A photograph of a person with long hair sitting at a desk, viewed from behind. On the desk is a laptop, a white coffee cup on a saucer, a pair of glasses, and a small gold-colored object. The desk is made of light-colored wood and is positioned in front of a window with a view of a building across a body of water. The lighting suggests it's either sunrise or sunset.

JM2050 – Natural Language Processing

Working with prompts and generative LLMs

October 17th, 2024 Prof.dr.ir. U. Kaymak

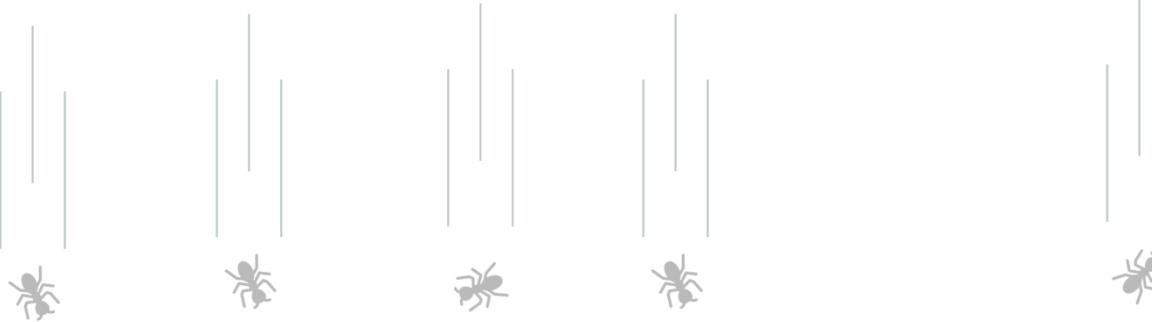
Based on slides from E. Rijcken

1

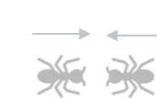
Can Large Language Models reason?

2





Suppose twenty ants are dropped at random locations on a 1-meter plank and begin walking in arbitrary directions at 1 meter per minute. If they reverse direction upon meeting and cannot pass each other, how long will it take at most for all ants to leave the plank?



3



JADS
Jheronimus
Academy
of Data Science

GPT 3.5



The maximum time it will take for all ants to leave the plank is 59 minutes.

Here's the reasoning:

1. When two ants meet, they each turn 180 degrees. This means they effectively swap positions.
2. As long as there are any ants facing each other, they will continue to meet and swap positions.
3. The worst-case scenario is that all ants initially face each other and form pairs that keep swapping positions.
4. After 1 minute, the ants that started at each end of the plank will have met in the middle and swapped positions. After 2 minutes, they will have returned to their original starting positions.
5. This pattern will repeat every 2 minutes, so after 59 minutes, they will have swapped positions 29 times, and no ants will be facing each other. All ants will have left the plank.

So, it takes at most 59 minutes for all ants to leave the plank in the worst-case scenario.

GPT 4



To determine the maximum time it will take for all ants to leave the plank, consider the following:

1. The worst-case scenario for an individual ant would be if it starts at one end of the plank and walks the entire length of 1 meter without any interruptions. This would take the ant exactly 1 minute.
2. However, the ants can bump into each other and change directions. But this doesn't increase the time for any individual ant. To see why, imagine labeling each ant with a unique number and then ignoring the fact that they reverse directions when they meet. Instead, pretend that when two ants meet, they pass right through each other and continue walking. The time it takes for any "labeled" ant to fall off the plank in this imaginary scenario will be the same as in the real scenario. That's because each ant's speed and total distance traveled remain the same.

So, regardless of how often ants bump into each other or where they start, the maximum time it takes for any ant to leave the plank is 1 minute.

Therefore, the answer to your question is 1 minute. After 1 minute, all the ants will have left the plank.

What is Reasoning?

Reasoning: the process of thinking about things in a logical way; opinions and ideas that are based on logical thinking (Oxford Learner's Dictionary)

Does the ant example demonstrate reasoning skills? How do we know the model did not just learned this problem from the web? How can one test reasoning skills?

20 questions game

A game that requires a process of logical deductive reasoning

Rules

- **Thinker:** One player selects a secret object (typically "animal, vegetable, or mineral").
- **Questions:** Guesser asks up to 20 yes-or-no questions.
- **Answers:** Thinker responds truthfully to each question.
- **Guessing:** Players can guess the object, but it counts as one question.

Winning

- Correct guess within 20 questions: guesser wins.
- No correct guess after 20 questions: thinker wins.

The screenshot shows a conversational interface between a user and a AI agent. The user's messages are on the left, and the AI's responses are on the right. The AI is playing a 20-question game to guess an object.

User message:

Lets play 20 questions. I want you to think of an object, and its my role to guess what you are thinking about. Ready?

AI response:

Sure, I'm ready! I've "thought" of an object. Go ahead and start asking your questions! Remember, I'll answer with "yes," "no," or "sometimes/it depends."

User message:

Is it alive?

AI response:

No.

User message:

Is it created by humans?

AI response:

Yes.

User message:

Is it typically used indoors?

AI response:

Yes.

User message:

Is it typically used in houses?

AI response:

Yes.

User message:

Does it use electricity?

AI response:

Yes.

User message:

Continue this conversation

AI response:

Is it typically used for decoration?

AI response:

No.

AI response:

Is it used in the kitchen?

AI response:

Yes.

AI response:

Is it used to heat or cool things?

AI response:

Yes.

AI response:

Is it used to heat things?

AI response:

Yes.

AI response:

Is it an oven?

AI response:

Yes! Well done! It is an oven. That was impressive detective work. Would you like to play again or is there something else I can assist you with?

JADS Jheronimus Academy of Data Science logo

It works the other way too



How did you come up with oven?



When playing the "20 questions" game, I don't actually "think" or "imagine" objects the way humans do. Instead, when you expressed interest in playing the game, I selected a general category or type of object based on the provided context. As you started asking questions, I used the logic and information built into my programming to provide accurate answers based on the category I "selected."

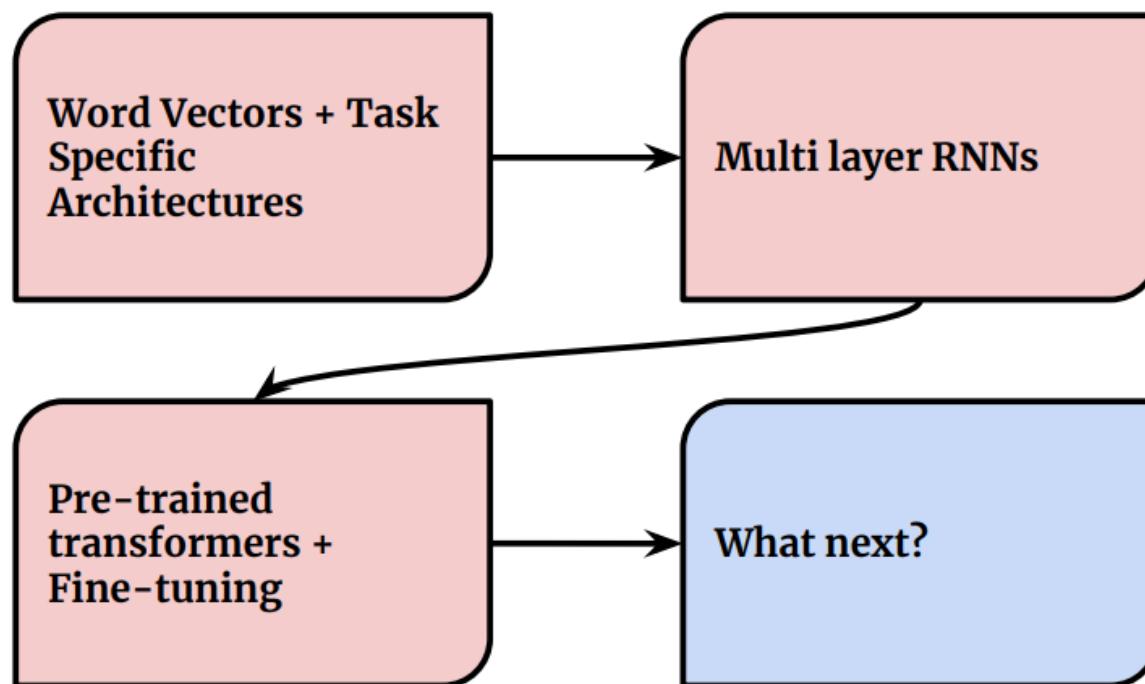
When you chose to play and the topic was an object, I essentially designated a data point or concept from my vast database. The oven was that concept in this case. Your questions helped narrow down to that specific answer.

How did we get here?

9



Shifting paradigms in NLP



10

Source: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>

Often, it is often impractical to fine-tune a very large model (1/3)

Practical Issues

- Need large task-specific datasets for fine-tuning
- Collect data for task A → Fine-tune to solve task A → Repeat for task B
→ Repeat for task C → and so on ...
- End up with many “copies” of the same model .

11

Source: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>



Often, it is often impractical to fine-tune a very large model (2/3)

Potential to exploit spurious correlations (Overfitting)

- Large models fine-tuned on very narrow task distributions
- Evidence suggests: models overfit to training distributions and don't generalize well outside of it (Evidence: [Hendricks et al. 2020](#), [Yogatama et al. 2019](#), [McCoy et al. 2019](#))
- Models are good on datasets, not so good at the underlying task ([Gururangan et al. 2018](#), [Niven et al. 2019](#))

12

Source: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>



Often, it is often impractical to fine-tune a very large model (3/3)

Humans don't need large supervised datasets

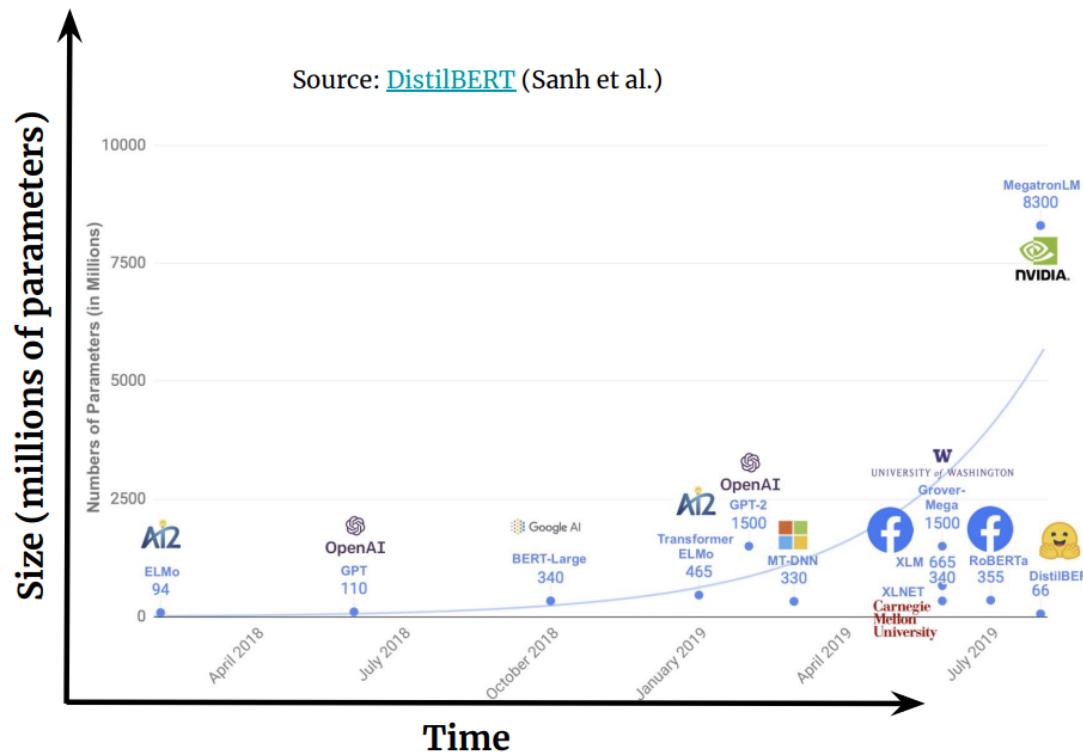
- Humans can learn from simple directives
- Allows humans to mix and match skills + switch between tasks easily
- Hope is for NLP systems to one day function with the same fluidity!

13

Source: <https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec04.pdf>



LM Landscape pre GPT-3

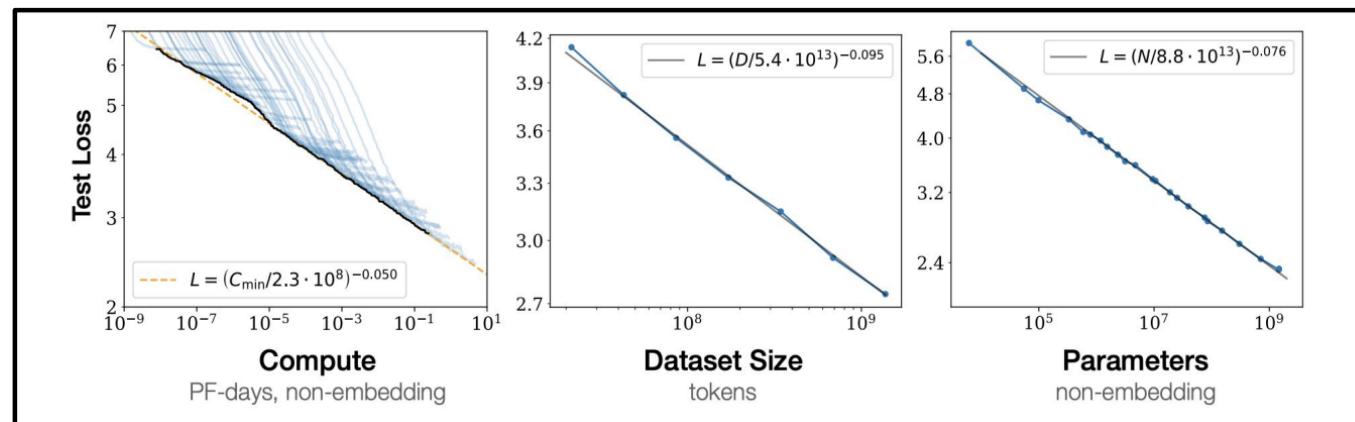


14

JADS
Jheronimus
Academy
of Data Science

Scaling Laws for Neural Language Models – Kaplan et al. 2020

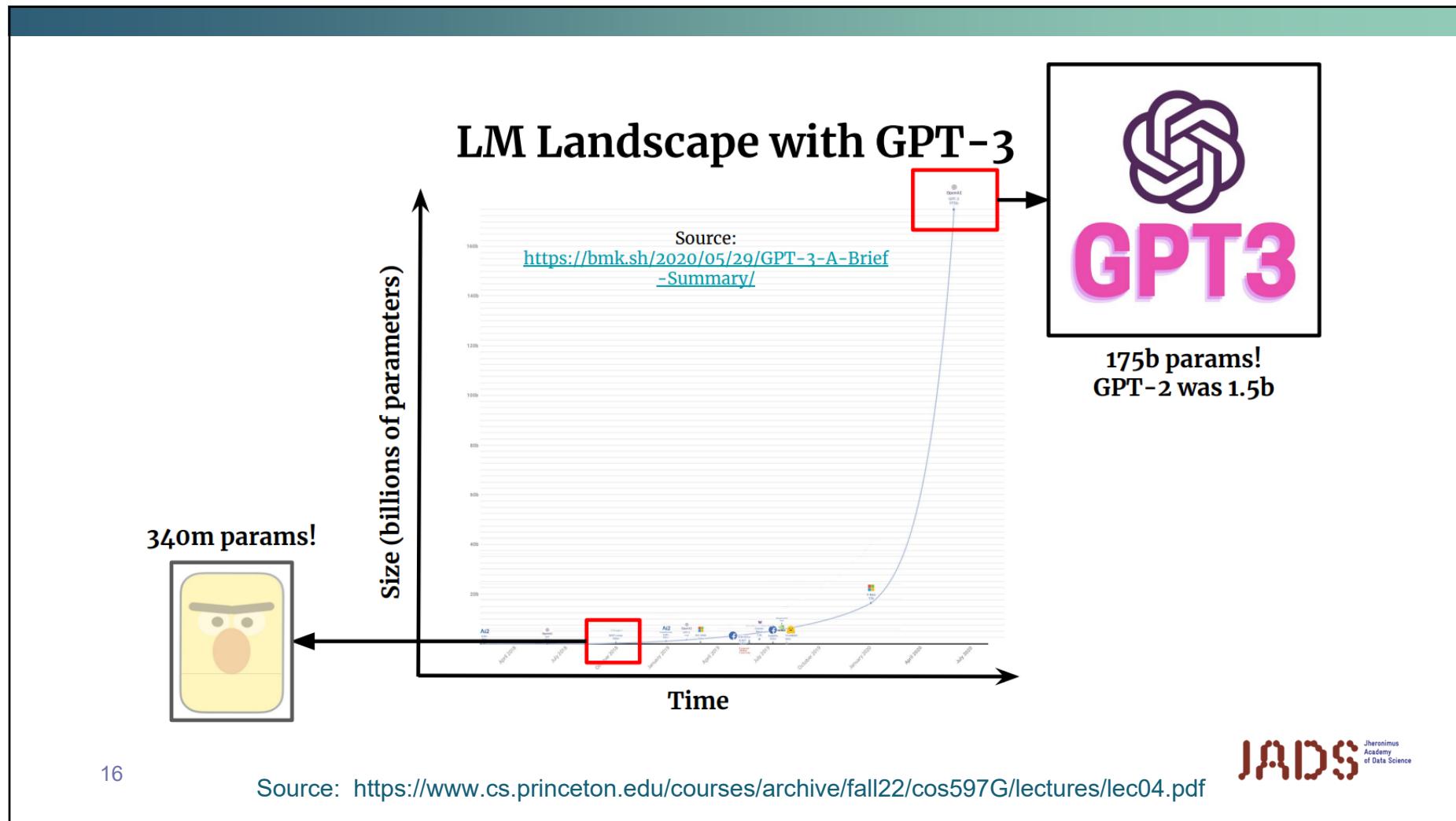
- Loss scales as a power-law with model size, dataset size and the amount of compute used for training.
- Other architectural details, such as network width/depth have minimal effects.
- Larger models are more sample efficient: they learn more effectively from a smaller amount of data)



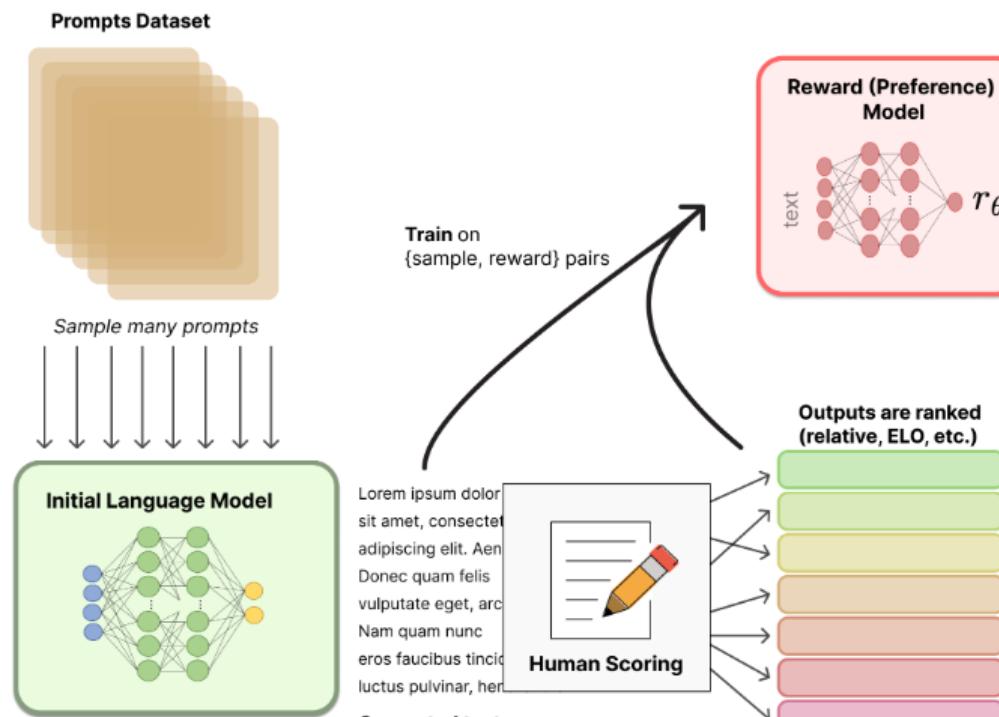
15

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.





ChatGPT uses Reinforcement Learning from Human Feedback for fine-tuning



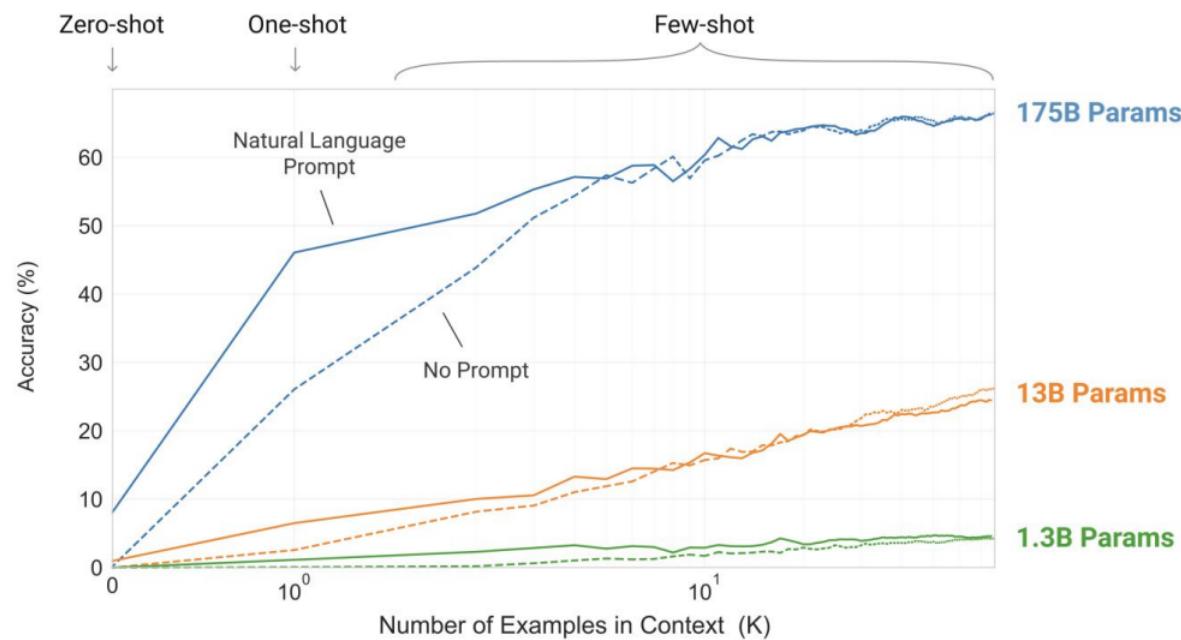
Source:
<https://huggingface.co/blog/rhf>

17

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

JADS
Jheronimus
Academy
of Data Science

Larger models learn better In-Context



18

Source: Brown, Tom, Mann, B. et al. "Language Models are Few-Shot Learners." In Larochelle H. et al. (eds.), Advances in Neural Information Processing Systems, Vol. 33, p. 1877 – 1901, 2020.

Prompts and prompt engineering

- **Prompts** involve instructions and context passed to a language model to achieve a desired task
- **Prompt engineering** is the practice of developing and optimizing prompts to efficiently use language models (LMs) for a variety of applications
 - Prompt engineering is a useful skill for AI engineers and researchers to improve and efficiently use language models

What is prompt engineering?

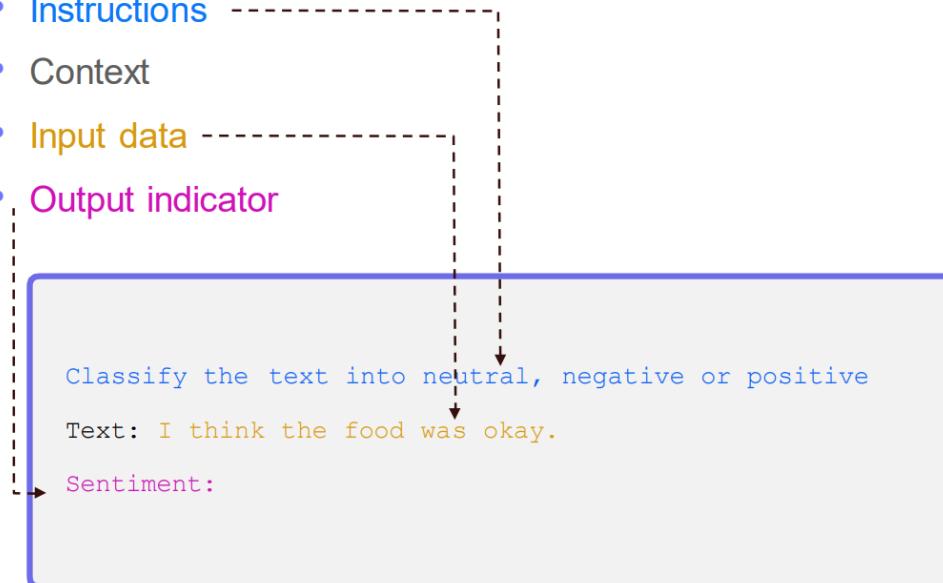
Prompt engineering is a process of creating a set of prompts, or questions, that are used to guide the user toward a desired outcome. It is an effective tool for designers to create user experiences that are easy to use and intuitive. This method is often used in interactive design and software development, as it allows users to easily understand how to interact with a system or product..

19

Source: <https://www.cse.iitd.ac.in/~mausam/courses/col772/spring2023/lectures/22-promptengg.pdf>

Elements of a prompt

- A prompt is composed with the following components:
 - Instructions
 - Context
 - Input data
 - Output indicator



Zero-shot vs. Few-shot learning

Zero-shot:

Simply feed the task text to the model and ask for results.

Few-shot learning:

Presents a set of high-quality demonstrations, each consisting of both input and desired output, on the target task.

Zero-shot

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



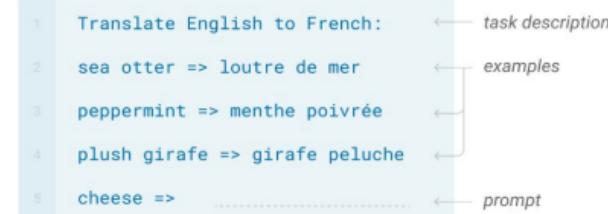
One-shot

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



Few-shot

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



Brown et al., 2020

Few-shot learning considerations

Pro:

As the model first sees good examples, it can better understand human intention and criteria for what kinds of answers are wanted.

Con:

However, it comes at the cost of more token consumption and may hit the context length limit when input and output text are long.



Prompt Engineering

JADS
Jheronimus
Academy
of Data Science

Six strategies for getting better results

1. Include details in your query to get more relevant answers
2. Ask the model to adopt a persona
3. Use delimiters to clearly indicate distinct parts of the input
4. Specify the steps required to complete a task
5. Provide examples
6. Specify the desired length of the output

1. Include details in your query to get more relevant answers

In order to get a highly relevant response, make sure that requests provide any important details or context. Otherwise you are leaving it up to the model to guess what you mean.

Worse	Better
How do I add numbers in Excel?	How do I add up a row of dollar amounts in Excel? I want to do this automatically for a whole sheet of rows with all the totals ending up on the right in a column called "Total".
Who's president?	Who was the president of Mexico in 2021, and how frequently are elections held?
Write code to calculate the Fibonacci sequence.	Write a TypeScript function to efficiently calculate the Fibonacci sequence. Comment the code liberally to explain what each piece does and why it's written that way.
Summarize the meeting notes.	Summarize the meeting notes in a single paragraph. Then write a markdown list of the speakers and each of their key points. Finally, list the next steps or action items suggested by the speakers, if any.

2. Ask the model to adopt a persona

Tactic: Ask the model to adopt a persona

The system message can be used to specify the persona used by the model in its replies.

SYSTEM When I ask for help to write something, you will reply with a document that contains at least one joke or playful comment in every paragraph.

USER Write a thank you note to my steel bolt vendor for getting the delivery in on time and in short notice. This made it possible for us to deliver an important order.

3. Use delimiters to clearly indicate distinct parts of the input

Delimiters like triple quotation marks, XML tags, section titles, etc. can help demarcate sections of text to be treated differently.

USER Summarize the text delimited by triple quotes with a haiku.

"""insert text here"""

[Open in Playground ↗](#)

SYSTEM You will be provided with a pair of articles (delimited with XML tags) about the same topic. First summarize the arguments of each article. Then indicate which of them makes a better argument and explain why.

USER <article> insert first article here </article>

<article> insert second article here </article>

[Open in Playground ↗](#)

4. Specify the steps required to complete a task

Some tasks are best specified as a sequence of steps. Writing the steps out explicitly can make it easier for the model to follow them.

SYSTEM Use the following step-by-step instructions to respond to user inputs.

Step 1 - The user will provide you with text in triple quotes. Summarize this text in one sentence with a prefix that says "Summary: ".

Step 2 - Translate the summary from Step 1 into Spanish, with a prefix that says "Translation: ".

USER """"Insert text here"""

5. Provide examples

Providing general instructions that apply to all examples is generally more efficient than demonstrating all permutations of a task by example, but in some cases providing examples may be easier. For example, if you intend for the model to copy a particular style of responding to user queries which is difficult to describe explicitly. This is known as "few-shot" prompting.

SYSTEM Answer in a consistent style.

USER Teach me about patience.

ASSISTANT The river that carves the deepest valley flows from a modest spring; the grandest symphony originates from a single note; the most intricate tapestry begins with a solitary thread.

USER Teach me about the ocean.

6. Specify the desired length of the output

You can ask the model to produce outputs that are of a given target length. The targeted output length can be specified in terms of the count of words, sentences, paragraphs, bullet points, etc. Note however that instructing the model to generate a specific number of words does not work with high precision. The model can more reliably generate outputs with a specific number of paragraphs or bullet points.

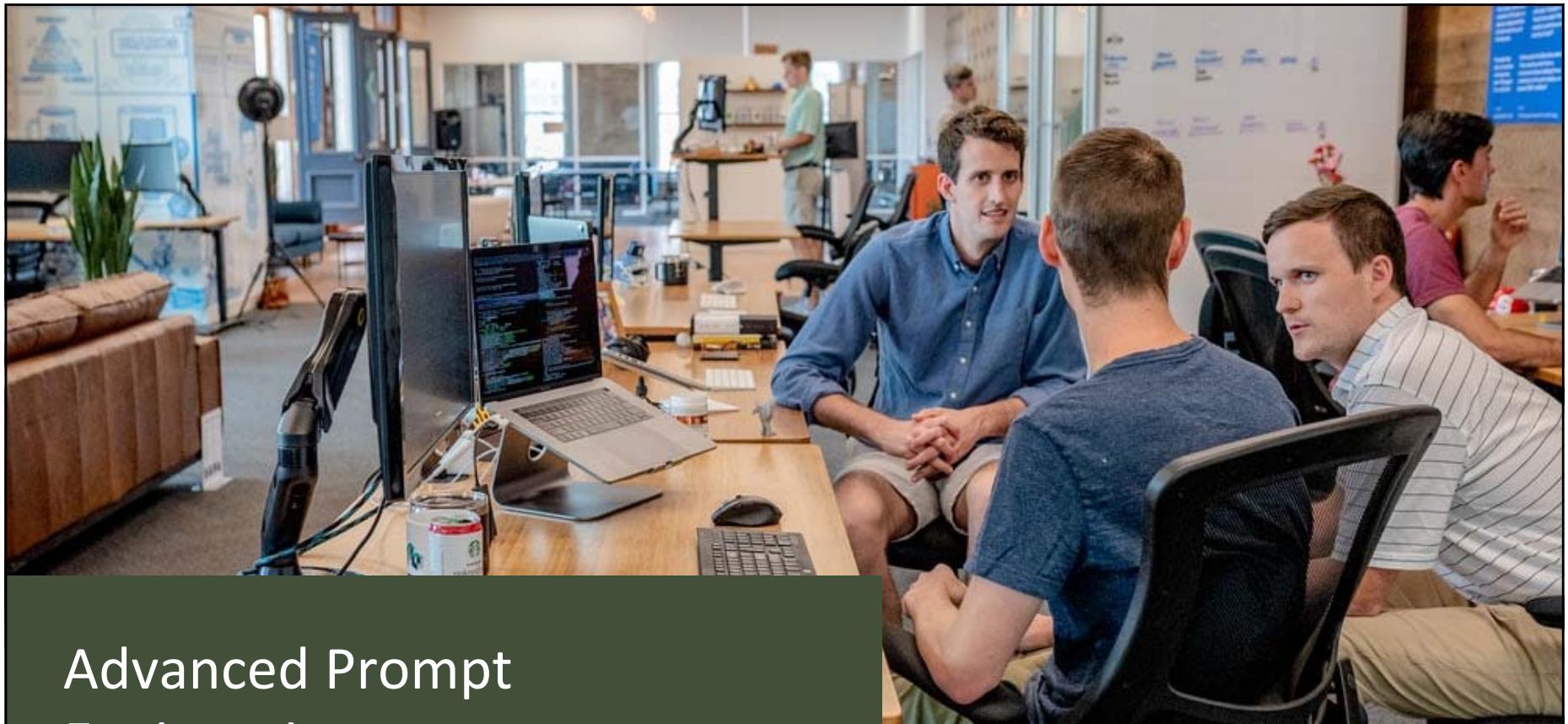
USER Summarize the text delimited by triple quotes in about 50 words.

"""insert text here"""

[Open in Playground ↗](#)

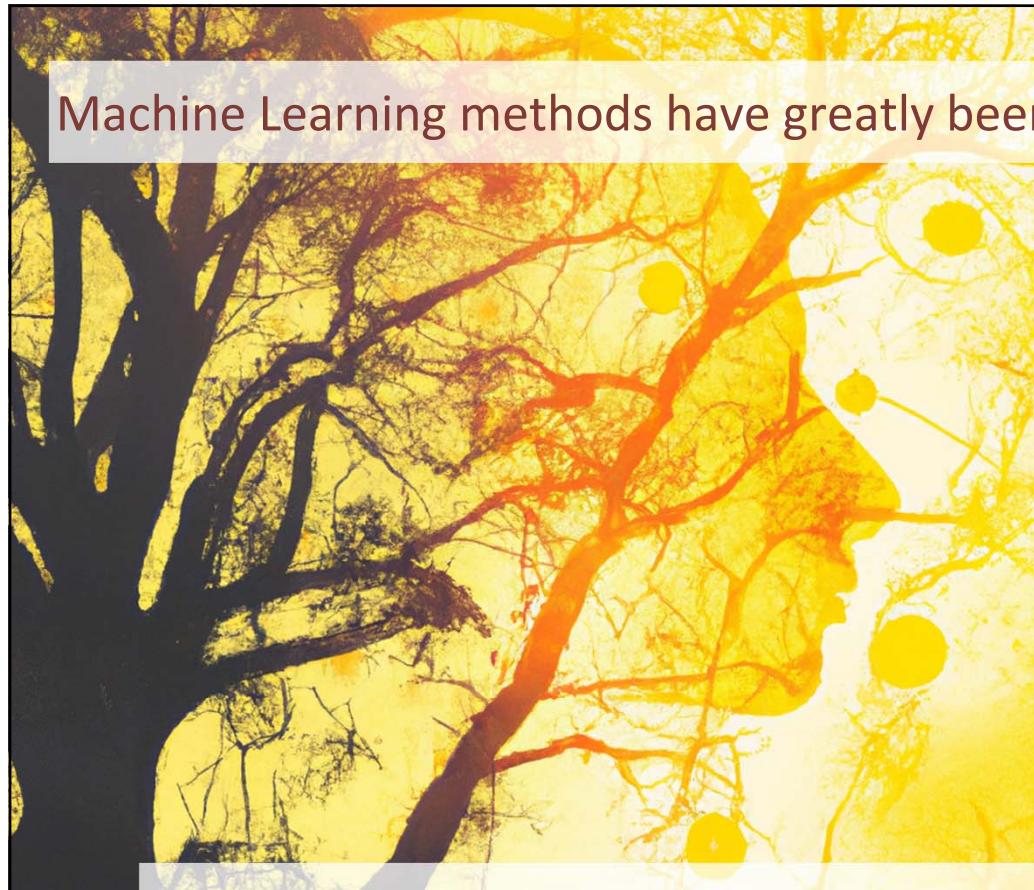
USER Summarize the text delimited by triple quotes in 2 paragraphs.

"""insert text here"""



Advanced Prompt Engineering

JADS
Jheronimus
Academy
of Data Science



Machine Learning methods have greatly been inspired by natural phenomena

- Neural Networks
- Genetic Algorithms
- Particle Swarm Optimization
- Ant Colony Optimization
- Bee algorithms
- Fuzzy sets

Many prompting techniques are inspired by the human-thought process

JNDS

This image was created by DALL-E 2

Create an image that shows how inspiration for artificial intelligence is drawn from nature phenomena

Generate

Bad prompt: very vague



39

JADS
Jheronimus
Academy
of Data Science

Combining a vague and specific description in a prompt

Create an image of a robot with a thinking box on top of his head which shows natural phenomena that it can use as inspiration for getting smarter

Generate



40

JADS
Jheronimus
Academy
of Data Science

Advanced prompt engineering techniques

- Few shot prompts
- Chain of thought reasoning
- Effective graph of thought <https://arxiv.org/abs/2305.16582>
- Self-consistency <https://openreview.net/forum?id=1PL1NIMMrw>
- Knowledge generation prompting
- ReACT

Few-Shot Prompting

JADS
Jheronimus
Academy
of Data Science

Few-Shot Prompts

Min, S. et al. (2022). study *how* the model learns and *which* aspects of the demonstrations contribute to end task performance.

Key findings:

- Ground truth in prompts matters little. Instead, other key aspects of the demonstrations are key drivers in end task performance:
 1. The label space
 2. The distribution of the input text
 3. The overall format of the sequence.

⁴³ Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Few-Shot Prompts – Ground truth matters little

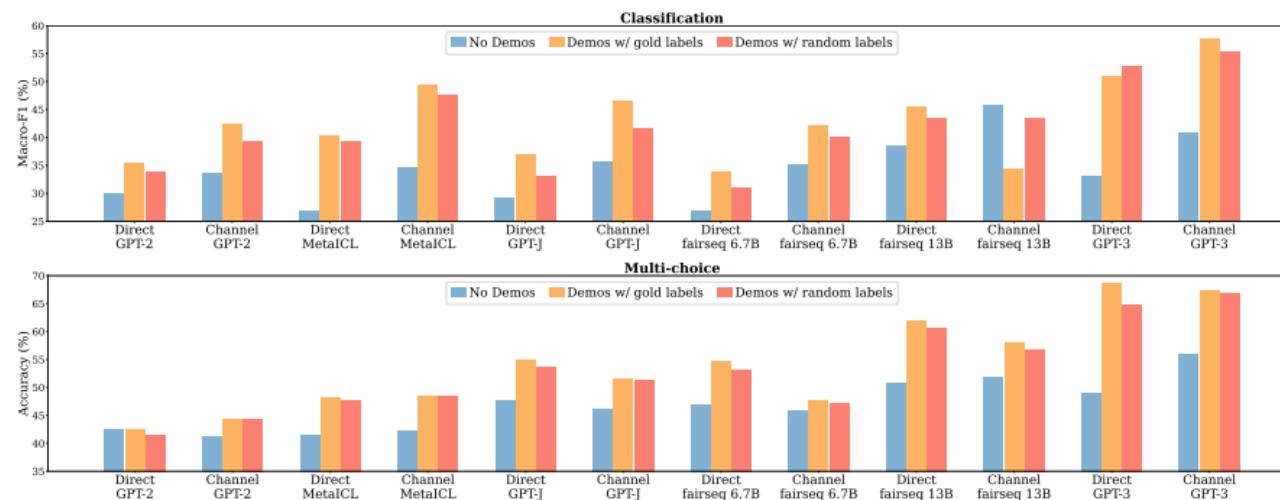


Figure 3: Results when using no-demonstrations, demonstrations with gold labels, and demonstrations with random labels in classification (top) and multi-choice tasks (bottom). The first eight models are evaluated on 16 classification and 10 multi-choice datasets, and the last four models are evaluated on 3 classification and 3 multi-choice datasets. See Figure 11 for numbers comparable across all models. **Model performance with random labels is very close to performance with gold labels** (more discussion in Section 4.1).

44 Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Few-Shot Prompts – Ground truth matters little

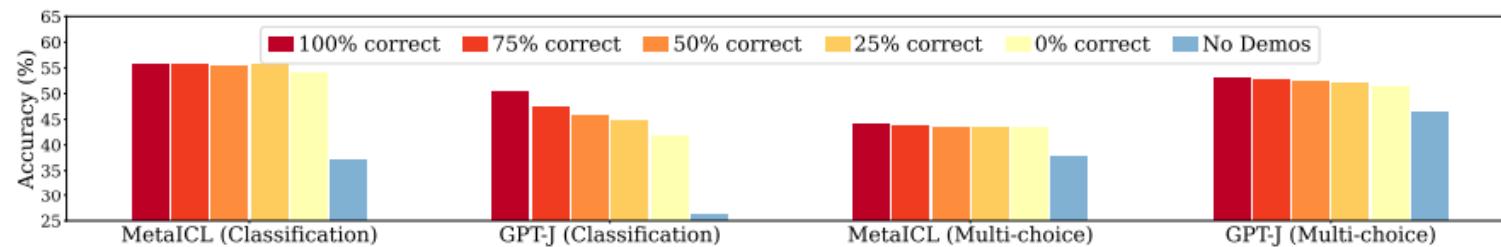


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

45

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Few-Shot Prompts – Ground truth matters little

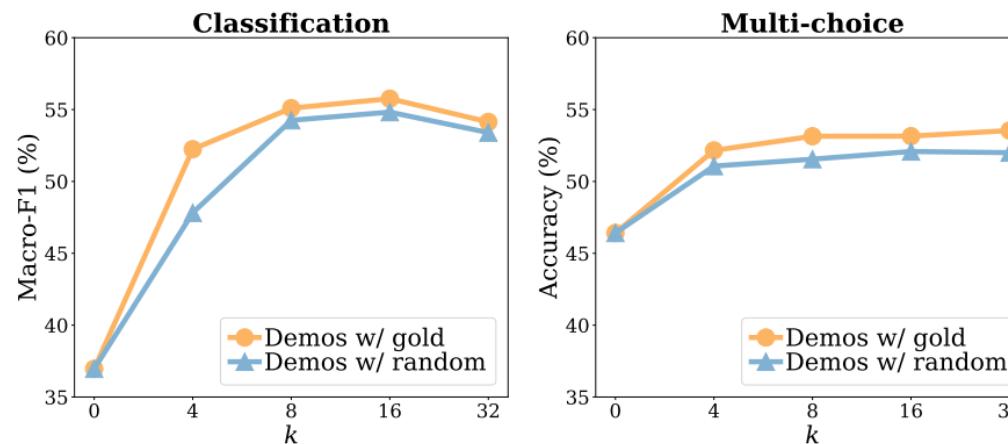


Figure 5: Ablations on varying numbers of examples in the demonstrations (k). Models that are the best under 13B in each task category (Channel MetaICL and Direct GPT-J, respectively) are used.

46 Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Few-Shot Prompts – The label space matters

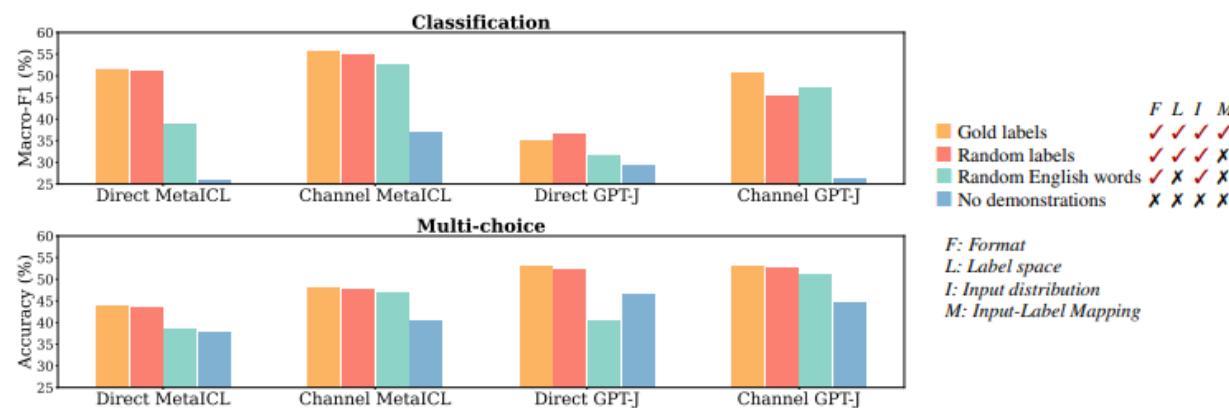


Figure 9: Impact of the label space. Evaluated in classification (top) and multi-choice (bottom). The impact of the label space can be measured by comparing ■ and ■. The gap is significant in the direct models but not in the channel models (discussion in Section 5.2).

47

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Few-Shot Prompts – The distribution of the input space matters

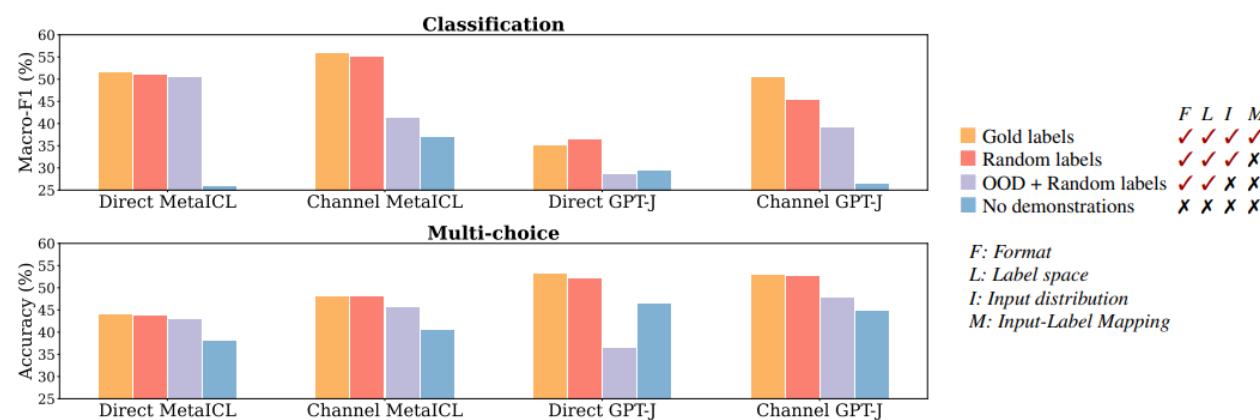


Figure 8: Impact of the distribution of the inputs. Evaluated in classification (top) and multi-choice (bottom). The impact of the distribution of the input text can be measured by comparing █ and █. The gap is substantial, with an exception in Direct MetaICL (discussion in Section 5.1).

48

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Few-Shot Prompts – The overall format of the sequence

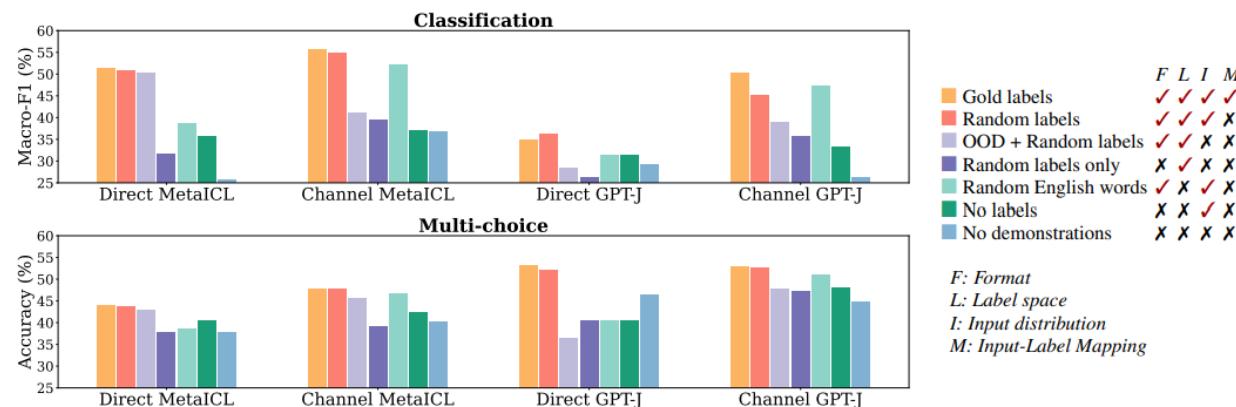


Figure 10: Impact of the format, i.e., the use of the input-label pairs. Evaluated in classification (top) and multi-choice (bottom). Variants of demonstrations without keeping the format (■ and ■) are overall not better than no demonstrations (□). Keeping the format is especially significant when it is possible to achieve substantial gains with the label space but without the inputs (■ vs. ■ in Direct MetaICL), or with the input distribution but without the labels (■ vs. ■ in Channel MetaICL and Channel GPT-J). More discussion in Section 5.3.

49 Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). "Rethinking the role of demonstrations: What makes in-context learning work?". In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 11048 – 11064.

Chain-of-Thought

JADS
Jheronimus
Academy
of Data Science

Chain of Thought Reasoning

Goal: enhance reasoning capabilities of large language models.

Motivation:

1. Techniques for arithmetic reasoning can benefit from generating natural language rationales that lead to the final answer.

Limitation: costly to create a large set of high-quality rationales.

2. Large language models offer the exciting prospect of in-context few-shot learning.

Limitation: traditional few-shot prompting methods work poorly on tasks requiring reasoning abilities, and often do not improve substantially with increasing language model scale

⁵¹ Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.



Chain of Thought Reasoning

Goal: enhance reasoning capabilities of large language models.

Motivation:

1. Techniques for generating rationales that lead to the final answer.

Limitation:

CoTR combines the strengths of these two ideas while avoiding their limitations.

2. First, techniques for generating rationales that lead to the final answer.

Limitation: traditional few-shot prompting methods work poorly on tasks requiring reasoning abilities, and often do not improve substantially with increasing language model scale

⁵² Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.



Chain of Thought Reasoning

Chain of thought =

a series of intermediate natural language reasoning steps that lead to the final output

Traditional few-shot prompts format:

<input, output>

CoT few-shot prompts format:

<input, *chain of thought*, output>

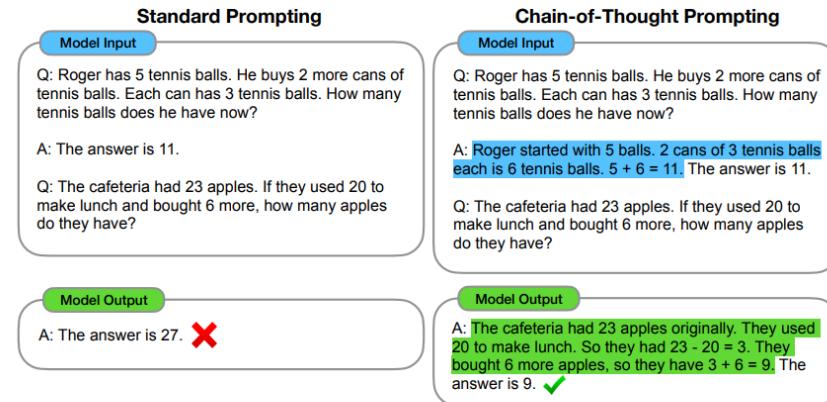


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

53

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.



CoT-reasoning has several attractive properties

1. Allows models to **decompose multi-step problems** into intermediate steps, which means that additional computation can be allocated to problems that require more reasoning steps.
2. A chain of thought provides an **interpretable window into the behavior** of the model, suggesting how it might have arrived at a particular answer and providing opportunities to debug where the reasoning path went wrong
3. Can be used for tasks such as **math word problems**, **commonsense reasoning**, and **symbolic manipulation**, and is potentially applicable (at least in principle) to any task that humans can solve via language.
4. Can be **readily elicited in sufficiently large off-the-shelf language models** simply by including examples of chain of thought sequences into the exemplars of few-shot prompting.

⁵⁴ Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.

Examples of CoT triples

Math Word Problems (free response) <p>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?</p> <p>A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p>	Math Word Problems (multiple choice) <p>Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788</p> <p>A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).</p>	CSQA (commonsense) <p>Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock</p> <p>A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).</p>
StrategyQA <p>Q: Yes or no: Would a pear sink in water?</p> <p>A: The density of a pear is about 0.6 g/cm³, which is less than water. Thus, a pear would float. So the answer is no.</p>	Date Understanding <p>Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?</p> <p>A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.</p>	Sports Understanding <p>Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."</p> <p>A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.</p>
SayCan (Instructing a robot) <p>Human: How would you bring me something that isn't a fruit?</p> <p>Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar. Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().</p>	Last Letter Concatenation <p>Q: Take the last letters of the words in "Lady Gaga" and concatenate them.</p> <p>A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.</p>	Coin Flip (state tracking) <p>Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?</p> <p>A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.</p>

55

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.

CoT reasoning is effective on large-scale models

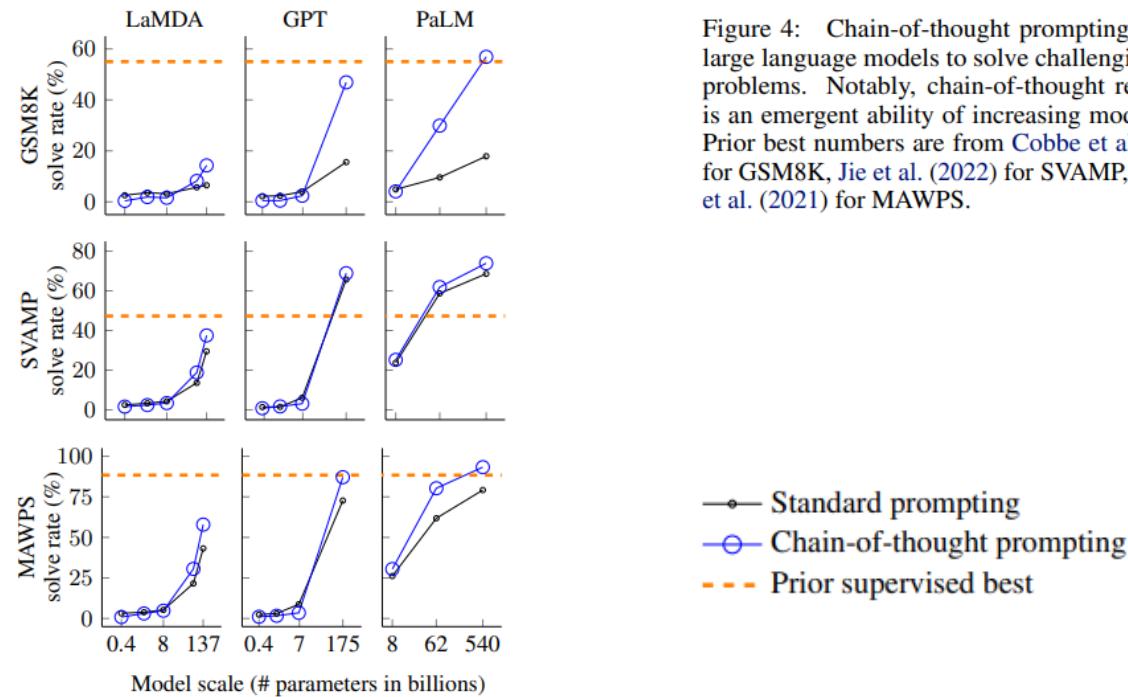


Figure 4: Chain-of-thought prompting enables large language models to solve challenging math problems. Notably, chain-of-thought reasoning is an emergent ability of increasing model scale. Prior best numbers are from Cobbe et al. (2021) for GSM8K, Jie et al. (2022) for SVAMP, and Lan et al. (2021) for MAWPS.

56

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.

Experimenting with textual reasoning tasks

- Motivation:** Prior CoT reasoning largely targets symbolic reasoning tasks with a very different structure, such as math word problems.
- Approach:** Investigate the nature of the explanations that LLMs generate and whether they can improve few-shot in-context learning for **textual reasoning tasks**. Evaluate on two axes:
1. **Factuality**: whether the explanation is correctly grounded in the input.
 2. **Consistency**: whether the explanation entails the final prediction.

57

Ye, X., & Durrett, G. (2022). The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35, 30378-30392.



LLMs can be unreliable

The analysis suggests that

- LLMs' internal “reasoning” does not always align with explanations that it generates, as shown by our consistency results.
- More concerning, the explanations might not be factually grounded in the provided prompt. This shortcoming should caution against any deployment of this technology in practice: because the explanations are grammatical English and look very convincing, they may deceive users into believing the system's responses even when those responses are incorrect.

Self-Consistency Sampling

JADS
Jheronimus
Academy
of Data Science

Self-Consistency Sampling

- Problem CoT:** The language model generates text token by token. Yet, it doesn't know how to end, and it might go down a wrong path when choosing the next word.
- Intuition:** Complex reasoning tasks typically admit multiple reasoning paths that reach a correct answer. The more deliberate thinking and analysis is required for a problem (Evans, 2010), the greater the diversity of reasoning paths that can recover the answer.
- Idea:** sample multiple, diverse reasoning paths through few-shot CoT, and use the generations to select the most consistent answer.

61

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR).



Self-Consistency approach

1. **Identify the problem:** Define the problem or question for which you require LLM's assistance. Make sure it is clear and specific.
2. **Create multiple prompts:** Develop various prompts that approach the problem from different angles or perspectives. Each prompt should provide a unique reasoning path for the AI to follow.
3. **Generate responses:** Submit the prompts to LLM and obtain the responses generated by the model.
4. **Evaluate consistency:** Analyze the generated responses to determine their coherence, relevance, and consistency. This step may involve comparing the responses to each other, looking for common themes or patterns, and checking for internal logical consistency.
5. **Select the best response:** Based on the evaluation, choose the most consistent and accurate response as the final answer.

62

Source: <https://www.promptengineering.org>

Self-Consistency sampling example 1/2

Prompt:

When I was 6 my sister was half my age. Now I'm 70 how old is my sister?

Output:



63

Source: <https://www.promptengineering.org>

JADS
Jheronimus
Academy
of Data Science

Self-Consistency sampling example 2/2

The following prompt, provide multiple examples of a similar problem and how we would work to come up with the correct answer.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted. So, they must have planted $21 - 15 = 6$ trees. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are 3 cars in the parking lot already. 2 more arrive. Now there are $3 + 2 = 5$ cars. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Leah had 32 chocolates and Leah's sister had 42. That means there were originally $32 + 42 = 74$ chocolates. 35 have been eaten. So in total they still have $74 - 35 = 39$ chocolates. The answer is 39.

Q: When I was 6 my sister was half my age. Now I'm 70 how old is my sister?

A:

Output 1:

When I was 6 my sister was half my age, so she was 3. Now I am 70, so she is $70 - 3 = 67$. The answer is 67.

Majority vote

Output 2:

When the narrator was 6, his sister was half his age, which is 3. Now that the narrator is 70, his sister would be $70 - 3 = 67$ years old. The answer is 67.

Output 3:

When I was 6 my sister was half my age, so she was 3. Now I am 70, so she is $70/2 = 35$. The answer is 35.

Knowledge Generation Prompting

JADS
Jheronimus
Academy
of Data Science

Knowledge Generation Prompting

- Intuition:** Perhaps models' commonsense reasoning performs better when prompts are incorporated with external knowledge.
- Idea:** Generate useful knowledge from a language model, then provide the knowledge as an input prompt that is concatenated with a question.
- Approach:** For a given question:
1. *Knowledge generation*: let language models generate knowledge statements conditioned on the question.
2. *Knowledge Integration*: integrate the knowledge from step 1 into the decision process of a language model used for inference.
- Advantage:** This approach does not require access to a structured knowledge base or joint finetuning for knowledge integration.

Knowledge Generation Prompting

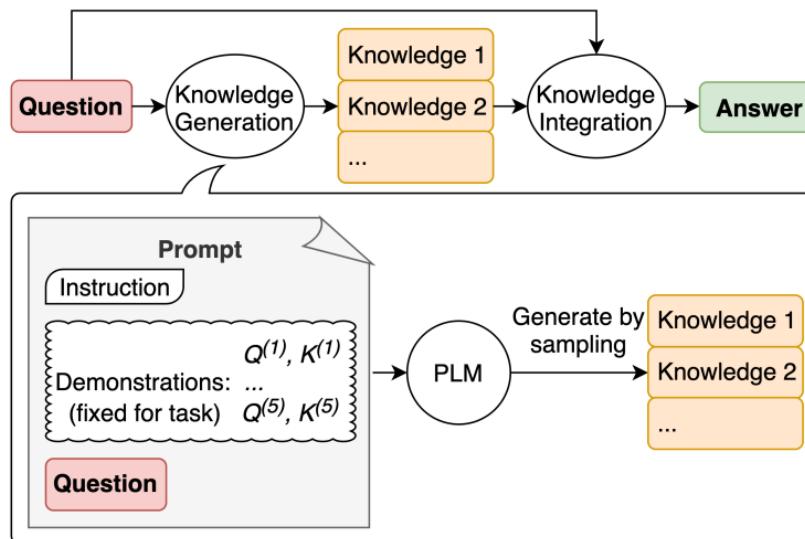


Figure 1: Generated knowledge prompting involves (i) using few-shot demonstrations to generate question-related knowledge statements from a language model; (ii) using a second language model to make predictions with each knowledge statement, then selecting the highest-confidence prediction.

Examples where prompting with generated knowledge rectifies model prediction

Dataset	Question / Knowledge	Prediction	Score
NumerSense	the word children means [M] or more kids.	one	0.37 0.35
	<i>The word child means one kid.</i>	two	0.91
CSQA	She was always helping at the senior center, it brought her what?	feel better	0.97 0.02
	<i>People who help others are usually happier.</i>	happiness	0.98
CSQA2	Part of golf is trying to get a higher point total than others.	yes	1.00 0.00
	<i>The player with the lowest score wins.</i>	no	1.00
QASC	Sponges eat primarily	cartilage	0.95 0.00
	<i>Sponges eat bacteria and other tiny organisms.</i>	krill and plankton	0.99

Table 1: Examples where prompting with generated knowledge rectifies model prediction. Each section shows the correct answer in green, the incorrect answer in red, and the prediction scores from the inference model that only sees the question (top) and the same model that sees the question prompted with the given knowledge (bottom).

Tree of Thoughts

- Motivation:** For complex tasks that require exploration or strategic lookahead, traditional or simple prompting techniques fall short.
- Idea:** Combine the strengths of chain-of-thought prompting and self-consistency.
- Approach:** For a given question:
1. Generate multiple initial thoughts, representing the initial steps of problem-solving.
2. Query the AI to self-criticize each of these thoughts. Let it evaluate how well each thought aligns with the problem-solving objective (this could involve ranking each thought or assigning scores to them).
3. Upon reviewing the critiques, discard the less suitable thoughts and expand on the remaining thoughts.

69

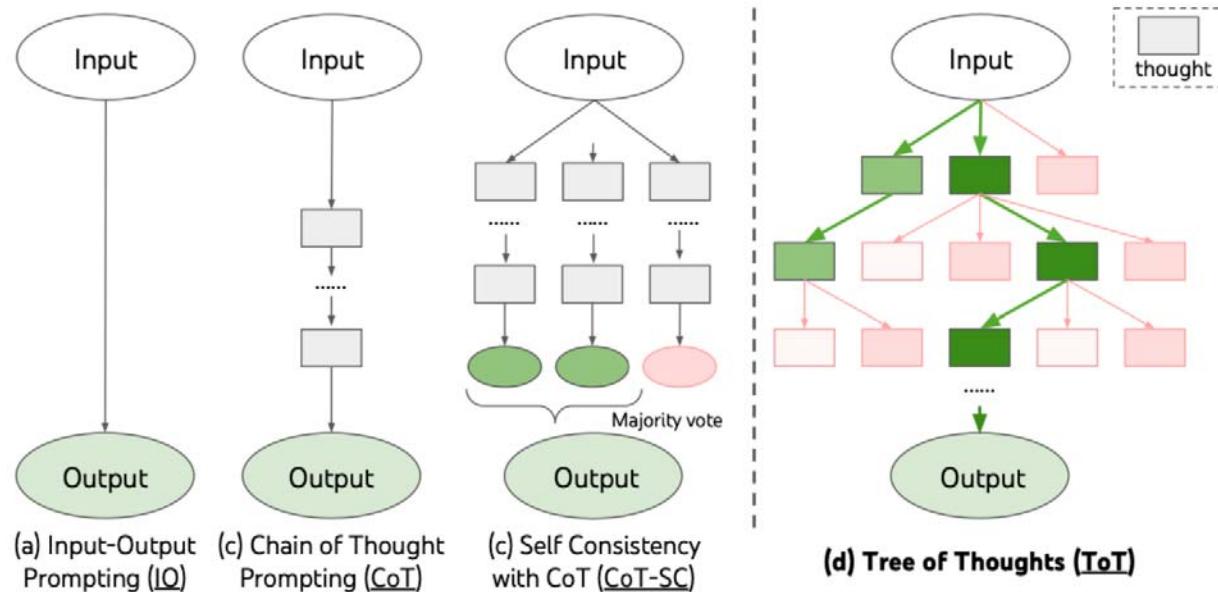
Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). "Tree of thoughts: Deliberate problem solving with large language models". In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS'23)*, p. 11809 – 11822.



Tree of Thoughts

JADS
Jheronimus
Academy
of Data Science

Tree of Thoughts



71

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). "Tree of thoughts: Deliberate problem solving with large language models". In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS'23)*, p. 11809 – 11822.

JADS
Jheronimus
Academy
of Data Science

Tree of Thoughts example

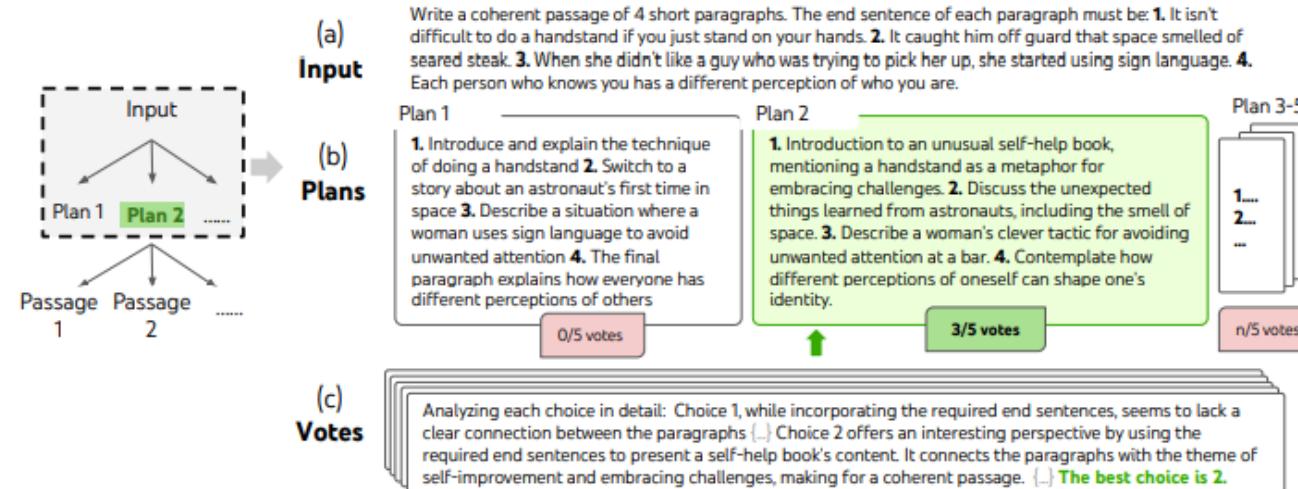


Figure 4: A step of deliberate search in a randomly picked Creative Writing task. Given the input, the LM samples 5 different plans, then votes 5 times to decide which plan is best. The majority choice is used to consequently write the output passage with the same sample-vote procedure.

Tree of Thoughts surpasses CoT prompts significantly

Method	Success
IO prompt	7.3%
CoT prompt	4.0%
CoT-SC ($k=100$)	9.0%
ToT (ours) ($b=1$)	45%
ToT (ours) ($b=5$)	74%
IO + Refine ($k=10$)	27%
IO (best of 100)	33%
CoT (best of 100)	49%

Table 2: Game of 24 Results.

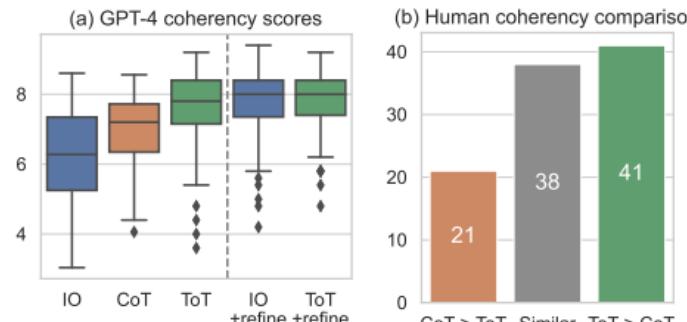


Figure 5: Creative Writing results.

Method	Success Rate (%)		
	Letter	Word	Game
IO	38.7	14	0
CoT	40.6	15.6	1
ToT (ours)	78	60	20
+best state	82.4	67.5	35
-prune	65.4	41.5	5
-backtrack	54.6	20	5

Table 3: Mini Crosswords results.

73

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). "Tree of thoughts: Deliberate problem solving with large language models". In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS'23)*, p. 11809 – 11822.

Limitations and considerations of ToT

- Disadvantage:** Computational requirements: at worst, ToT has an exponential number of calls with increasing lengths per call.
- Considerations:** Deliberate search/reasoning methods might not be necessary for many existing tasks.
- However, as LMs are being deployed for more real-world decision making applications, more complex tasks could emerge and present new opportunities to study these research questions.

74

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). "Tree of thoughts: Deliberate problem solving with large language models". In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS'23)*, p. 11809 – 11822.



Retrieval-Augmented Generation

JADS
Jheronimus
Academy
of Data Science

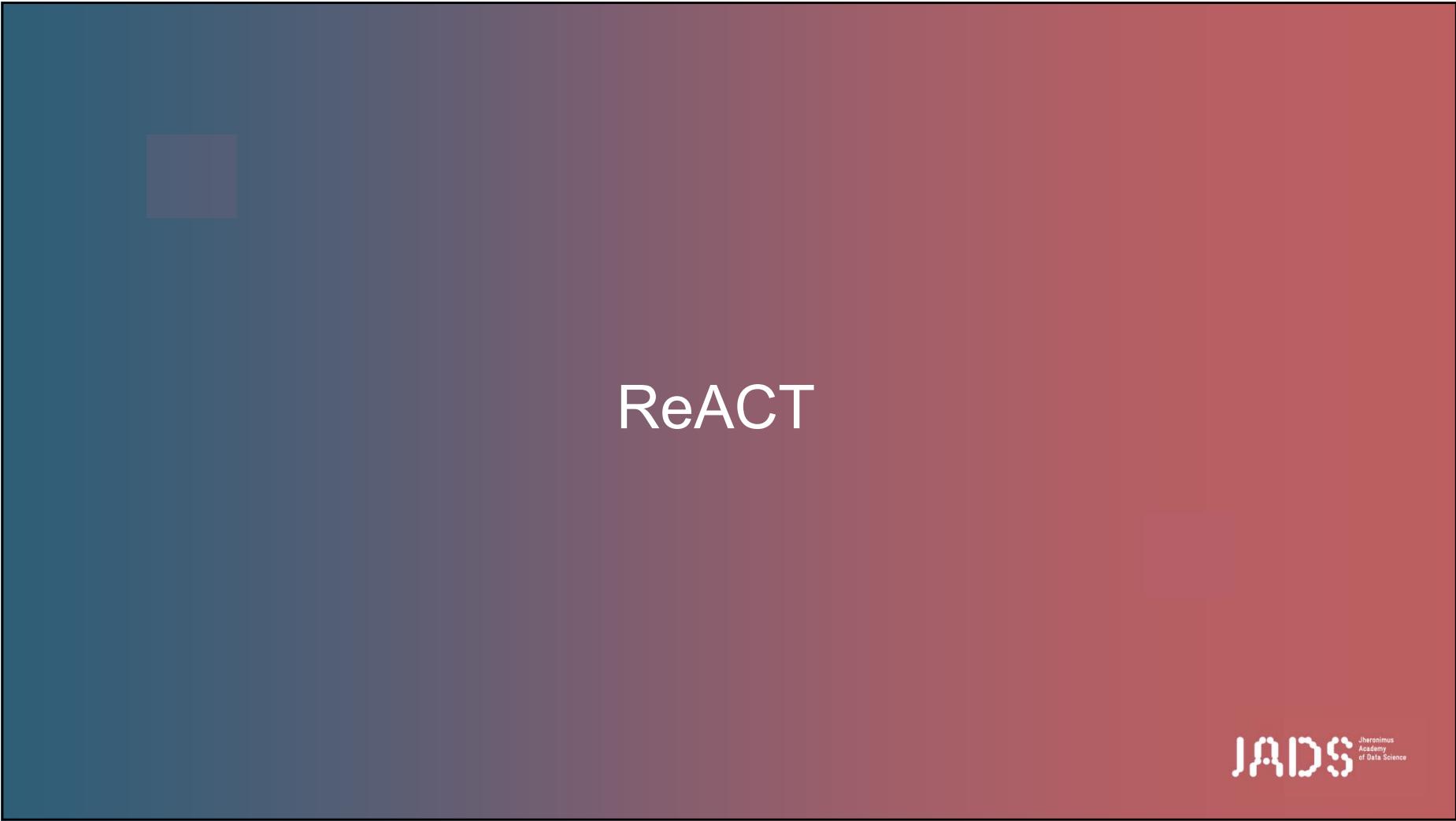
Retrieval Augmented Generation (RAG)

- Motivation:** For more factual consistency, improved reliability and mitigating ‘hallucinations’ some complex tasks may benefit from additional background knowledge.
- Idea:** Combine an information retrieval component with a text generator model.
- Approach:** Given an input:
1. Retrieve a set of relevant/supporting documents given a source (e.g. Wikipedia)
2. Concatenate the documents as context with the original input prompt.

76

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.





ReACT

JADS
Jheronimus
Academy
of Data Science

ReAct: Reasoning and Acting

Motivation:

LLMs have been used separately to get impressive performance across language tasks (e.g. chain-of-thought prompting) and action (e.g. plan generation). It would be interesting to combine both.

Action planning complex

Consider a general setup of an **agent interacting with an environment** for task solving.

At time step t , an agent receives an **observation** $o_t \in \mathcal{O}$ from the environment and **takes action** $a_t \in \mathcal{A}$ following some **policy** $\pi(a_t|c_t)$ where $c_t = (o_1, a_1, \dots, o_{t-1}, a_{t-1}, o_t)$ is the context to the agent.

Learning a policy is challenging when the mapping $c_t \mapsto a_t$ is highly implicit and requires extensive computation.

78

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. In the proceedings of the International Conference on Learning Representations 2023.



ReAct: Reasoning and Acting

Approach: Augment the agent's action space to $\hat{\mathcal{A}} = \mathcal{A} \cup \mathcal{L}$

Where:

- \mathcal{L} the space of language,
- $\hat{a}_t \in \mathcal{L}$ an action in the language space, referred to as a thought.
 - * Does **not** affect the external environment.
 - * A thought \hat{a}_t aims to compose useful information by reasoning over the current context c_t , and update the context $c_{t+1} = (c_t, \hat{a}_t)$ to support future reasoning or acting.

Datasets: * FEVER - a fact verification benchmark where each claim is annotated SUPPORTS, REFUTES or NOT ENOUGH INFO.

79

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. In the proceedings of the International Conference on Learning Representations 2023.



ReAct example

Action space:

Wikipedia web API with three types of actions:

- 1) **search[entity]** – returns the first 5 sentences from page.
- 2) **lookup[string]** – returns the next sentence in the page.
- 3) **finish[answer]** – finish task with answer.

Purpose: simulate how humans would interact with Wikipedia.

Prompts:

ReAct-format trajectories to use as few-shot exemplars in the prompts. Each trajectory consists multiple thought-action-observation steps, where free-form thoughts are used for various purposes.

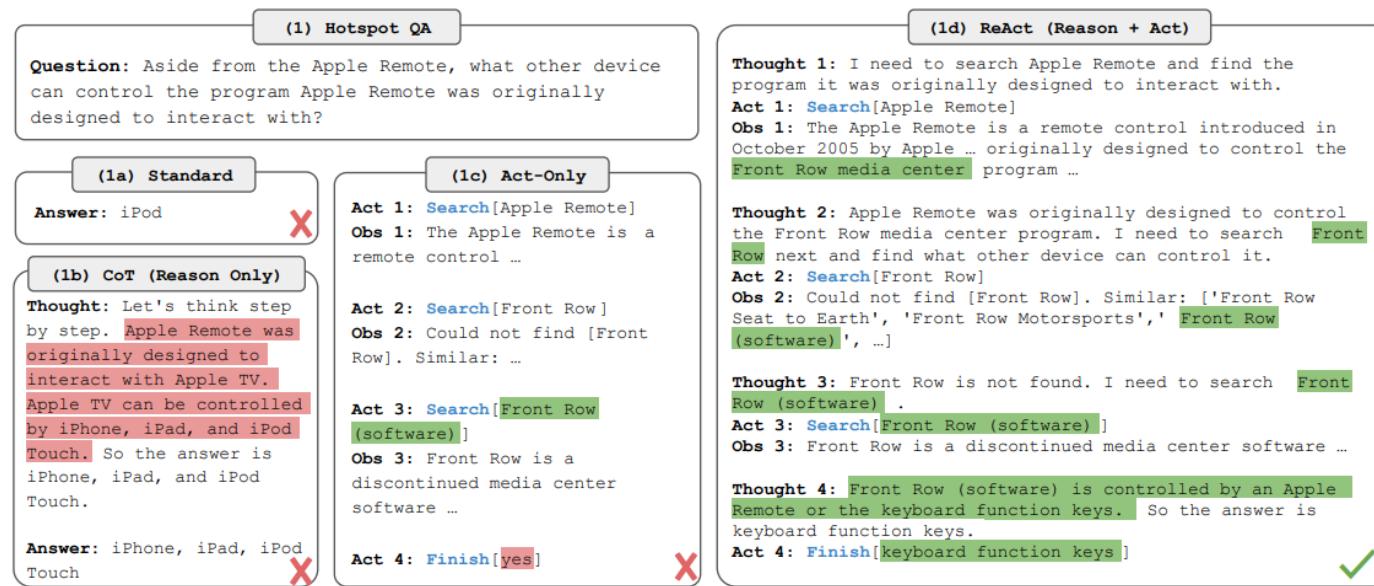
- 1) to decompose questions,
- 2) extract information observations,
- 3) perform commonsense,
- 4) arithmetic reasoning,
- 5) guide search reformulation,
- 6) synthesize the final answer.

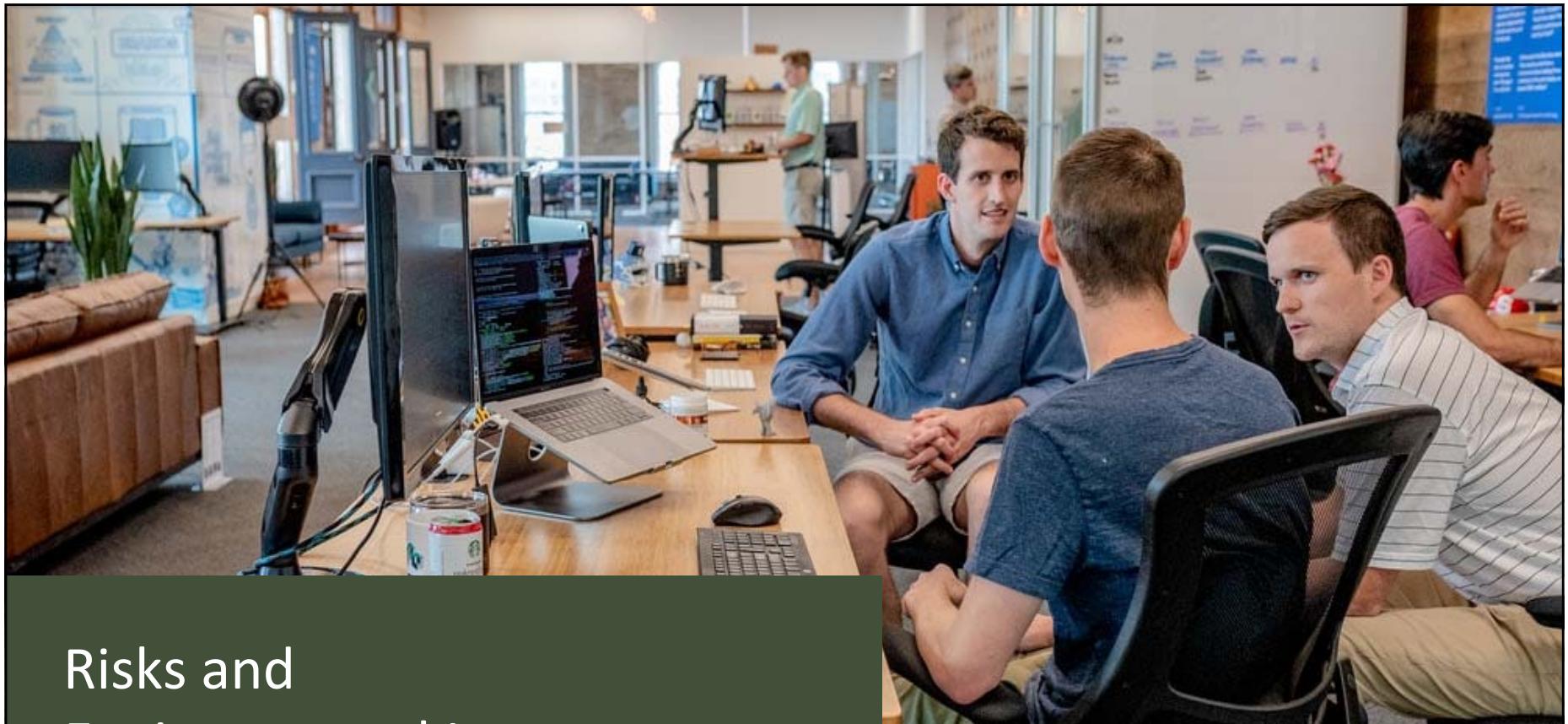
80

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. In the proceedings of the International Conference on Learning Representations 2023.



ReAct example





Risks and Environmental Impact

JADS
Jheronimus
Academy
of Data Science

LLMs have been prone to displaying hallucinations, toxicity, and biased stereotypes

Hallucinations: The phenomenon where the model generates text that is **incorrect, nonsensical, or not real**.

Toxicity: The generation of harmful, offensive, or inappropriate content by these models.

Bias: The manifestation of societal biases and prejudices in the generated text. Biases include:

1. gender biases,
2. cultural biases,
3. political biases.

Environmental impact

BERT (Bidirectional Encoder Representations from Transformers)

- Google, 2019
- Parameters: 300 million
- Training on a GPU is roughly equivalent to a trans-American flight

BLOOM

- Hugging Face, 2022
- Training: 25 tons of carbon dioxide emissions (30 flights between London and New York)
- But less than equivalent LLMs because uses nuclear energy

GPT-3

- Training: 500 tons of carbon dioxide emissions (600 flights)

ChatGPT costs OpenAI approximately \$700,000 a day to operate.

94

<https://unu.edu/macau/blog-post/unsustainability-chatgpt-impact-large-language-models-sustainable-development-goals#:~:text=Secondly%2C%20environmental%20impacts%20by%20such,ands%20large%20clusters%20computing%20infrastructure.>



Lack of Explainability in Document Classification?

Traditional data science: Explainability was typically about feature importance, sensitivity analysis, model-specific explainers (e.g. LIME or SHAP).

Human explainability: Based on reasoning, not on ‘parameters in the brain’.

Large Language Models: For document classification, ChatGPT can classify documents in a zero-, or few-shot manner and reason about its decision.



• Questions?