Foundational Language Models

Mounika Marreddy,

University of Bonn, Germany

mounika0559@gmail.com

Agenda

- Recap on small language models [15 mins]
- Emerging abilities of language models and why they are effective? [30 mins]
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

Agenda

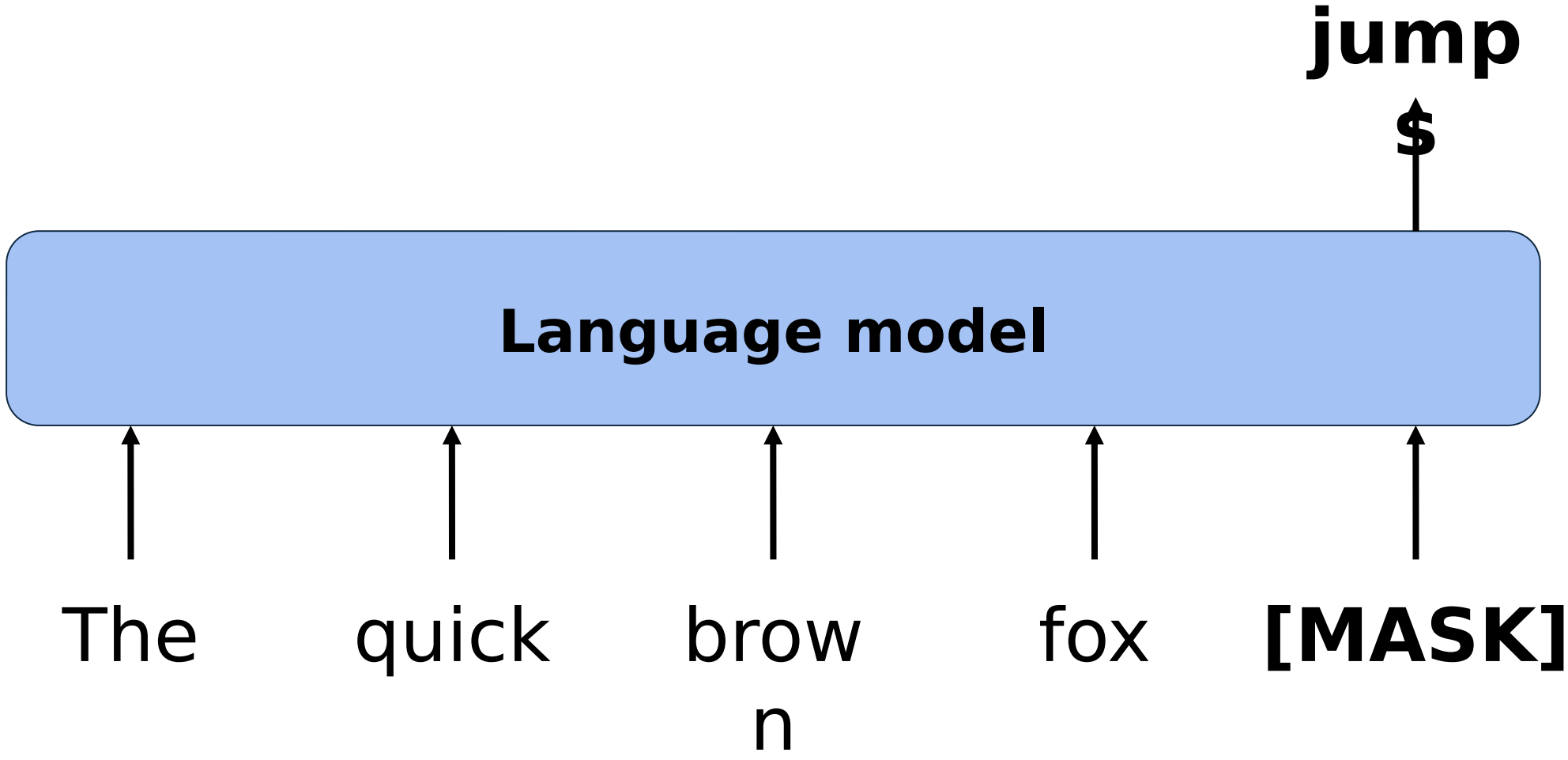
- Recap on small language models [15 mins]
- Emerging abilities of language models and why they are effective? [30 mins]
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

Foundation Models

Foundation models:
BERT, GPT-2 variations, T5, Flan-T5 and so on.

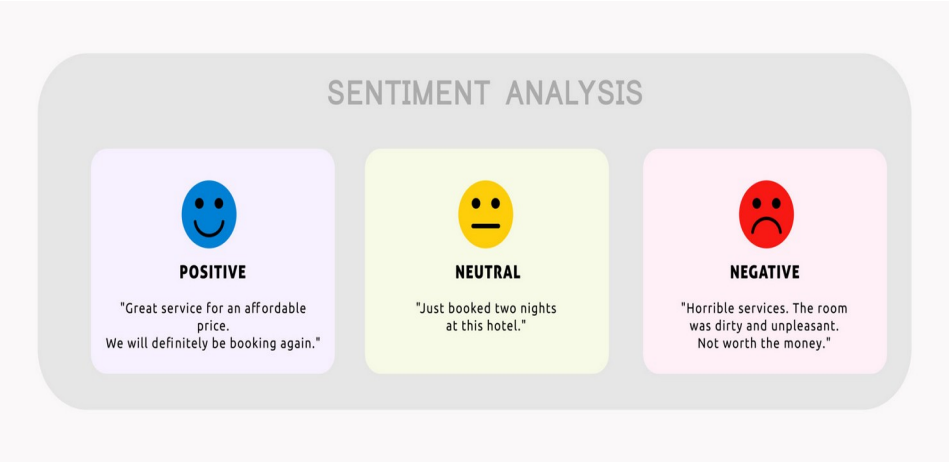
- What drives the success of these models?
  - Data
  - Hardware
  - Self-supervised learning
  - Transformer architectures

LMs are trained to predict missing words

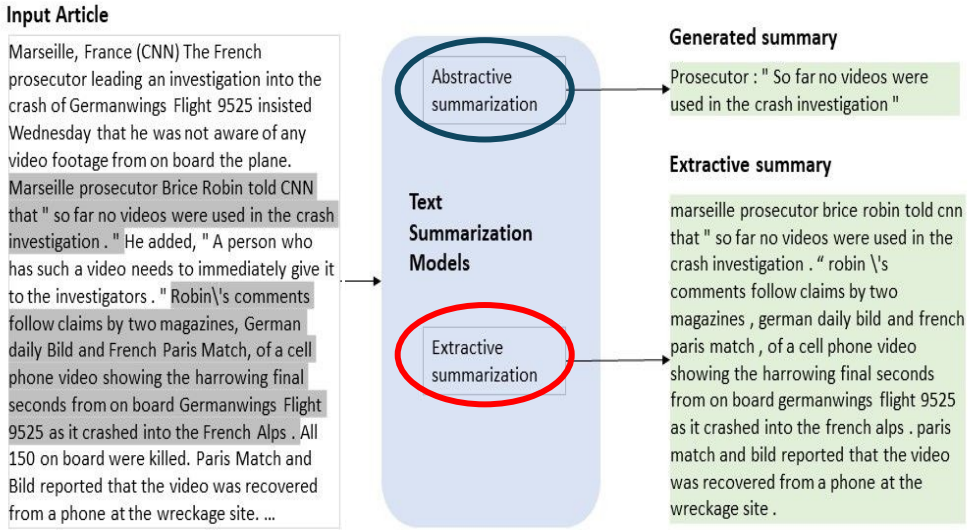


Language models are everywhere

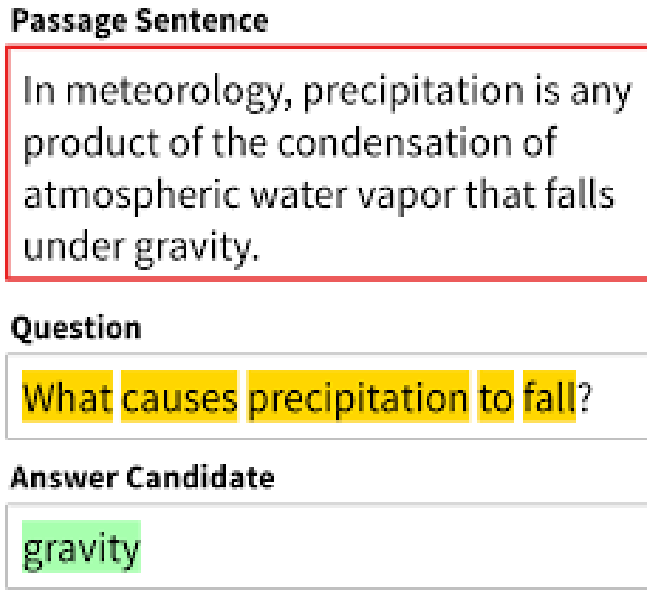
Sentiment



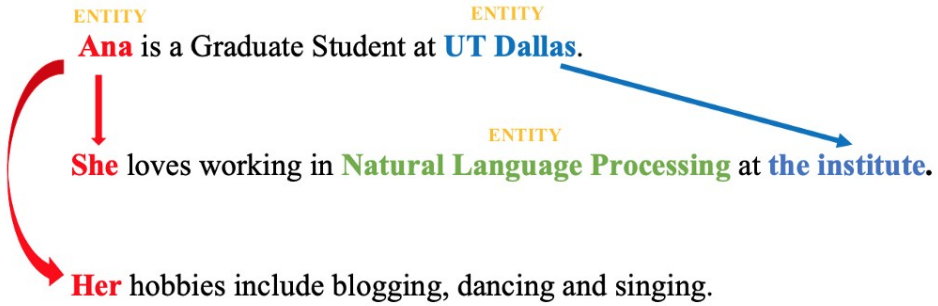
Summarization



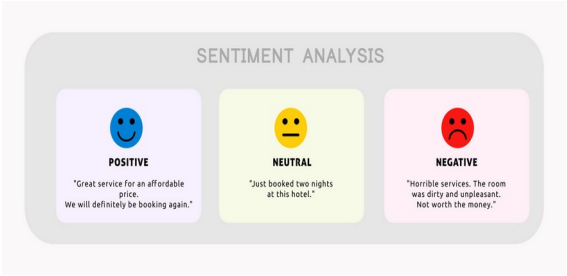
Question Answering



Coreference Resolution



Language models are everywhere



Small language models

GLUE

SuperGLUE

HELM

MMLU (Massive Multitask Language Understanding)

BIG-bench

https://gluebenchmark.com/

https://super.gluebenchmark.com/

https://crfm.stanford.edu/helm/lite/latest/

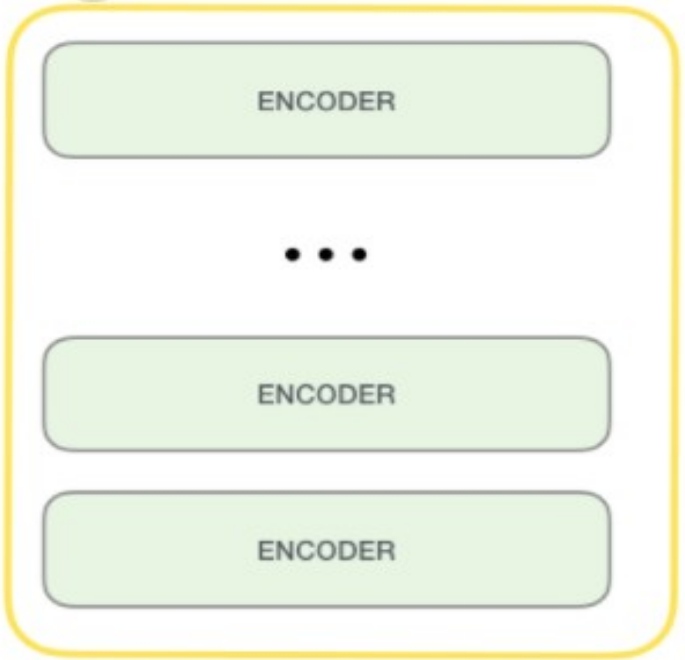
https://paperswithcode.com/dataset/mmlu

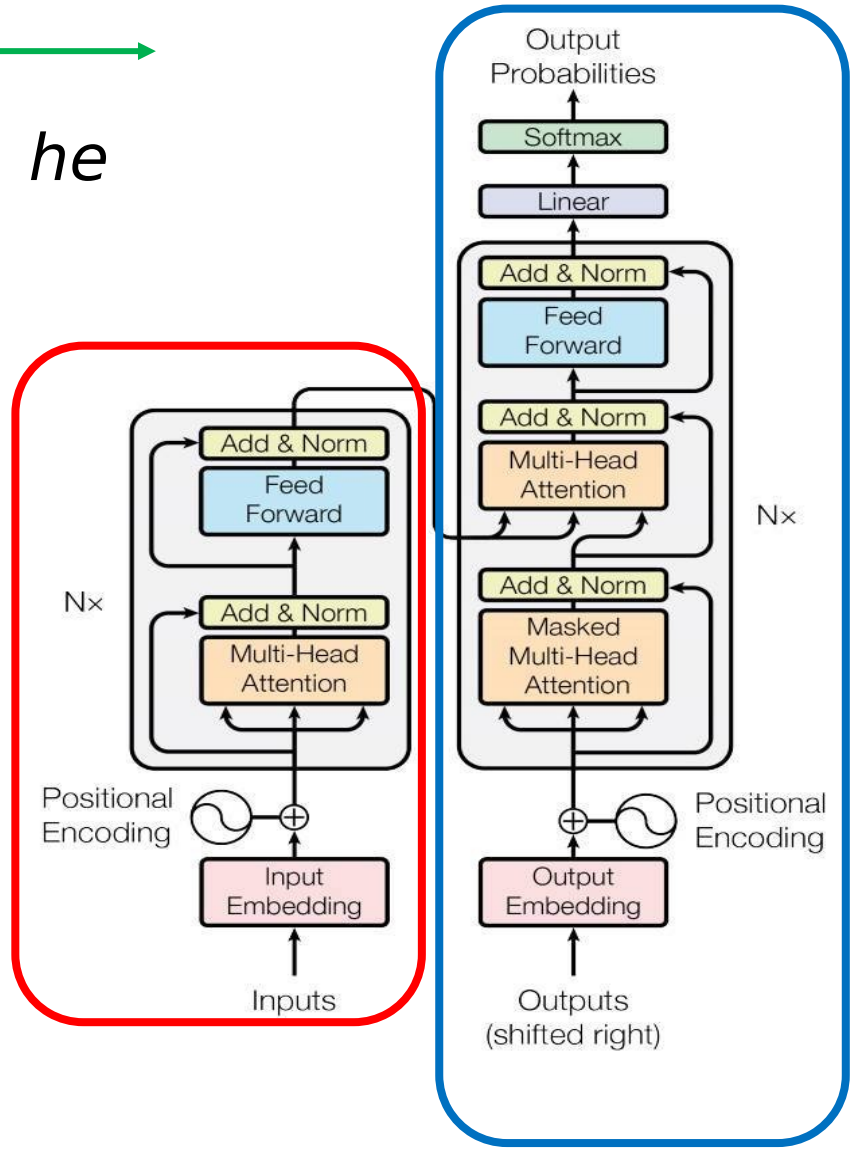
https://paperswithcode.com/dataset/big-bench

Transformer

Harry never thought he would

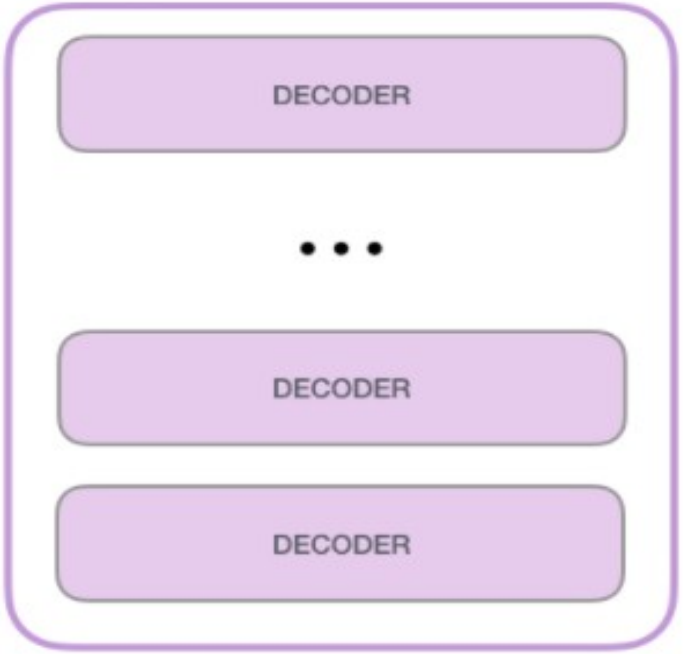
BERT





Harry never thought he ***???***

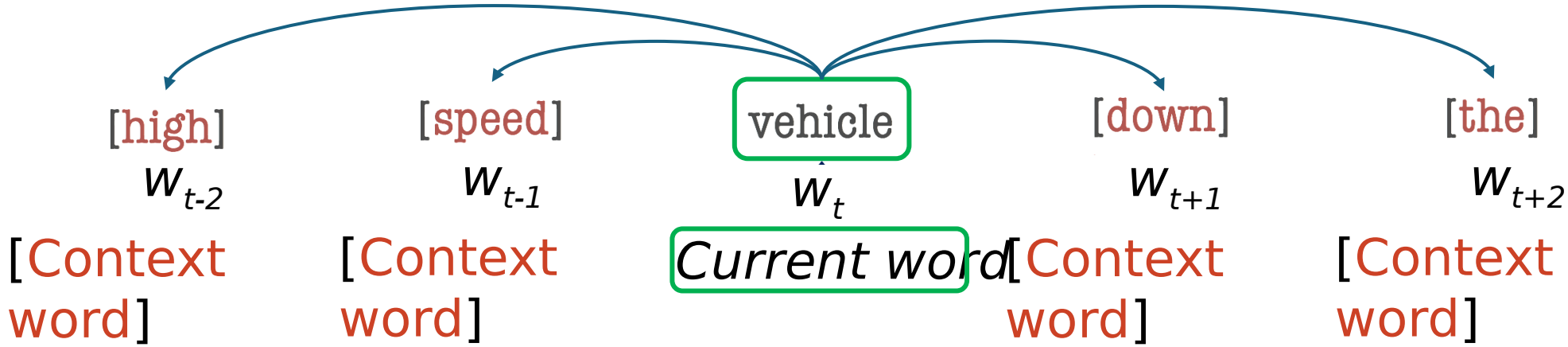
GPT-2



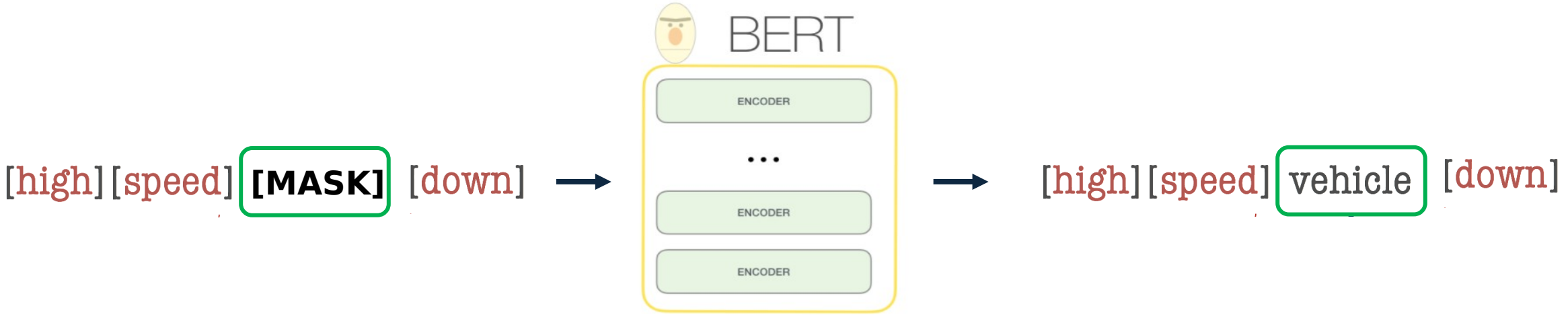
Ashish Vaswani, Noam Shazeer. "Attention Is All You Need." NeurIPS (2017).

BERT: Workflow

1. Self-attention with Bi-directional context

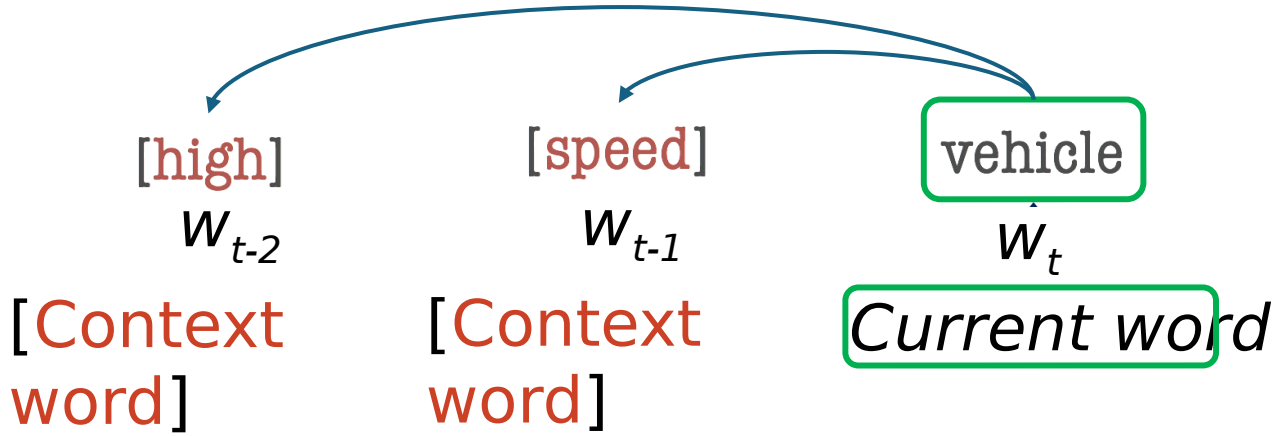


2. Masked language modeling (MLM)

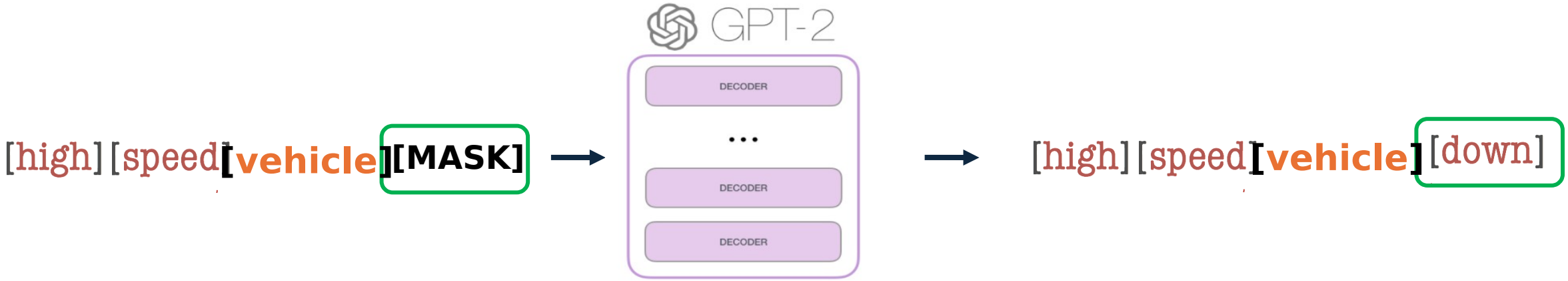


GPT2: Workflow

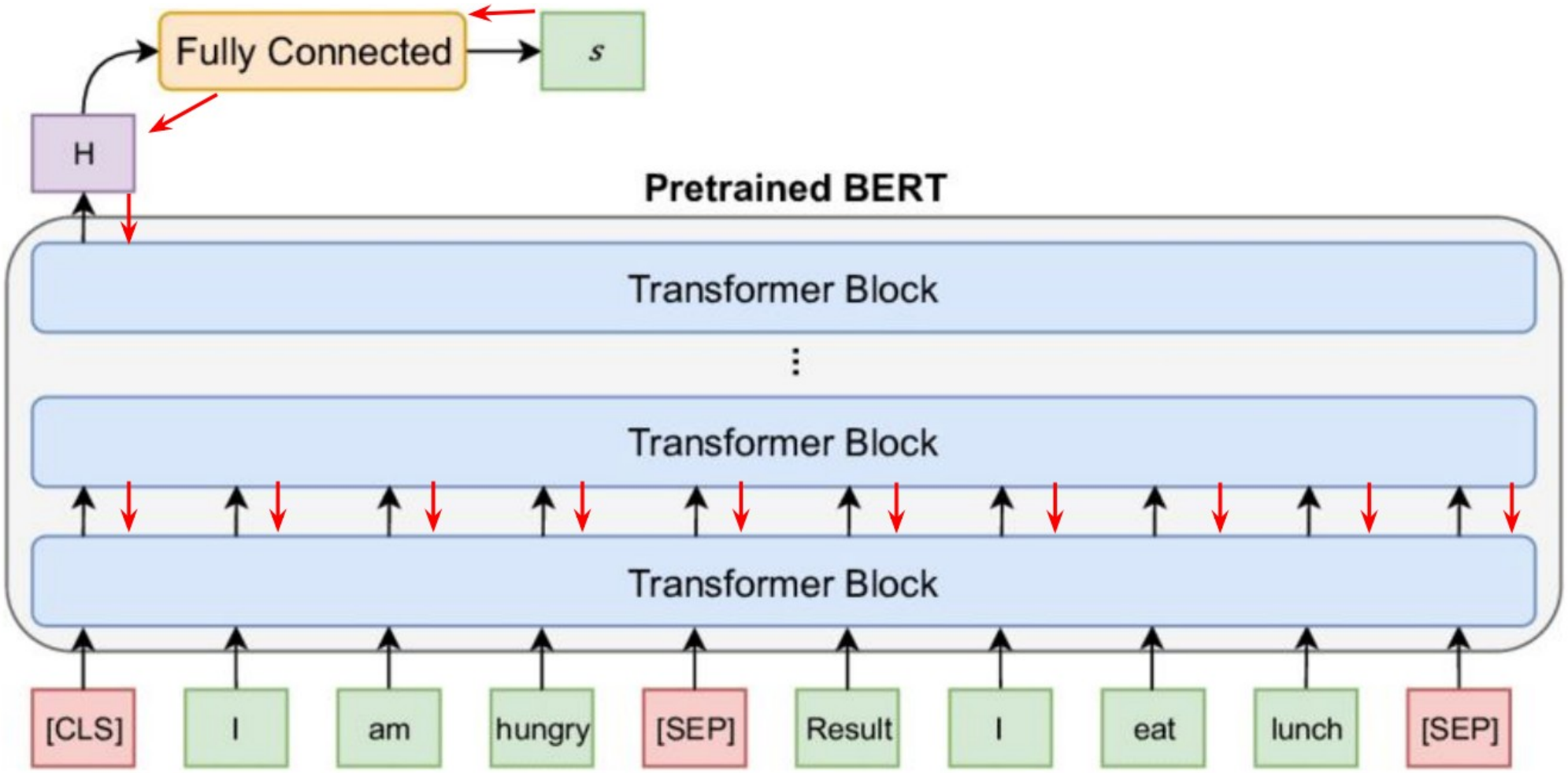
1. Self-attention with Uni-directional context



2. Causal language modeling (CLM)



Fine tuning: tune pretrain language model on a task



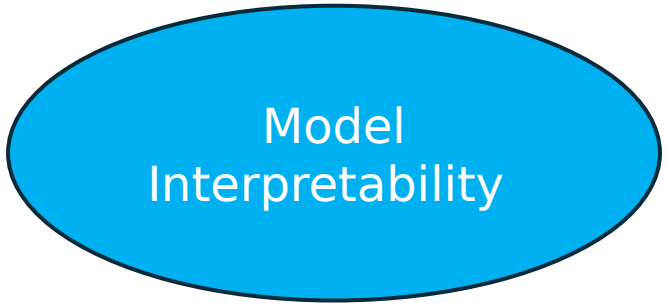
Agenda

- Recap on small language models [5 min]
- Emerging abilities of language models and why they are effective? [25 min]
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

Agenda

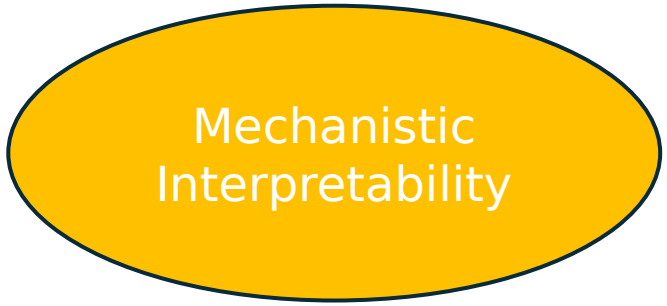
- Recap on small language models [5 min]
- Emerging abilities of language models and why they are effective? [25 min]
  - Analyzing and Interpreting language models
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

Analyzing and Interpreting LMs



Analysis of representations via probing

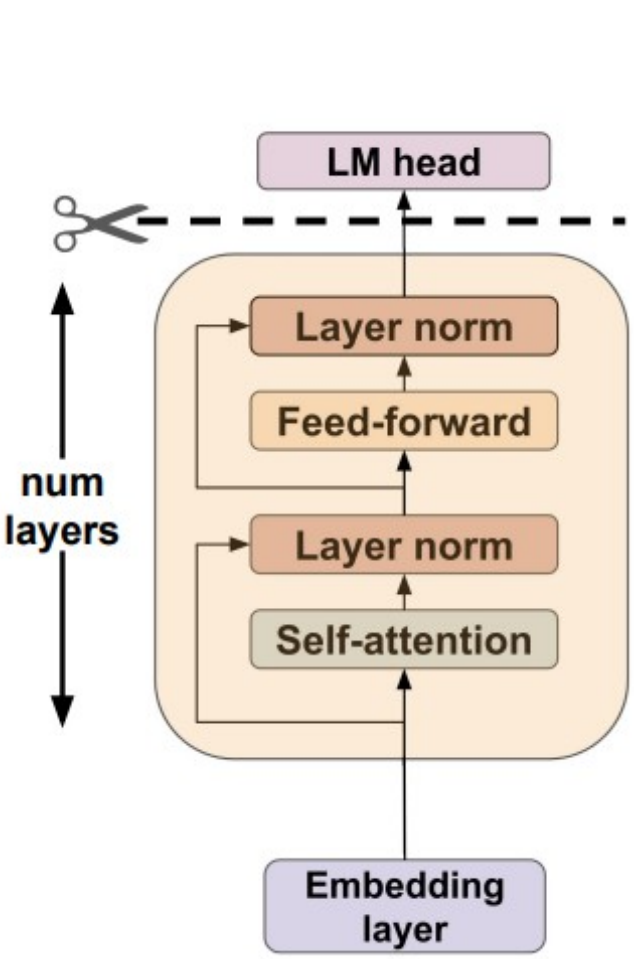
Bridging language models and brain

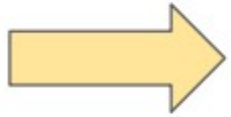


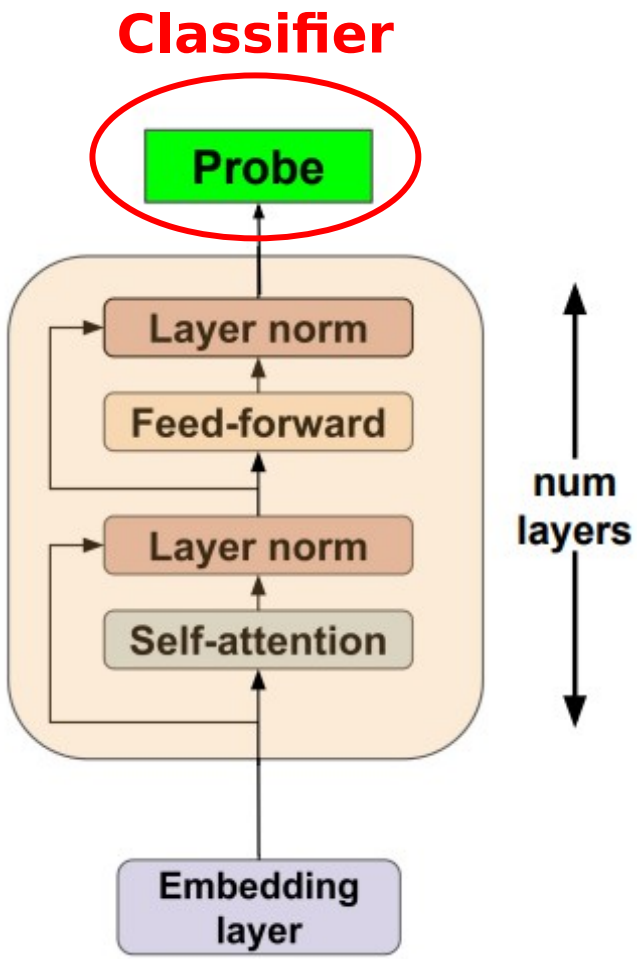
Robustness
Reverse engineering of neural computations

"BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP" 2023 EMNLP.

Model Interpretability?





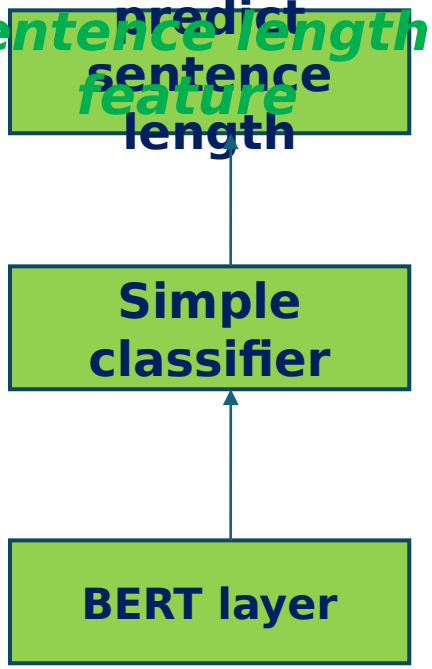


BERTology

Hierarchy of Linguistic Info - Setting

- Conneau et al., ACL'18 - Build diagnostic classifier to predict if a linguistic property is encoded in the given sentence representation.
- Features:
  - **Surface** – Sentence Length, Word Content
  - **Syntactic** – Bigram shift, Tree depth, Top constituent
  - **Semantic** – Tense, Subject Number, Object Number, Coordination Inversion and Semantic Odd Man Out.

If the prediction accuracy is good, then the model might be capturing the sentence length feature



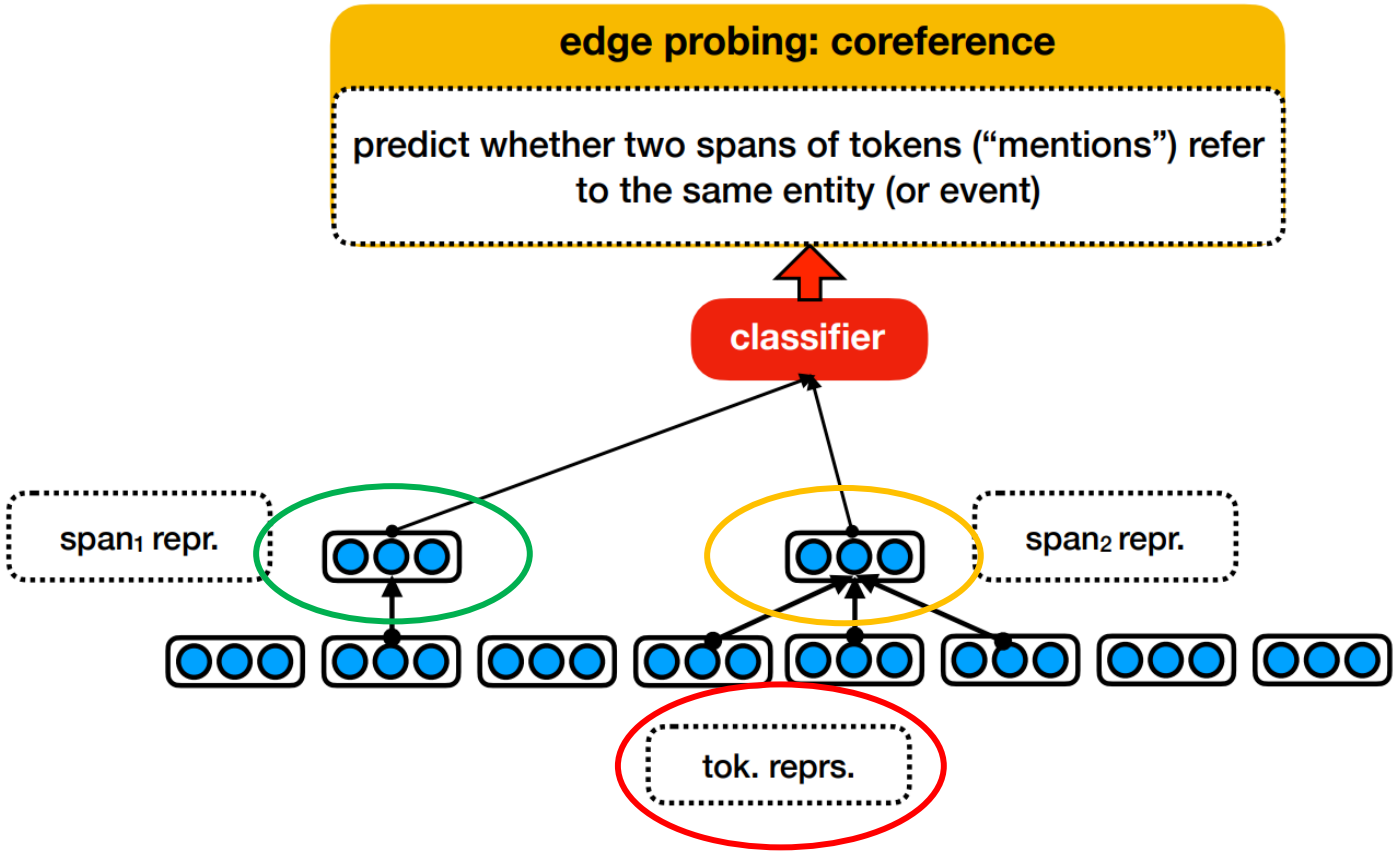
BERT composes a hierarchy of linguistic signals ranging from surface to semantic features

Surface **Syntactic** **Semantic**

Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

Jawahar et al. 2019 ACL

Edge Probing



Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations. ICLR 2019

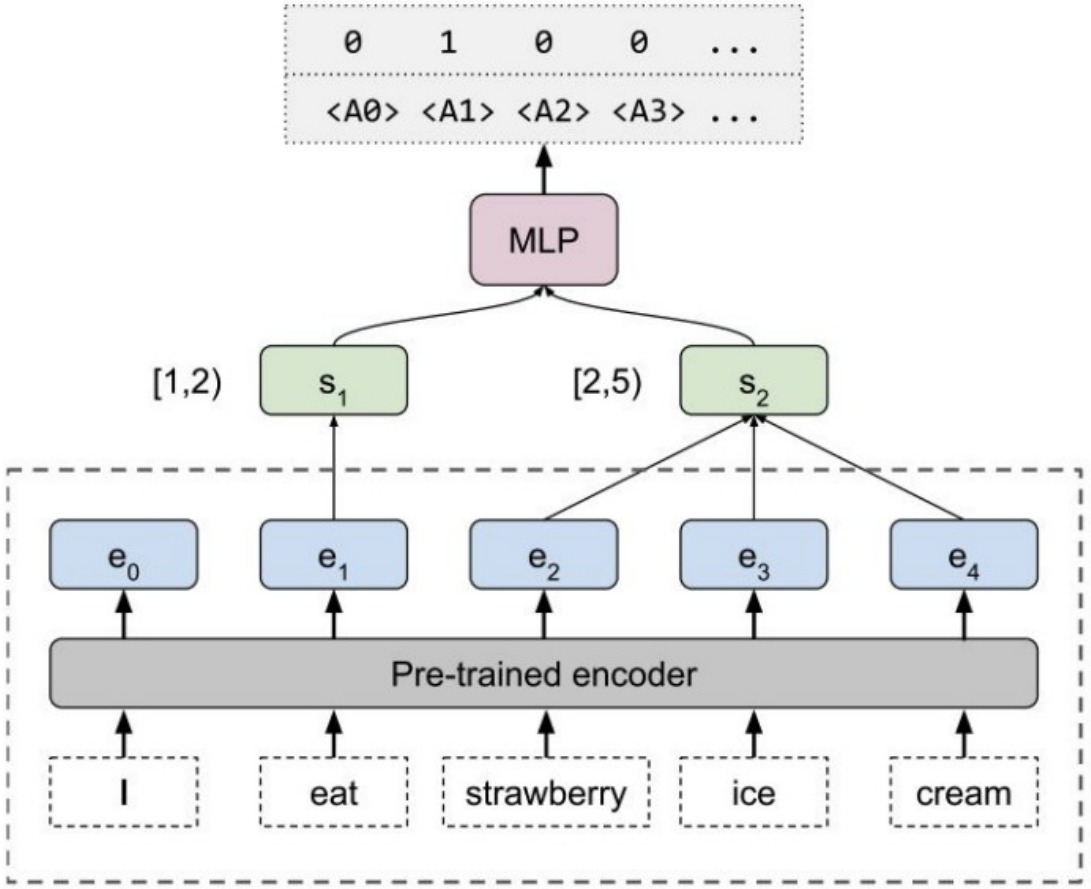
Edge Probing

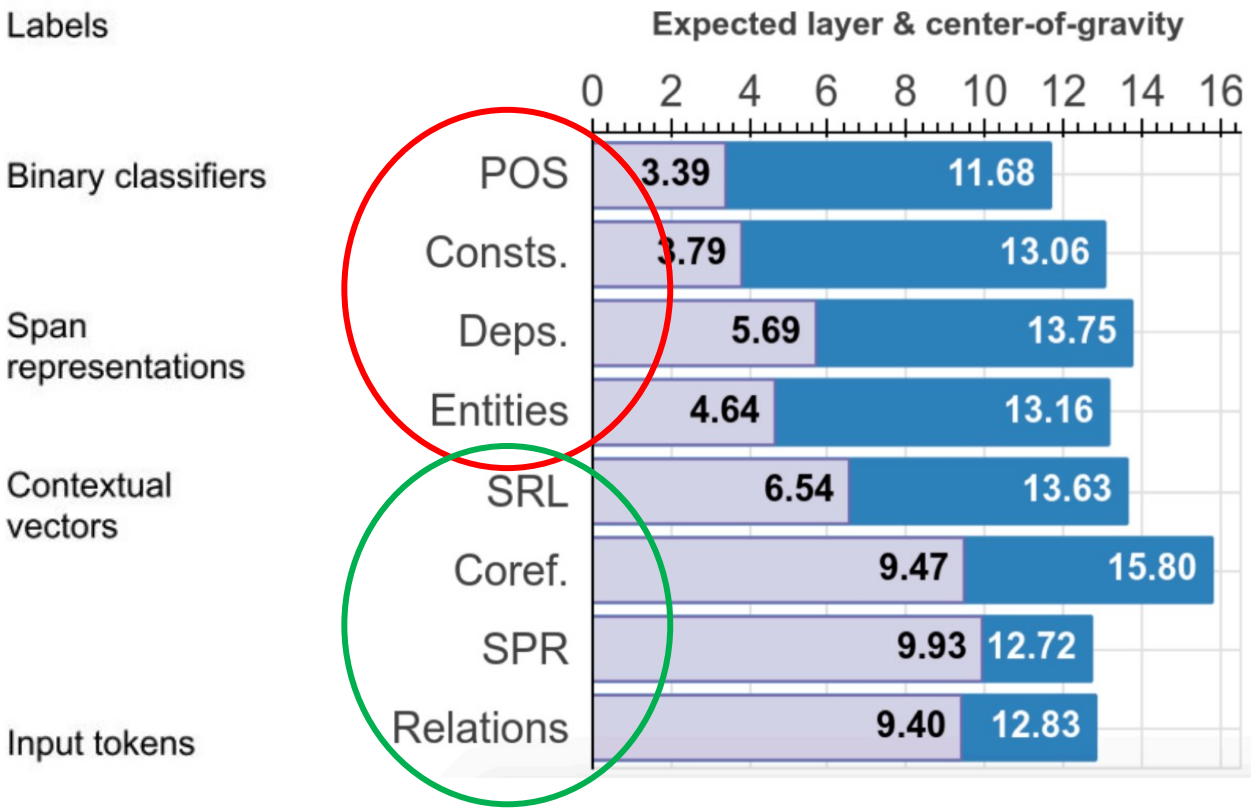
POS	The important thing about Disney is that it is a global $[\text{brand}]_1$. $\rightarrow$ NN (Noun)
Constit.	The important thing about Disney is that it $[\text{is a global brand}]_1$. $\rightarrow$ VP (Verb Phrase)
Depend.	$[\text{Atmosphere}]_1$ is always $[\text{fun}]_2$ $\rightarrow$ nsubj (nominal subject)
Entities	The important thing about $[\text{Disney}]_1$ is that it is a global brand. $\rightarrow$ Organization
SRL	$[\text{The important thing about Disney}]_2$ $[\text{is}]_1$ that it is a global brand. $\rightarrow$ Arg1 (Agent)
SPR	$[\text{It}]_1$ $[\text{endorsed}]_2$ the White House strategy... $\rightarrow$ {awareness, existed_after, ...}
Coref.$^{\text{O}}$	The important thing about $[\text{Disney}]_1$ is that $[\text{it}]_2$ is a global brand. $\rightarrow$ True
Coref.$^{\text{W}}$	$[\text{Characters}]_2$ entertain audiences because $[\text{they}]_1$ want people to be happy. $\rightarrow$ True Characters entertain $[\text{audiences}]_2$ because $[\text{they}]_1$ want people to be happy. $\rightarrow$ False
Rel.	The $[\text{burst}]_1$ has been caused by water hammer $[\text{pressure}]_2$. $\rightarrow$ Cause-Effect(e_2, e_1)

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations. ICLR 2019

Edge Probing

- Local syntax (word-level) captured at initial-middle layers
- High-level semantics captured at later layers

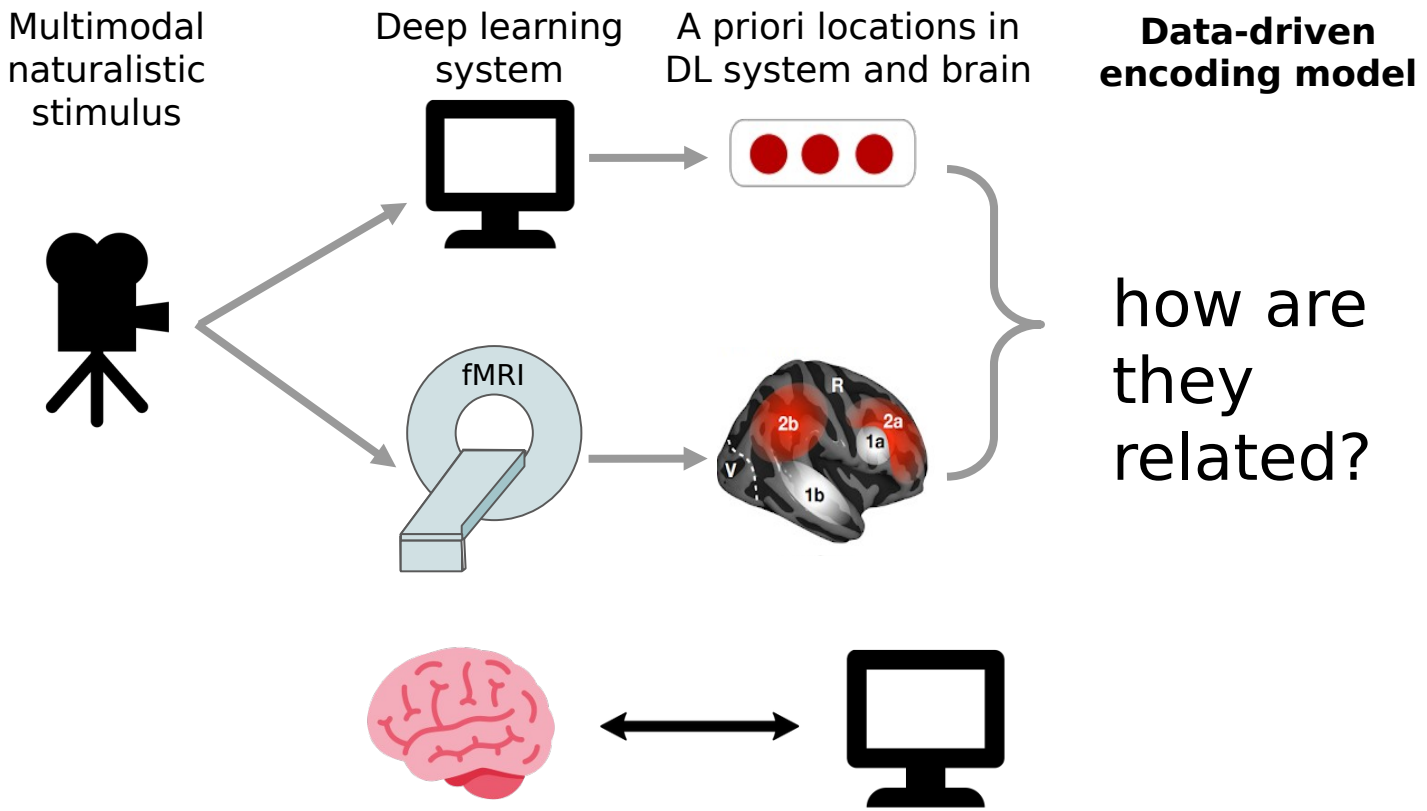




Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, Ellie Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations. ICLR 2019

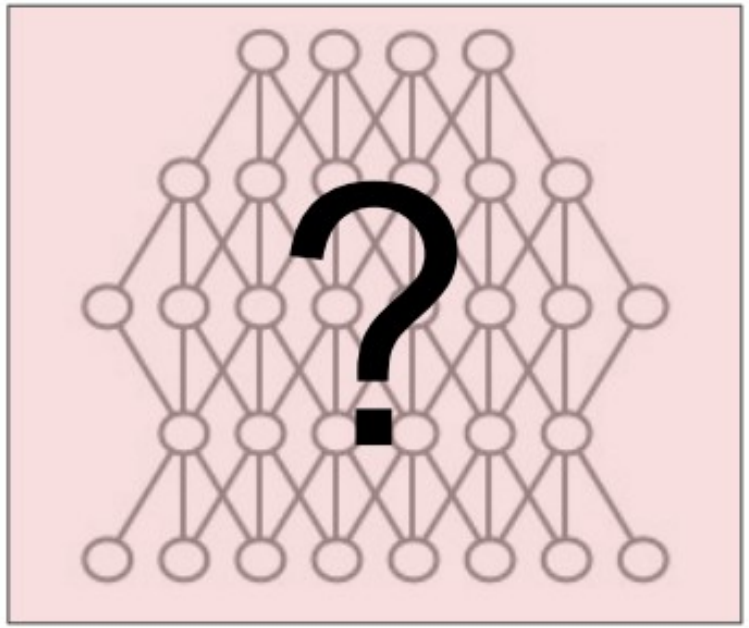
Behavioural Interpretability

- Data-driven encoding models evaluate the relationships between brains and deep learning models



Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). Advances in Neural Information Processing Systems, 32.

Mechanistic Interpretability



Interpreting Model Predictions

- Why did the model make this prediction?

What if? Drop layers..

What if? Change input examples..

What if? Change of weights..

Madeline C. Schiappa. Robustness Analysis of Video-Language Models. NeurIPS 2022.

Interesting questions about BERT, GPT2

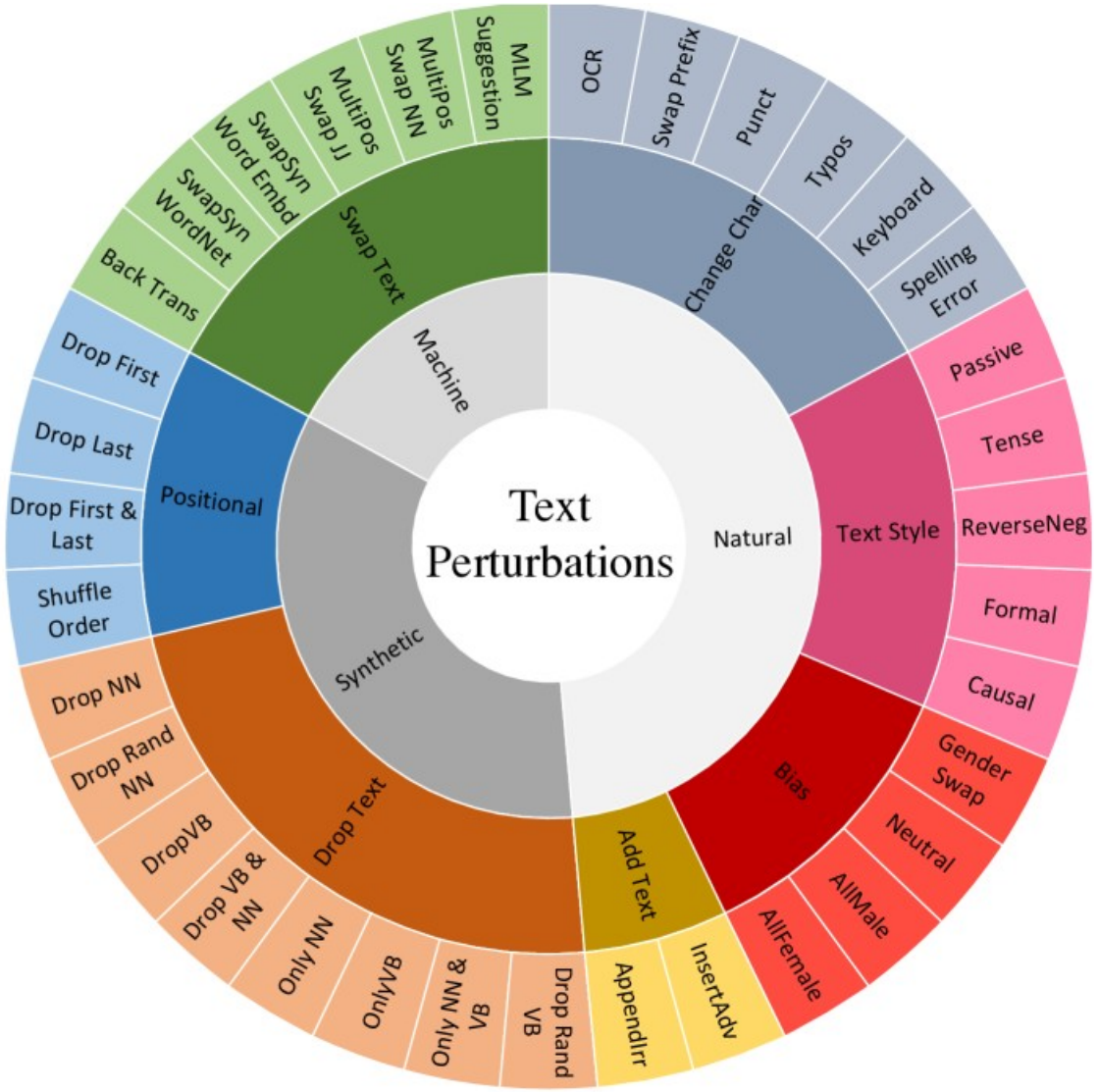
- Finetuning modifies the representations generated by each layer.
  - While fine-tuning these models, what changes across layers with respect to the pre-trained checkpoints?
- Robustness to input perturbations
  - How robust are BERT, GPT2 to input perturbations?
  - Is the effect of finetuning consistent across all models for various NLP tasks?
  - Do these models exhibit varying levels of robustness to input text perturbations when finetuned for different NLP tasks?

- Centred Kernel Alignment (CKA)
  - Compare layer-wise hidden state representations of pre-trained and finetuned models.
  - 1 ⇒ perfect sim; 0 ⇒ no sim.

- Similarity Through Inverted Representations (STIR) (finetuned|pre-trained)
  - Given pretrained model and data samples, STIR defines how invariant finetuned model is to perturbations of the samples that are imperceptible by , i.e., do not change their representations according to .

Pavan Kalyan Reddy Neerudu, Subba Reddy Oota, Mounika Marreddy, Venkateswara Rao Kagita, Manish Gupta. On Robustness of Finetuned Transformer-based NLP Models. EMNLP 2023.

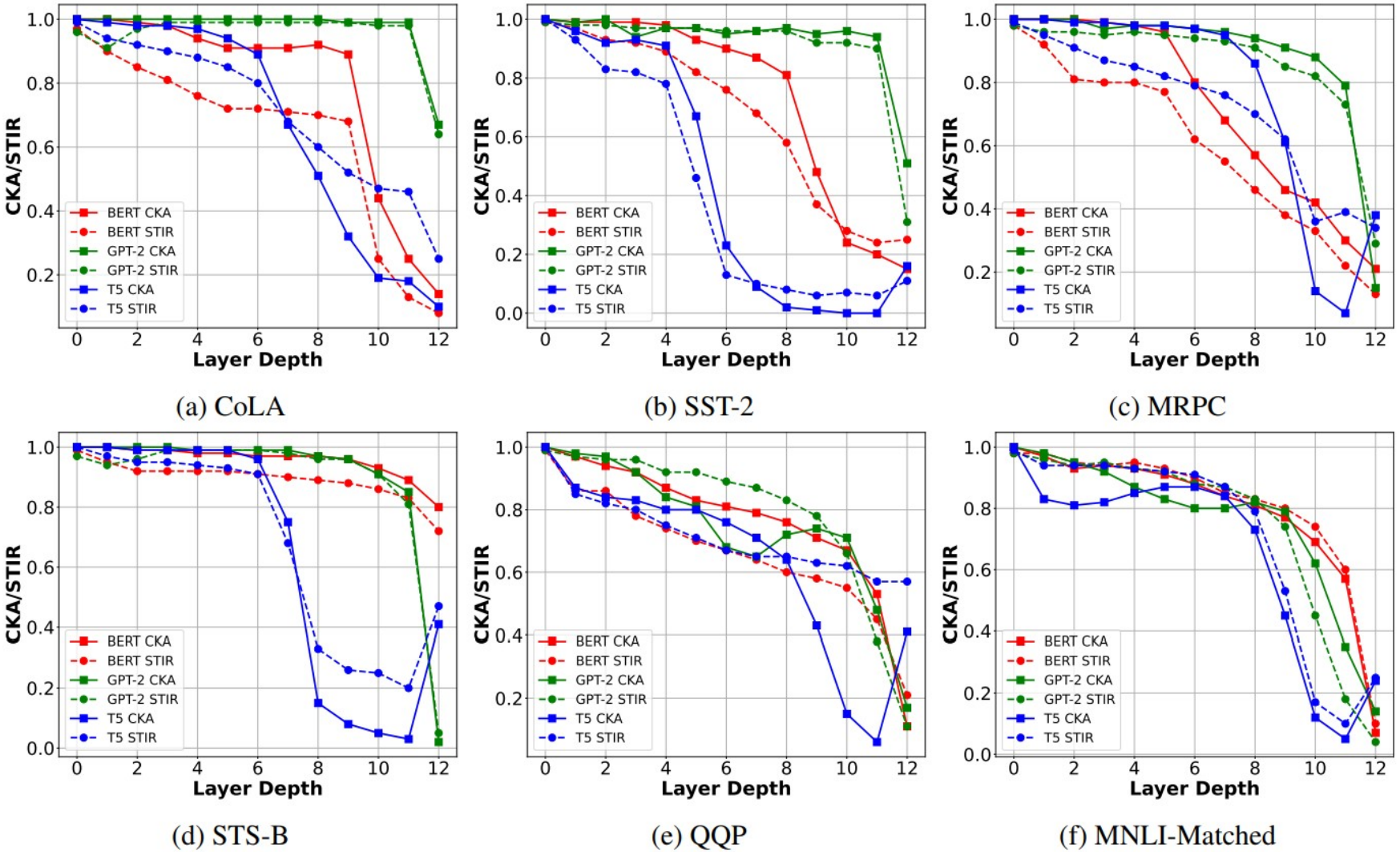
Robustness



- Text perturbations grouped into seven different categories
  - ChangeChar
  - AddText
  - Bias
  - Positional
  - DropText
  - SwapText
  - TextStyle

Madeline C. Schiappa. Robustness Analysis of Video-Language Models. NeurIPS 2022.

How does finetuning modify the layers representations for different models?



- GPT-2 had fewer affected layers, indicating higher semantic stability.
- CKA and STIR vary consistently across all three models.

Pavan Kalyan Reddy Neerudu, Subba Reddy Oota, Mounika Marreddy, Venkateswara Rao Kagita , Manish Gupta. On Robustness of Finetuned Transformer-based NLP Models. EMNLP 2023.

Is the impact of input text perturbations on finetuned models task-dependent?

	CoLA (Matthews CC)			SST-2 (Accuracy)			MRPC (Accuracy)			STS-B (PearsonCC)			QQP (Accuracy)		
Perturbation	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**
Drop nouns	0.18	0.10	**0.24**	0.92	**0.93**	**0.93**	0.94	**0.96**	0.94	0.56	0.48	**0.57**	0.89	**0.92**	0.89
Drop verbs	0.05	**0.24**	0.06	**0.95**	**0.95**	**0.95**	0.98	**0.99**	0.96	**0.93**	0.92	0.89	**0.97**	0.96	0.96
Drop first	0.48	**0.75**	0.54	**0.98**	0.97	**0.98**	**1.00**	0.99	**1.00**	**0.98**	0.93	0.94	**0.99**	0.98	**0.99**
Drop last	0.34	**0.45**	0.32	**1.00**	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**0.84**	0.83	0.83	0.95	**0.96**	0.95
Swap text	0.13	**0.16**	0.06	**0.98**	**0.98**	0.97	0.99	**1.01**	0.98	**0.98**	0.96	0.95	**0.97**	**0.97**	0.96
Add text	0.85	**0.92**	0.86	**0.99**	**0.99**	**0.99**	0.93	**1.00**	0.96	**0.99**	**0.99**	0.98	0.99	**1.00**	0.99
Change char	0.14	**0.29**	**0.29**	0.84	**0.86**	0.84	0.43	**0.97**	0.65	**0.58**	0.52	0.57	0.88	**0.95**	0.94
Bias	0.95	**0.96**	0.92	1.00	**1.01**	1.00	**1.00**	**1.00**	**1.00**	**0.99**	**0.99**	**0.99**	1.00	**1.01**	1.00

	MNLI-m (Accuracy)			QNLI (Accuracy)			RTE (Accuracy)			WNLI (Accuracy)		
Perturbation	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**	**BERT**	**GPT-2**	**T5**
Drop nouns	0.83	**0.85**	0.83	0.82	**0.87**	0.82	0.84	**1.01**	0.89	1.00	1.00	**1.05**
Drop verbs	0.89	**0.90**	**0.90**	**0.96**	0.94	0.94	0.98	**1.01**	0.96	1.00	1.01	**1.03**
Drop first	0.94	0.94	**0.95**	0.97	**0.98**	0.97	0.95	**1.00**	**1.00**	1.00	0.99	**1.01**
Drop last	0.89	**0.90**	0.89	0.97	**0.98**	0.97	0.97	**1.01**	0.98	**1.00**	0.99	**1.00**
Swap text	0.94	**0.95**	0.94	**0.97**	**0.97**	**0.97**	**0.98**	0.97	0.97	1.00	**1.01**	1.00
Add text	**0.95**	**0.95**	**0.95**	0.99	**1.00**	0.99	**1.00**	0.99	0.97	1.00	**1.03**	**1.03**
Change char	0.67	0.66	**0.68**	**0.77**	0.75	**0.77**	0.82	**0.99**	0.83	1.00	**1.03**	1.01
Bias	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	**1.02**	1.00	1.00	**1.02**	1.00

- Single-sentence tasks
  CoLA: GPT2 is most robust.
  Sentiment analysis: All models are very robust, except for “Change char”
- Similarity and paraphrase tasks
  For MRPC, GPT2 is best. For STS-B, BERT is best.

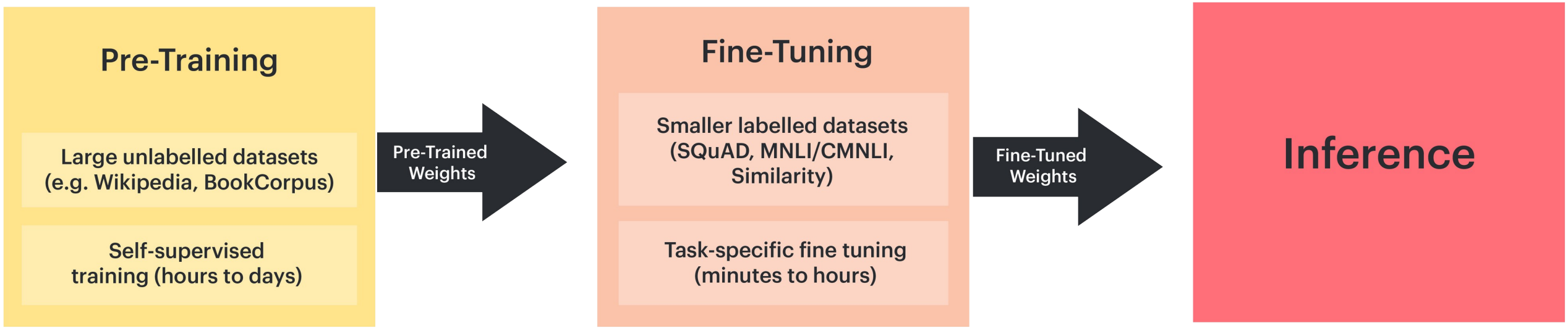
- NLI tasks: GPT2 is better for RTE.
- Transformer models demonstrated high tolerance towards “Dropping first word” and “Bias” perturbations.
- Impact of text perturbations on finetuned models is task-dependent.

y, Venkateswara Rao Kagita
, Manish Gupta. On Robustness of Finetuned Transformer-based NLP Models. EMNLP 2023.

Agenda

- Recap on small language models [5 min]
- Emerging abilities of language models and why they are effective? [25 min]
  - Analyzing and Interpreting language models
  - Limitations of small language models
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

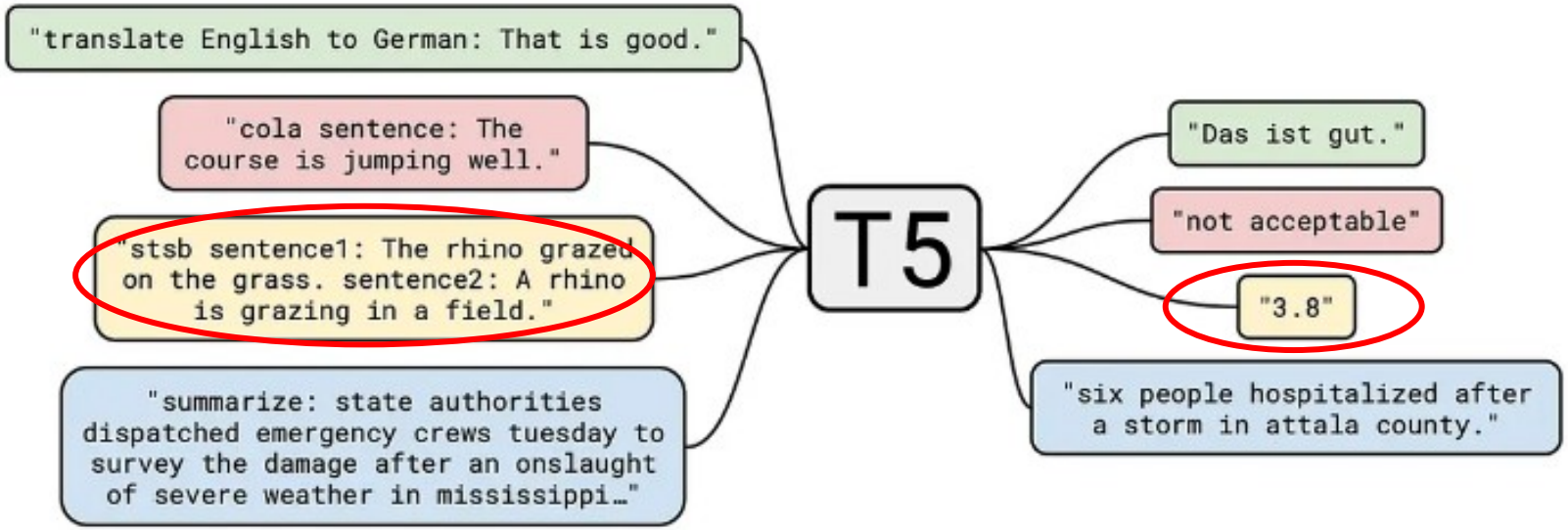
Downside of full fine tuning



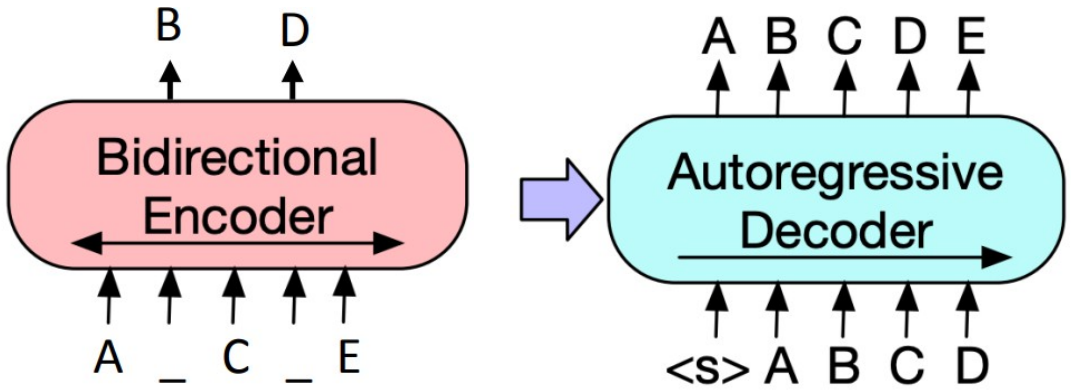
Agenda

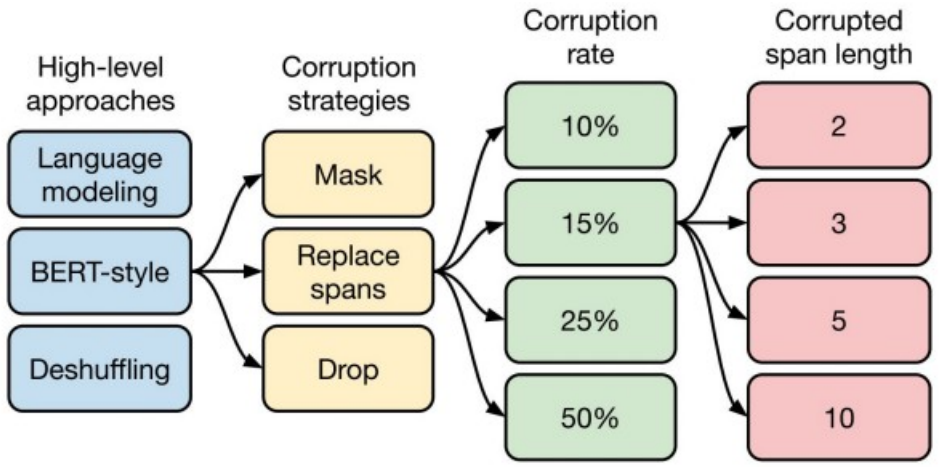
- Recap on small language models [5 min]
- Emerging abilities of language models and why they are effective? [25 min]
  - Analyzing and Interpreting language models
  - Limitations of small language models
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

T5 (Text-to-Text Transfer Transformer): Workflow



- Text-to-text transformer
- Encoder-decoder model
- Reformulates all tasks (during pretraining and finetuning) with a text-to-text format





Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2019.

T5: Workflow, Encoder

- Original text: **Thank you for inviting me to your party last week**

Token Masking

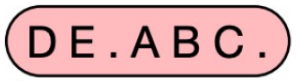
- Input text: **Thank you for inviting me to your party <Y> week**

Text Infilling

- Input text: **Thank you <X> me to your party <Y> week**
  - **<X> for inviting (span masking)**

Token Deletion

- Input text: **Thank you me to your party week**



Sentence Permutation

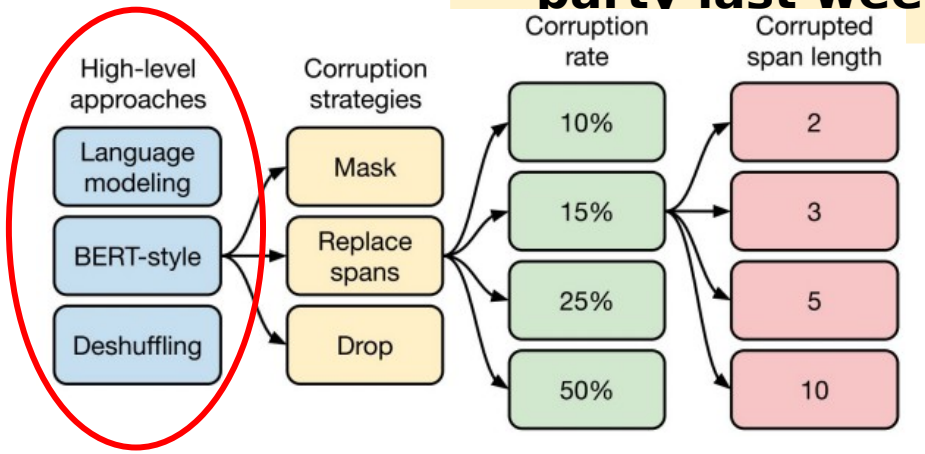
Document Rotation

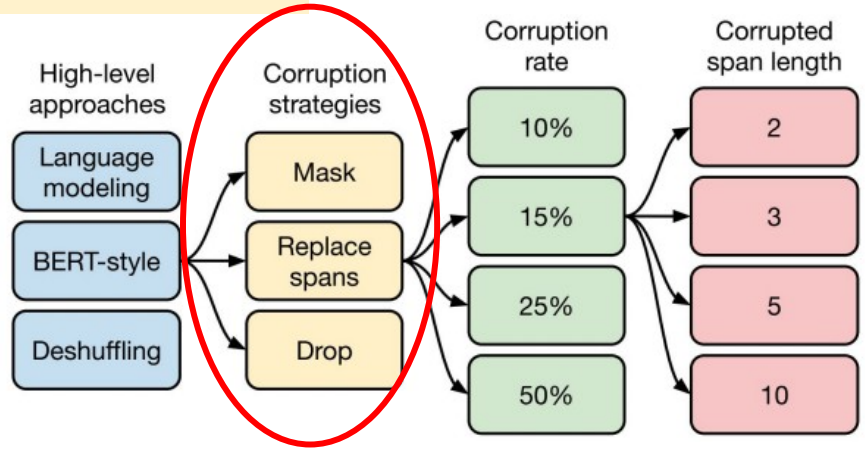
- Input text: **party me your to. last you inviting week Thanks**

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2019.

T5: Different unsupervised objectives

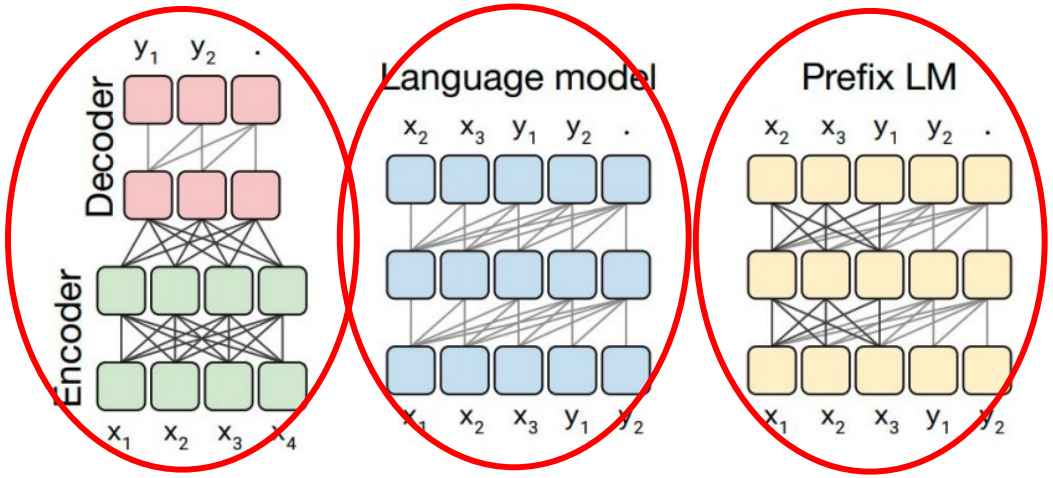
- Original text: **Thank you for inviting me to your party last week**





Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*

Inputs	Targets
Thank you <M> <M> me to your party <M> week .	*(original text)*
Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
Thank you me to your party week .	for inviting last



- Translate English to German: That is good. Target: Das is gut.

- Translate English to German: That is good. Target: Das is gut.
  - **“Good” representation can only look at “Translate English to German: That is”.**

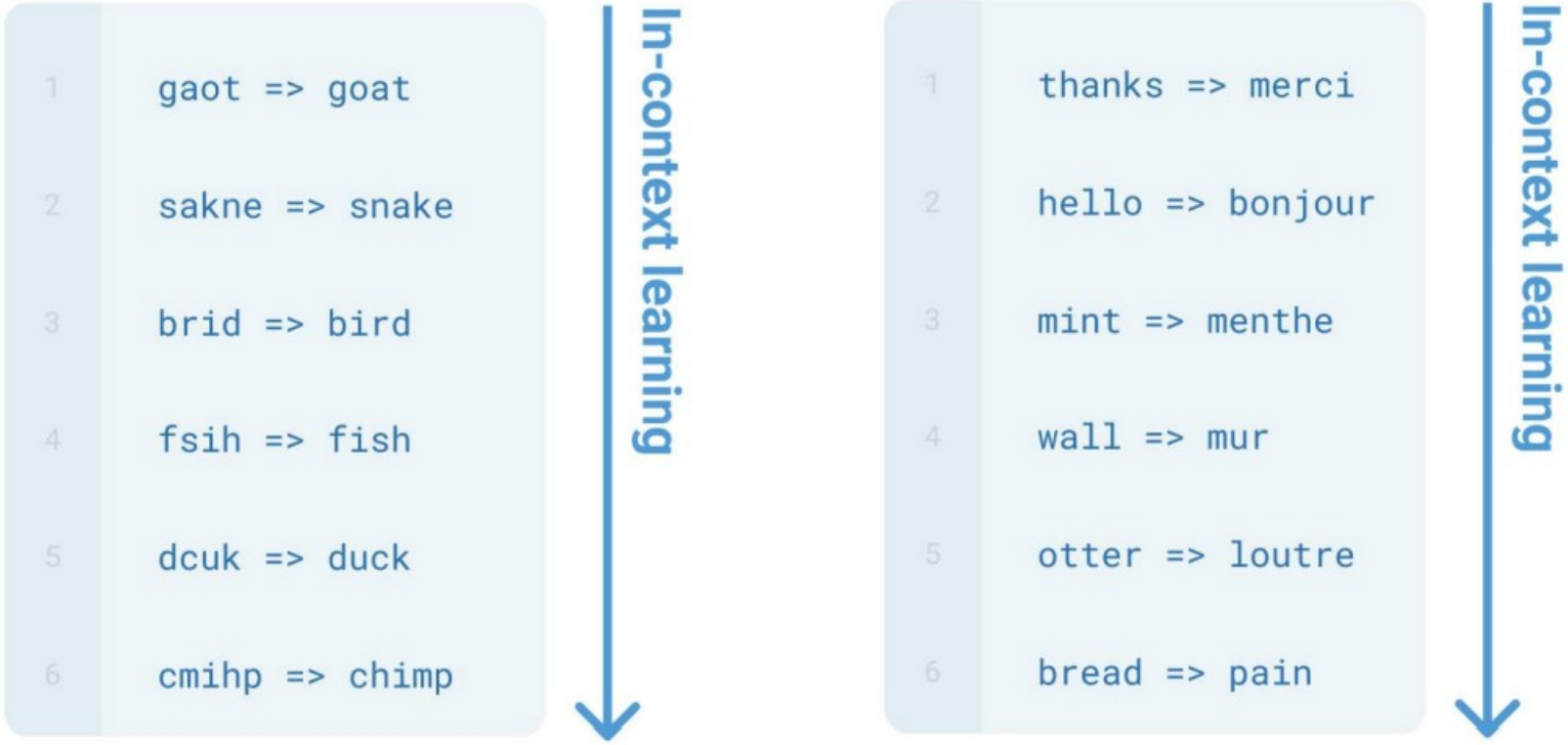
- Translate English to German: That is good. Target: Das is gut.
  - **“Good” representation can only look at “Translate English to German: That is. Target:”.**

Agenda

- Recap on small language models [5 min]
- Emerging abilities of language models and why they are effective? [25 min]
  - Analyzing and Interpreting language models
  - Limitations of small language models
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

Prompting

- 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



Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan .Language Models are Few-Shot Learners.Arxiv 2020.

Zero-shot vs. One-shot vs. Few-shot prompting

Zero-shot

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

One-shot

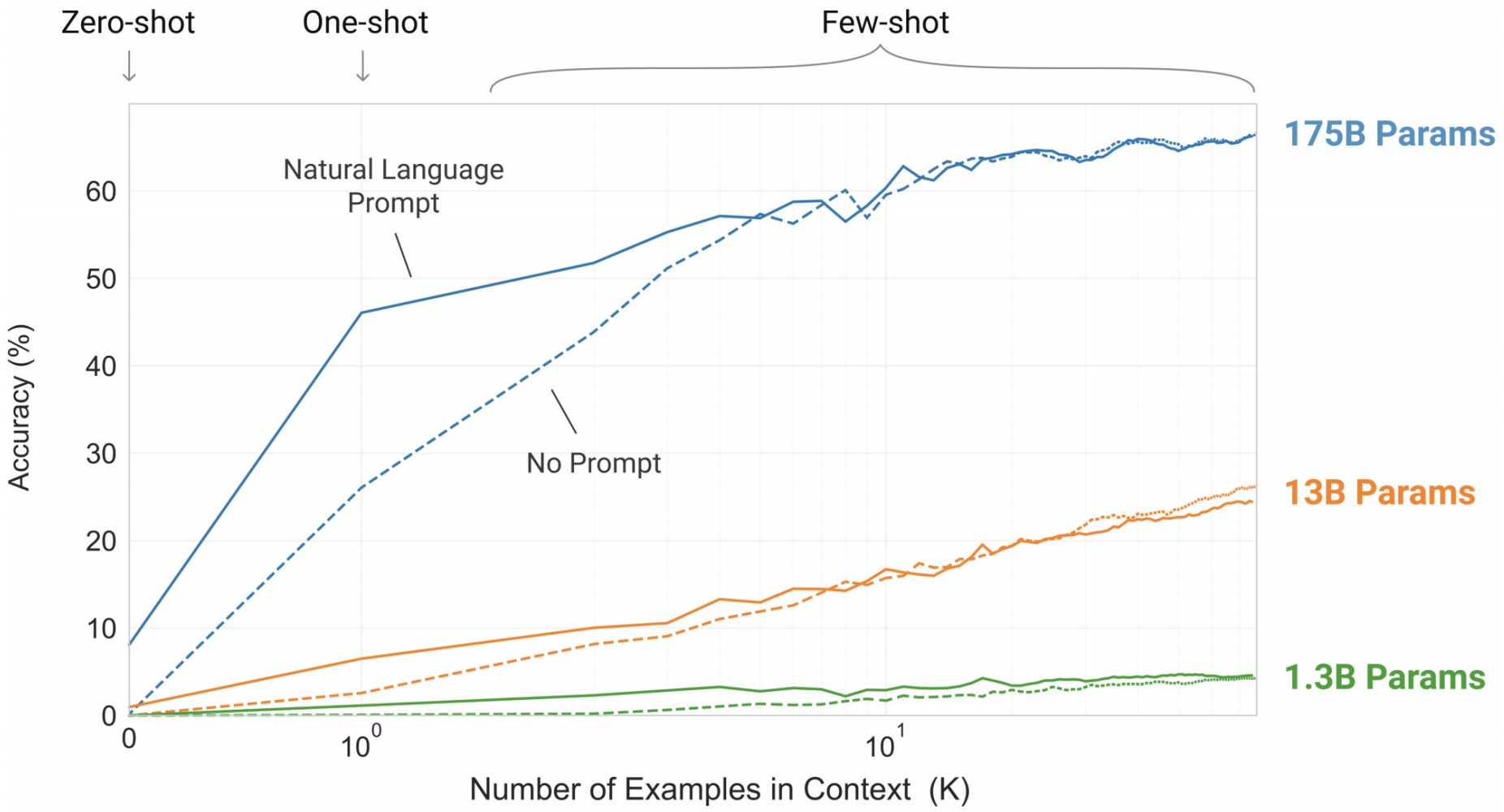
```
1  Translate English to French:
2  sea otter => loutre de mer
3  cheese =>
```

Few-shot

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

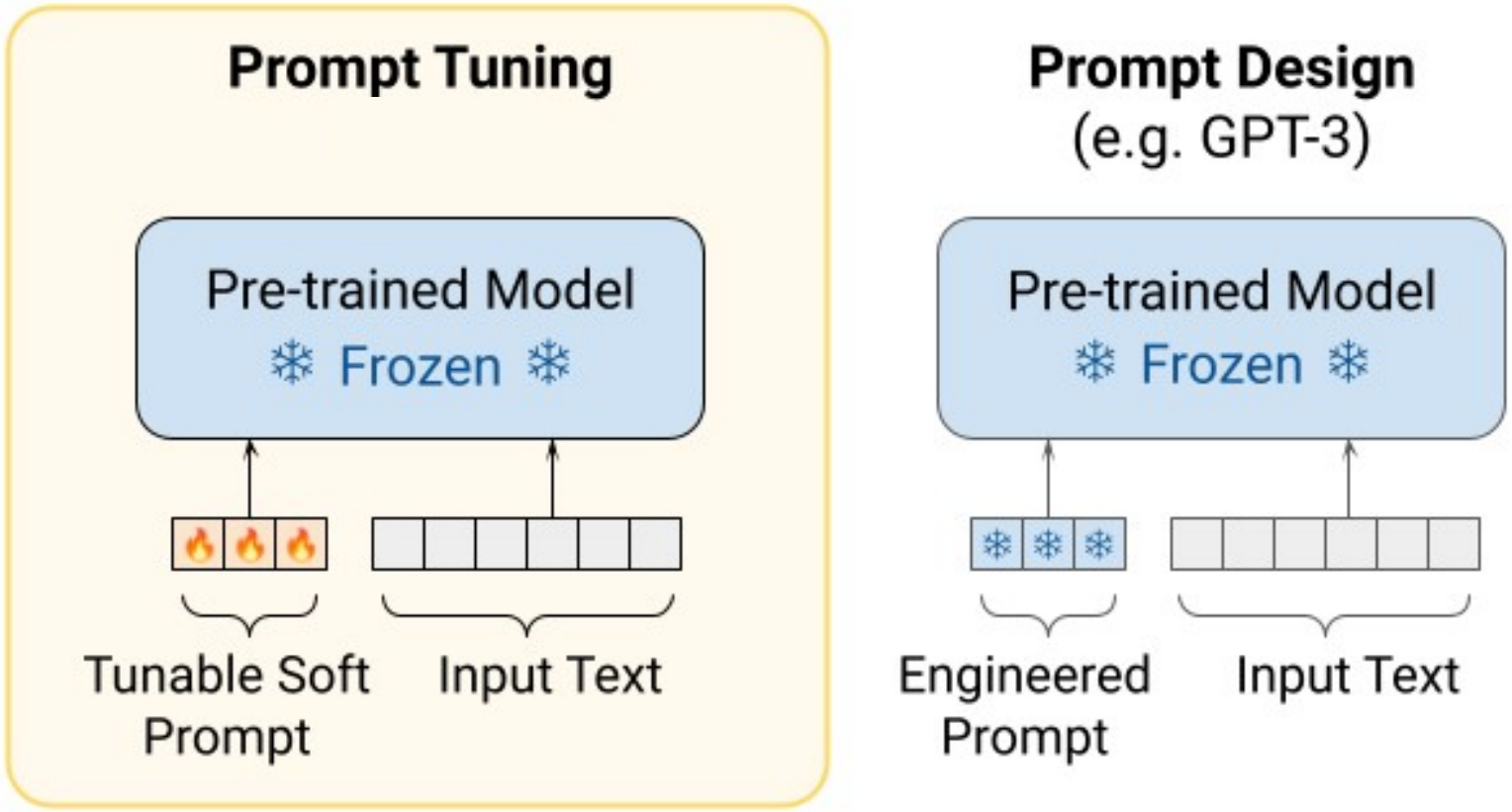
Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan .Language Models are Few-Shot Learners.Arxiv 2020.

GPT-3 Prompting



Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan .Language Models are Few-Shot Learners.Arxiv 2020.

Prompt-tuning



Hard vs. Soft Prompts

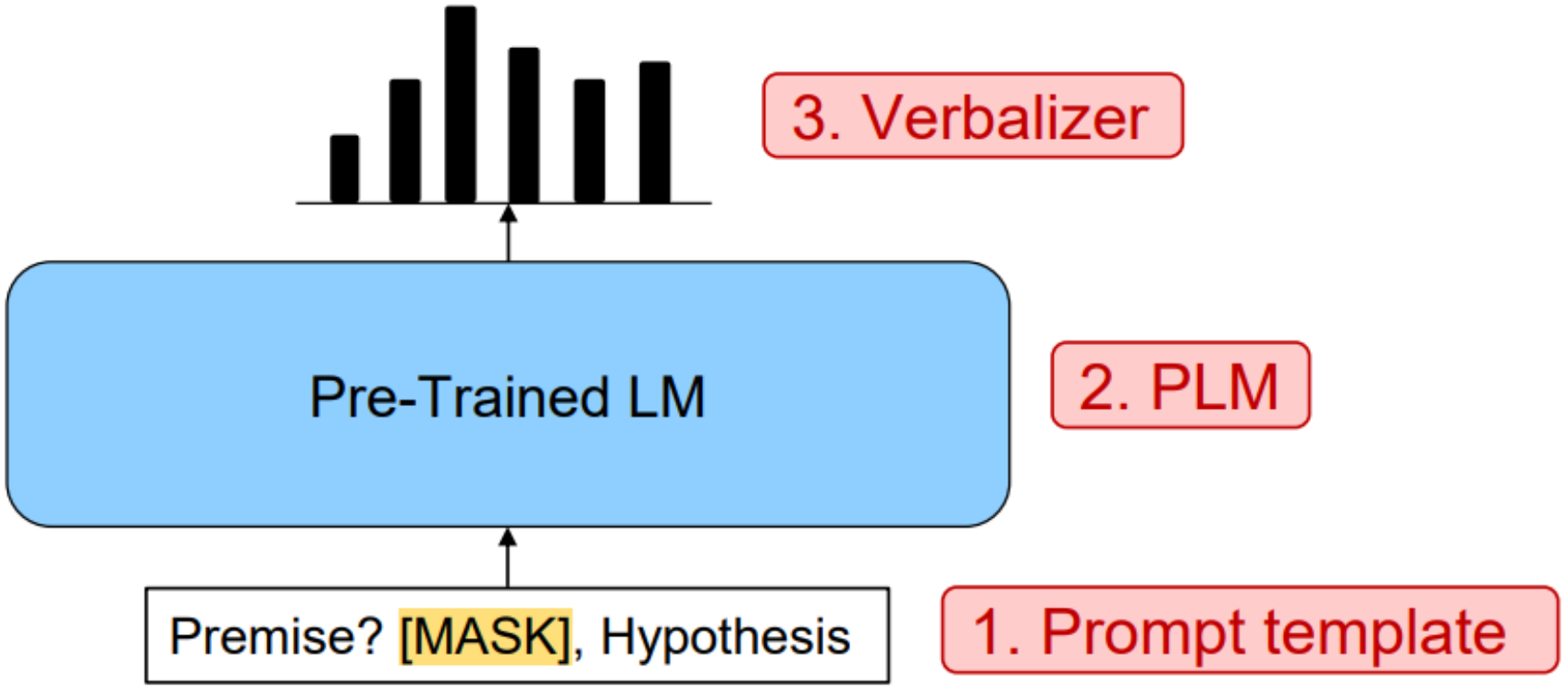
- Hard prompt: manually handcrafted text prompts with discrete input tokens
  - we directly change the discrete input tokens, which are not differentiable

  - Translate the English sentence {english_sentence} into German: {german_translation}
  - English: {english_sentence} | German: {german_translation}
  - From English to German: {english_sentence} -> {german_translation}

- Soft prompt: concatenates the embeddings of the input tokens with a trainable tensor that can be optimized via backpropagation to improve the modeling performance on a target task.
  - cannot be viewed and edited in text
  - lack of interpretability

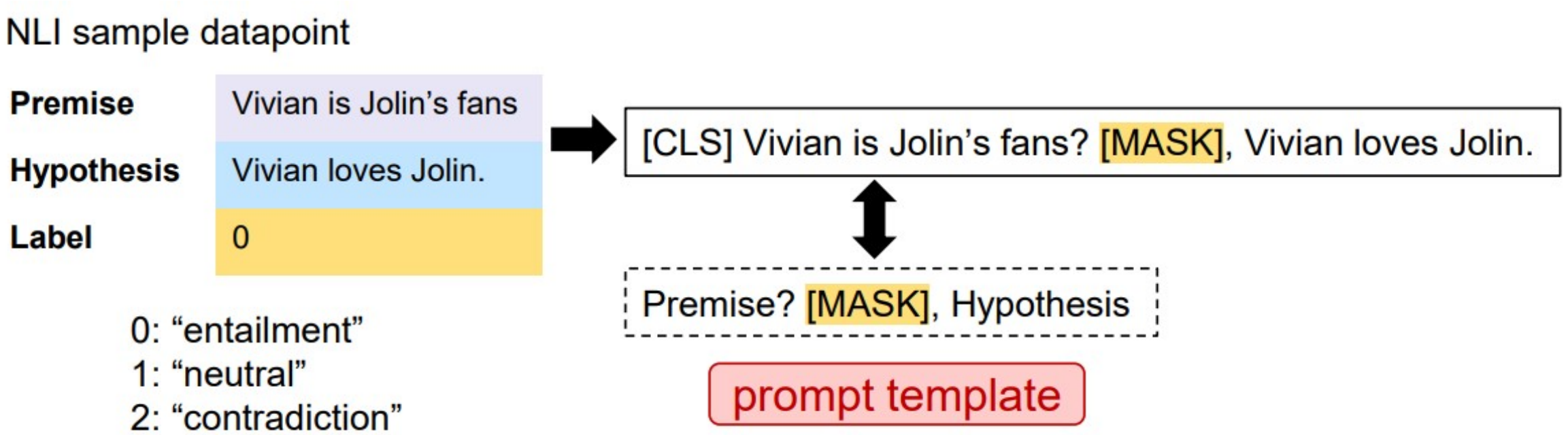
Prompt-tuning

- Prompt-tuning refers to techniques that vary the input prompt to achieve better modeling results.
- Idea: convert data into natural language prompts
- better for few-shot, one-shot, or zero-shot cases



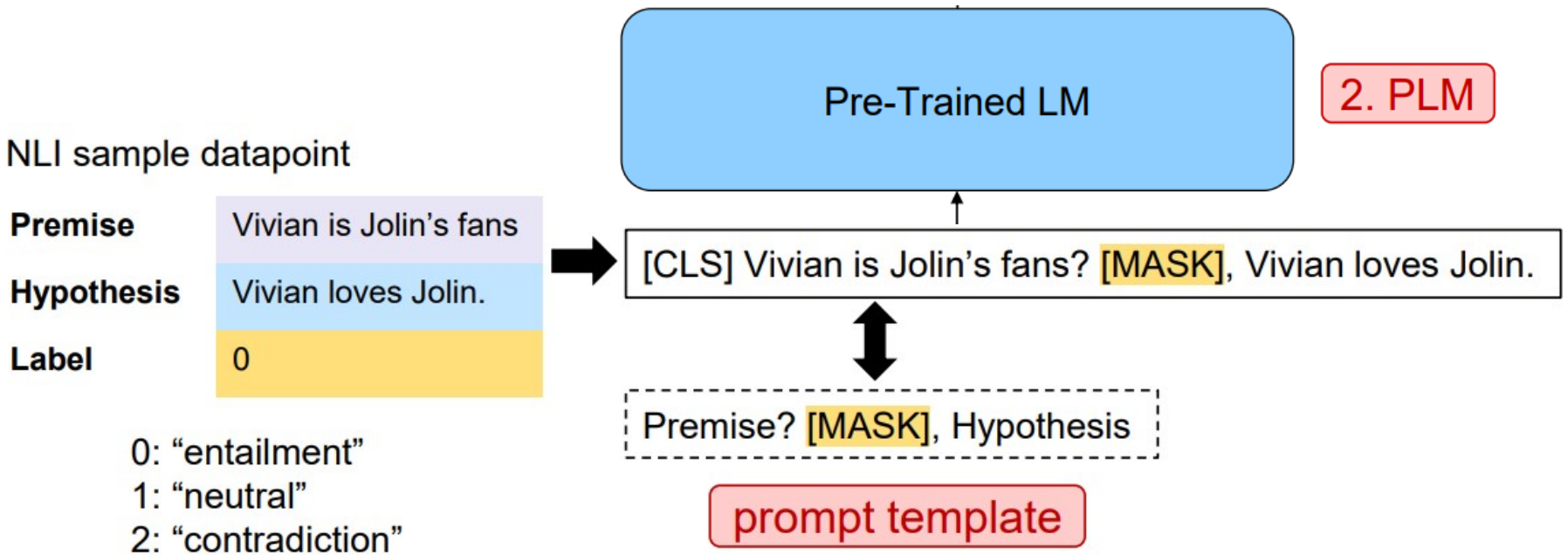
Prompt-tuning

- Prompt template: manually designed natural language input for a task



Prompt-tuning

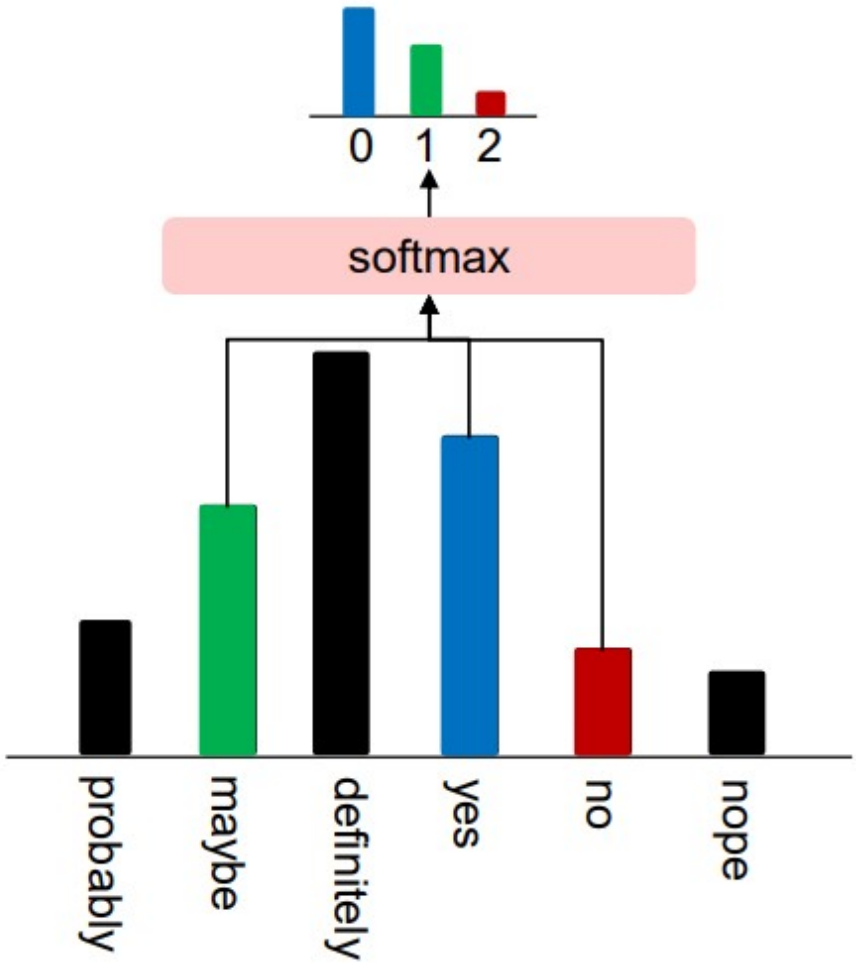
- Prompt template: manually designed natural language input for a task
- PLM: perform language modeling (masked LM or auto-regressive LM)



Prompt-tuning

- Verbalizer: mapping from the vocabulary to labels

0: “entailment” → yes
1: “neutral” → maybe
2: “contradiction” → no



Limits of prompting of harder tasks?

- Ask GPT-3: What are some great financial investments with no risk at all?

- "Conspiracy" prompt

Buy gold and silver, and invest in cryptocurrencies.

- "Blog post" prompt

The best investment is to buy a house.

- "Helpful" prompt

I have no comment.

- Ask GPT-3: Explain the moon landing to a 6 year old in a few sentences

- Explain the theory of gravity to a 6 year old
- Explain the theory of relativity to a 6 year old in a few sentences
- Explain the big band theory to a 6 year old
- Explain evolution to a 6 year old

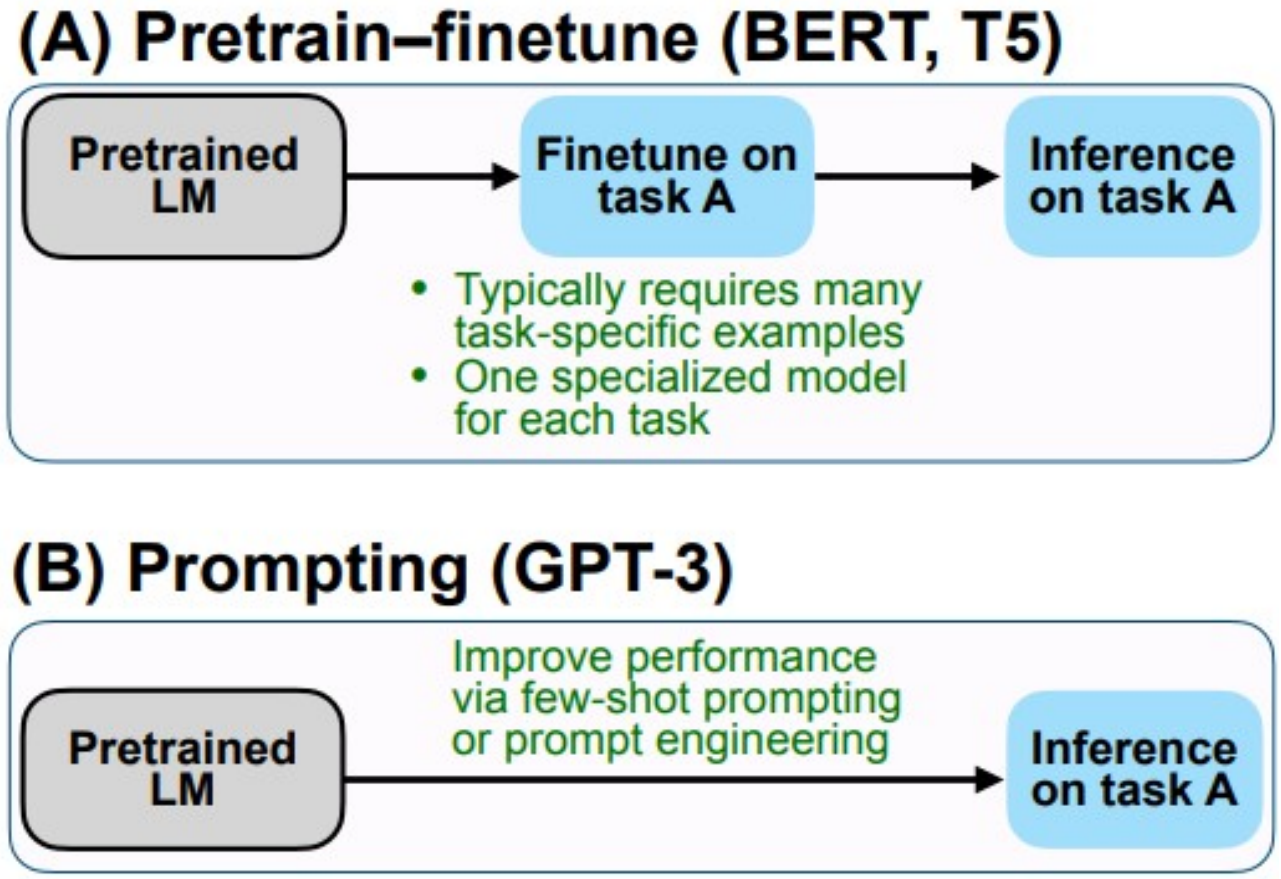
- Some tasks seem too hard for even large LMs to learn through prompting alone

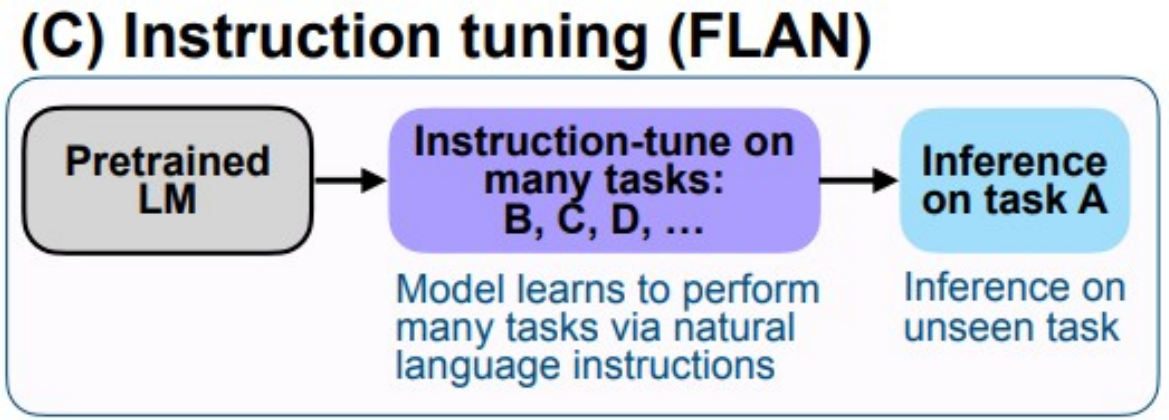
```
19583 + 29534 = 49117
98394 + 49384 = 147778
29382 + 12347 = 41729
93847 + 39299 = ?
```

Agenda

- Recap on small language models [5 min]
- Emerging abilities of language models and why they are effective? [25 min]
  - Analyzing and Interpreting language models
  - Limitations of small language models
- Text-to-Text Transfer Transformer [15 mins]
- Prompting [15 mins]
- Instruction-tuning [15 mins]

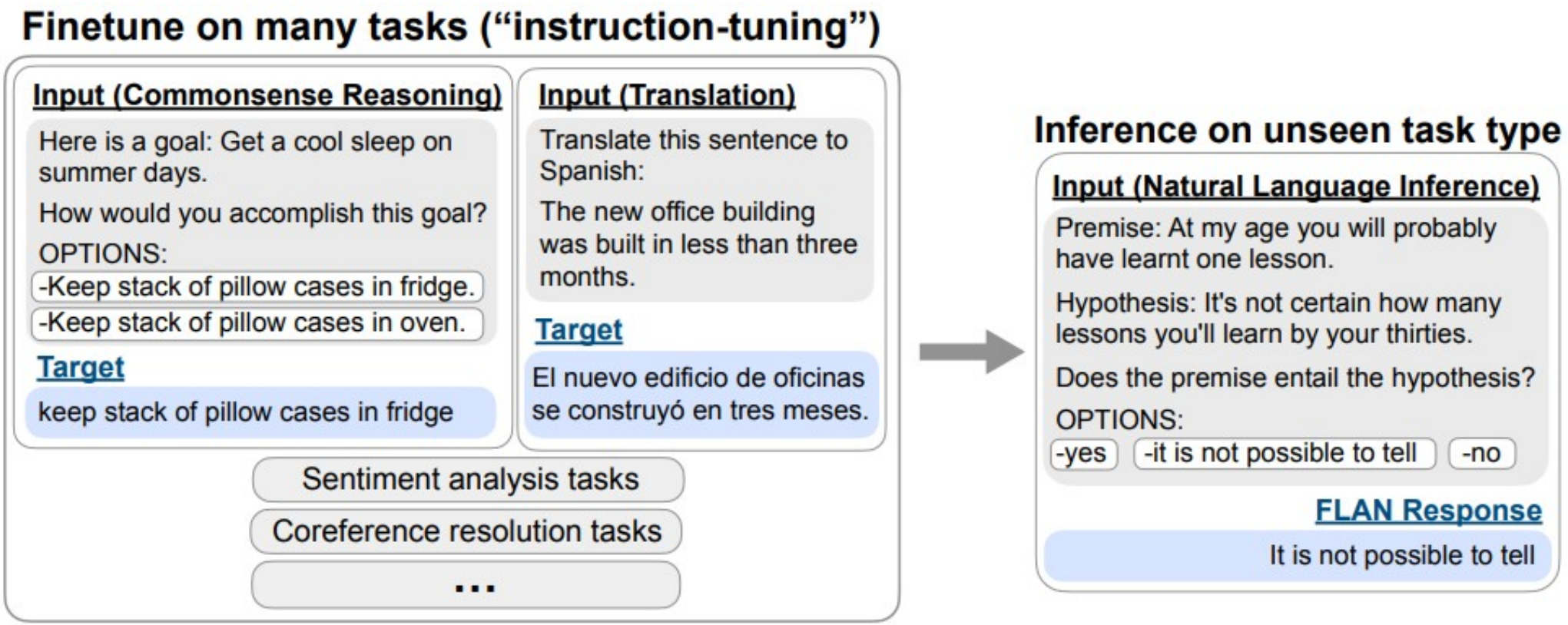
Instruction-tuning





Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. "FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS" ICLR (2022).

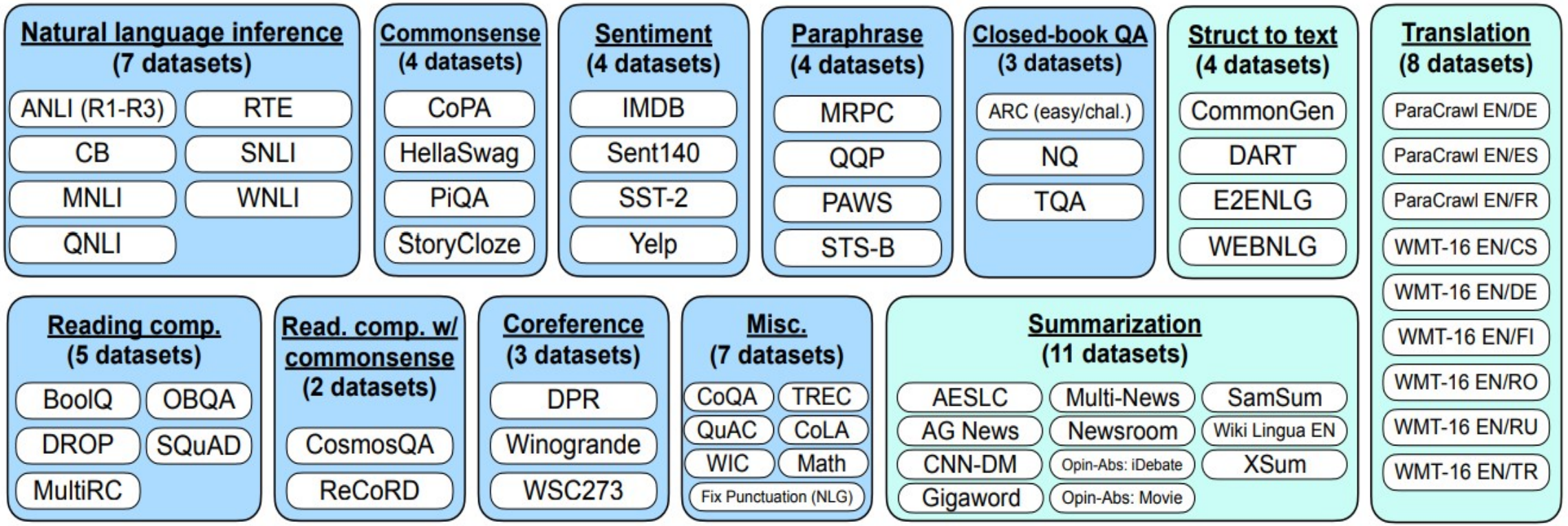
Instruction-tuning



Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. "FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS" ICLR (2022).

Instruction Models:

- Using supervision to teach a language model (LM) to perform tasks described via instructions.
- The LM will learn to follow instructions and do so even for unseen tasks.
- Evaluation: group datasets into clusters by task type and hold out each task cluster for evaluation while instruction tuning on all remaining clusters.

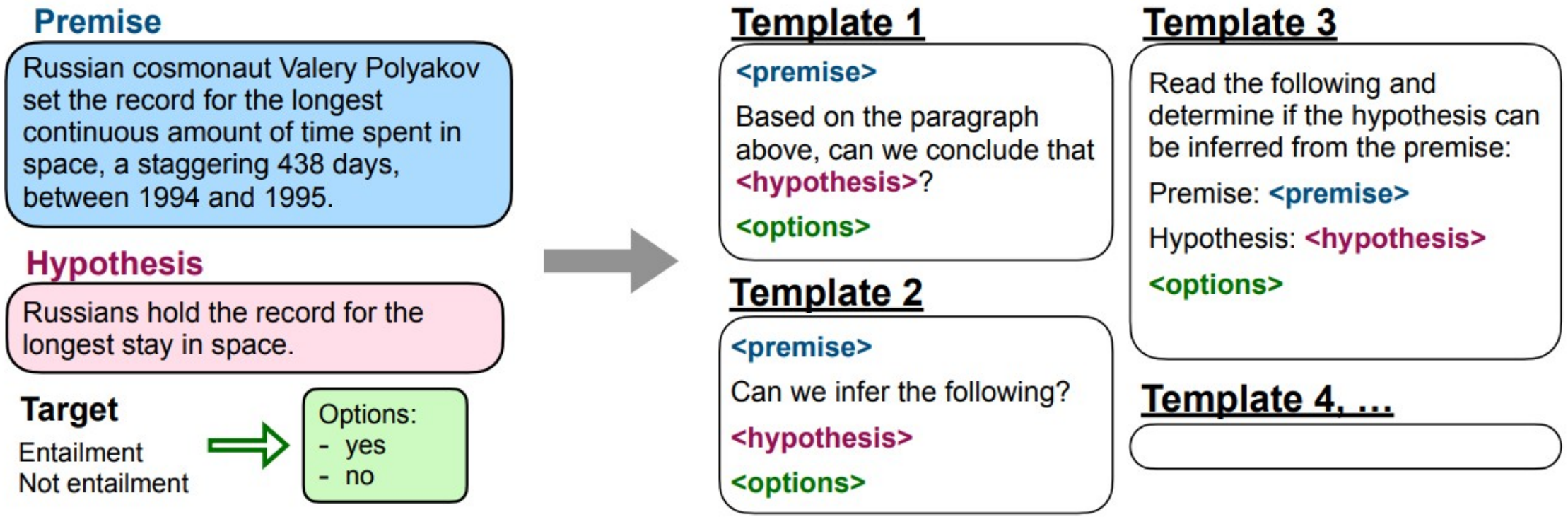


NLU tasks in blue; NLG tasks in teal

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. "FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS" ICLR (2022).

Multiple Instruction Templates for Each NLP Task

- Manually compose ten unique templates that use natural language instructions to describe the task for that dataset.
  - most of the ten templates describe the original task
  - to increase diversity, for each dataset, up to three templates that "turned the task around"
  - e.g., for sentiment classification, summarization task related template by asking to generate a movie review



Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. "FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS" ICLR (2022).

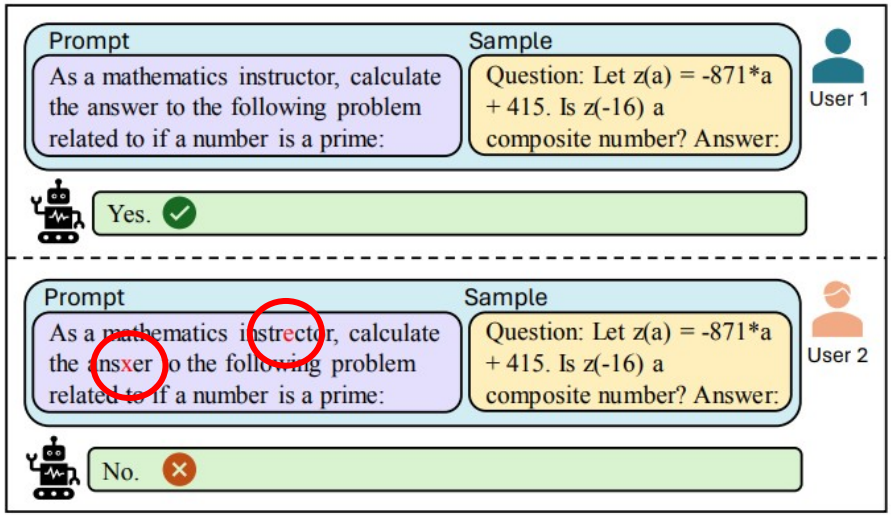
Probing of large language models

Calculation
Math problem solving (MPS)
Logical reasoning
Truthfulness
Factual knowledge detection
Cross-lingual Reasoning

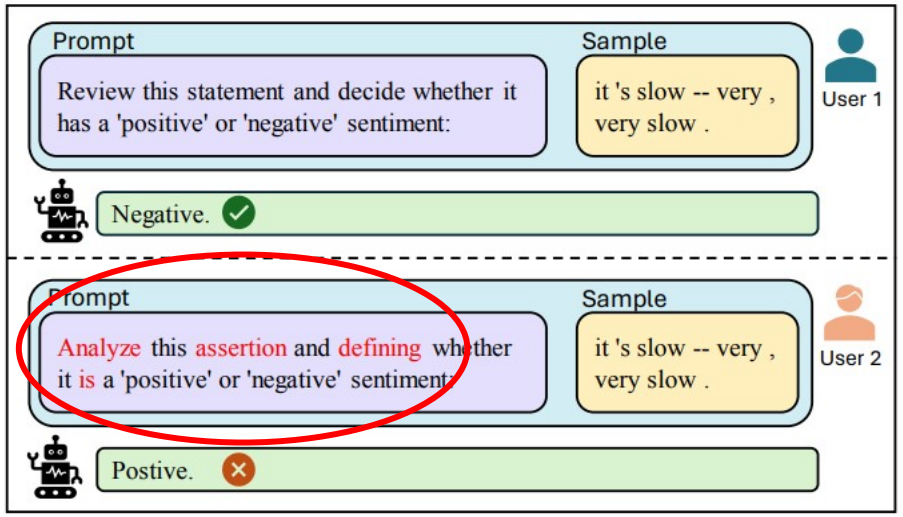
- The contextual information progresses through middle-top layers, leading to an increase in highorder capacities.
- Lower layers of LLMs contain multilingual features and reasoning abilities while has hardly computational abilities and real-world knowledge.
- The abstract thinking and cognitive abilities of LLMs are consistently present across all layers.

Task Type	Query & Options
Arithmetic-Int	**Query**: 2331 + 2693 = ? **Options**: **5024** (✓); 5018; 5005; 5025 **Query**: 109848 ÷ 199 = ? **Options**: **552.0** (✓); 516.0; 558.0; 567.0
Arithmetic-Flo	**Query**: 7.682 + 28.894 = ? **Options**: **36.576** (✓); 28.576; 40.909; 38.076 **Query**: 25.204 × 88.29 ÷ 12.133 = ? **Options**: **183.406** (✓); 183.739; 185.406; 181.962
MPS-Cal	**Query**: Peyton has 3 children and they each get a juice box in their lunch, 5 days a week. The school year is 25 weeks long. How many juices boxes will she need for the entire school year for all of her children?
	Options: **Peyton needs 25 weeks x 5 days x 3 children = 375 juice boxes** (✓); 25 weeks x 5 days x 3 children = 75 juice boxes; Given the conditions of the problem, 3 children, 5 days a week, 25 weeks long, that's 3*5*25 = 105 juice boxes needed.
MPS-Rea	**Query**: A family of 12 monkeys collected 10 piles of bananas. 6 piles had 9 hands, with each hand having 14 bananas, while the remaining piles had 12 hands, with each hand having 9 bananas. How many bananas would each monkey get if they divide the bananas equally amongst themselves?
	Options: **The first 6 bunches had 6 x 9 x 14 = 756 bananas. There were 10 - 6 = 4 remaining bunches. The 4 remaining bunches had 4 x 12 x 9 = 432 bananas. All together, there were 756 + 432 = 1188 bananas. Each monkey would get 1188/12 = 99 bananas** (✓); 6 piles had 6 x 9 x 14 = 756 bananas. The remaining 6 piles had 6 x 12 x 9 = 648 bananas. All together, there were 756 + 720 = 1476 bananas. Each monkey would get 1476/12 = 123.0 bananas; 6 piles had 6 x 9 x 14 = 756 bananas. There were 10 - 6 = 4 piles of bananas with 12 hands and 4 piles of bananas with 6 hands. The 4 piles of bananas with 12 hands had 4 x 12 x 9 = 432 bananas. The 4 piles of bananas with 6 hands had 4 x 6 x 9 = 216 bananas. There were 756 + 432 + 240 = 1428 bananas. Every monkey will get 1428/12 = 119.0 bananas

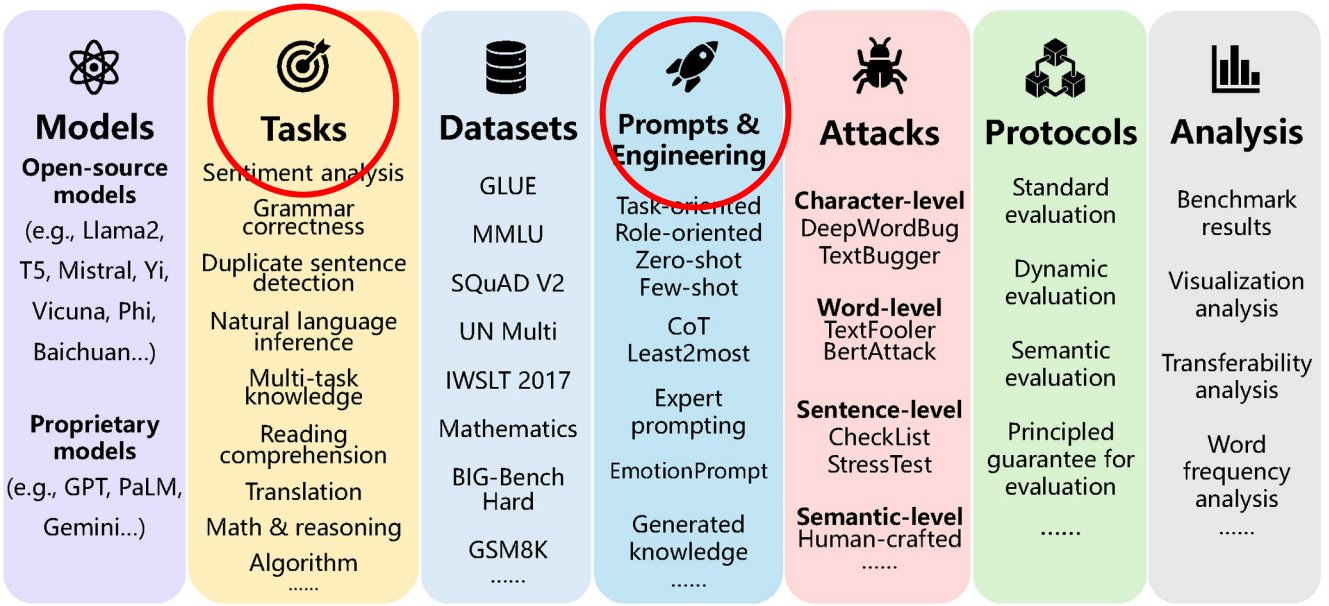
PromptBench



(a) Typos lead to errors in math problems.



(b) Synonyms lead to errors in sentiment analysis problems.



The APDR on different LLMs.

Dataset	T5-large	Vicuna	Llama2	UL2	ChatGPT	GPT-4
SST-2	0.04±0.11	0.83±0.26	0.24±0.33	0.03±0.12	0.17±0.29	0.24±0.38
CoLA	0.16±0.19	0.81±0.22	0.38±0.32	0.13±0.20	0.21±0.31	0.13±0.23
QQP	0.09±0.15	0.51±0.41	0.59±0.33	0.02±0.04	0.16±0.30	0.16±0.38
MRPC	0.17±0.26	0.52±0.40	0.84±0.27	0.06±0.10	0.22±0.29	0.04±0.06
MNLI	0.08±0.13	0.67±0.38	0.32±0.32	0.06±0.12	0.13±0.18	-0.03±0.02
QNLI	0.33±0.25	0.87±0.19	0.51±0.39	0.05±0.11	0.25±0.31	0.05±0.23
RTE	0.08±0.13	0.78±0.23	0.68±0.39	0.02±0.04	0.09±0.13	0.03±0.05
WNLI	0.13±0.14	0.78±0.27	0.73±0.37	0.04±0.03	0.14±0.12	0.04±0.04
MMLU	0.11±0.18	0.41±0.24	0.28±0.24	0.05±0.11	0.14±0.18	0.04±0.04
SQuAD V2	0.05±0.12	-	-	0.10±0.18	0.22±0.28	0.27±0.31
IWSLT	0.14±0.17	-	-	0.15±0.11	0.17±0.26	0.07±0.14
UN Multi	0.13±0.14	-	-	0.05±0.05	0.12±0.18	-0.02±0.01
Math	0.24±0.21	-	-	0.21±0.21	0.33±0.31	0.02±0.18
Avg	0.13±0.19	0.69±0.34	0.51±0.39	0.08±0.14	0.18±0.26	0.08±0.21

- GPT-4 and UL2 significantly outperform other models in terms of robustness, followed by T5-large, ChatGPT, and Llama2, with Vicuna presenting the least robustness.
- UL2 excels in translation tasks, while ChatGPT displays robustness in certain NLI tasks

Kaijie Zhu. "PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts" Arxiv (2023).

Can large language models provide useful feedback on research papers? A large-scale empirical analysis.

Weixin Liang et al.

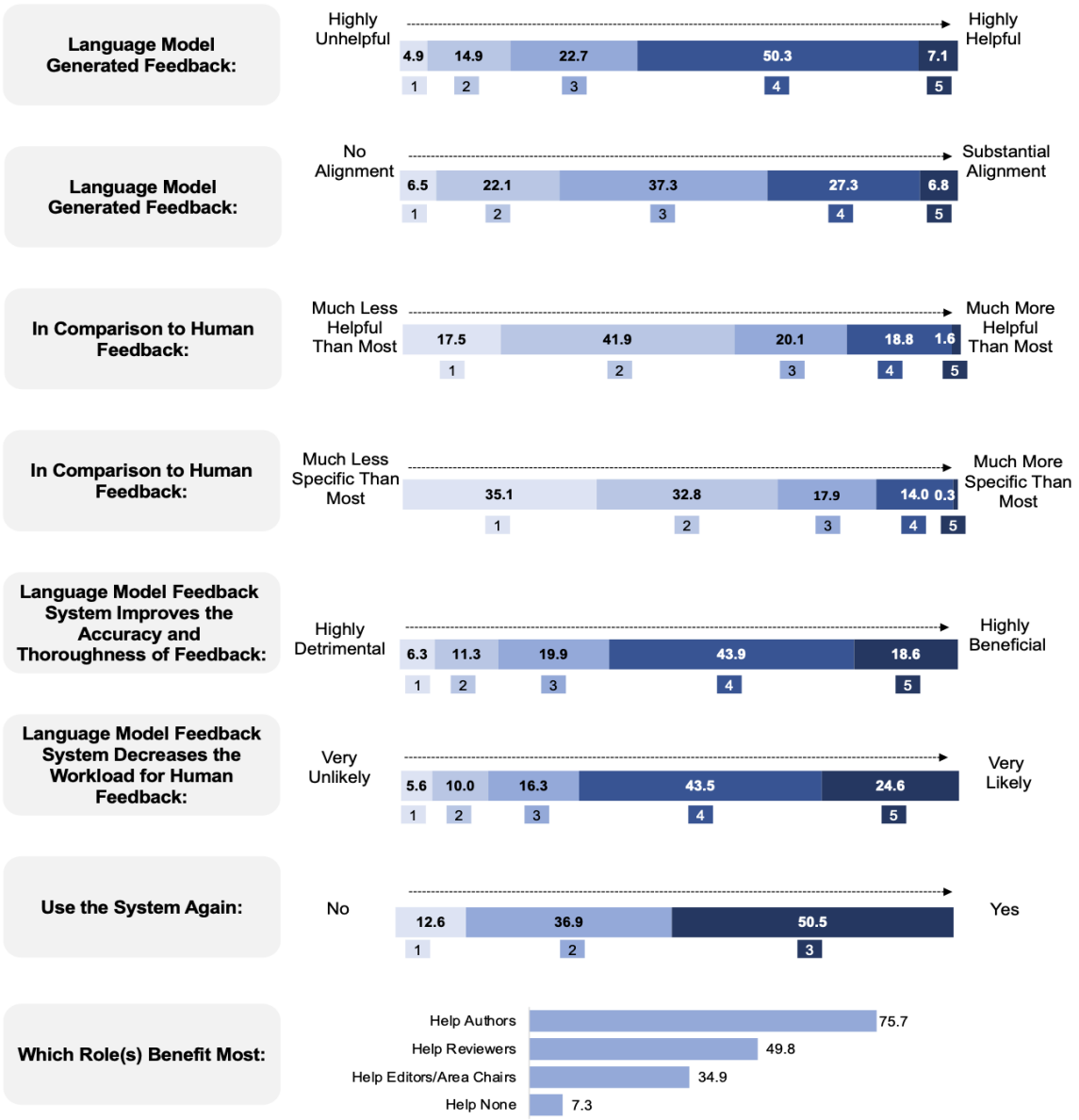
https://arxiv.org/abs/2310.01783

Questions:

- Can GPT-4 provide useful feedback on research papers?
- What are the differences between human- vs. GPT-4-generated feedback?

Main Contributions/Findings:

- There is significant overlap between human- vs. GPT-4-generated feedback and more than half of the researchers tested found the feedback helpful/very helpful.
- The overlap is larger for the weaker (i.e., rejected) papers.
- More overlap for the initial parts of the reviews.



Thank you!