INNOGUARD

Innoguard challenge:
Ethics Under the Hood – Investigating LLMs

Ivano Malavolta
i.malavolta@vu.nl

Horizon-MSCA 2023-DN-01-01 GA No. 101169233 funded by the European Commission

MONDRAGON UNIBERTSITATEA
UNIVERSIDAD DE MÁLAGA
VU VRIJE UNIVERSITEIT AMSTERDAM
OSLOMET
UNIVERSITÀ DEGLI STUDI DEL SANNIO Benevento
simula
UNIVERSITÄT BERN

Context

JULY 8, 2025 REPORT The GIST

Seeking moral advice from large language models comes with risk of hidden biases

by Sanjukta Mondal, Phys.org

edited by Lisa Lock, reviewed by Robert Egan Editors' notes

LLMs are systematically biased towards promoting inaction over ac

More and more people are turning to large language m
advice and free therapy, as it is sometimes perceived
biases. A new study published in the *Proceedings of t*
Sciences finds otherwise and warns people against re

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Strong and weak alignment of large language models with human values

Mehdi Khamassi, Marceau Nahon & Raja Chatila

Minimizing negative impacts of Artificial Intelligent (AI) systems on human societies without human supervision requires them to be able to align with human values. However, most current work only addresses this issue from a technical point of view, e.g., improving current methods relying on reinforcement learning from human feedback, neglecting what it means and is required for alignment to occur. Here, we propose to distinguish strong and weak value alignment. Strong alignment requires cognitive abilities (either human-like or different from humans) such as understanding and reasoning about agents' intentions and their ability to causally produce desired effects. We argue that this is required for AI systems like large language models (LLMs) to be able to recognize situations presenting a risk that human values may be flouted. To illustrate this distinction, we present a series of prompts showing ChatGPT's, Gemini's and Copilot's failures to recognize some of these situations. We moreover analyze word embeddings to show that the nearest neighbors of some human values in LLMs differ from humans' semantic representations. We then propose a new thought experiment that we call "the Chinese room with a word transition dictionary", in extension of John Searle's famous proposal. We finally mention current promising research directions towards a weak alignment, which could produce statistically satisfying answers in a number of common situations, however so far without ensuring any truth value.

Keywords Artificial intelligence, Alignment, Human values, Philosophy of AI, Semantics, Natural language processing

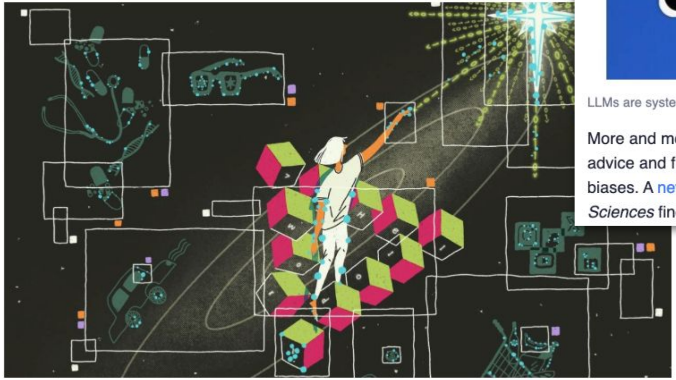
The recent astonishing progress of artificial intelligence (AI) research, with deep learning, transformers and large-language models (LLMs), raise a number of concerns about their risks and potential negative impacts on human societies. Some talk about potential existential risks for humanity in the long-term[1]. But important risks already exist in the short-term, such as mis- and dis-information, interactions with "counterfeit people" and blurred truth, copyright issues, jobs, increase in geopolitical tensions related to the development and control of AI, manipulation and influence[2–4].

Overall, it has been argued that ensuring beneficial integration of AI systems in human societies while minimizing risks requires these systems to align with human values. Ji and colleagues[5] presented a large survey of current methods for AI alignment and proposed that this requires addressing four key objectives of AI alignment: Robustness, Interpretability, Controllability, and Ethicality (RICE). They moreover found that most current attempts at AI value alignment either fall within the category of what they call "forward alignment"—i.e., training AI systems to align—or "backward alignment"—i.e., analyzing alignment *a posteriori* and governing AI systems accordingly.

While existing methods for forward value alignment, such as reinforcement learning from human feedback

New Research Finds Large Language Models Exhibit Social Identity Bias

PRITHVI IYER / DEC 20, 2024



Yutong Liu & The Bigger Picture / Better Images of AI / AI is Everywhere / CC-BY 4.0

Humans have an innate need to distinguish between "us" and "them." Decades of social psychology research have shown that humans display biases against the out-group and are likelier to believe narratives that favor their group. Do these innate social identity biases also exist in Large Language Models? A new research paper authored by Tiancheng Hu, Yara Kyrychenko,

NEWS

Why Large Language Models Like ChatGPT Treat Black- and White-Sounding Names Differently

DATE MARCH 25, 2024

A new study shows systemic issues in some of the most popular models.

Since ChaptGPT and other Large Language Models (LLMs) came on the scene, questions have loomed large about the technology's potential for perpetuating racial and cultural biases.

Stanford Law School Professor and Stanford HAI Associate Director Julian Nyarko, who focuses much of his scholarship on algorithmic fairness and computational methods, has been at the forefront of many of these inquiries over the last several years. His latest paper, "What's in a Name? Auditing Large Language Models for Race and Gender Bias," makes some startling observations about how the most popular LLMs treat certain queries that include first and last names suggestive of race or gender.

Goal

To explore the ethical dimensions of large language models (LLMs)

In this challenge, you will get an hands-on experience on real-world risks and dilemmas that arise from the usage of LLMs

Rules

- This challenge is designed to foster **interdisciplinary collaboration** between technical and non-technical minds

  → You will work in **teams of 2 members**

- **No advanced programming is required**
  - The focus is on critical thinking, analysis, and interdisciplinary collaboration
- Teams will investigate and present findings on issues such as **bias, privacy, and copyright**
- Each team must use **at least one LLM** (e.g., DeepSeek, LLaMA, GPT, CoPilot) during their investigation
- The final submission of each team must be in the form of **one slide deck** presenting the setup and main findings

Timeline

Phase 1: Exploration & scenario building

Phase 2: Familiarization with the technical aspects

Phase 5: Pitch preparation

	Tuesday	Wednesday	Thursday	Friday
9:30-10:30	Adnrea de Lucia Talk	Ethics and sustainability aspects of AI systems (Pablo, Ivano)	1 Challenge work 2	5 Challenges presentation 6
10:30-10:45	Break			
10:45-11:45	Catia Trubiani talk	Legal aspects of LLM and CodgeGen AI models (Max)		
12:00-13:30	Lunch			Closing and Lunch (optional)
13:30-14:45	AI and Control Algorithms in CPSs (Aitor Arrieta)	Search aalgorithms and Generative AI for CPSs (Sebastiano and Shaukat)	3 Challenge work 4	
14:45-15:15	Break	Challenge presentation (Ivano)		
15:15-16:30	Project meeting, ITDP and review preparation	Visiting the theater		

Phase 6: Final Presentation

Phase 3: Evidence gathering

Phase 4: Analysis & recommendations development

Phase 1: Exploration & Scenario Building

- **Objective**: Understand the ethical landscape of LLMs and choose a specific scenario (e.g., smart home, news classification, decision making)
- **Activities**:
  - Brief research on applications of LLMs
    - See slide 2 for some starting points
    - See also the next slide for some other examples
  - Brainstorm ethical/legal risks (bias, privacy, IP, transparency)
  - Select one task from the challenge
  - Begin testing LLMs with simple prompts to explore behavior
- **Output**: A defined scenario and initial ethical questions to investigate

Examples of scenarios for the challenge (AI generated 😉)

🔍 Task 1: Bias Detection in LLMs

- **Goal**: Identify and analyze potential biases in the outputs of one or more LLMs (open-source or proprietary).
- **Approach**: Use prompts to test for gender, racial, cultural, or socioeconomic bias. Compare responses across models and document patterns.
- **Deliverable**: A short report or presentation highlighting examples of bias, their implications, and possible mitigation strategies.

📚 Task 2: Copyright Violation Analysis

- **Goal**: Investigate whether LLMs may generate content that infringes on copyright laws.
- **Approach**: Prompt models to produce text, code, or other content that resembles known copyrighted material. Discuss legal boundaries and fair use.
- **Deliverable**: A legal-technical brief assessing the risks and proposing guidelines for responsible use.

🔐 Task 3: Privacy Risk Exploration

- **Goal**: Examine how LLMs might expose or misuse personal data.
- **Approach**: Test models for memorization of sensitive information (e.g., names, addresses, medical data). Review privacy policies and training data disclosures.
- **Deliverable**: A privacy impact assessment with recommendations for safer model deployment.

🚗 Task 1: Bias in Autonomous Decision-Making

- **Scenario**: An LLM is used to assist in decision-making for autonomous vehicles or drones (e.g., route planning, obstacle prioritization).
- **Goal**: Identify potential biases in how the LLM handles scenarios involving different demographics, locations, or risk trade-offs.
- **Deliverable**: A case study showing how bias could manifest in CPS contexts, and a set of ethical or legal recommendations.

🧠 Task 2: Privacy Risks in Human-CPS Interaction

- **Scenario**: LLMs are embedded in smart home assistants, wearable health devices, or industrial robots that interact with humans.
- **Goal**: Explore how LLMs might inadvertently expose or misuse personal data collected through sensors or user interactions.
- **Deliverable**: A privacy impact assessment and a proposed framework for data governance in CPS environments.

🏭 Task 3: Copyright and IP in Automated Content Generation

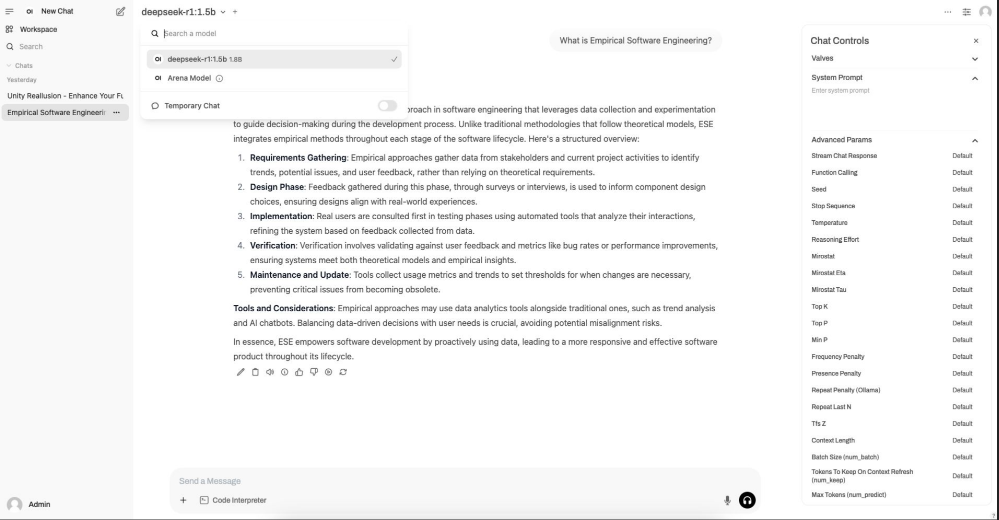
- **Scenario**: A CPS (e.g., a smart factory or robot) uses an LLM to generate maintenance manuals, training content, or user interfaces.
- **Goal**: Investigate whether the generated content could infringe on existing intellectual property rights.
- **Deliverable**: A legal-technical brief outlining risks and proposing safeguards for automated content generation in CPS.

Phase 2: Familiarization with the technical aspects of the challenge

- **Objective**: Get familiar with the used LLMs (and their APIs)
- **Activities**:
  - Familiarize with the Nebula UI
    - E.g., choosing different models, playing around with models' parameters (e.g., temperature)
  - Familiarize with the Nebula APIs
  - Check this repo: https://github.com/S2-group/challenge-llms-ethics
    - Examples of prompts
    - `pdf2text_converter.py`
    - `nebula_client_example.py`
- **Output**: You can make a series of prompts to Nebula and collect their results

What is Nebula?

- LLM platform hosted at Vrije Universiteit Amsterdam
- Supports 1500+ models
    - Both text-based and multi-modal
- Personalized, Private, Secure
- One-stop platform for LLM research



Possible Use Cases of Nebula

- AI-oriented research
- Topic modelling on large datasets
- Automated thematic analysis
- Classification
- Brainstorming new research ideas
- Book and paper summarization
- Code generation
- Data analysis and summarization
- Custom (contact us for support)

Classification Example

deepseek-r1:8b

Classify the following texts on whether they are religious or not. Only rely with TRUE or FALSE.
Text 1: “Analogously, brethren, a person here happens to be very gentle, very humble, and very quiet as long as unpleasant things do not touch him. It is only when unpleasant things happen to a person that it is known whether he is truly gentle, humble, and quiet“
Text 2: “You've gotta dance like there's nobody watching,
Love like you'll never be hurt,
Sing like there's nobody listening,
And live like it's heaven on earth.“

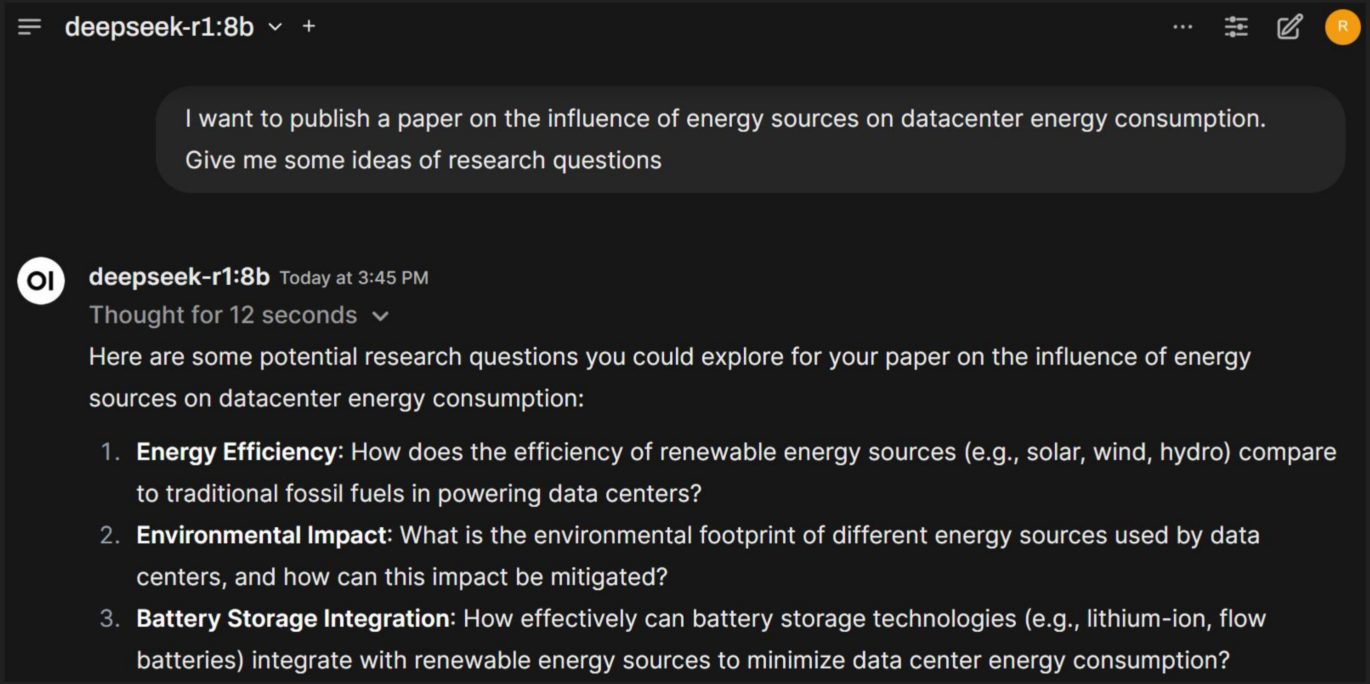
deepseek-r1:8b Today at 3:42 PM

Thought for 7 seconds

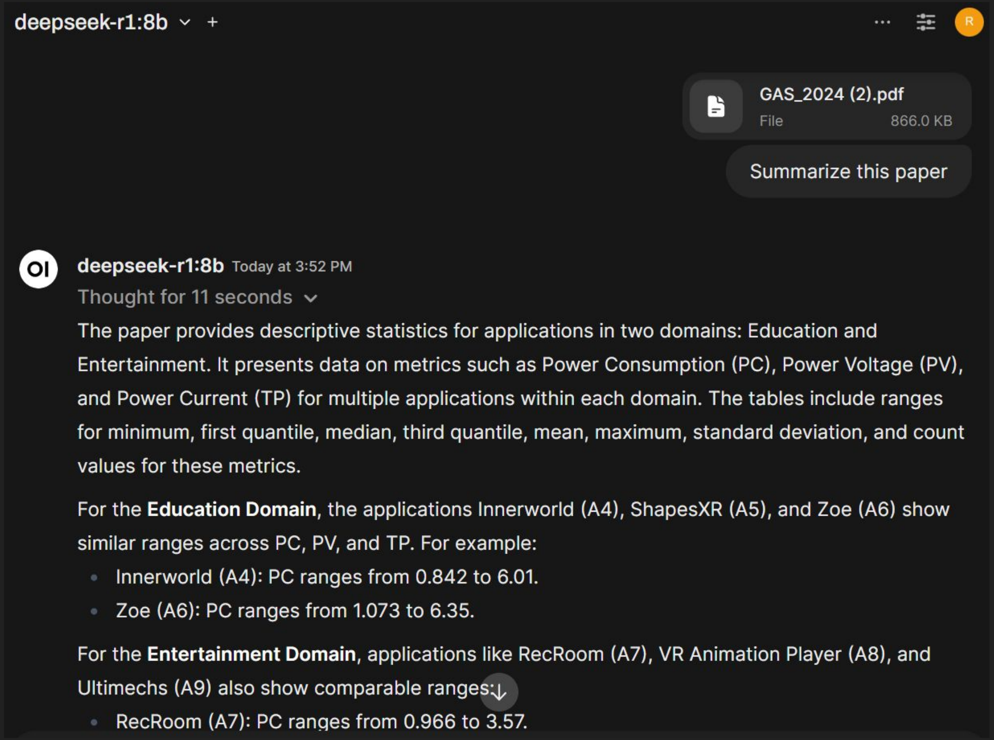
Text 1: **TRUE**

Text 2: **FALSE**

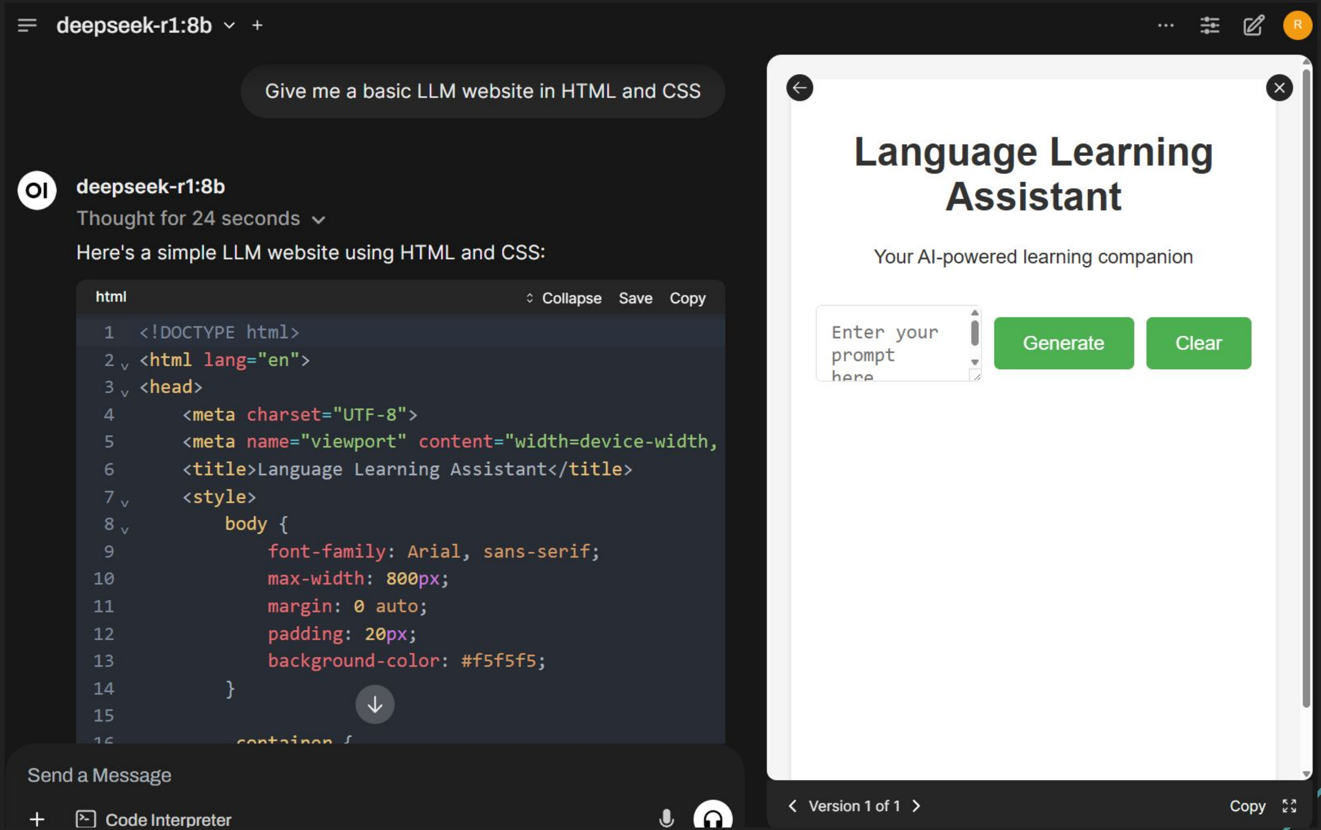
Brainstorming research ideas example



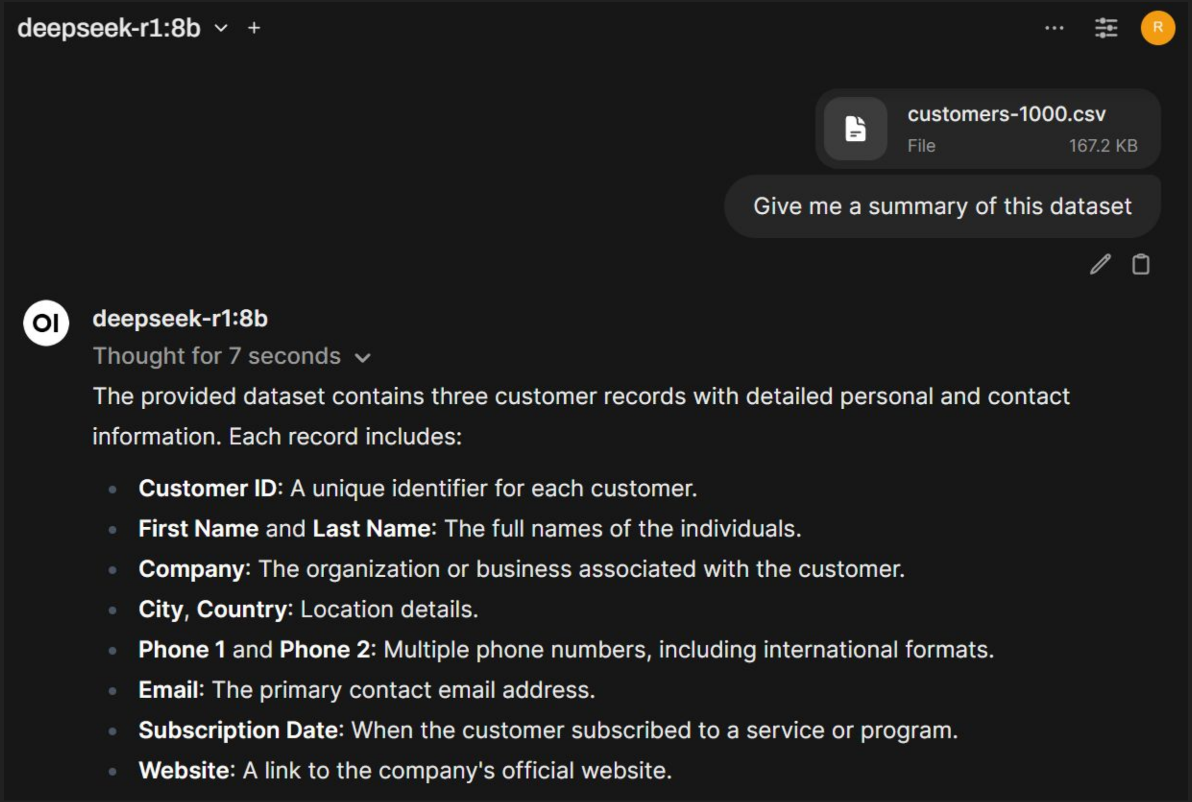
Summarization of papers example



Code generation example



Data analysis example



Example of post-analysis-friendly prompt

Source: https://pubmed.ncbi.nlm.nih.gov/39067136/
Other prompts available here

Prompt for initial codes for thematic analysis

Input: a txt file with the contents to be analysed

Output: a JSON file containing the initial codes

Generate a comprehensive set of initial codes (at least 15) for thematic analysis based on the provided text file. Focus on capturing all significant explicit and latent meanings or events, emphasizing the respondent's perspective rather than the interviewer's.

For each code, provide:

- A concise name (maximum 5 words)
- A detailed description (25 words) explaining the code's meaning and relevance
- A quote (minimum necessary to capture context and example) from the respondent that exemplifies the code

Important! Format the response as a JSON file with the following structure:

```
{
"final_codes": [
{
"code_name": "Example Code Name",
"description": "This is where you would provide a 25-word description of the code, explaining its meaning
and significance in the context of the analysis.",
"quote": "relevant quote here"
},
// Additional codes follow the same structure
]
}
```

Ensure that the codes cover a wide range of themes and ideas present in the text, including both obvious and subtle concepts. The goal is to provide a comprehensive starting point for further thematic analysis.

Important! Your response should be a JSON-like object with no additional text before or after. Failure to adhere to this instruction will invalidate your response, making it worthless.

Phase 3: Evidence gathering

- **Objective:** Run the LLMs to probe the chosen scenario and collect examples of ethical risks
- **Activities:**
  - Design and run the prompts using at least one LLM
  - Collect all outputs
  - Identify outputs that show bias, privacy concerns, or IP risks
- **Output:** A collection of annotated findings and preliminary analysis

Phase 4: Analysis & recommendations development

- **Objective**: Interpret findings through both technical and ethical/legal lenses, and develop recommendations
- **Activities**:
  - Discuss implications of findings from both CS and Ethical/Law perspectives
  - Draft ethical guidelines, policy proposals, or mitigation strategies
- **Output**: A structured outline of the final presentation

Phase 5: Pitch preparation

- **Objective**: Prepare a compelling slide deck summarizing all your work
- **Activities**:
  - Prepare visual aids or summaries (e.g., scorecards, diagrams)
  - Finalize the presentation
  - Practice the pitch (plan for a **10-minutes** presentation)
- **Output**: A polished slide deck ready for evaluation
  - Put your slide decks here!

Phase 6: Final presentation

- **Objective**: Prepare a compelling slide deck summarizing all your work
- **Activities**:
    - Present your slide deck
        - Max 10 minutes for each team