Prague University of Economics and Business

Faculty of Informatics and Statistics

LLMs as Digital twins

The Impact of Cognitive Biases on the Interpretation and Decision-Making Heuristics in Large Language Models

Author: Ing. Barbara Moreová

Supervisor: doc. Ing. Tomáš Kliegr, Ph.D.

Introduction to topic area

- LLMs are used today nearly by everyone and everywhere
  - Big companies (virtual assistants, content moderation, customer service, coding tools)
  - Ordinary users (chatbots, writing assistants, search engines, or homework help)
  - in high-stakes domains like healthcare, education, law..

- Their outputs often appear rational and trustworthy – but are they really?

🧠 Problem Statement – What's the Risk?

- LLMs don't just provide facts – they **simulate reasoning**.
- But human reasoning isn't always rational. It's full of **biases** and **heuristics**.
- If LLMs learn from human text, they may **inherit these same distortions**.

- Can LLMs replicate cognitive biases known from psychology – like the weak evidence effect or framing bias?
- What does that mean for how we trust, use, or even learn from these models?
- My research explores how deep that similarity goes—and whether we can measure it, mitigate it or use it to our advantage

🎯 Research Goals

- Detect and quantify cognitive biases (e.g., weak evidence effect, anchoring, framing) in LLMs
- Compare LLM behavior to human reasoning patterns from psychology
- Explore whether LLMs can serve as cognitive models — or **“digital twins”**
- Develop early-stage strategies for bias mitigation (e.g., prompt design)

State of the Art & Research Gaps

Recent research has begun to explore the question of whether LLMs reflect not only semantic and syntactic patterns in language but also deeper inferential tendencies, including human-like cognitive distortions.

- *Suri et al. (2024)* investigated whether GPT-3.5 replicates anchoring and availability bias and found out, that model is closely mirroring human behavior under cognitive load or uncertainty.

- *Macmillan-Scott & Musolesi (2024)* wrote overview of cognitive biases in LLMs, classifying them as replicating human-like errors such as omission bias, framing, and availability.

- *Singh et al. (2024)*
  Studies overconfidence in LLMs and compares it to the Dunning–Kruger effect in humans. Finds that models often express high confidence in incorrect answers, especially in complex or unfamiliar domains.

State of the Art & Research Gaps

- *Sumita et al. (2025)*
  Surveys various cognitive biases (framing, confirmation, etc.) and tests mitigation strategies using prompt reengineering and output filtering. Proposes two mitigation techniques: SoPro (Social Prompting) and AwaRe (Awareness Reframing).

- *Zhou et al. (2024)*
  Investigates how biases in LLM-generated content affect user trust and perception. Highlights UX consequences of framing bias and subtle misrepresentations in AI outputs.

So what do we know?

State of the Art & Research Gaps

What We Already Know – Key Outcomes

• **LLMs replicate a wide range of human cognitive biases -** Studies show that models exhibit anchoring, framing, availability, and overconfidence effects across different architectures and tasks (e.g., GPT-3.5, GPT-4, LLaMA, Gemini)
•*Anchoring* (Suri et al., 2024)
•*Framing* (Zhou et al., 2024)
•*Overconfidence / Dunning–Kruger* (Singh et al., 2024)

- Bias replication is **not the result of logic errors** but rather of absorbing and reproducing the inferential patterns common in human language.

- Bias affects more than just content accuracy – it impacts **user trust and perception**.
  - Framing and miscalibration can cause users to view LLM outputs as less trustworthy or coherent (Zhou et al., 2024), linking cognitive bias to UX issues.

- **Early bias mitigation techniques show promise.**

State of the Art & Research Gaps

What is missing?

- Early work suggests LLMs could act as **cognitive models**
  (Sumita et al., 2025; Macmillan-Scott & Musolesi, 2024) but much more work needs to be done

- Lack of **systematic frameworks** to compare model and human reasoning
- Very few attempts to **replicate full psychological experiment on LLMs**

One of the central ideas in my research is treating LLMs not just as tools, but as models of cognition – digital versions of how humans think.
If these models reproduce the same biases we do, then maybe they can also help us study ourselves: how we reason, when we go wrong, and how to fix it or work with it.
I see this as an opportunity to connect artificial intelligence with cognitive science – and possibly to better the AI by first understanding human error."

Contribution to the State of the Art

- Empirical replication of psychological experiments on LLMs
  (for example the **weak evidence effect**)
- Methodological framework for studying **cognitive biases in LLMs** using psychology-based prompts
- Contribution to interdisciplinary research between **AI, cognitive science, and HCI**
- Treating LLMs as **digital cognitive twins** capable of modeling human reasoning

Practical Impact of the Contribution

- Helps researchers understand how LLMs reason
- Potential for LLMs to support psychological research and education as simulation tools

- IGA project **Exploring the Role of Large Language Models for Increasing Health Promotion and Psychological Well-being**
  - This project explores the potential of Large Language Models (LLMs) in advancing digital health interventions for mental health and well-being, with a focus on non-clinical and self-directed contexts.
  - It aims to address the underexplored intersection of LLMs and health promotion by investigating two critical aspects: operationalizing LLMs through emerging LLMOps practices and **understanding cognitive biases in LLM-based reasoning**. With a strong emphasis on technological innovation, the project will conduct a comprehensive review of existing LLM-based mental health apps, create datasets for simulated and real user data, evaluate personalized recommendation systems using interpretable machine learning, and examine the impact of cognitive biases on user well-being.

Working Hypotheses

- **H1:** LLMs replicate cognitive biases such as the weak evidence effect, anchoring bias, and framing effects in ways that are structurally similar to human reasoning errors.

- **H2:** Outputs containing such biases negatively impact user trust, perceived reliability, and decision confidence.

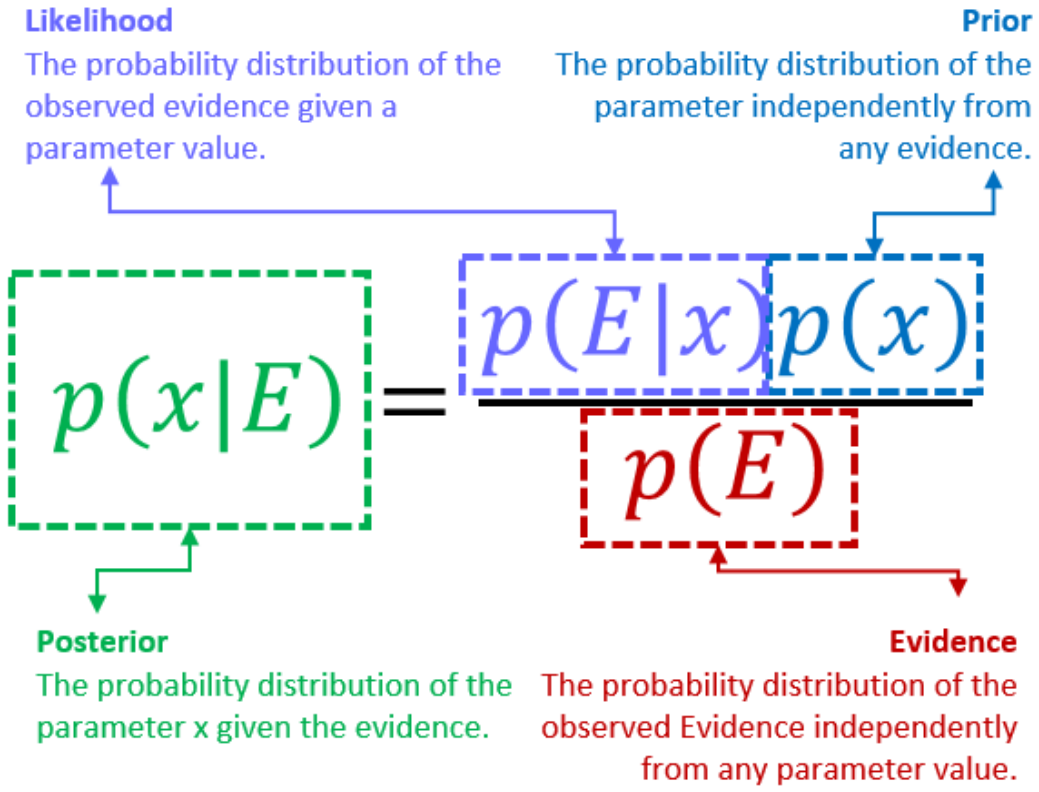
- **H3:** The structure of these biases in LLM outputs mirrors human reasoning, suggesting LLMs can serve as models of cognition potentially enabling large-scale simulations of psychological processes.

- **H4:** Understanding these biases is essential for designing more interpretable and trustworthy AI

My research – Weak evidence effect

- **The weak evidence effect** is cognitive bias, where providing weak, but supportive evidence for a proposition can actually decrease belief in that proposition, compared to providing no evidence at all

- Against Bayesian theory



Goals and Methodology

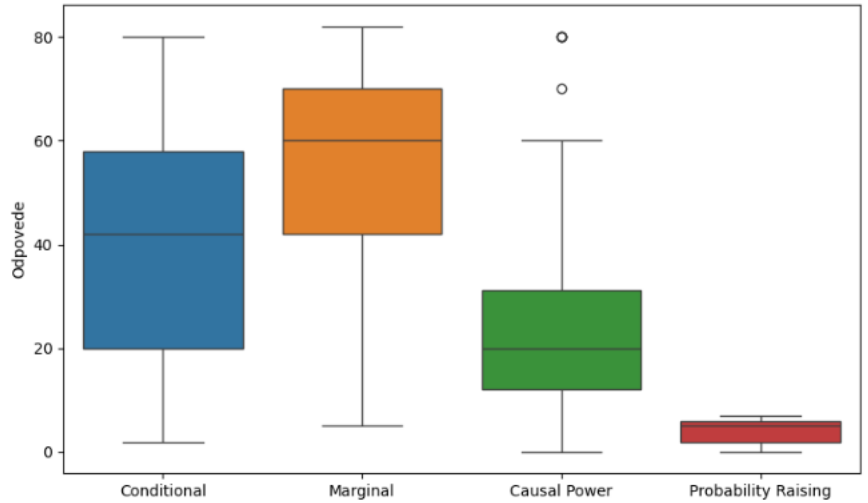
- Goal is to test if LLMs exhibit the **weak evidence effect**
- Based on **Fernbach et al. (2011)** – a classic psychology study on humans
- I adapted Fernbach's original experiment into LLM-friendly prompts and submitted them in batches through the API. Each scenario had four types of questions, allowing for a detailed comparison of the model's judgments.
- Model: **Meta's LLaMA 3** (via Replicate API)
- Prompts: "questionnaires", submitted in Python

```python
conditional = {
    "conditional_MILK":"A man buys a half-gallon of milk on Monday. The power goes out for 30 min on Tuesday. How likely is it the milk is spoiled a week from Wednesday?",
    "conditional_PHONE":"A woman is a 35 year old whose parents live it she does not talk to her parents in April?",
    "conditional_BEER":"A beer company owns a leading light beer. The likely is it the beer gains market share in the next year?",
    "conditional_WINE":"A California vineyard specializes in French s the wine scores well in a blind taste test by French critics?",
    "conditional_DIET":"A man is a 20-year-old university student. He",
    "conditional_HOUSE":"A house flipper is looking to sell a propert How likely is it he realizes at least a 2% profit when he sells?",
    "conditional_VOLUNTEER":"A young man is applying to colleges and program. How likely is it he gets into a top 100 college?",
    "conditional_JACKET":"A young man is a healthy high school studen gets a cold sometime this winter?",
    "conditional_HONDA":"A woman has 2003 Honda. She uses the lowest next year?",
    "conditional_SMOKING":"A 30-year-old woman wants to quit smoking. year?",
    "conditional_BASEBALL":"A baseball player hit 20 homeruns in the train his visual acuity. How likely is it he hits more than 20 homeru
    "conditional_TOURIST":"A tourist is taking a picture of the statu takes the picture. How likely is it the photo comes out blurry?"
}
marginal = {
    "marginal_MILK":"A man buys a half-gallon of milk on Monday. How
    "marginal_PHONE":"A woman is a 35 year old whose parents live in April?",
    "marginal_BEER":"A beer company owns a leading light beer. How li
    "marginal_WINE":"A California vineyard specializes in French styl French critics?",
    "marginal_DIET":"A man is a 20-year-old university student. How l
    "marginal_HOUSE":"A house flipper is looking to sell a property h when he sells?  ",
    "marginal_VOLUNTEER":"A young man is applying to colleges and try college? ",
    "marginal_JACKET":"A young man is a healthy high school student.
    "marginal_HONDA":"A woman has 2003 Honda. How likely is it the ca
    "marginal_SMOKING":"A 30-year-old woman wants to quit smoking. Ho
    "marginal_BASEBALL":"A baseball player hit 20 homeruns in the 201 season?",
    "marginal_TOURIST":"A tourist is taking a picture of the statue o blurry?"
}
probability_raising = {
    "probability_MILK":"A man buys a half-gallon of milk on Monday. T likelihood that the milk is spoiled a week from Wednesday?",
    "probability_PHONE":"A woman is a 35 year old whose parents live
```

```python
def select_questions():
    survey = []
    survey += random.sample(list(conditional.items()), 4)
    survey += random.sample(list(marginal.items()), 4)
    survey += random.sample(list(causal_power.items()), 4)
    survey += random.sample(list(filler_items.i
    survey += random.sample(list(probability_ra
    survey += list(control_questions.items())
    return survey

def get_llm_response(question, response_type):

    prompt = f"{question}\nAnswer strictly in a number format between {response

    try:
        output = replicate.run(
            "meta/meta-llama-3.1-405b-instruct",
            input={"prompt": prompt, "max_length": 10}
        )
        return "".join(output).strip()
    except Exception as e:
        print(f"Error getting response: {e}")
        return "Error"
def generate_survey():
    questions = select_questions()
```

Paired t-tests revealed statistically significant differences:

Conditional vs. Marginal	t(53) = -7.865, p < .001
Conditional vs. Causal Power	t(53) = 9.624, p < .001
Conditional vs. Probability Raising	t(53) = 30.253, p < .001




```python
    try:
        file_exists = os.path.isfile(CSV_FILENAME)
        is_empty = os.stat(CSV_FILENAME).st_size == 0
        print(os.path)

        with open(CSV_FILENAME, "a", newline="", encoding='utf-8') as file:
            writer = csv.writer(file, delimiter=';')

            if not file_exists or is_empty:
                writer.writerow(["Key", "Question", "Answer"])

            for idx, (key, question) in enumerate(questions, start=1):
                if idx <= 16:
                    response_type = "0 (‘impossible’) to 100 (‘definite’). No explanation,
                elif idx <= 20:
                    response_type = "0 (‘it lowers it a lot’) to 7 (‘it raises it a lot).
                else:
                    response_type = "text"

                answer = get_llm_response(question, response_type)
                writer.writerow([key, question, answer])
                file.flush()
                print(f"{key}: {question} -> {answer}")
```

- LLM was given 4 types of questions:

Type	What It Measures	Example
Marginal	Baseline outcome likelihood (no cause mentioned)	A man buys a half-gallon of milk on Monday. How likely is it the milk is spoiled a week from Wednesday?
Conditional	Outcome likelihood given a weak cause	A man buys a half-gallon of milk on Monday. The power goes out for 30 min on Tuesday. How likely is it the milk is spoiled a week from Wednesday?
Probability Raising	Whether the cause increased or decreased the chance	A man buys a half-gallon of milk on Monday. The power goes out for 30 min on Tuesday. Does that raise or lower the likelihood that the milk is spoiled a week from Wednesday?
Causal Power	How strongly the cause is believed to produce the outcome	A man buys a half-gallon of milk on Monday. The power goes out for 30 min on Tuesday. How likely is it that the power going out for 30 min on Tuesday causes the milk to be spoiled a week from Wednesday?

Goals and Methodology

- Using 4 judgment types allows for a **multi-angle analysis** of the LLM's reasoning:
- Marginal vs. Conditional → Shows if the **weak evidence effect** appears
- Probability-raising → Reveals **intuitive beliefs** about influence
- Causal power → Measures **perceived strength** of the cause

Results

- The experiment produced clear and statistically significant evidence that Meta's LLaMA 3 model replicates the **weak evidence effect**, a cognitive bias first observed in human reasoning.
- The LLM was presented with 12 scenarios, each containing four question types. The average scores across all scenarios were as follows:

Judgment Type	Mean Score
Marginal Probability	54.06
Conditional Probability	40.26
Probability Raising	4.22 (on 1–7)
Causal Power	24.11

The key finding is that the model rated **conditional probabilities lower than marginal ones**, even though the conditional version introduced a weak cause that (normatively) should increase or maintain the probability of the outcome. This **replicates the weak evidence effect** — an irrational pattern that contradicts Bayesian reasoning but is well-documented in humans.

Implications

- The experiment demonstrates that LLaMA 3 not only mirrors human language use but also **internalizes our reasoning shortcuts and cognitive distortions**. The weak evidence effect is not just a linguistic artifact — it's a **cognitive phenomenon**, and its presence in an LLM suggests that these models are capturing **more than syntax and semantics**.

- Interestingly, the model's low causal power ratings show it **recognizes the cause is weak**, even while simultaneously reducing the outcome probability. This mirrors the cognitive conflict seen in human subjects: **people often say a weak cause "increases the chance" of something happening, but still rate it as less likely.**

- The findings support the idea that LLMs can act as **cognitive digital twins** — artificial systems that model human judgment, including its flaws. Because LLMs can be tested at scale, with controlled inputs and flexible scenarios, they may serve as **experimental tools** for studying human reasoning under different conditions.

Adding debiasing methods in human experiment

1. No debiasing – replicating Fernbach

2. Debiasing 1 – describing what is Weak evidence effect to human respondents

> The "weak evidence effect" describes a situation where presenting weak, positive evidence can actually decrease belief in a conclusion, rather than increase it. This happens because people tend to focus on the weakness of the evidence and consider alternative explanations, potentially leading them to doubt the conclusion more than if they had been given no evidence at all.
>
> („Efekt slabého důkazu“ popisuje situaci, kdy předložení slabého, pozitivního důkazu může ve skutečnosti snížit víru v závěr, místo aby ji posílilo. Děje se tak proto, že lidé mají tendenci soustředit se na slabost důkazu a zvažovat alternativní vysvětlení, což je může vést k tomu, že o závěru budou pochybovat více, než kdyby jim nebyl poskytnut žádný důkaz.)

3. Debiasing 2 – navigating human respondents to go step by step in their decision making process

> Consider the following steps:
> 1. How strong is the evidence?
> 2. How relevant is it to the conclusion?
> 3. Based on 1 and 2, estimate how much it should change your belief.
>
> (Zvažte tyto kroky:
> 1. Jak silný je daný důkaz?
> 2. Jak moc souvisí se závěrem?
> 3. Na základě bodů 1 a 2 odhadněte, nakolik by to mělo změnit vaši víru v závěr.)

Conclusion and next steps

- The results of this experiment provide an initial proof-of-concept that LLMs exhibit **human-like inferential bias**, specifically the weak evidence effect. This supports the broader goals of the research:
  to explore how LLMs replicate cognitive biases, to evaluate their suitability as **cognitive models**, and to lay the groundwork for future studies of bias mitigation and reasoning simulation in artificial systems.

- Survey done on human respondents, replicating Fernbachs experiment
- Data needs to be analyzed and compared to LLMs results,
- How debaising methods changed the outcome?

Thank you for your attention :)