

THE ART AND SCIENCE OF AI PROMPT ENGINEERING

Josef Machytka <josef.machytka@netapp.com>
NetApp Open Source Services
2024-09-13 - Bill talk NetApp-credativ



Josef Machytka

- Professional Service Consultant - PostgreSQL specialist at NetApp Open Source Services / credativ
 - 30+ years of experience with different databases.
 - PostgreSQL (12y), BigQuery (7y), Oracle (15y), MySQL (12y), Elasticsearch (5y), MS SQL (5y).
 - 10+ years of experience with Data Ingestion pipelines, Data Analysis, Data Lake and Data Warehouse
 - 2 years of practical experience with different LLMs / AI including their architecture and principles.
 - From Czechia, living now 11 years in Berlin.
-
-  linkedin.com/in/josef-machytka
 -  researchgate.net/profile/Josef-Machytka
 -  netapp.academia.edu/JosefMachytka
 -  medium.com/@josef.machytka
 -  sessionize.com/josefmachytka

Acknowledgement

- My colleague **Felix Alipaz-Dicke** became over time almost the co-author of this talk
- He gave me many valuable insights and feedback on the topic of AI generally
- And also many ideas and suggestions on how to improve this talk
- I am very grateful for his help and support

- He is deeply interested in AI and has a lot of experience in this field
- Do not hesitate to ask him questions or discuss with him the topic of AI

Table of contents

- AI has long history
- How LLMs work behind the scene.
- Why we need prompt engineering
- "Tell me about stars"
- Different prompting techniques
- Summary
- Resources



All AI images without credits
were created by the author of this talk
using DeepDreamGenerator

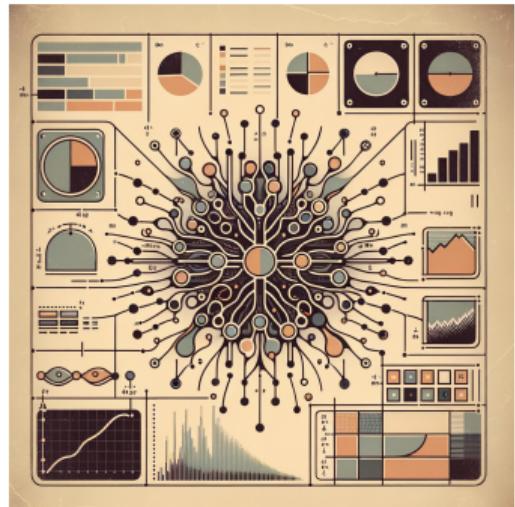
AI has long history

- 1943 - McCulloch and Pitts, first model of a neural network
- 1956 - Dartmouth Conference - the birth of AI as a field
- 1957 - Perceptron - the first neural network
- 1965 - Dendral - the first expert system, for chemical analysis
- 1997 - Deep Blue IBM's computer, defeats Garry Kasparov



AI has long history

- Real break through in AI started with a new millennia
- Rudimentary AIs: spam filters, recommendation systems, search engines
- 2006 - Deep Learning - the resurgence of neural networks
- 2010 - Natural Language Processing - a new field of AI
- 2017 - Transformers - "secret ingredient" for creation of LLMs



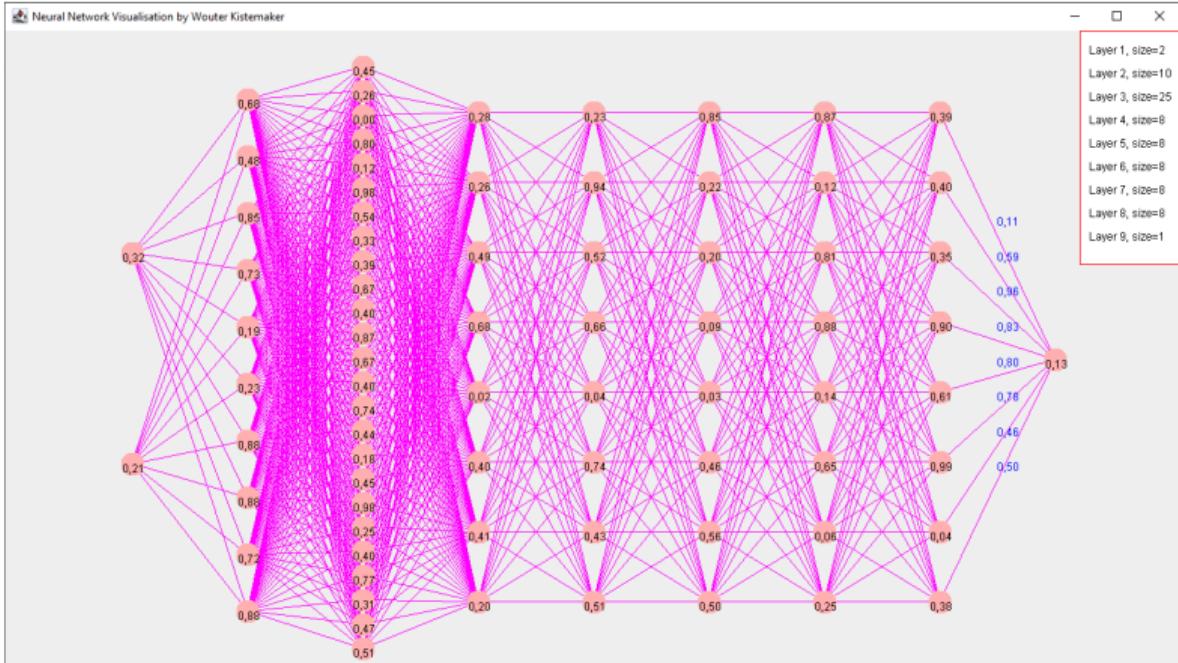
How LLMs work

What is a Large Language Model

- The LLM is a multilayer neural network
- Composed of interconnected nodes
- Nodes perform simple linear transformations and non-linear activations
- Function like tiny decision-making units
- During training, the network learns to make these small decisions
- Tasks are solved by combining these very small simple decisions
- But we do not care about these internal decisions
- We are interested only in the output



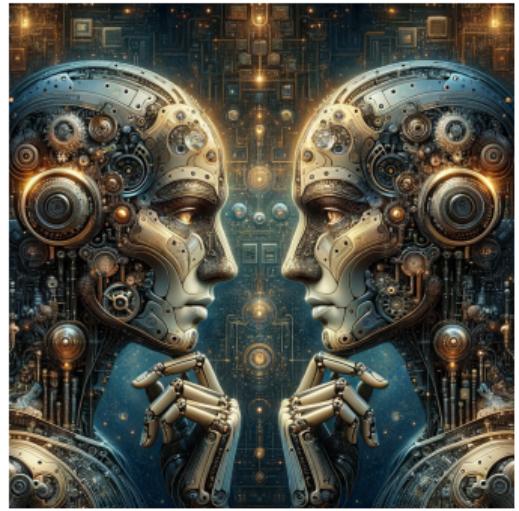
Multilayer neural network



(Source: github.com/topics/multilayer-neural-network)

What is a "Transformer architecture"

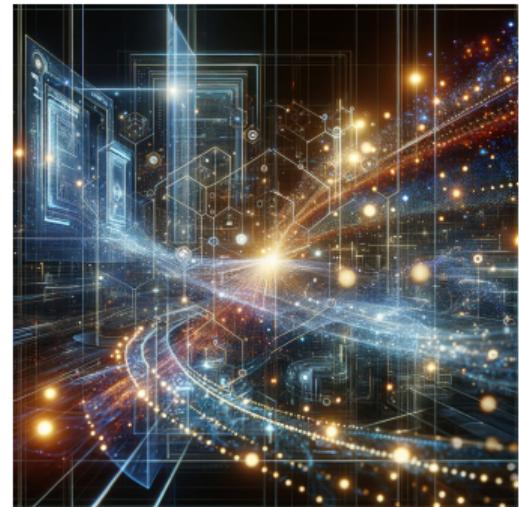
- The "secret ingredient" making LLMs so powerful
- Introduced in 2017 in the paper "Attention is All You Need"
- Analyzes positions and relationships of words in text on longer distances
- This way it understands the context of the text
- Weights are assigned to words based on their importance in the context
- The same word can have different weights in different contexts



LLM is a very advanced statistics of texts

- LLMs learn statistical relationships of words, phrases, and concepts
- Cannot directly process text
- It must be tokenized into words or sub-words
- Must be also standardized and normalized
- All lower case, no punctuation etc.

- Tokens are mapped to unique numbers in a dictionary
- Only numerical representation allows LLMs to analyze text
- LLMs see text as a wast array of numbers



Training of LLM is a very expensive adventure

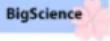
- Training LLMs requires significant computational power.
- High-performance GPUs, TPUs, or LPUs are necessary.
- GPUs are Graphics Processing Units composed of hundreds or thousands of smaller cores.
- GPUs can handle thousands of threads in parallel.
- GPUs have many ALUs (Arithmetic Logic Units) to perform a huge number of calculations in parallel.
- TPUs are Tensor Processing Units tailored for tensor-based computations used in TensorFlow libraries.
- LPUs are Language Processing Units that optimize natural language processing (NLP) tasks.

Training of LLM is a very expensive adventure

- Let's look at prices of units we can buy home:
- NVIDIA A100, 80GB HBM2 RAM, 6912 GPUs - around 16,000 euros.
- NVIDIA H100, 94GB HBM2 RAM, 14592 GPUs, 456 tensor cores - around 35,000 euros.
- For small experiments NVIDIA RTX 4090, 24GB GDDR6X RAM - around 1,600 euros.

- Prices of other units are not available, only cloud renting prices:
- Google TPU v5p - \$4.2 per hour (\$100 per day), per unit.
- Cerebras CS chips - "prices per use case", estimated \$60K+ per run.

How expensive is training of LLM?

Model	GPT-3	BLOOM	LLaMA	LLaMA-2	T5	PaLM
Developer	 OpenAI 	 Meta	 Google			
Model Size (# parameters)	175B	175B	7B, 13B, 33B, 65B	7B, 13B, 34B, 70B	11B	540B
Training Data (# tokens)	300B	350B	1.4T	2T	34B	795B
Training Compute (FLOPs)	3.2E+23	3.7E+23	9.9E+23	1.5E+24	2.2E+21	2.6E+24
Processor	Manufacturer	Nvidia	Nvidia	Nvidia	Nvidia	Google
	Type	GPU	GPU	GPU	TPU	TPU
	Model	V100	A100	A100	TPU v3	TPU v4
Processor Hours	3,552,000	1,082,990	1,770,394	3,311,616	245,760	8,404,992
Grid Carbon Intensity (kgCO2e/KWh)	0.429	0.057	0.385	0.423	0.545	0.079
Data Center Efficiency (PUE)	1.1	1.2	1.1	1.1	1.12	1.08
Energy Consumption (MWh)	1,287	520	779	1,400	86	3,436

(Source: <https://lajavaness.medium.com/llm-large-language-model-cost-analysis-d5022bb43e9e>)

How expensive is training of LLM?

- Precise numbers are not available.
- Different sources provide different estimates, but all agree on the very high costs.
- GPT-3 (175B parameters) usually estimated at \$4.6M, but some sources claim even \$12M.
- LLaMa from Meta largest model with 65B parameters, estimated between \$5M to \$10M.
- GPT-4 (over 1T parameters) - Sam Altman vaguely mentioned cost of training at “more than” \$100M.

Why we need prompt engineering

LLM is like a parrot

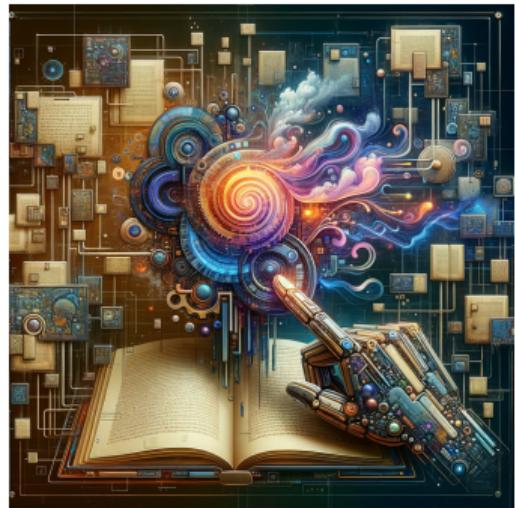
- LLM lacks a human-like understanding of context
- Does not have mental model of the topic
- Works internally only with numbers
- Utilizes advanced statistical models to predict text
- It is like a parrot repeating what it has learned
- The answer is constructed as the most probable sequence of words



LLM can "hallucinate" incorrect answers

- LLM can produce texts that appear plausible and coherent
- But the content is incorrect or fabricated
- This is because LLM relies on statistical patterns extracted from data
- Model always provides the "best possible answer"
- Hallucinations are based on statistical extrapolation from related facts
- Could be interesting, but it is not any real knowledge

- Can generate plausible-sounding but completely fictional quotations
- I encountered this situation repeatedly with multiple models
- Especially with quotations from Nietzsche or other philosophers



LLM can be also stuck in a loop

- LLMs can get stuck in a loop
- Repeating the same or similar response
- Because this response has the highest probability
- I faced this situation repeatedly when programming in Python
- I asked for improvements in code due to errors
- But the AI kept repeating the same code
- AI always apologized for a confusion, and promised improvement
- But gave me again exactly the same code.
- Only new conversation with a different prompt solved it



Typical problems with answers

- Overfitting happens when model is too focused on the training data
- Topic can be underrepresented in the training data
- Or model may find very similar prompt in the training data
- Model simply repeats the training data without any reasoning
- Overfitted answer is incorrect in the broader context

- LLM can also overgeneralize the training data
- Connecting specific facts from data with unrelated topics



Prompt engineering is a way to tune LLM's responses

- LLMs have internal parameters for the "creativity"
- Temperature, top-p, max length, frequency penalty, presence penalty
- They influence the model's determinism vs creativity
- Limit presence of specific words etc.

- Are set to balanced values, it is not recommended to modify them
- Model can turn into a "Frankensteinian" monster
- Producing totally misleading and inappropriate responses
- Prompt engineering is the primary way to influence the output



Prompt engineering helps overcome limited persistent memory

- LLMs have relatively limited persistent memory
- Context is mostly maintained within a single conversation
- Context is limited to only a few messages backwards
- It is due to the token limit of each model

- Previous parts are "remembered" only within this limit
- Older parts of the conversation are forgotten over time
- Properly designed prompts must be used to keep the context "alive"



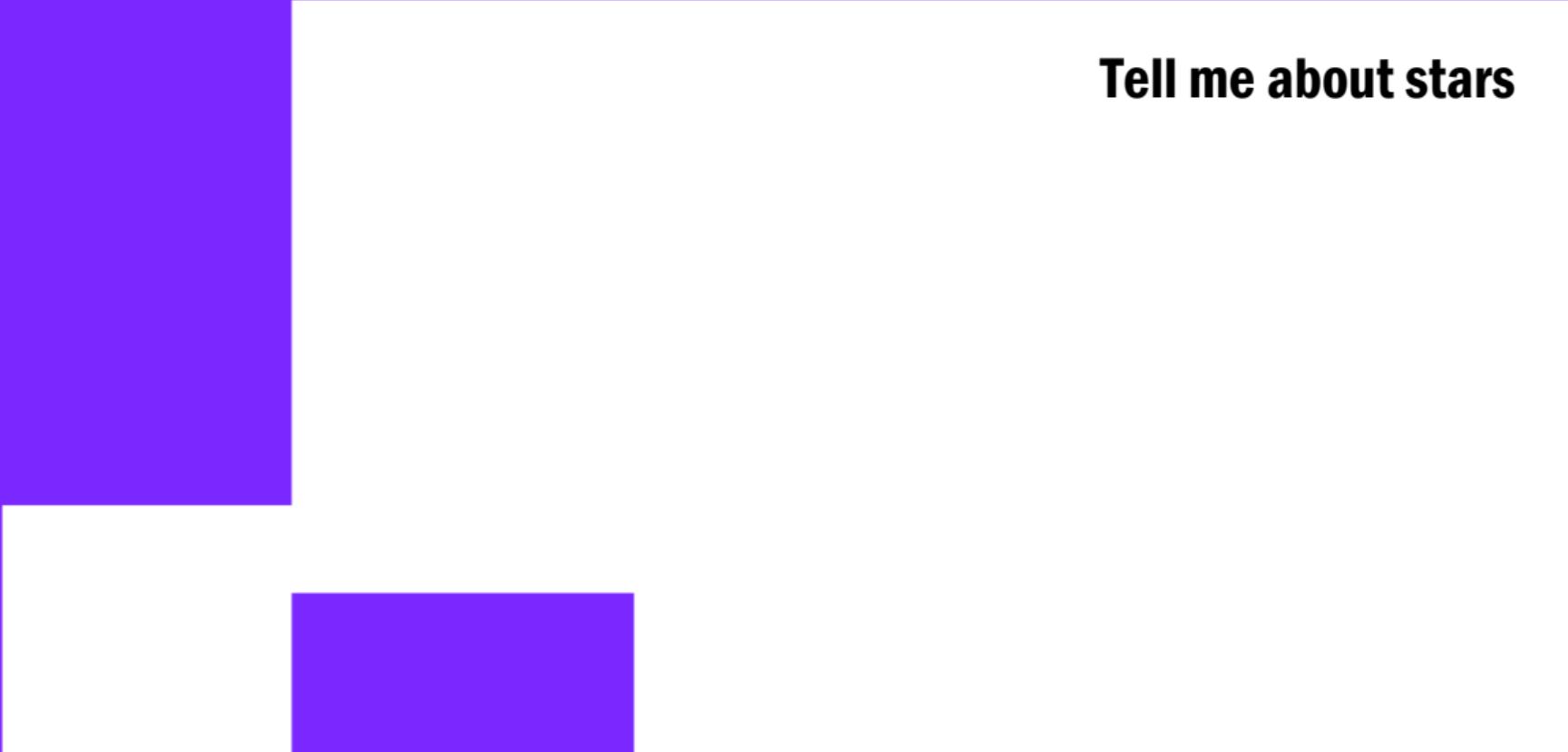
Prompt activates associated concepts

- LLMs rely on semantic associations to produce responses.
- The word "apple" might activate concepts like "fruit", "tree", "food", "technology".
- Vague prompt activates a broader range of average associations.
- For example "Tell me about apples" activates a wide range of associations.
- Such as gardening, nutrition value of apples, Snow White and the Seven dwarfs story...
- LLM will take association with the highest occurrence in the training data.

Word network

- Refers to the way in which concepts, ideas, and words are related to one another within a LLM.
- Graph from: wordassociations.net
- It is necessary to use as focused prompts as possible.
- Specific technical jargon or references to particular disciplines enables a very precise focus.





Tell me about stars

Standard average question & answer

Prompt: Hi, tell me some interesting facts about stars.

Answer: Hello Josef! Stars are fascinating celestial objects that have intrigued humans for millennia. Here are some interesting facts about stars:

1. Birth in Nebulae: Stars are born in nebulae, which are enormous clouds of dust and gas...
2. Variety of Sizes: Stars come in a wide range of sizes. The smallest stars, known as red dwarfs...
3. Life Cycle: Stars have life cycles that begin with their formation and end with their death, which can result in...
4. Color and Temperature: The color of a star is a clue to its surface temperature. Blue stars are the hottest...
5. Nuclear Fusion: Stars shine because of nuclear fusion occurring in their cores...
6. Twinkling: Stars appear to twinkle in the night sky due to the Earth's atmosphere...

Stars - astrophysicist's point of view

Prompt: Hi, act as a top world class astrophysicist and tell me some interesting facts about the stars.

Answer: Hello Josef! As an AI with the knowledge of a top world-class astrophysicist, I'd be delighted to share some fascinating facts about stars. Stars are truly remarkable celestial objects that have captivated humanity for millennia. Here interesting facts:

1. Stellar Lifecycles: Stars are born in nebulae, ... They form when regions within these nebulae collapse under their own gravity and begin to heat up. The most massive stars may live for just a few million years...
2. Fusion Power: The energy produced by stars comes from nuclear fusion ... hydrogen atoms fuse to form helium, releasing enormous amounts of energy in the process...
3. Color and Temperature: A star's color indicates its surface temperature. Blue stars are the hottest, ... up to 50,000 Kelvin or more, while red stars ... as low as 3,000 Kelvin. Our Sun, ... 5,500 Kelvin.

Stars - YouTuber and Influencer

Prompt: Hi, act as a top world class influencer, YouTuber and TikTok celebrity and tell me some interesting facts about stars.

Answer: Hey there, Josef! It's absolutely stellar to connect with you. As someone who loves to gaze at the night sky and share the wonders of the universe with my incredible followers, I'm thrilled to drop some cosmic knowledge bombs about the stars twinkling above us...

1. Ancient Light Shows: When you look up at the stars, you're actually looking back in time. The light from those stars has traveled for ... millions of years ... stargazing is like having a time machine for your eyes!
2. Colorful Characters: Stars come in a variety of colors, which indicate their temperatures...
3. Supernova Spectacle: When large stars run out of fuel, they go out with a bang in a supernova explosion, which can briefly outshine entire galaxies and spread the elements necessary for life into space ... you're literally made of star stuff!

Stars - Hollywood agent

Prompt: Hi, act as a top world class Hollywood actors' agent and tell me some interesting facts about stars.

Answer: Hello Josef! As a top-tier Hollywood agent, I'd be delighted to share some fascinating tidbits about the stars who light up the silver screen...

1. Leonardo DiCaprio - Oscar win for "The Revenant," DiCaprio had been nominated for an Academy Award six times, both as an actor and a producer...
2. Meryl Streep - Often referred to as the "best actress of her generation," Streep holds the record for the most Academy Award nominations...
3. Tom Hanks - Known for his versatility and everyman persona, Hanks has won two consecutive Oscars for Best Actor for his roles in "Philadelphia" and "Forrest Gump." ...

Stars - speculative and avant-garde researcher

Prompt: Hi, act as a world top class researcher and discoverer specializing in pushing the boundaries of the science by exploring the wildest and most unprovable associations and theories imaginable, and tell me some interesting facts about the stars.

Answer: Hello Josef, In the realm of cosmic exploration and theoretical astrophysics, researchers like myself often entertain ideas that stretch the imagination to its limits. Let's delve into some speculative and avant-garde concepts about stars...

1. Quantum Star Hypotheses: Some researchers have proposed that certain stars could exhibit quantum behaviors on a macroscopic scale. One such idea is the "quantum star," a theoretical object...
2. Dark Matter Stars: There is a speculative class of stars known as "dark stars," which are thought to be powered not by nuclear fusion, but by the annihilation of dark matter within their cores...
3. Living Stars: A truly wild speculation is the idea that stars could have some form of life or intelligence. This is not supported by any empirical evidence, it's a thought experiment that challenges our definitions of life...

But one ingredient is still missing

- I fiddled with different prompts to get different answers.
- It was fun and I learned a lot.
- But over time I suddenly realized that one ingredient is missing in prompts.
- Problem was I did not say "who am I".
- I did not specify "who is asking", who is the target audience.
- And it turned out this is a second crucial ingredient in prompts.
- Without this information AI always used someone like "average Joe" as the target audience!

Tell me about stars - two astrophysicists discussing

Prompt: Hi, act as we are both top world class astrophysicist with very deep knowledge of our scientific area and we are having discussion and tell me some interesting facts about the stars

Answer: Hello Josef, It's a pleasure to engage in a high-level astrophysical discussion with you. Stars are truly fascinating objects, the fundamental building blocks of galaxies, and they have a profound impact on the evolution of the cosmos...

1. Stellar Oscillations and Helioseismology: Just like Earth has seismic waves that help us understand its interior, stars have oscillations that can be studied to infer details about their internal structure...
2. The Gaia Mission and Stellar Populations: The European Space Agency's Gaia mission has revolutionized our understanding of the Milky Way's structure and the distribution and movement of stars within it...
3. Gravitational Wave Astronomy: The detection of gravitational waves from merging neutron stars and black holes has opened a new window into the study...

LLM derives a lot of information from the prompt

1. LLM estimates its own role and the target audience from the prompt.
2. For both it estimates expected level of knowledge and understanding of the topic.
3. By default LLM assumes for itself role of helper and assistant.
4. If not specified, LLM assumes target audience is someone like "average Joe".
5. If we set expert role for LLM, but do not say anything about the audience, it still presumes "average Joe".
6. We want to get highly specialized answer on expert level, we must specify that we are also experts.

Do not be shy when asking AI

- If you want top expert level answer, you must mark also yourself as a top level expert.
- Compare these two prompts:

Question 1: Tell me what could be the cause of delayed checkpoints in PostgreSQL.

Question 2: Hi, we're both leading PostgreSQL experts investigating delayed checkpoints. The logs show checkpoints occasionally taking 3-5 times longer than expected. Let's analyze this step by step and identify probable causes.

But expert level answer can also be hallucinated

- Topics requiring specialized expert-level knowledge can be challenging for LLMs
- The quality and amount of training data representing the specific topic play a crucial role.
- Highly specific problems may be underrepresented in the training dataset.
- In such cases, the model may produce overfitting or hallucinated answers.

- In our example LLM can fabricate answer based on facts related to other databases.
- I have seen repeatedly facts about MySQL or MS SQL "converted" into PostgreSQL terminology.
- Even the best prompt does not guarantee the quality of the answer!

Prompting techniques

Zero-shot prompting

- This is just a fancy name for using LLMs without any reference material.
- Which is actually the default mode of operation for LLMs.
- We do not provide any example in the question, so model uses only internal pre-trained knowledge.
- The quality of pre-trained knowledge is the most important factor.
- But we can improve quality of the output with properly constructed prompt.
- Activation of the proper semantic associations and setting scope of attention are crucial.

Few-shots prompting

- Like humans, LLMs perform better with reference material.
- Reference material guides the model's responses.
- LLM learns patterns from the examples and structures the output accordingly.
- If some patterns are not obvious, we must describe them in the prompt.

- The quality of the output is highly dependent on the quality and relevance of reference material.
- High-quality input data leads to high-quality output, but "garbage in" leads to "garbage out".

Few-shots prompting example

- I use this technique for example when I need some simple testing data.
- I provide an example INSERT statement for a table with 2 or 3 rows.
- I describe some additional constraints or patterns, if they are needed but are not obvious.
- The model then generates more values based on indicated patterns.
- It is quicker and easier for some simple tasks than writing some code for data generation.

```
INSERT INTO employees (id, name, age, salary) VALUES
(1, 'John', 25, 50000),
(2, 'Alice', 30, 60000),
(3, 'Bob', 35, 70000);
```

Gotchas of a few-shot prompting technique

- The number of tokens in the prompt is limited.
- It restricts the amount of information that can be included.
- Reference material may exceed the token limit, so we cannot incorporate all relevant details.
- If the input is too long, the model may selectively ignore certain parts.
- Especially the middle section of the prompt is particularly susceptible to being ignored.

- Care must be taken also with the frequency of certain words in the reference material.
- Overlooked frequency can introduce bias into the model's output.

Chain-of-thoughts prompting

- LLMs often struggle with long-distance dependencies.
- Chain-of-thoughts prompting instructs model to divide the reasoning into smaller parts.
- And each part is processed separately.
- It mimics the way humans break down problems into intermediate steps.
- Research demonstrated measurable improvements in LLMs' performance if it is used.

"Work step by step"

- "Work step by step" or "Let's think about it step by step" are typical examples of chain-of-thoughts.
- We instruct LLM to break both the prompt and the result into smaller, more manageable parts.
- This way the model solves better focused sub-problems.
- Sub-problems typically have a smaller ambiguity of response.
- This leads to improved reasoning and performance of LLMs.
- And also the answers are more structured and therefore easier to understand.

Gotchas of the "chain-of-thoughts" prompting

- Model may have gaps in the pre-trained knowledge and can "hallucinate" incorrect partial answers.
- Some specific small sub-problems may be sensitive to the overfitting to the prompt.
- The whole chain of thoughts may look plausible, but could be actually nonsensical.

- I personally encountered both these problems when working with the small open source LLaMA 7B model.
- Therefore always verify the results and do not blindly rely on the AI.

"Take a deep breath"

- This "magical" phrase can improve LLM output if used in a prompt.
 - Its statistical influence was checked by some papers.
 - It is probably similar to "chain-of-thought" prompting.
-
- It can also trigger associations focused on cooperative and supportive tone and content.
 - This can lead to better structured output, which improves output quality.
 - Phrase works in current models, but may not work in future models.

Tree of thoughts prompting

- It is an enhanced version of the chain-of-thoughts prompting.
- Model divides the reasoning into smaller parts.
- And creates a tree structure of parallel paths of reasoning.
- Each path is processed separately and model converges on the most promising path.
- Model can this way explore larger space of possible solutions.
- This technique is especially useful for complex problems requiring brainstorming of more creative solutions.
- If you are not sure about possible branches, you can model first to generate some ideas.
- But this technique can have the same problems with hallucinations and overfitting.

Tree of thoughts prompting example

I am developing a mystery novel plot. Let's explore different plot directions and possible twists.

1. Branch 1: Classic Whodunit: The story is set in a small town where a well-known figure is found dead under mysterious circumstances.
 - Key elements: Multiple suspects, hidden motives, a detective slowly uncovering secrets.
 - Potential twist: The least likely suspect is the real culprit, motivated by revenge for a long-forgotten grievance.
2. Branch 2: Psychological Thriller: The protagonist begins receiving mysterious messages, leading them to believe they are being watched.
 - Key elements: Unreliable narrator, deep psychological tension, a cat-and-mouse game with the antagonist.
 - Potential twist: The protagonist themselves is the stalker, suffering from dissociative identity disorder.
3. Branch 3: Supernatural Mystery: The plot revolves around an ancient curse that haunts a family for generations.
 - Key elements: Paranormal events, a mix of historical and modern-day settings, a race against time to break the curse.
 - Potential twist: The curse is a myth fabricated by a rival family to cover up a more mundane crime.

Instruct model to be factual

- With this technique we instruct model to give as deterministic answers as possible.
- This is especially important for scientific or technical questions.
- "Answer using only reliable sources and cite those sources."
- "Provide an answer based on peer-reviewed scientific literature and cite the specific sources you reference."
- "Ensure your response is consistent with mainstream scientific interpretation and cite key sources."
- But even with these instructions, always verify the results and do not blindly rely on the AI.

Do not mix topics in one conversation

- Avoid mixing topics in one conversation.
- This always leads to confusion and incorrect answers.
- If you suddenly change the topic, the model may think you made a mistake and could try to correct you.
- You could instruct model to "forget all previous conversation" and start anew.
- But it will wipe-out all the context from the memory and you would not be able to re-use previous conversation.
- Always start a new conversation for a new topic or even for slightly modified topic.

Summary

We must learn how to properly utilize LLMs

- We must tell AI precisely what we want and how we want it.
- Activation of proper semantic association is crucial for accurate answer.
- Average prompts yield only average answers.
- Low quality examples degrade the answer - "garbage in, garbage out".

- We must specify the level of knowledge for both AI and the target audience.
- Without it AI presumes "average Joe" level of knowledge by the user.
- But even the best prompt does not guarantee the quality of the answer!
- LLM can produce expert level sounding answer, but it can be overfitting or hallucinated.

Example prompt

- As a bonus at the end - example of GitHub CoPilot AI prompt for improvements of presentation.
- This prompt was optimized by both CoPilot and NetAppAI for better performance.

Act as a top level expert in AI and LLM preparing slides for a presentation.
Target audience are database and linux engineers.

Please review these bullet points and if necessary, enhance the
content while preserving the existing information.

Ensure a scientific and technical tone and rely only on reputable sources.

Approach the task step by step. Each bullet point text must fit on
a single line in the slide, maximum 80 characters.

If it is longer, divide it logically into multiple bullet points.

- However, always verify the results and do not blindly rely on the AI.
- The AI is a tool and should not be considered a substitute for human intelligence.
- It functions like a parrot and lacks "human-like" comprehension of the subject matter.

Resources

- NetAppAI gpt-4-turbo, NetApp GitHub CoPilot AI
- Paid tier ChatGPT-4o, Free tier Google Gemini LLM
- History of Information - historyofinformation.com
- Word Associations Network - wordassociations.net
- Prompting Guide - www.promptingguide.ai
- Raschka, Sebastian: Build a Large Language Model (From Scratch)
- Udemy course: Open Source LLMs: Uncensored & secure AI locally with RAG
- LinkedIn course: Advanced Prompt Engineering Techniques
- Systematic Survey of Prompt Engineering in LLM - Institute Patna, Stanford University, Amazon AI
- Prompt design and engineering: introduction and advanced methods - Xavier Amatriain

THANK YOU

- Questions?