DMI WINTER SCHOOL
2025

What Kind Of Sociologist Is Your LLM?

Whose Voice Matters:

Biases in LLMs Recommending Sociological Authors for a research topic

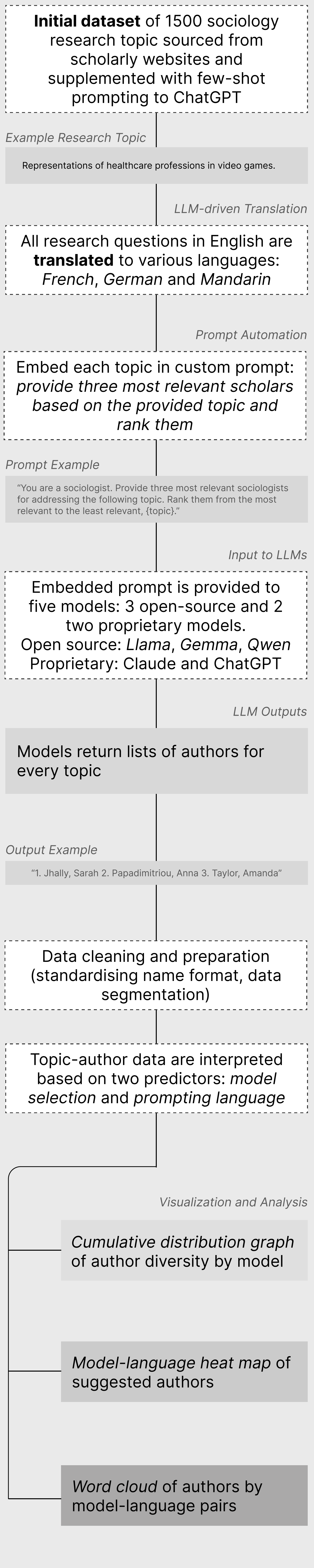
Goal: Evaluate the biases and capabilities of LLMs to recommend sociologist authors

This project is situated within the broader debates on the **capabilities and limitations of generative artificial intelligence** (LLMs) in disciplines where interpretation, critique, and epistemological diversity are central, such as **sociology and the humanities**. While LLMs have already been extensively evaluated in fields like medicine, law, and finance, these analyses overlook disciplines where quantification is less directly applicable, and where **cultural and epistemological biases** have a more significant impact. We lack sufficient knowledge regarding the reliability of conversational AIs in sociology and the humanities.

Methodology: Comparison of the output of different LLMs in different languages to detect possible biases

In addition to exploring the ability of AI chatbots to support sociologists in their academic work, we wanted to **evaluate the biases** and outcomes of different LLMs in English, German, and Mandarin in terms of author recommendations. In this way, we brought a global and cultural lens to the project. After generalizing the output of all LLMs under each respective language, we were able to compare the suggested sociologist authors per country and see if there was a difference in the AI's recommendation.

Methodology



Initial dataset of 1500 sociology research topic sourced from scholarly websites and supplemented with few-shot prompting to ChatGPT

Example Research Topic

Representations of healthcare professions in video games.

LLM-driven Translation

All research questions in English are **translated** to various languages: *French*, *German* and *Mandarin*

Prompt Automation

Embed each topic in custom prompt: *provide three most relevant scholars based on the provided topic and rank them*

Prompt Example

"You are a sociologist. Provide three most relevant sociologists for addressing the following topic. Rank them from the most relevant to the least relevant, {topic}."

Input to LLMs

Embedded prompt is provided to five models: 3 open-source and 2 two proprietary models.
Open source: *Llama*, *Gemma*, *Qwen*
Proprietary: Claude and ChatGPT

LLM Outputs

Models return lists of authors for every topic

Output Example

"1. Jhally, Sarah 2. Papadimitriou, Anna 3. Taylor, Amanda"

Data cleaning and preparation (standardising name format, data segmentation)

Topic-author data are interpreted based on two predictors: *model selection and prompting language*

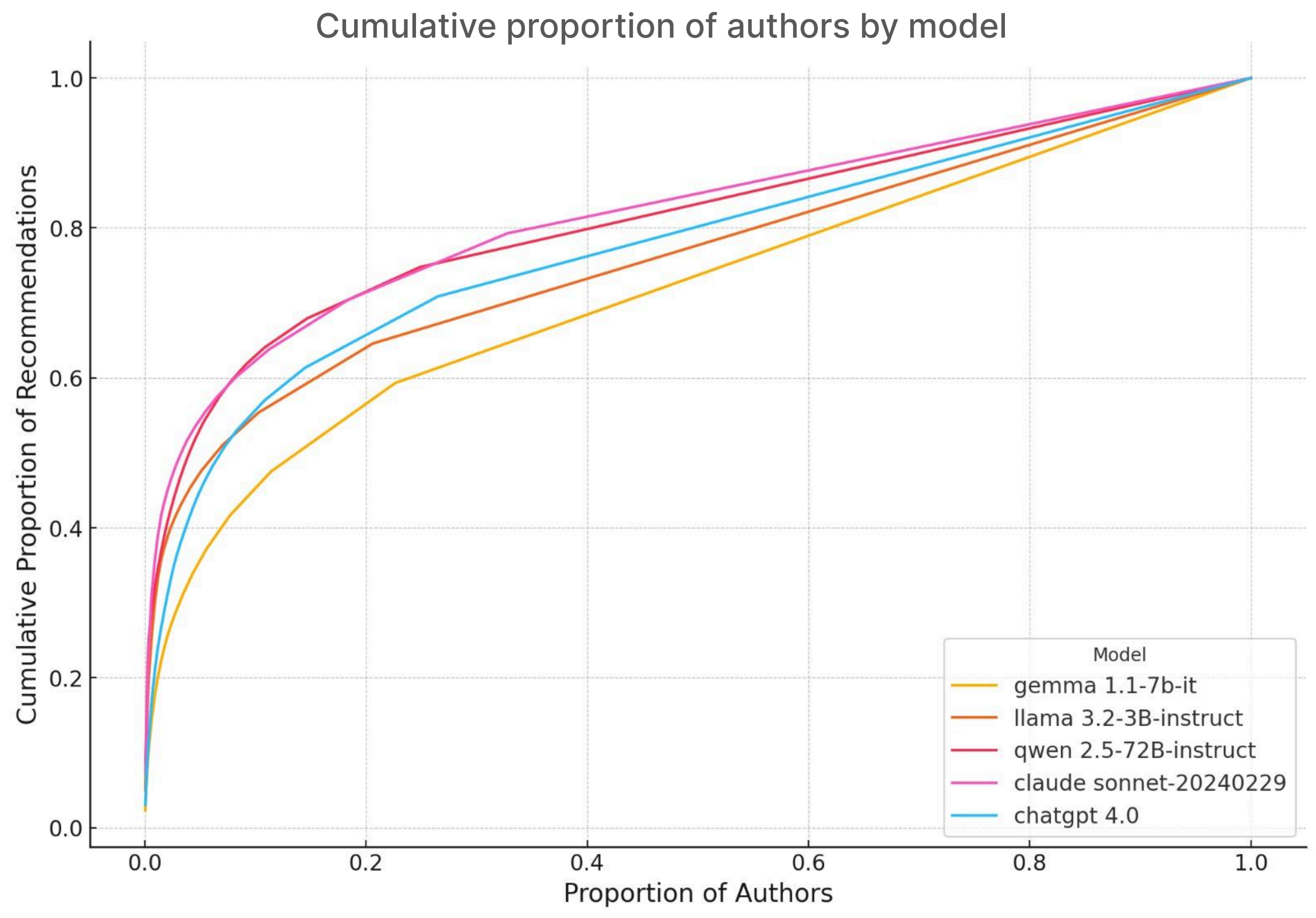
Visualization and Analysis

Cumulative distribution graph of author diversity by model

Model-language heat map of suggested authors

Word cloud of authors by model-language pairs

The size of a model does not automatically guarantee better diversity in recommendations of sociologists



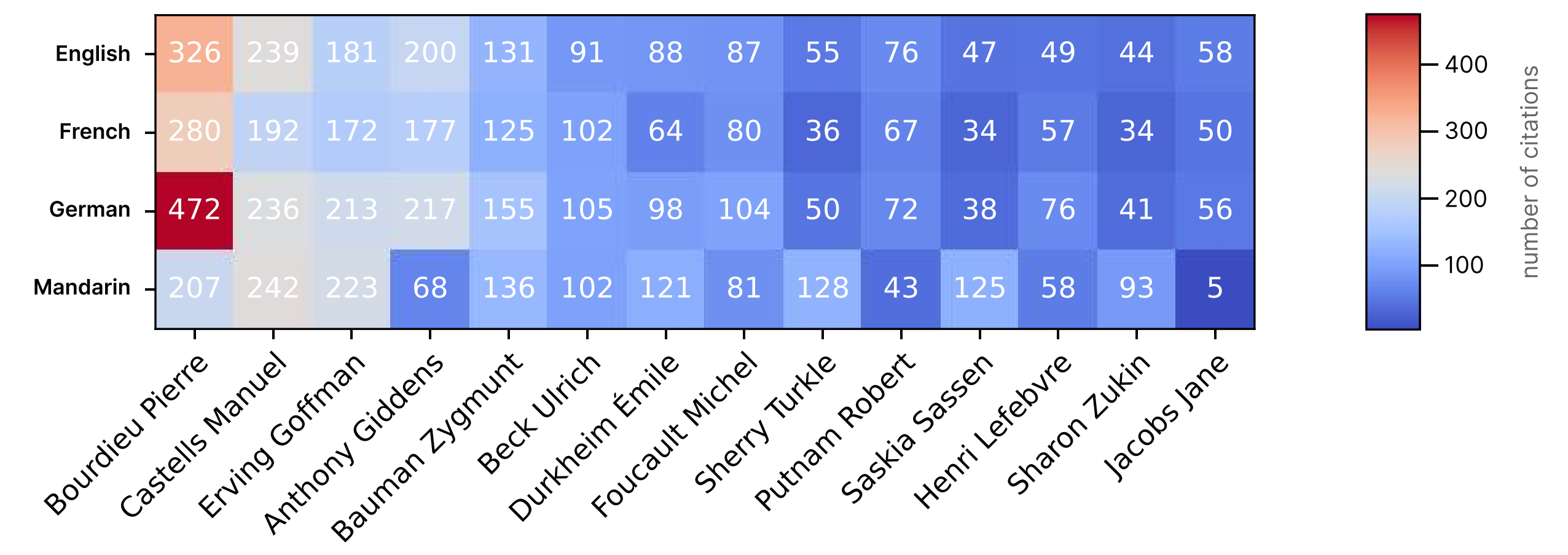
The cumulative distribution curves highlight a fascinating paradox:

Model size does not directly determine diversity in recommendations. Smaller models (less parameters), like *Gemma 1.1-7b-it* (7B), can rival or outperform larger ones, such as *Qwen 2.5-72B-instruct* (72B), in diversity due to factors like training data quality and optimization strategies. *Gemma's* gradual curve reflects broader exploration, while *Qwen* and *Claude* focus on dominant figures, influenced by their data or design priorities.

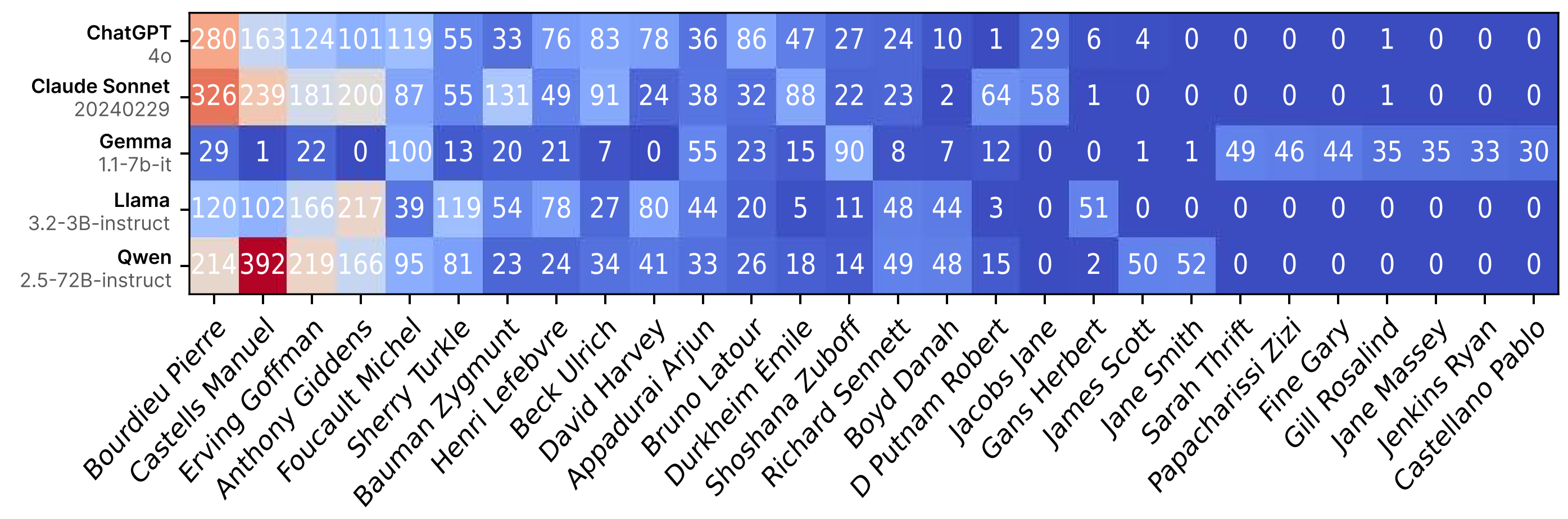
Even *ChatGPT 4o* (~175B) balances concentration and diversity effectively, showing that size supports versatility but is not the sole determinant. This highlights that well-optimized smaller models can excel when aligned with specific goals.

Same topic, different authors: The LLM model utilized influences the diversity of recommended sociological authors, but offers limited observable differences in outputs across languages

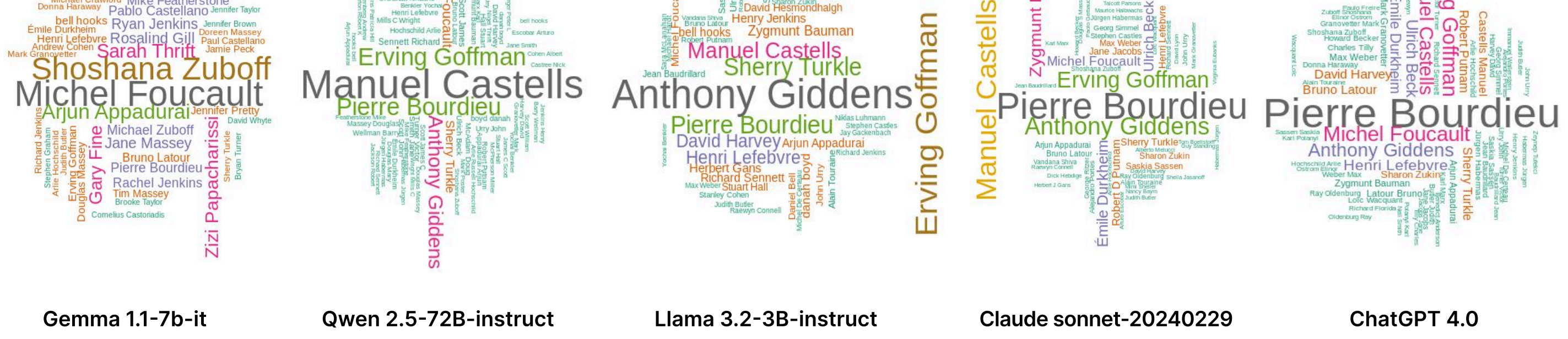
Heat Map: Number of Author Citations per Language
Claude Sonnet-20240229 (Anthropic)



Heat Map: Number of Author Citations per LLM
English Prompt: ChatGPT 4o, Claude Sonnet, Gemma, Llama, Qwen



Word Cloud Analysis: LLM Model Comparison
English Prompt: ChatGPT 4o, Claude Sonnet, Gemma, Llama, Qwen



Gemma 1.1-7b-it   Qwen 2.5-72B-instruct   Llama 3.2-3B-instruct   Claude sonnet-20240229   ChatGPT 4.0

This graph displays the frequency in citation across the top 14 sociologist authors within *Claude Sonnet-20240229*, inclusive across each four languages (**English, French, German, and Mandarin**). Our findings conclude that **Pierre Bourdieu** is consistently highly cited and most recommended across each language, with a noticeable peak in German where the difference with other authors is particularly significant. This suggests the prompt language does not have a significant influence in the output of each LLM. Although the LLM mainly recommended western well-known authors, a diversity of theoretical approaches and a diversity of classical and contemporary approaches can be found across the 14 scholars. The emphasis on Western sociologists may reflect the limitations of the training data and may underestimate non-Western perspectives and theories.

This heat map reveals fascinating insights into how different LLMs rank authors in English. **Pierre Bourdieu** emerges as the most frequently recommended author across models, closely followed by **Manuel Castells**, aligning with patterns observed in the word cloud analysis. However, the standout finding is *Gemma's* distinct behavior: it consistently suggests a broader and more diverse range of authors compared to the other models, as reflected by the lower and more varied rankings in the chart. This highlights *Gemma's* unique approach to author recommendations, offering greater diversity and breaking away from the concentrated focus of its counterparts. However, a closer look at the profiles of the authors recommended by Gemma reveals that some of them are not social scientists. For example, we could not find any information on Sarah Thrift as a sociologist. So, while *Gemma* suggested a broader list of authors, its accuracy cannot be verified.

The word clouds highlight the *overrepresentation* of **Pierre Bourdieu** across most models, particularly in *Claude* and *ChatGPT*, where classical figures clearly dominate. *Gemma*, however, stands out by diversifying its suggestions, including less established and more contemporary authors who explore interdisciplinary and emerging fields. These findings underscore the importance of balancing focus and diversity in author recommendations, as also reflected in the heat map analysis. While this diversity suggests an attempt to escape the concentration on dominant figures, it raises critical questions: does *Gemma* truly mitigate the bias of overrepresentation, or does it reflect a different bias, oriented towards less academic or contingent contributions? This observed diversity could stem from an uncontrolled bias linked to its training corpus or optimization objectives. In seeking to avoid focusing on dominant figures, Gemma risks shifting toward dispersion, potentially reducing its academic relevance.

FACILITATORS

Bilel BENBOUZID (Université Gustave Eiffel), Alexis PERRIER (Université Gustave Eiffel), Carlos ROSAS HINOSTROZA, Irène GIRARD, Noé DURANDARD

PARTICIPANTS

Eldi BICARI, Ahmed DIAKITE, Eusebie HUYSMAN, Kanako INOUE, Kevin JIN, Jana RESKE, Oswin SPRINGER, Justine XU, Yue YU, Katherine ZDANOWSKI

Université Gustave Eiffel
ESIEE PARIS
Global Digital Cultures. University of Amsterdam