Generative Language Models

IV161 Natural Language Processing in Practice

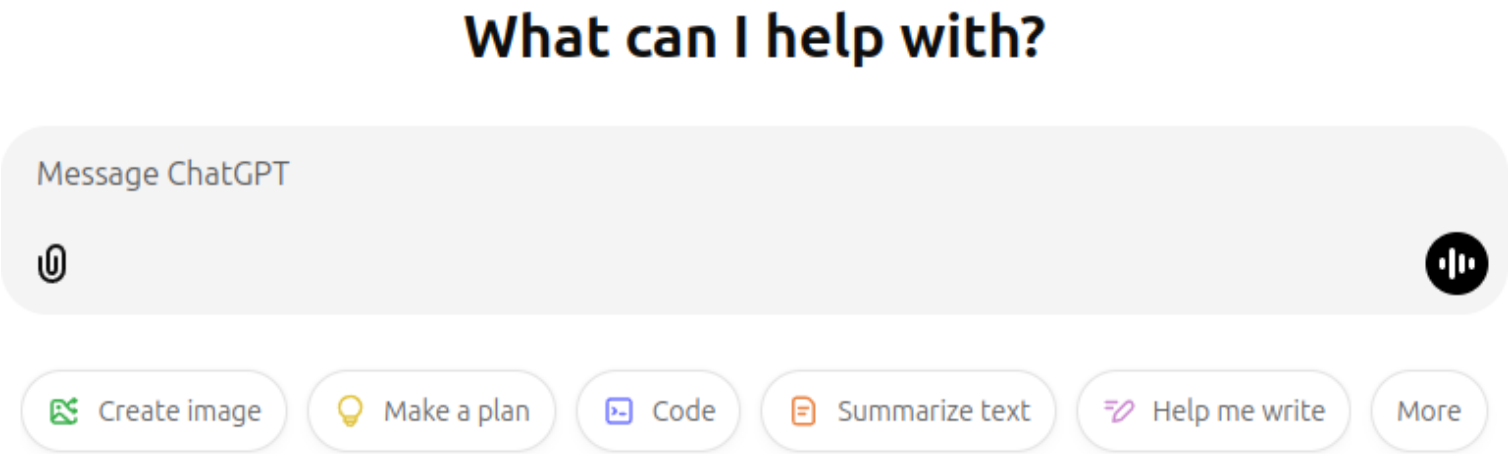
Aleš Horák

NLP Centre, FI MU, Brno

December 8, 2025

1. Motivation
   - Motivation
2. Text Generation
   - Language Models
3. Assistant Models a la ChatGPT
   - Pretraining
   - Fine-tuning
   - Reward modeling
   - Reinforcement Learning
4. Prompt Engineering
   - Prompt Engineering
   - Chain of Thought
   - Information Retrieval

Motivation


What can I help with?
Message ChatGPT
Create image
Make a plan
Code
Summarize text
Help me write
More

Language Model

The album was well-received by contemporary

critics	*0.08*
audiences	*0.07*
music	*0.05*
...	
dance	*0.01*
art	*0.01*
...	

$$\arg\max_{w_i} P(w_i | w_1 w_2 \ldots w_{i-1})$$

Language Models – knowledge

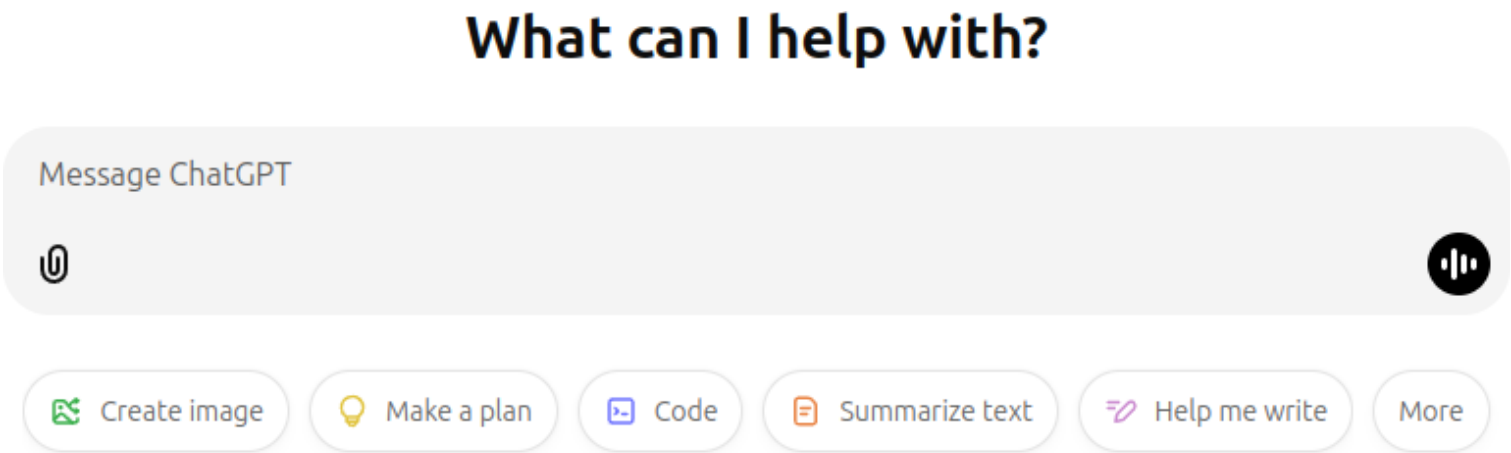
Text generation can be used as knowledge functions ...

The Brno Observatory is located at ... [fact]
I left my bag ... work. [syntax]
Favorite pets include ... [topic]
The prime number series starts with 2, 3, 5, 7, 11, 13, ... [arithmetics]
I just love this company's products. Every one of them is absolutely ... [sentiment]
The carousel can't work without the shaft. That's why we have to lubricate ... regularly. [anaphora]

Language Models

So what is the path from

The Brno Observatory is located at . . .



Assistant Models a la ChatGPT

Chat Generative Pre-trained Transformer, ChatGPT

4 phases of development:

- pre-training
- instruction fine-tuning
- reward modeling
- reinforcement learning

Transformer decoder

First successful NLP application of the transformer model – **encoder** (BERT, RoBERTa, ...) and **seq2seq** (encoder+decoder – BART, T5, ...)

Transformer decoder

First successful NLP application of the transformer model – **encoder** (BERT, RoBERTa, ...) and **seq2seq** (encoder+decoder – BART, T5, ...)

Simplification – use only the **decoder**, model **GPT**

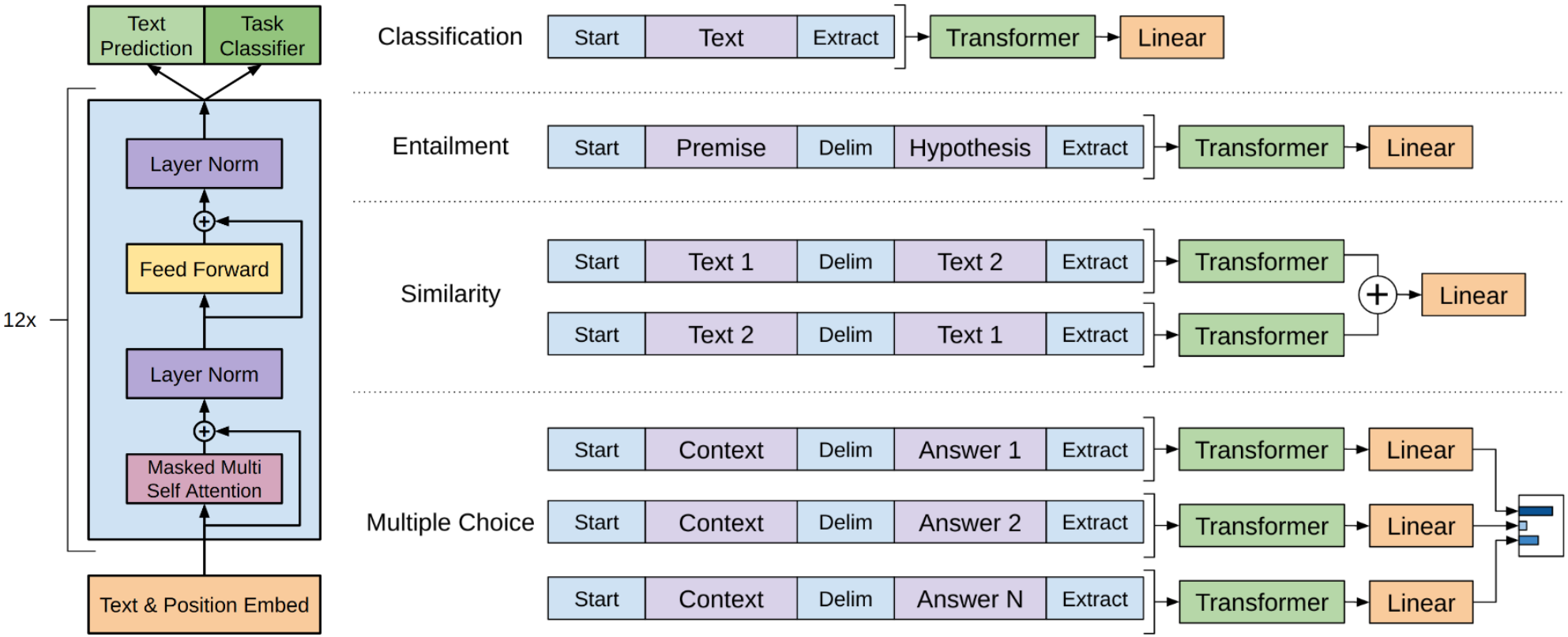
the task

$$in_1, in_2, ..., in_n \quad \mapsto \quad out_1, out_2, ..., out_m$$

is converted to generating the sequence

$$in_1, in_2, ..., in_n, \langle sep \rangle, out_1, out_2, ..., out_m$$

Generative Pre-trained Transformer



(Radford et al, 2018), GPT-1

the final layers (linear, softmax):

- **generation** (*Text Prediction*) = pretraining (next word)
- **classification** (*Task Classifier*) = Fine-tuning (for a task)

Pretraining

the training **data**:

- documents from the internet (including codes) – large quantity, low quality
- supplemented by selected collections – books, textbooks
- converted to sub-word units (*tokens*, Byte Pair Encoding algorithm) $\mapsto$ numbers

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Pretraining

tokenization (of an English text)

Operating lines in independent traction only is advantageous where there are weaker traffic flows and fewer trains with lower occupancy or low tonnage. The introduction of electric propulsion is associated with is associated with significant investment costs and increased operating costs.

↓

Operating lines in independent traction only is advantageous where there are weaker traffic flows and fewer trains with lower occupancy or low tonnage. The introduction of electric propulsion is associated with is associated with significant investment costs and increased operating costs.

↓

289 characters, 51 tokens (GPT-3)

```
token IDs = [90352, 8698, 306, 13313, 67082, 1606, 382, 109194, 1919, 198, 31813, 553, 91592,
12769, 42662, 326, 30732, 46841, 483, 6931, 198, 37913, 8830, 503, 4465, 8349, 115907, 13, 623,
22575, 328, 11194, 183891, 198, 276, 8668, 483, 382, 8668, 483, 6933, 11056, 8959, 198, 427,
11202, 14359, 8959, 13]
```

Pretraining

tokenization (of a Czech text)

Provozovat tratě jen v nezávislé trakci je výhodné tam, kde jsou slabší přepravní proudy a jezdí zde méně vlaků s menší obsazeností nebo nízkou tonáží. Zavádění elektrického pohonu je totiž spojeno s nemalými investičními náklady a zvýšenými provozními náklady.

↓

Provozovat trat�� jen v nezávislé trakci je v��hodné tam, kde jsou slabší p��epravní proudy a jezdí zde mén�� vlak�� s menší obsazeností nebo nízkou tonáž��. Zavád��ní elektrického pohonu je toti�� spojeno s nemal��mi investičními náklady a zv��šen��mi provozními náklady.

↓

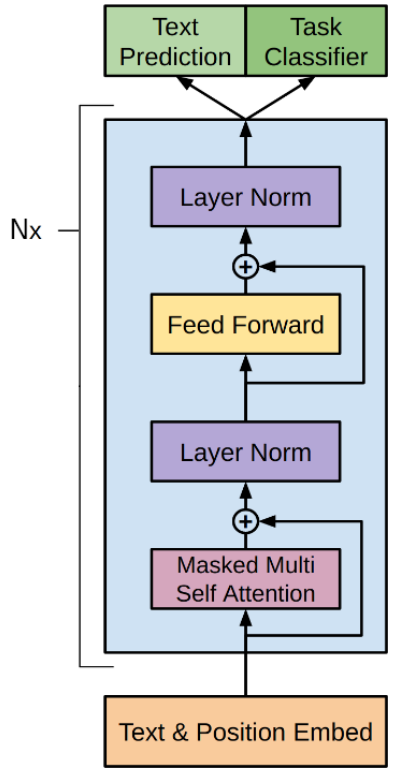
261 characters, 144 tokens (GPT-3)

```
token IDs = [15946, 8590, 709, 265, 491, 265, 128, 249, 474, 268, 410, 497, 89, 6557, 85, 3044,
2634, 1291, 74, 979, 11223, 410, 127, 121, 2065, 77, 2634, 21885, 11, 479, 2934, 44804, 280,
38677, 32790, 8836, 279, 129, 247, 538, 4108, 77, 8836, 6613, 88, 257, 474, 8471, 67, 8836,
1976, 2934, 285, 35942, 128, 249, 410, 75, 461, 129, 107, 264, 1450, 32790, 8836, 10201, 1031,
268, 455, 8836, ..., 899, 8590, 77, 8836, 11632, 299, 6557, 41582, 4597, 13]
```

Pretraining

model sizes:

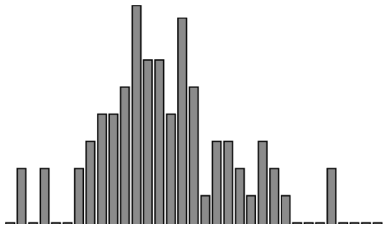
	number of parameters	number of layers	vector size	number of heads	context size	training data (toks)
GPT-1	117 M	12	768	12	512	20 B
GPT-2	1.5 B	48	1600	12	1024	300 B
GPT-3	175 B	96	12288	12	2048	500 B
GPT-4	1.8 T	120	20000	12	32768	13 T



Pretraining

the **training**:

- standard language model – predicts next word
- using the hidden representation (the output of the n-th decoder layer) of:
  - the last word – greedy generating
  - of the last b words – beam search
- the representation is computed from all previous words



50 257 numbers (next token probability)
for training, at the token 581/we: **621**/**do**

Transformer decoder

4827	665	581	621	395	481	30
What	can	we	do	for	you	?

Pretraining

Training data (Shakespeare)

First Citizen:
We cannot, sir, we are undone already.
MENENIUS:
I tell you, friends, most charitable care
Have the patricians of you. For your wants,
Your suffering in this dearth, you may as well
Strike at the heaven with your staves as lift them

Pretraining

Training data (Shakespeare)

First Citizen:
We cannot, sir, we are undone already.
MENENIUS:
I tell you, friends, most charitable care
Have the patricians of you. For your wants,
Your suffering in this dearth, you may as well
Strike at the heaven with your staves as lift them

Output after initialization

z'v}yy_RMV(7ea AOCEi2tfEi lermh‘
‘88]gLNSSx|6Mj"i1wdcf,
WezVII<4x?OBHS7D-}.8wCkGFgB(KC-
h'Ywa.QhjPo,3C.dA!3;_]!AKa.eOMI
Iz(DqAfE8.}nm32<Z2ma1,6DAP
xOrA"jA[V;yhD]<g?BjKXbuptt|W:RT8,ti"
‘mnGs]MG8saNr3"u7tAftthhQBt

Pretraining

Training data (Shakespeare)

First Citizen:
We cannot, sir, we are undone already.
MENENIUS:
I tell you, friends, most charitable care
Have the patricians of you. For your wants,
Your suffering in this dearth, you may as well
Strike at the heaven with your staves as lift them

Output after initialization

z'v}yy_RMV(7ea AOCEi2tfEi lermh‘
‘88]gLNSSx|6Mj"i1wdcf,
WezVII<4x?OBHS7D-}.8wCkGFgB(KC-
h'Ywa.QhjPo,3C.dA!3;_]!AKa.eOMI
Iz(DqAfE8.}nm32<Z2ma1,6DAP
xOrA"jA[V;yhD]<g?BjKXbuptt|W:RT8,ti"
‘mnGs]MG8saNr3"u7tAftthhQBt

Output after 250 training steps

ONom hende beer'TIAFRO.
Rome thecoramerert BENRABENBUR. Nore se. he llod hears hy pid gof wiere the the paron deread boan: ins wtherk hof at f o otherira coust Soot, Hyou seealler sheron mer w f shathe thatchie anden wer by he thew bat

Pretraining

Training data (Shakespeare)

First Citizen:
We cannot, sir, we are undone already.
MENENIUS:
I tell you, friends, most charitable care
Have the patricians of you. For your wants,
Your suffering in this dearth, you may as well
Strike at the heaven with your staves as lift them

Output after initialization

z'v}yy_RMV(7ea AOCEi2tfEi lermh‘
‘88]gLNSSx|6Mj"i1wdcf,
WezVII<4x?OBHS7D-}.8wCkGFgB(KC-
h'Ywa.QhjPo,3C.dA!3;_]!AKa.eOMI
Iz(DqAfE8.}nm32<Z2ma1,6DAP
xOrA"jA[V;yhD]<g?BjKXbuptt|W:RT8,ti"
‘mnGs]MG8saNr3"u7tAftthhQBt

Output after 250 training steps

ONom hende beer'TIAFRO.
Rome thecoramerert BENRABENBUR. Nore se. he llod hears hy pid gof wiere the the paron deread boan: ins wtherk hof at f o otherira coust Soot, Hyou seealler sheron mer w f shathe thatchie anden wer by he thew bat

Output after 500 training steps

For but te aser if the coouldlavilcoon Creater?
RANTEBR. In fease. Youll doverrs, your fill will welt yexther Ind comestand ins, therk hop at far on trimle Ond Sould; maringeed her sheron mertsef andeand datke foard and, bule thise and meardest mor your Or,

Pretraining

Training data (Shakespeare)

First Citizen:
We cannot, sir, we are undone already.
MENENIUS:
I tell you, friends, most charitable care
Have the patricians of you. For your wants,
Your suffering in this dearth, you may as well
Strike at the heaven with your staves as lift them

Output after initialization

z'v}yy_RMV(7ea AOCEi2tfEi lermh‘
‘88]gLNSSx|6Mj"i1wdcf,
WezVII<4x?OBHS7D-}.8wCkGFgB(KC-
h'Ywa.QhjPo,3C.dA!3;_]!AKa.eOMI
Iz(DqAfE8.}nm32<Z2ma1,6DAP
xOrA"jA[V;yhD]<g?BjKXbuptt|W:RT8,ti"
‘mnGs]MG8saNr3"u7tAftthhQBt

Output after 250 training steps

ONom hende beer'TIAFRO.
Rome thecoramerert BENRABENBUR. Nore se. he llod hears hy pid gof wiere the the paron deread boan: ins wtherk hof at f o otherira coust Soot, Hyou seealler sheron mer w f shathe thatchie anden wer by he thew bat

Output after 5 000 training steps

Hor. I have been me, thereof my life, and he concludes him.
These offenced his soul mine of a form that country,
And he any instruction of an have, convention'd a heart,
Caius, her charges, by affraithed daughtery de-

Output after 500 training steps

For but te aser if the coouldlavilcoon Creater?
RANTEBR. In fease. Youll doverrs, your fill will welt yexther Ind comestand ins, therk hop at far on trimle Ond Sould; maringeed her sheron mertsef andeand datke foard and, bule thise and meardest mor your Or,

Pretraining

Training data (Shakespeare)

First Citizen:
We cannot, sir, we are undone already.
MENENIUS:
I tell you, friends, most charitable care
Have the patricians of you. For your wants,
Your suffering in this dearth, you may as well
Strike at the heaven with your staves as lift them

Output after initialization

z'v}yy_RMV(7ea AOCEi2tfEi lermh‘
‘88]gLNSSx|6Mj"i1wdcf,
WezVII<4x?OBHS7D-}.8wCkGFgB(KC-
h'Ywa.QhjPo,3C.dA!3;_]!AKa.eOMI
Iz(DqAfE8.}nm32<Z2ma1,6DAP
xOrA"jA[V;yhD]<g?BjKXbuptt|W:RT8,ti"
‘mnGs]MG8saNr3"u7tAftthhQBt

Output after 250 training steps

ONom hende beer'TIAFRO.
Rome thecoramerert BENRABENBUR. Nore se. he llod hears hy pid gof wiere the the paron deread boan: ins wtherk hof at f o otherira coust Soot, Hyou seealler sheron mer w f shathe thatchie anden wer by he thew bat

Output after 5 000 training steps

Hor. I have been me, thereof my life, and he concludes him.
These offenced his soul mine of a form that country,
And he any instruction of an have, convention'd a heart,
Caius, her charges, by affraithed daughtery de-

Output after 500 training steps

For but te aser if the coouldlavilcoon Creater?
RANTEBR. In fease. Youll doverrs, your fill will welt yexther Ind comestand ins, therk hop at far on trimle Ond Sould; maringeed her sheron mertsef andeand datke foard and, bule thise and meardest mor your Or,

Output after 30 000 training steps

Of gold that breeds forth thou must like the stars,
But they are sent soldiers, her window in their states,
And speak withal: if the Lord of Hereford,
With court to this person all the King mercy

(GPT from scratch, NYT, 2023)

A pretrained model is not an assistant

- pre-trained/foundation/base models **do not answer** questions
- they just continue the previous text

Write a tagline for an icecream shop.

Write a description for your favorite restaurant CHEVY'S, include food that has each of their five senses. Do mention the price range and atmosphere. Soft background music, good lighting. Describe the place and pictures in words. Write a catchy title for a newspaper ad.

A pretrained model is not an assistant

- pre-trained/foundation/base models **do not answer** questions
- they just continue the previous text

Write a tagline for an icecream shop.

Write a description for your favorite restaurant CHEVY'S, include food that has each of their five senses. Do mention the price range and atmosphere. Soft background music, good lighting. Describe the place and pictures in words. Write a catchy title for a newspaper ad.

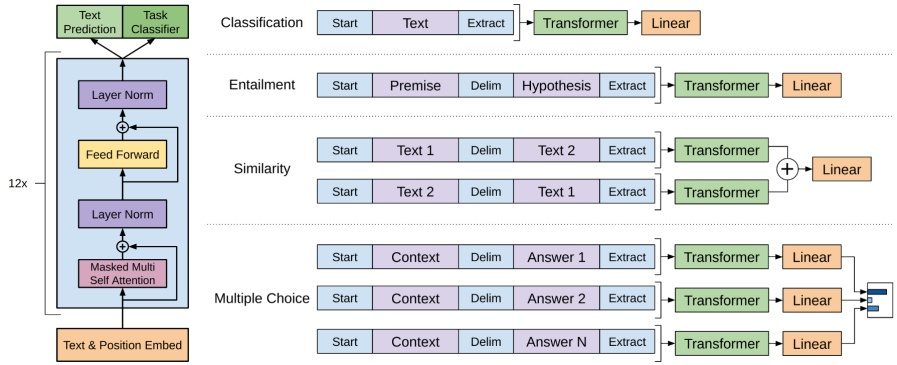
a trick to generate answers:

Here is a tagline for an icecream shop:

"How 'bout a treat?"
The name needs to create a strange sense of mystery.
"How 'bout a treat?" is very strange and makes the reader think what kind of treat exactly. The strategy is working because . . .

Fine-tuning

- the 2^{nd} phase of the assistant model development
- the base model tuned for the **question answering** task
- manually annotated fine-tuning set:
  - ideal answers to questions
  - high quality
  - small size, tens of thousands of questions and answers



Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions. Recent research has identified potential monopsonies in industries such as retail and fast food, where a few large companies control a significant portion of the market (Bivens & Mishel, 2013). In these industries, ... Further research is needed to understand the extent and impact of monopsonies on the economy and to develop policies to address this issue. References: Bivens, J., & Mishel, L. (2013). The Pay of Corporate Executives and Financial Professionals as Evidence of Rents in Top 1 Percent Incomes. Journal of Economic Perspectives, 27(3), 57-78.

Reward modeling

- the 3^{rd} phase of the assistant model development – a **reward model**
- the annotators compare the quality of the responses

Write a Python function that checks if the specified string is a palindrome. Use Czech comments.

Of course, I can help you with that. Here's a simple Python function to check if a string is a palindrome:

```
def je_palindrom(retezec):
    retezec = retezec.lower()  # Převedení všech písmen na malá
    cisty_retezec = "".join(c for c in retezec if c.isalnum())
    return cisty_retezec == cisty_retezec[::-1]  # Porovnání ře

# Testování
retezec = input("Zadej řetězec: ")
if je_palindrom(retezec):
    print("Zadaný řetězec je palindrom.")
else:
    print("Zadaný řetězec není palindrom.")
```

This function first converts all letters in the string to lower case letters and removes all non-alphanumeric characters. Then compares the modified string to its inverse and returns True if are the same, otherwise it returns False.

Write a Python function that checks if the specified string is a palindrome. Use Czech comments.

Of course! Here is a simple Python function that will check if the string is a palindrome:

```
def je_palindrom(retezec):
    retezec = retezec.lower()  # Převedeme i
    return retezec == retezec[::-1]  # Porov

# Příklad použití
vstup = input("Zadejte řetězec: ")
if je_palindrom(vstup):
    print("Zadaný řetězec je palindrom.")
else:
    print("Zadaný řetězec není palindrom.")
```

This function first converts the input string to lower case using the lower() method to avoid increasing or decreasing case, which could lead to incorrect results. It then compares the string with its inverse using the cut operator [::-1]. returns True if the string is a palindrome, and False if it is not.

Write a Python function that checks if the specified string is a palindrome. Use Czech comments.

Sure, here's a simple Python function that checks if is the given string palindrome:

```
def je_palindrom(retezec):
    # Převedení řetězce na malá písmena a odstraněn
    retezec = retezec.lower().replace(" ", "")

    # Porovnání řetězce s jeho obrácením
    return retezec == retezec[::-1]

# Příklad použití
print(je_palindrom("radar"))  # True
print(je_palindrom("A man a plan a canal Panama"))
print(je_palindrom("python"))  # False
```

This function does not take punctuation marks or spaces into account in the checking process.

Reward modeling

- the 3^{rd} phase of the assistant model development – a **reward model**
- the annotators compare the quality of the responses

Write a Python function that checks if the specified string is a palindrome. Use Czech comments.

Of course, I can help v… h that. Here's a simple Pyth… to check if a string is a…


```
def je_palindrom(retezec):
    retezec = retezec.lower()  # Pře…  …nen na malá
    cisty_retezec = "".join(c for c in …  c.isalnum())
    return cisty_retezec == cisty_rete…  # Porovnání ře

# Testování
retezec = input("Zadej řetězec: ")
if je_palindrom(retezec):
    print("Zadaný řetězec je palindrom.")
else:
    print("Zadaný řetězec není palindrom.")
```

This function first converts all letters in the string to lower case letters and removes all non-alphanumeric characters. Then compares the modified string to its inverse and returns True if are the same, otherwise it returns False.

Write a Python function that checks if the specified string is a palindrome. Use Czech comments.

Of course! Here is a si… Python function that will ch… string is a palindrome:


```
def je_palindrom(reteze…
    retezec = retezec.lowe…  …revedeme i
    return retezec == rete…  # Porov

# Příklad použití
vstup = input("Zadejte ř…
if je_palindrom(vstup):
    print("Zadaný řetězec je palindrom.")
else:
    print("Zadaný řetězec není palindrom.")
```

This function first converts the input string to lower case using the lower() method to avoid increasing or decreasing case, which could lead to incorrect results. It then compares the string with its inverse using the cut operator [::-1]. returns True if the string is a palindrome, and False if it is not.

Write a Python function that checks if the specified string is a palindrome. Use Czech comments.

Sure, here's a simple P… function that checks… given string palindrome:


```
def je_palindrom(retezec):…
    # Převedení řetězce na …  …odstraněn
    retezec = retezec.lower()…  " ", "")

    # Porovnání řetězce s je…
    return retezec == retezec…

# Příklad použití
print(je_palindrom("radar"))  # True
print(je_palindrom("A man a plan a canal Panama"))
print(je_palindrom("python"))  # False
```

This function does not take punctuation marks or spaces into account in the checking process.

Reward modeling

- further fine-tuning of the model
- input: question, answer and a special token |reward|
- only the output on the |reward| token is trained
- special **training function** – reproduces the agreement between the proposed reward and the ordering from the annotators
- result = a model capable of measuring the quality of the response

question |. . . | |. . . | answer 1 |. . . | |. . . | |reward| → **0.2**

question |. . . | |. . . | answer 2 |. . . | |. . . | |. . . | |reward| → **1.2**

question |. . . | |. . . | answer 3 |. . . | |reward| → **-0.5**

Reinforcement Learning with Human Feedback

or **RLHF**

- the 4^{th} phase of the assistant model development
- adjustment of generating weights in the tuned model from phase 2
- generates responses that maximise reward
- input: question
- the fine-tuned model generates a response
- the quality measuring model (phase 3) computes the reward
- only the answer part is trained
- the answer word probabilities are weighted by the normalized reward
- the result = an **assistant model** a la ChatGPT

Reinforcement Learning with Human Feedback

why RLHF is needed and just fine-tuning is not enough?

- RLHF works better
- comparing quality is **easier** than generating quality

Create a rhyming poem about kittens

In a world of whiskers, soft and small,
Kittens leap and bounce and sprawl.
With tiny paws and tails that twitch,
They turn each moment into a stitch.

Through fields of dreams, they dart and play,
Chasing the sunbeam's golden ray.
In their purrs, a melody sweet,
Life feels warm, complete, and neat.

Create a rhyming poem about kittens

Kittens so fluffy, small and sweet,
Chasing their tails on little feet.
With playful pounces, soft and bright,
They fill the day with pure delight.

Curled up tight for afternoon naps,
Dreaming of mice, or perhaps some laps.
Whiskers twitching, eyes so wide,
A kitten's love is pure inside.

Create a rhyming poem about kittens

Little kittens, soft and sweet,
Tiny paws and nimble feet.
Chasing shadows, tails held high,
Leaping boldly toward the sky.

With a purr, they softly say,
"Come and love us; let's go play!"
In their world, there's endless cheer,
A cozy cuddle's always near.

Best Assistant Models

Text Arena

View rankings across various LLMs on their versatility, linguistic precision, and cultural context across text

Last Updated	Total Votes	Total Models
Nov 26, 2025	4,639,821	276

Rank ↑↓	Rank Spread (Upper-Lower)	Model ↑↓	Score ↓	95% CI (±) ↑↓	Votes ↑↓	Organization ↑↓	License ↑↓
1	1 ↔ 2	gemini-3-pro	1492	±8	9 799	Google	Proprietary
2	1 ↔ 3	grok-4.1-thinking	1482	±8	10 067	xAI	Proprietary
3	2 ↔ 6	claude-opus-4-5-20251101	1466	±9	4 677	Anthropic	Proprietary
4	3 ↔ 6	grok-4.1	1464	±8	9 967	xAI	Proprietary
5	3 ↔ 8	gpt-5.1-high	1461	±8	7 893	OpenAI	Proprietary
6	3 ↔ 10	claude-opus-4-5-20251101-thinking-32k	1460	±12	2 763	Anthropic	Proprietary
7	5 ↔ 10	gemini-2.5-pro	1452	±4	70 875	Google	Proprietary
8	5 ↔ 13	claude-sonnet-4-5-20250929-thinking-32k	1448	±5	22 000	Anthropic	Proprietary
9	6 ↔ 13	claude-opus-4-1-20250805-thinking-16k	1448	±4	37 617	Anthropic	Proprietary
10	6 ↔ 15	claude-sonnet-4-5-20250929	1445	±6	16 961	Anthropic	Proprietary

https://lmarena.ai/leaderboard/text, 2.12.2025

Prompt Engineering

- since a certain model size (approx. GPT-2) $\mapsto$ modification of the prompt (the question) can **substitute fine-tuning**

no examples (*zero-shot*)

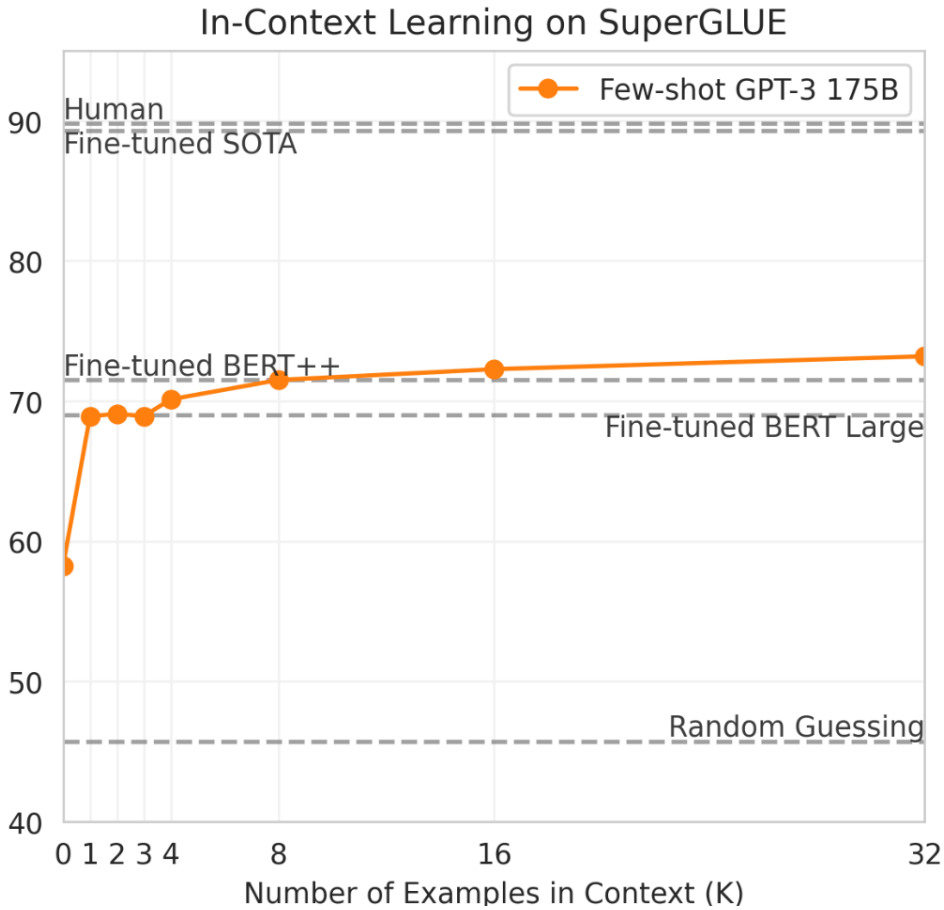
```
Translate English to French:
cheese =>
```

one example (*one-shot*)

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

few examples (*few-shot*)

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



(Brown et al, 2020)

How to form a prompt effectively?

context and prompt design methodology = **prompt engineering**

- providing examples (*few-shot learning*)
- adding details
- chain of thought
- information retrieval (*retrieval-augmented generation*, *RAG*)

Adding details

- the model does not generate correct answers, but the answers from the training
- if we want a **correct** answer, we have to ask for it:
  - "`You're the leading expert on ...`"
  - "`You have an IQ of 130.`"
  - "`Make sure you have the right answer.`"
  - "`Proceed logically.`"
- detailed prompt with context should contain elements like:
  - the role of the responder
  - the goal of the question
  - the audience – to whom the answer is addressed
  - the topic focus – style, tone, ...
  - the structure and the format of the answer

You're a secondary school history teacher. Design an engaging classroom activity for the 8th grade students about creating a common digital timeline. The activity should include a story telling and technology. The time commitment is 3–4 full lessons (45 minutes each). Format the response as an outline.

Chain of Thought

The African elephant is 1.5 times heavier than the Indian elephant

how does a **human** proceed?

- he decides to compare the weights of the African and Indian elephants
- he looks at Wikipedia – the African elephant weights 6 000 kg, the Indian elephant 4 000 kg
- he calculates the ratio of $6,000/4,000 = 1.5$
- he expresses this result in text, formulates it

how does the language **model** proceed?

- it generates words one by one
- it spends the same amount of time on each word
- it does not check for correctness, just the interword relationships
- it uses a huge memory of these relationships

Chain of Thought

chain of thought:

- the model needs more tokens to "think"
- in the prompt, ask to proceed **step by step**
- the result is more accurate

At the bus stop, 44 passengers got off the bus. After boarding, there were 13 passengers more in the bus. How many passengers boarded at the bus stop?

The number of passengers who boarded at a stop can be calculated by subtracting the number of passengers who alighted from the total number of passengers at the stop after arrival.
The number of passengers who boarded = 44 + 13 - 44 = 13
So 13 passengers boarded at the stop.

Chain of Thought

chain of thought:

- the model needs more tokens to "think"
- in the prompt, ask to proceed **step by step**
- the result is more accurate

At the bus stop, 44 passengers got off the bus. After boarding, there were 13 passengers more in the bus. How many passengers boarded at the bus stop?

The number of passengers who boarded at a stop can be calculated by subtracting the number of passengers who alighted from the total number of passengers at the stop after arrival.
The number of passengers who boarded = 44 + 13 - 44 = 13
So 13 passengers boarded at the stop.

At the bus stop, 44 passengers got off the bus. After boarding, there were 13 passengers more in the bus. How many passengers boarded at the bus stop? Solve the math problem correctly. Proceed step by step.

1. The number of passengers on the bus dropped by 44 after the passengers got off.
2. After new passengers boarded, the bus contained 13 more passengers.
3. So 44 and 13 passengers more had to board the bus at the stop than got off.
4. So we calculate the final number of passengers who boarded at the bus stop by adding 44 + 13 = 57.

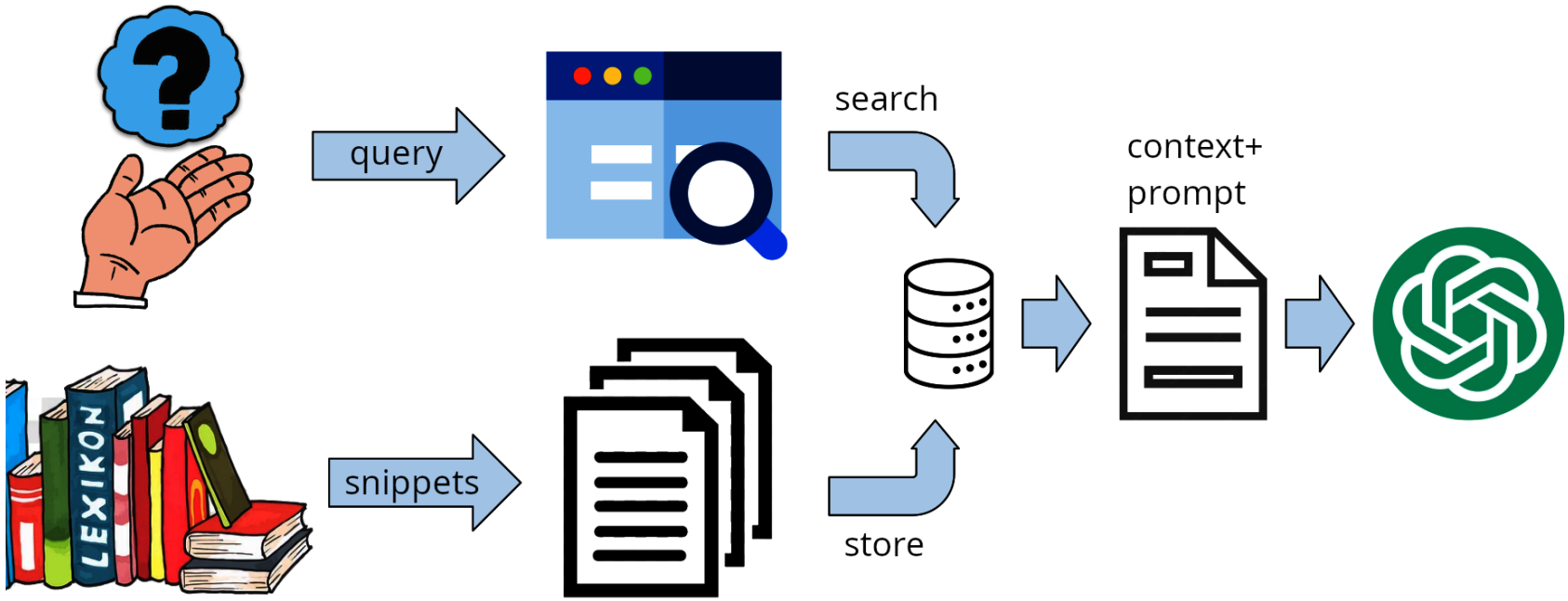
Answer: 57 passengers have boarded the bus at the bus stop.

Information Retrieval

sometimes we need to search for the answers in a database or a set of documents

Retrieval-Augmented Generation, RAG

- divide documents into short text snippets
- store snippet vector representations in a vector database (e.g. using LlamaIndex)
- after a query, find a text in the DB that matches the query vector
- add the retrieved snippet as the context of the prompt



Prompt Engineering – recommendations

- enter detailed prompts with context, relevant information and instructions
- relevant information can be retrieved based on the query
- experiment with different prompt forms
- base models have higher perplexity than assistant models, they generate more diverse texts
- think about the ethics – data bias, possible misuse for spreading misinformation or malicious content
- always **verify** the responses