CS224C: NLP for CSS

Deep Learning Highlights for CSS

Diyi Yang

Stanford CS

Lecture Overview

- BERT for Classification
- Prompting LLMs
- Using Prompting in CSS

BERT for Classification

Bidirectional encoder representations from Transformers

Context is the key

p(play | Elmo and Cookie Monster play a game .)

$\neq$

p(play | The Broadway play premiered yesterday .)

BERT demonstrated strong performances on a wide range of NLP tasks!

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Masked Language Modeling

Mask out k% of the input words, and then predict the masked words (k=15%)

Input: The man went to the [MASK] . He bought a [MASK] of milk .

Labels: [MASK] = store; [MASK] = gallon.

Next Sentence Prediction

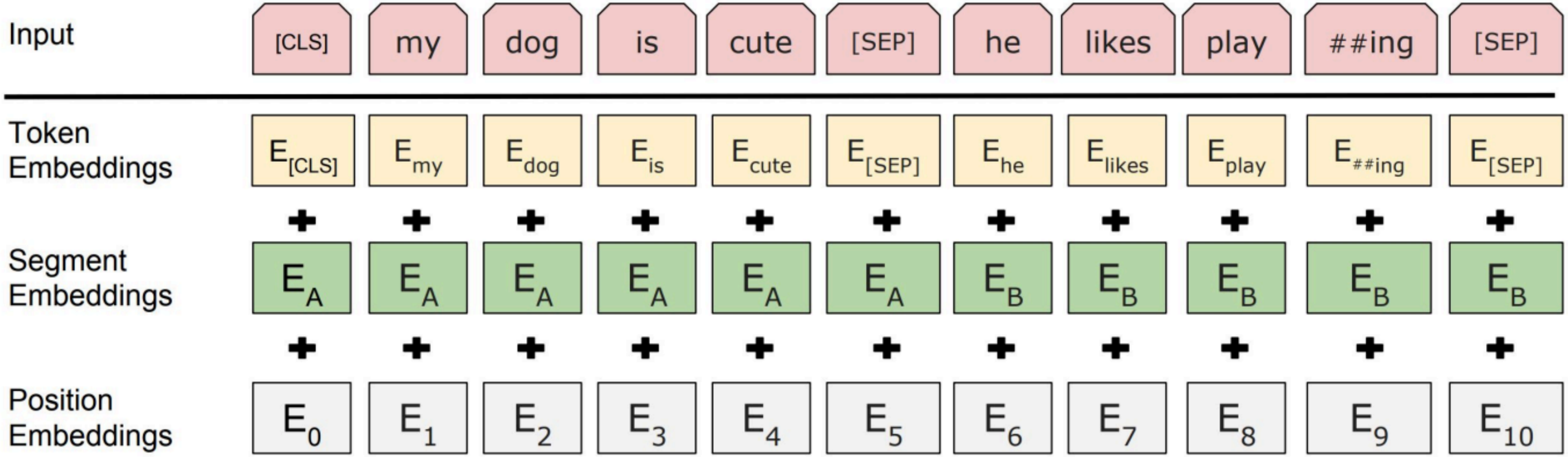
To learn relationship between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Input Representation

Each token is the sum of three embeddings

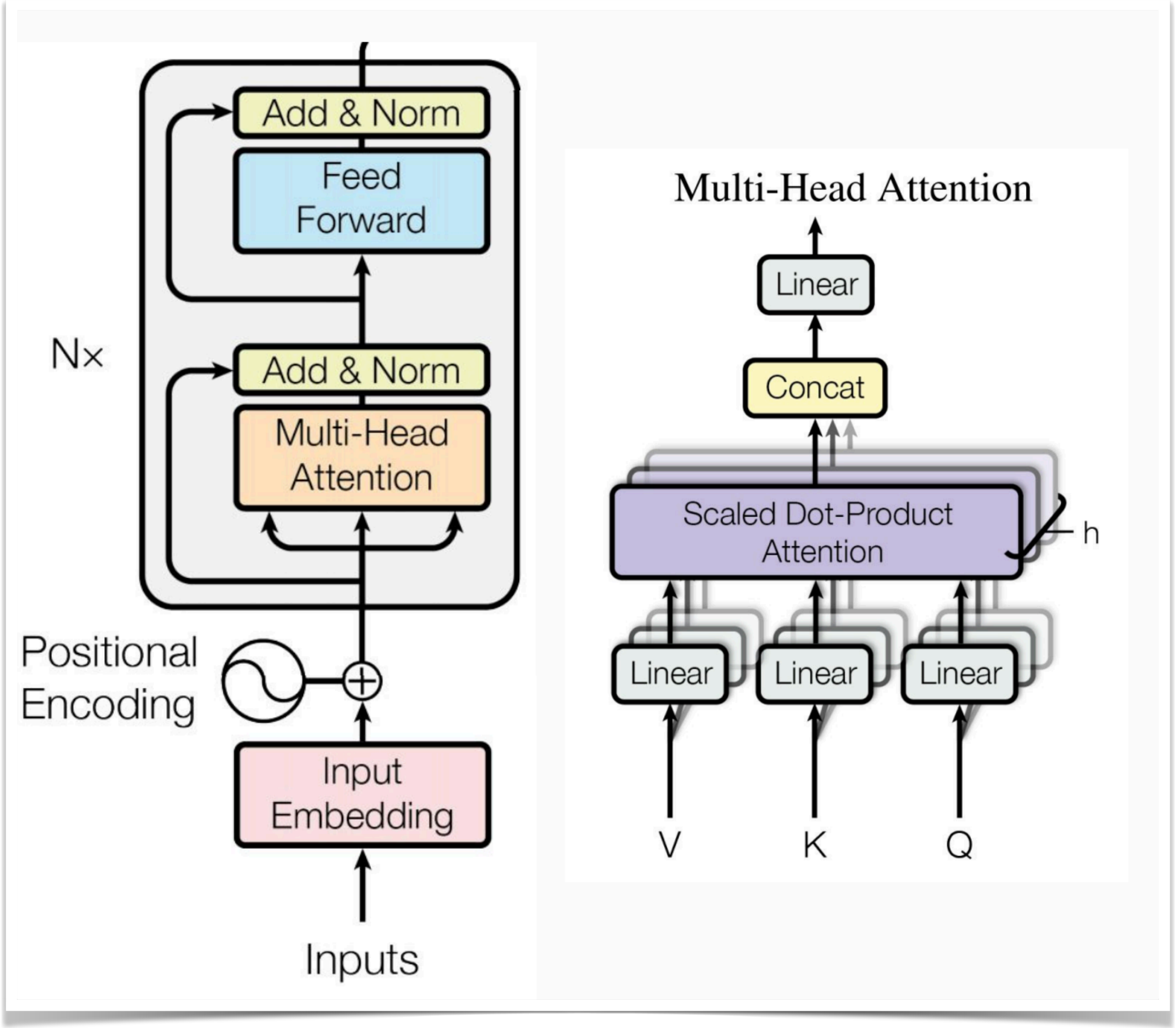


Model Architecture: Transformer

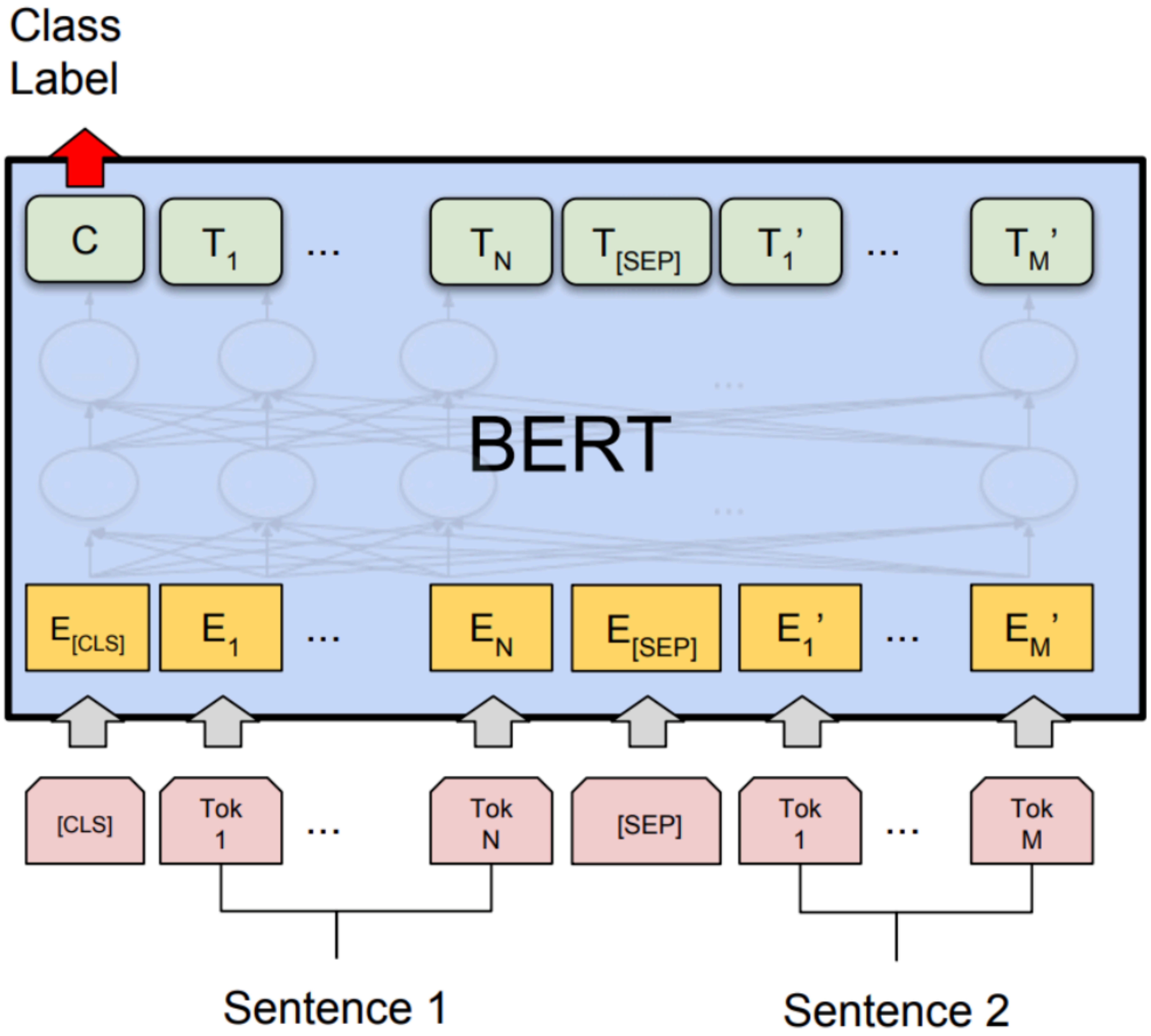
Multi-headed self attention to model context

Feed-forward layers to compute non-linear hierarchical features

Positional embeddings to allow model to learn relative positioning



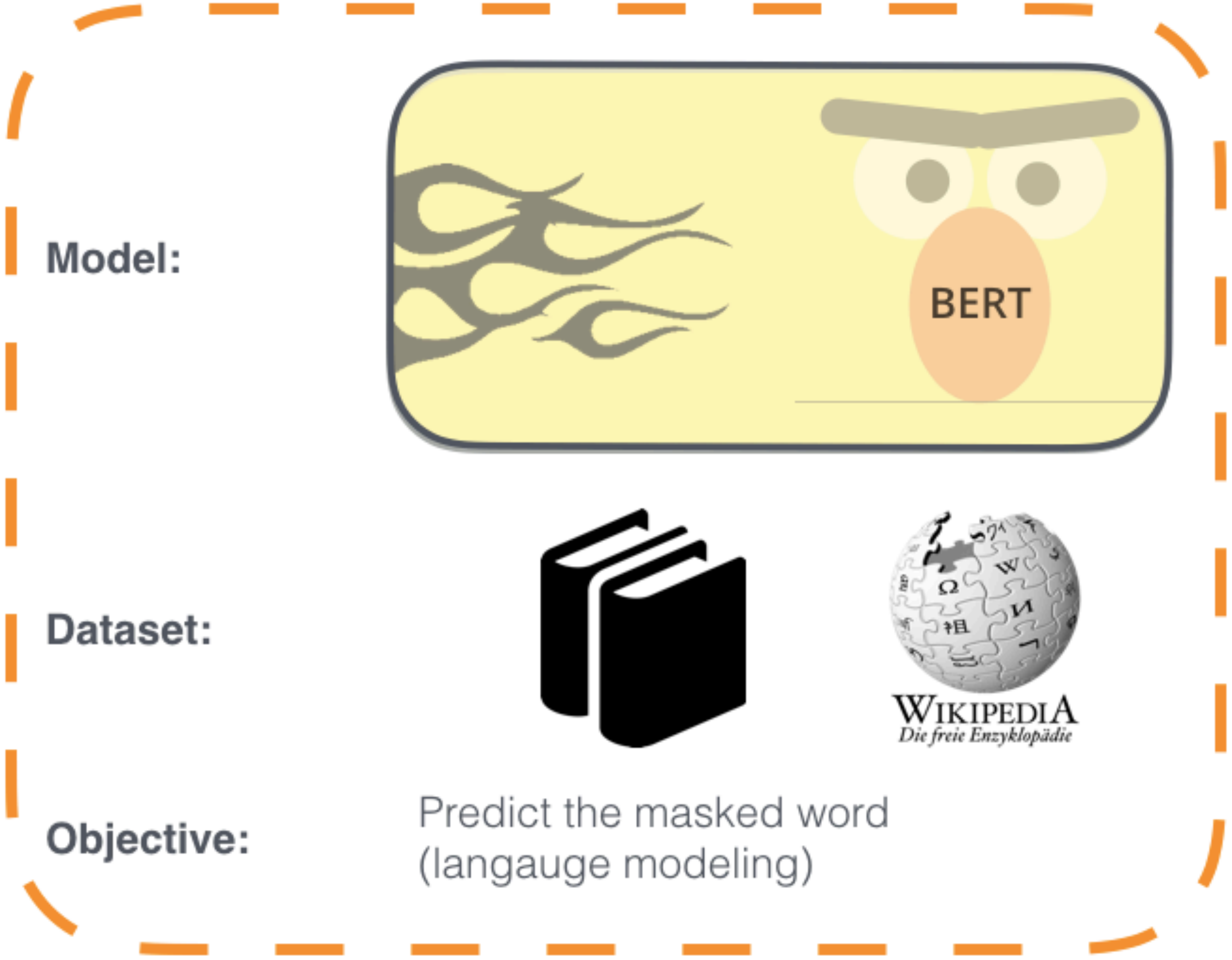
Link: https://nlp.stanford.edu/seminar/details/jdevlin.pdf


Class
Label
C
T_1
...
T_N
$T_{[SEP]}$
T_1'
...
T_M'
BERT
$E_{[CLS]}$
E_1
...
E_N
$E_{[SEP]}$
E_1'
...
E_M'
[CLS]
Tok 1
...
Tok N
[SEP]
Tok 1
...
Tok M
Sentence 1
Sentence 2

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

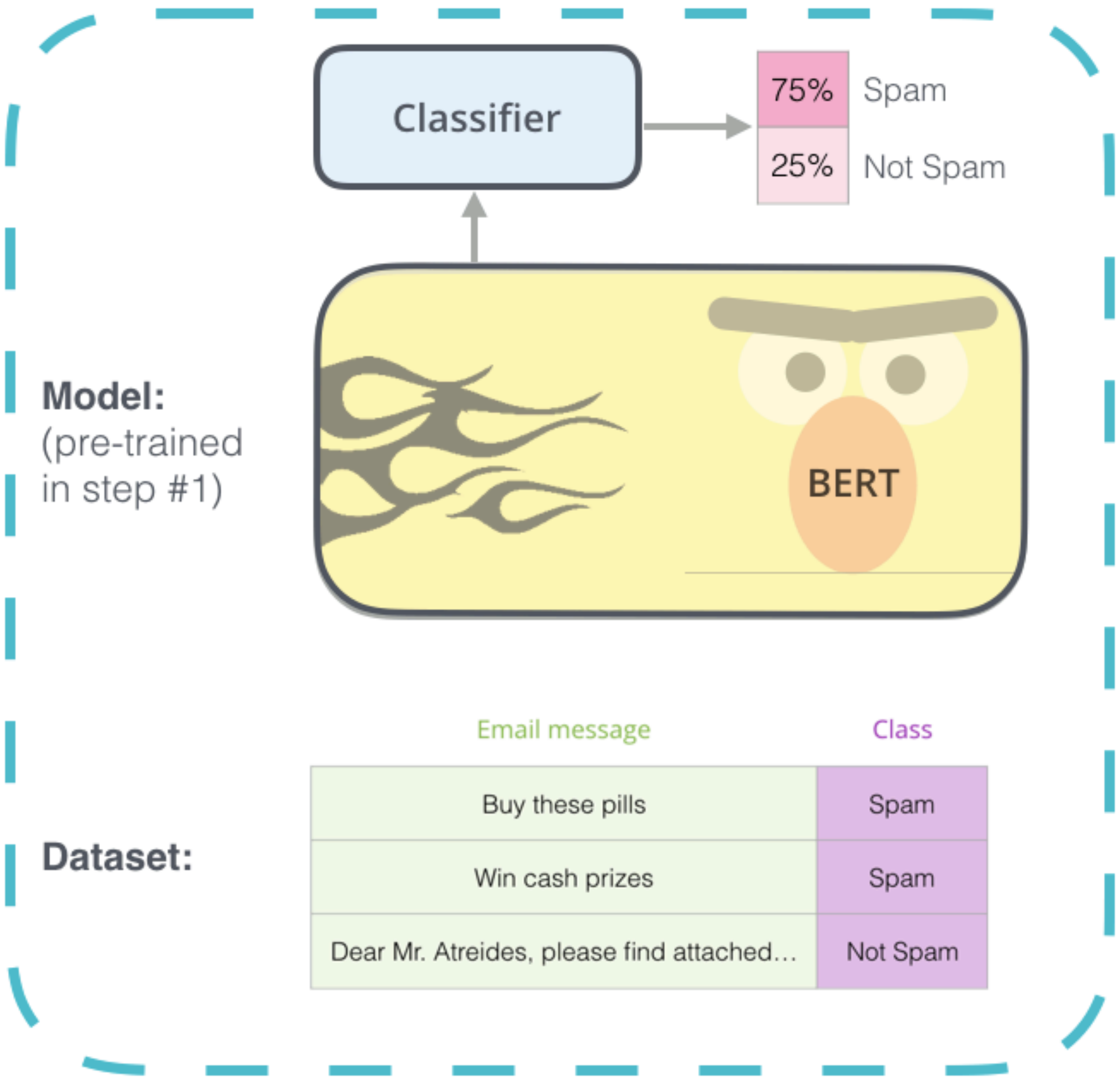
The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step



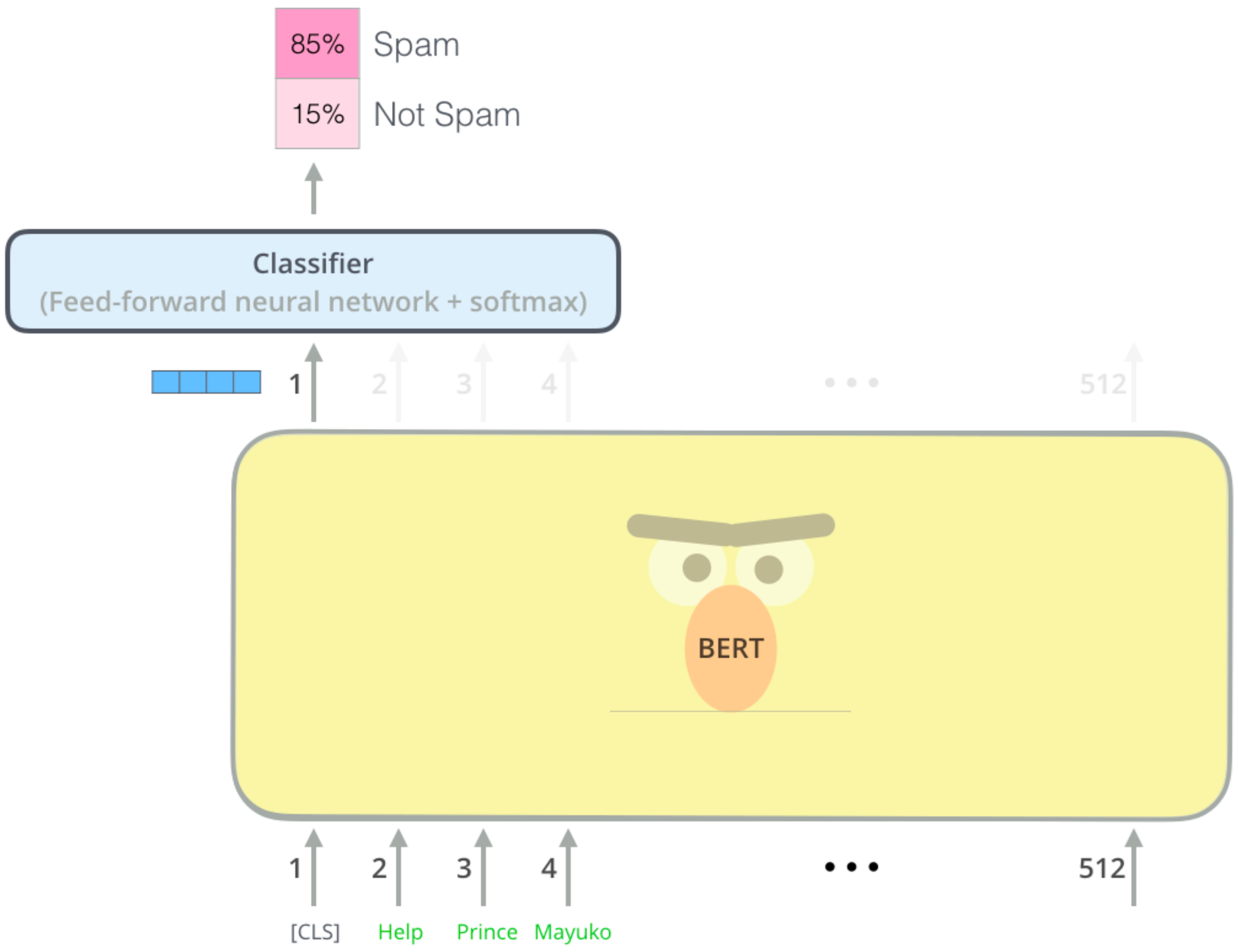
2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step



https://jalammar.github.io/illustrated-bert/

How to use BERT for Classification
(e.g., sentiment, fact-checking, rumors)



Pros and Cons of BERT for CSS

- Strong prediction performance
- Fine-tuning on top of pertained representations
- Prediction and representation can be hard to interpret
- Subject to biases in these learned representations
- Require computational resources

Lecture Overview

- BERT for Classification
- **Prompting LLMs**

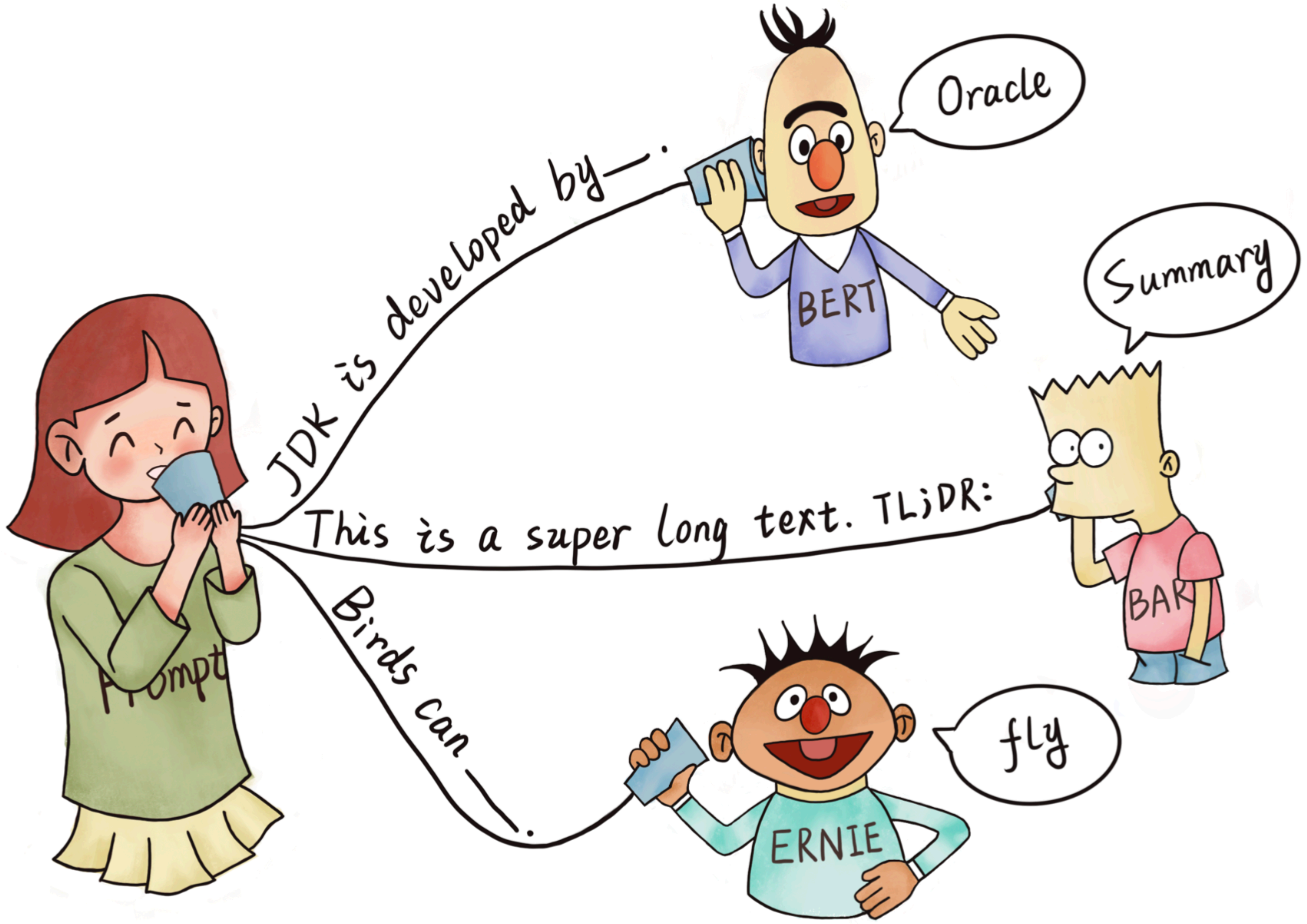
Prompt for LLMs

Fine-tuning LLMs (e.g., GPT-3 175B) is often not feasible due to its large size

Prompts (or **in-context learning**) were then introduced and used

Prompting

Prompting: encourage a pre-trained model to make particular predictions by providing a "prompt" specifying the task to be done.



Liu Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *arXiv 2021*

Intuition of Prompting

Sentiment	The value I got was the sum total of the popcorn and the drink. Overall, it was a _boring_ movie!
World knowledge	Peking University is located in _Beijing_, China.
Syntactic categories	I put _the_ fork down on the table.
Coreference	The woman walked across the street, checking for traffic over _her_ shoulder.
Semantic categories	I went to the ocean to see the fish, turtles, seals, and _crabs_.
Reasoning	Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the _kitchen_.

LAMA Benchmark, from Petroni, Fabio, et al. "Language models as knowledge bases?." *EMNLP 2019*

Slides adjusted from John Hewitt, Stanford CS 224n

Intuition of Prompting

Sentiment

World knowledge

Syntactic categories

Coreference

Semantic categories

Reasoning

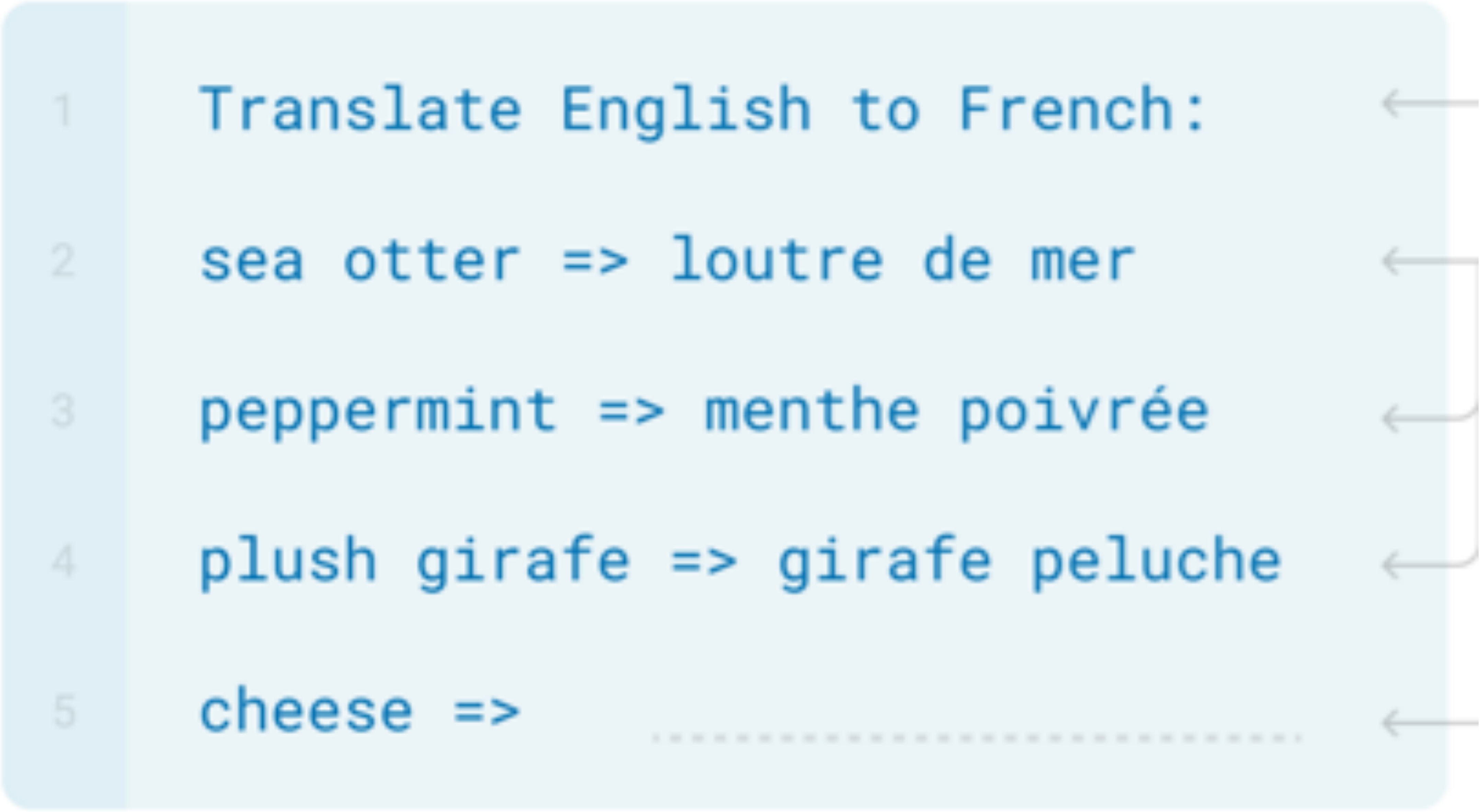
Pre-trained models learn **various types of knowledge**.

The knowledge is useful **across NLP tasks.**

These knowledge **can be surfaced** with “templates” (prompt).

Prompting

Zero/few-shot Prompting



[Brown et al., 2020]

Traditional Fine-tuning

One model, N tasks (e.g., GPT-3, 175B)

Sentiment analysis classifier

```
Decide whether a Tweet's sentiment is
positive, neutral, or negative.

Tweet: "I loved the new Batman movie!"
Sentiment: Positive
```

Q&A

Targeted world knowledge

```
Chatbot: I am a ML/AI language model tutor
You: What is a language model?
Chatbot: A language model is a statistical
model that describes the probability of a
word given the previous words.
```

Translation / NL2code

```
Create a SQL request to find all users who
live in California and have over 1000
credits: SELECT * FROM users WHERE
state='CA' AND credits > 1000;
```

Semantic of query

Syntax of code

Summarization

```
A neutron star is the collapsed core of a massive
supergiant star, which had a total mass of between 10
and 25 solar masses, possibly more if the star was
especially metal-rich.[1] Neutron stars are the
smallest and densest stellar objects, excluding black
holes and hypothetical white holes, quark stars, and
strange stars.[2] Neutron stars have a radius on the
order of 10 kilometres (6.2 mi) and a mass of about 1.4
solar masses.[3] They result from the supernova
explosion of a massive star, combined with
gravitational collapse, that compresses the core past
white dwarf star density to that of atomic nuclei.

TL;DR: A neutron star is the collapsed core of a
massive supergiant star. These ultra-dense objects are
incredibly fascinating due to their strange properties
and their potential for phenomena such as extreme
gravitational forces and a strong magnetic field.
```

The semantic of TL;DR

GPT-3 API, from OpenAI: https://beta.openai.com/examples

Emergent Zero-shot Learning

GPT-2 beats SoTA on language modeling benchmarks with no task-specific fine-tuning

Context: “Why?” “I would have thought you’d find him rather dry,” she said. “I don’t know about that,” said Gabriel. “He was a great craftsman,” said Heather. “That he was,” said Flannery.
Target sentence: “And Polish, to boot,” said ______.
Target word: Gabriel

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**

LAMBADA (language modeling w/ long discourse dependencies)
[Paperno et al., 2016]

Emergent few-shot learning

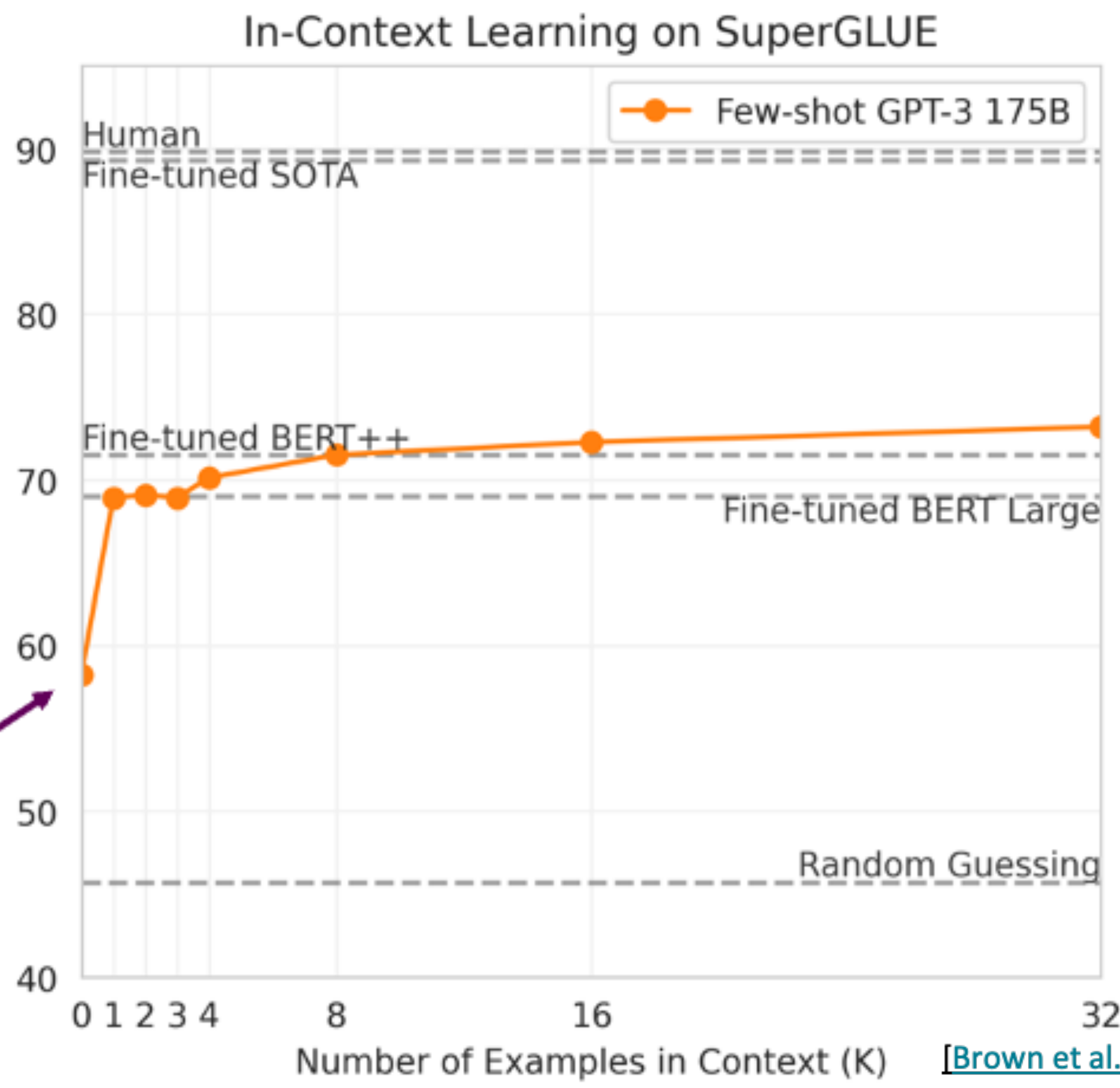
Specify a task by simply prepending examples of the task before your example

Also called in-context learning, to stress that no gradient updates are performed when learning a new task

Emergent few-shot learning

Zero-shot

```
Translate English to French:
cheese =>
```

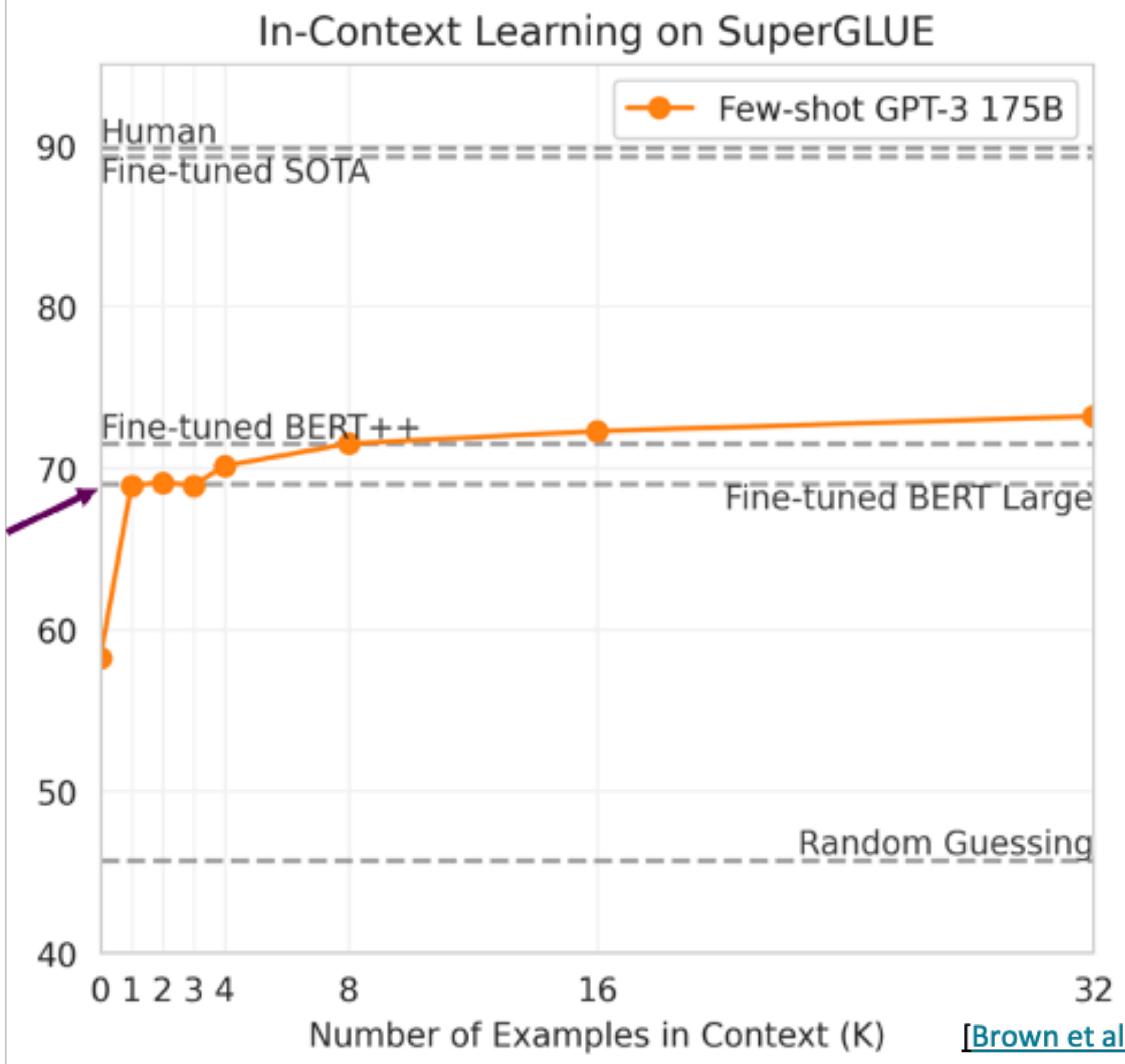


[Brown et al., 2020]

Emergent few-shot learning

One-shot

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

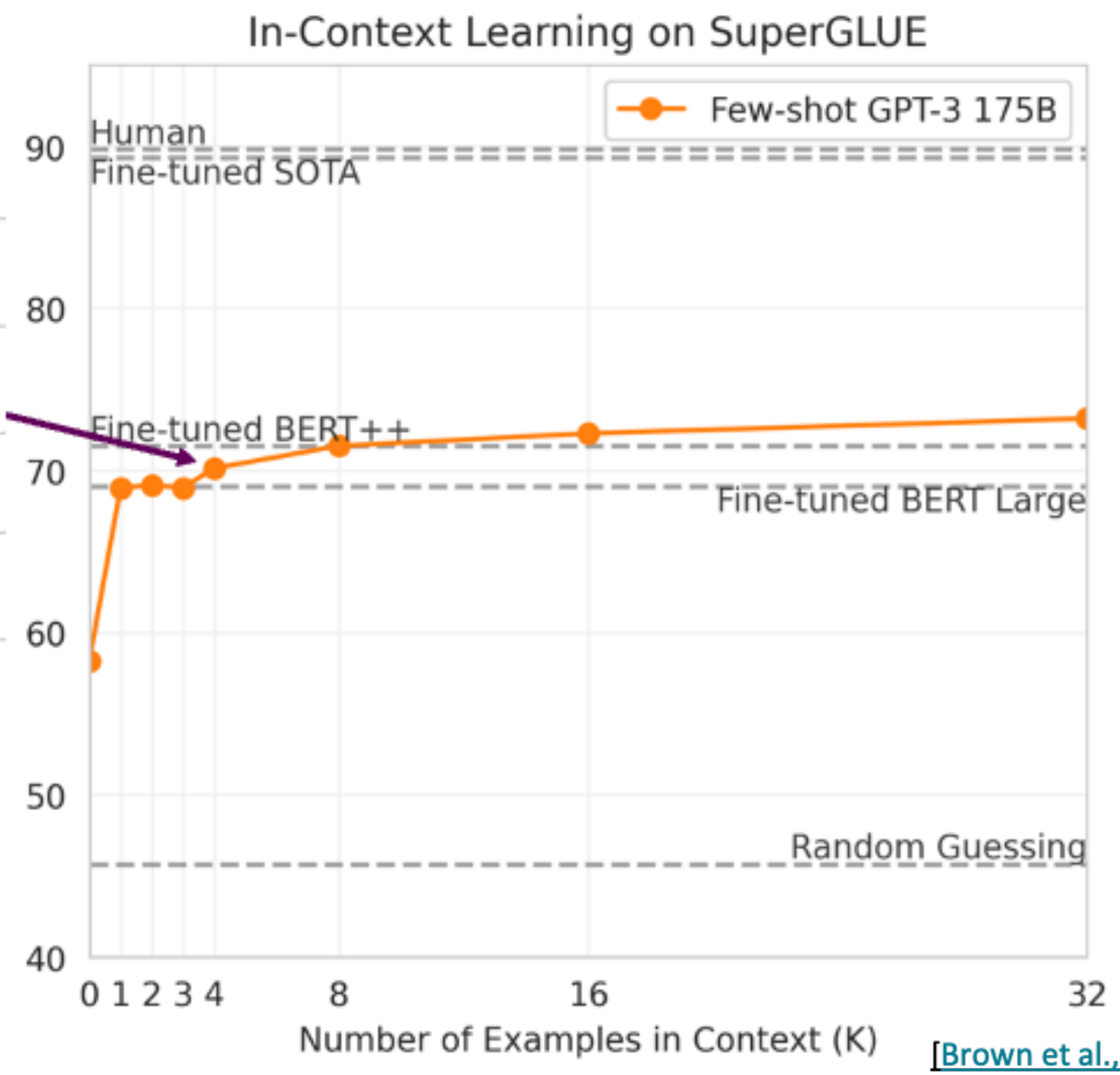


[Brown et al., 2020]

Emergent few-shot learning

Few-shot

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```



[Brown et al., 2020]

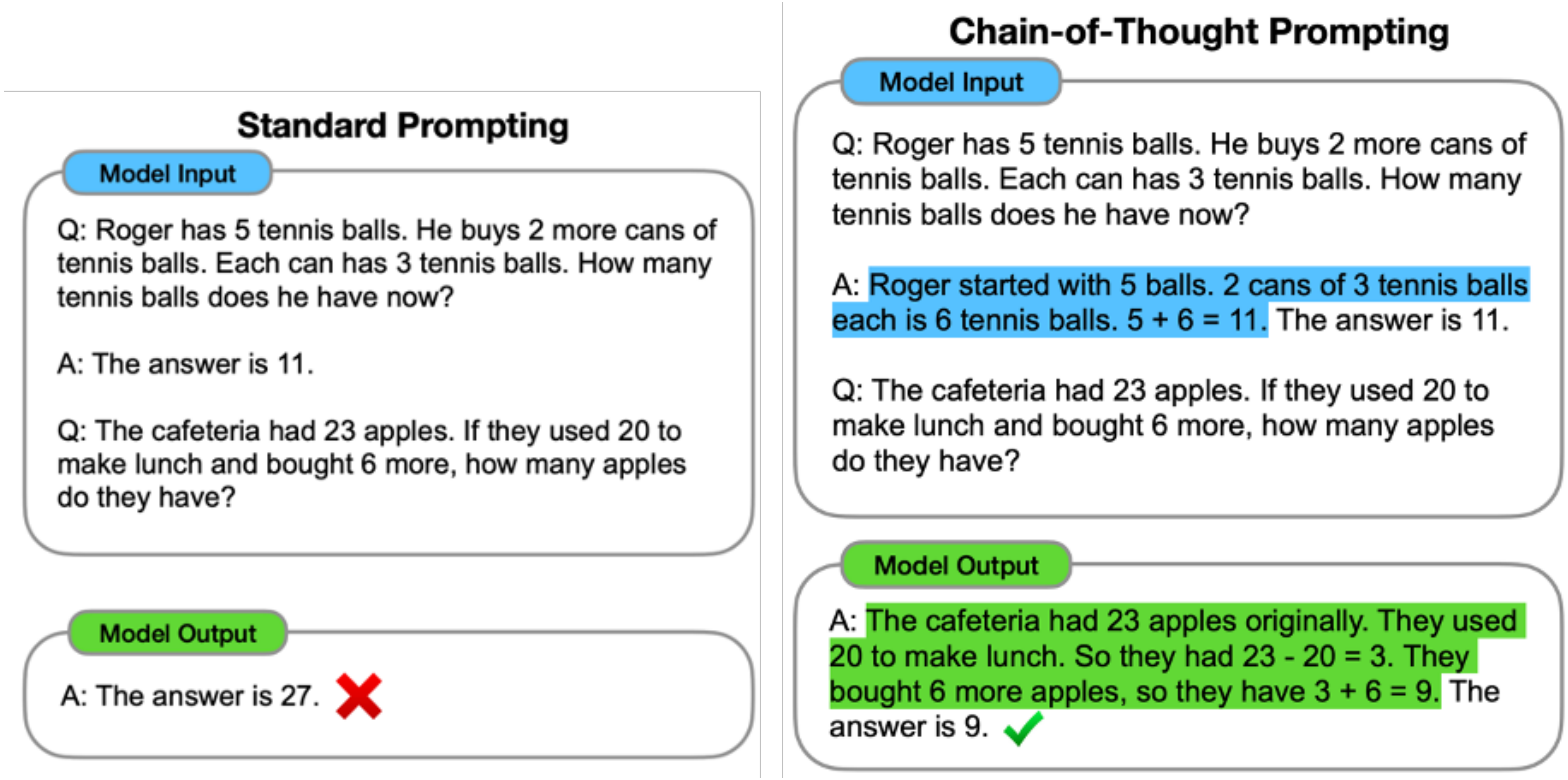
Limits of Prompting for Harder Tasks

Some tasks seem too hard for even large LMs to learn through prompting alone.
Especially tasks involving **richer, multi-step reasoning.**
(Humans struggle at these tasks too!)

```
19583 + 29534 = 49117
98394 + 49384 = 147778
29382 + 12347 = 41729
93847 + 39299 = ?
```

Solution: change the prompt!

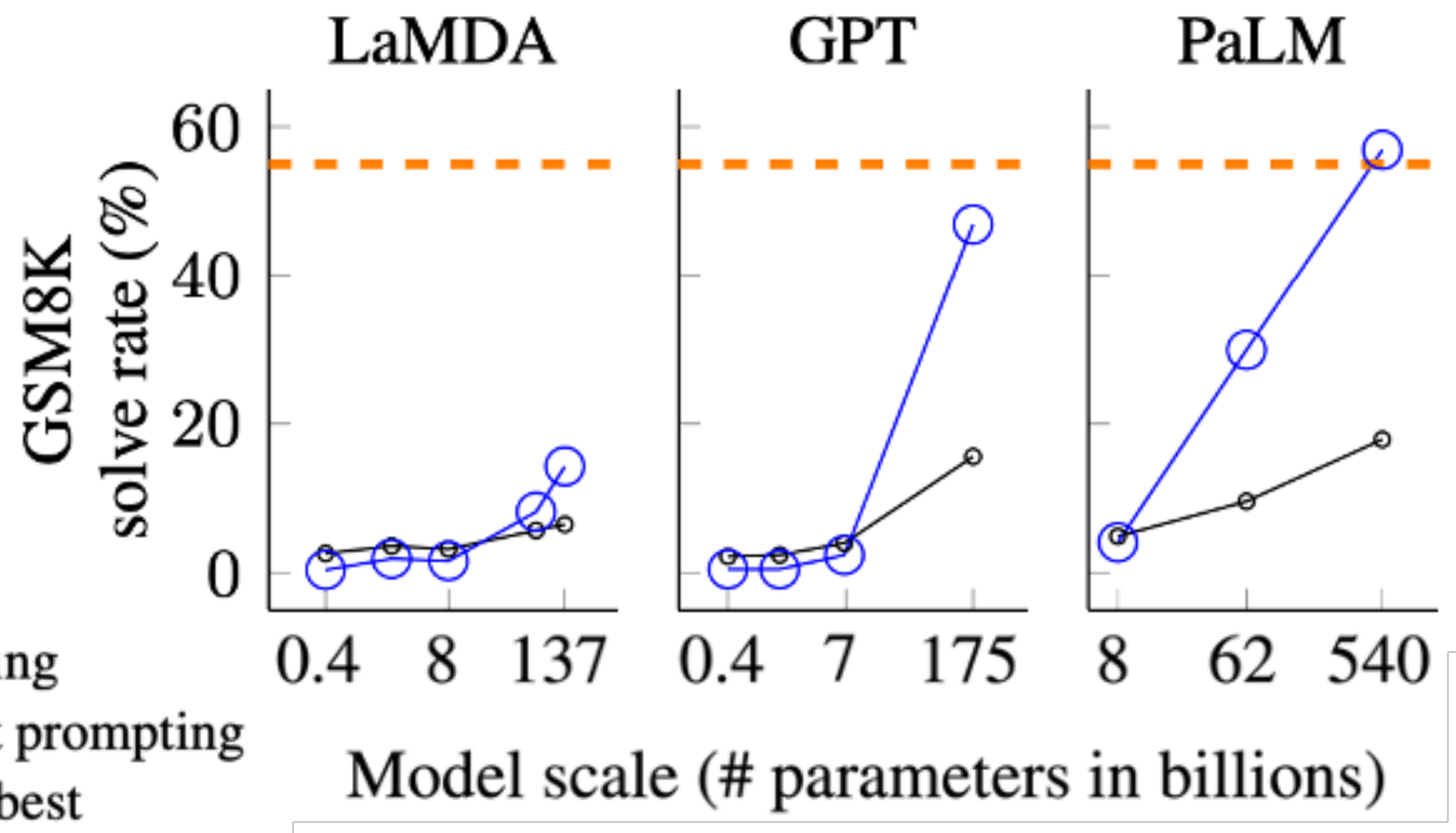
Chain-of-thought Prompting



[Wei et al., 2022; also see Nye et al., 2021]

Chain-of-thought prompting is an emergent property of model scale

Middle school math word problems



[Wei et al., 2022; also see Nye et al., 2021]

Chain-of-thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Do we even need examples of reasoning? Can we just ask the model to reason through things?

[Wei et al., 2022; also see Nye et al., 2021]

Zero-shot Chain-of-thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.** There are 16 balls in total. Half of the balls are golf balls. That means there are 8 golf balls. Half of the golf balls are blue. That means there are 4 blue golf balls. ✓

[Kojima et al., 2022]

Zero-shot COT prompting

	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7

Greatly outperforms zero-shot

Manual CoT still better

[Kojima et al., 2022]

Zero-shot Chain-of-thought prompting

1	LM-Designed	Let’s work this out in a step by step way to be sure we have the right answer.	**82.0**
2	Human-Designed	Let’s think step by step. (*1)	78.7
3		First, (*2)	77.3
4		Let’s think about this logically.	74.5
5		Let’s solve this problem by splitting it into steps. (*3)	72.2
6		Let’s be realistic and think step by step.	70.8
7		Let’s think like a detective step by step.	70.3
8		Let’s think	57.5
9		Before we dive into the answer,	55.7
10		The answer is after the proof.	45.7
-		(Zero-shot)	17.7

[Zhou et al., 2022; Kojima et al., 2022]

Self-Consistency Further Improves Reasoning!

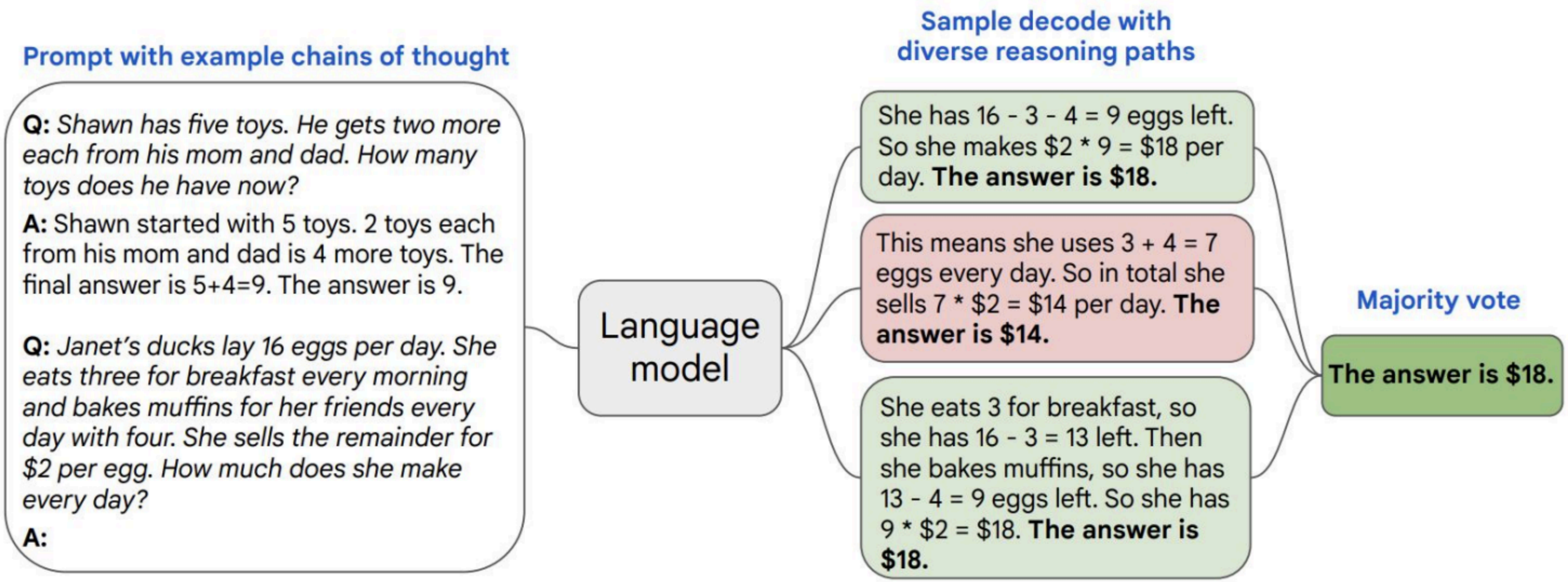


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using example chains of thought; (2) sample from the language model's decoder to generate a diverse set of reasoning paths; and (3) choose the most consistent answer using the majority/plurality vote.

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. "Self-consistency improves chain of thought reasoning in language models." arXiv preprint arXiv:2203.11171 (2022).

Hallucination

What is the world record for crossing the English Channel entirely on foot?

[redacted]

Hallucination

No fact check, e.g., summarizing a non-existent news article.

No explicit reasoning mechanism, leading to stupid mistakes

Easy to be manipulated, when the prompt is contaminated.

Downside of Prompt-based Learning

- **Inefficiency:** The prompt needs to be processed every time the model makes a prediction.

- **Poor performance:** Prompting generally performs worse than fine-tuning [Brown et al., 2020].

- **Sensitivity** to the wording of the prompt [Webson & Pavlick, 2022], order of examples [Zhao et al., 2021; Lu et al., 2022], etc.

- **Lack of clarity** regarding what the model learns from the prompt. Even random labels work [Zhang et al., 2022; Min et al., 2022]

Lecture Overview

- BERT for Classification
- Prompting LLMs
- **Using Prompting in CSS**

Prompting for CSS

Can Large Language Models Transform Computational Social Science?

Caleb Ziems*
Stanford University

William Held
Georgia Institute of Technology

Omar Shaikh
Stanford University

Jiaao Chen
Georgia Institute of Technology

Zhehao Zhang
Dartmouth College

Diyi Yang**
Stanford University

Are LLMs feasible tools for CSS?

emotion recognition
humor recognition
politeness recognition
empathy classification
issue frames extraction
ideology detection
agent framing
relationship dynamics
event extraction
power relations identification
social role detection
dialect feature identification

Psychology
Political Science
Literature
History
Sociology
Linguistics

Are LLMs feasible tools for CSS?

- emotion recognition
- humor recognition
- politeness recognition
- empathy classification
- issue frames extraction
- ideology detection
- agent framing
- relationship dynamics
- event extraction
- power relations identification
- social role detection
- dialect feature identification

- **CARER** (Saravia et al. 2018)
- **r/Jokes + Pun of the Day** (Weller and Seppi 2019)
- **Stanford Politeness Corpus** (Danescu-Niculescu-Mizil et al., 2013)
- **EPITOME** (Sharma et al., 2020)
- **SemEval-2016 Stance Dataset** (Mohammad et al., 2016)
- **Ideological Books Corpus** (Gross et al., 2013)
- **Article Bias Corpus** (Baly et al. 2020)
- **WikiEvents** (Li et al., 2021)
- **Hippocorpus** (Sap et al., 2020)
- **Wikipedia Talk Pages** (Danescu-Niculescu-Mizil et al. 2012)
- **CMU Movie Corpus** (Bamman et al. 2013)
- **Indian English Minimal Pairs** (Demszky et al. 2019)

utterance-level ***conversation-level*** ***document-level***

Zero-Shot Prompted LLMs

- Autoregressive
  - GPT-3
    - text-ada
    - text-babbage
    - text-curie
    - text-davinci
      - 001
      - 002 (more / better data)
      - 003 (+ RLHF)
      - ChatGPT
    - GPT-4
- Encoder-Decoder
  - Flan-T5 (instruction-tuned)
    - small 80M
    - base 250M
    - large 780M
    - XL 3B
    - XXL 11B

Note: GPT model sizes are estimates

Fine-Tuned LLMs

- MLM (Classification)
  - RoBERTa-large
- Encoder-Decoder (Generation)
  - T5

Prompt Engineering

Best Practice: multiple choice

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

Prompt Engineering: multiple choices

Best Practice: multiple choice (Hendrycks et al. 2021)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation

B: Trustworthy

Prompt Engineering: newlines

Best Practice: newlines (see Inverse Scaling Prize)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation ↩

B: Trustworthy ↩

Prompt Engineering: Give instructions

Best Practice: giver instructions after the context (Child et al. 2019)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation

B: Trustworthy

giving instructions or questions after the context

Prompt Engineering: Clarify the expected output

Best Practice: multiple choice (Hendrycks et al. 2021)

White House Ousts Top Climate Change Official

Which of the following describes the above news headline?

A: Misinformation

B: Trustworthy

Constraint: Answer with only the option above that is most accurate and nothing else.

Prompt Engineering: Request Structured Output

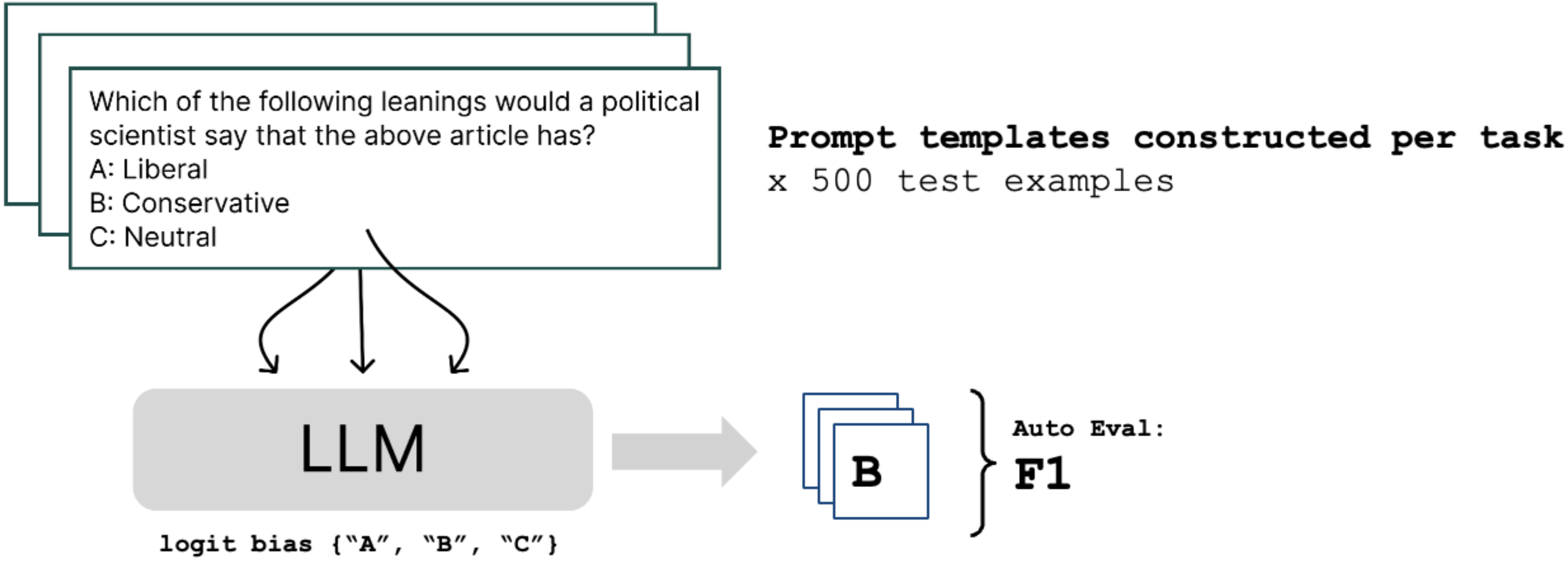
Best Practice: request structured responses in JSON format (see MiniChain)

{‘Victim’: ‘BLANK’, ‘Place’: ‘BLANK’, ‘Killer’: ‘BLANK’, ‘MedicalIssue’: ‘Blank’}

Replace the BLANKs with the extracted information about the event in <tgr>. Leave the keys of the JSON unchanged.

JSON Output:

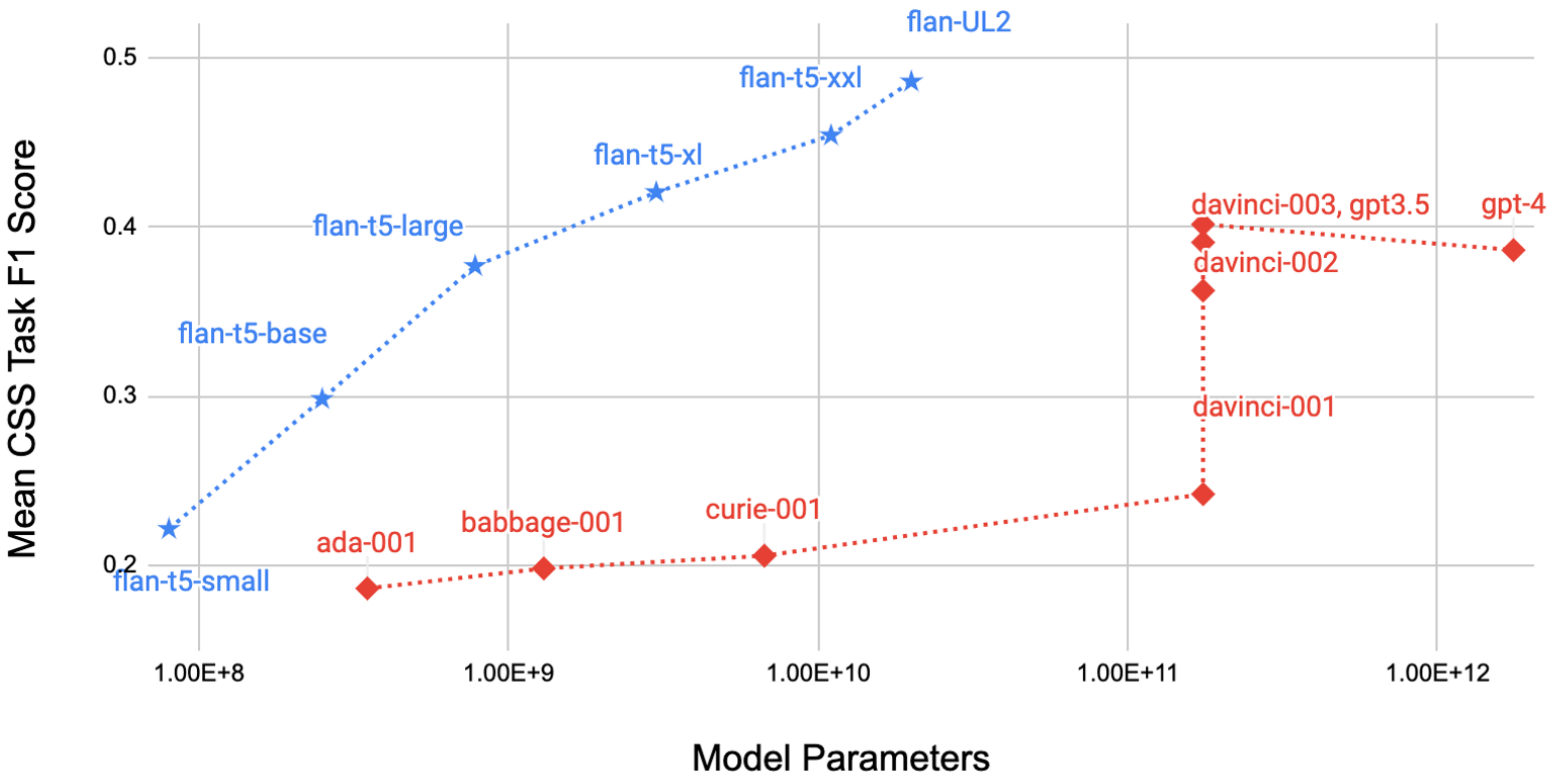
Classification Evaluation



Data \ Model	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Utterance Level Tasks																
Dialect	3.3	3.0	0.2	4.5	23.4	24.8	30.3	32.9	0.5	0.5	1.2	9.1	17.1	14.7	11.7	23.2
Emotion	16.7	71.6	19.8	63.8	69.7	65.7	66.2	70.8	6.4	4.9	6.6	19.7	36.8	44.0	47.1	50.6
Figurative	25.0	99.2	16.6	23.2	18.0	32.2	53.2	62.3	10.0	15.2	10.0	19.4	45.6	57.8	48.6	17.5
Humor	49.5	73.1	51.8	37.1	54.9	56.9	29.9	56.8	38.7	33.3	34.7	29.2	29.7	33.0	43.3	61.3
Ideology	33.3	64.8	18.6	23.7	43.0	47.6	53.1	46.4	39.7	25.1	25.2	23.1	46.0	46.8	43.1	60.0
Impl. Hate	16.7	62.5	7.4	14.4	7.2	32.3	29.6	32.0	7.1	7.8	4.9	9.2	18.4	19.2	16.3	3.7
Misinfo	50.0	81.6	33.3	53.2	64.8	68.7	69.6	77.4	45.8	36.2	41.5	42.3	70.2	73.7	55.0	26.9
Persuasion	14.3	52.0	3.6	10.4	37.5	32.1	45.7	43.5	3.6	5.3	4.7	11.3	21.6	17.5	23.3	56.4
Sem. Chng.	50.0	62.3	33.5	41.0	56.9	52.0	36.3	41.6	32.8	38.9	41.3	35.7	41.9	37.4	44.2	21.2
Stance	33.3	36.1	25.2	36.6	42.2	43.2	49.1	48.1	18.1	17.7	17.2	35.6	46.4	41.3	48.0	76.0
Conversation Level Tasks																
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4
Document Level Tasks																
Event Arg.	22.3	65.1	–	–	–	–	–	–	–	–	8.6	8.6	21.6	22.9	22.3	23.0
Event Det.	0.4	75.8	9.8	7.0	1.0	10.9	41.8	50.6	29.8	47.3	47.4	44.4	48.8	52.4	51.3	14.8
Ideology	33.3	85.1	24.0	19.2	28.3	29.0	42.4	38.8	22.1	26.8	18.9	21.5	42.8	43.4	44.7	51.5
Tropes	36.9	-	1.7	8.4	13.7	14.6	19.0	28.6	7.7	12.8	16.7	15.2	16.3	26.6	36.9	44.9

Data \ Model	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Utterance Level Tasks																
Dialect	3.3	3.0	0.2	4.5	23.4	24.8	30.3	32.9	0.5	0.5	1.2	9.1	17.1	14.7	11.7	23.2
Emotion	16.7	71.6	19.8	63.8	69.7	65.7	66.2	70.8	6.4	4.9	6.6	19.7	36.8	44.0	47.1	50.6
Figurative	25.0	99.2	16.6	23.2	18.0	32.2	53.2	62.3	10.0	15.2	10.0	19.4	45.6	57.8	48.6	17.5
Humor	49.5	73.1	51.8	37.1	54.9	56.9	29.9	56.8	38.7	33.3	34.7	29.2	29.7	33.0	43.3	61.3
Ideology	33.3	64.8	18.6	23.7	43.0	47.6	53.1	46.4	39.7	25.1	25.2	23.1	46.0	46.8	43.1	60.0
Impl. Hate	16.7	62.5	7.4	14.4	7.2	32.3	29.6	32.0	7.1	7.8	4.9	9.2	18.4	19.2	16.3	3.7
Misinfo	50.0	81.6	33.3	53.2	64.8	68.7	69.6	77.4	45.8	36.2	41.5	42.3	70.2	73.7	55.0	26.9
Persuasion	14.3	52.0	3.6	10.4	37.5	32.1	45.7	43.5	3.6	5.3	4.7	11.3	21.6	17.5	23.3	56.4
Sem. Chng.	50.0	62.3	33.5	41.0	56.9	52.0	36.3	41.6	32.8	38.9	41.3	35.7	41.9	37.4	44.2	21.2
Stance	33.3	36.1	25.2	36.6	42.2	43.2	49.1	48.1	18.1	17.7	17.2	35.6	46.4	41.3	48.0	76.0
Conversation Level Tasks																
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4
Document Level Tasks																
Event Arg.	22.3	65.1	–	–	–	–	–	–	–	–	8.6	8.6	21.6	22.9	22.3	23.0
Event Det.	0.4	75.8	9.8	7.0	1.0	10.9	41.8	50.6	29.8	47.3	47.4	44.4	48.8	52.4	51.3	14.8
Ideology	33.3	85.1	24.0	19.2	28.3	29.0	42.4	38.8	22.1	26.8	18.9	21.5	42.8	43.4	44.7	51.5
Tropes	36.9	-	1.7	8.4	13.7	14.6	19.0	28.6	7.7	12.8	16.7	15.2	16.3	26.6	36.9	44.9

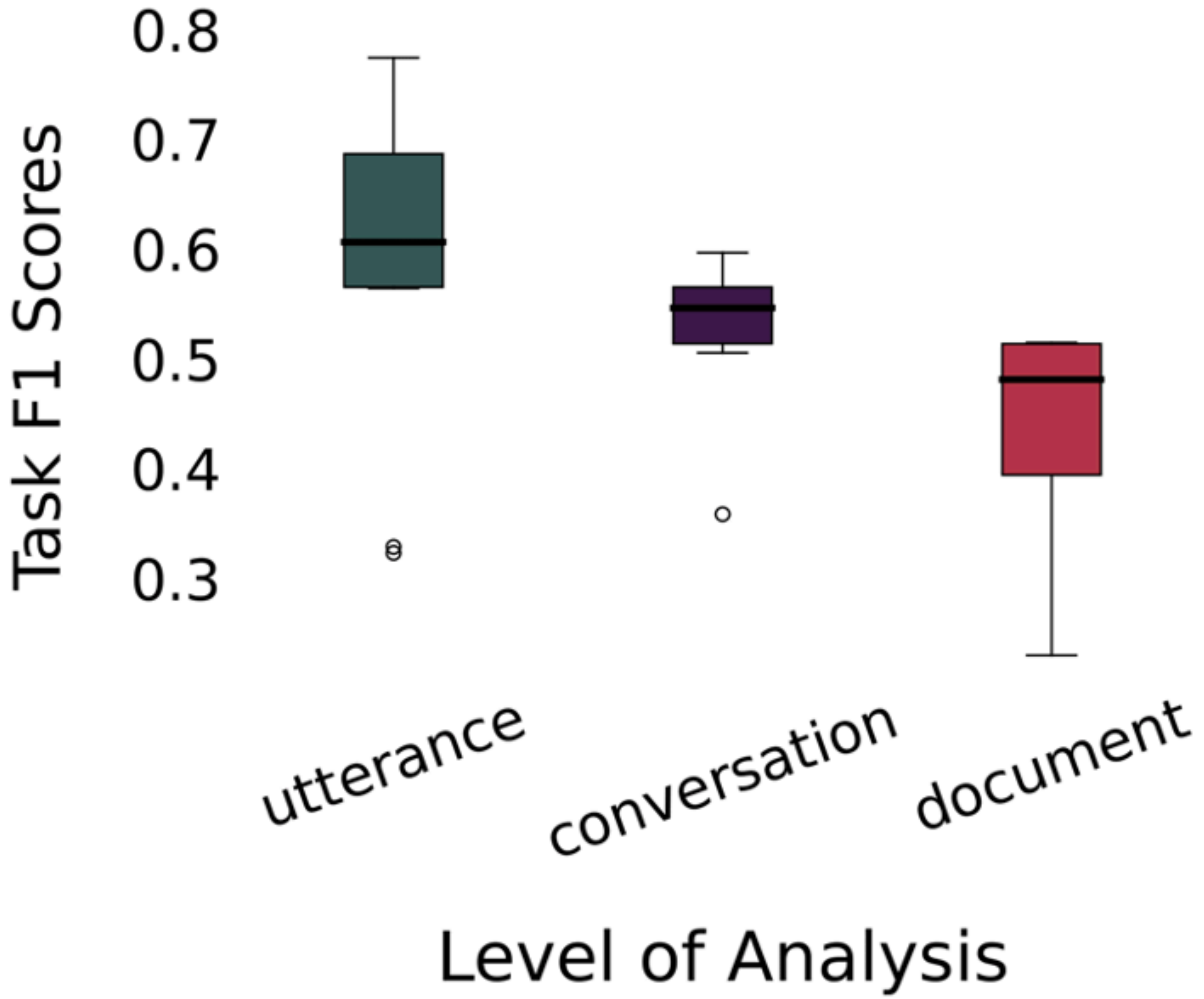
How does model size affect CSS tasks?



Are LLMs better adapted for some subfields?

Performance
is ***not tied to academic discipline***

but rather by
the ***complexity*** of
the ***input***



Lecture Overview

- BERT for Classification
- Prompting LLMs
- Using Prompting in CSS