in the prompt.

Bertsch et al (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

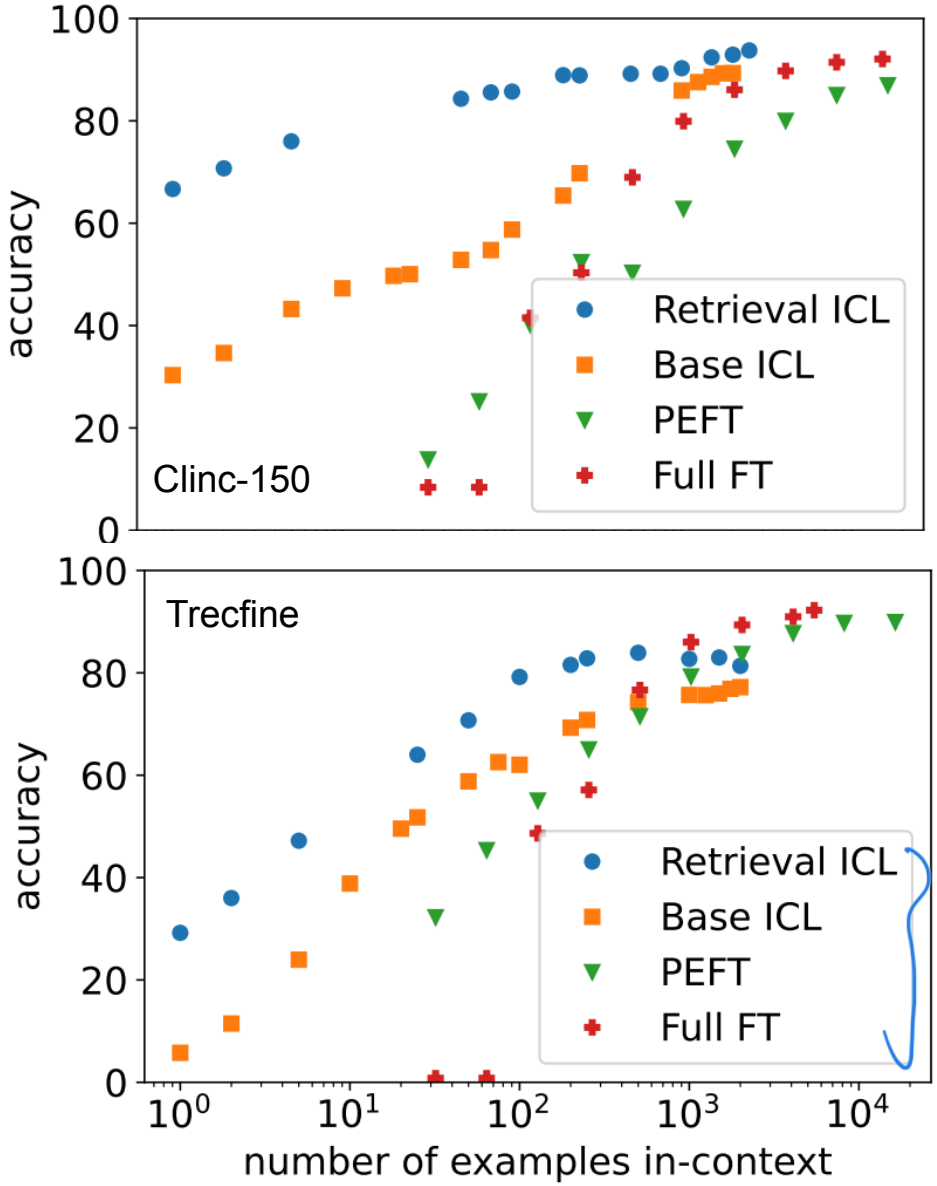
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

How does this interact with the number of examples?



Retrieval ICL:
Find the most relevant examples in your training data and use them in the prompt.

Compare with other

Bertsch et al (2024)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

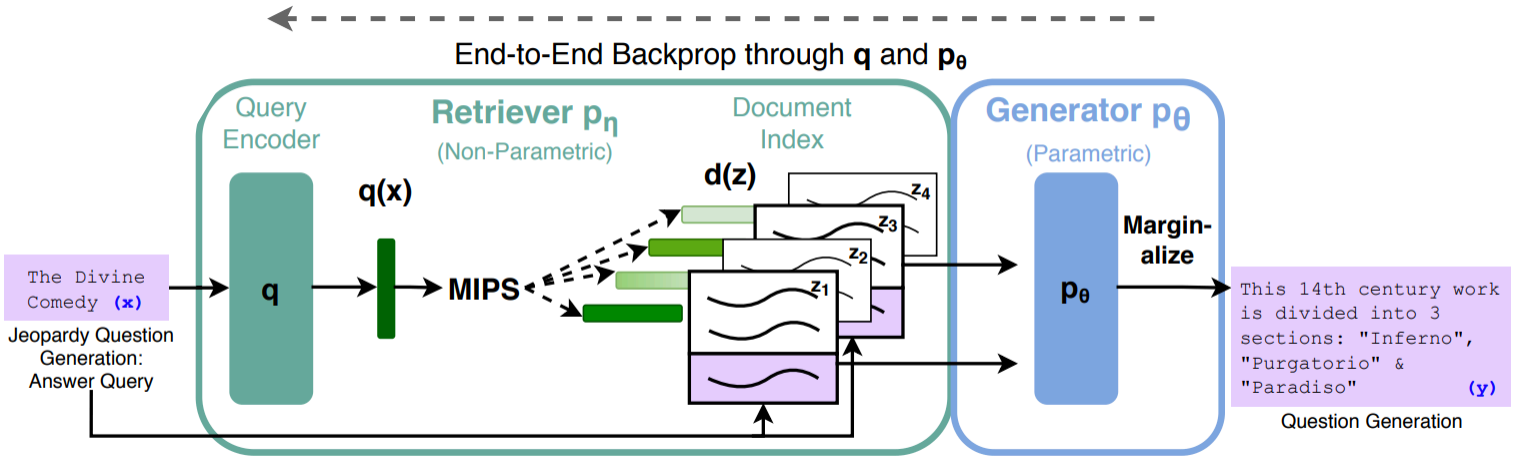
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

RAG: Retrieval Augmented Generation



retrieve relative info

Lewis et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

REALM

Unlabeled text, from pre-training corpus $(\mathcal{X})$

The [MASK] at the top of the pyramid (x)

Gus et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

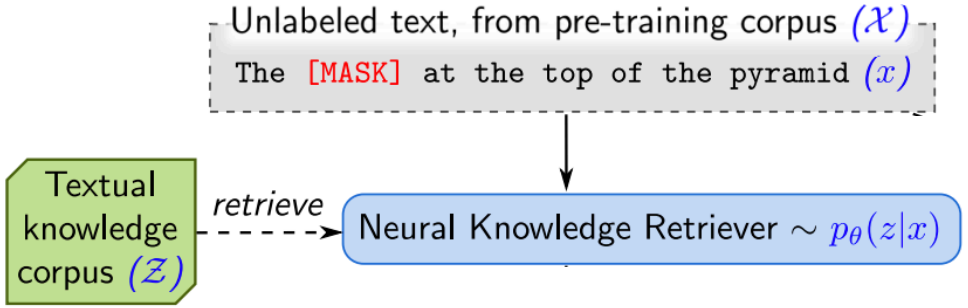
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

REALM



Gus et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

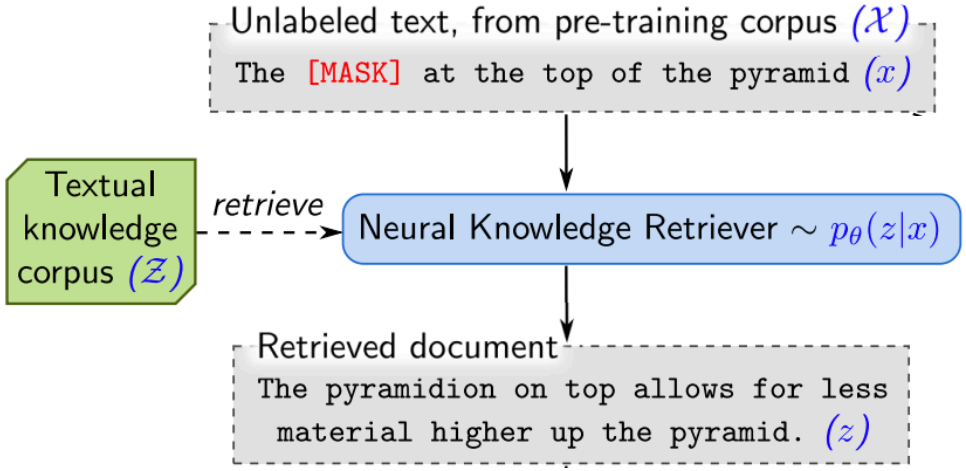
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

REALM



Gus et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

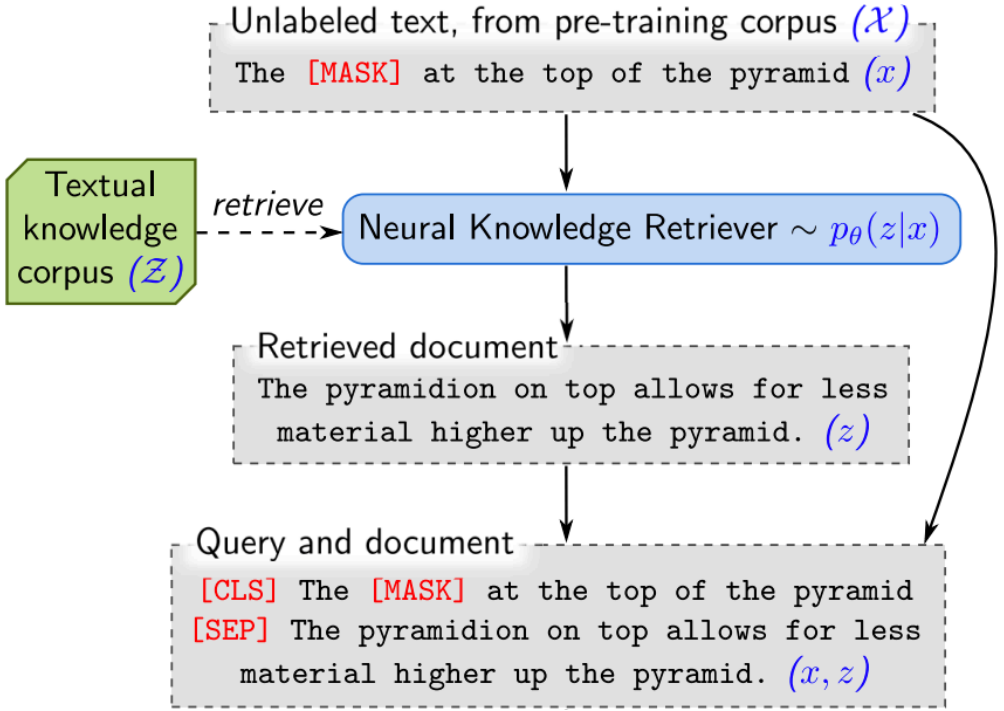
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

REALM



Gus et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

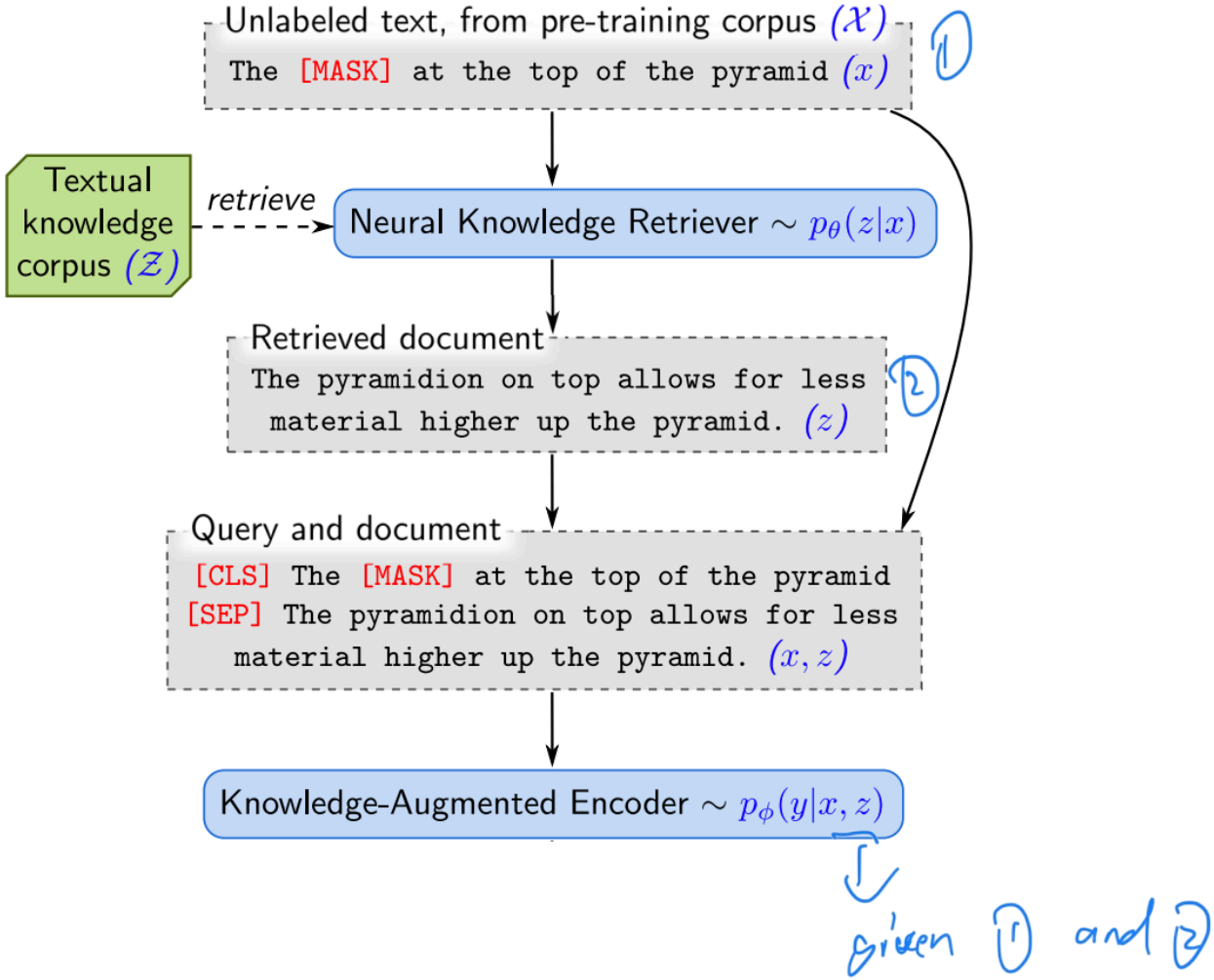
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

REALM



Gus et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

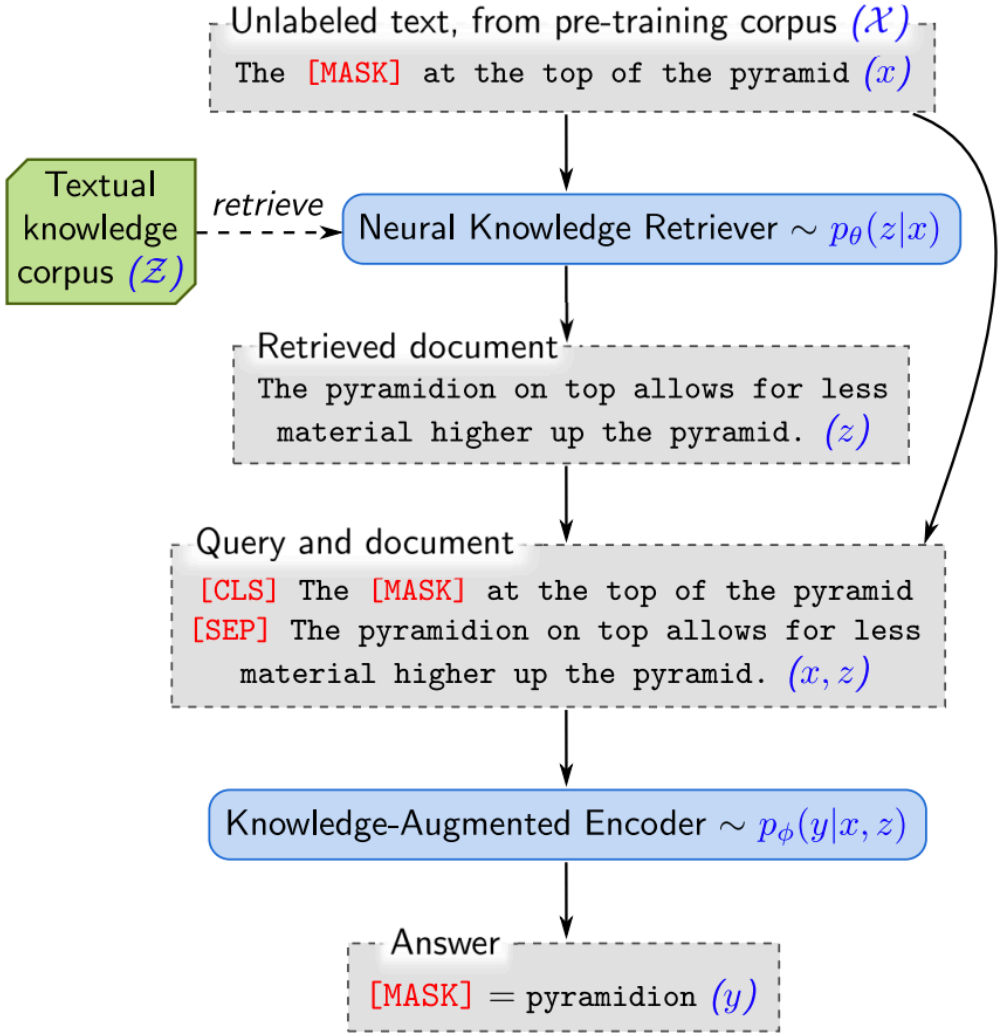
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

REALM



Gus et al (2020)

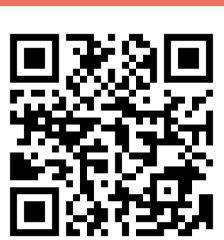
COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

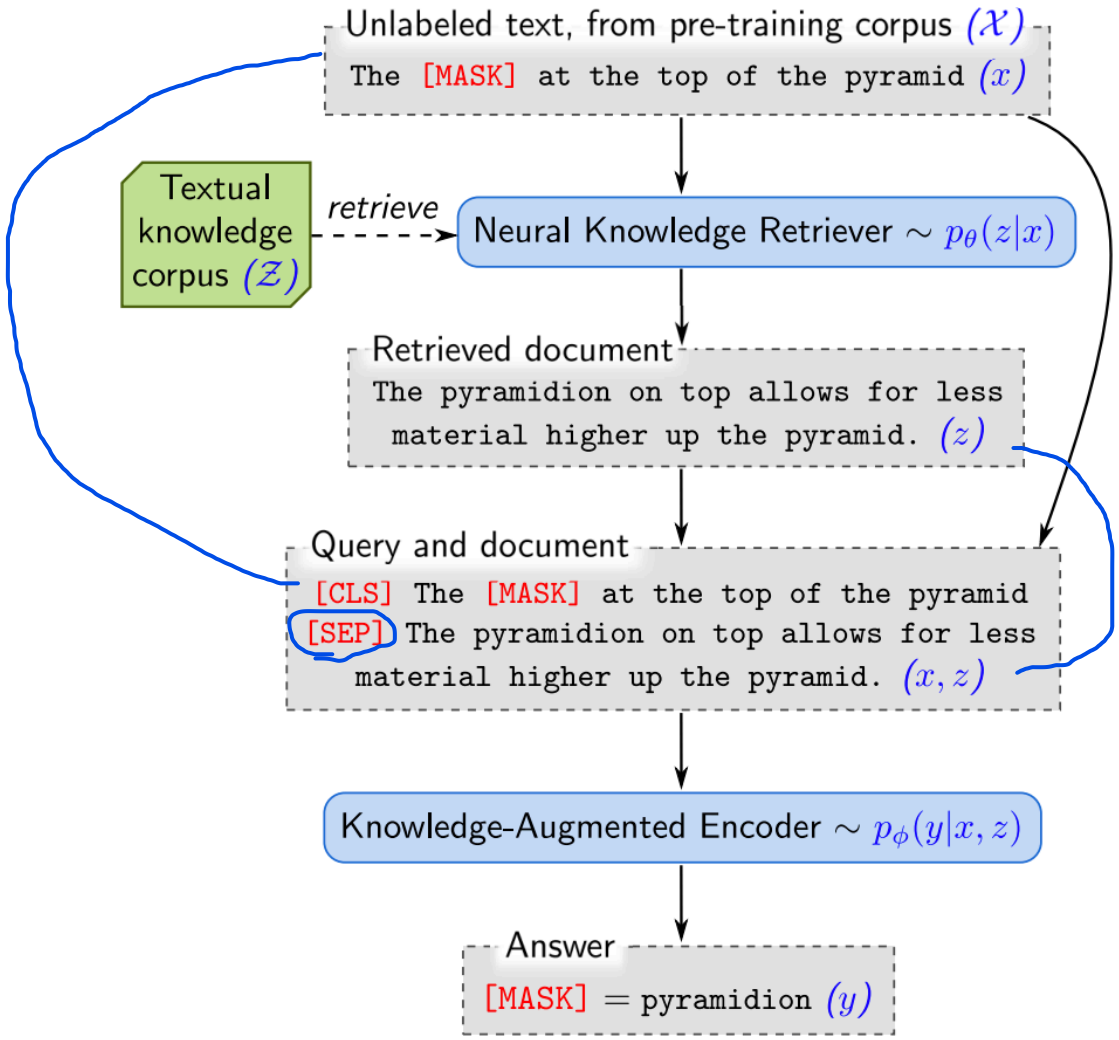
Retrieval Augmentation

Workshop Preview



menti.com 2376 2478

REALM



Gus et al (2020)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

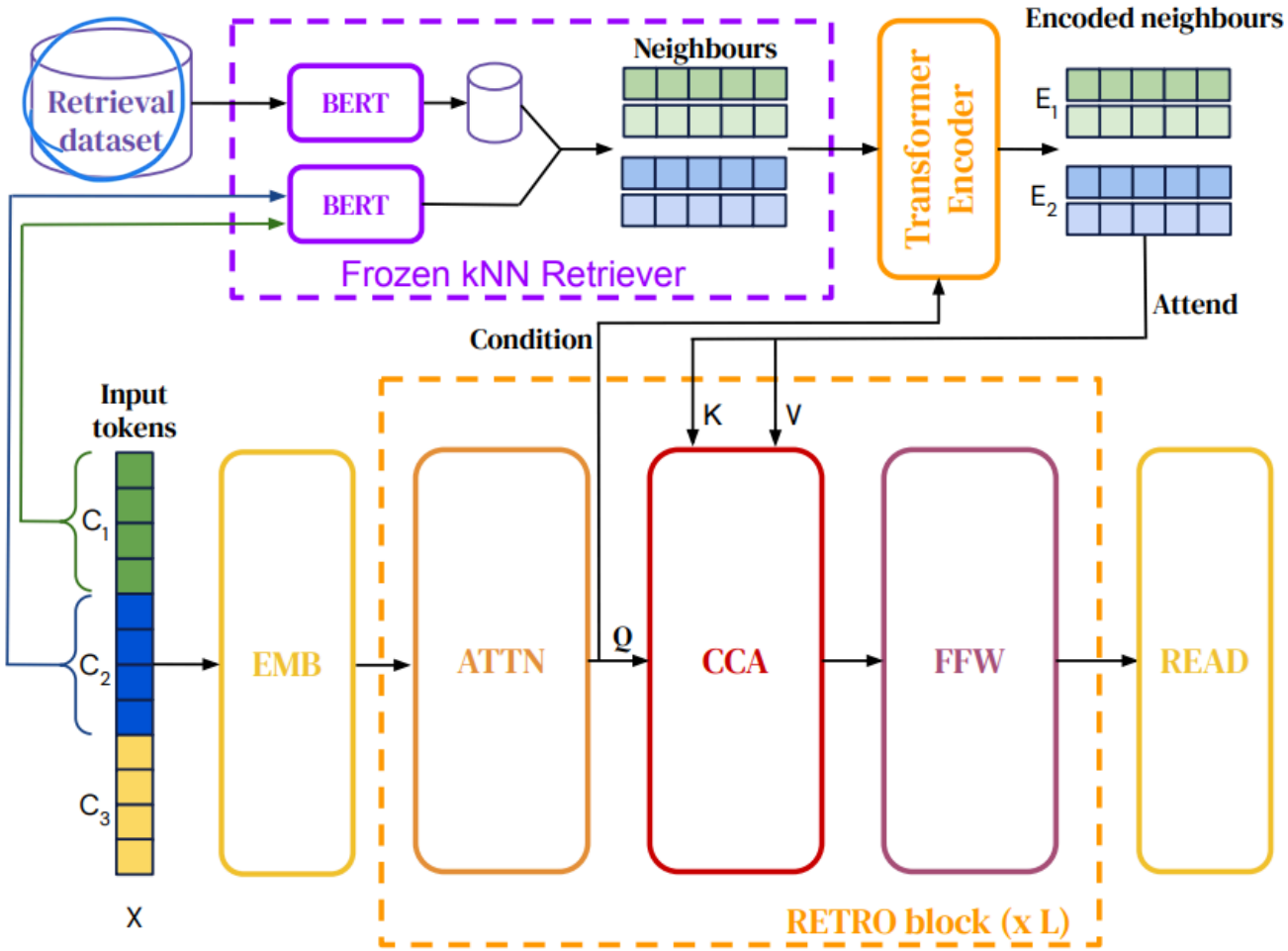
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

RETRO



Borgeaud et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

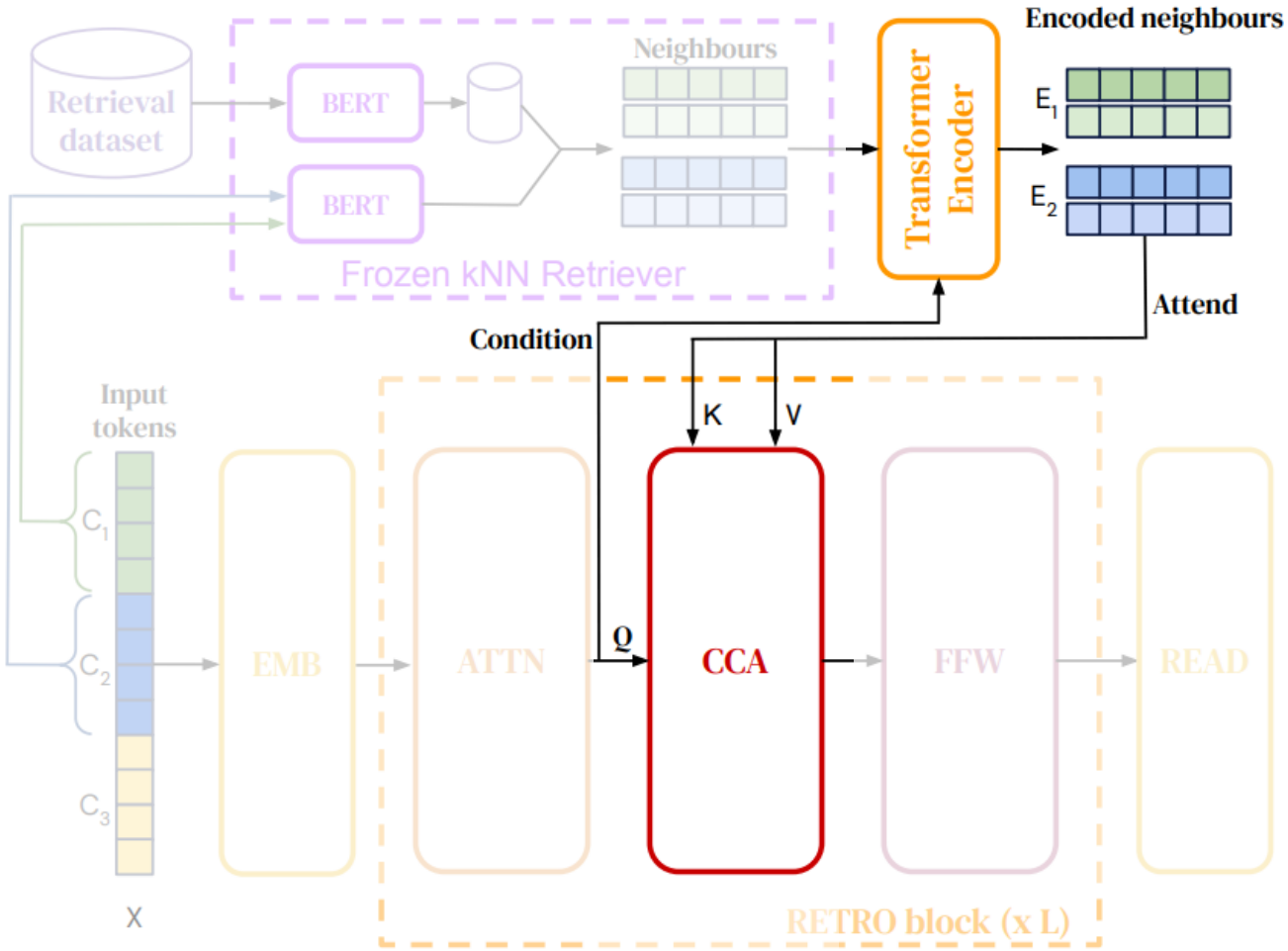
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

RETRO



Borgeaud et al (2022)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

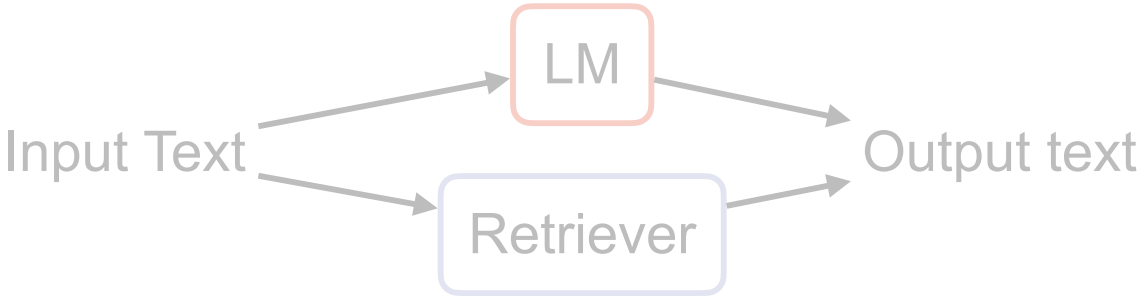
Retrieval Augmentation

Workshop Preview

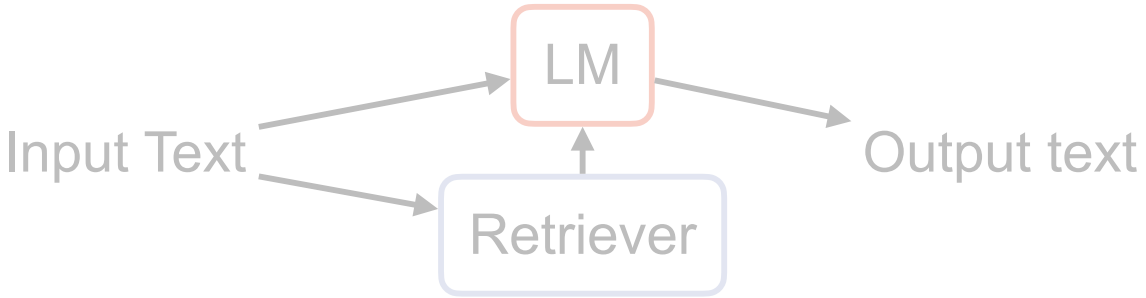
menti.com 2376 2478

Three general forms

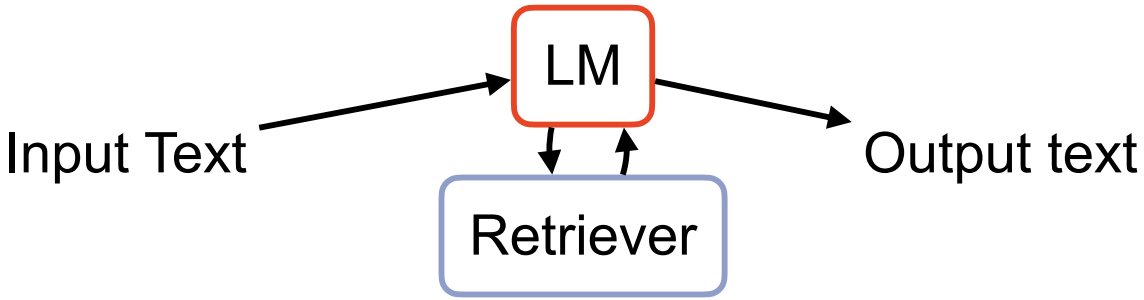
Parallel Interaction



Sequential Single Interaction



Sequential Multiple Interaction



COMP 4446 / 5046
Lecture 9, 2025

Efficiency

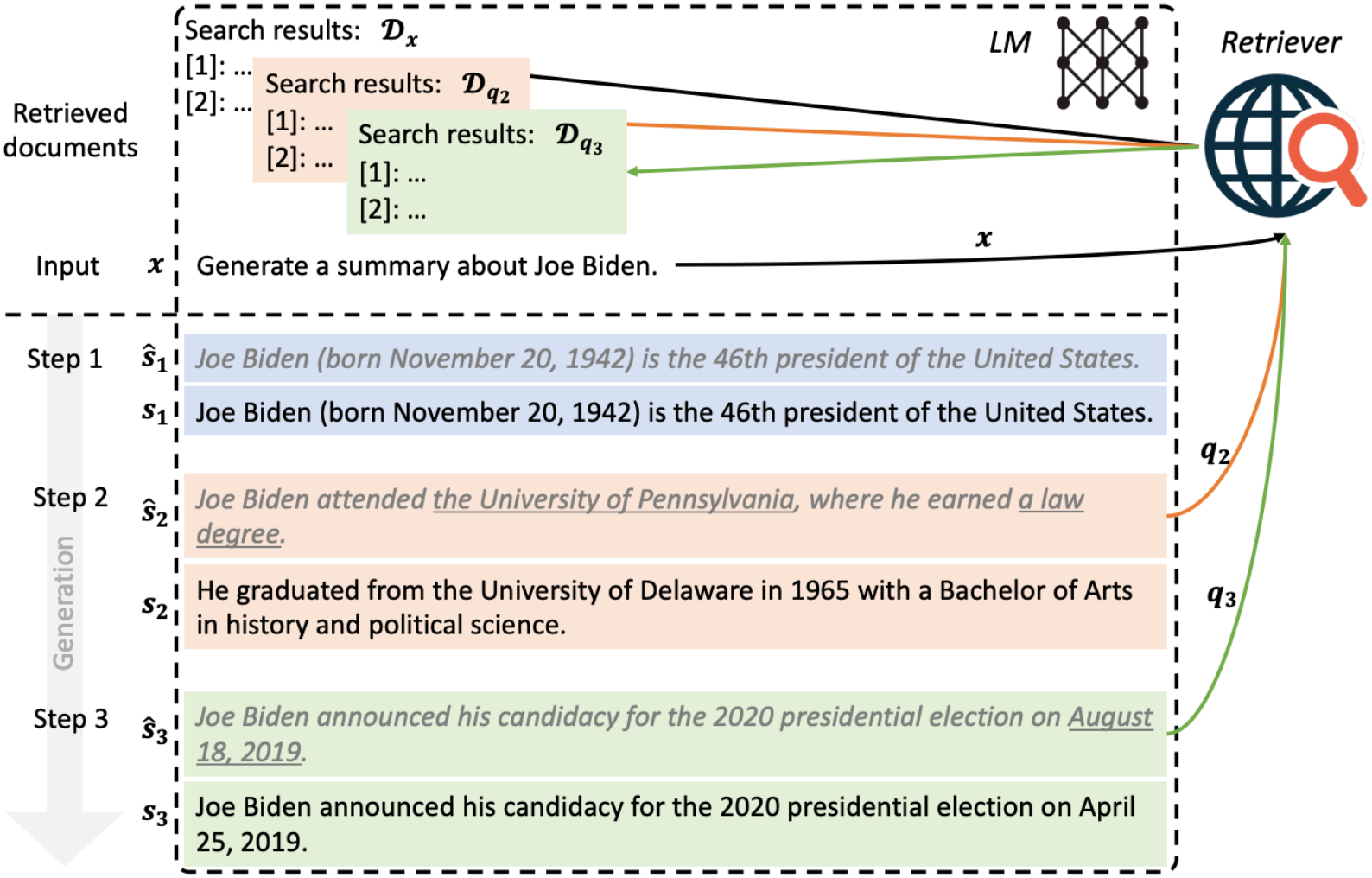
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

FLARE



Jiang et al (2023)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

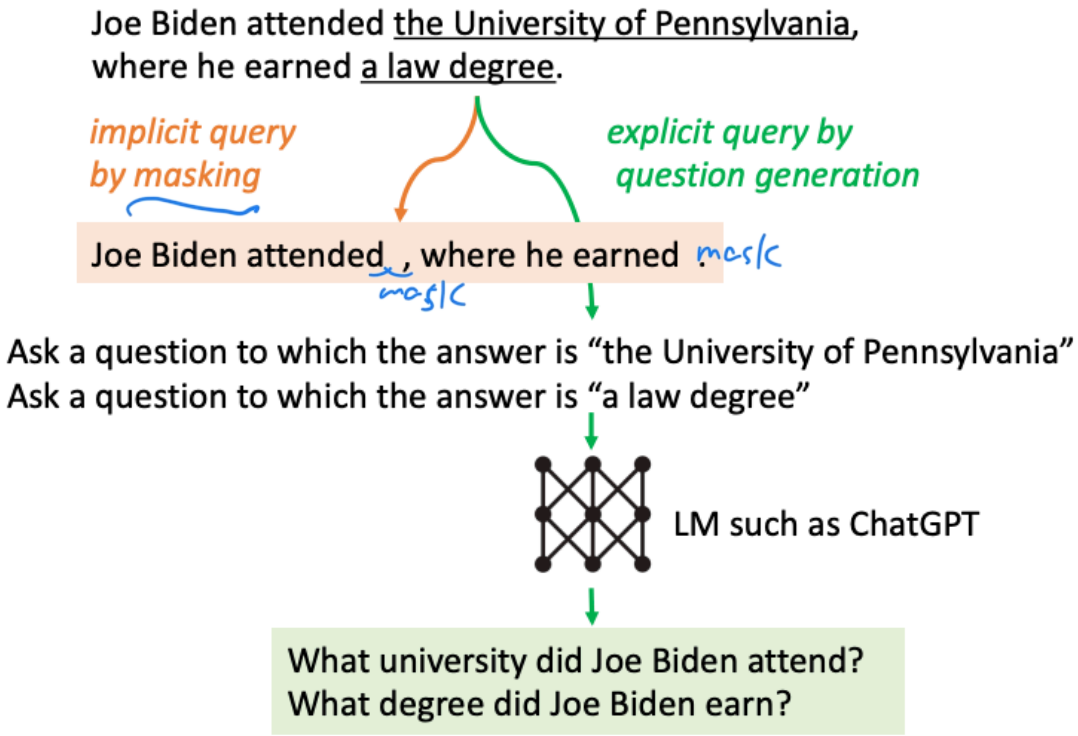
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

FLARE



Jiang et al (2023)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

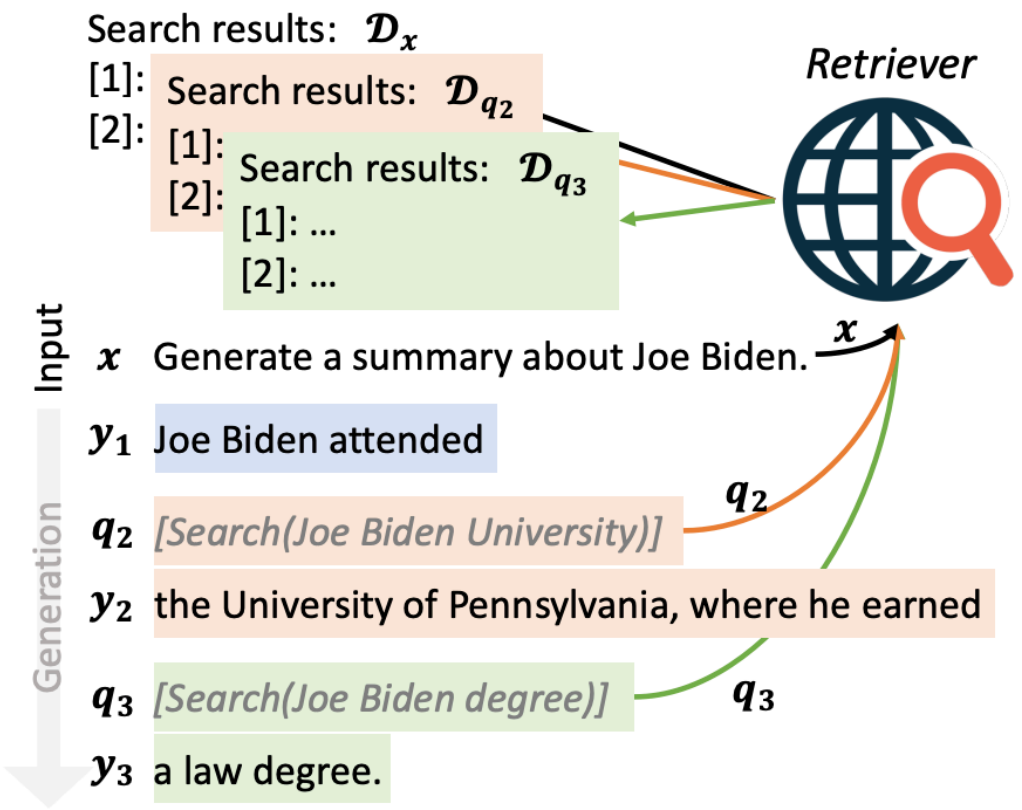
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

FLARE



Jiang et al (2023)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

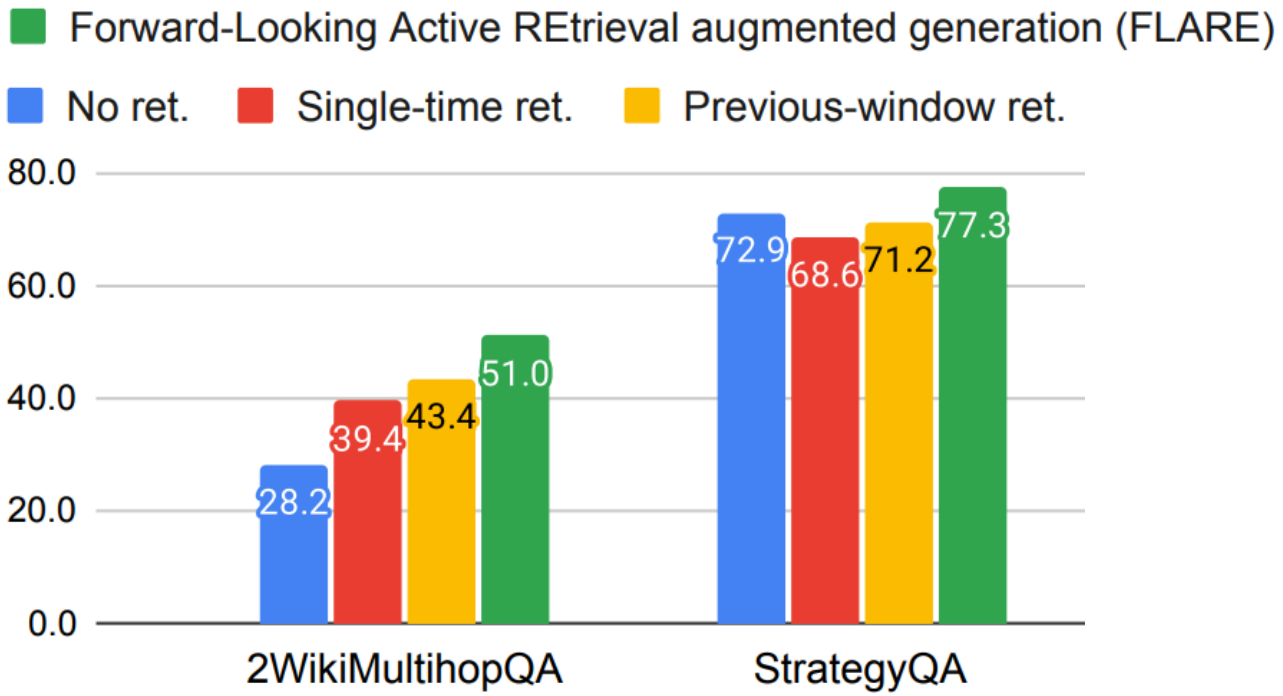
In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

FLARE



Jiang et al (2023)

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Recap: Retrieval Augmentation

General Idea: Retrieve data from a text collection or database and then use that in the generation process somehow.

Retrieval: Most common approach is to calculate similarity of vector representations between items in the database and some aspect of input.

Use: A variety of methods! Good to know about:

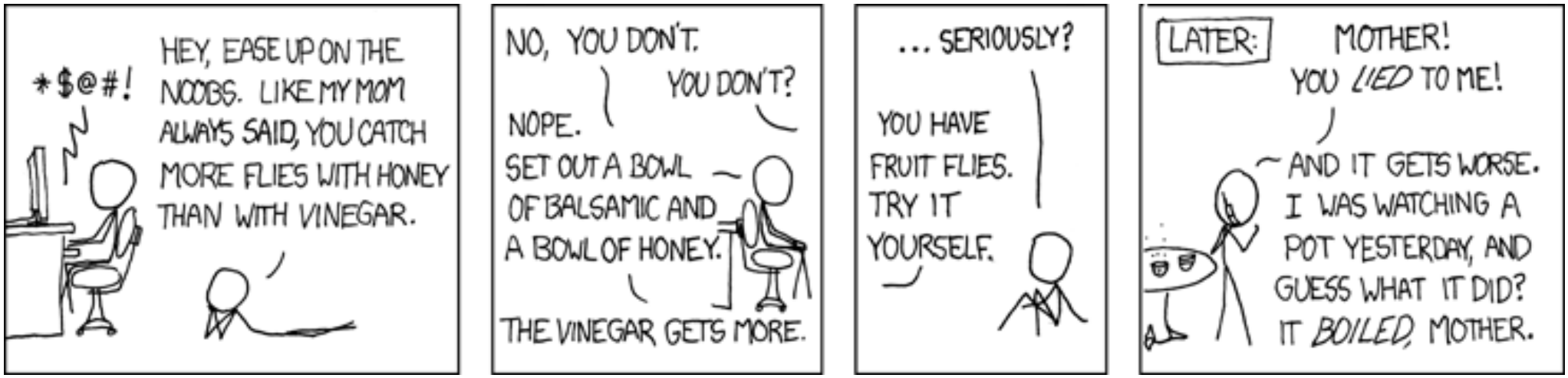
- Treating retrieval as an LM itself and doing model ensembling
- Provide the retrieved content in the prompt
- Interactive methods (related to Agents, which we will see later)

Jiang et al (2023)

3 minute Break - stretch and visit Menti

menti.com
2376 2478

Flies



[I don't know about houseflies, but we definitely caught a lot of fruit flies with our vinegar bowl. Hooray science!]

Source: https://xkcd.com/357/

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Workshop Preview

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Pinecone - a vector database widely used for RAG

Pre-work - Important set up, getting API keys!

COMP 4446 / 5046
Lecture 9, 2025

Efficiency

In-Context Learning

Retrieval Augmentation

Workshop Preview

menti.com 2376 2478

Muddy Card

Open shortly, closes at 7:05pm

https://saipll.shinyapps.io/student-interface/

If you do not wish to participate in the study, use the Ed form instead

Go to Ed → Lessons → Muddy Cards Lecture 9