PARTICIPANTS Alan ANGELUCI, Fabiola FENILI, Cyrille Anicet Bio KOUMA, Dounia PIHAN, Nathalie KOUBAYOVA, Renée RIDGWAY

FACILITATORS Irène GIRARD, Carlos ROSAS HINOSTROZA

What Kind Of Sociologist Is Your LLM?
Task C: Exploring Theory And Method Approaches Through Prompt Engineering

```
Evaluate which would be the best theoretical approach the
list to apply to the research project.

Here's the theoretical approaches list:
- Marxism (short description)
- Structuralism (short description)
- Functionalism (short description)
- Critical Theory (short description)
- Constructivism (short description)
- Interactionism (short description)

Choose one option from the list I gave that is most adecuate
to execute the research project you just read.
```

This research evaluated several LLMs—including open-source models such as Mistral 7B-v.0.3, Qwen 2.5-7B, and Llama 3.1-8B, alongside proprietary models Claude and ChatGPT—employing topic modeling and prompt engineering to analyze the research methods and sociological theoretical approaches present in 900 French Sociology PhD thesis abstracts.

Research Questions

- Which sociological methods do the models mostly identify in these models? In which ways do they exhibit a preference for qualitative or quantitative methods?
- What theoretical approaches are suggested by the models?
- Which methods and theoretical approaches are more frequently used?
- Where does those biases come from?

PROTOCOL

STEP 1.

Extracting:
Data was extracted from French Sociology PhD theses repo (theses.fr), for the period from 1986 to 2023. Cleaning data reduced the dataset from 5.6K to 900 English entries.

STEP 2.

Exploring:
Topic modeling was carried out using Claude in order to verify the diversity of sociological subjects. Abstracts were analyzed to obtain two approaches: the main research methods created using ChatGPT and theoretical frameworks based on Wikipedia's "sociological theories".

STEP 3.

Prompting/modelling:
Prompts were created, tested and refined using five LLMs: Llama, Qwen, Mistral, ChatGPT and Claude for the two experiments.
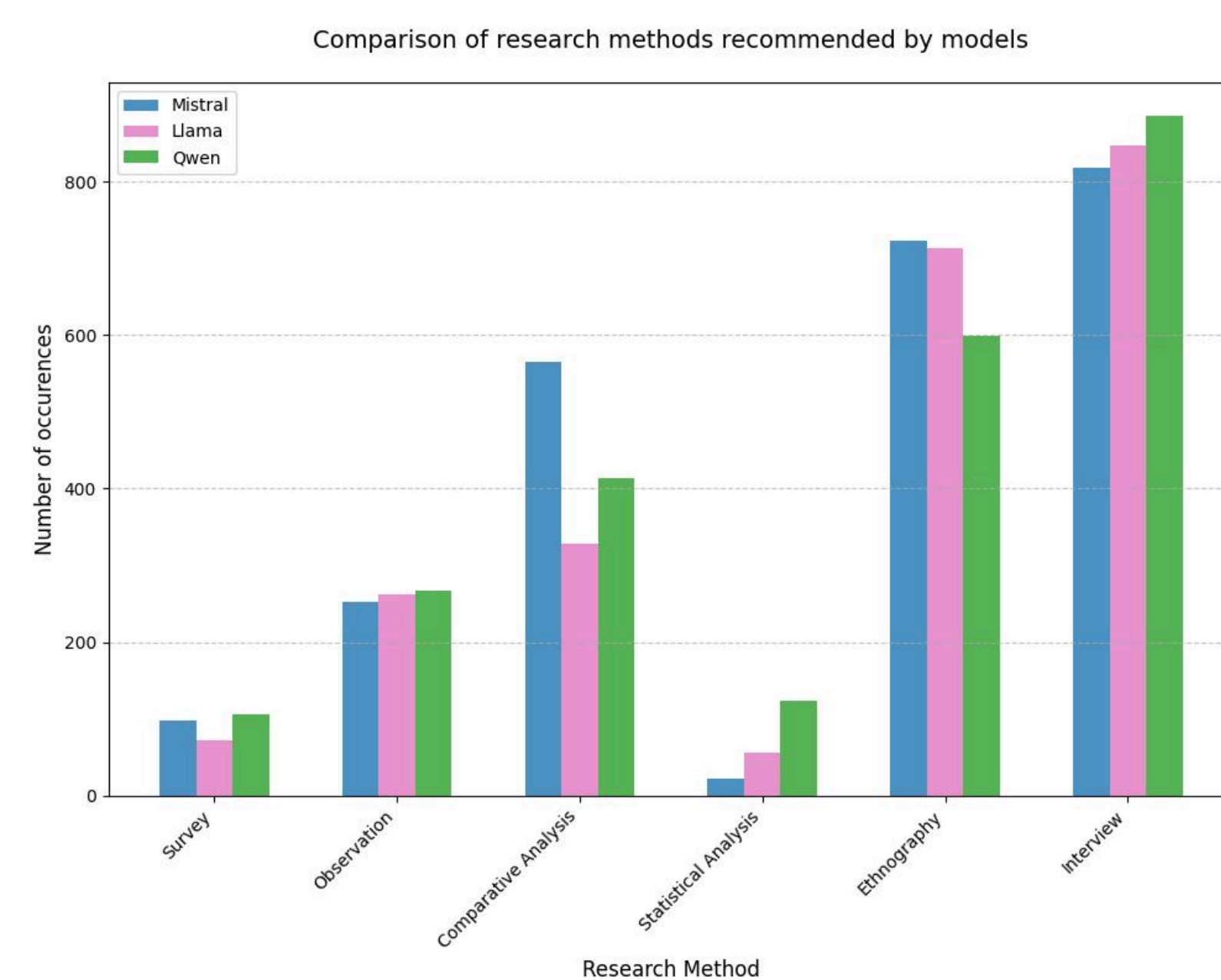
STEP 4.

Visualizing/analyzing:
Statistical analyses evidenced through data visualization showed comparisons among research methods (1) and theoretical approaches (2 and 3). To explore the bias of the results, adversarial testing was conducted (4).
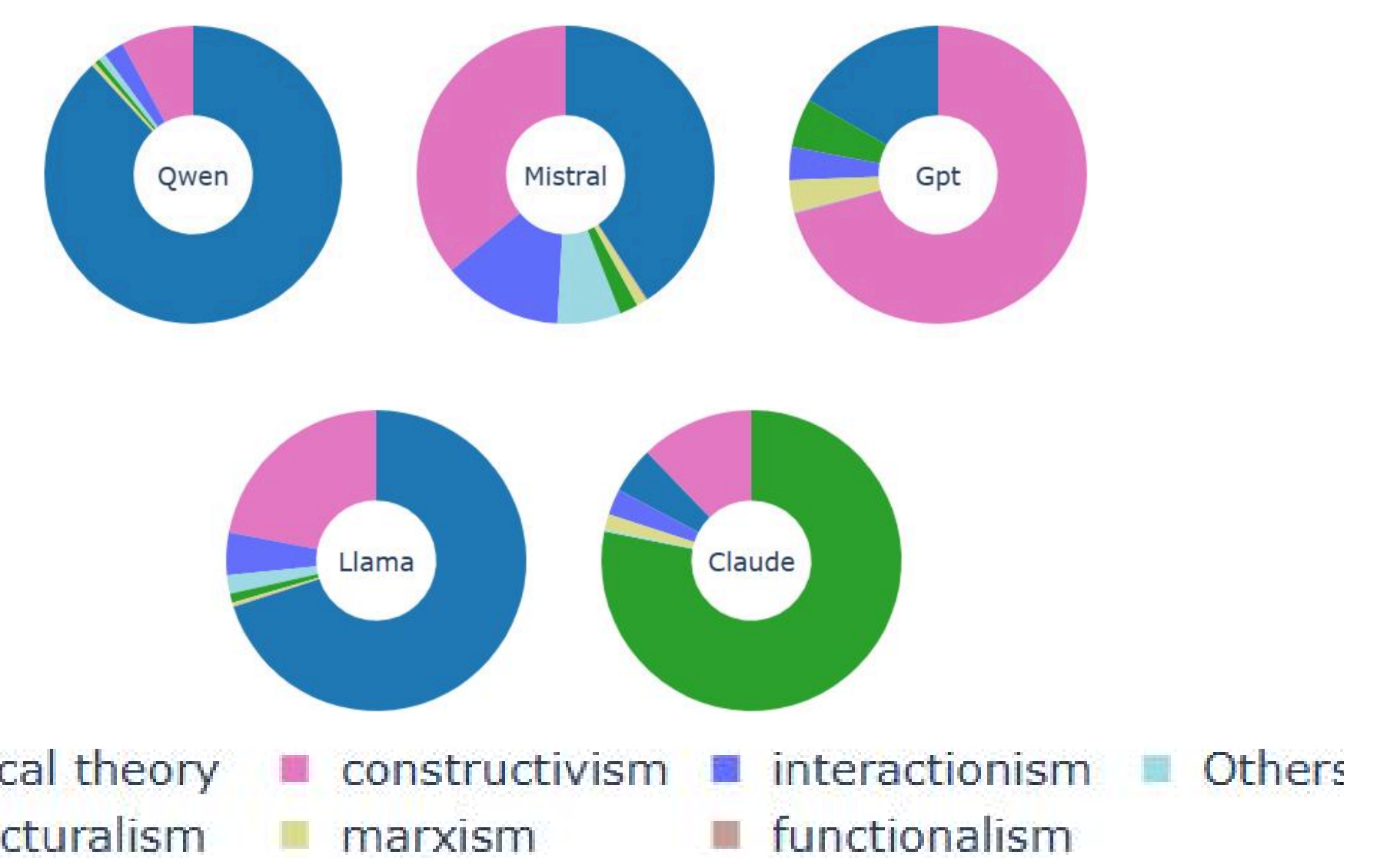
CODE DOCUMENTATION

1. The prevalence of qualitative methods

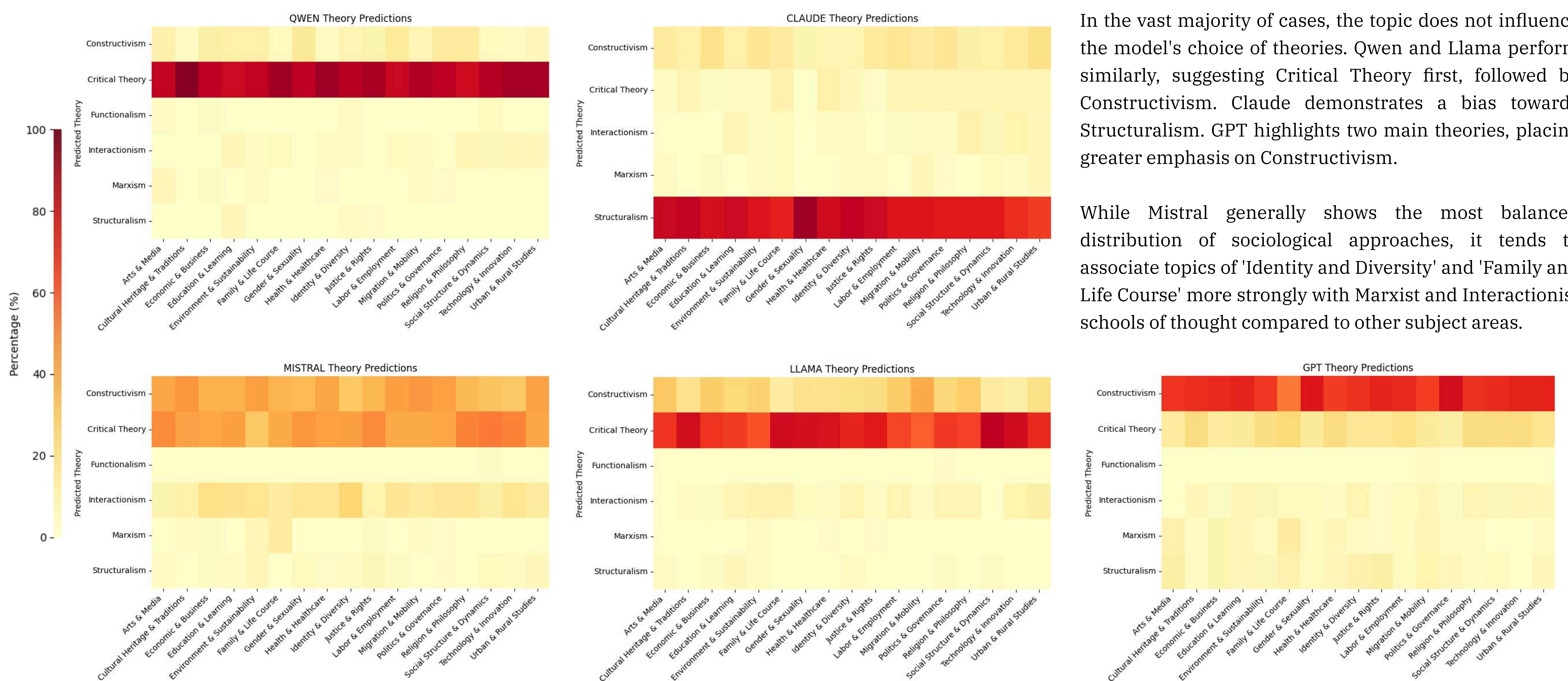Comparison of research methods recommended by models

The three models identify the two qualitative methods: interviews and ethnography, as the most commonly used. Mistral, however, significantly emphasizes comparative analyses, while the results for other methods remain consistent across the models. Quantitative methods are almost never mentioned by any models, although it is known that the term "mixed method" did not figure in the data set, hence it is not included in the list.

2. Mistral: the most diverse LLM for theoretical approach

Legend: critical theory, constructivism, interactionism, Others, structuralism, marxism, functionalism

When asked about the dominant theoretical approaches of each abstracts, the LLMs show variation across frequencies. Open-source LLMs such as Qwen consistently prioritize Critical Theory as the most frequently suggested framework, with a strong preference for this theory, while Llama, and Mistral exhibit a more balanced distribution. On the contrary, a proprietary model like GPT is strongly inclined to Constructivism, while Claude favors Structuralism.

3. Sociological Theories: biases, balances, and sensitivities

QWEN Theory Predictions, CLAUDE Theory Predictions, MISTRAL Theory Predictions, LLAMA Theory Predictions, GPT Theory Predictions
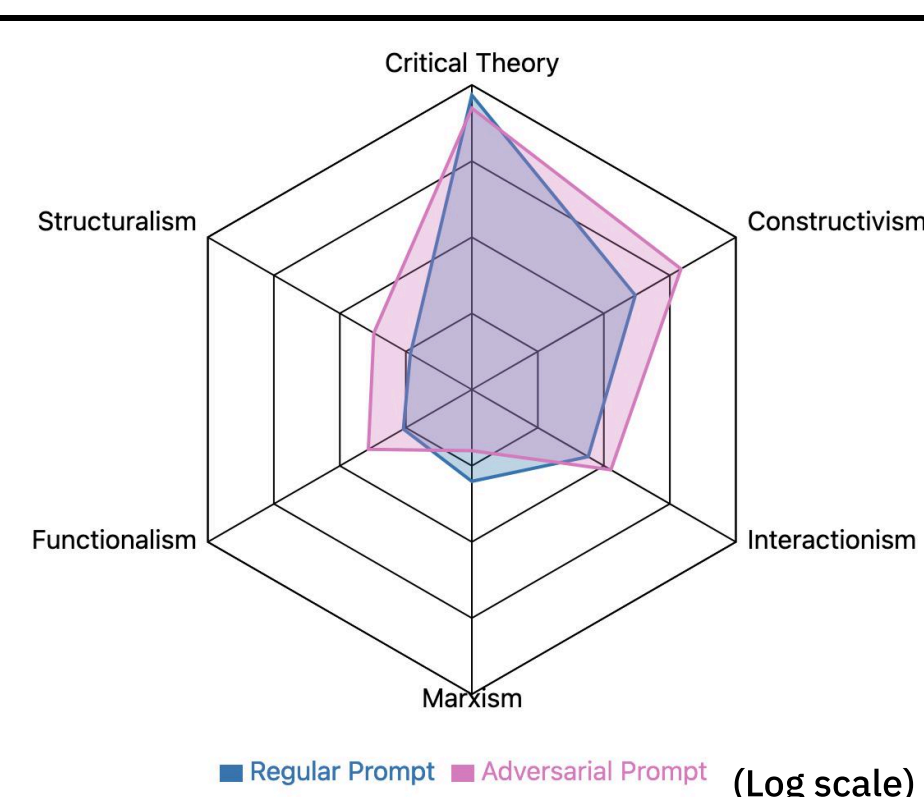
In the vast majority of cases, the topic does not influence the model's choice of theories. Qwen and Llama perform similarly, suggesting Critical Theory first, followed by Constructivism. Claude demonstrates a bias towards Structuralism. GPT highlights two main theories, placing greater emphasis on Constructivism.

While Mistral generally shows the most balanced distribution of sociological approaches, it tends to associate topics of 'Identity and Diversity' and 'Family and Life Course' more strongly with Marxist and Interactionist schools of thought compared to other subject areas.

4. Exploring the bias in the model

Adversarial testing was applied to Qwen, the model with the most pronounced inclination (Critical Theory). The descriptions of the theoretical options in the prompt were intentionally altered to their complete opposites. Additionally, the labels were mixed while keeping the original descriptions. The results still demonstrated strong tendency towards Critical Theory, suggesting that this orientation originates directly from the training data. This appears to compel the model to select the Critical Theory option as soon as those words are recognized, even when the remainder of the prompt provides an unrelated description of the approach.

Regular Prompt, Adversarial Prompt (Log scale)

5. Preliminary Results: bigger is not better

- Research across various AI models has demonstrated that despite differences in their scale, architecture, training approaches, open-source/proprietary, age and country of origin, they consistently exhibit polarizing decision-making patterns. Common sense would lead us into thinking that bigger models are more balanced but GPT and Claude show otherwise.
- Based on the above experiments, it is hard to analyze cultural biases. Claude, GPT and Llama are American however they behave very differently. Llama and Qwen can however be compared, even though Qwen is Chinese. Above all the LLMs interrogated, Mistral has shown more balance and diversity in the theoretical approach.
- The last experiment gives us a huge insight : the source of the bias seems to be located in the training corpus.
- Questions remain: more testing is needed with other models (i.e. GPT, Claude), both approaches. and adversarial testing.