



Applied Generative AI in Climate Research. Prompt Engineering with ChatGPT-4

Wegener Center Seminars.
02.11.2023. Graz

Christopher Pollin

<https://chpollin.github.io>

Zentrum für Informationsmodellierung, Graz

<https://informationsmodellierung.uni-graz.at/de/>

Digital Humanities Craft OG

www.dhcraft.org



ChatGPT-4 Advanced Data Analysis

You are a data scientist specialising in climate system research. I have a NetCDF file containing multidimensional data relevant to the climate system. Your task is to analyse this data set in depth. First, make an initial assessment of the data by contextualising its relevance in the field of climate system research. Then explore the data to identify patterns or anomalies. Discuss, critique, and then implement meaningful data visualizations. Finally, discuss your findings and implications for climate system research.

Take a deep breath and let's think step by step.

Persona Modelling
Context
Task
Chain of Thought
Prompting
Zero-Shot Approach

Contextualising

Initial Assessment and Contextualization

In the field of climate change, multi-dimensional scientific data is often represented in NetCDF format. The columns you mentioned—time, latitude, longitude, and air temperature—form a crucial part of climate models and observational data sets.

1. **Time:** Knowing the temporal range allows researchers to assess trends over time.
2. **Latitude & Longitude:** These are geographical coordinates that can help identify regional impacts.
3. **Air Temperature:** This is a key variable to understand climate patterns and anomalies.

These types of data can serve multiple purposes, such as:

1. **Monitoring Climate Change:** Long-term changes in air temperature can be a strong indicator of climate change.
2. **Regional Analysis:** Geo-specific data can help in understanding how different regions are impacted differently.
3. **Model Validation:** Such data sets are often used to validate climate models.

Initial Assessment & Contextualization

The data in question contains four key variables: time, latitude, longitude, and air temperature measured in Kelvin. This type of data is highly relevant to climate change research for several reasons:

1. **Time-Series Analysis:** Having time-based data allows for tracking changes and trends in air temperature over time. This can provide insights into whether the climate is warming or cooling.
2. **Geographical Variability:** With latitude and longitude information, we can examine how air temperature varies across different regions. This helps in understanding the spatial distribution of climate change impacts.
3. **Model Calibration & Validation:** Such datasets are often used to calibrate or validate climate models, helping to improve their accuracy.
4. **Policy Implications:** Recognizing patterns or anomalies can inform policy decisions aimed at mitigating the effects of climate change.

Data Exploration

Next, we'll explore the data to identify patterns, trends, or anomalies. For this, we'll first load the data and then perform some basic statistics and visualizations.

Let's start by loading the NetCDF file and inspecting its contents.

Finished working Show work ▾

Great, we've successfully loaded the NetCDF file. Here's what we have:

- **Dimensions:**

- Time: 2920 time points
- Latitude: 25 unique values
- Longitude: 53 unique values

- **Variables:**

- `lat`: Latitude, measured in degrees north
- `lon`: Longitude, measured in degrees east
- `air`: Air temperature, measured in degrees Kelvin

- **Global Attributes:**

- The data is from an NMC initialized reanalysis, specific to the 0.9950 sigma level values.

Data Exploration Plan

1. **Basic Statistics:** We'll calculate summary statistics for the air temperature.
2. **Time Series Analysis:** Look at how air temperature has changed over time.
3. **Spatial Analysis:** Examine geographical variations in air temperature.
4. **Anomaly Detection:** Identify any unusual patterns or outliers.

Reading the data

Python + Xarray Library

Data Exploration

We will start by loading the NetCDF file and performing some basic exploratory data analysis (EDA) to understand the structure, check for missing values, and identify any initial patterns or anomalies.

First, I'll load the data.

Finished working Show work ▾

The dataset contains three dimensions:

- **Latitude (lat):** 25 unique values ranging from 15.0 to 75.0.
- **Longitude (lon):** 53 unique values ranging from 200.0 to 330.0.
- **Time (time):** 2920 unique timestamps, ranging from January 1, 2013, to December 31, 2014, with a 4x daily frequency.

The data variable is:

- **Air Temperature (air):** This is a 3D array with dimensions corresponding to time, latitude, and longitude.

The dataset also contains metadata attributes like `Conventions`, `title`, `description`, `platform`, and `references`, which provide additional context.

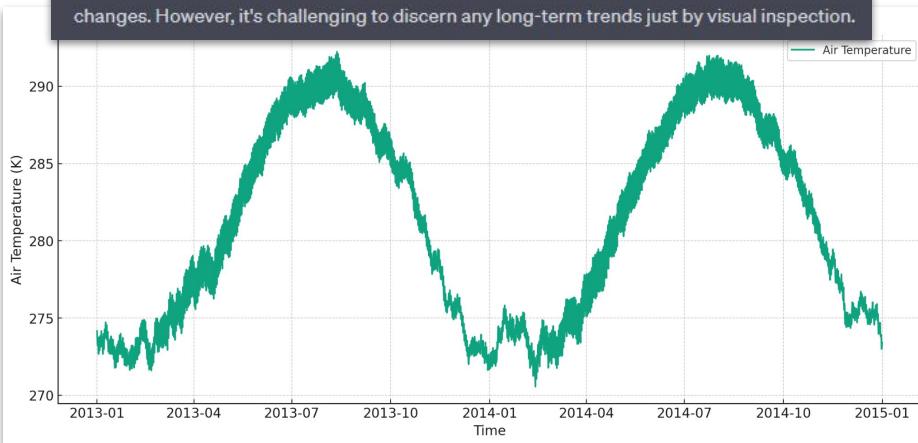
Here are the summary statistics for the air temperature data:

- **Mean Temperature:** 281.26 K
- **Median Temperature:** 285.20 K
- **Standard Deviation:** 16.32 K
- **Minimum Temperature:** 221.0 K
- **Maximum Temperature:** 317.4 K

The mean and median are fairly close, suggesting that the data is somewhat symmetrically distributed. The standard deviation indicates variability, and the minimum and maximum values give us a sense of the range.

Here's the time series plot representing the average air temperature over time. This plot takes the mean air temperature across all latitudes and longitudes at each time point.

From the plot, it appears that there is a cyclical pattern, likely corresponding to seasonal changes. However, it's challenging to discern any long-term trends just by visual inspection.

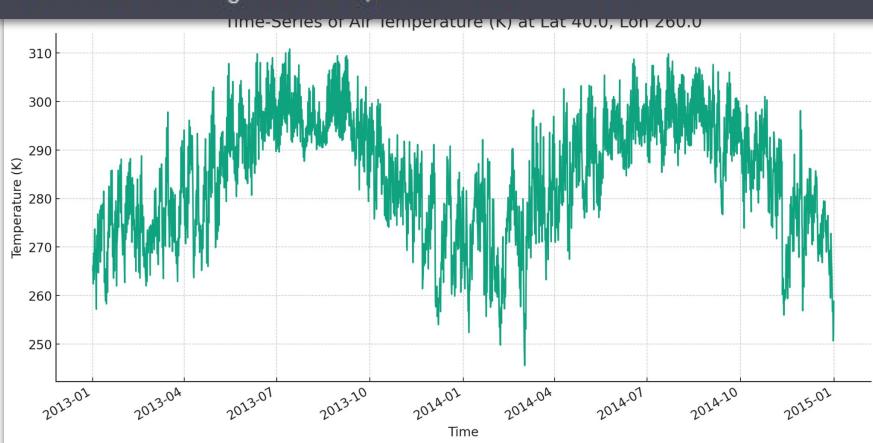


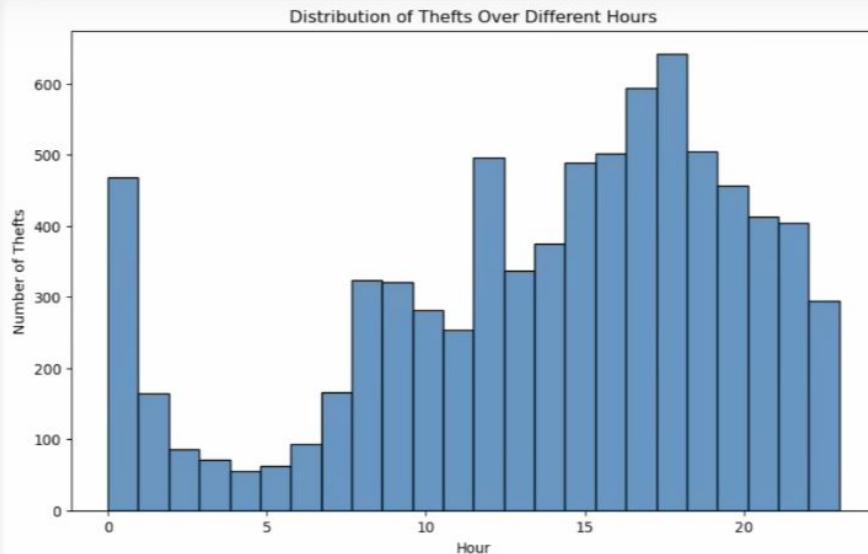
Calculation and data visualisation

The summary statistics for the air temperature data are as follows:

- **Mean Temperature:** Approximately 281.26 K
- **Standard Deviation:** Approximately 16.32 K
- **Min Temperature:** 221.0 K
- **Max Temperature:** 317.4 K

Time-Series Plot: The plot for a specific latitude and longitude (Lat 40.0, Lon 260.0) shows variations in air temperature over time. While the data is too limited to make broad conclusions about long-term trends, the fluctuations are notable.





From the plot, we can see that thefts tend to occur more frequently in the afternoon and evening hours, with a peak around 18:00 (6 PM). The number of thefts decreases significantly during the late night and early morning hours.



Why are there so many bike thefts right after midnight?



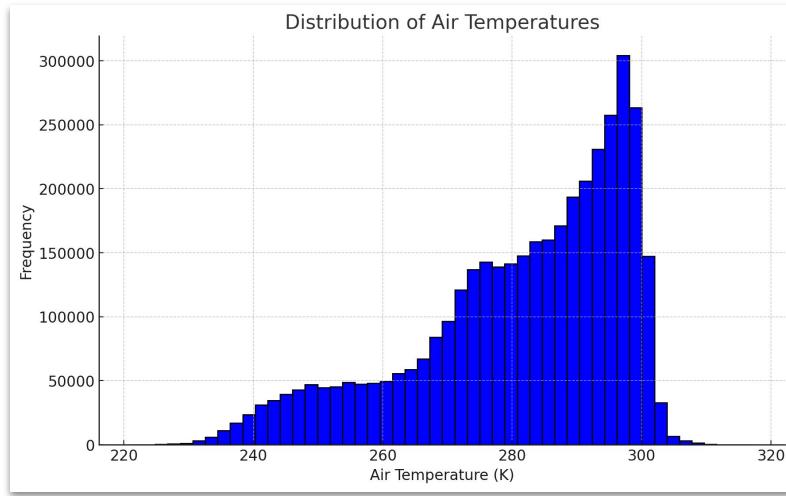
The spike in bike thefts right after midnight could be due to several reasons:

1. **Less Surveillance:** There are usually fewer people around after midnight, which might give thieves more opportunities to steal bikes without being noticed.
2. **More Bikes Left Unattended:** People often leave their bikes locked up outside for the night, making them easy targets for thieves.
3. **Reporting Bias:** Sometimes, if the exact time of the theft is not known, it might be reported as having occurred at the start of the day, which is midnight.
4. **Data Entry Errors:** If the time of the theft was not recorded, it might be entered as 00:00 by default, leading to an overrepresentation of thefts at midnight.

It's important to note that these are just possible explanations. To understand the exact reasons, we would need more detailed data and possibly additional information that might not be ■

GPT-4 not only creates visualisations, it also lets you talk about them.

Exciting: GPT-4 detects "data entry errors"!



Finished working Hide work ^

python

```
# Extract air temperature data as a flat array for the histogram
flat_air_temp_data = data['air'].values.flatten()

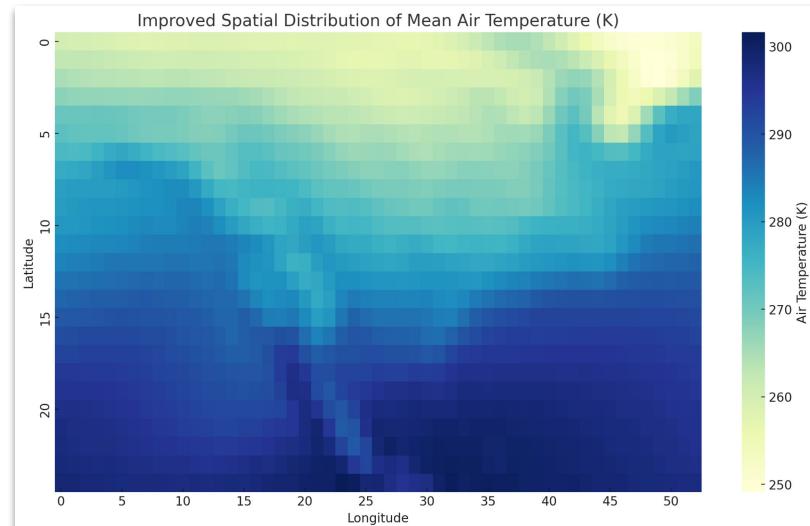
# Create a histogram
plt.figure(figsize=(10, 6))
plt.hist(flat_air_temp_data, bins=50, color='blue', edgecolor='black')
plt.xlabel('Air Temperature (K)')
plt.ylabel('Frequency')
plt.title('Distribution of Air Temperatures')
plt.grid(True)
plt.show()
```

Copy code

Finished working Hide work ^

python

```
# Improved Heatmap with colorblind-friendly color scheme and reduced tick labels
plt.figure(figsize=(14, 8))
ax = sns.heatmap(spatial_data, xticklabels=5, yticklabels=5, cmap="YlGnBu",
                  center=280)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title('Improved Spatial Distribution of Mean Air Temperature (K)')
plt.show()
```



Initial Assessment and Data Loading

- The dataset provided is in NetCDF format, commonly used for storing multi-dimensional scientific data.
- The dataset has four key variables: time, latitude, longitude, and air temperature (in Kelvin).

Summary Statistics

- Mean Temperature: 281.26 K
- Median Temperature: 285.20 K
- Standard Deviation: 16.32
- Minimum Temperature: 22
- Maximum Temperature: 31

Histogram

Original

- Showed the frequency distribution of air temperatures.

Improved Version

- Used the Freedman-Diaconis rule to optimize the number of bins.
- Converted the y-axis to show density instead of frequency.
- Annotated outliers in the temperature distribution.

Interpretation

- The close values of mean and median suggest a relatively stable dataset.
- The standard deviation indicates moderate variability in temperatures.

Time Series Plot

Original

- Showed a line plot of the mean temperature over time.

Improved Version

- Introduced more visible markers for specific data points.
- Annotated the plot to highlight seasonal peaks.

Overall Interpretation

- Seasonal Variations:** The data shows clear signs of seasonal variations but does not provide sufficient evidence of long-term trends in air temperature.
- Spatial Distribution:** There's a clear geographic pattern where temperatures are warmer at lower latitudes and cooler at higher latitudes.
- Temperature Distribution:** The temperature values are largely normally distributed with some outliers, corroborating the summary statistics.

By examining the data through multiple lenses—time, space, and distribution—we gain a comprehensive understanding of the dataset's characteristics and implications. This multi-faceted approach is crucial for any rigorous analysis, especially in a complex field like climate science.

“Interpretation” and summary

1. Summary Statistics Interpretation

- Mean Temperature:** The average temperature is about 281.26 K. This is a critical baseline number; deviations or trends away from this mean could signal climate anomalies.
- Standard Deviation:** The standard deviation is about 16.32 K, which indicates the extent of variability around the mean. A higher standard deviation could indicate more significant fluctuations, possibly due to extreme weather events.

2. Seasonal Trends

- Cyclical Patterns:** The seasonal trend plot indicated a cyclical pattern in temperatures at the selected location (Lat 40.0, Lon 260.0). The temperature seems to rise and fall in a predictable manner each year.
- Amplitude:** The amplitude of these cycles can be an indicator of seasonality strength. Any increase in amplitude over time could signify more extreme seasons, potentially due to climate change.

3. Extreme Values Analysis

- Frequency Distribution:** The histogram showed a relatively normal distribution of temperatures but with some instances of extremes.
- Outliers:** These outliers, if increasing over time, could signify a rising number of extreme weather events, a hallmark of climate change.

4. Geographical Heatmaps

- Latitudinal Gradient:** The spatial heatmaps for different time points show a clear latitudinal gradient in temperatures, with colder temperatures towards the poles and warmer temperatures towards the equator. This is a well-known climatological pattern.
- Temporal Changes:** The multiple heatmaps over different time points could be used to assess whether this latitudinal gradient is shifting over time, which would be a significant finding. For instance, if higher latitudes are warming more rapidly than lower ones, that could be indicative of polar amplification, a well-documented climate phenomenon.

“Exploring the jungle of terms”

(“Klarheit im
Begriffssdschungel”)

Midjourney: ai, two male programmer with laptops holding hands in front of a dense jungle, footpath in the jungle, landscape, cinematic lighting, Geof Darrow style
--weird 50 --chaos 10 --quality 0.25 --ar 16:9

Large Language Models (LLM)

A "large language model" is a machine learning model trained on large amounts of text data (and resource consumption!) to understand and generate human-like text based on given prompts.

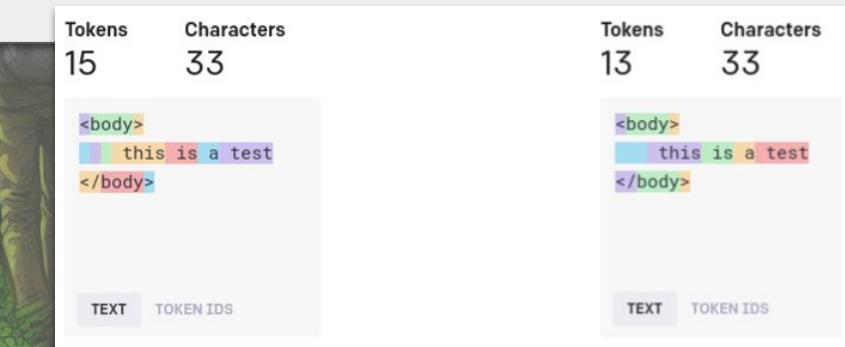
- GPT-3.5, GPT-4, BERT, ...
- Gemini (Google), Gobi (OpenAI) (multimodal models coming soon)
- Open Source: LLAMA 2, Mistral, Orca, ...

Token & Embedding

"A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text [...] (so 100 tokens ~ = 75 words)".

<https://platform.openai.com/tokenizer>

The tokens are represented as a sequence of numbers in a multidimensional vector space. The mapping into the vector space (= embedding) represents the "meaning" of text.





What is being studied in climate research?



Climate research is a vast and multidisciplinary field that aims to understand the Earth's climate system and how it responds to natural and human-induced changes. Here's a breakdown of some key areas of study within climate research, along with a brief explanation of each:

1. **Climate Modeling:** This involves the use of computer simulations to predict future climatic changes based on various scenarios. These models take into account atmospheric, oceanic, and terrestrial processes to simulate how the climate might evolve over time.
2. **Atmospheric Studies:** This encompasses the study of the Earth's atmosphere, including the composition, processes, and interactions of various atmospheric layers. Researchers in this area study phenomena like the greenhouse effect, ozone depletion, and atmospheric circulation patterns.

ChatGPT

... a GPT optimized for conversation and a web application with plugins.

Prompt

... is the natural language input provided to the model (e.g. LLM) to which the model responds.

Prompt Engineering

... is the process of designing, refining, and optimizing prompts to effectively communicate user intent to an LLM.



What is being studied in climate research?



Climate research is a vast and multidisciplinary field that studies the Earth's climate system and how it responds to natural and human-made influences. It involves a breakdown of some key areas of study within climate science, each focusing on different aspects of the system.

- Climate Modeling:** This involves the use of computer models to predict future climate changes based on various scenarios. These models take into account oceanic, atmospheric, and terrestrial processes to simulate how the climate might change over time.
- Atmospheric Studies:** This encompasses the study of the Earth's atmosphere, including the composition, processes, and interactions of various atmospheric layers. Researchers in this area study phenomena like the greenhouse effect, ozone depletion, and atmospheric circulation patterns.



You are a climate science expert. Create a table of the most important topics in climate science with short and concise summaries. Focus on providing an overview and data analysis of the field. Use a neutral, academic tone. Lets think step by step and return the table at the end.



Certainly. When creating a table to summarize the most important topics in climate science, it's important to think about the scope, depth, and audience for this information. Given that you're interested in an overview and data analysis, the focus will be on breadth rather than depth for each topic.

- Topic Selection:** The first step is to identify the key topics that should be included. Climate science is a multi-disciplinary field that encompasses physical sciences, biological sciences, social sciences, and more. Therefore, the topics should reflect this diversity while



Prompt Engineering is ... Context, Context, Context

Topic	Summary	Data Analysis
Global Warming	Increase in Earth's average surface temperature due to human activities like burning fossil fuels.	Global average temperature has risen ~1.2°C since pre-industrial levels.
Greenhouse Gas Emissions	Gases like CO ₂ , CH ₄ , and N ₂ O that trap heat in the atmosphere.	About 43 billion metric tons of CO ₂ -equivalent were emitted in 2019.
Ice Melt & Sea Level Rise	Melting polar ice contributes to rising sea levels.	Sea levels have risen about 8-9 inches since 1880.
Ocean Acidification	CO ₂ absorption by oceans changes their pH level.	Ocean surface pH has decreased from ~8.25 to ~8.14 since the early 20th century.
Climate Models	Computational simulations to predict climate change outcomes.	Models predict a temperature rise of 1.5–4.5°C by 2100 depending on mitigation efforts.
Biodiversity Loss	Climate change negatively affects ecosystems and species.	1 million species at risk of extinction due to climate change and other factors.
Extreme Weather Events	Increase in frequency and intensity of storms, heatwaves, etc.	3-fold increase in extreme weather events since 1980.
Human Health Impacts	Adverse effects on human health due to climate change.	An estimated 250,000 additional deaths per year from 2030 to 2050 due to climate change.
Climate Policy & Governance	Laws and policies to mitigate climate change.	Paris Agreement aims to limit global warming to below 2°C above pre-industrial levels.
Social & Economic	Social inequality and economic repercussions of climate change.	By 2050, climate change could cost the global economy \$7.9 trillion.

Basics of Prompt Engineering - Principles and Techniques

Main components of a prompt

Prompt Engineering Principles

What is ChatGPT good at?

- **Summarisation** “Summarize the following text”
- **Classification** “Classify the following text”
- **Distillation** “Remove all the noise from the following text”
- **Extraction** “Extract the dates from the following text”
- **Translation** “Translate the following text to <language>”
- **Text generation** “Write a <type> about <topic>.”
- **Answering** “Explain <topic> to me”

What is ChatGPT good at?

- **Feedback** “Give me (critical) feedback on the following text”.
- **Improvement** “Improve the following text|code”.
- **Analysis** “Analyse the following data set”

... with prompt engineering you can improve your results 1/2

- **Summarisation**
“Summarize the following text”
“Summarize the following text in three sentences.”
- **Classification**
“Classify the following text”
“Classify the following text into one of the following categories: Business, Technology, Entertainment.”
- **Translation**
“Translate the following text to <language>”
“Translate the following text to <language>, maintaining the tone and context.”
- **Text generation**
“Write a <type> (e.g., essay, article) about <topic>.”
“Compose a <type> about <topic>, emulating the writing style of <style/author>.”

... with prompt engineering you can improve your results 1/2

- **Answering**
 - “Explain <topic> to me”
 - “Explain <topic> to me in layman's terms, focusing on its practical applications.”
 - “Explain <topic> to me like I'm 10 years old (and show me why it's cool or useful)”
- **Feedback**
 - “Give me (critical) feedback on the following text”.
 - “Provide constructive criticism on the following text, highlighting areas for improvement.”
- **Improvement**
 - “Improve the following text|code”.
 - “Improve the following text|code for readability and efficiency.”

Prompt Engineering Principles

1/2

- **Specificity and Clarity:**

Prompts should be clearly and unambiguously formulated to avoid inaccurate or undesirable results.

- **Allow Time for "Thinking":**

It's essential to give large language models sufficient time to process information.

- **Use Context and Examples:**

Providing context and examples can enhance the quality and relevance of the model's responses.

- **Iterative Approach:**

The development of prompts often requires repeated adjustments, making it important to maintain an open attitude and a readiness to refine the prompts based on the received answers.

Prompt Engineering Principles

2/2

- **Understanding GPT's Capabilities:**

The model excels at summarizing, deriving information, converting data into various formats, generating ideas, and expanding texts.

- **Use Explicit Constraints:**

Introducing clear limits or guidelines into the prompt can help direct the way the model responds.

- **Avoid Overloading:** Overly complex or too many tasks at once can be problematic for the model, leading to imprecise or incomplete answers. It's often advisable to break down such requests into more manageable segments.

- **Multimodal Consideration:**

As AI advances, models aren't just text-based. Considering how prompts work in multimodal models (combining text, image, sound, etc.) can be another principle.

Good to know about ChatGPT-4

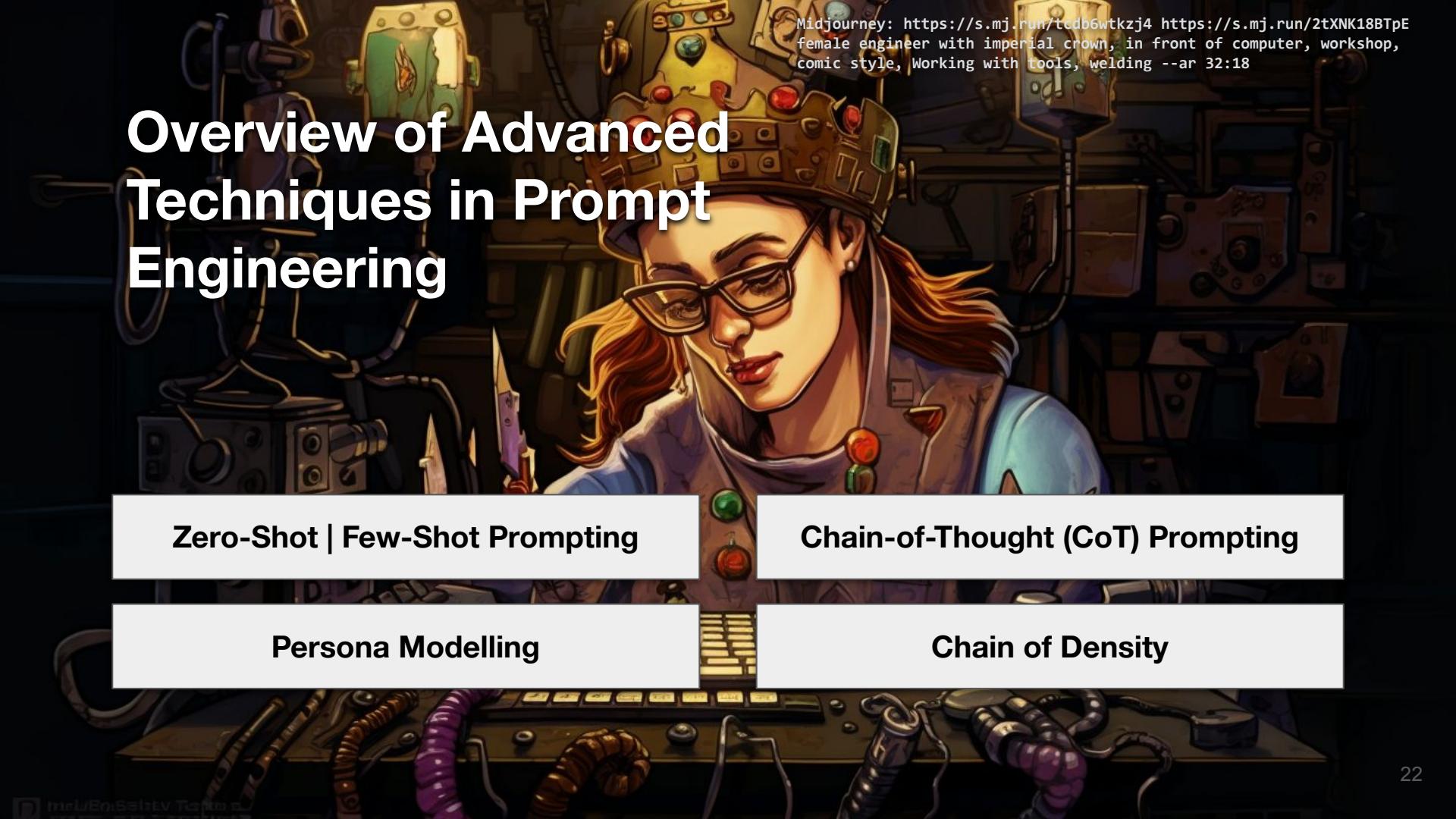
1/2

- **GPT-4's Limitations in Knowledge and Understanding:** GPT-4 doesn't truly "know" or "understand" things in the way humans do. It predicts the next most likely token based on its training.
- **Prompt sensitivity:** Minor (linguistic) alterations in the prompt can already significantly impact the model's performance.
- **GPT-4's "Attention Span":** The model has a token limit, which means that the combined input and output should not surpass 32,000 tokens. Exceeding this limit means that some information will be lost. The context window is extremely important!
- **"Context Contamination":**
The context or style in which a question is asked can influence the model's subsequent answers within the same conversation (*Explain <topic> to me like I'm 10 years old*).

Good to know about ChatGPT-4

2/2

- **GPT-4's Limitations in Calculation and Logic:** The model isn't inherently good at counting, arithmetic, or certain logical operations.
- **Regenerating Answers:** It can be beneficial to regenerate the model's responses, potentially leading to different or more suitable outputs.
- **Safety Protocols:** With AI models, especially those that interact with users, safety precautions to prevent misleading or harmful outputs are vital.



Midjourney: <https://s.mj.run/tcdb6wtkj4> <https://s.mj.run/2tXNK18BTpE>
female engineer with imperial crown, in front of computer, workshop, comic style, Working with tools, welding --ar 32:18

Overview of Advanced Techniques in Prompt Engineering

Zero-Shot | Few-Shot Prompting

Chain-of-Thought (CoT) Prompting

Persona Modelling

Chain of Density

Zero-Shot Prompting

Classify the text into neutral, negative or positive.

Text: I think the vacation is okay.

Sentiment:



Classify the text into neutral, negative or positive.

Text: I think the vacation is okay.

Sentiment:



Sentiment: Neutral

Few-Shot Prompting

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:



The provided text appears to be setting up examples to describe made-up words, "whatpu" and "farduddle." To continue the pattern provided, an example sentence using the word "farduddle" could be:

"The children began to farduddle with excitement when they were told they were going to the amusement park."

Chain-of-Thought (Few-Shot Prompting)

Standard Prompting

Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

Chain-of-Thought Prompting

Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. 

Chain-of-Thought - Example (Zero-Shot Prompting)

You are an expert in visualising climate-related data. Analyse the data very carefully and discuss possible ways of visualising the data.

Take a deep breath and think step by step.



Careful Data Analysis for Visualization

Step 1: Understand the Dimensions and Variables

The dataset has three dimensions: time, latitude, and longitude, and it primarily focuses on air temperature. Understanding the dimensions is critical because it informs us about the types of visualizations that can be effectively used.

Step 2: Consider the Research Context

The dataset is used for climate system research, often aimed at studying long-term trends, seasonal cycles, and spatial variations. The choice of visualizations should facilitate these kinds of analyses.

Step 3: Identify Key Metrics and Patterns

Before visualizing, it's crucial to identify what we're trying to showcase or investigate. This could range from average temperatures and variances to more complex metrics like anomalies over time or geographical hotspots.

“Take a deep breath”

Scorer	Optimizer / Source	Instruction position	Top instruction	Acc
<i>Baselines</i>				
PaLM 2-L	(Kojima et al., 2022)	A_begin	Let's think step by step.	71.8
PaLM 2-L	(Zhou et al., 2022b)	A_begin	Let's work this out in a step by step way to be sure we have the right answer.	58.8
PaLM 2-L		A_begin	Let's solve the problem.	60.8
PaLM 2-L		A_begin	(empty string)	34.0
text-bison	(Kojima et al., 2022)	Q_begin	Let's think step by step.	64.4
text-bison	(Zhou et al., 2022b)	Q_begin	Let's work this out in a step by step way to be sure we have the right answer.	65.6
text-bison		Q_begin	Let's solve the problem.	59.1
text-bison		Q_begin	(empty string)	56.8
<i>Ours</i>				
PaLM 2-L	PaLM 2-L-IT	A_begin	Take a deep breath and work on this problem step-by-step.	80.2
PaLM 2-L	PaLM 2-L	A_begin	Break this down.	79.9
PaLM 2-L	gpt-3.5-turbo	A_begin	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
PaLM 2-L	gpt-4	A_begin	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5
text-bison	PaLM 2-L-IT	Q_begin	Let's work together to solve math word problems! First, we will read and discuss the problem together to make sure we understand it. Then, we will work together to find the solution. I will give you hints and help you work through the problem if you get stuck.	64.4
text-bison	text-bison	Q_end	Let's work through this problem step-by-step:	68.5
text-bison	gpt-3.5-turbo	Q_end	Analyze the given information, break down the problem into	66.5

Chain-of-Density Prompting

<https://arxiv.org/pdf/2309.04269.pdf>

You will generate increasingly concise entity-dense summaries of the above article.

Repeat the following 2 steps 5 times:

Step 1: Identify 1-3 informative entities (delimited) from the article which are missing from the previously generated summary.

Step 2: Write a new denser summary of identical length which covers every entity and detail from the previous summary plus the missing entities.

A missing entity is:

Relevant: to the main stories.

Specific: descriptive yet concise (5 words or fewer).

Novel: not in the previous summary.

Faithful: present in the article.

Anywhere: located in the article.

Guidelines:

The first summary should be long (5-10 sentences, ~500 words), yet highly non-specific, containing little information beyond the entities marked as missing. Use overly verbose language and fillers (e.g., “this article discusses”) to reach ~500 words. Make every word count. Rewrite the previous summary to improve flow and make space for additional entities. Make space with fusion, compression, and removal of uninformative phrases like “the article discusses”. The summaries should become highly dense and concise, yet self-contained, e.g., easily understood without the article. Missing entities can appear anywhere in the new summary. Never drop entities from the previous summary. If space cannot be made, add fewer new entities. Remember: Use the exact same number of words for each summary.”

Text:

...

...

Data - Information - Knowledge. An information science perspective

The ubiquity of data, information and knowledge in our society can hardly be overstated. These concepts are fundamental to all scientific disciplines and, as a result, there are many definitions of these terms. While this thesis cannot cover all the discussions of these concepts in different fields, it will focus primarily on information science and how it relates to working with historical information in information systems. However, it should be noted that there is no universal definition of information, even within the field of information science (Kuhlen et al. 2023, 3-7).

A key publication exploring the concepts of data, information, and knowledge is (Favre-Bulle 2001). The author offers a comprehensive discussion of these terms and how they relate to other concepts such as syntax, semantics, perception, and cognitive systems. The importance of context is also highlighted, which is particularly relevant for the discussion of historical information. Other works that provide an overview of different disciplinary perspectives on the term “information” include (Ott 2004; Capurro 2000). These authors emphasise that definitions of information vary across different disciplines. In a more general approach, (Meadows 2019) provides an analysis of data, information, and knowledge.

(Kuhlen et al. 2023) covers various facets of information science, including but not limited to definitions, methods and theories, as well as relevant domains such as archives, libraries and information visualisation. The textbook also includes practical applications of information science.

Chain of Density - Example

<https://chat.openai.com/share/9b562d4c-8ff1-4752-98f7-7d738abb370d>

First Summary (5-10 sentences, ~500 words):

This article offers an in-depth exploration into the complexities and nuances of the concepts of data, information, and knowledge, with a primary focus on the field of information science. The text traverses a landscape that includes discussions on the myriad definitions these terms can have, depending on the discipline or context in which they are being used. The article cites various scholars and their works to bolster its argument, such as (Favre-Bulle 2001) who offers a comprehensive discussion on how these terms relate to concepts like syntax, semantics, and cognitive systems. Another noteworthy point the article delves into is the significance of context in interpreting historical information, which is particularly relevant to the field of information science. Other academic works that provide insights into the disciplinary perspectives on these terms are also highlighted. For example, (Kuhlen et al. 2023) covers a wide range of topics in information science, including definitions, methods, theories, and practical applications. In addition, the text briefly touches upon the concept of the DIKW Hierarchy, also known as the Knowledge Pyramid, as a model that helps in defining information and related terms. This hierarchy has undergone various adaptations over time to include new concepts like human agency and complexity. Different forms of representation for the DIKW model are discussed, such as a pyramid shape and a field with x and y axes. Essentially, the article serves as an extensive overview and an academic resource on the concepts of data, information, and knowledge, drawing on a plethora of academic works to offer a well-rounded view.

Entities added: (Favre-Bulle 2001), (Kuhlen et al. 2023), DIKW Hierarchy

Third Summary (~500 words):

Focusing on the intricate landscape of data, information, and knowledge, particularly in information science, this article synthesizes academic contributions from experts like (Favre-Bulle 2001) and (Kuhlen et al. 2023). The former elaborates on the interplay between these key terms and other frameworks such as syntax and semantics. The latter discusses a range of subjects including the importance of historical data, and practical applications in archives and libraries. Within this scope, the article also addresses the evolving DIKW Hierarchy, which is instrumental in delineating these terms. The model has been adapted to encompass new dimensions like human agency and complexity. Alongside, different visual formats of the DIKW model like pyramids and fields are highlighted to elucidate the hierarchical nature of the terms. In essence, the article serves as a nuanced, interdisciplinary guide to understanding the complexities of data, information, and knowledge.

Entities added: syntax and semantics, practical applications, new dimensions

Second Summary (~500 words):

Delving into the complex web of data, information, and knowledge, this article focuses mainly on their relevance to information science. The text refers to scholars like (Favre-Bulle 2001) who discuss these concepts' relationships to other cognitive and linguistic frameworks like syntax and semantics. It also brings to light the significance of context in interpreting historical data, a topic elaborated by (Kuhlen et al. 2023). They offer a multi-faceted view that encompasses definitions, methodologies, theories, and practical domains like archives and libraries. Furthermore, the DIKW Hierarchy or Knowledge Pyramid, is presented as a model for understanding these terms. This model has evolved to include concepts such as human agency and complexity. Various visual representations of the DIKW model are also discussed, including pyramid and field forms, all aimed at illustrating the hierarchical relationships among data, information, and knowledge. The article thus serves as a comprehensive resource that pulls from a wide array of academic contributions to provide a well-rounded understanding of these key concepts.

Entities added: historical

Fourth Summary (~500 words):

Centering on the complexities of data, information, and knowledge within information science, the article draws from seminal works by (Favre-Bulle 2001) and (Kuhlen et al. 2023). It covers the intricate relationship between these terms and frameworks like syntax and semantics, as well as their practical applications in areas like archives and libraries. Importantly, the article sheds light on historical data's relevance, explaining how context is crucial for interpretation. The evolving DIKW Hierarchy is discussed, incorporating new dimensions such as human agency and complexity. Various visual depictions of the DIKW model, like pyramids and fields, are also featured, clearly mapping out the hierarchical relationships among the key terms. Ultimately, the article serves as a nuanced guide to the interdisciplinary world of data, information, and knowledge.

Entities added: context for interpretation, visual depictions, interdisciplinary world

Key Findings

<https://ki.uni-graz.at>
<https://chpollin.github.io/GM-DH/>

- Emergent Capabilities of GPT-4
 - In-context learning - quickly assimilate and apply truly new information
 - Logical Reasoning via Chain-of-Thought Prompting
- ChatGPT is an expert (!) tool that can help with various research tasks.
- You need to know how to use LLM, ChatGPT (+ plugins), as well as prompting techniques.
- This technology is still in its early stages of development. Multimodal models such as Gemini (Google) and Gobi (OpenAI) will be (much) better.
- Open source models are also getting better. These are the LLMs we want to use in research (but we should also learn from the LLMs of the tech giants, because their models will be even (much?!?) better.
- It is a new paradigm of research: it is about implementing new workflows and tools to stabilise and improve the results of LLMs for our specific use cases.