Turkish Political Compass Test of LLMS

Ahmet Ertuğrul Hacıoğlu, Ceyhun Sonyürek - Advisors: Abdullatif Köksal, Suzan Üsküdarlı

Computer Engineering, Bogazici University

Problem Statement

To investigate the degree to which large language models (LLMs) align with cultural values and demographic diversity with particular emphasis on the context of Turkey.

Motivation

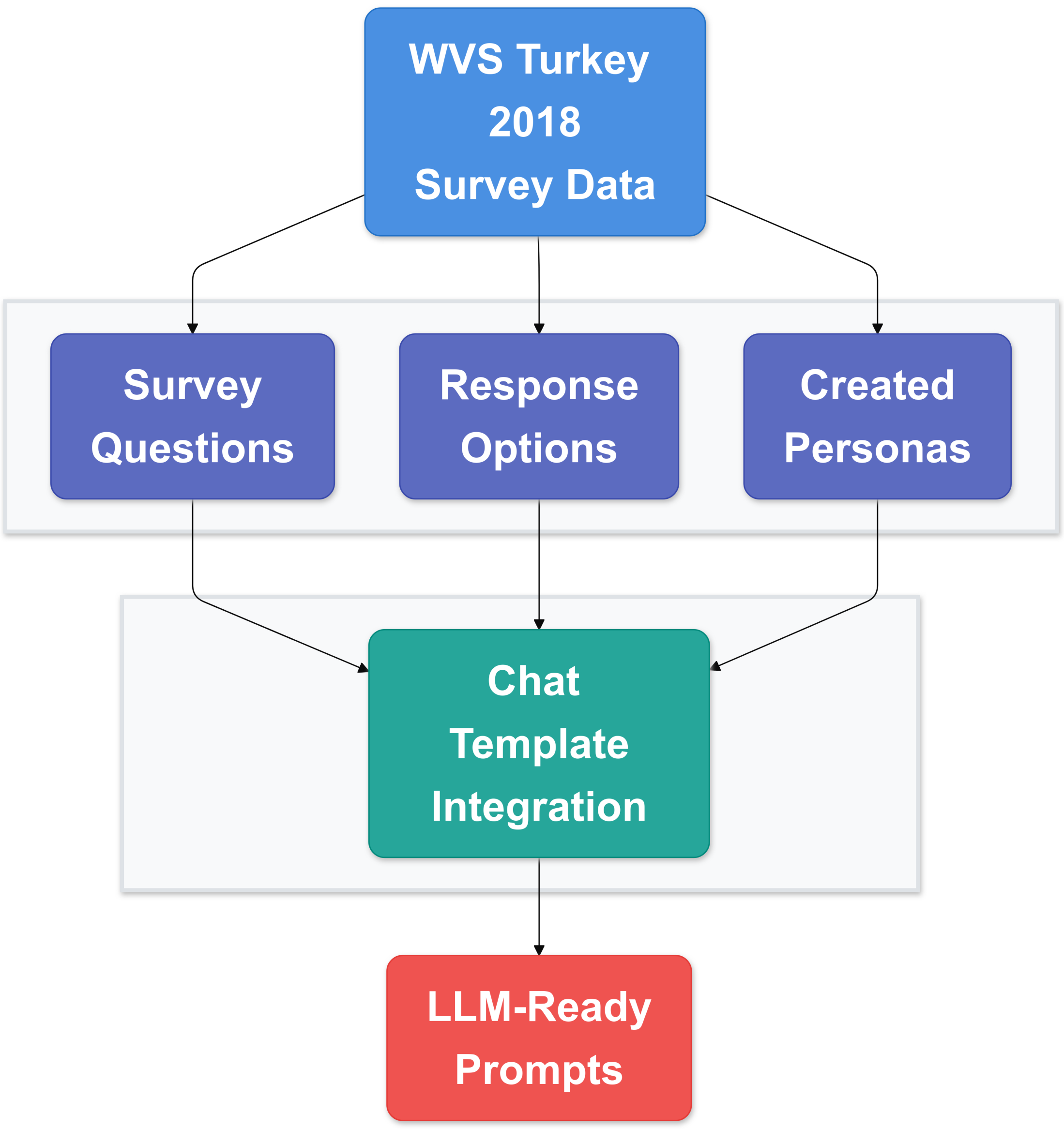
We test LLM models' understanding of diverse Turkish political perspectives. This reveals how well these global models grasp cultural nuances beyond their data origins.

State of Art

Prior research shows LLMs exhibit Western-centric biases[1], especially from the USA and Europe, and their political stance outputs are sensitive to prompt variations[2].

Dataset

- World Values Survey (WVS) Wave 7 Turkey dataset (collected in 2018)[3].
- Created value-centric prompts to feed to Gemma-2-2b, Gemma-2-9b and Gemma-2-27b language models.



Creation of personas from WVS respondents' demographic attributes

Metric Used

We use **Kullback-Leibler (KL) Divergence** to measure how closely model predictions match real survey responses. Lower **KL Divergence** means better alignment between model and reality, helping us identify where the model understands different demographic groups.

Methodology

- **Data Preprocessing Before Experiments:** Extract demographic attributes and survey responses from WVS dataset. Then, create Turkish personas based on real demographic distributions while maintaining statistical validity.

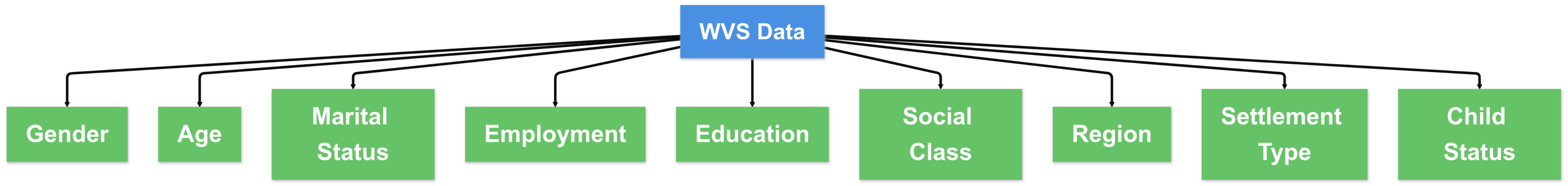
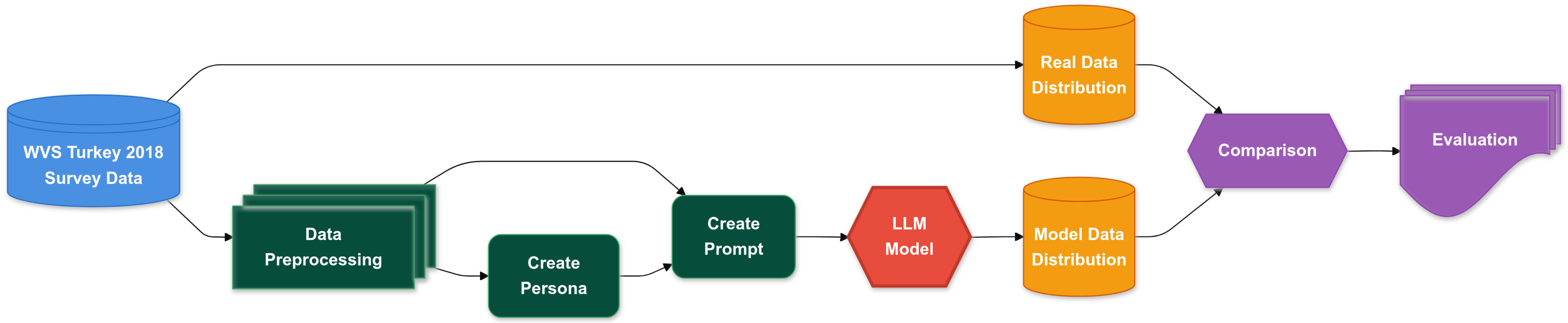


Diagram of persona creation

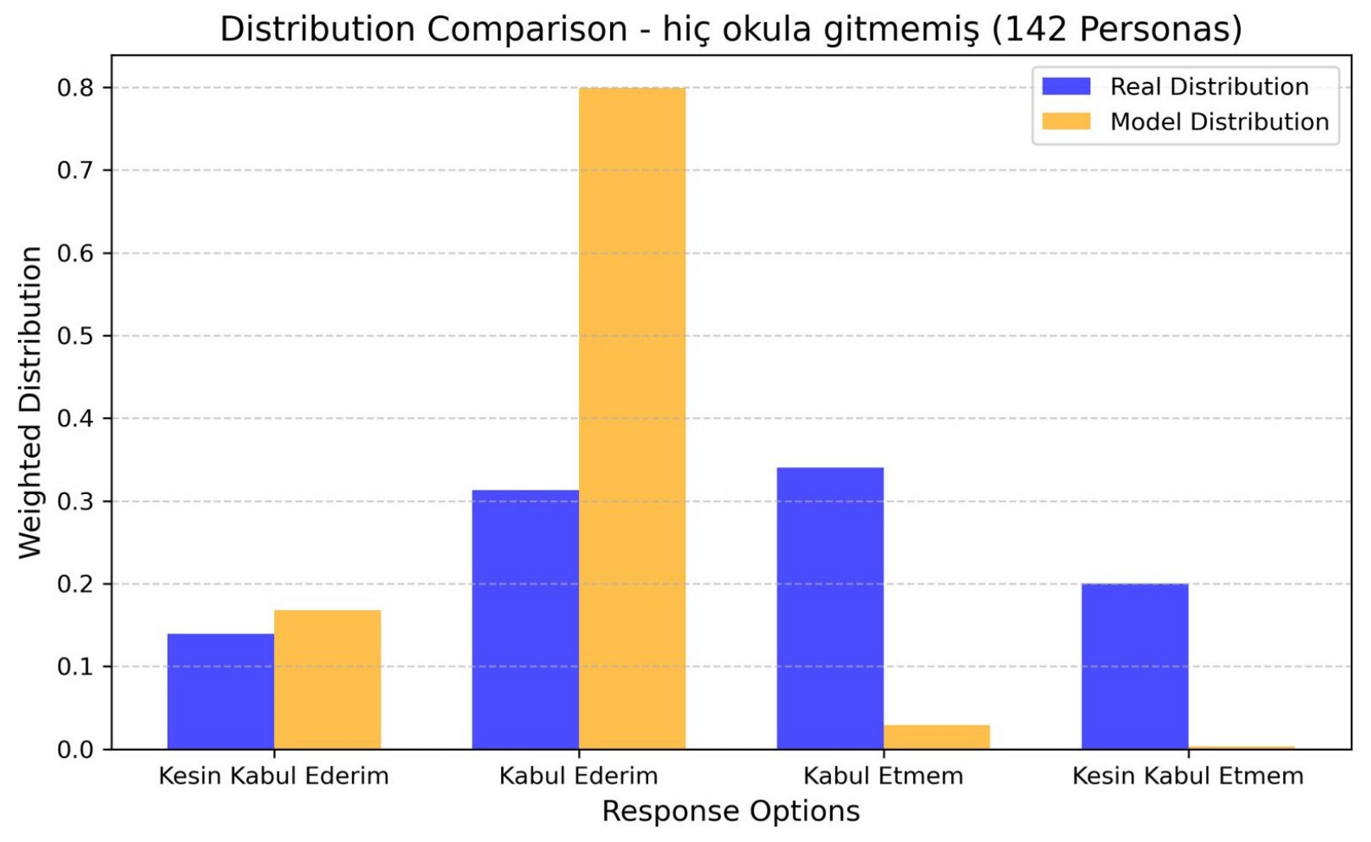
- **LLM Prompting Experiments:** Combine generated personas with survey questions using specialized chat templates for LLM interaction. Then, compare LLM responses against original survey distributions to assess cultural understanding.

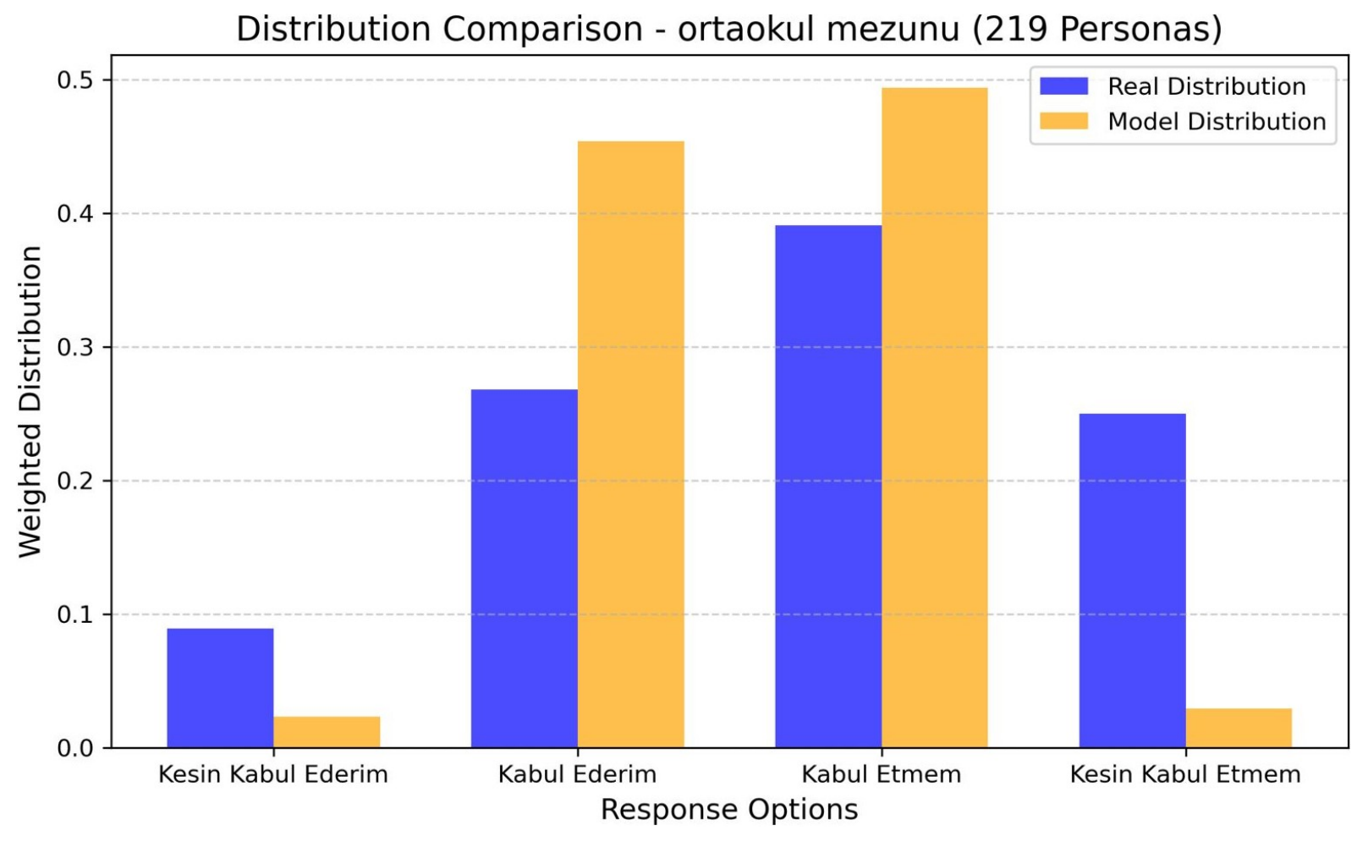


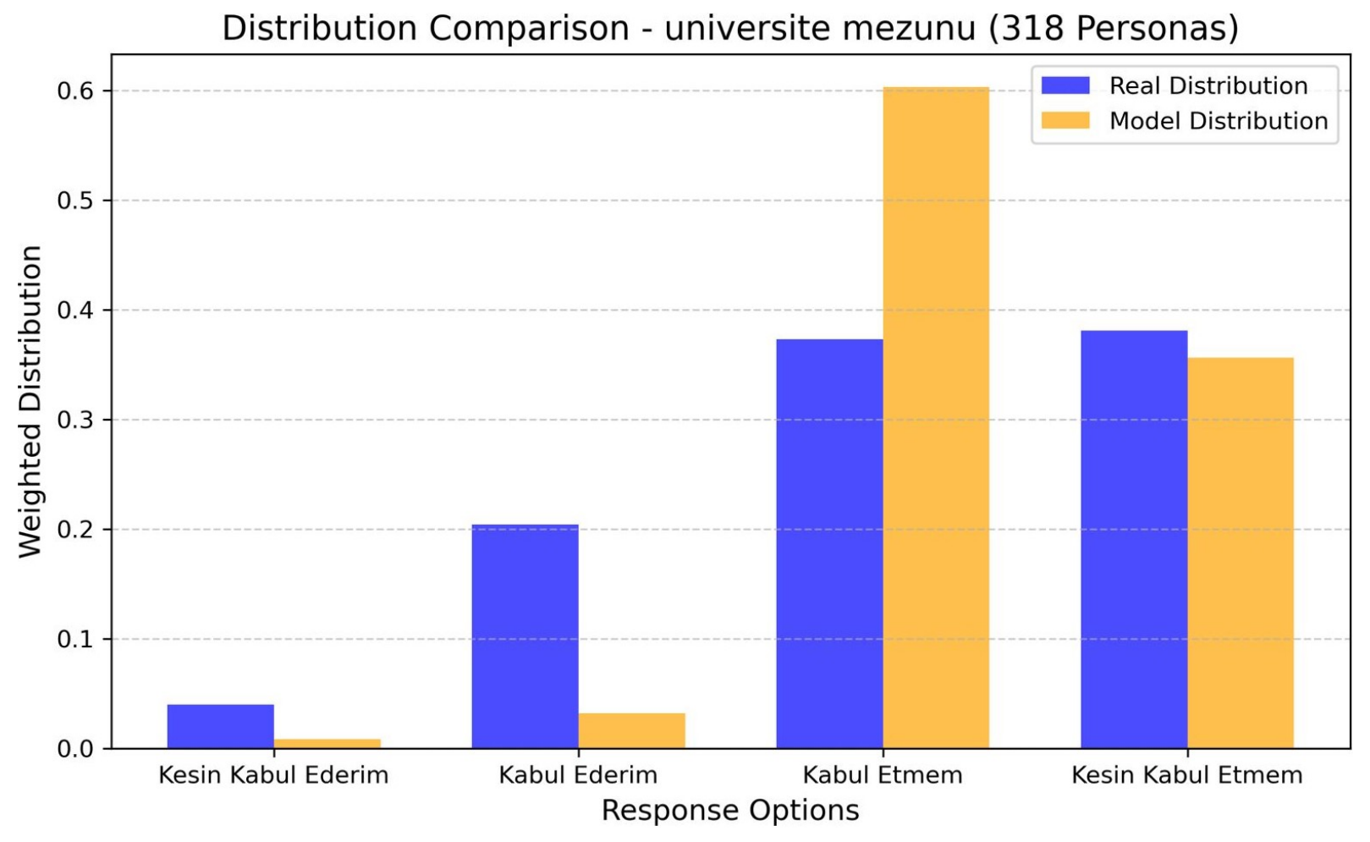
Example Analysis

Question: "Şimdi size bazı görüşler okuyacağım. Bunların her biri hakkındaki fikrinizi 'kesinlikle kabul ederim', 'kabul ederim', 'kabul etmem' veya 'kesinlikle kabul etmem' şeklinde belirtiniz. Üniversite eğitimi, kız çocuktan çok erkek çocuk için önemlidir."

Aspect	Persona Sample 1	Persona Sample 2
Description	Aysel, 19 yaşında bekâr bir kadın, çocuğu yok, şu anda öğrenci, İstanbul şehrinde kent merkezinde yaşayan, kendi sosyal sınıfını orta sınıfın üst kısmında olarak tanımlayan, oğrenci birisidir.	Sevim, 35 yaşında evli bir kadın, çocuk sahibi, hiç okula gitmemiş, İstanbul şehrinde kent merkezinde yaşayan, kendi sosyal sınıfını alt sınıf olarak tanımlayan, ev kadını birisidir.
Real Distribution	Kesinlikle Kabul Ederim: 0% Kabul Ederim: 33% Kabul Etmem: 0% Kesinlikle Kabul Etmem: 67%	Kesinlikle Kabul Ederim: 0% Kabul Ederim: 0% Kabul Etmem: 0% Kesinlikle Kabul Etmem: 100%
Model Distribution(gemma-2-27b)	Kesinlikle Kabul Ederim: 0% Kabul Ederim: 0% Kabul Etmem: 0.5% Kesinlikle Kabul Etmem: 99.5%	Kesinlikle Kabul Ederim: 4.2% Kabul Ederim: 95.2% Kabul Etmem: 0.5% Kesinlikle Kabul Etmem: 0%
KL Divergence	7.04	23.03
Alignment	✓High Alignment: Model agrees with real distribution showing correct understanding of educated persona's view	× Significant Misalignment: Model fails to capture the disagreement shown in real distribution, suggesting stereotype bias







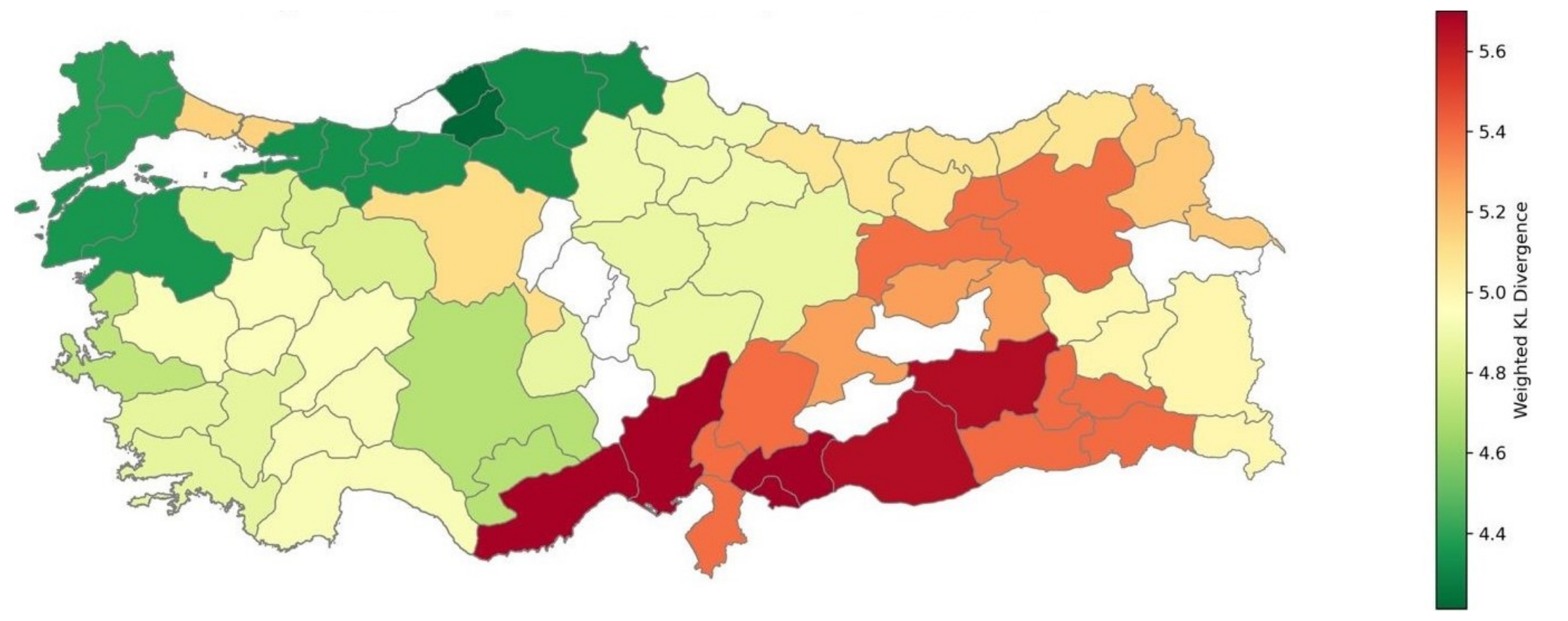
Results

- **Attribute Sensitivity:** Specific cultural or demographic attributes in personas significantly influence model predictions for some of given questions.
- **Model Scaling Effects:** Larger models are more accurate at predicting the most probable answers of real respondents. However, they exhibit lower entropy, leading to more decisive but potentially incorrect predictions by a wide margin.

Metric	Gemma 2B	Gemma 9B	Gemma 27B
Exact Match (%)	32.58%	37.89%	41.56%
Weighted Avg. Entropy	0.56	0.39	0.31

Table 1: Exact Match Percentages and Entropy Across Models

- **Stereotype Bias:** Certain persona attributes, especially persona's living region, consistently result in higher KL Divergence between their related real respondents, indicating systematic stereotypical biases in model predictions for specific demographic groups.



KL Divergence across Turkey's regions

- **Model Robustness:** Model responses remained consistent when tested with different prompt formats (letters, numbers, and their reversed orders), showing only minor variations in KL divergence compared with real distribution, which confirms the stability of models.

Format	Non-Reversed	Reversed	Difference
Numeric (KL div.)	4.73	4.81	+0.08
Letter (KL div.)	4.71	4,69	-0.02

Table 2: 27B Model Results Compared with Results of Real Respondents

Future Work

Our next steps will aim to try newer LLM models and involve an ontological representation of a system meant for collecting data. We will also examine specific issues how language models deal with cultural and political issues across various countries.

References

[1] E. Durmus, K. Nyugen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, and L. Lovitt. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. arXiv preprint arXiv:2306.16388.

[2] P. Röttger, V. Hofmann, V. Pyatkin, M. Hinck, H. R. Kirk, H. Schütze, and D. Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. arXiv preprint arXiv:2402.16786.

[3] https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp