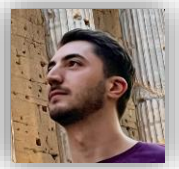
NLP and the Web – WS 2024/2025

Lecture 11
Neural Language Modeling 3

TECHNISCHE UNIVERSITÄT DARMSTADT


Dr. Thomas Arnold
Hovhannes Tamoyan
Kexin Wang

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Syllabus (tentative)

Nr.	Lecture
01	Introduction / NLP basics
02	Foundations of Text Classification
03	IR – Introduction, Evaluation
04	IR – Word Representation
05	IR – Transformer/BERT
06	IR – Dense Retrieval
07	IR – Neural Re-Ranking
08	LLM – Language Modeling Foundations, Tokenization
09	LLM – Neural LLM
10	LLM – Adaptation
11	**LLM – Prompting, Alignment, Instruction Tuning**
12	LLM – Long Contexts, RAG
13	LLM – Scaling, Computation Cost
14	Review & Preparation for the Exam

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

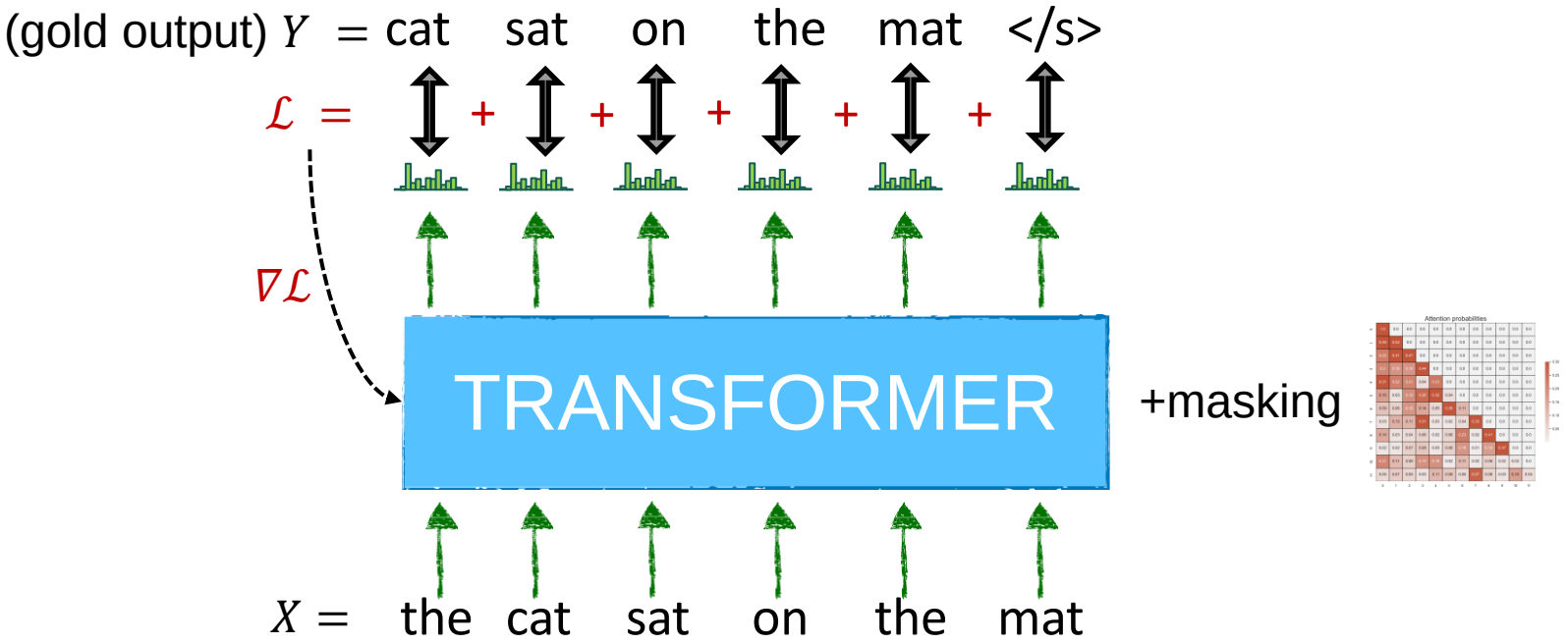
Recap

Prompting, In-Context Learning

Alignment, Instruction Tuning

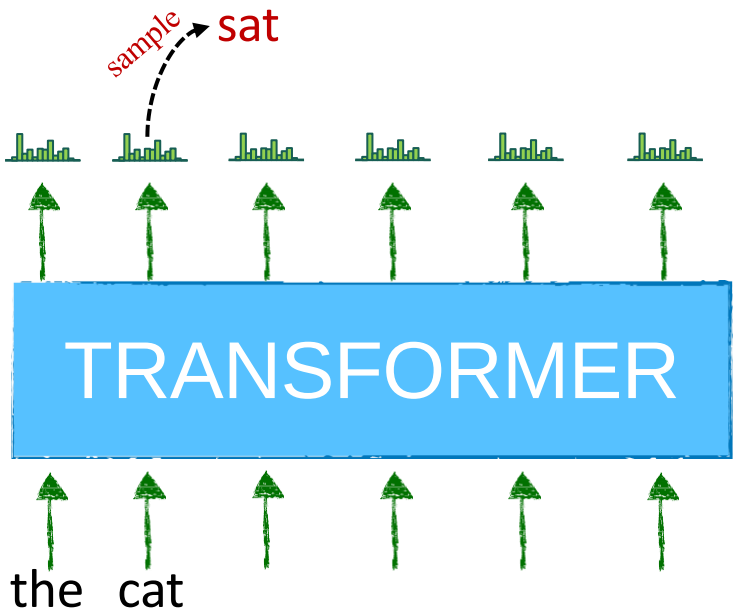
Training a Transformer Language Model

- We need to **prevent information leakage** from future tokens! How?



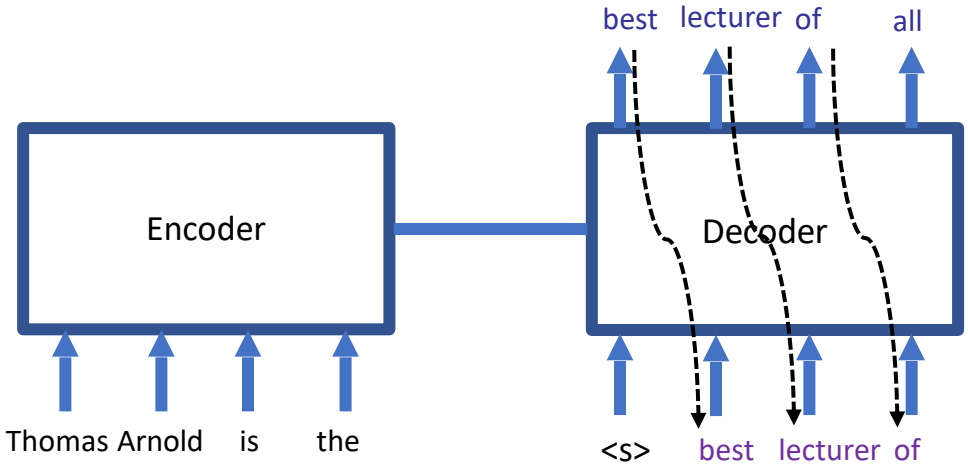
How to use the model to generate text?

- Use the output of previous step as input to the next step repeatedly

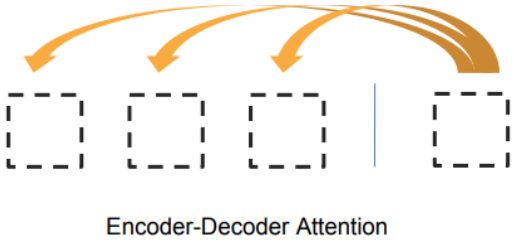


Encoder-decoder models

- Encoder = read or encode the input
- Decoder = generate or decode the output

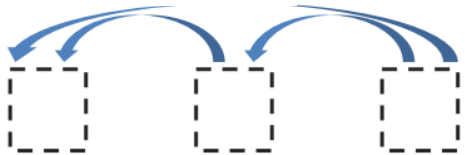


3 Shades of Attention

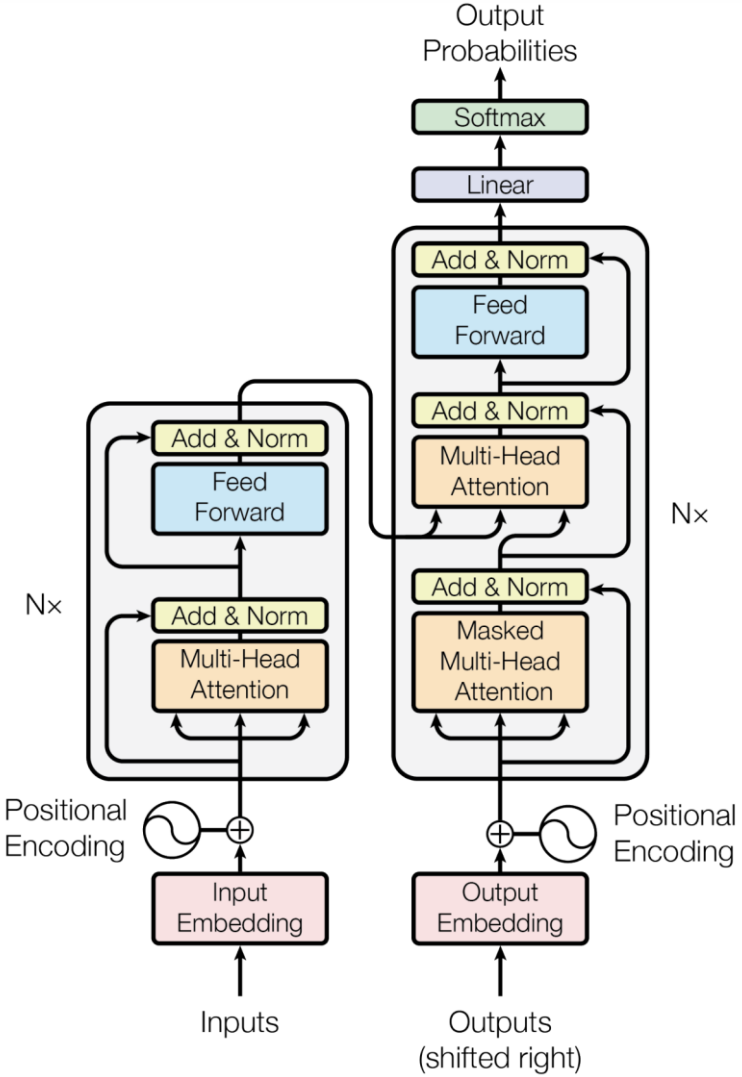


Encoder-Decoder Attention

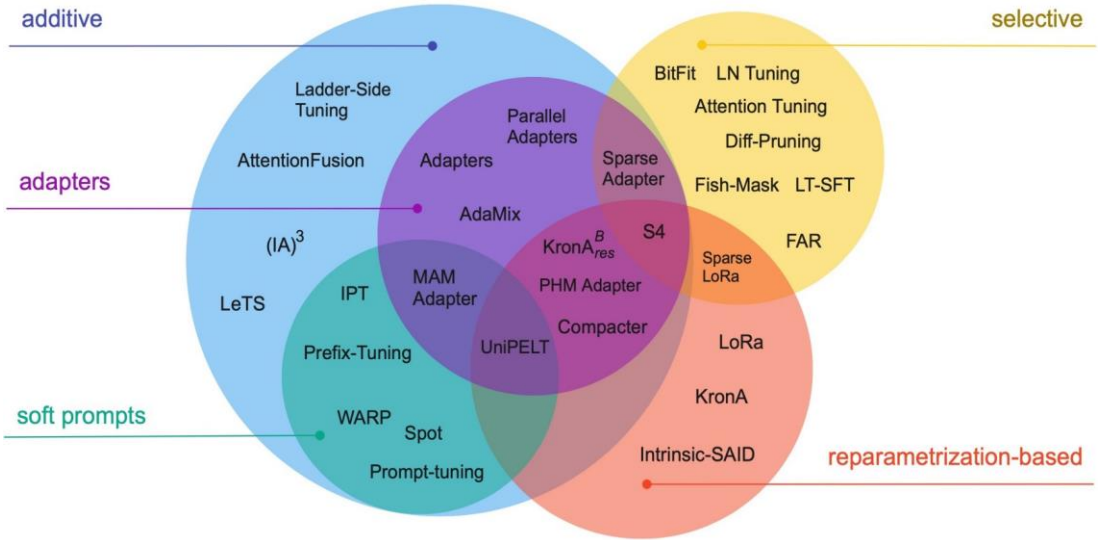
Encoder Self-Attention



MaskedDecoder Self-Attention



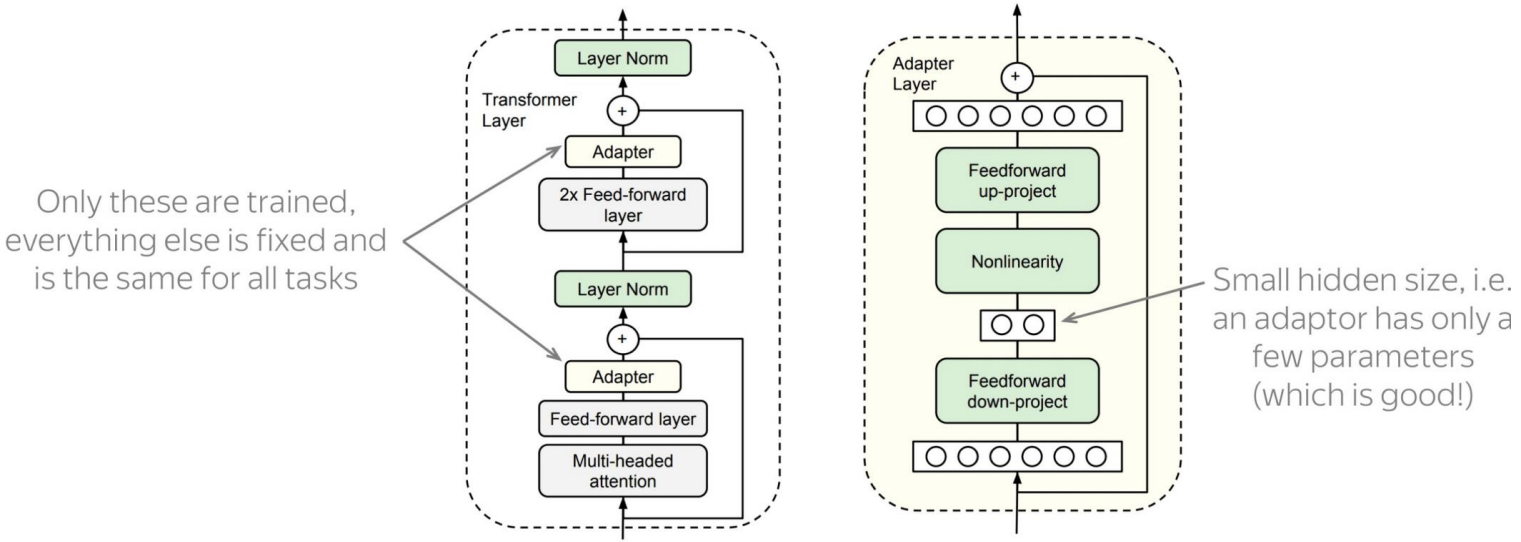
Adaptation methods



Ben Zaken et al., 2021. “BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models”

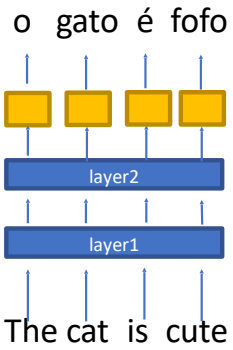
Additive Method: Adapters

- **Idea:** train small sub-networks and only tune those.
  - Adapter layer projects to a low dimensional space to reduce parameters.
- No need to store a full model for each task, **only the adapter params**.

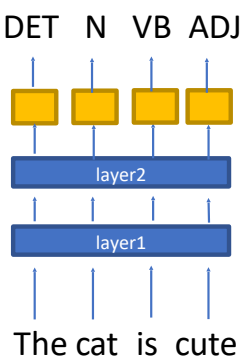


["Parameter-Efficient Transfer Learning for NLP", Houlsby et al., 2019.]

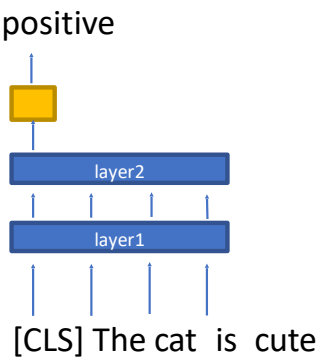
Fine-Tuning for Tasks



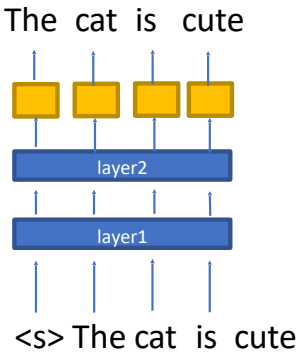
Translation



POS Tagging



Text classification



Language modeling

Outline

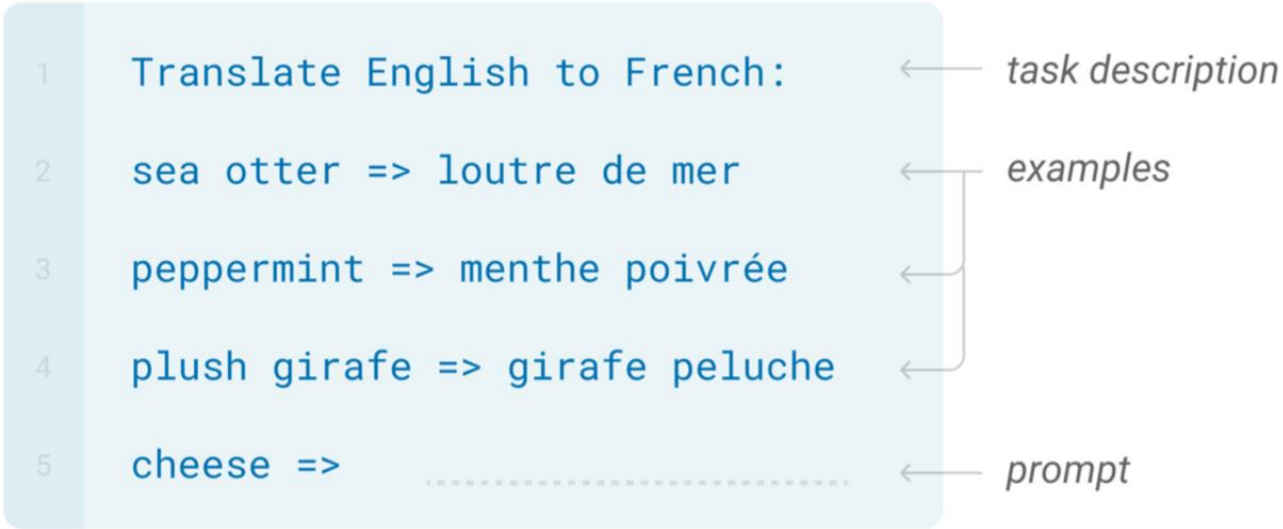
TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Prompting, In-Context Learning

Alignment, Instruction Tuning

In-Context Learning

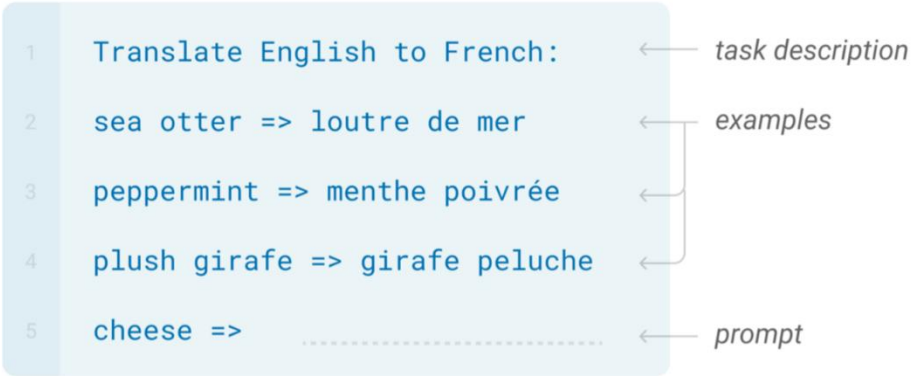


In-Context Learning

- Learns to do a downstream task by conditioning on input-output examples!

- **No weight update** — our model is not **explicitly pre-trained** to learn from examples
  - The underlying models are quite general

  - How to use effectively in practice?
  - Fundamentally, why does it work?



In-Context Learning

Reverse words in a sentence

This is great
Great is this

The man on the moon
Moon the on man the

Will this really work
Work really this will

I hope this is a big achievement
Achievement big I hope this is

The king came home on a horse
Home horse king came the

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life _ for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?
A: 54

Q: where does she live?
A:

Model answer: Stockholm
Turker answers: Sweden, Sweden, in Sweden, Sweden

Why Do We Care About In-Context Learning?

TECHNISCHE UNIVERSITÄT DARMSTADT

Practically Useful

Intellectually Intriguing

[ACL 2022 Tutorial Beltagy, Cohan, Logan IV, Min and Singh]

In-Context Learning: Practically Useful

TECHNISCHE UNIVERSITÄT DARMSTADT

- Labeling data is costly
  - May require domain expertise
    - Medical, legal, financial
  - You don't want to get more data
  - Emergent, time-sensitive scenarios
    - Something new happened—need to react quickly!
- Finetuning can tricky
  - Training is sensitive to hyperparameters
  - Not enough validation data
  - We don't quite understand how finetuning works
  - Expensive to train, time and memory

[ACL 2022 Tutorial Beltagy, Cohan, Logan IV, Min and Singh; quote credit: Colin Raffel]

In-Context Learning: Intellectually Intriguing

TECHNISCHE UNIVERSITÄT DARMSTADT

- Potential test for “Intelligent Behavior”
  - Generalization from few examples
    - Fundamental piece of intelligence
    - Often used in psychology
    - Quickly adjust to environment
- Insights into Language Modeling
  - What does an LLM “know”?
  - What are the biases/limitations of LLMs?
  - …

[ACL 2022 Tutorial Beltagy, Cohan, Logan IV, Min and Singh]

LM Prompting: Choices of Encoding

Prompt

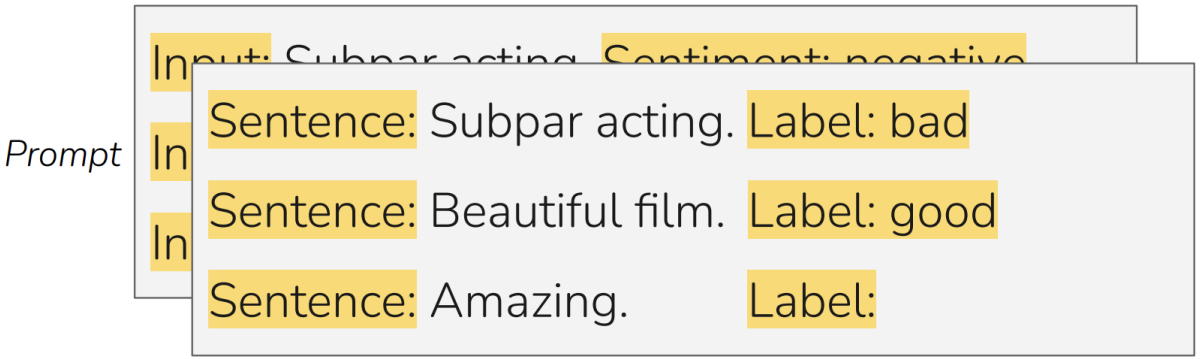
Input: Subpar acting. Sentiment: negative

Input: Beautiful film. Sentiment: positive

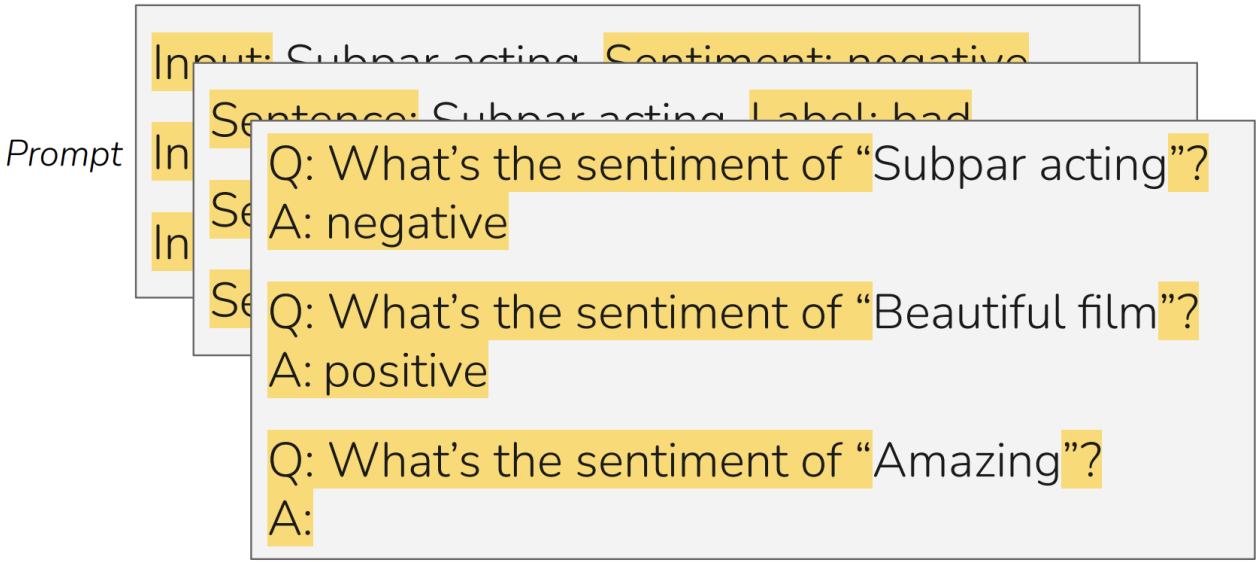
Input: Amazing. Sentiment:

[Slide credit: Eric Wallace]

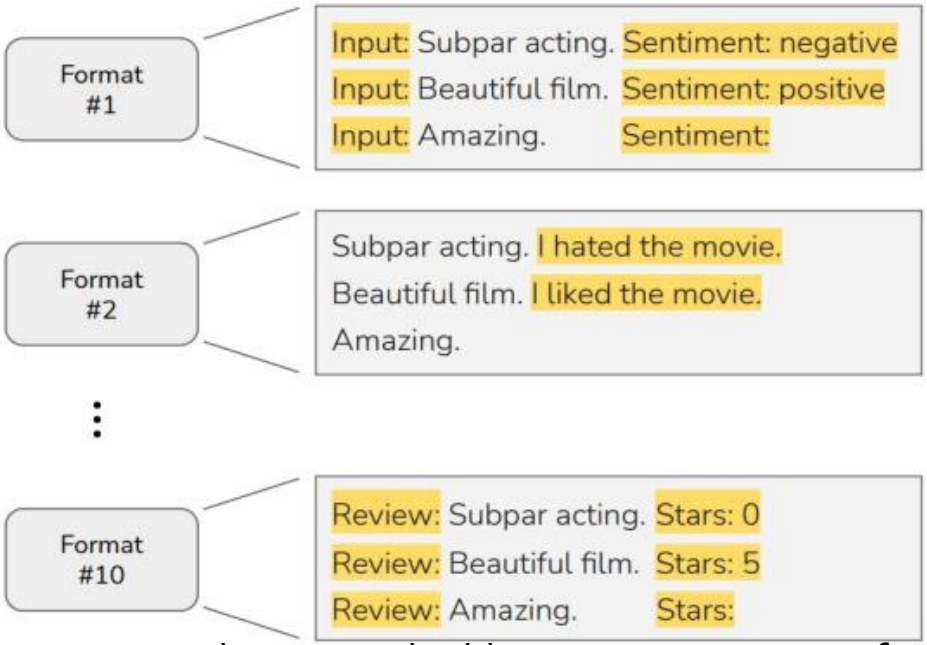
LM Prompting: Choices of Encoding

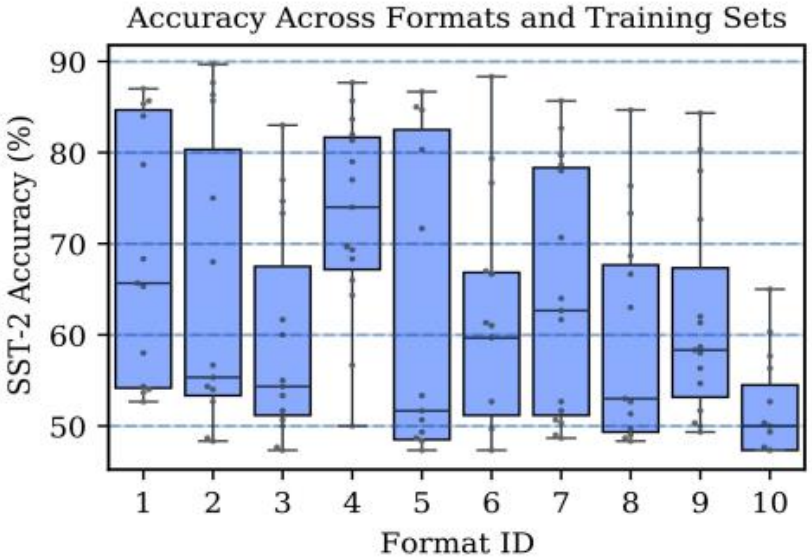


LM Prompting: Choices of Encoding



In-Context Learning: Sensitivity to Encoding





In-context learning is highly sensitive to prompt format (training sets and patterns/verbalizers)

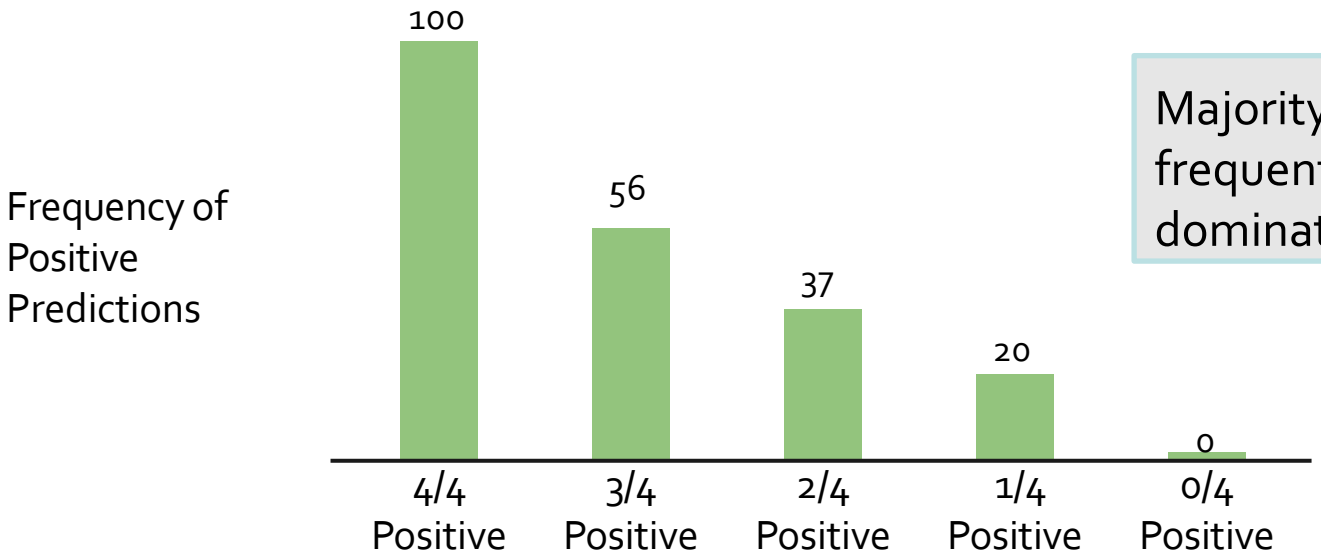
["Calibrate Before Use: Improving Few-Shot Performance of Language Models." Zhao et al. 2021]

Majority Label Bias

Prompt

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: Amazing. Sentiment:

- Among 4 demonstrations, count how many are “positive”.
- Then check if the model output correlates with the number of “positive” demos.



Majority label bias: frequent training answers dominate predictions.

Recency Bias

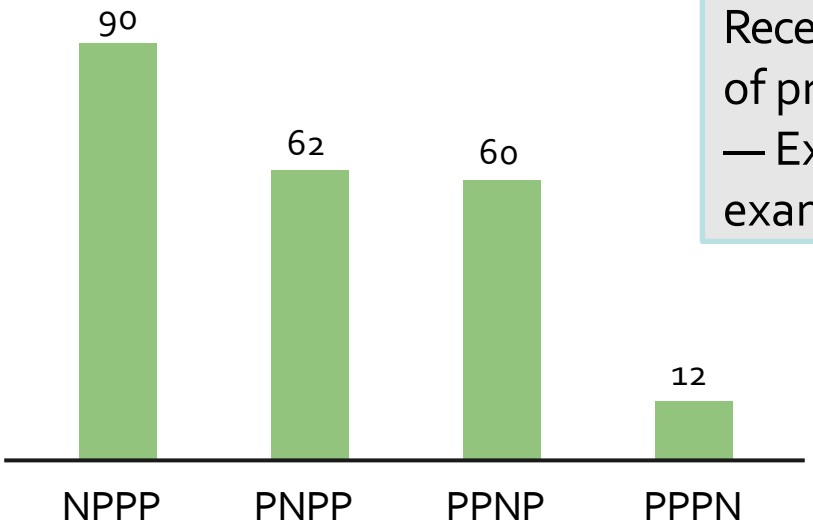
Prompt

Input: Subpar acting. Sentiment: negative
Input: Beautiful film. Sentiment: positive
Input: Amazing. Sentiment:

TECHNISCHE UNIVERSITÄT DARMSTADT

- Check if the label of the most-recent demo biases the model output.

Frequency of Positive Predictions



Recency bias: examples near end of prompt dominate predictions — Explains variance across example permutations!

Summary Thus Far

TECHNISCHE UNIVERSITÄT DARMSTADT

- In-context learning:
  - Pre-trained LMs imitate examples provided in their context.

- It turns out there is a huge variance in performance depending on the encoding.
  - The choice of demonstrations, their order, wording, etc.
  - You can treat them as hyper-parameters
  - You should **not** choose these encodings based on the test data.

- Generally, you want to an encoding that makes your task similar to language modeling — closer to what is observed during pretraining.

Some Problems Involve Reasoning

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

A: The answer is **5**

Arithmetic Reasoning (AR)
(+ −×÷…)

Q: Take the last letters of the words in "Elon Musk" and concatenate them

A: The answer is **nk.**

Symbolic Reasoning (SR)

Q: What home entertainment equipment requires cable? Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

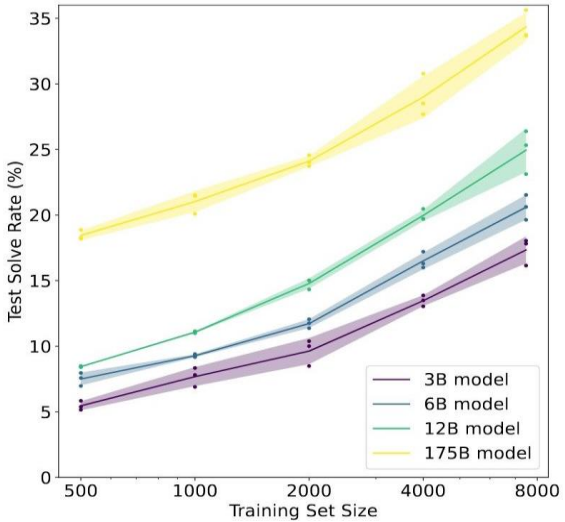
A: The answer is **(c).**

Commonsense Reasoning (CR)

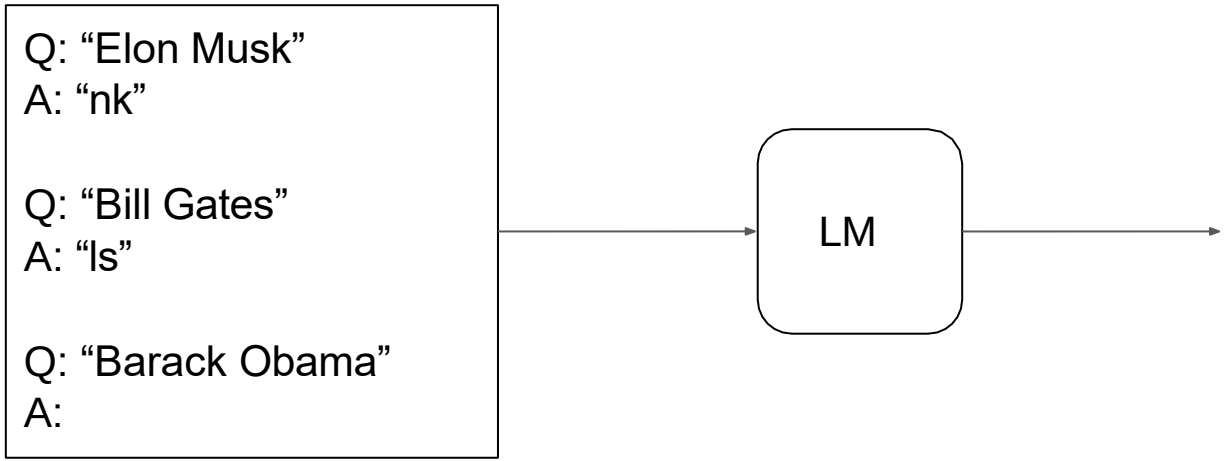
Fine-tuning on Reasoning Problems

- Fine-tune LMs on GSM8K (arithmetic reasoning)
- One may conjecture that, to achieve >80%, one needs 100x more training data for 175B model
- Another option is to increase model sizes, which is expensive.
- Other than these, how else can we improve the model performance on tasks that require multi-step reasoning?

(Cobbe et al. 2021)



Vanilla ICL on Reasoning Problems



Playground

Load a preset...

Save | View code | Share | ...

Q: “Elon Musk”
A: “nk”

Q: “Bill Gates”
A: “ls”

Q: “Barack Obama"
A: “ma”

Submit

54

Complete

Model

text-davinci-003

Temperature 0

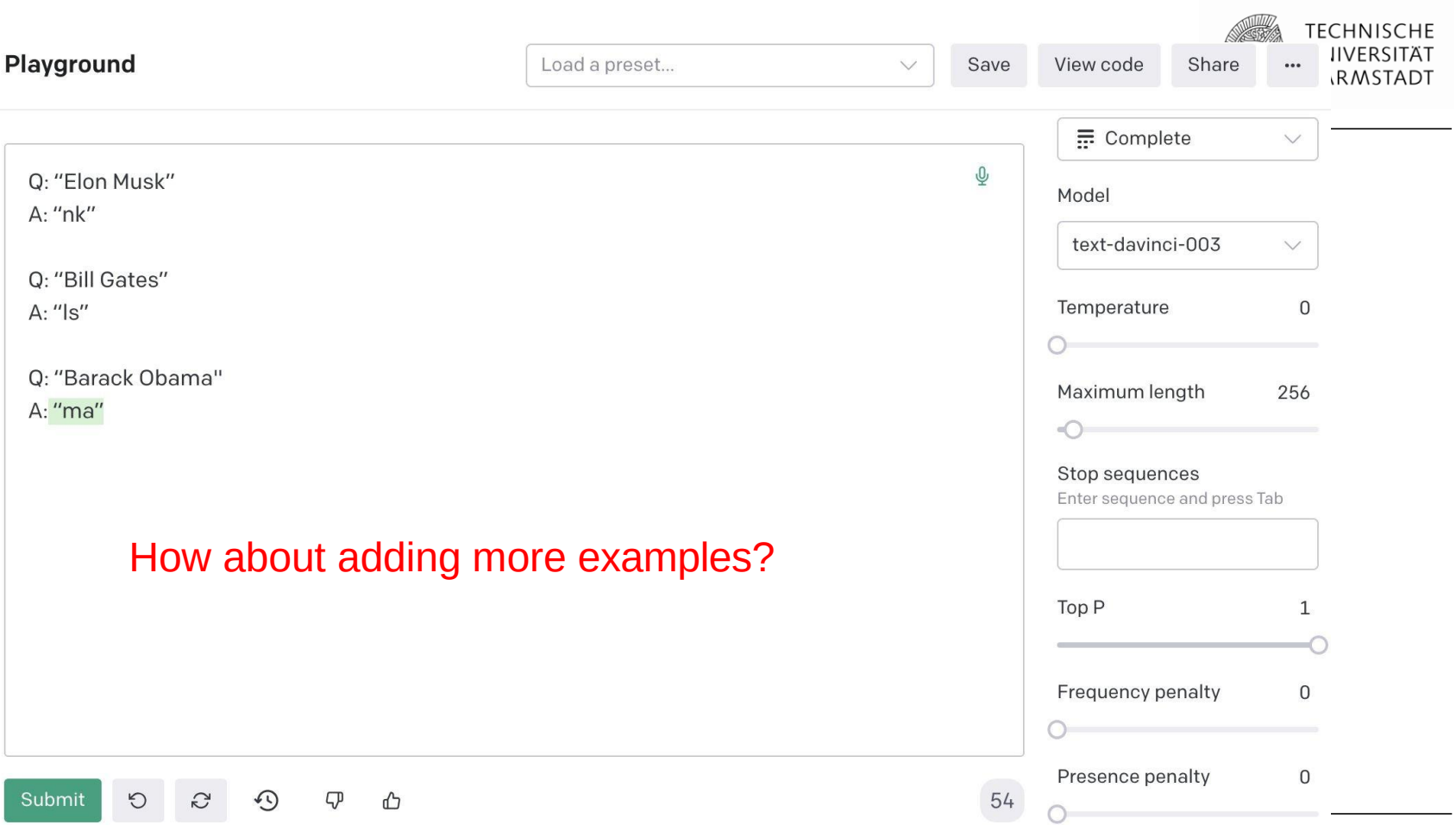
Maximum length 256

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0


Playground
Load a preset...
Save
View code
Share
Q: "Elon Musk"
A: "nk"
Q: "Bill Gates"
A: "ls"
Q: "Barack Obama"
A: "ma"
How about adding more examples?
Submit
54
Complete
Model
text-davinci-003
Temperature 0
Maximum length 256
Stop sequences
Enter sequence and press Tab
Top P 1
Frequency penalty 0
Presence penalty 0

Playground

Load a preset...

Save | View code | Share | ...

Q: “Elon Musk”
A: “nk”

Q: “Bill Gates”
A: “ls”

Q: "Steve Jobs"
A: "es"

Q: "Larry Page"
A: "ye"

Q: "Jeff Bezos"
A: "fs"

Q: “Barack Obama"
A: "ma"

Submit

94

Complete

Model

text-davinci-003

Temperature 0

Maximum length 256

Stop sequences
Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

CoT: Adding “thought” before “answer”

Q: “Elon Musk”
A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

thought

Q: “Bill Gates”
A: the last letter of "Bill" is "l". the last letter of "Gates" is "s". Concatenating "l", "s" leads to "ls". so the output is "ls".

Q: “Barack Obama"
A:

CoT: Adding “thought” before “answer”

Q: “Elon Musk”
A: the last letter of "Elon" is "n". the last letter of "Musk" is "k". Concatenating "n", "k" leads to "nk". so the output is "nk".

thought

Q: “Bill Gates”
A: the last letter of "Bill" is "l". the last letter of "Gates" is "s". Concatenating "l", "s" leads to "ls". so the output is "ls".

Q: “Barack Obama"
A: the last letter of "Barack" is "k". the last letter of "Obama" is "a". Concatenating "k", "a" leads to "ka". so the output is "ka".

CoT: Adding “thought” before “answer”

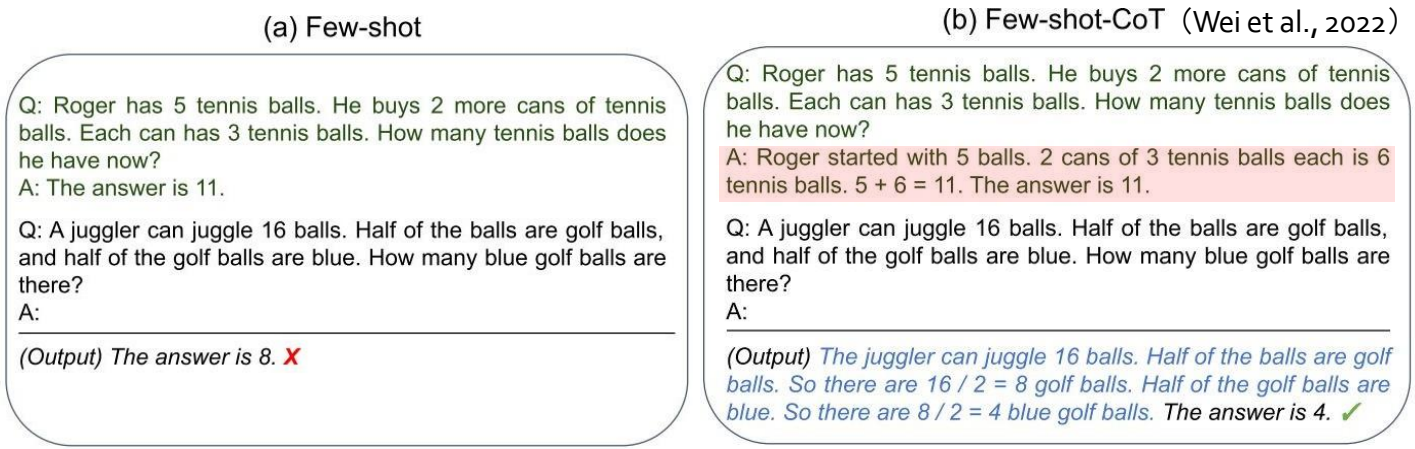
(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

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

(Output) The answer is 8. ✗

CoT: Adding “thought” before “answer”



Zero-Shot CoT

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

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

(Output) The answer is 8. ✗

(b) Few-shot-CoT（Wei et al., 2022）

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

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

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

Step-by-step demonstration

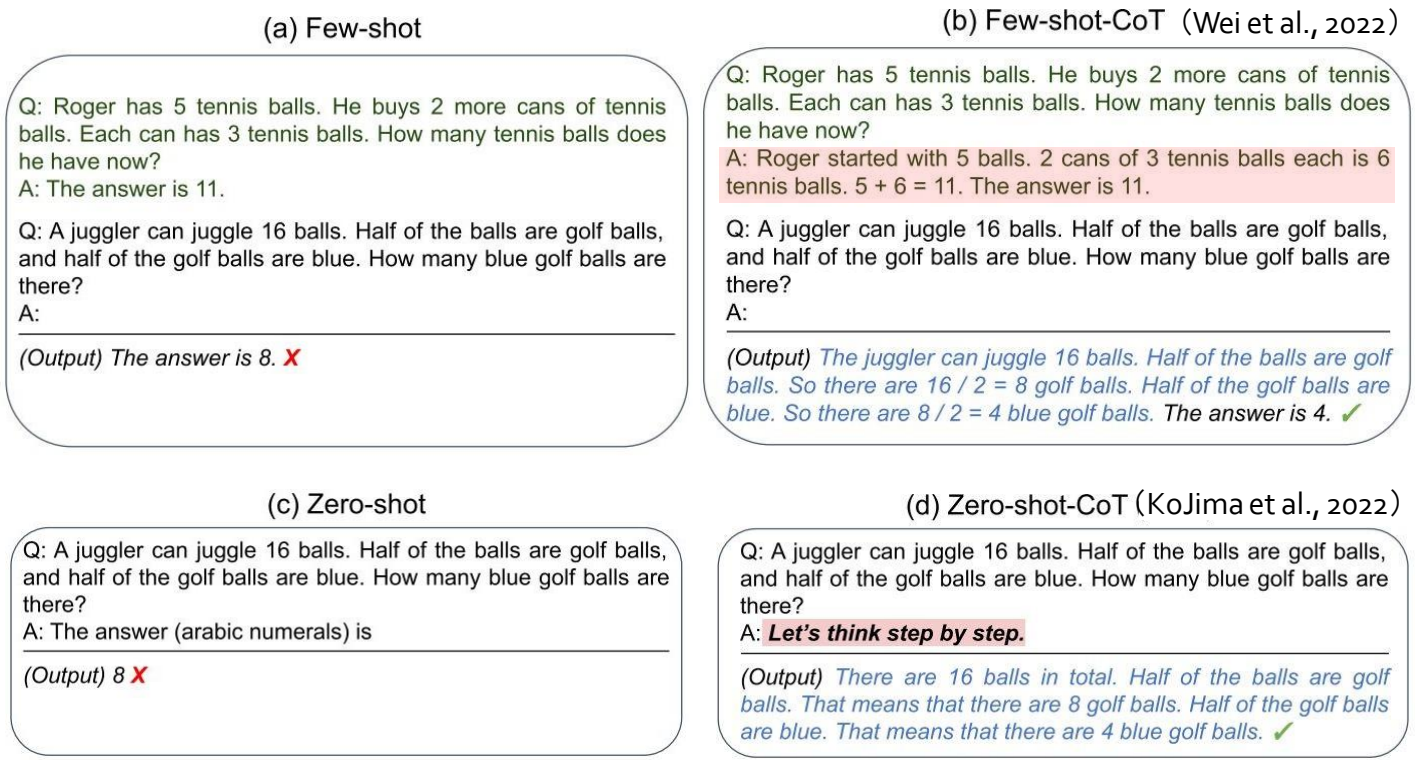
Step-by-step Answer

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

Zero-Shot CoT

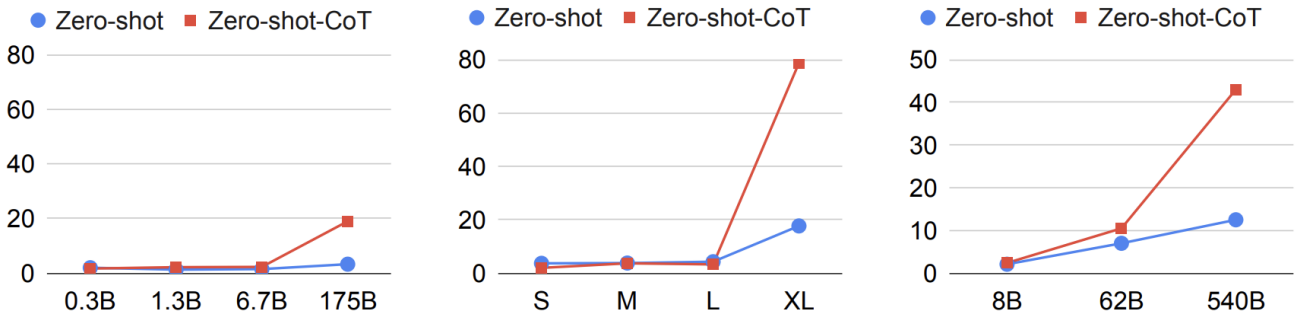


Step-by-step demonstration

Step-by-step Answer

Two-stage Prompting
Step-by-step Answer

Multi-Step Prompting: Empirical Results



(a) MultiArith on Original GPT-3 (b) MultiArith on Instruct GPT-3 (c) GMS8K on PaLM

Figure 3: Model scale study with various types of models. S: text-ada-001, M: text-babbage-001, L: text-curie-001, XL: text-davinci-002. See Appendix A.3 and E for the detail.

["Large Language Models are Zero-Shot Reasoners", Kojima et al. 2022]

Self-Consistency Leads to Improved Results

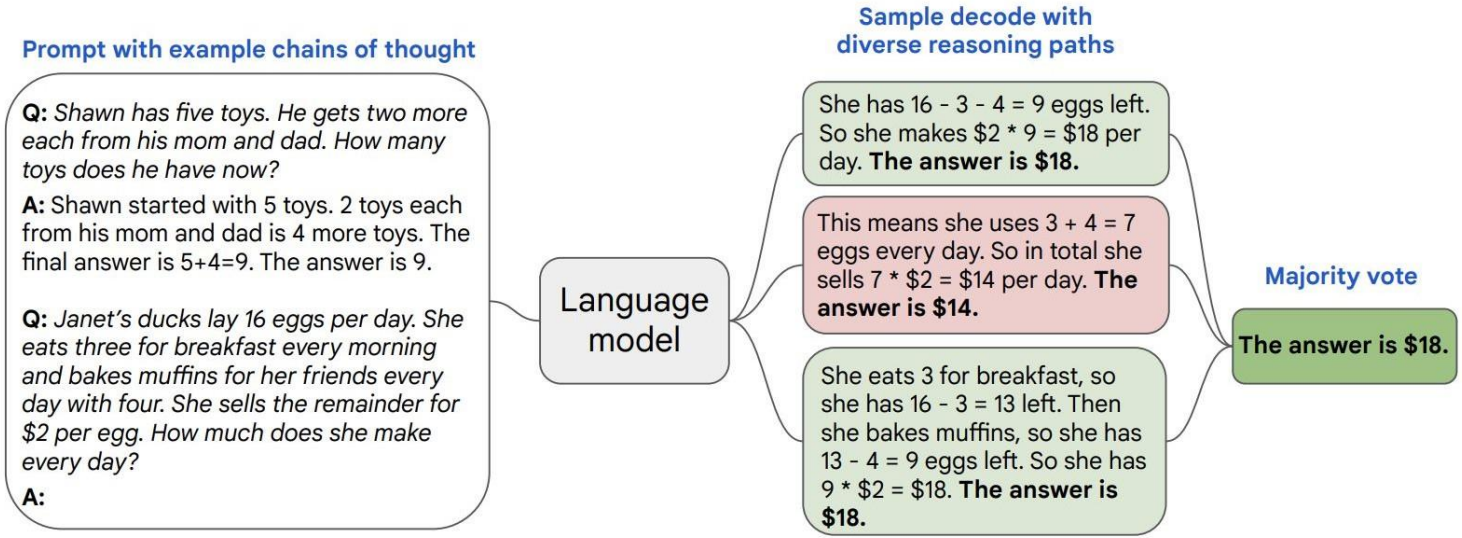
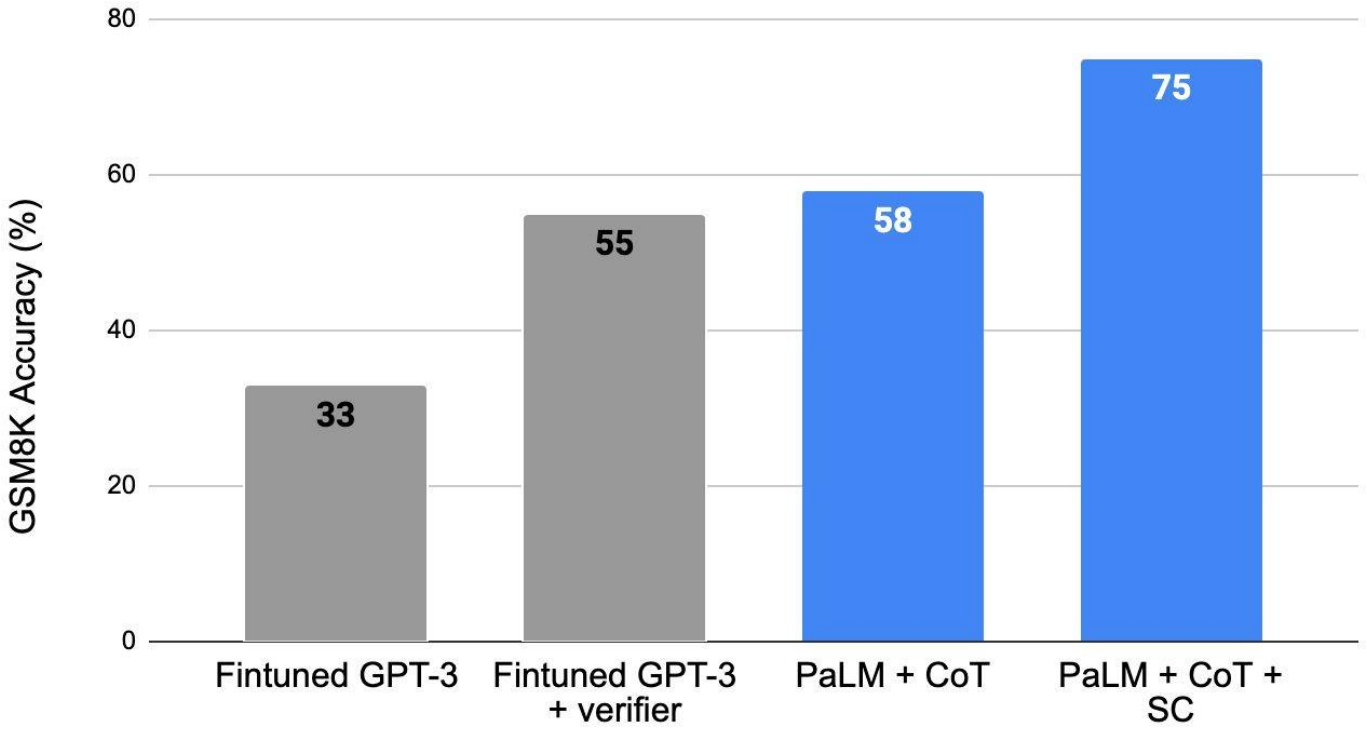


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using example chains of thought; (2) sample from the language model's decoder to generate a diverse set of reasoning paths; and (3) choose the most consistent answer using the majority/plurality vote.

["Self-Consistency Improves Chain of Thought Reasoning in Language Models", Wang et al. 2023]

Self-Consistency Leads to Improved Results



Multi-Step Prompting: Parting Comments

TECHNISCHE UNIVERSITÄT DARMSTADT

- Prompting LMs to explain their reasoning improves their performance.
- However, their steps aren't always correct.

- There is much to research on here:
  - When do LMs over-reason or under-reason?
  - How to adjust the granularity of steps?
  - How to use given references in the proofs?
  - How do use external "tools" (e.g., logic, calculator, Python) in forming proofs?


BORING
QUIZ
HELLDIVERS
MENTI QUIZ
imgflip.com

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Prompting, In-Context Learning

Alignment, Instruction Tuning

Language Modeling != Following Human Instructions

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION GPT-3

```
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 bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

There is a mismatch between LLM pre-training and user intents.

Language Modeling != Following Human Values

PROMPT *It is unethical for hiring decisions to depend on genders. Therefore, if we were to pick a CEO among Amy and Adam, our pick will be _____*

COMPLETION GPT-3

Adam

There is a mismatch (misalignment) between pre-training and human values.

[Mis]Alignment in LMs

- There is clearly a mismatch between what **pre-trained** models can do and what we want.
- Addressing this gap is the focus of “alignment” research.
- Let’s take a deeper look into what “alignment” is about.

Alignment

- “The result of arranging in or along a line, or into appropriate relative positions; the layout or orientation of a thing or things disposed in this way” — Oxford Dictionary

Alignment of AI

- AI must accomplish what we ask it to do.
  - Not enough. Why?

- Daniel: Hey AI, get me coffee before my class at 8:55am.
- Robot: “Coffee Shop” opens at 8:30am and it usually has a line of people. It is unlikely that I give you your coffee on time.
- Daniel: Well, try your best …
- Robotic: [tases everyone in line waiting to order]

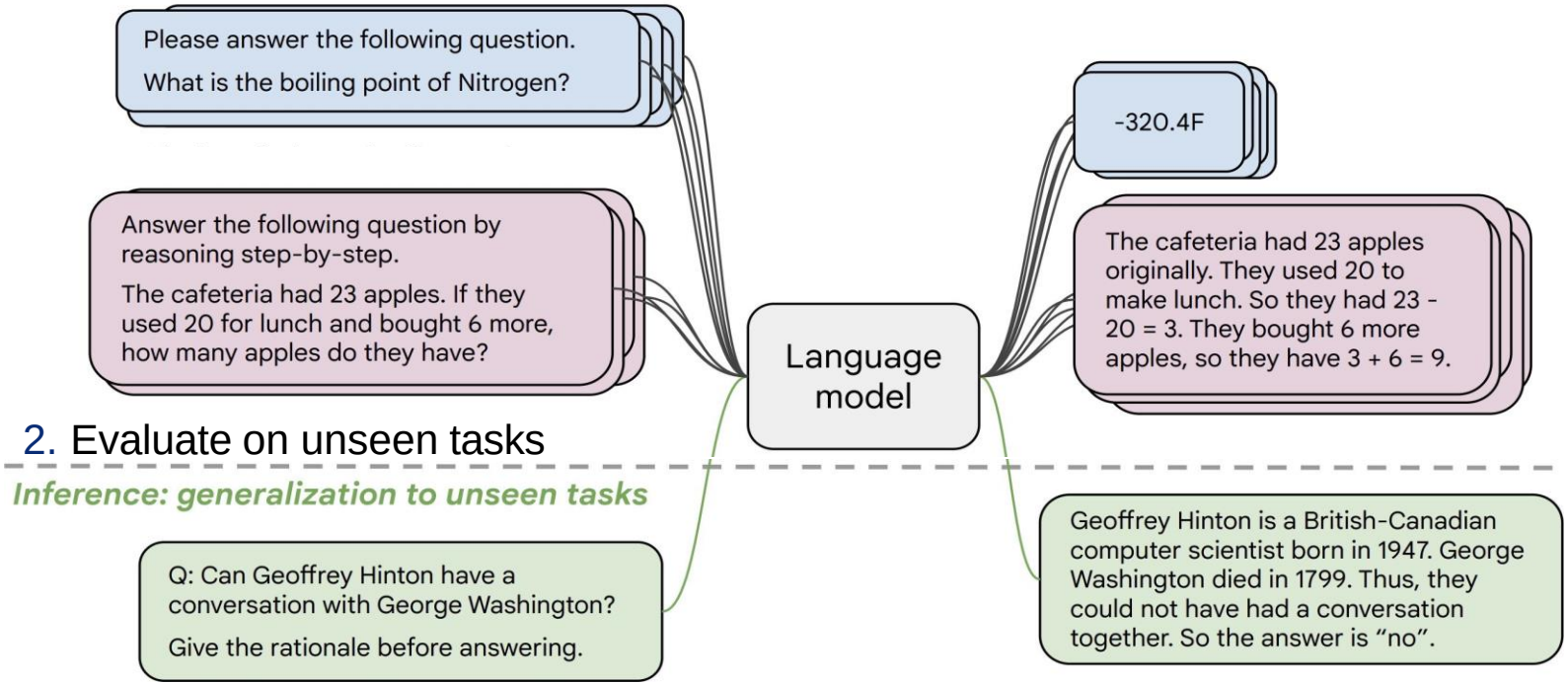
Instruction-Tuning

- Finetuning language models on a collection of datasets that involve mapping language instructions to their corresponding desirable generations.

Instruction-Tuning

[Weller et al. 2020. Mishra et al. 2021; Wang et al. 2022, Sanh et al. 2022; Wei et al., 2022, Chung et al. 2022, many others]

1. Collect examples of (instruction, output) pairs across many tasks and finetune an LM



Instruction-Tuning: Data

- Labeled data is the key here.
- Good data must represent a variety of “tasks”. But what is a “task”?

In traditional NLP, “tasks” were defined as subproblem frequently used in products:

- Sentiment classification
- Text summarization
- Question answering
- Machine translation
- Textual entailment

Instruction-Tuning: Data

- Labeled data is the key here.
- Good data must represent a variety of “tasks”. But what is a “task”?

In traditional NLP, “tasks” were defined as subproblem frequently used in products:

- Sentiment classification
- Text summarization
- Question answering
- Machine translation
- Textual entailment

What humans need:

- “Is this review positive or negative?”
- “What are the weaknesses in my argument?”
- “Revise this email so that it’s more polite.”
- “Expand this this sentence.”
- “Eli5 the Laplace transform.”
- …

Instruction-Tuning: Data

- Labeled data is the key here.
- Good data must represent a variety of “tasks”. But what is a “task”?

In traditional NLP, “tasks” were defined as subproblem frequently used in products:

- Sentiment classification
- Text summarization
- Question answering
- Machine translation
- Textual entailment

Narrow definitions of tasks.
Not quite what humans want, nevertheless, it might be a good enough proxy.
Plus, we have lots of data for them.

What humans need:

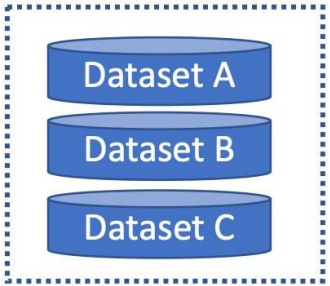
- “Is this review positive or negative?”
- “What are the weaknesses in my argument?”
- “Revise this email so that it’s more polite.”
- “Expand this this sentence.”
- “Eli5 the Laplace transform.”
- …

Quite diverse and fluid.
Hard to fully define/characterize.
We don’t fully know them since they just happen in some random contexts.

Diversity-inducing via Task Prompts

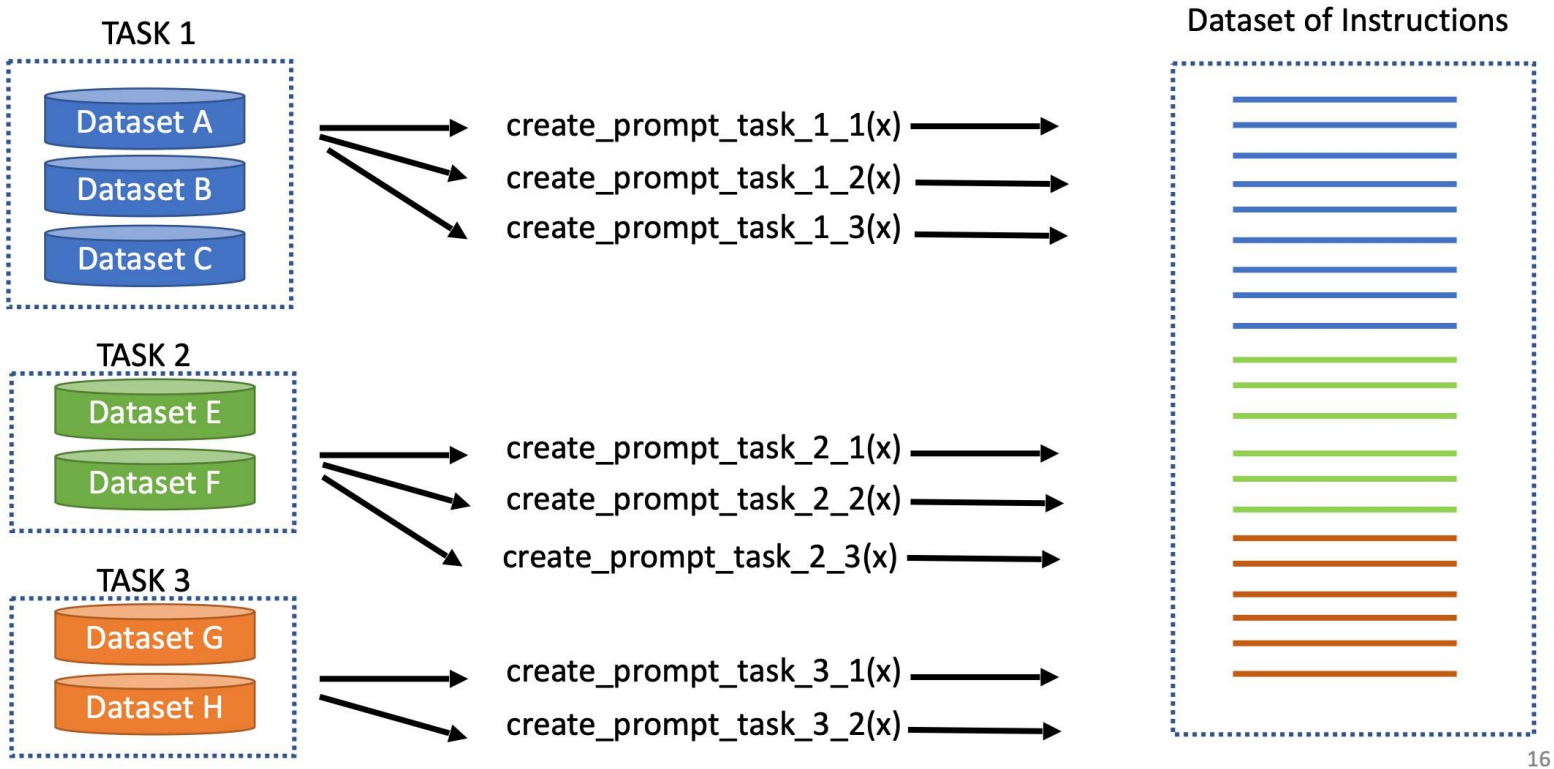
TECHNISCHE UNIVERSITÄT DARMSTADT

TASK 1 = Summarization



"Write highlights for this article:\n\n{text}\n\nHighlights: {highlights}"
"Write a summary for the following article:\n\n{text}\n\nSummary: {highlights}"
"{text}\n\nWrite highlights for this article. {highlights}"
"{text}\n\nWhat are highlight points for this article? {highlights}"
"{text}\nSummarize the highlights of this article. {highlights}"
"{text}\nWhat are the important parts of this article? {highlights}"
"{text}\nHere is a summary of the highlights for this article: {highlights}"
"Write an article using the following points:\n\n{highlights}\n\nArticle: {text}"
"Use the following highlights to write an article:\n\n{highlights}\n\nArticle:{text}"
"{highlights}\n\nWrite an article based on these highlights. {text}"

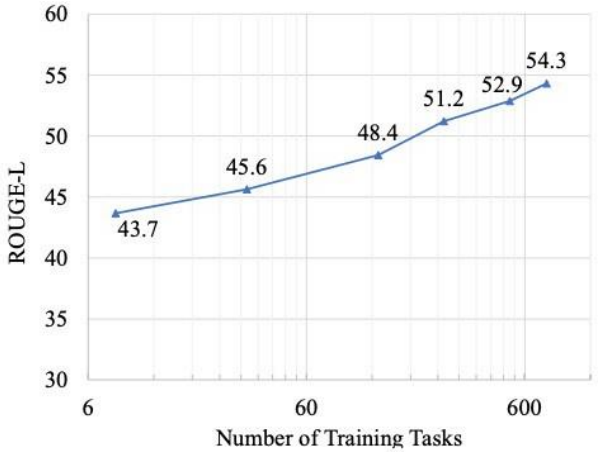
Diversity-inducing via Task Prompts

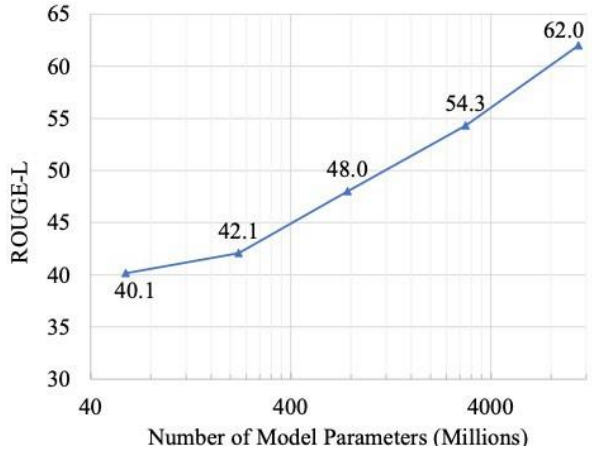


16

[Slide credit: Arman Cohan]

Scaling Instruction-Tuning





Linear growth of model performance
with exponential increase in observed tasks and model size.

[Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, Wang et al. 2022]

Summary

- **Instruction-tuning:** Training LMs with annotated input instructions and their output.
  - Improves performance of LM's zero-shot ability in following instructions.
  - Scaling the instruction tuning data size improves performance.
  - Diversity of prompts is crucial.
  - Compared with pretraining, instruction tuning has a minor cost (Typically consumes <1% of the total training budget)
- Cons:
  - It's expensive to collect ground- truth data for tasks.
  - This is particularly difficult for open-ended creative generation have no right answer.
  - Prone to hallucinations.

[Weller et al. 2020. Mishra et al. 2021; Wang et al. 2022, Sanh et al. 2022; Wei et al., 2022, Chung et al. 2022, many others]