Making Generative AI Better for You:

Fine-tuning and Experimentation for Custom Research Solutions

Shane Storks (he/him)

PhD Candidate, Computer Science and Engineering

Situated Language and Embodied Dialogue (SLED) Lab

MIDAS Generative AI Tutorial Series

November 29, 2023

MICHIGAN INSTITUTE FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

AI LAB

Large Language Models (LLMs)

LLMs like ChatGPT and GPT-4 have recently gained popularity due to their impressive language understanding and reasoning capabilities, making them useful assistants for a variety of language tasks.

You
When is the next tutorial session based on the below information?

Transforming Your Research with Generative AI – Tutorials
MIDAS and the Michigan AI Laboratory will jointly offer a training series to researchers across research fields as they start incorporating Generative AI in their research. Each will consist of lectures, demonstrations, and hands-on tutorials for using Generative AI in research.

PAST EVENTS
10/2/23: Writing, Planning and Literature Review: Enhancing Professional Productivity with Generative AI

10/18/23: Code Smarter, Not Harder: Harnessing Generative AI for Research Programming Efficiency

10/27/23: Integrating Generative Image AI into Your Research Workflow

...

ChatGPT
The next tutorial session in the series "Transforming Your Research with Generative AI" is titled "Making Generative AI Better for You: Fine-tuning and Experimentation for Custom Research Solutions." This session is scheduled for Wednesday, November 29, 2023, from 2:30 PM to 4:30 PM. It will be held on the 10th Floor of Weiser Hall, located at 500 Church Street, Ann Arbor. The prerequisites for this session include some experience with machine learning.

How can we customize them and apply them to empirical research?

https://chat.openai.com/
OpenAI. GPT-4 Technical Report. arXiv: 2303.08774.

Role of LLMs in Research

- LLMs can be helpful *assistants* for tasks like writing and coding, but they can do so much more!
- They can also be useful to automate aspects of:
  - Data annotation
  - Domain-specific content generation
  - Any language-based applications
- May not perform well at specialized tasks like these out of the box
- **How can we *customize* LLMs to adapt them to various specialized language tasks?**

MIDAS Using Generative AI for Scientific Research User Guide

Outline

- The Road to LLMs
- Fine-Tuning LLMs
- Prompting LLMs

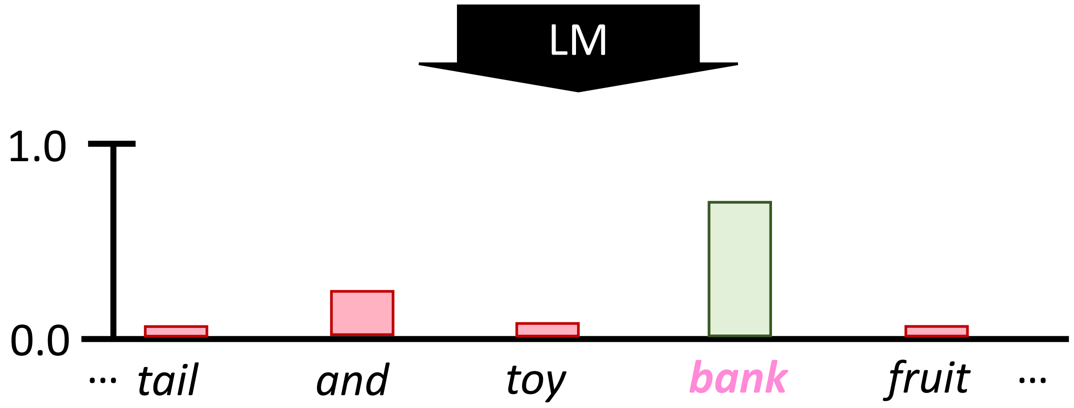
Outline

- **The Road to LLMs**
- Fine-Tuning LLMs
- Prompting LLMs

Language Models (LMs)

$$p(w_n | w_1, w_2, \ldots, w_{n-1})$$

*Jack needed some **money**, so he went and shook his **piggy** ____*



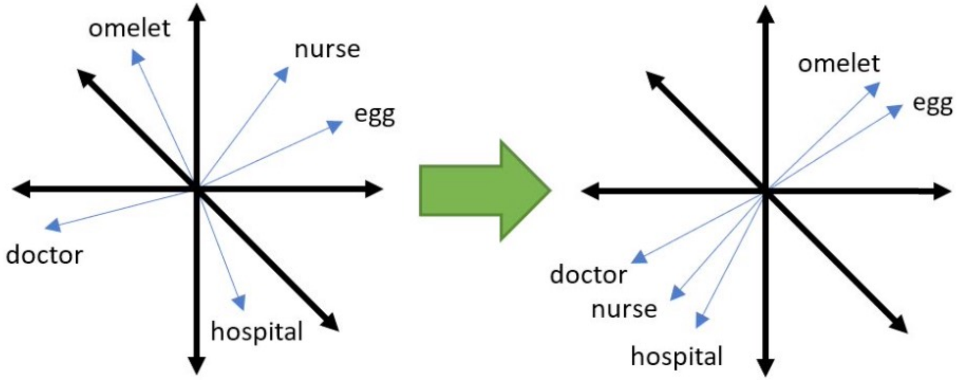
$p($____$|$*Jack needed … shook his piggy*$)$

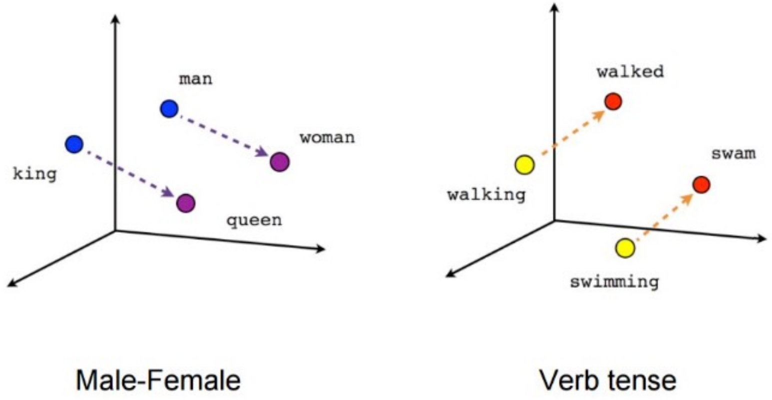
(dreamstime)

$p(\textit{Jack needed …shook his piggy bank}) \approx p(\textit{bank}|\textit{Jack needed…shook his piggy}) \cdot p(\textit{piggy}|\textit{Jack needed…shook his}) \cdot p(\textit{his}|\textit{Jack needed…shook}) \cdot \cdots$

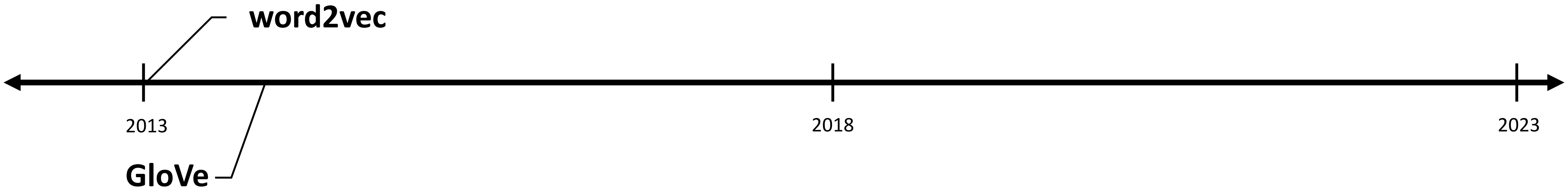
Minsky, M. (2000). Commonsense-based interfaces. In *Commun. ACM*, 43(8): p. 66-73.

Vector-Based Word Embeddings



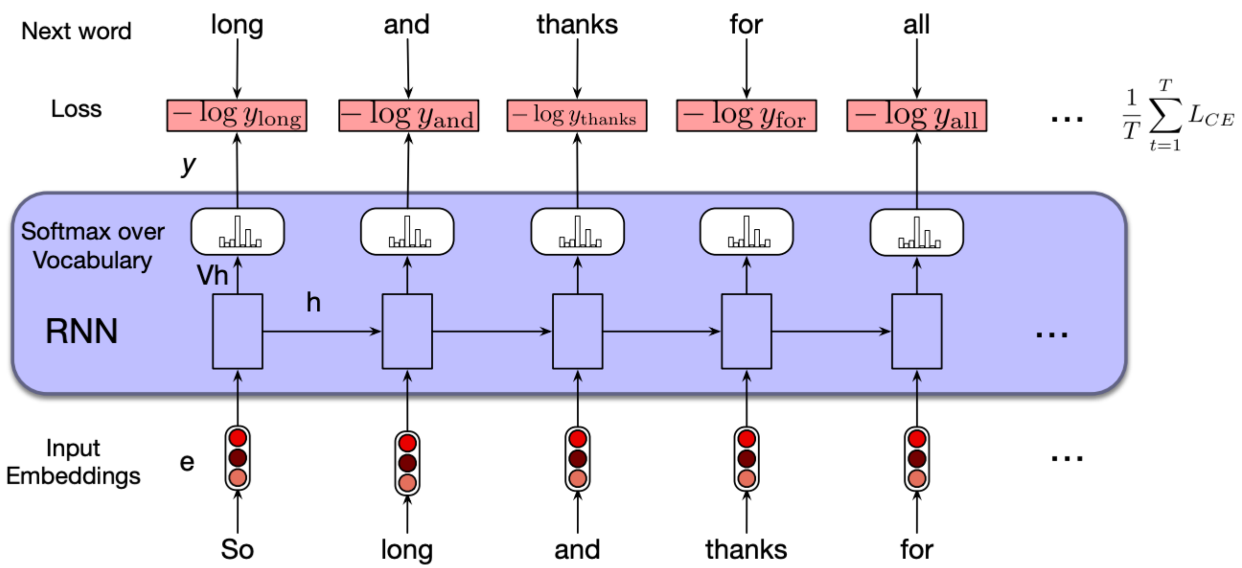


(Image from TensorFlow docs)

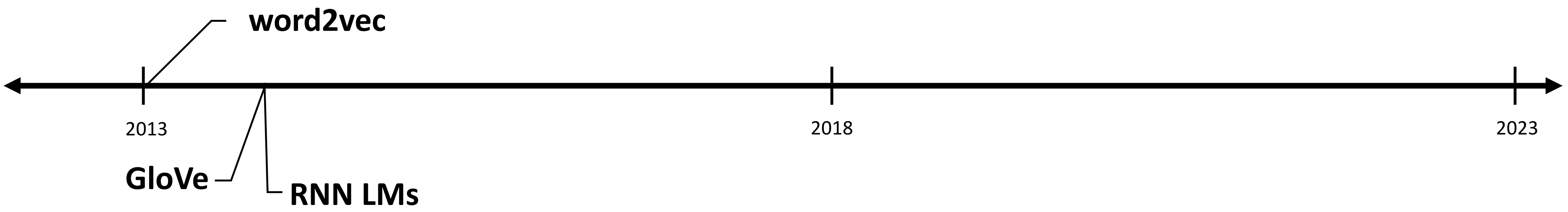


Tomas Mikolov, Kai Chen, Greg Corrado, & Jeffrey Dean. (2013). "Efficient Estimation of Word Representations in Vector Space." *International Conference on Learning Representations 2013.*
Tomas Mikolov, Ilya Sutskever, Kai Chen, et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality." *Advanced in Neural Information Processing Systems 26.*
Jeffrey Pennington, Richard Socher, & Christopher Manning. (2014). "GloVe: Global Vectors for Word Representation." *2014 Conference on Empirical Methods in Natural Language Processing.*

Representing Sequences of Words

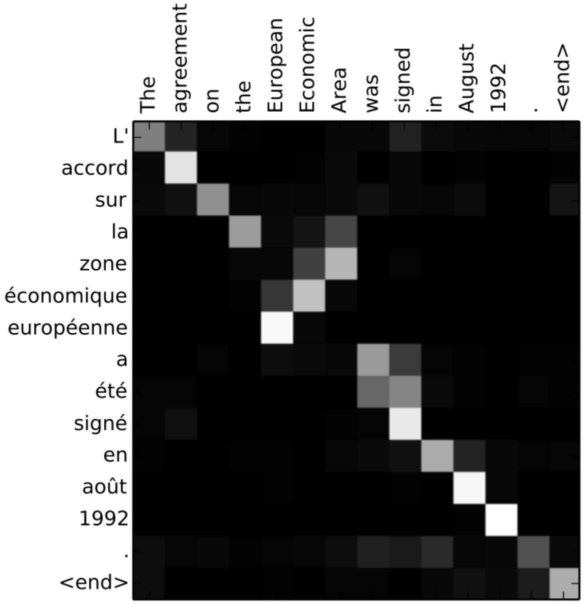


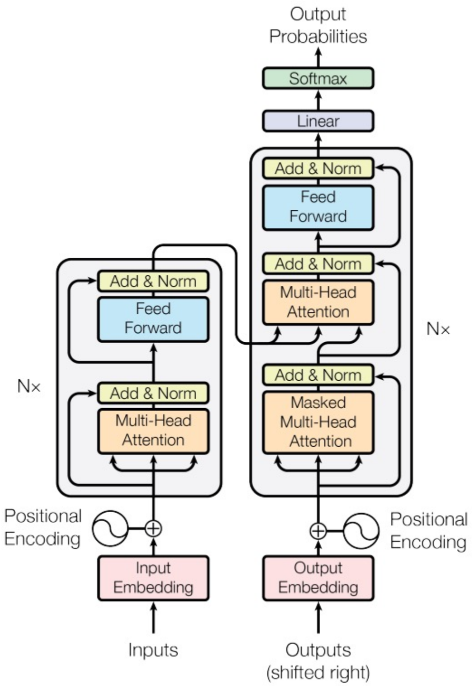
(figure from Jurafsky and Martin)

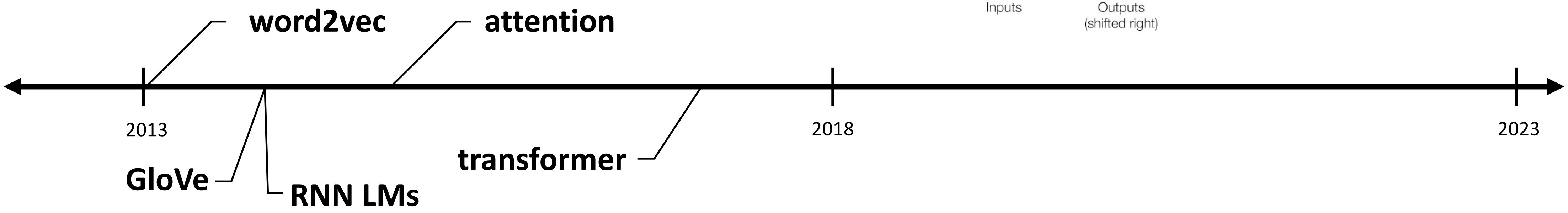


Ilya Sutskever, Oriol Vinyals, & Quoc V. Le. (2014). Sequence to Sequence Learning with Neural Networks. arXiv: 1409.3215.
Rafal Jozefowicz, Wojciech Zremba, & Ilya Sutskever. (2015). An Empirical Exploration of Recurrent Network Architectures. *ICML 2015.*

Attention and Transformers



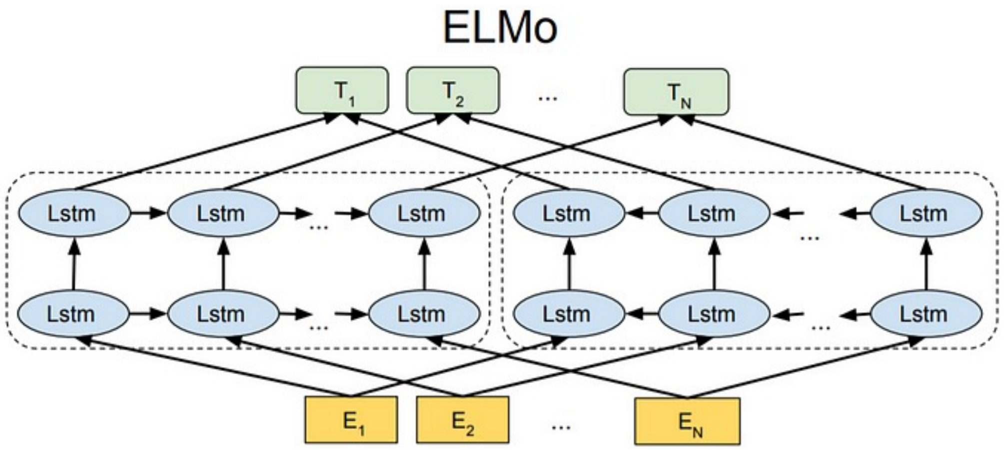




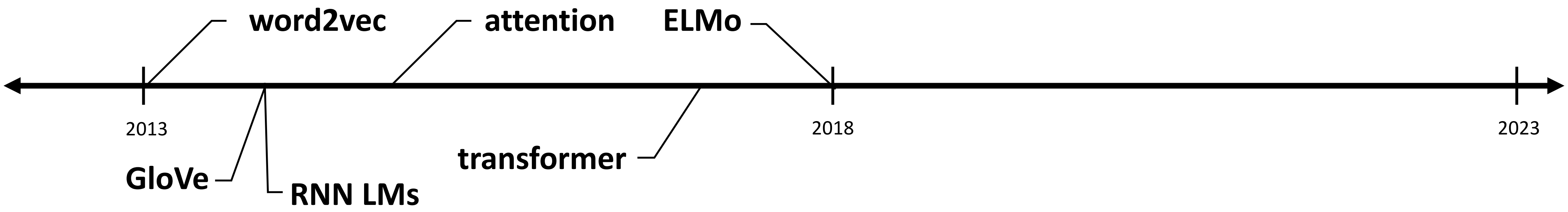
Dzmitry Bahdanau, Kyunghyun Cho, & Yoshua Bengio. (2015). ”Neural Machine Translation by Jointly Learning to Align and Translate.” *International Conference on Learning Representations 2015.*
Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. (2017). “Attention is All You Need.” *Advances in Neural Information Processing Systems 30.*

Contextual Language Representations

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
ELMo	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

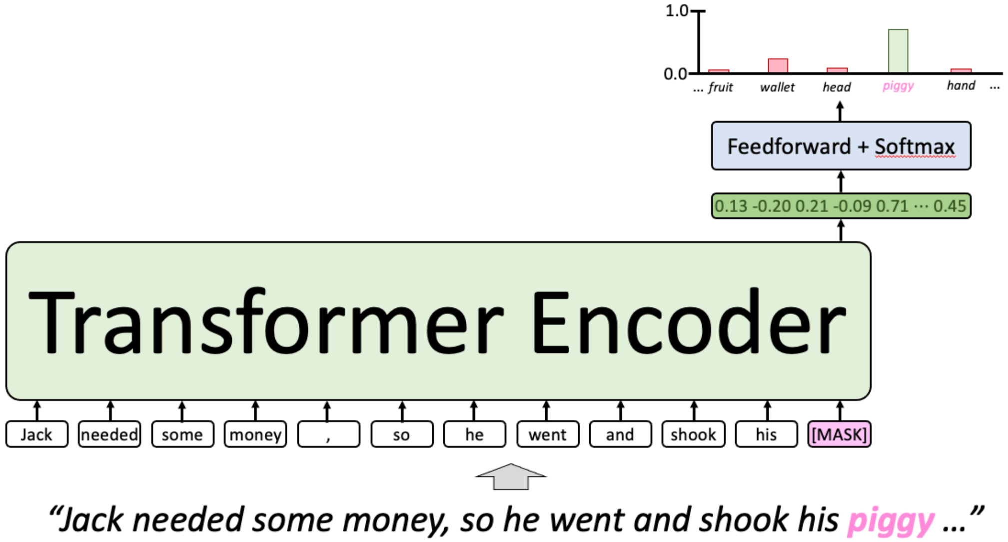


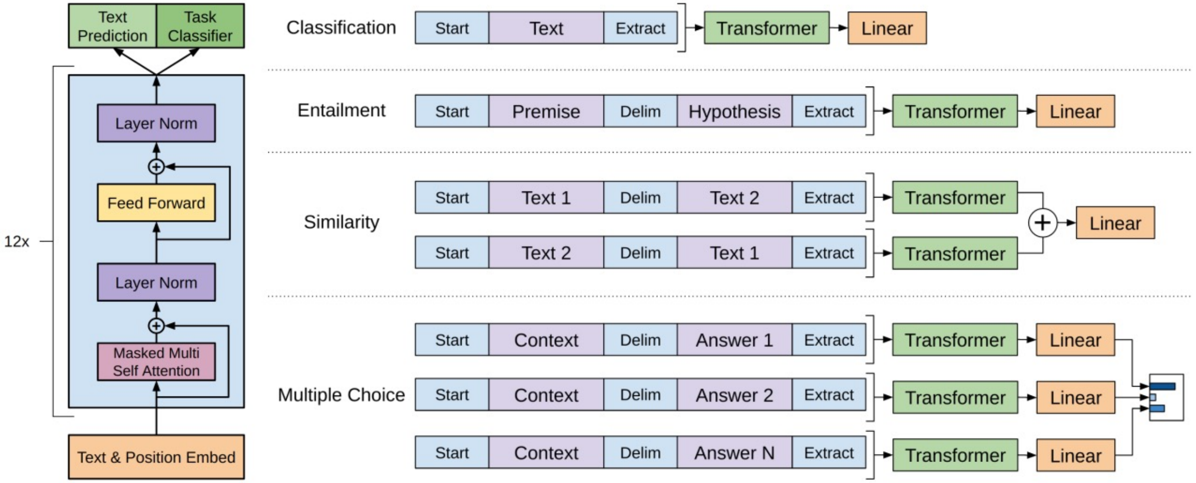
(Karan Purohit)

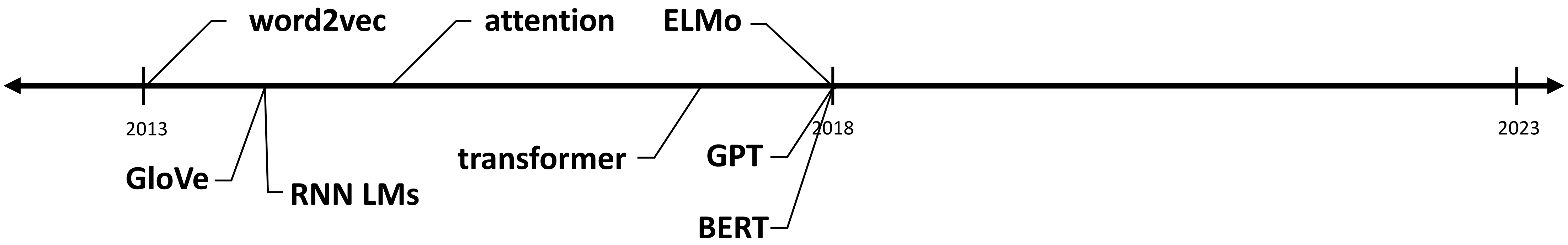


Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. (2018). "Deep Contextualized Word Representations." *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Self-Supervision and Transfer Learning in LMs



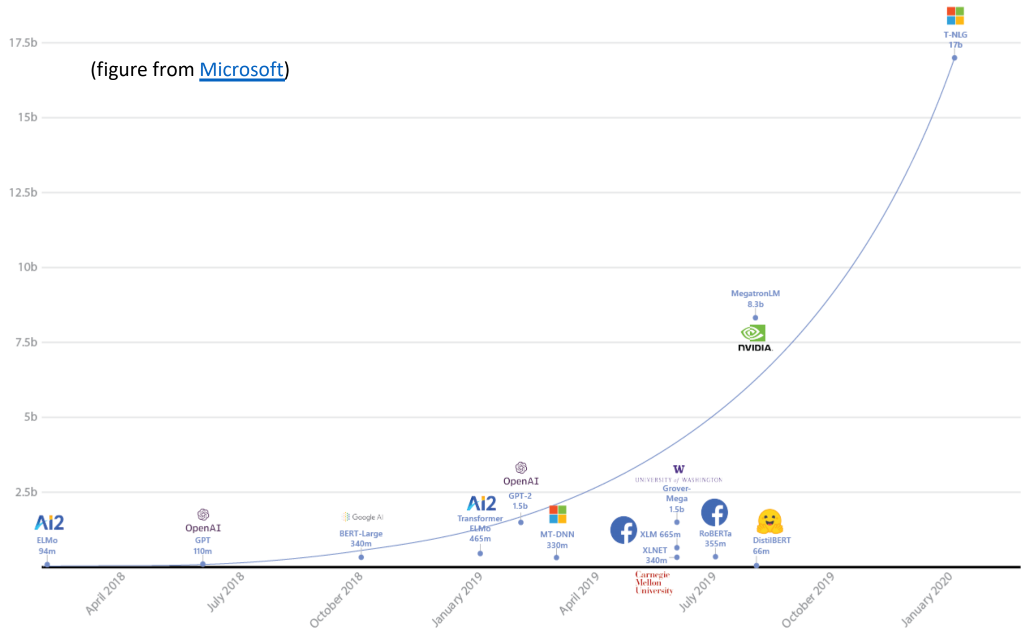


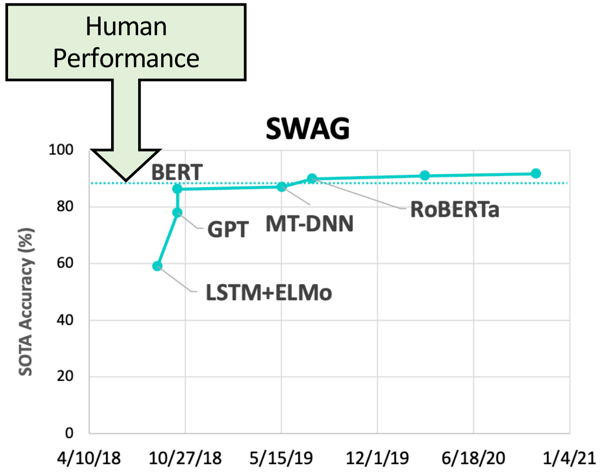


Alec Radford, Karthik Narasimhan, Tim Salimans, & Ilya Sutskever. (2018). "Improving Language Understanding by Generative Pre-Training."

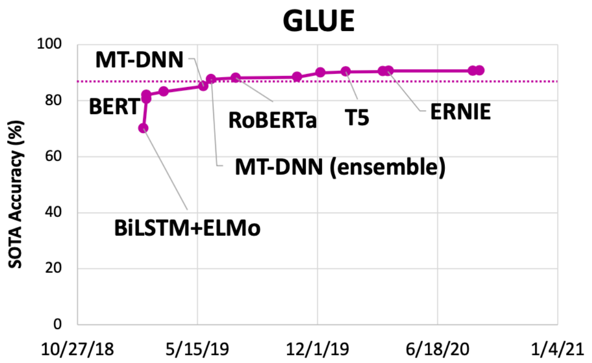
Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2018). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Bigger Data & Bigger Models -> LLMs

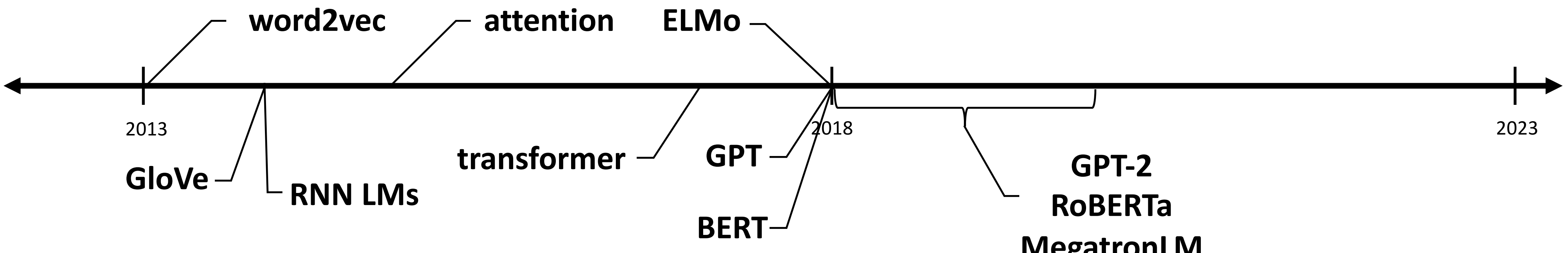




https://leaderboard.allenai.org/swag/submissions/public



https://gluebenchmark.com/leaderboard



GPT-2
RoBERTa
MegatronLM
Turing-NLG
…

Alec Radford, Jeff Wu, Rewon Child, et al. (2019). "Language Models are Unsupervised Multitask Learners."
Yinhan Liu, Myle Ott, Naman Goyal, et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692.
Mohammad Shoeybi, Mostofa Patwary, Raul Puri, et al. (2019). Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv: 1909.08053.
Corby Rosset. (2020). Turing-NLG: A 17-billion-parameter language model by Microsoft.

Prompting & In-Context Learning

Zero-shot

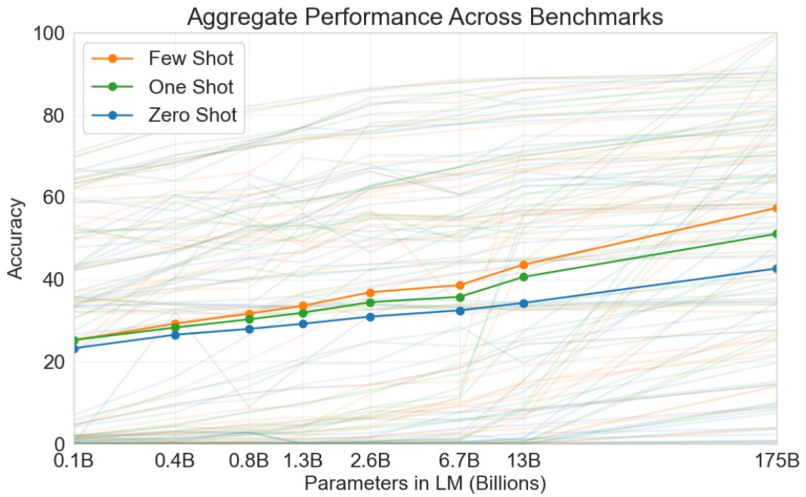
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

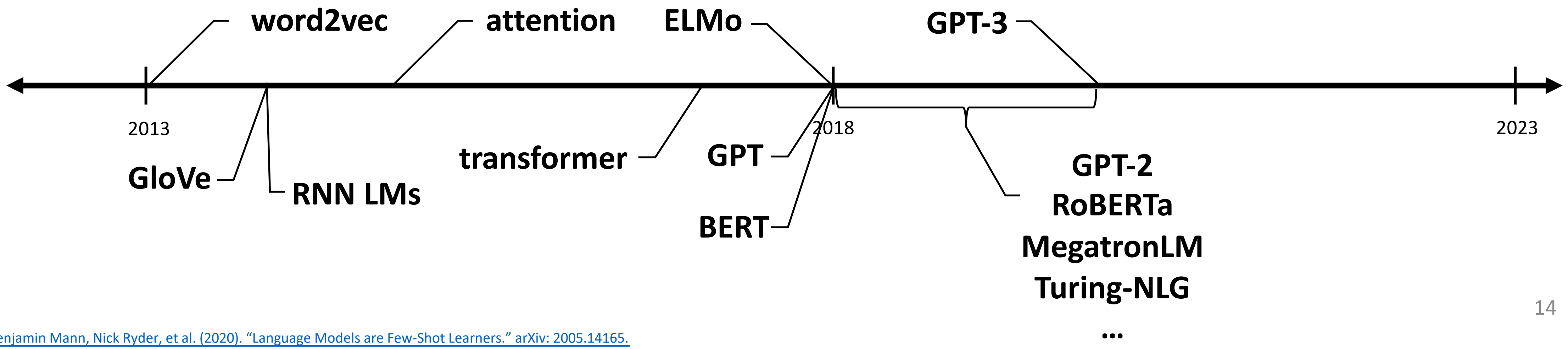
```
1  Translate English to French:     ← task description
2  cheese =>                         ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

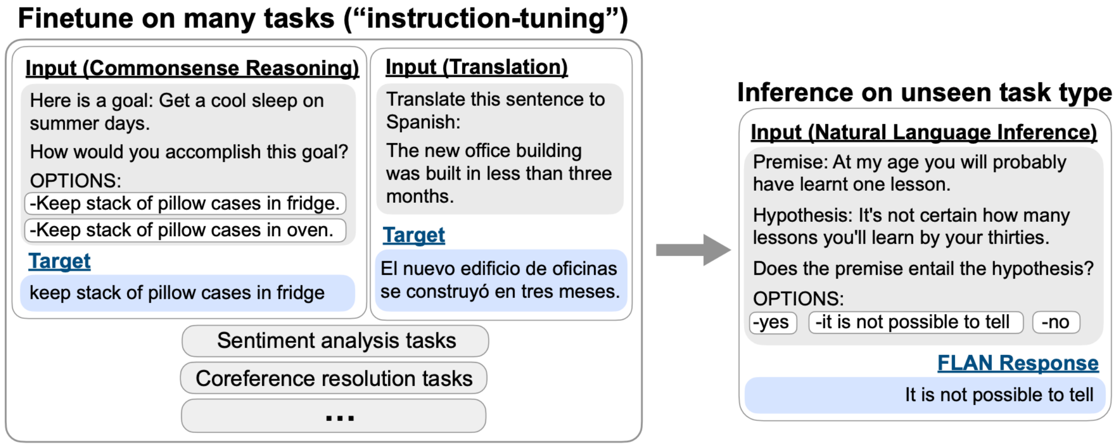
```
1  Translate English to French:     ← task description
2  sea otter => loutre de mer        ← examples
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>                         ← prompt
```

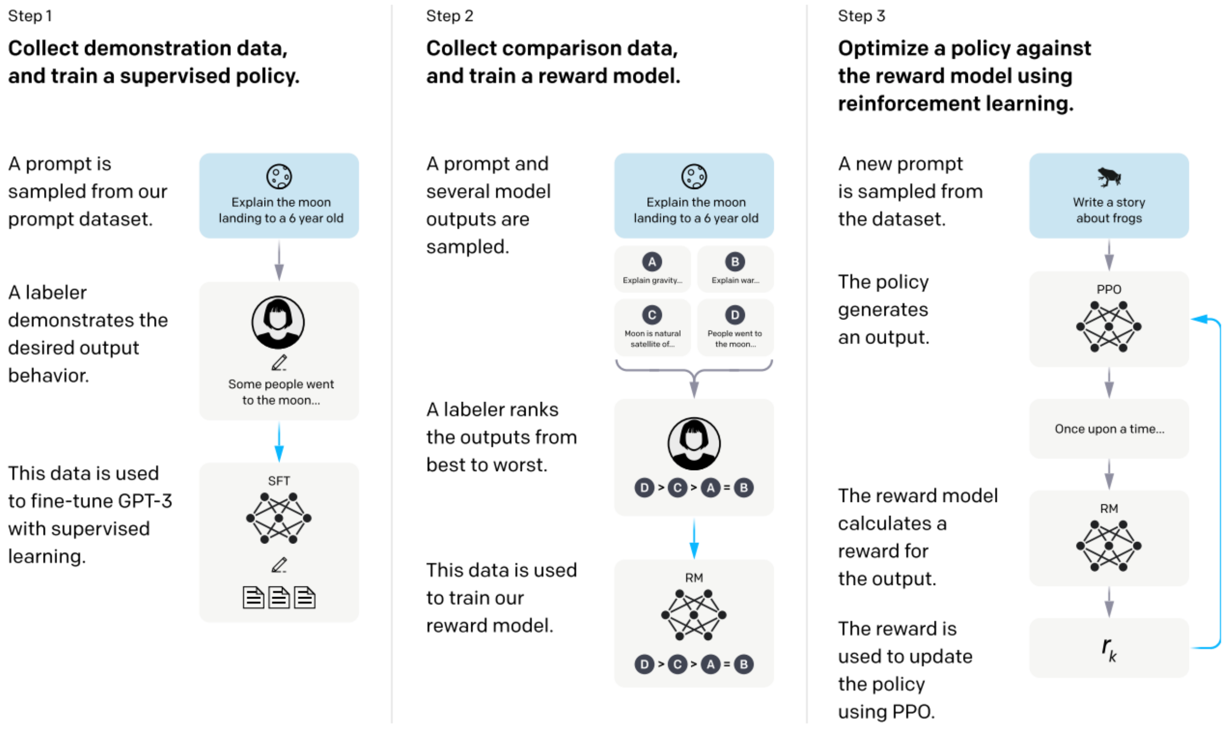


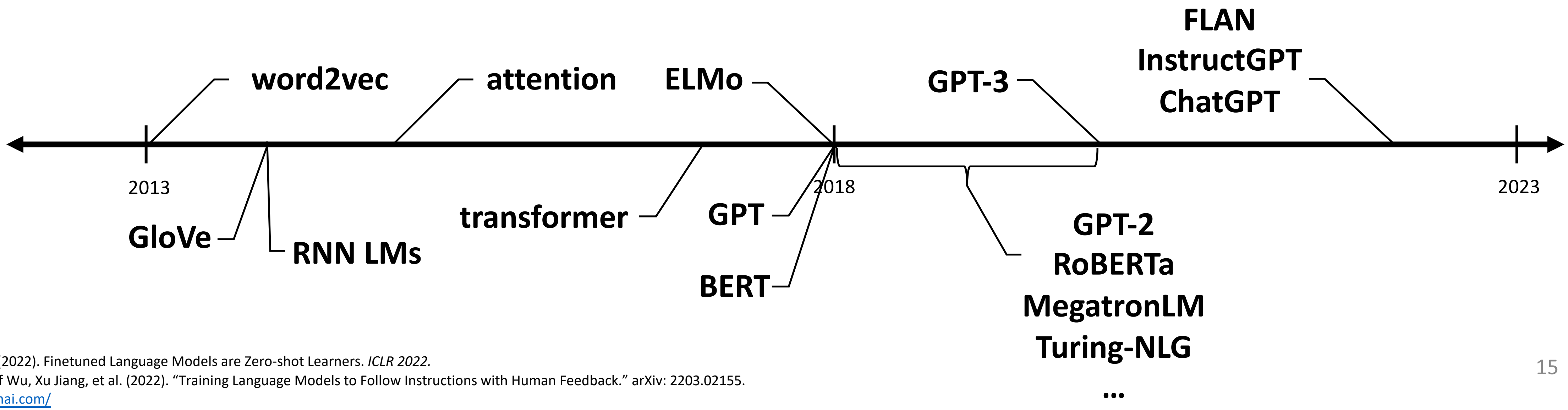


Tom B. Brown, Benjamin Mann, Nick Ryder, et al. (2020). "Language Models are Few-Shot Learners." arXiv: 2005.14165.

Instruction Tuning

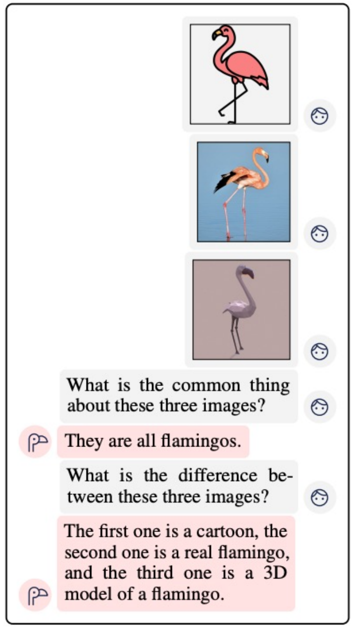






Jason Wei, et al. (2022). Finetuned Language Models are Zero-shot Learners. *ICLR 2022.*
Long Ouyang, Jeff Wu, Xu Jiang, et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." arXiv: 2203.02155.
https://chat.openai.com/

Vision & Multimodality

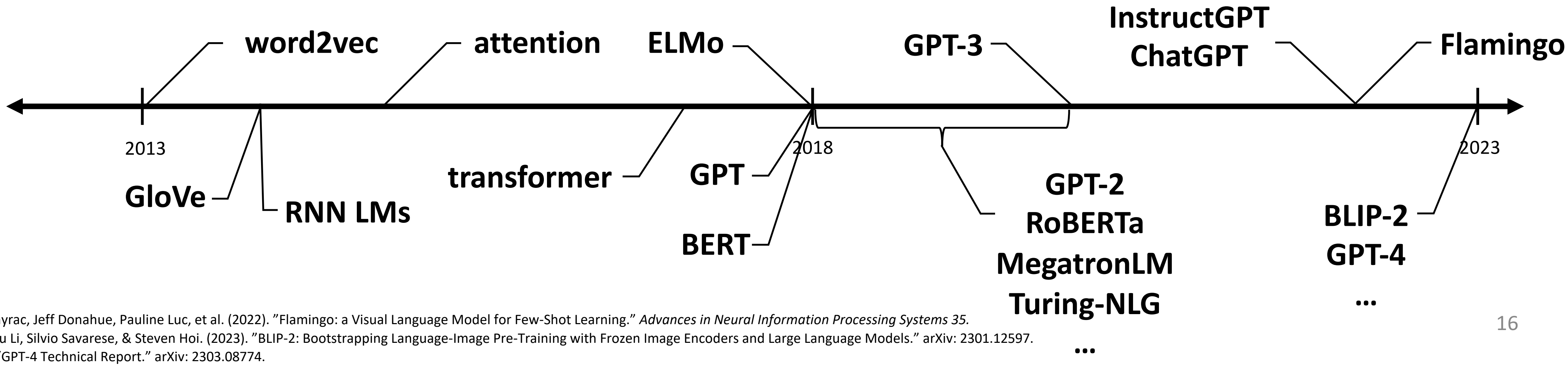


GPT-4 visual input example, Extreme Ironing:

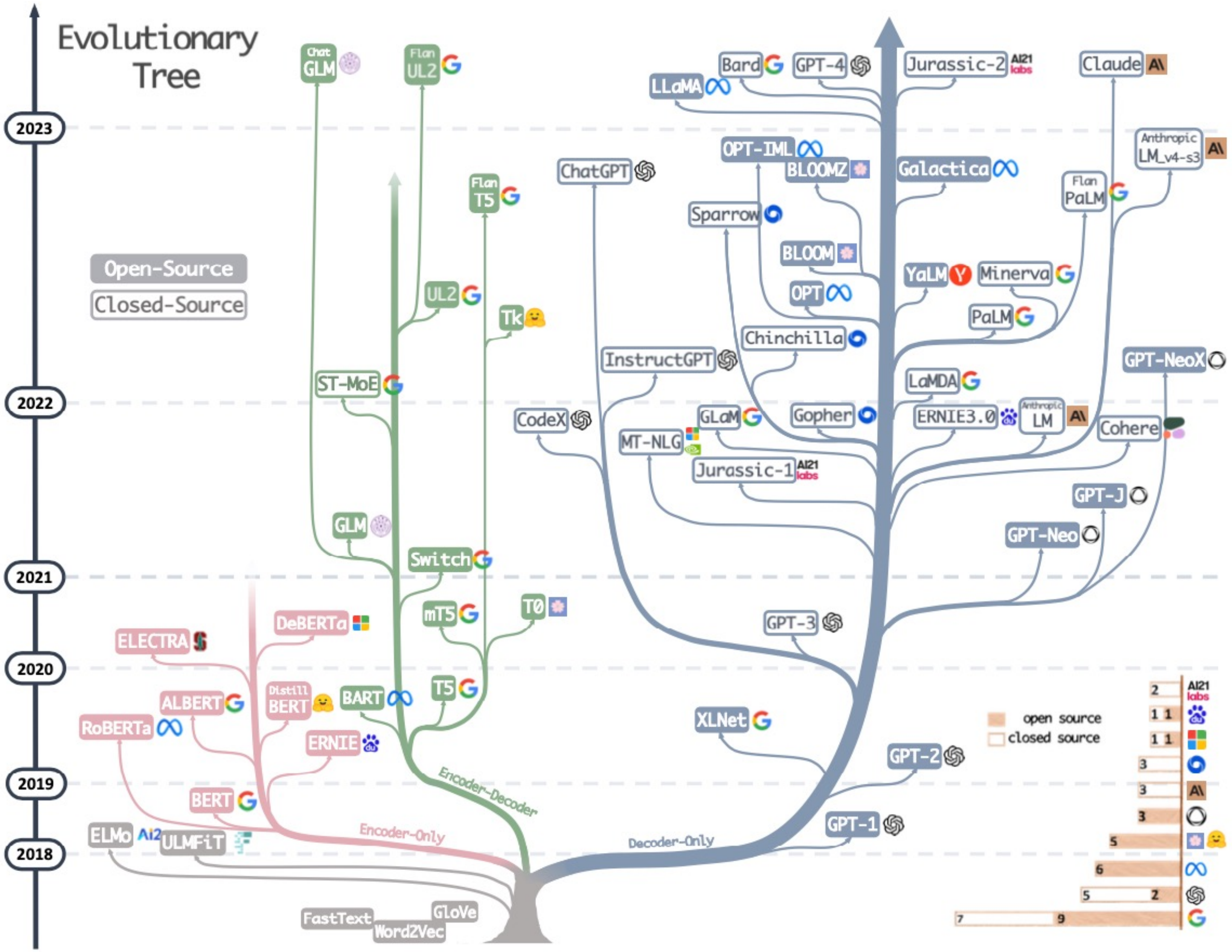
User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.



Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. (2022). "Flamingo: a Visual Language Model for Few-Shot Learning." *Advances in Neural Information Processing Systems 35.*
Junnan Li, Dongxu Li, Silvio Savarese, & Steven Hoi. (2023). "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models." arXiv: 2301.12597.
OpenAI. (2023). "GPT-4 Technical Report." arXiv: 2303.08774.

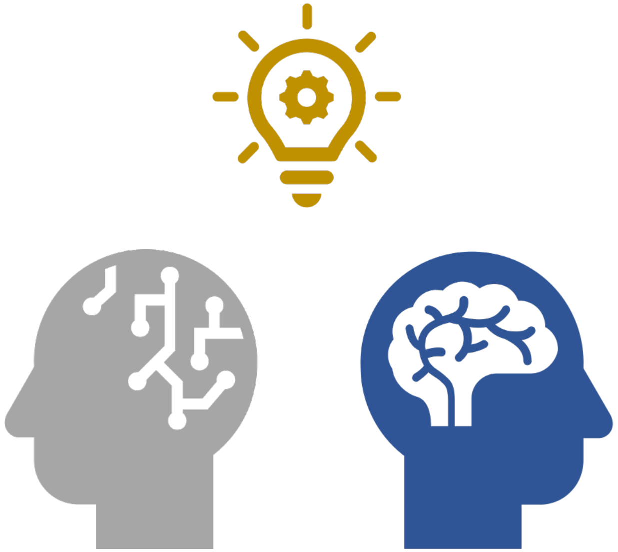


17

[Jingfeng Yang, Hongye Jin, Ruixiang Tang, et al. 2023. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. arXiv: 2304.13712.]

Limitations of LLMs

- Despite these advancements and impressive capabilities, LLMs have some key limitations that cause undesirable behaviors
- In order to effectively and responsibly apply them in research, we need to be mindful of these limitations!



Limitations of LLMs: Spurious Cues

Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

How does the story end?

Karen became good friends with her roommate.	Karen hated her roommate.
	

[Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., & Smith, N.A. (2017). The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In CoNLL 2017.]
[Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P. & Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In NAACL 2016.]

Limitations of LLMs: Data Contamination

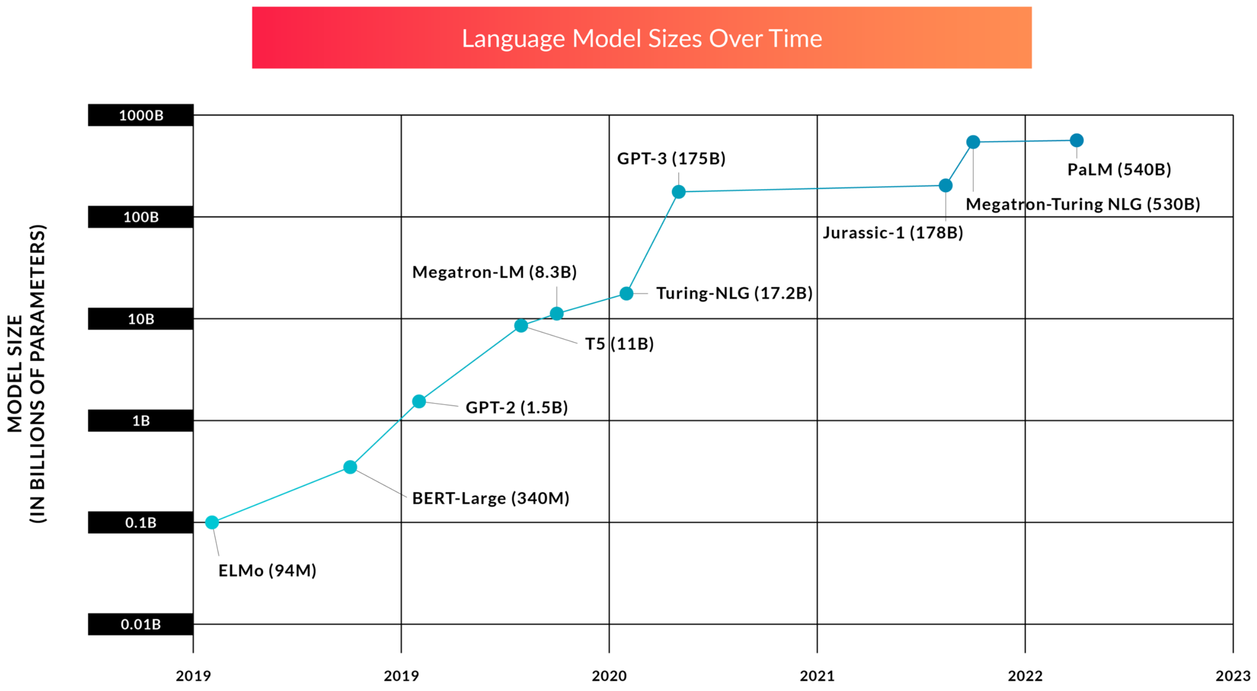
- LLMs have seen so much data in pre-training
- They may have been trained on benchmark datasets…
- Training on the test data is not an objective evaluation!

4 Measuring and Preventing Memorization Of Benchmarks

Since our training dataset is sourced from the internet, it is possible that our model was trained on some of our benchmark test sets. Accurately detecting test contamination from internet-scale datasets is a new area of research without established best practices. While it is common practice to train large models without investigating contamination, given the increasing scale of pretraining datasets, we believe this issue is becoming increasingly important to attend to.

Inbal Magar & Roy Schwartz. (2022). Data Contamination: From Memorization to Exploitation. In *ACL 2022*.
W. Shi, A. Ajith, M. Xia, et al. (2023). Detecting Pretraining Data from Large Language Models. arXiv: 2310.16789.
T.B. Brown, B. Mann, N. Ryder, et al. (2020). Language Models are Few-Shot Learners. arXiv: 2005.14165.

Limitations of LLMs: Interpretability



(figure from Vinay Iyengar)

GPT-4 Technical Report

OpenAI*

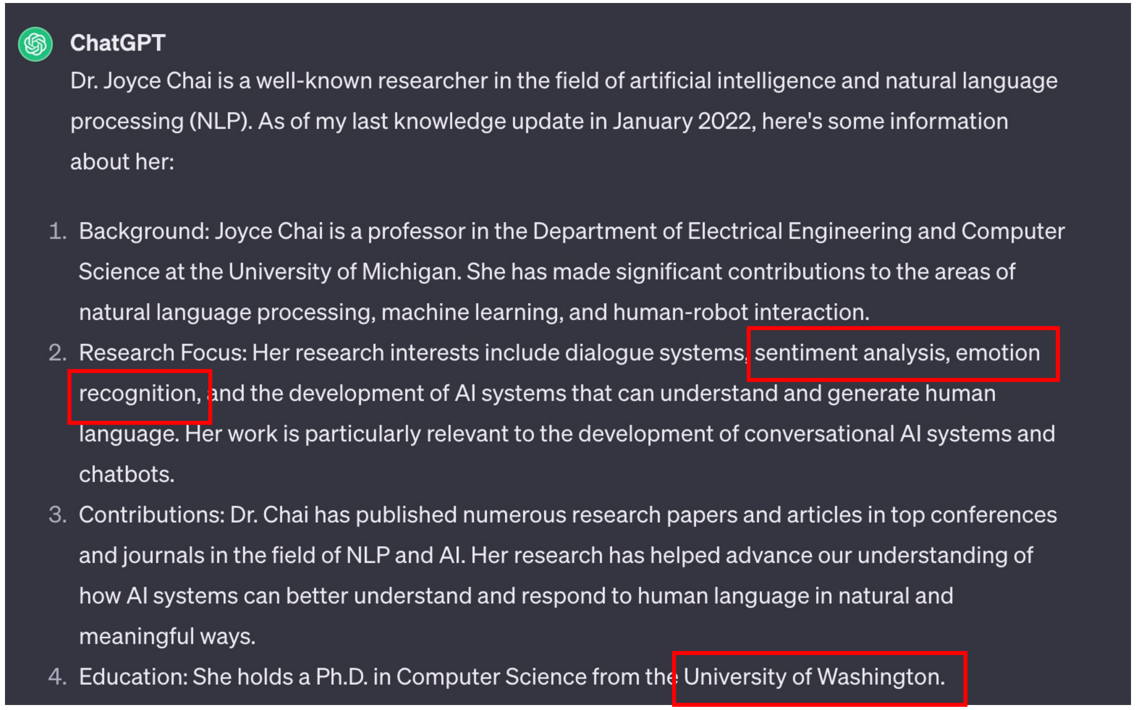
...

2 Scope and Limitations of this Technical Report

This report focuses on the capabilities, limitations, and safety properties of GPT-4. GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [40]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.

Limitations of LLMs: Hallucination

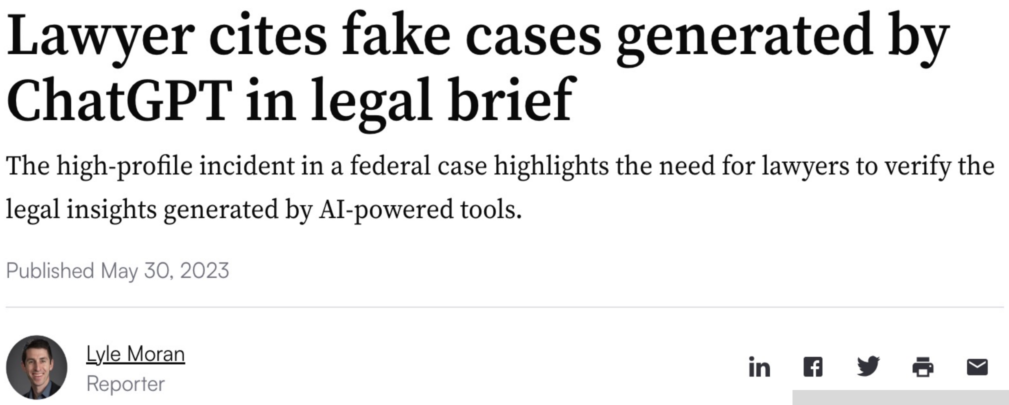
- **Hallucination:** generation of text that is factually incorrect, nonsensical, unfaithful to inputs, or otherwise incoherent



ChatGPT

Dr. Joyce Chai is a well-known researcher in the field of artificial intelligence and natural language processing (NLP). As of my last knowledge update in January 2022, here's some information about her:

1. Background: Joyce Chai is a professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. She has made significant contributions to the areas of natural language processing, machine learning, and human-robot interaction.
2. Research Focus: Her research interests include dialogue systems, sentiment analysis, emotion recognition, and the development of AI systems that can understand and generate human language. Her work is particularly relevant to the development of conversational AI systems and chatbots.
3. Contributions: Dr. Chai has published numerous research papers and articles in top conferences and journals in the field of NLP and AI. Her research has helped advance our understanding of how AI systems can better understand and respond to human language in natural and meaningful ways.
4. Education: She holds a Ph.D. in Computer Science from the University of Washington.



Lawyer cites fake cases generated by ChatGPT in legal brief

The high-profile incident in a federal case highlights the need for lawyers to verify the legal insights generated by AI-powered tools.

Published May 30, 2023

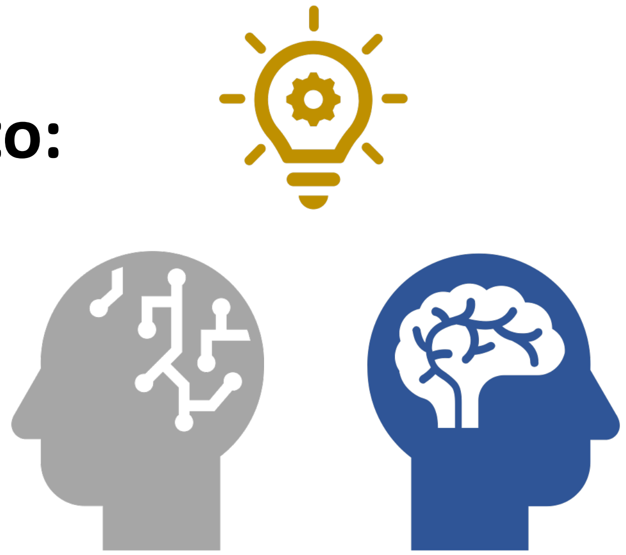
Lyle Moran
Reporter

[Legal Dive](#)

Z. Jio, N. Lee, R. Frieske, et al. (2023). Survey of Hallucination in Natural Language Generation. In *ACM Computing Surveys*, 55.
https://chat.openai.com/

Summary

- LLMs' are remarkably useful for many language tasks, but these limitations make them impossible to trust consistently
- Verifying LLM outputs is important:
  - Automated metrics
  - Human evaluation
- **We must be mindful that LLMs are primarily trained to:**
  - **Generate fluent-sounding language (pre-training)**
  - **Satisfy users' requests (instruction-tuning)**



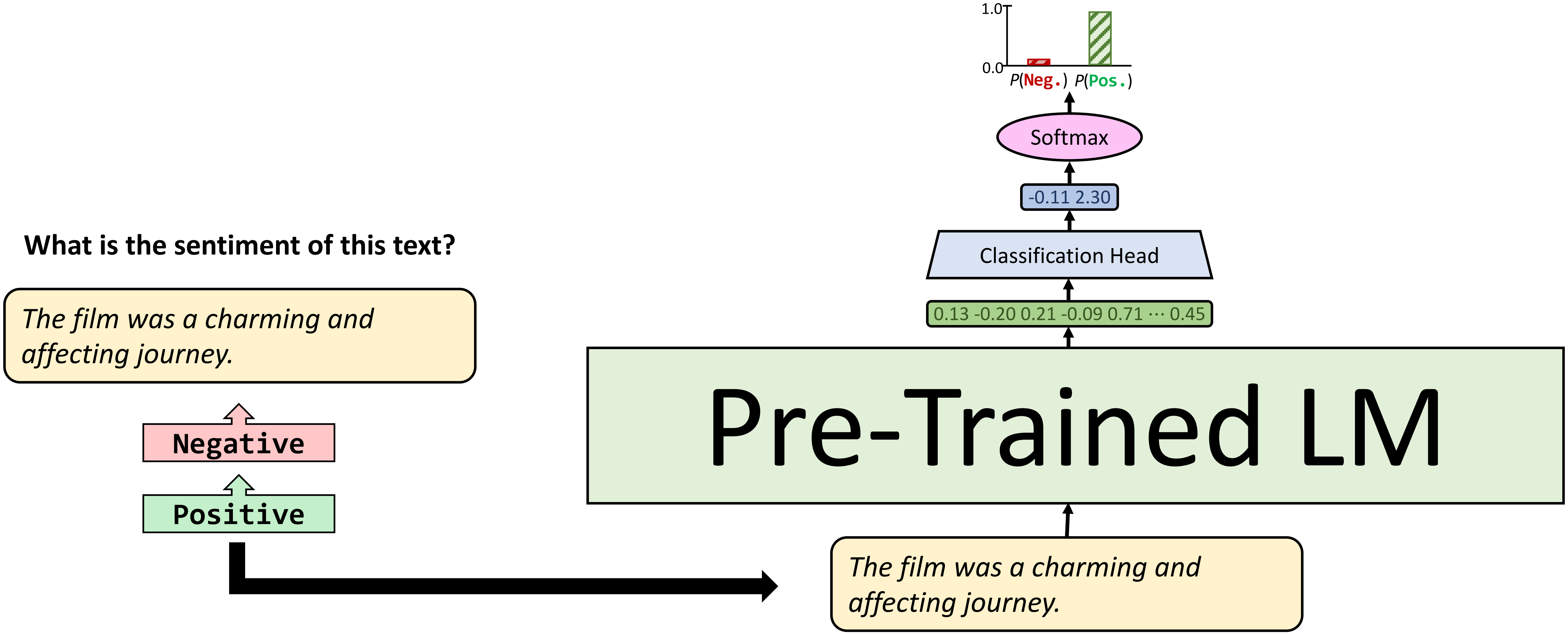
2 Ways to Customize LLMs

Fine-Tuning:	**Prompting:**
Small hardware requirements	Larger hardware requirements
Host locally (private, more flexible)	Best LMs behind proprietary APIs
Optimized for specific task	Requires prompt engineering
Technical skills, engineering effort	User-friendly language interface
Large amount of training data	No training data needed
Hard to adapt once trained	Generalizable and adaptable

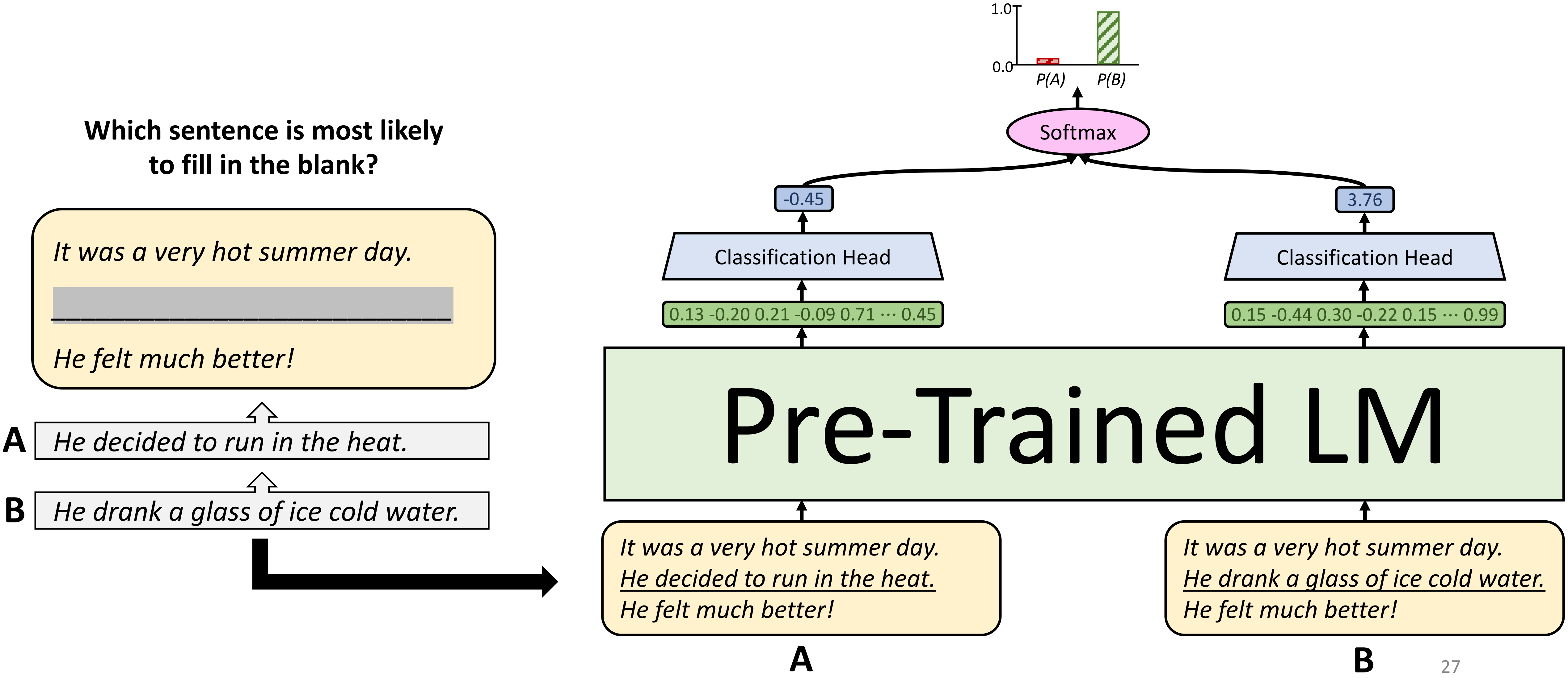
Outline

- The Road to LLMs
- **Fine-Tuning LLMs**
- Prompting LLMs

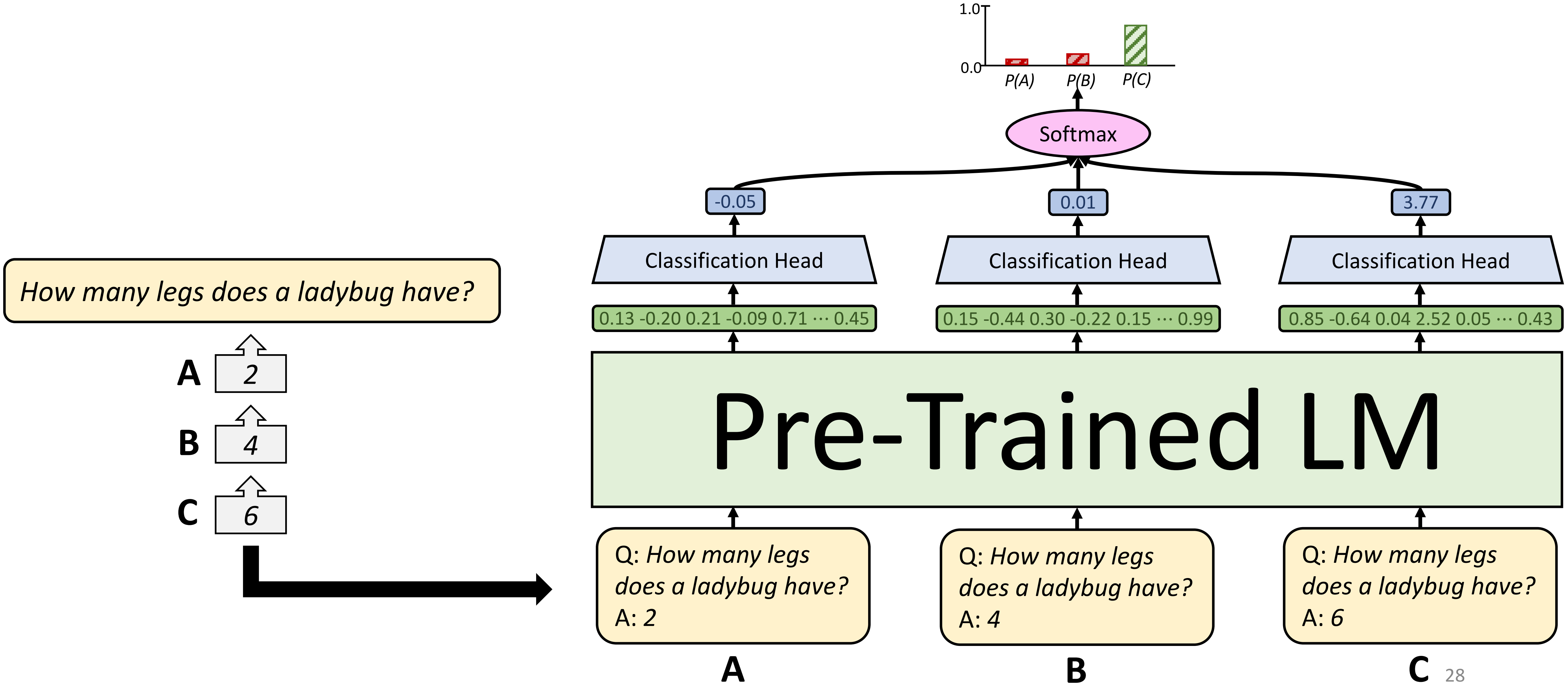
Fine-Tuning: Text Classification



Fine-Tuning: Multiple Choice Completion

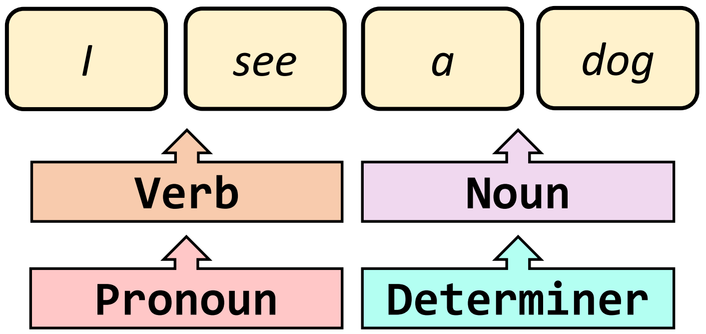


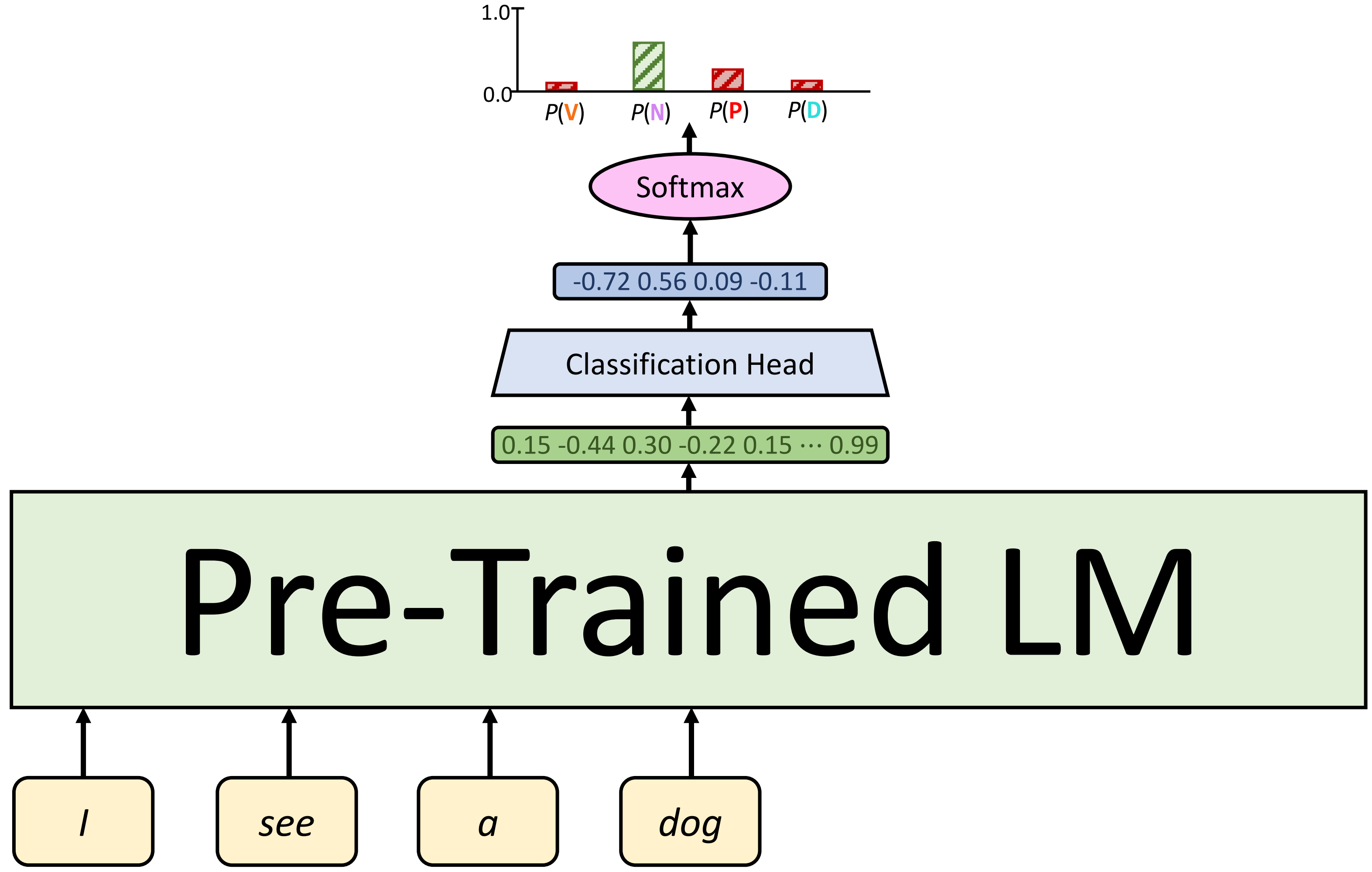
Fine-Tuning: Multiple Choice QA



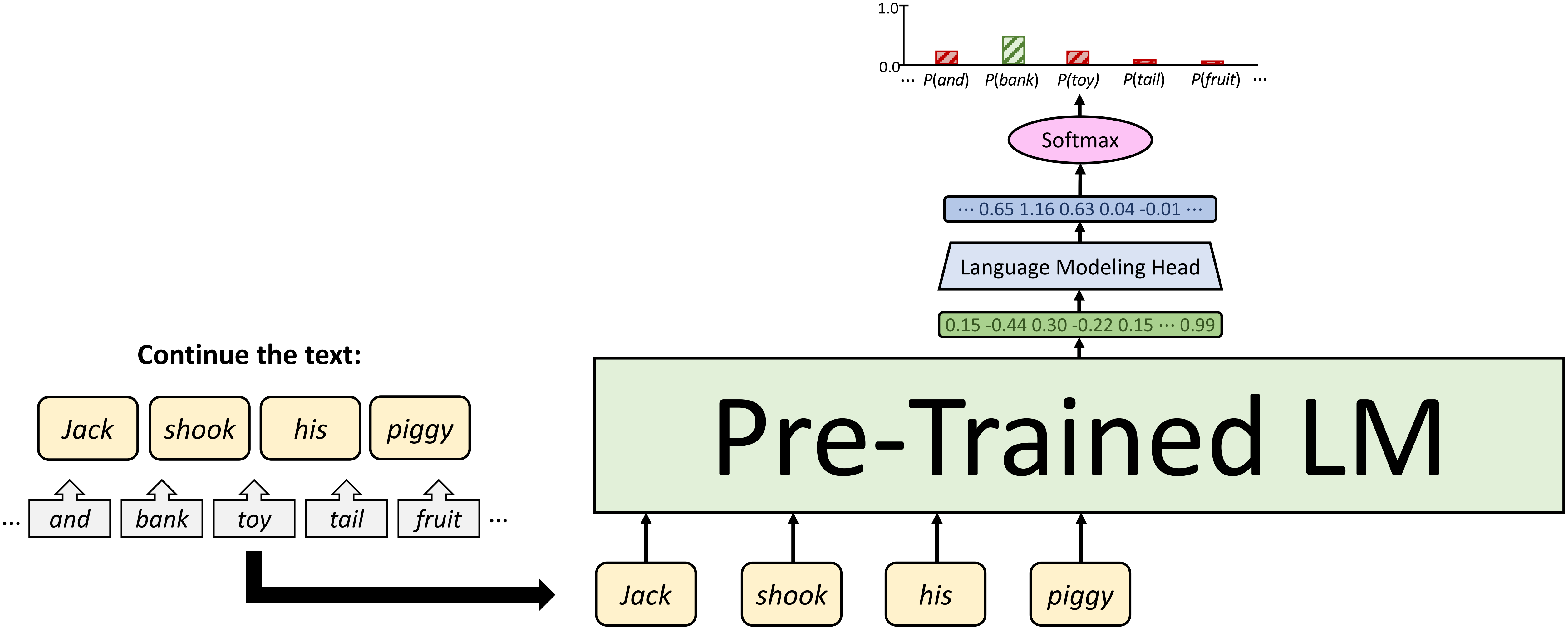
Fine-Tuning: Token Classification

Label each token with its part of speech (POS):





Fine-Tuning: Text Generation

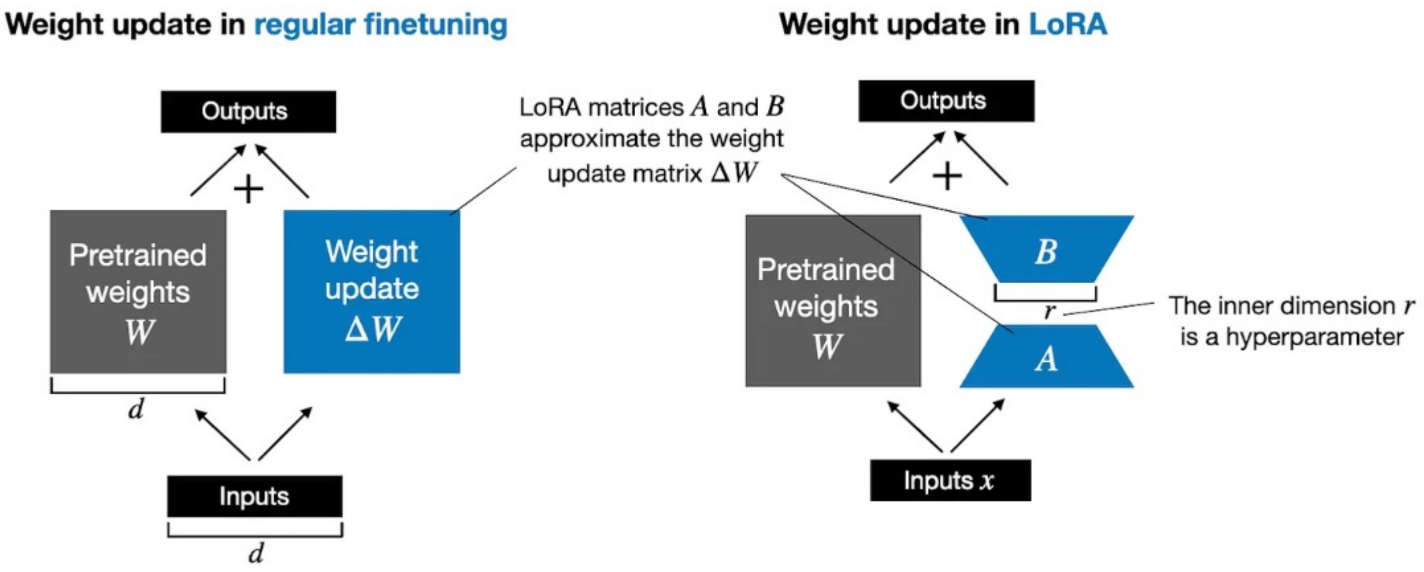


Parameter-Efficient Fine-Tuning (PEFT)

- While fine-tuning LMs is generally more feasible when we have less available compute, there are still some problems:
  - Fine-tuning on a large amount of data can take a long time
  - The size of LM we can fine-tune is limited by compute
  - Updating all weights of the LM during fine-tuning is expensive and inefficient
- Creates a need for **parameter-efficient fine-tuning (PEFT)** methods!

Low-Resource Adaptation (LoRA)

- Instead of updating weights W directly during fine-tuning, learn the weight update ΔW
- Approximate ΔW by a decomposition AB:
  - Reducing the number of learned parameters
  - Faster training with lower GPU memory requirements!
  - Pre-compute ΔW in deployment for fast inference



(figure from Sebastian Raschka)

Edward Hu, Yelong Shen, Phillip Wallis, et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv: 2106.09685.

Outline

- The Road to LLMs
- Fine-Tuning LLMs
- **Prompting LLMs**

Prompting LMs

To customize an LLM for your problem through prompting, need to make a few choices (**prompt engineering**):

1. Prompt template
2. Answer mapping
3. In-context demonstration

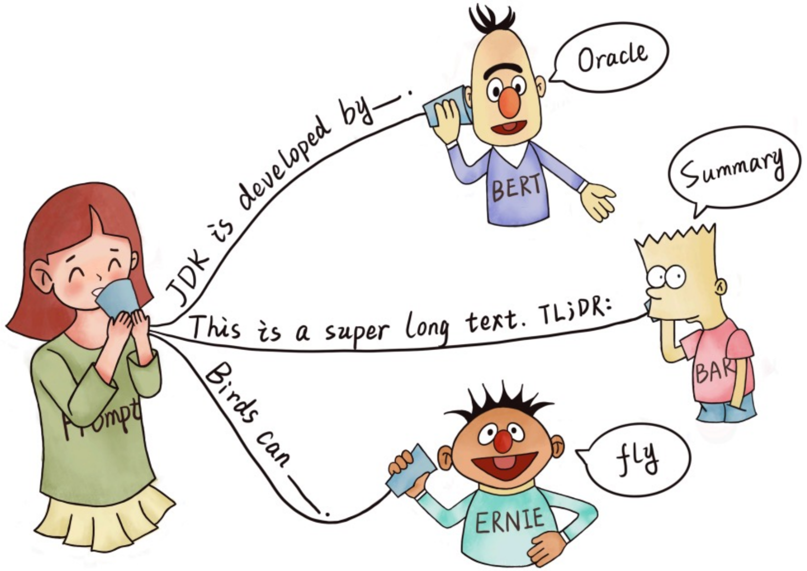
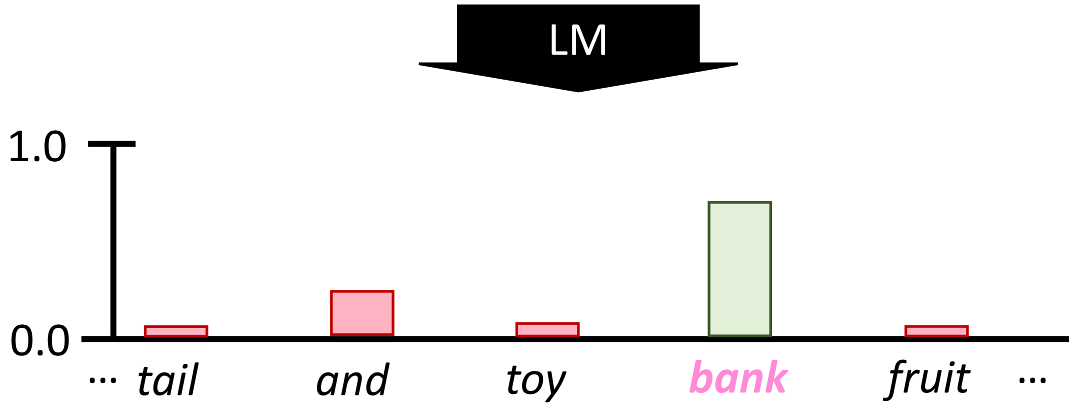


Figure credit: P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, & G. Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv: 2107.13586.

Language Models (LMs)

$$p(w_n | w_1, w_2, \dots, w_{n-1})$$

Jack needed some ***money****, so he went and shook his* ***piggy*** ____



$p($____$|$*Jack needed … shook his piggy*$)$

(dreamstime)

$p(\textit{Jack needed …shook his piggy bank}) \approx p(\textit{bank}|\textit{Jack needed…shook his piggy}) \cdot p(\textit{piggy}|\textit{Jack needed…shook his}) \cdot p(\textit{his}|\textit{Jack needed…shook}\ ) \cdot \cdots$

Minsky, M. (2000). Commonsense-based interfaces. In *Commun. ACM*, 43(8): p. 66-73.

Prompt Templates

If filling a blank from a few possible choices, can use a **cloze prompt**:

Task	Inputs ([X])	Template	Answer ([Z])

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, & G. Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv: 2107.13586.
Sheng Zhang, Xiaodong Liu, Jingjing Liu, et al. 2018. ReCoRD: Bridging the Gap Between Human and Machine Commonsense Reading Comprehension. arXiv: 1810.12885.

Prompt Templates

When completing a prompt or generating text, use a **prefix prompt**:

Task	**Inputs ([X])**	**Template**	**Answer ([Z])**

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, & G. Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv: 2107.13586.

Prompt Templates

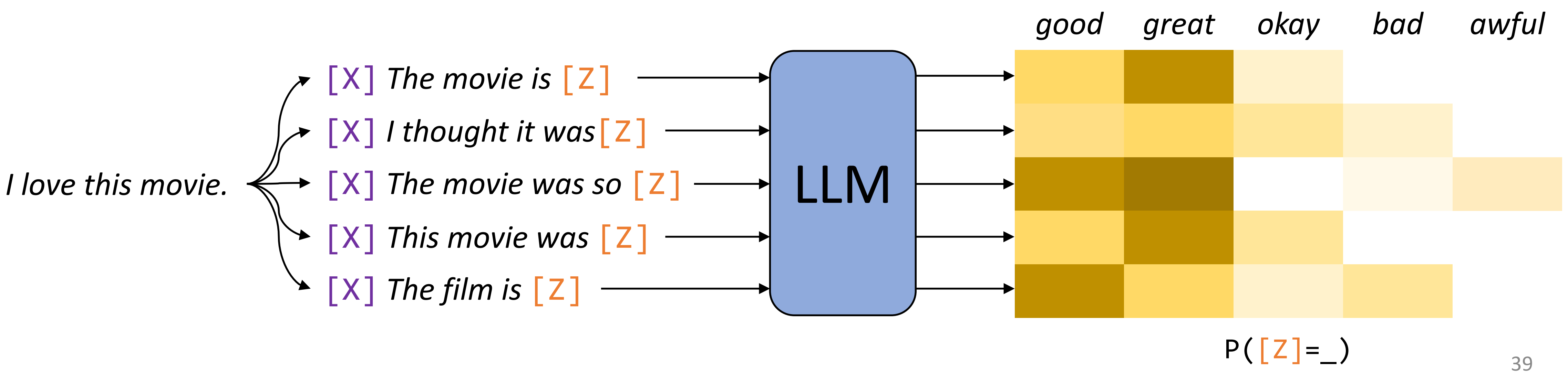
When completing a prompt or generating text, use a **prefix prompt**:

Task	Inputs ([X])	Template	Answer ([Z])

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, & G. Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv: 2107.13586.

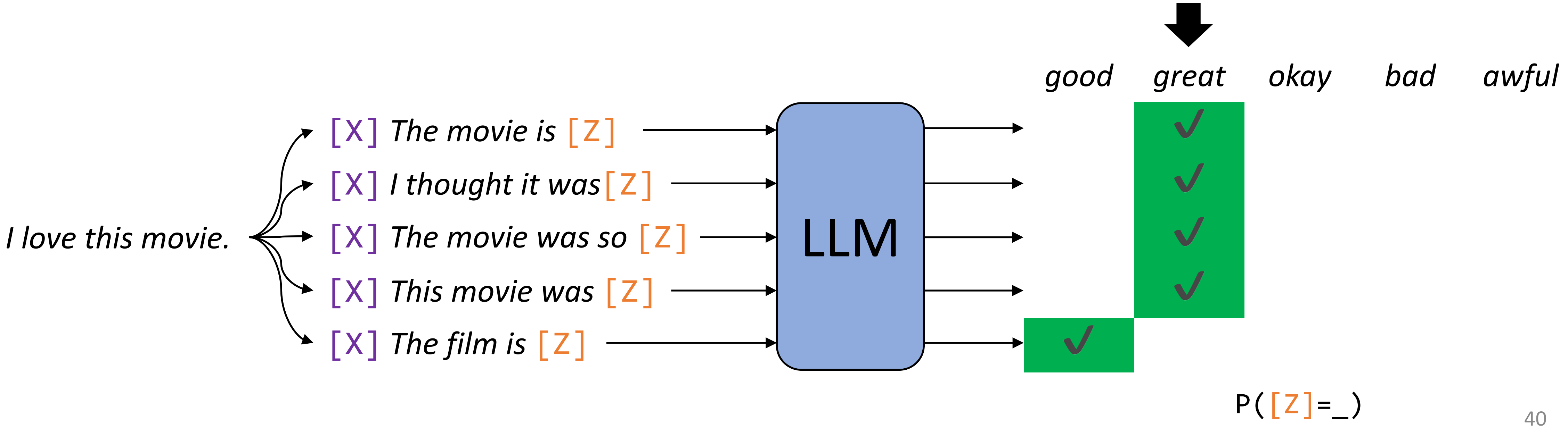
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
  - Trial and error
  - Ensembling templates



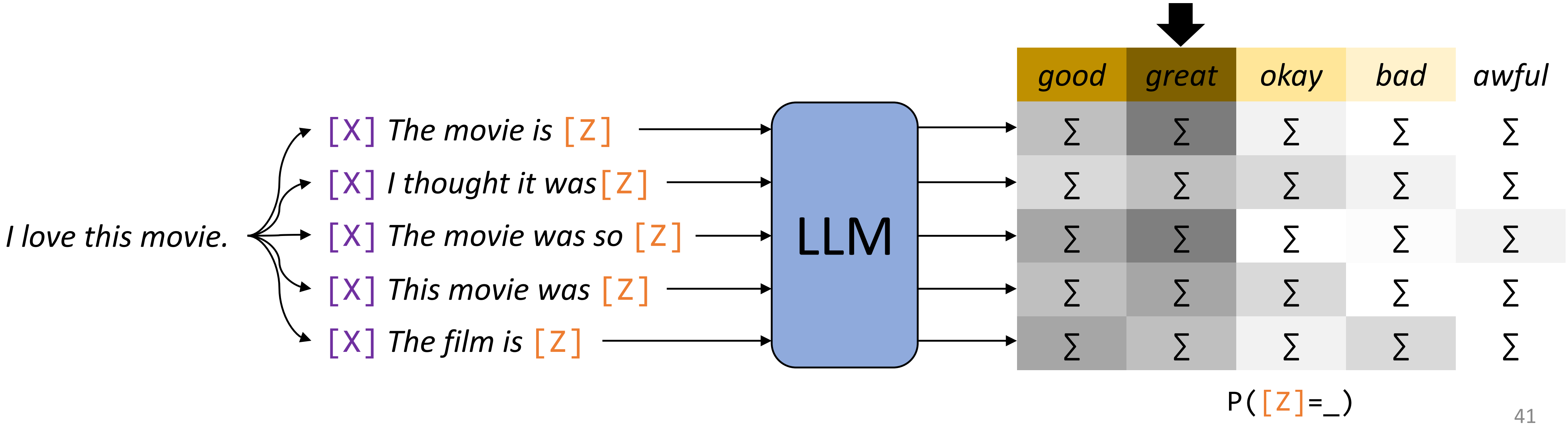
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
  - Trial and error
  - Ensembling templates



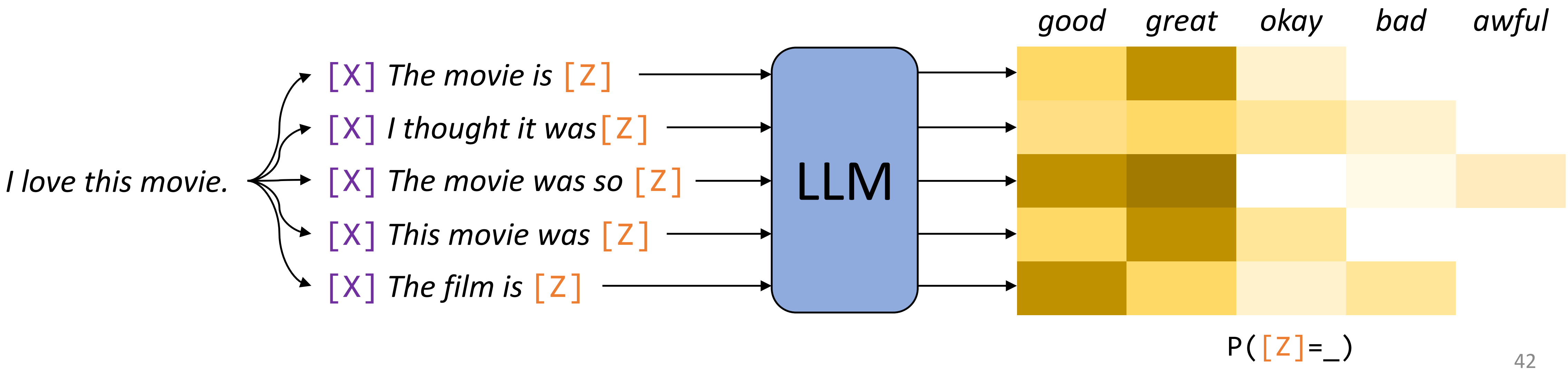
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
  - Trial and error
  - Ensembling templates



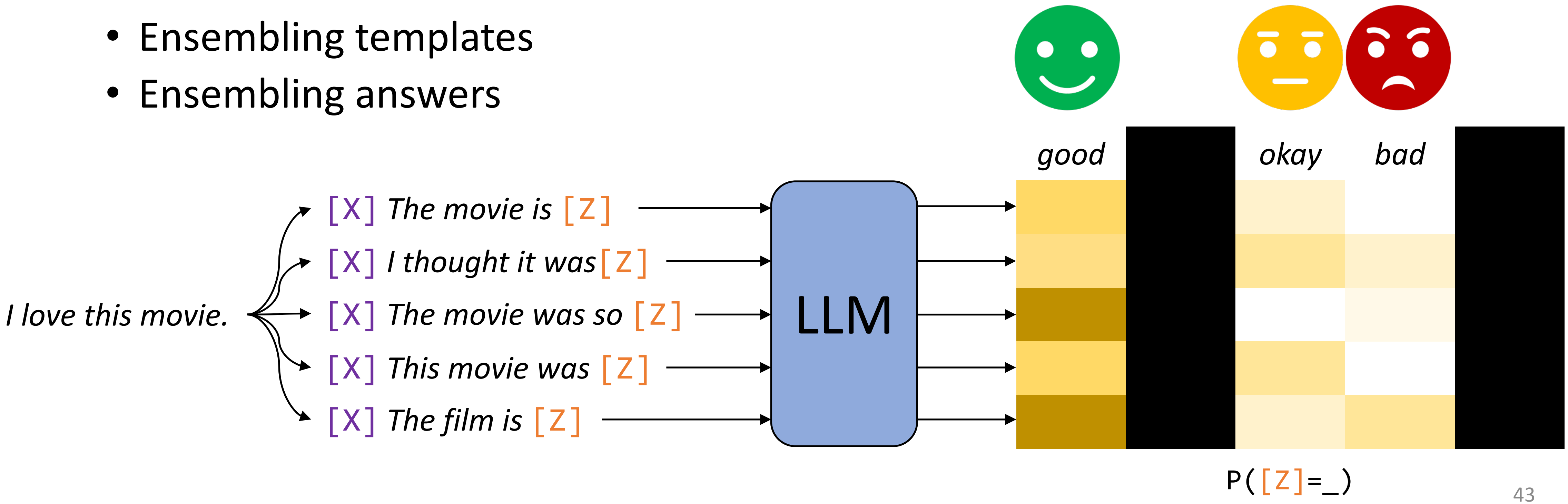
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
  - Trial and error
  - Ensembling templates
  - Ensembling answers



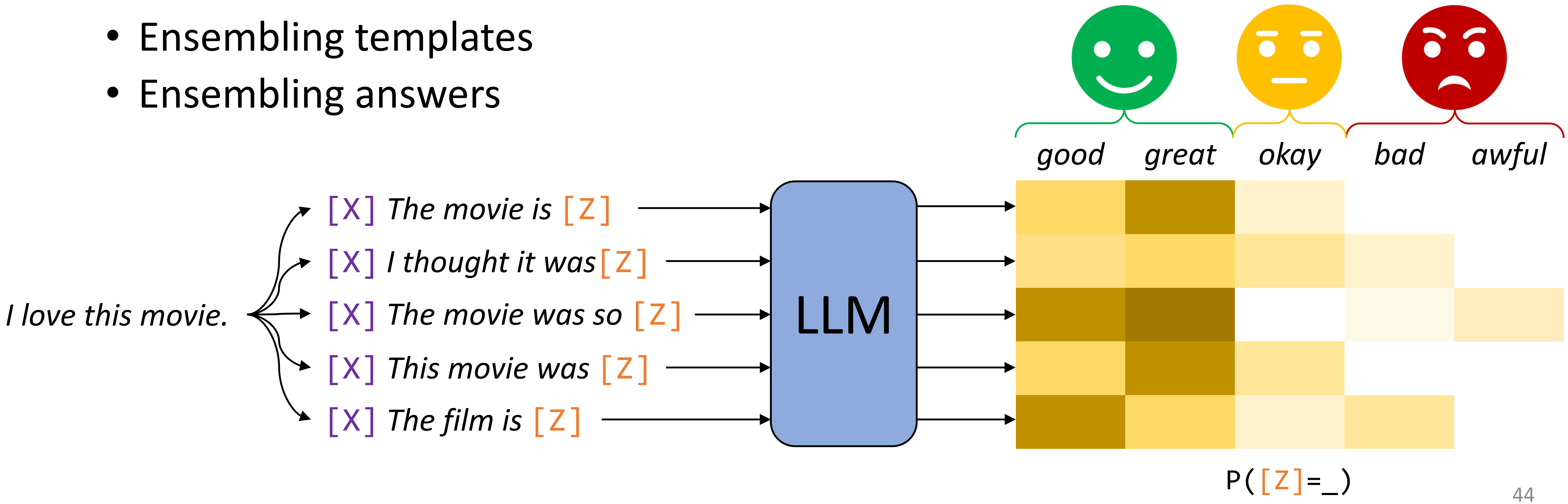
Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
  - Trial and error
  - Ensembling templates
  - Ensembling answers



Finding the Best Template and Answers

- Different prompts can yield different results
- May take extra work to find the best prompt
  - Trial and error
  - Ensembling templates
  - Ensembling answers



Managing Randomness in LLMs

- LLM decoding algorithms may incorporate some randomness by default to increase the diversity of generation
- Some solutions:
  - Generate multiple times and average results
  - Greedy decoding

In-Context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
cheese => ....................    ← prompt
```

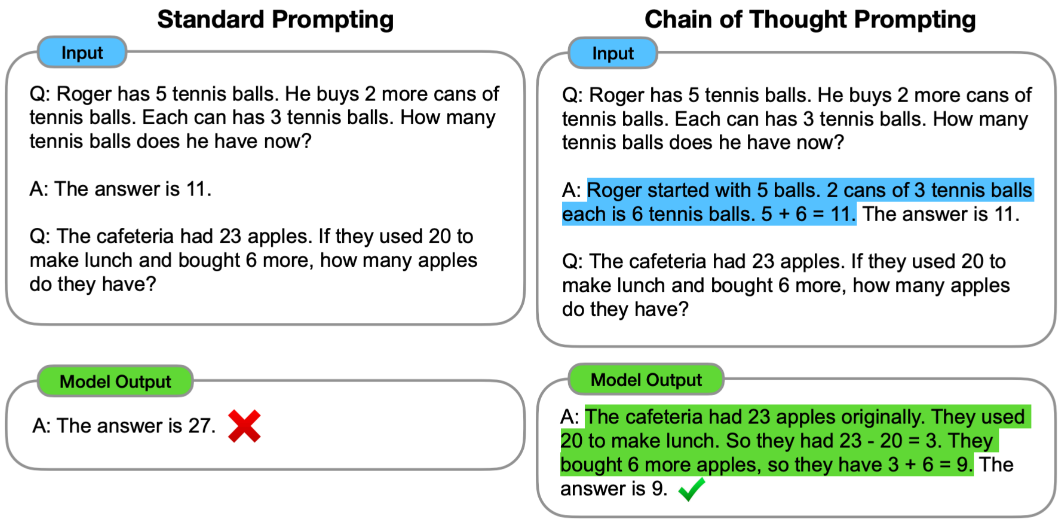
Few-shot

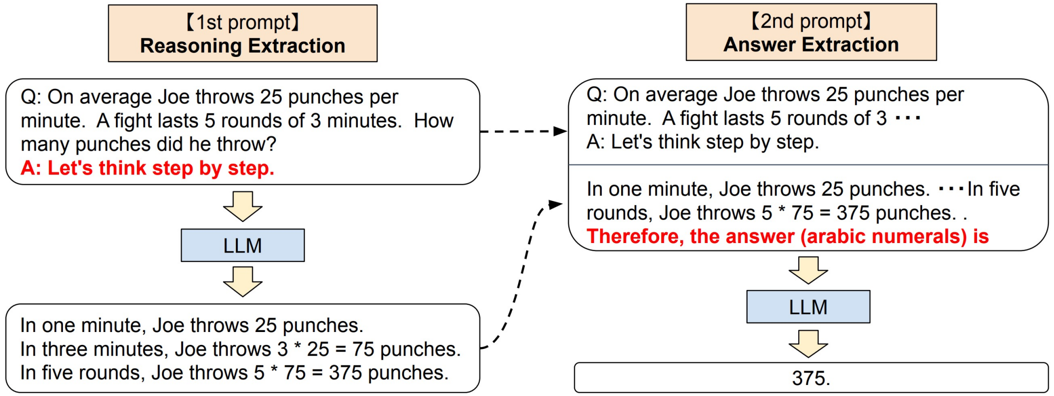
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:          ← task description
sea otter => loutre de mer            ← examples
peppermint => menthe poivrée          ←
plush girafe => girafe peluche        ←
cheese => ....................        ← prompt
```


Tom B. Brown, Benjamin Mann, Nick Ryder, et al. (2020). “Language Models are Few-Shot Learners.” arXiv: 2005.14165.

Chain-of-Thought Prompting





Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems 35.*
Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems 35.*

Next: ***From Theory to Practice!***

I’m on the job market for academic and industry positions!

@shanestorks

www.shanestorks.com