

**SYNTHETIC RESPONDENTS: EVALUATING LARGE
LANGUAGE MODELS AS SUBSTITUTES FOR HUMAN
SURVEY PARTICIPANTS**

ALP UÇAR

ÖZYEGİN UNIVERSITY

JANUARY, 2026

SYNTHETIC RESPONDENTS: EVALUATING LARGE LANGUAGE MODELS AS SUBSTITUTES FOR HUMAN SURVEY PARTICIPANTS

**by
Alp Uçar**

Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in
Data Science

Advisor: Assoc. Prof. Enis Kayış

Graduate School of Science and Engineering
Özyeğin University
İstanbul

January, 2026

SYNTHETIC RESPONDENTS: EVALUATING LARGE LANGUAGE MODELS AS SUBSTITUTES FOR HUMAN SURVEY PARTICIPANTS

Approved by:

Assoc. Prof. Enis Kayış, Advisor
Department of Industrial Engineering
Özyeğin University

Assoc. Prof. Reyhan Aydoğan
Department of Artificial Intelligence and Data
Engineering
Özyeğin University

Asst. Prof. Mehmet Yasin Ulukuş
Department of Industrial Engineering
Istanbul Technical University

Approval Date: December 29, 2025

To my wife, my family, and to the giants upon whose shoulders I stand...



DECLARATION OF ORIGINALITY

I hereby declare that I am the sole author of this thesis and that this is the true copy of my thesis, including the final revisions, approved by my thesis committee. All data and information have been obtained, produced, and presented in accordance with the rules of research ethics and principles of academic honesty. As required by these rules, to the best of my knowledge I have acknowledged ideas, thoughts, and any copyrighted material in accordance with the standard referencing rules. I certify that any part of this thesis has not been submitted for a degree or diploma in another educational institution.

Alp Uçar

ABSTRACT

This thesis evaluates Large Language Models (LLMs) as "synthetic participants" in opinion surveys, addressing rising costs and biases in traditional data collection. Using 10,000 profiles from Wave 7 of the World Values Survey, ten LLM architectures were prompted to respond to seven trust-related questions. Performance was assessed via exact-match accuracy, mean absolute error (MAE), distributional divergence, and feature importance alignment, comparing LLM outputs to actual human participants and supervised machine learning models. Findings show that LLMs achieve a predictive accuracy that outperforms chance but lags behind specifically-trained supervised models. While errors were directionally accurate, divergence tests revealed a consistent "majority class fixation," where LLMs overrepresented dominant views and suppressed minority perspectives. Furthermore, while both machine learning models trained on human responses and those trained on LLM outputs rely on macro-institutional indicators, the first group is further shaped by cultural identity (ethnicity, language), whereas the second group relies on socio-economic status (income, social class). Consequently, LLMs function as "global average smoothers"—capturing broad trends but failing to replicate the heterogeneity required for high-fidelity social research. While promising for exploratory studies, LLMs remain an inappropriate substitute for human data in contexts requiring representational accuracy for minority or cross-cultural viewpoints.

Keywords: Large Language Models, Synthetic Data, Computational Social Science, Survey Simulation, Silicon Sampling

ÖZET

Bu tez, geleneksel veri toplama süreçlerindeki artan maliyet ve yanlılık sorunlarına bir çözüm arayışı olarak Büyük Dil Modellerini (LLM) anketlerdeki "sentetik katılımcılar" rolüyle değerlendirmektedir. Çalışma kapsamında Dünya Değerler Araştırması'nın (WVS) 7. dalgasından alınan 10.000 profil kullanılarak, on farklı LLM, güvenle ilgili yedi soruyu yanıtlamaya yönlendirilmiştir. LLM performansı; tam eşleşme doğruluğu, ortalama mutlak hata, dağılımsal sapma ve değişken önemlerinin uyumu metrikleri üzerinden, gerçek insan yanıtları ve gözetimli makine öğrenmesi modelleriyle karşılaştırılarak analiz edilmiştir. Bulgular, LLM tahminlerinin rastlantısal eşiğin üzerinde doğruluk gösterdiğini, ancak gözetimli makine öğrenmesi modellerinin gerisinde kalabildiğini göstermektedir. Hatalar yönsel olarak tutarlı olsa da, sapma testleri LLM'lerin baskın görüşleri aşırı temsil ederek ve azınlık perspektiflerini bastırarak ve bir "çoğunluk sınıfa saplanma" sergilediğini ortaya koymaktadır. Ayrıca, her ne kadar hem insan yanıtları üzerinde eğitilmiş makine öğrenmesi modellerinin hem de LLM çıktıları üzerinde eğitilmiş olanların Hükümet Etkinliği, İnsani Gelişmişlik Endeksi, Gayrisafi Yurt İçi Hasıla gibi makro göstergelere öncelik verdiği görülse de, ilk grubun aynı zamanda etnik köken ve dil gibi kültürel kimliğe ilişkin özelliklere önem verdiği, ikinci grubun ise gelir ve sosyal sınıf gibi sosyo-ekonomik faktörleri baz aldığı gözlemlenmiştir. Sonuç olarak LLM'ler, genel demografik eğilimleri yakalayan ancak detaylı sosyal araştırmalar için gereken heterojenliği yansıtmayan "ortalama temsilciler" olarak işlev görmektedir. Çalışma, keşifsel çalışmalar için vaat edici olsalar da incelenen LLM'lerin; azınlık veya kültürlerarası perspektifler için temsil doğruluğu gerektiren bağlamlarda insan verisinin yerini tutmak için henüz yeterli olmadığını göstermektedir.

Anahtar Kelimeler: Büyük Dil Modelleri, Sentetik Veri, Hesaplamalı Sosyal Bilimler, Anket Simülasyonu, Yapay Örneklem

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor, Assoc. Prof. Enis Kayış, for his invaluable guidance, support, and patience throughout this research journey.

I am also grateful to the World Values Survey Association for providing the open-access microdata that made this research possible. Furthermore, I acknowledge the developers and researchers in the artificial intelligence community; their groundbreaking work in Large Language Models has opened the new frontiers in computational social science that this thesis explores.

On a more personal note, I wish to honor the memory of Mustafa Kemal Atatürk, founder of the Republic of Türkiye. I am deeply indebted to his vision, which built the modern institutions that paved the way for me to pursue science and engineering today. My admiration for his foresight only deepens as I progress in my academic career.

My deepest thanks go to my family—especially my mother and my two sisters. You taught me how to stand up whenever I fall, a lesson I had to relearn daily during the more challenging phases of this work. I also owe a unique debt of gratitude to my canine sister, Şans; her 18 years of companionship taught me that intelligence is not an exclusively human trait—an insight that became indispensable to this study. Likewise, I thank Ceviz, our late neighbourhood squirrel, for the simple joy of pure existence.

Finally, and most importantly, I thank my beloved wife and best friend. You are the unstoppable force to my immovable object. Your support has been so profound that this work belongs to you as much as it does to me.

TABLE OF CONTENTS

DECLARATION OF ORIGINALITY	iv
ABSTRACT	v
ÖZET	vi
ACKNOWLEDGEMENTS	vii
LIST OF TABLES	xi
LIST OF FIGURES	xiii
LIST OF ACRONYMS AND ABBREVIATIONS	xiv
1 INTRODUCTION.....	1
2 LITERATURE REVIEW	6
2.1 Traditional Methods Considered for Survey Simulation	6
2.1.1 Statistical Inference and Opinion Modeling	6
2.1.2 Structural Synthetic Data Generation	8
2.1.3 The Data-Availability Paradox	10
2.2 Survey Simulation Using LLMs	11
2.2.1 Concept of "Silicon Sampling" and Semantic Generalization	11
2.2.2 Methodology for Prompting and Persona Conditioning	12
2.2.3 Results Indicating Consistency with Human Data	13
2.2.4 Systematic Challenges	14
3 PROBLEM DEFINITION AND RESEARCH QUESTIONS	16
3.1 Scientific Goal and Research Questions	16
3.1.1 Research Question 1: Representational Fidelity	17
3.1.2 Research Question 2: Comparative Performance and Feature Im- portance.....	18
3.2 Contributions of This Thesis	20
3.2.1 Integrated Evaluation of Representational Fidelity	20
3.2.2 Benchmarking Against Traditional Machine Learning Models.....	20

3.2.3	Assessment of Demographic Conditioning Capacity	21
3.2.4	Exploration of Feature Importance Alignment	21
4	METHODOLOGY	22
4.1	Experiment Design.....	22
4.1.1	Survey Selection Criteria	22
4.1.2	Question Selection	24
4.1.3	Feature Selection	26
4.1.4	Response Sub-sampling and Representativeness Test.....	27
4.2	Methodology for Research Question 1: Representational Fidelity	29
4.2.1	Dataset Preparation	29
4.2.2	Prompt Construction	30
4.2.3	Model Selection and Prompting Procedure	30
4.2.4	Response Collection and Dataset Integration	32
4.2.5	Evaluation Design and Metrics Used	32
4.3	Methodology for Research Question 2: Comparative Performance and Feature Importance	34
4.3.1	Data Preparation, Preprocessing, and Feature Engineering	35
4.3.2	Model Training on Human Survey Responses.....	40
4.3.3	Predictive Performance Evaluation.....	42
4.3.4	Model Training on LLM-Generated Responses	44
4.3.5	Feature Importance Analysis	45
4.3.6	Robustness Validation: Bootstrap Analysis.....	47
5	RESULTS.....	51
5.1	Results for Research Question 1: Representational Fidelity	51
5.1.1	Individual-Level Accuracy Results.....	51
5.1.2	Aggregate-Level Distributional Alignment	56
5.2	Results for Research Question 2: Comparative Performance and Feature Importance	78
5.2.1	Predictive Performance Evaluation.....	78
5.2.2	Feature Importance Analysis	87

6	DISCUSSION AND LIMITATIONS	101
6.1	Discussion	101
6.2	Limitations and Ethical Considerations	110
7	CONCLUSIONS	114
	REFERENCES	117
	APPENDICES	122
	Appendix A — REPRESENTATIVENESS TEST RESULTS	123
	Appendix B — SAMPLE SURVEY PROMPT STRUCTURE	125
	Appendix C — CATEGORICAL GROUPING MAPPINGS	127

LIST OF TABLES

Table 5.1:	Binomial test results comparing LLM accuracy to random baseline. P-values test whether accuracy significantly exceeds random guessing (one-tailed test). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$	53
Table 5.2:	Exact match accuracy (%) by model and question. Models ordered by overall performance (ascending, left to right).	54
Table 5.3:	Mean Absolute Error (MAE) Analysis. Lower values indicate better performance.	55
Table 5.4:	Jensen-Shannon divergence analysis (lower values indicate better alignment).....	57
Table 5.5:	Relative Cost-Efficiency Analysis.	60
Table 5.6:	Gender differences in trust responses: Human vs LLM. Values represent mean trust scores (1 = complete trust, 4 = no trust at all). Positive differences indicate women express higher trust than men.	70
Table 5.7:	Age differences in trust responses: Human vs LLM. Values represent mean trust scores for youngest (18-30) and oldest (60+) age groups. Negative differences indicate increasing trust with age, since lower numbers indicate higher reported trust levels.....	71
Table 5.8:	Country-level variation in trust responses: Human extremes vs. corresponding LLM predictions. LLM columns show the predicted values for the specific countries identified as human extremes. Lower scores indicate higher trust levels.	72
Table 5.9:	Randomness control experiment: Summary statistics for three stages of prompt structure. Values represent response frequency (%), Chi-square (χ^2) goodness-of-fit against a uniform distribution, and Shannon Entropy (H). Theoretical maximum entropy for uniform randomness is $H = 2.0$	77
Table 5.10:	5-fold cross-validation results for Machine Learning (ML) models trained on human responses. Validation accuracy and standard deviation shown for each model, averaged across all seven questions.	79
Table 5.11:	Majority-class and chance-level baseline comparisons. Majority baseline represents accuracy from always predicting the most frequent class.	80

Table 5.12: Predictive accuracy comparison: five machine learning models trained on human responses (5-fold stratified cross-validation), compared to the average and best-performing Large Language Models (LLMs) for each question. Accuracy values represent validation performance averaged across folds.	81
Table 5.13: Paired t-test results comparing ML model accuracy against LLM average accuracy across seven questions. Positive mean difference indicates LLM outperforms ML; negative indicates ML outperforms LLM. Cohen's d provides standardized effect size: small (0.2), medium (0.5), large (0.8).	84
Table 5.14: MAE comparison. Lower values indicate better ordinal proximity. For binary Q57, MAE ranges 0-1; for ordinal Q58-Q63, MAE ranges 0-3. ..	86
Table 5.15: Feature importance alignment: Spearman rank correlation (ρ) between feature importances from models trained on human responses versus LLM responses. Higher values indicate stronger alignment.	87
Table 5.16: Top-10 feature overlap between human vs LLM-response trained models. Values represent number of features (out of 10) appearing in both models' top-10 lists.	90
Table 5.17: Complete top-10 feature comparison for Random Forest across all seven questions.	93
Table 5.18: Complete bootstrap validation results: 500-iteration analysis of feature importance correlation stability.	96
Table A.1: Representativeness Test: 10,000-Respondent Subsample vs. Full WVS Wave 7 Dataset. All variables show negligible effect sizes (Cramér's V or KS < 0.1) and non-significant p-values (> 0.05), confirming subsample representativeness.	123
Table C.1: Ethnicity Grouping Mapping. Pattern matching was performed on lowercase, normalized strings using Python regex.	127
Table C.2: Language Grouping Mapping	128

LIST OF FIGURES

Figure 4.1:	Illustrative prompt structure used for LLM response generation, demonstrating the mapping of demographic attributes to survey questions.....	50
Figure 5.1:	Exact-match accuracy analysis	52
Figure 5.2:	Jensen-Shannon divergence between human and LLM response distributions across all model-question combinations. Darker colors indicate higher divergence (poorer alignment).	58
Figure 5.3:	Human Response Distributions.	62
Figure 5.4:	Q61 - Human vs Average LLM Response Distributions.	63
Figure 5.5:	Q61 - Detailed Response Distributions by LLM.	64
Figure 5.6:	Q58 - Detailed Response Distributions by LLM.	65
Figure 5.7:	Modal Concentration Factor (MCF) analysis. (Left) MCF by question; (Right) Average MCF by model.	67
Figure 5.8:	Heatmap of Modal Concentration Factors. Colors are mapped on a logarithmic scale capped at 50 to maintain contrast.	68
Figure 5.9:	Response Entropy Across Experimental Stages.	75
Figure 5.10:	Frequency Distribution of "Random" Choices across three experimental stages.....	76
Figure 5.11:	Performance comparison: XGBoost vs. best individual LLM vs. LLM average across seven questions.	83
Figure 5.12:	Feature importance alignment: scatter plots comparing human vs LLM-response trained models. Spearman correlations shown in panel titles.	89
Figure 5.13:	Top-10 Feature Importance Comparisons.	91
Figure 5.14:	Bootstrap distributions of Spearman correlations for selected question-model combinations across 500 iterations. Vertical lines indicate 95% confidence interval boundaries.....	99

LIST OF ACRONYMS AND ABBREVIATIONS

ABM	: Agent-Based Modeling
API	: Application Programming Interface
CI	: Confidence Interval
DT	: Decision Tree
ESS	: European Social Survey
GAN	: Generative Adversarial Networks
GDP	: Gross Domestic Product
GNI	: Gross National Income
GPT	: Generative Pre-Trained Transformer
GSS	: General Social Survey
HDI	: Human Development Index
JS	: Jensen–Shannon
JSON	: Javascript Object Notation
KS	: Kolmogorov–Smirnov
LGB	: LightGBM
LLM	: Large Language Model
LR	: Logistic Regression
MCF	: Modal Concentration Factor
MAE	: Mean Absolute Error
ML	: Machine Learning
MRP	: Multilevel Regression with Poststratification
RF	: Random Forest
SAE	: Small Area Estimation
SDG	: Synthetic Data Generation
VAE	: Variational Autoencoder
WVS	: World Values Survey
XGB	: XGBoost

1. INTRODUCTION

Access to high-quality data is a fundamental requirement for evidence-based decision-making across a wide range of disciplines, including but not limited to policy-making, institutional planning, marketing, and academic research [1]. Among many entities such as governments, businesses, and research institutions depend on regularly collected survey data to evaluate public opinion, consumer behavior, and societal trends [2]. The value of such data extends beyond simple descriptive statistics; reliable survey responses enable decision-makers to understand complex social phenomena, track behavioral changes over time, and identify emerging patterns that might otherwise remain hidden in aggregate statistics [3, 4]. For example, survey data plays a crucial role in shaping electoral strategies, health campaigns, and social welfare policies in the public sector [5]. Political parties rely on polling data to understand voter sentiment, identify swing demographics, and craft targeted messaging that resonates with specific groups [6]. Public health officials use survey responses to assess vaccination attitudes, track disease awareness, and design interventions that address community-specific concerns [5].

In the private sector, on the other hand, such data supports product development, market segmentation, and brand positioning efforts [7]. Companies invest heavily in understanding consumer preferences, purchase intentions, and satisfaction levels—insights that directly inform strategic decisions worth millions of dollars [8].

Finally, for academic researchers, longitudinal and cross-sectional survey datasets are often central to theory testing and model development [9]. These datasets enable researchers to test hypotheses about human behavior, validate theoretical frameworks, and contribute to cumulative knowledge in many fields of research such as psychology, sociology, economics, and political science [10]. The societal relevance of accurate data has intensified in the digital age, where public sentiment shifts rapidly across social media platforms, news cycles accelerate, and stakeholders expect real-time insights to inform

agile decision-making [11].

Yet despite its indispensable role, the data collection process remains resource-intensive and fraught with methodological challenges. Traditional survey research often requires weeks or months of fieldwork, substantial operational costs for survey administration, and significant participant incentives to ensure adequate response rates [12]. The cost of large-scale national surveys can reach up to the magnitude of millions of dollars, while specialized studies targeting hard-to-reach populations may require even greater investments in recruitment and retention strategies [13]. Beyond financial costs, the temporal dimension poses its own challenges: by the time survey results are collected, cleaned, and analyzed, the phenomena being studied may have already evolved, particularly in fast-moving domains like political opinion or consumer trends. Surveys also remain vulnerable to systematic biases: nonresponse bias occurs when non-participants differ systematically from participants [12], while social desirability bias and inaccurate self-reporting further compromise data quality. Moreover, data protection regulations (e.g. General Data Protection Regulation and California Consumer Privacy Act) add compliance burdens that increase costs and limit data retention.

Recent advances in artificial intelligence, particularly the development of LLMs, have facilitated various language-based tasks that were once thought to be beyond the capabilities of such systems [14]. These models represent a qualitative leap in natural language processing capabilities, moving beyond earlier rule-based systems and simple statistical models to architectures that can process and generate human-like text at scale [15, 14]. Trained on vast corpora of text data—often encompassing billions of words from books, websites, academic papers, and social media—LLMs demonstrate remarkable abilities to generate coherent responses, summarize complex information, and adapt across diverse domains with minimal task-specific tuning [14]. Unlike many traditional machine learning models that require extensive labeled datasets for each new task,

LLMs exhibit strong transfer learning capabilities [14]. In addition to their generalization prowess, LLMs exhibit emergent capabilities that exceed the expectations of both the user and often their developers [16]. For example, latest multimodal LLMs can identify geographic locations from images using subtle architectural and vegetation cues, commonly referred to Geoguessing, despite no targeted training on this specific task [17]. This instance demonstrates that LLMs may acquire and apply complex domain knowledge in unexpected ways – raising the broader question of: Can LLMs excel at similarly complex cognitive tasks, such as simulating human opinion formation, when guided by structured prompts and contextual cues?

This idea that LLMs could replicate individual-level survey responses by simulating human opinion rests on the assumption that human behavior, at least in the aggregate, is statistically encoded in the language data used to train these models. If models can synthesize visual and cultural cues to infer locations, as in Geoguessing, they might similarly integrate demographic cues to infer opinion patterns encoded in their training data. If LLMs can generate survey responses that align with real people matching those profiles, they might serve as cost-effective "synthetic participants", in a process called "silicon sampling" [18, 19]. If successful, this could provide a low-cost, scalable alternative to conventional survey methods, with potential applications in market research, policy testing, and social analysis. For instance, politicians could assess the likely public reception of a proposed policy before it reaches parliament; beverage companies might explore flavor preferences among core consumer segments without conducting field studies; and social media agencies could estimate which influencer is most likely to incite trust among a target audience – all by using carefully constructed artificial personas in place of traditional human respondents.

However, this hypothesis is subject to critical limitations. First, the opacity of training data for proprietary models (e.g., OpenAI models released after GPT-3) makes it

difficult to ensure that specific survey instruments or datasets were not part of pretraining, potentially introducing data leakage or memorization bias [20]. Second, the risk of the models “parroting” the memorized responses stands as a risk that might undermine any “correct” predictions made by the model [21]. Lastly, there is also another risk that even if LLMs successfully replicate survey responses, the models themselves might be prone to the same biases that plague the responses of real survey participants, or worse, if the training data over-represents certain viewpoints, model simulations might favor those perspectives, potentially excluding real individuals from the research process or misrepresenting minority groups [22].

Given these concerns, this study aims to empirically investigate whether synthetic survey responses generated by LLMs can accurately and representatively simulate the responses of real individuals. Unlike earlier exploratory studies with a single model, this investigation compares ten LLM architectures and investigates how and why their responses differ from human data. The research fills a specific gap in the literature: while prior research has demonstrated that LLMs can generate plausible survey responses in aggregate [18, 23], few studies have systematically compared LLM performance against traditional machine learning methods or examined whether LLMs capture the same demographic-opinion associations as human responses.

The investigation is structured around two interconnected research questions: (1) Can LLMs accurately simulate the responses of real individuals when prompted with specific demographic personas? (2) How do LLMs compare to traditional machine learning models in terms of predictive accuracy when both are trained on identical demographic features, and do they rely on similar demographic signals with human respondents when generating responses (feature importance)? By examining alignment at individual and aggregate levels, it is aimed to determine whether LLMs can serve as valid proxies for human opinion across questions, demographics, and response patterns.

The remainder of this thesis builds from conceptual foundations through empirical investigation to integrated synthesis. Chapter 2 reviews the literature on synthetic data generation, opinion modeling, and LLM applications in social science. Chapter 3 defines the research problem and presents the specific research questions and hypotheses. Chapter 4 describes the methodology, including survey design, data collection procedures, LLM prompting, and model specifications. Chapter 5 presents empirical results, evaluating both LLM representational fidelity and comparative performance against traditional machine learning. Chapter 6 synthesizes findings, discusses limitations and ethical considerations, while Chapter 7 finalizes the thesis by presenting the conclusions.

2. LITERATURE REVIEW

Since the main research question of this study is whether LLMs can effectively replace human participants in surveys, it is necessary to review existing literature that directly explores this possibility. At the same time, this question touches upon a broader methodological idea: one could, in principle, generate synthetic data and train models to estimate how people might respond to survey questions. In this sense, LLM-based survey simulation represents a more recent, model-driven approach to existing methods considered for survey simulation, where both the data and the reasoning process are embedded within the model itself. A review of prior work on opinion modeling, synthetic data generation, and LLM-based survey simulation is therefore essential to frame this study and to understand the implications of using these models in place of actual respondents.

2.1 Traditional Methods Considered for Survey Simulation

Prior to the emergence of large-scale generative language models, efforts to simulate or augment survey data relied predominantly on statistical inference and structural learning paradigms. While serving different purposes from estimating results for subgroups with lower data density (e.g. minorities) to preserving privacy, these approaches require a pre-existing "ground truth" dataset. This section categorizes such methods into two broad streams: predictive opinion modeling (inference based on demographic features) and structural synthetic data generation (distributional learning to mimic statistical properties of real-world data).

2.1.1 *Statistical Inference and Opinion Modeling*

To address the scarcity of survey data within specific subgroups or geographies, researchers utilize a class of model-based techniques collectively known as Small Area

Estimation (SAE). SAE methods leverage statistical models to "borrow strength" from related areas or auxiliary data sources—typically administrative records or census data—to produce reliable estimates where direct survey samples are too sparse or unrepresentative for design-based inference [24]. While early SAE approaches relied on synthetic estimators that uniformly applied national averages to local demographics, modern applications in opinion research predominantly employ regression-based methods such as Multilevel Regression with Poststratification (MRP). First popularized in political science by Park, Gelman, and Bafumi, MRP refines the SAE framework by estimating a unit-level hierarchical model that predicts individual opinion based on demographic and geographic covariates, subsequently weighting these predictions by their known population frequencies [25]. As Hanretty explains, this process allows researchers to impute missing values in sparse datasets by assuming that the conditional relationships between demographic features and opinions remain stable across varying contexts [26]. Systematic comparisons demonstrate that this approach consistently outperforms alternative methods like disaggregation, yielding more accurate and reliable subnational estimates even when using single national surveys [27]. However, like all model-based SAE techniques, MRP remains fundamentally constrained by the requirement that researchers correctly specify these stable relationships ex-ante.

Parallel to the regression-based methods, Agent-Based Modeling (ABM) attempts to simulate opinion formation through bottom-up interaction dynamics [28]. Conceptually, in ABM frameworks, virtual agents represent individual decision-makers whose behaviors evolve over time, based on simple, rule-based interactions with their environment or other agents. These agents operate according to explicit micro-rules that define their behavioral patterns and decision-making procedures [29]. For instance, Gao et al. developed a platform that simulates voting behavior based on historical patterns, demographic data, and macroeconomic indicators, successfully predicting outcomes in both

the 2020 general election in Taiwan and several states in the United States without relying on survey data [30]. Their study demonstrates that ABM can, in principle, generate population-level outputs comparable to survey-based forecasts.

However, both approaches face significant methodological constraints. While ABM enables exploration of hypothetical what-if scenarios, the methodology encounters substantial calibration challenges: learning realistic agents' behavioral patterns requires explicitly specifying decision rules and parameters that are often difficult to estimate empirically without extensive prior data [29]. Similarly, MRP and related regression-based areal interpolation methods encounter specification requirements, necessitating researchers to correctly identify demographic correlations and functional forms [31]. Both MRP and ABM thus share a fundamental limitation: the specification problem. These models require that researchers explicitly identify relevant variables and functional relationships *ex-ante* (e.g., defining the specific correlation between income and voting behavior). Consequently, if the underlying demographic drivers of an opinion are unknown, non-linear, or complex, these models are prone to significant mis-specification bias.

2.1.2 Structural Synthetic Data Generation

While not developed specifically for survey-related tasks, structural Synthetic Data Generation (SDG) methods are relevant to this study in that they attempt to generate new samples that preserve key statistical characteristics of real datasets. Unlike the predictive orientation of MRP or ABM, these methods focus on creating artificial datasets to address data access, privacy, and data-sharing concerns. The underlying motivation, in principle, aligns with the broader aim of simulating human-like responses without relying on direct participation.

Synthetic data generation centers on producing datasets that retain the statistical

properties of sensitive records (such as medical or financial data) while preventing reidentification [32]. Contemporary SDG relies heavily on deep generative architectures, specifically Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which learn to generate new data points by modeling the underlying data distribution [33, 34]. While these techniques have seen practical use in domains such as economics and healthcare, the underlying generative principle of constructing plausible synthetic observations from learned distributions has potential implications for survey simulation as well.

However, the fidelity of structural synthetic data is frequently compromised by systematic biases. Empirical evaluations of synthetic data generators using nonparametric methods demonstrate that certain approaches struggle to preserve complex multivariate relationships between variables [35]. More critically, Van Breugel et al. demonstrate that deep generative models exhibit high failure rates in "low-density regions" of the data space—effectively omitting minority groups and outliers to optimize for the statistical mean [36]. Since low-density regions often correspond to underrepresented groups in the population, synthetic data generators may systematically fail to capture the diversity of minority subpopulations [36]. By smoothing over these irregularities, structural synthetic data risks amplifying pre-existing biases in the source data and presenting an artificially homogenized representation of the population.

While these methods have not yet been widely applied for survey simulation, their ability to create statistically grounded yet artificial human data points makes them conceptually adjacent to LLM-based simulation, albeit without the semantic generalization capabilities that modern language models offer.

2.1.3 The Data-Availability Paradox

While the methods discussed above are effective for interpolating within existing distributions or anonymizing sensitive records, they are not capable of addressing the core challenge of extrapolating to new contexts; a limitation this thesis defines as the Data-Availability Paradox.

As explained in the previous section, both predictive modeling (e.g. MRP) and structural generation (e.g. GANs) are fundamentally interpolative methodologies. They function by mimicking the mathematical topology of a specific training dataset. For instance, to generate synthetic data regarding a population’s preferences, a GAN must first be trained on an empirical dataset containing those specific preferences. Similarly, to predict a demographic group’s voting intention, an MRP model requires a seed survey establishing the baseline correlation between demographics and voting behavior.

This creates a circular dependency for survey simulation: the generation of synthetic data is predicated on the prior existence of the very data one seeks to simulate. Consequently, these methods are rendered unsuitable for zero-shot research inquiries, such as predicting public reaction to a novel policy proposal or analyzing consumer preferences for a product category that has not yet been surveyed. Finally, these methods lack semantic generalization capabilities; their utility is strictly confined to the structural parameters provided in their training set.

To resolve this paradox, a methodological shift is required: moving from systems based on structural data replication to those capable of semantic generalization, the ability to infer likely opinions based on broader cultural, linguistic, and contextual associations. This necessitates the investigation of LLMs as an alternative generative framework.

2.2 Survey Simulation Using LLMs

The inherent limitations of the traditional, data-dependent methods require a fundamental paradigm shift: a transition from generative models trained on structured numerical matrices to those derived from unstructured human narratives. LLMs could facilitate this transition by leveraging the vast semantic and sociodemographic associations encoded within their pre-training corpora, effectively digitizing the aggregate of human discourse. In contrast to GANs, which rely on specific datasets to interpolate missing values, LLMs arguably be viewed as probabilistic distillates of human belief systems. This architecture, in theory, enables "zero-shot" simulation, where response patterns are generated via generalized semantic inference rather than direct interpolation of training examples. This shift from structural interpolation to semantic generalization constitutes the theoretical basis for deploying LLMs as synthetic survey participants.

2.2.1 *Concept of "Silicon Sampling" and Semantic Generalization*

The theoretical justification for utilizing LLMs in social scientific inquiry is rooted in the "distributional hypothesis"—the linguistic axiom positing that semantic similarity correlates with contextual proximity [37]. By scaling this principle across billions of parameters, LLMs implicitly encode not merely linguistic syntax, but also the complex sociodemographic correlations and cultural associations present in the training data [38]. This capacity arises from the models' foundational nature: by learning statistical regularities from massive text corpora, they ingest a "worldview" that reflects the biases, stereotypes, and social structures embedded in human discourse [39].

Argyle et al. formalized this methodological approach as "silicon sampling," proposing that LLMs can function as proxies for human populations by conditionally sampling from the model's high-dimensional latent space. The underlying mechanism is stochastic next-token prediction: the model does not retrieve a stored database record but

rather computes the most probable continuation of a sequence based on provided context. This architecture permits combinatorial generalization [18]. To provide a theoretical example, an LLM may be able to link "triathletes" to "isotonic beverage preferences" to simulate a response; although it has never encountered a specific survey linking the two; yet, it can synthesize learned representations of triathlete lifestyle attributes and beverage product characteristics to infer a probabilistically consistent preference.

Aher et al. expanded this framework through "Turing Experiments," demonstrating that LLMs can replicate robust findings from behavioral economics and psychology, including framing effects and prospect theory [23]. These findings suggest that LLMs do not simply reproduce textual patterns but simulate the cognitive heuristics and biases characteristic of human decision-making, thereby offering a viable substrate for modeling opinion dynamics in zero-shot contexts.

2.2.2 Methodology for Prompting and Persona Conditioning

The implementation of "silicon sampling" requires precise methodological intervention to steer the model from a generalized "internet mean" toward specific demographic viewpoints. This is primarily achieved through demographic conditioning and persona-based prompting.

Research by Lutz et al. indicates that explicit sociodemographic conditioning—wherein the model is assigned a structured biographical profile (e.g., "You are a 34-year-old female nurse residing in a rural constituency")—significantly outperforms implicit priming strategies [40]. By embedding these constraints within the system prompt, researchers restrict the model's generation to a subspace of tokens statistically associated with that specific persona. This facilitates the generation of heterogeneous response distributions that approximate the diversity of real-world populations. However, the construction of these personas introduces a critical granularity trade-off. Wang et al. observe

that while "rich" personas—those integrating multiple intersectional attributes such as age, race, and socioeconomic status—theoretically enhance simulation fidelity, they concurrently increase the risk of activating reductive stereotypes [19].

As important as prompting is, stochastic hyperparameters such as "temperature", which governs the entropy of token selection, also require attentive calibration. Because temperature governs the entropy of the token selection process, it directly influences the variance of the simulated population. Low temperatures may induce artificial consensus (mode collapse), while excessive entropy risks incoherent output, necessitating a delicate balance to mirror the variance observed in human survey data.

2.2.3 Results Indicating Consistency with Human Data

Empirical evaluations of this paradigm have yielded substantial evidence regarding aggregate-level alignment. The most comprehensive assessment to date, the World-ValuesBench by Zhao et al., evaluated the capacity of foundational models, including GPT-3.5 and Mixtral, to reproduce response distributions from the World Values Survey (WVS) [41]. The study established that LLMs successfully replicate macroscopic sociodemographic trends, such as the correlation between Gross Domestic Product (GDP) per capita and secular-rational values, effectively reproducing the "cultural map" of human societies [41]. Corroborating this, Bisbee et al. demonstrated that LLMs could generate feeling thermometer scores (i.e., a rating scale where respondents indicate how warm or cold they feel toward a target group) with averages closely matching American National Election Study data, though with significantly reduced variation [42]. These results suggest that for high-level social signals—opinions exhibiting strong covariance with major demographic indicators—LLMs can serve as effective "coarse-grain" simulators, providing a cost-efficient mechanism for hypothesis generation and survey piloting.

2.2.4 *Systematic Challenges*

Despite the demonstrated aggregate alignment, the validity of LLMs as individual-level survey participants is compromised by three systematic failures: algorithmic caricaturization, data contamination, and homogenization.

First, LLMs exhibit systematic representational biases when simulating demographic subgroups. Lee et al. identified algorithmic bias in climate opinion simulation, observing that models systematically underestimated global warming concerns among Black Americans—a distortion likely stemming from training data that inadequately represents minority perspectives [43]. Complementing this, Bai et al. (2024) demonstrate that even "explicitly unbiased" models harbor deep implicit stereotype biases, systematically associating specific demographic groups (e.g., Black or Muslim identities) with negative concepts (e.g., "guilt" or "weapons") and gendered roles (e.g., men with science, women with arts) [44]. These findings suggest that LLMs may rely on demographic stereotypes and prejudices rather than capturing authentic within-group heterogeneity, which can be interpreted as being prone to algorithmic caricaturization.

Second, the integrity of simulation is threatened by data contamination and memorization. As highlighted by Balloccu et al., the opacity of pre-training corpora for proprietary models complicates the distinction between genuine simulation and mere retrieval [20]. If an LLM has retained the World Values Survey within its training weights, high performance metrics may represent overfitting or "leakage" rather than genuine sociological inference.

Finally, homogenization and majority class fixation persist as significant structural limitations. Santurkar et al. found that while LLMs can be steered towards broad demographic personas, they frequently fail to capture the distributional variance of these groups, instead collapsing to a "modal" opinion, as in the example of a simulated Democrat showing 99% approval for Joe Biden [45]. This structural tendency is reinforced by

the inherent "truth-convergence" bias observed by Chuang et al. where LLM agents in social simulations drift naturally toward factual consensus and mainstream views, failing to maintain the diverse or polarized "tail" opinions characteristic of authentic human dynamics unless explicitly forced via confirmation bias parameters [46].



3. PROBLEM DEFINITION AND RESEARCH QUESTIONS

The previous chapter demonstrated that LLMs possess a promising — yet still imperfect — capacity to simulate human survey responses. While current literature highlights their semantic generalization abilities and partial alignment with empirical data, it also reveals persistent challenges related to bias, variance collapse, and insufficient methodological validation. Taken together, these findings suggest that although LLMs can approximate certain macroscopic patterns in human data, it remains unclear whether they can do so reliably, systematically, and for the right underlying reasons. In particular, the literature review identified several gaps in the existing research on LLM-based opinion simulation: (1) limited evaluation of both individual-level accuracy and aggregate-level distributional alignment within unified frameworks; (2) insufficient comparison against traditional machine learning methods to assess whether LLMs offer genuine methodological advantages; and (3) lack of systematic analysis of feature importance alignment to understand whether LLMs capture the same demographic-opinion associations as human responses.

This chapter defines the research problem and presents the specific research questions and hypotheses that structure this inquiry, explicitly linking these questions to the identified gaps. The investigation addresses whether LLMs can serve as valid synthetic survey respondents by accurately simulating human responses to opinion questions.

3.1 Scientific Goal and Research Questions

This thesis aims to empirically evaluate whether synthetic survey responses generated by LLMs can accurately and representatively simulate those of real individuals across diverse opinion domains. The investigation is structured around two interconnected research questions that address both the fundamental capability of LLMs to replicate human

responses and their potential advantages over existing methods.

3.1.1 *Research Question 1: Representational Fidelity*

The first research question is whether LLMs can accurately simulate the responses of real individuals when prompted with specific demographic personas, in terms of both individual-level accuracy and aggregate-level distributional alignment. This inquiry is structured around two specific hypotheses:

3.1.1.1 Hypothesis 1: Individual-Level Accuracy

It is hypothesized that large language models can reproduce human survey responses at the individual level when prompted with feature-rich demographic personas. Specifically, when provided with an identical demographic profile corresponding to a real survey respondent, an LLM is expected to generate responses that match the original individual’s answers with accuracy above chance-level baselines.

Testing this hypothesis requires prompting multiple LLM architectures with the same set of demographic personas and survey questions, and then evaluating the degree to which their outputs are accurate. Specifically, each model’s responses are to be compared to the corresponding individual human responses to compute exact-match accuracy.

3.1.1.2 Hypothesis 2: Distributional Representation

It is hypothesized that the aggregate distributions of LLM responses closely approximate those observed in real survey data. Beyond individual accuracy, this hypothesis tests whether the collective pattern of LLM responses, when aggregated across many demographic personas, matches the distribution of actual survey responses across the same opinion dimensions.

To test this hypothesis, Jensen–Shannon (JS) divergence is to be employed as the

primary metric, given its symmetric formulation and bounded range between 0, which indicates perfect alignment, and 1, showing complete divergence. For each question, the empirical distribution of real responses is compared against the aggregated LLM response distribution across all generated personas. Lower JS divergence values would indicate stronger alignment, supporting the hypothesis that LLMs can replicate not only individual responses but also the broader statistical properties of human opinion distributions.

3.1.2 Research Question 2: Comparative Performance and Feature Importance

The second research question is how well LLMs perform compared to traditional machine learning models in terms of predictive accuracy when both are given identical demographic features and whether they rely on similar demographic signals with human respondents when generating responses (feature importance).

This question positions the capabilities of LLMs within the broader methodological landscape, determining whether LLMs offer practical advantages over other established ML methods. The analysis focuses on two dimensions: (1) comparative predictive accuracy between LLMs and traditional ML models, and (2) feature importance alignment—whether the same demographic factors predict responses in both human and LLM data, and whether the directions and magnitudes of associations align. This inquiry is structured around two specific hypotheses:

3.1.2.1 Hypothesis 3: Predictive Performance

It is hypothesized that LLMs achieve comparable predictive accuracy to traditional ML models trained on identical demographic features. This hypothesis tests whether the computational complexity and resource requirements of LLMs are justified by superior predictive performance, or whether simpler traditional ML approaches achieve comparable or superior results.

Testing this hypothesis involves training a set of traditional machine learning models (e.g., Logistic Regression, Random Forest, XGBoost) on the same demographic features used to construct the LLM personas, and comparing their predictive accuracies to those of the LLMs. Predictive accuracy is evaluated based on the models' ability to correctly reproduce individual survey responses, allowing for a direct comparison of performance across approaches.

3.1.2.2 Hypothesis 4: Feature Importance Alignment

It is hypothesized that feature importance rankings for models trained on LLM-derived responses correlate positively with those from models trained on human responses. This hypothesis tests whether LLMs capture the same demographic-opinion associations as human response patterns, or whether they prioritize different demographic factors, which would indicate systematic differences in how LLMs encode demographic-opinion relationships.

This analysis captures a fundamental aspect of representational fidelity: whether the same demographic factors that shape human responses also influence LLM-generated answers, and whether the direction and relative strength of these associations align. For example, if women express higher levels of trust than men in human responses, this gender-based pattern is expected to be reflected in the LLM-generated data as well. Such alignment would support the notion that LLMs internalize genuine sociological structures rather than relying on superficial statistical shortcuts. Conversely, misalignment would indicate that LLMs may encode biases or idiosyncratic heuristics that diverge from real-world population dynamics.

Testing this hypothesis involves training interpretable surrogate models (e.g., logistic regression or tree-based models) on both human responses and LLM-generated responses using identical demographic inputs. Since five interpretable ML models were

trained on human responses to test Hypothesis 3, a second set of ML models are trained on the LLM-generated responses to compare the feature importances. Feature importances derived from these models would then be compared to assess alignment between the demographic-opinion associations captured by each dataset. To quantify similarity in feature importance rankings, Spearman’s rank correlation coefficient is to be employed. A strong positive correlation would suggest that LLMs encode similar patterns as human respondents, whereas divergence would indicate representational mismatch.

3.2 Contributions of This Thesis

This thesis aims to contribute to the emerging literature on LLM-based opinion simulation and synthetic data generation by addressing key gaps identified in previous research. By addressing the planned research aims, the thesis aspires to inform the ongoing methodological discussion on the role of LLMs in computational social science, identifying both the opportunities and limitations of using synthetic responses for survey research.

The planned contributions are structured along the following dimensions:

3.2.1 Integrated Evaluation of Representational Fidelity

Prior studies have often focused either on individual-level accuracy or on aggregate-level distributional alignment, but rarely both in tandem. This thesis seeks to consider multiple dimensions of fidelity to human data, allowing for a more holistic assessment of LLM performance in simulating survey responses.

3.2.2 Benchmarking Against Traditional Machine Learning Models

In contrast to evaluating LLMs in isolation, this study aims to compare their performance with established machine learning approaches using identical input features.

This comparative lens is intended to clarify whether LLMs offer tangible methodological advantages over simpler, interpretable alternatives.

3.2.3 Assessment of Demographic Conditioning Capacity

By constructing demographic personas with a rich set of attributes, the study will explore whether LLMs can simulate responses conditioned on complex, multi-dimensional demographic profiles. This contributes to understanding the limits of LLM generalization in realistic social science settings.

3.2.4 Exploration of Feature Importance Alignment

The thesis intends to analyze whether the demographic factors that influence human responses also shape LLM-generated answers in similar ways. This comparison is expected to provide insights into how well LLMs internalize demographic-opinion associations found in real-world data.

4. METHODOLOGY

This chapter outlines the methodology used to investigate the research questions defined in the previous chapter. It begins by presenting the shared foundations of the study, including survey selection, question selection, feature selection, and response sub-sampling. These elements establish the empirical basis upon which all subsequent analyses are built. Following this, the chapter proceeds to detail the methodological frameworks for evaluating LLM representational fidelity (Research Question 1) and for benchmarking LLM performance against traditional machine learning models, and comparing the feature importances between the models trained on human data and the models trained on LLM-generated responses (Research Question 2).

4.1 Experiment Design

The following subsections detail the criteria for survey selection, the justification for focusing on specific opinion questions, the rationale behind the chosen demographic attributes, along with the reasoning and methodology behind the sub-sampling of survey responses and the representativeness test conducted.

4.1.1 *Survey Selection Criteria*

The choice of survey dataset is a critical decision for this study, as it directly shapes the realism, complexity, and generalizability of the LLM evaluation. To ensure that the findings would be robust, globally relevant, and methodologically sound, the selection process followed five core criteria: (1) geographic and cultural diversity, to capture a wide range of sociopolitical contexts; (2) demographic richness, to support nuanced persona construction; (3) open-access microdata availability, to ensure replicability and transparency; (4) strong methodology, including standardized instruments and sampling

protocols; and (5) inclusion of attitudinal—not merely factual—questions, suitable for assessing opinion-based reasoning.

Given the requirement for a large-scale and methodologically established human baseline to evaluate LLM performance across global demographics, this study utilizes existing survey programs rather than primary data collection. Three internationally recognized survey programs were considered: the World Values Survey WVS [47], the European Social Survey (ESS) [48], and the General Social Survey (GSS) [49]. The ESS stands out for its methodological precision and cognitive testing of items but is geographically limited to Europe, restricting cultural generalizability. The GSS offers unmatched longitudinal depth but is confined to a single country, the United States, limiting cross-national variation. In contrast, the WVS Wave 7 (2017–2022) emerged as the most suitable dataset, offering the broadest global reach combined with methodological reliability and rich demographic coverage.

WVS Wave 7 includes over 90,000 respondents from 66 countries spanning all major world regions and income levels [47]. Its global design enables the evaluation of LLMs in truly heterogeneous settings—across languages, religions, political regimes, and economic conditions. Harmonized survey instruments, meticulous translation procedures (including back-translation and cognitive interviews), and documented national sampling methodologies further ensure comparability across contexts.

Another key strength of the WVS is its publicly available microdata, which contains over 30 demographic and socioeconomic variables per respondent. This granularity supports the construction of realistic and multidimensional demographic personas, moving beyond simple age–gender profiles. Such detail is crucial for assessing whether LLMs respond to complex demographic conditioning in ways that mirror real-world survey behavior.

4.1.2 Question Selection

From over 600 items included in the World Values Survey, this study focuses on seven questions measuring interpersonal and group-based trust among individuals (Q57–Q63). These items span multiple layers of social proximity, covering generalized trust (“most people can be trusted”), as well as trust in family members, neighbors, acquaintances, strangers, people of another religion, and people of another nationality.

While the WVS also includes a separate battery of institutional trust items—such as confidence in the press, courts, government, or political parties—these were deliberately excluded from this study. The analytical focus is limited to trust in people rather than institutions, based on conceptual and practical considerations specific to this study’s design. While both interpersonal and institutional trust are important dimensions of social attitudes, combining them could introduce unnecessary complexity and interpretative ambiguity. First, the concept of trust one feels towards social groups may not be aligned with the trust (i.e. confidence) they report towards institutions. Trust in institutions may be influenced by factors such as perceived legitimacy or political performance, whereas trust in individuals tends to reflect social proximity, group identity, and everyday moral reasoning. Given these differences, including both domains under a single construct would risk conflating distinct response patterns. Second, since the survey is conducted within a time-frame that spans several years, a change in the government, or a public incident resulting in a reduced confidence in these institutions carry the risk to skew the reported results. Therefore the confidence in institutions is considered to be too temporally unstable to include within the scope of this study. To maintain a coherent analytical scope and avoid overextension, this study focuses solely on interpersonal trust and excludes institutional trust items from the analysis.

Trust was selected as the substantive domain for several conceptual and methodological reasons:

First, trust is a foundational dimension of social cohesion and has been widely examined across sociology, political science, psychology, and economics [50, 51].

Second, trust items help mitigate memorization concerns—an important challenge in evaluating LLMs. Unlike political opinions or contemporary events that frequently appear in public discourse, the WVS trust questions are stable and longstanding survey measures. Although the general topic of trust may appear within LLM training data, the specific question wordings, scale formats, and demographic-conditioned response distributions are less likely to have been encountered verbatim—let alone at the level of how over 90,000 individual respondents answered each item across diverse demographic profiles. This reduces the risk that model outputs reflect memorized associations rather than internal reasoning processes.

Third, trust measures exhibit relatively consistent conceptual interpretation across societies while still capturing substantial cross-national variation. Unlike questions about specific policies or political figures that may have culture-specific meanings, the concept of trust represents a cross-culturally recognized construct. This universality provides a stable foundation for comparing LLM performance across different demographic groups without confounding cultural interpretation differences with model prediction errors. At the same time, the observed variation in trust across countries, religions, and social groups provides a rich empirical setting to test whether LLMs can generalize across diverse populations.

Fourth, the structure of the selected trust items—comprising six four-point ordinal scales and one binary indicator—is considered to present an analytically appropriate level of complexity. Compared to purely binary questions where random guessing achieves 50% accuracy, the predominant use of ordinal scales reduces the likelihood of chance-level performance and requires models to differentiate between graded attitudinal positions. Yet the scales remain simple enough to allow for transparent comparison

of distributions and the use of distance-based evaluation metrics across both binary and ordinal formats.

Finally, trust questions impose meaningful cognitive demands on both humans and LLMs. Responding requires implicit reasoning about social norms, intergroup relations, and perceived risk—processes that cannot be reduced to factual recall. For LLMs, these items require the activation and integration of demographic associations learned during pretraining. Their performance therefore provides insight into whether demographic-conditioning prompts elicit structured opinion patterns or whether model outputs reflect generic heuristics. The ability (or inability) of LLMs to simulate trust responses serves as an informative indicator of their broader capacity to generate realistic attitudinal data.

4.1.3 Feature Selection

This study uses demographic variables to condition LLM outputs in a way that mimics how surveys collect background information about respondents. To preserve the zero-shot nature of the task, only variables that correspond to observable or externally verifiable demographic traits were included. Subjective opinion variables—such as political orientation or social trust itself—were deliberately excluded from conditioning inputs to avoid circularity and overfitting.

All features were drawn from the “Demographics” section of the World Values Survey Wave 7 dataset. These include age, gender, household structure, education, employment, religion, ethnicity, immigrant status, and similar attributes. A total of 32 variables were retained. The only variable excluded was “year of birth” (Q261), since the survey was conducted over a span of five years (2017–2022), making it a temporally unstable proxy for age. Instead, the “age” variable (Q262) was used directly. No derived or composite variables were created.

The inclusion of a wide set of demographic variables is not meant to imply causal

interpretation. Rather, the aim is to simulate realistic and information-rich personas that mirror the types of background characteristics available to researchers in traditional survey settings. These features are understood as correlates or proxies for deeper social processes—e.g., age for life experience, education for class and cognition—rather than as standalone determinants of trust.

Crucially, these demographic traits are those that survey researchers could plausibly infer or collect without asking the respondent to express an opinion. This aligns with the goal of assessing whether LLMs can simulate plausible responses based solely on externally observable characteristics, not based on hints from other attitude items. The restriction helps ensure that the generation process remains a genuine test of demographic conditioning rather than a disguised form of response copying.

4.1.4 Response Sub-sampling and Representativeness Test

To enable tractable experimentation and reduce computational complexity, a representative subsample of 10,000 respondents was drawn from the full World Values Survey Wave 7 dataset, which originally includes 97,220 individuals. Sub-sampling was necessary due to the prohibitive cost of generating and evaluating synthetic responses across the entire corpus. However, to ensure the validity of results and generalizability of conclusions, careful attention was paid to the representativeness of the chosen subsample.

Before subsampling, all individual responses were assigned an ID between 1 to 97,220 as an index column, in the original order of the survey file. Therefore the individual responses could easily be matched in future applications. Subsequently, the subsample was created using simple random sampling without stratification, with a fixed seed value (`random_state=42`) to ensure reproducibility. The same procedure was applied consistently throughout the study, including during prompting and evaluation phases.

To verify that the 10,000-respondent subsample reflects the overall composition

of the WVS Wave 7 population, a set of statistical tests was conducted across the 32 demographic variables.

The representativeness assessment employed both categorical and continuous variable comparison methods. For the 29 categorical variables, chi-square tests of independence were conducted to compare frequency distributions between the subsample and full dataset. In addition, Cramér's V was computed as an effect size measure, indicating the magnitude of any differences observed. For the three continuous or ordinal variables (age, household size, income level), the Kolmogorov–Smirnov (KS) two-sample test was applied, with the KS statistic used to quantify distributional divergence. Full results are provided in the appendix (Appendix A).

Given the large underlying sample sizes, p-values were treated cautiously due to their sensitivity to minor deviations. Instead, emphasis was placed on effect sizes to determine practical equivalence. Standard interpretive thresholds were adopted: for Cramér's V, values below 0.1 were treated as negligible; for KS statistics, values below 0.1 indicated small differences.

Results confirmed that the subsample is highly representative across all 32 tested variables. All categorical variables (29/29) exhibited negligible effect sizes (Cramér's V < 0.1), and all continuous or ordinal variables (3/3) exhibited small differences (KS < 0.1). The largest observed discrepancy was for ethnicity (Q290), with a Cramér's V of 0.059—still well below the 0.1 threshold. All associated p-values exceeded 0.05, indicating no statistically significant differences under conventional criteria.

This high level of alignment suggests that the subsample closely mirrors the full dataset in terms of demographic structure. Consequently, analytical conclusions drawn from the subsample can be generalized to the broader WVS Wave 7 population without concern for sampling bias. The use of random sampling and extensive validation procedures ensures the methodological integrity of subsequent analyses and the robustness of

findings.

4.2 Methodology for Research Question 1: Representational Fidelity

This section outlines the methodology used to assess whether LLMs can accurately simulate the survey responses of real individuals when prompted with demographic personas, as posed in Research Question 1. The goal is to evaluate both (i) the accuracy of different model outputs at the individual level, and (ii) the alignment of aggregate response distributions with real-world survey data.

4.2.1 Dataset Preparation

To ensure compatibility between real and synthetic responses, the WVS Wave 7 dataset was preprocessed to retain only the relevant variables. Specifically, the dataset was reduced to 40 columns: 32 demographic features as described in the previous section, 7 trust-related survey items (Q57–Q63), and a unique respondent identifier as previously mentioned. All non-relevant variables were dropped.

Certain categorical variables—such as country codes or binary responses—were mapped to human-readable English labels for use in natural language prompts. For example, “1” and “0” responses were converted to “Yes” and “No”, and ISO country codes were replaced with full country names. The prompt design followed a consistent structure across all respondents and retained all demographic fields where valid data existed.

Missing or invalid demographic data codes (negative values ranging from -1 to -5) were converted to NaN in accordance with WVS documentation guidelines. These missing values were omitted from the prompts rather than replaced with null tokens, ensuring that only available information was passed to the model. All 10,000 respondents were retained in the final dataset, with missingness handled at the prompt level rather than via row exclusion.

4.2.2 Prompt Construction

A custom script was used to generate a prompt for each respondent in the dataset. Each prompt took the form of a short persona description followed by an instruction to answer the seven trust questions using the same scale as in the original survey. An illustrative example structure is shown below; a full version is provided in the appendix (Appendix B).

Prompts were generated in English to avoid introducing linguistic variation and to maintain uniformity across models. The wording was deliberately kept simple and consistent with the terminology used in the original survey documentation. The prompt was first tested with a smaller sample of 100 responses, where it yielded the expected JSON output in the desired format. During the early stages of this study, OpenAI’s Structured Outputs functionality—allowing model outputs to be natively constrained to JSON formats—was not yet available. Initial tests therefore relied on explicitly instructing the model within the prompt to “provide the answer in JSON format,” and response parsing verified that models complied reliably with this structure. To ensure methodological consistency and to avoid introducing prompt contamination or format-related confounds across models, the same instruction-based JSON formatting approach was retained throughout the remainder of the study, even after newer models introduced native structured output capabilities. As a result, prompt generation remained deterministic and uniform across all model families: every respondent with valid demographic values received a single fixed prompt with identical formatting constraints.

4.2.3 Model Selection and Prompting Procedure

To improve the generalizability of the individual-level accuracy and distributional alignment results, ten LLM architectures were selected based on diversity in size, training architecture, and public accessibility. These included:

- OpenAI Models: GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-5, GPT-5-mini, GPT-5-nano, GPT-4o-mini, GPT-o4-mini (reasoning)
- DeepSeek Models: Deepseek V3, Deepseek V3.1

The specific version numbers of these models are as follows: gpt-4.1-2025-04, gpt-4.1-mini-2025-04-14, gpt-4.1-nano-2025-04-14, gpt-5-2025-08-07, gpt-5-mini-2025-08-07, gpt-5-nano-2025-08-07, gpt-4o-mini-2024-07-18, GPT-o4-mini-2025-04-16, deepseek-chat-V3-0324, deepseek-chat-V3.1

Each model was prompted with the exact same set of 10,000 persona prompts. For Generative Pre-Trained Transformer (GPT)-based models, the OpenAI Batch Application Programming Interface (API) was used to efficiently submit prompts and collect responses. Batch requests were divided into ten separate files (each containing 1,000 prompts) to comply with API input limits. For DeepSeek models, responses were collected via standard API calls.

All LLM queries were executed without specifying a temperature parameter (which defaults to temperature = 1.0) due to three major reasons. First, the study aims to preserve the zero-shot nature of the task, and default settings represent the model’s intended behavior without researcher-imposed tuning. Second, adjusting the temperature would introduce systematic distortions: lowering it increases determinism and reduces output variance, but risks collapsing responses toward modal categories—precisely the types of variance reductions this study seeks to measure. Conversely, raising the temperature increases randomness and reduces internal coherence, which would artificially inflate noise unrelated to the underlying representational capability being assessed. Since most LLMs operate within a recommended range of 0–2, retaining the default 1.0 temperature provides a balanced setting that avoids both extremes. Third, OpenAI is found to have removed the parametric control of temperature in their reasoning models and all models

after GPT-5 as they find it is disruptive of the reasoning process incorporated into their newer models. Therefore using the default parameter proved to be the more future-proof approach within the scope of this study.

4.2.4 Response Collection and Dataset Integration

The LLM-generated responses were collected and stored alongside the original WVS responses. Each synthetic answer was matched by ID to its corresponding human response, and integrated as new columns in the master dataset using the following naming convention: Q57_modelname, Q58_modelname, etc.

This yielded a unified evaluation dataset containing real responses, demographic features, and LLM-generated responses from each model. This structured dataset enabled the direct comparison of exact-match accuracy and aggregate distributional similarity across models and questions.

4.2.5 Evaluation Design and Metrics Used

To assess the representational fidelity of LLM-generated survey responses, the evaluation framework adopts a two-tiered approach, examining both (i) individual-level accuracy and (ii) aggregate-level distributional alignment. Each evaluation tier is applied both at the question level and model level, enabling granular and comparative assessments across dimensions.

4.2.5.1 Individual-Level Evaluation

The primary metric for individual-level alignment is exact-match accuracy, defined as the proportion of cases in which a model’s predicted response exactly matches the real respondent’s answer. This provides a strict baseline for response-level fidelity, but has the limitation of treating all deviations equally—for instance, predicting “2” instead

of the true label “1” is penalized identically to predicting “4.”

To address this limitation, Mean Absolute Error (MAE) is computed between predicted and actual ordinal responses. By capturing the average magnitude of deviation without regard to direction, MAE provides a more informative assessment of how far a model’s predictions deviate from real-world responses on a continuous scale. Both exact-match accuracy and MAE are computed (i) per trust question (Q57-Q63), averaged across all models, and (ii) per model, averaged across all questions.

To benchmark the performance, exact-match accuracy results are compared against theoretical random baselines derived from the number of valid response options for each question. For instance, trust items with four ordinal options yield a baseline exact-match accuracy of 25%, while the binary generalized-trust question results in a 50% baseline for a naïve point of comparison, which allows for a meaningful assessment of whether LLMs exhibit any capacity to exceed chance-level performance in both accuracy and response realism.

4.2.5.2 Aggregate-Level Evaluation

Beyond individual predictions, it is assessed how well each model replicates the overall statistical distributions observed in real survey responses. For this purpose, JS divergence is computed between the empirical distribution of the model-generated responses and the actual response distribution in the WVS dataset. As in the individual-level analysis, JS divergence is evaluated both by question and by model.

4.3 Methodology for Research Question 2: Comparative Performance and Feature Importance

This section describes the methodology used to address Research Question 2, which aims to evaluate (i) whether LLM-generated survey responses can achieve predictive performance comparable to traditional machine learning models trained on human responses, and (ii) whether models trained on human and LLM-generated data prioritize similar demographic features when making predictions.

To systematically examine these aims, the methodological approach follows a two-part analytical strategy. First, traditional machine learning models are trained on human survey responses to establish a benchmark for predictive performance. Second, the same modeling framework is applied to LLM-generated responses to enable a systematic comparison of feature importance across human-based and LLM-based predictions.

This methodology is implemented through six sequential but conceptually grouped stages, described below. For clarity, each stage is briefly outlined below before being discussed in detail in the subsequent subsections.

The methodology consists of six stages: (i) Data preparation, preprocessing and feature engineering, including cleaning, encoding, high cardinality variable reduction, construction of derived features, multicollinearity control, feature scaling, and missing value imputation; (ii) Model training on human survey responses, using explainable machine learning models; (iii) Predictive performance evaluation, evaluating the performance of the trained ML models and comparing with the predictive accuracy of the LLM-generated responses; (iv) Model training on LLM-generated responses, applying the same modeling framework to enable direct comparison of learned decision structures; (v) Feature importance analysis, assessing and contrasting the demographic variables prioritized by models trained on human versus LLM-generated data; (vi) Robustness and stability validation, employing repeated random sampling and retraining procedures to

test the sensitivity of results to data splits.

4.3.1 Data Preparation, Preprocessing, and Feature Engineering

This stage focused on preparing the feature matrix used to train machine learning models in subsequent stages of the analysis. Given that Research Question 2 investigates both predictive performance and feature importance alignment, the goal here was to transform the raw demographic and survey data into a structured, interpretable, and robust input suitable for both accurate modeling and meaningful post-hoc explanation. The preprocessing process was designed to balance predictive performance with the interpretability required for feature importance comparison.

The starting point was the integrated dataset constructed for Research Question 1, which contained 10,000 respondents from the WVS. Each respondent was represented by 32 demographic variables and provided answers to seven trust-related questions (Q57-Q63). In addition, the dataset included corresponding responses generated by ten different LLMs for the same questions. This structure enabled all models to be trained on identical demographic inputs, with the only variation being the type of target variable—human or LLM response—thus supporting fair comparisons in both performance and feature importance analyses.

Demographic variables were encoded according to their measurement type. Binary variables such as gender, urban/rural residence, citizenship status, co-residence with parents/in-laws and chief-earner status were represented as 0/1 indicators. Ordinal variables including education levels, budgeting status, income deciles, and perceived social class were converted to numeric values consistent with WVS coding. Nominal variables such as marital status, employment status, and occupational categories were one-hot encoded to preserve categorical structure. These transformations yielded a comprehensive set of features reflecting respondents' socioeconomic and cultural profiles.

However, several variables exhibited very high cardinality: country of residence included 66 distinct values; each of the origin country fields (respondent, mother, father) had over 100 unique entries; ethnicity featured 210 categories; and native language spanned 131 different values. One-hot encoding these variables would have resulted in hundreds of sparse features, which could reduce the interpretability of models and the reliability of feature importance estimates. Since interpretability is central to Research Question 2, these variables were transformed into aggregated or macro-level indicators that preserved informative variation while reducing sparsity and redundancy.

Country-related variables were replaced with 12 macroeconomic and governance indicators sourced from the World Bank:

- Worldwide Governance Indicators: Control of Corruption, Government Effectiveness, Rule of Law, Voice and Accountability
- World Development Indicators: GDP per capita, Unemployment rate
- Human Development Indicators: Human Development Index (HDI), Life Expectancy, Expected Years of Schooling, Mean Years of Schooling, Gross National Income (GNI) per capita, GNI-HDI rank difference

All four country fields—survey location, respondent origin, mother’s origin, and father’s origin—were mapped to these indicators, resulting in 48 interpretable numeric features, a significantly lower number compared to the 377 binary country features had one-hot encoding been employed. This representation also enabled models to learn from underlying political and economic conditions, rather than from opaque country labels, since for instance, identifying a respondent as living in a country with high or low GDP per capita is more interpretable and analytically informative than merely indicating that they are from Germany, Uganda, or Switzerland.

Ethnicity and language variables were also consolidated. Ethnicities were grouped into nine broad categories, inspired by the United States Census Bureau standards and adapted for global applicability (Black/African, Caucasian/White, Central Asian, East-/SE Asian, Hispanic/Latino, Indigenous/Native, Middle Eastern/West Asian, South Asian, and Other). Languages were grouped into the seven most common in the dataset that corresponds to the majority in the dataset (Spanish, English, Arabic, Chinese, Russian, Portuguese, French), with the remaining 124 languages assigned to a generic "Other" category. Complete list of mappings configured can be found in the appendix (Appendix C).

An additional derived variable, `Economic_Migration_Delta`, was constructed to capture the economic mobility between respondents' ancestral origins and their current country of residence. Given the highly right-skewed distribution of GDP per capita across countries, GDP values were first log-transformed to reduce the influence of extreme values and to reflect the diminishing marginal utility of income commonly assumed in economic theory [52, 53]. The resulting variable, `Economic_Migration_Delta`, was defined as the log difference between GDP per capita in the country of residence and GDP per capita in the ancestral country:

$$\Delta_{EM} = \ln(G_{res}) - \ln(G_{anc}) = \ln\left(\frac{G_{res}}{G_{anc}}\right) \quad (4.1)$$

Positive values of (4.1) indicated upward economic migration, negative values indicated downward mobility, and values close to zero reflected minimal economic change across generations. This log-ratio formulation brought several advantages: it ensured symmetry, with proportional upward and downward movements expressed with equal

magnitude but opposite sign; it introduced scale independence, treating proportional changes similarly regardless of absolute income levels; and it helped reduce multicollinearity by summarizing relative economic trajectories in a single interpretable feature, while still allowing absolute economic conditions (e.g., current GDP per capita) to be modeled separately.

To further improve model stability and interpretability, a systematic multicollinearity analysis was conducted. Pairwise Pearson correlations¹ were computed across all features, and variables exceeding a correlation threshold of $r = 0.80$ were flagged for consolidation or removal. In each case, decisions were made to retain the most theoretically central or generalizable variable while excluding redundant alternatives.

Among GDP-related features, six variables exhibited extremely high correlations ($r = 0.85$ to 0.98). Of these, two were retained: Log_GDP_Res, capturing absolute economic context, and Economic_Migration_Delta, encoding the economic mobility. The remaining four—Res_gdp_cap_2022, Res_GNI_per_capita, Ancestral_GDP_Cap, and Log_GDP_Ancestral—were excluded due to their redundancy and overlap with the retained indicators.

A similar decision was made for governance-related indicators. Res_gov_eff_2022 (government effectiveness) and Res_law_2022 (rule of law) exhibited a correlation of $r = 0.94$. Since government effectiveness more directly captures state capacity in institutional settings, it was retained, while rule of law was excluded. In the social domain, IS_MARRIED was found to be highly correlated ($r = 0.89$) with EDU_SPOUSE, a variable representing the respondent's spouse's education level and imputed as zero for unmarried individuals. Because EDU_SPOUSE implicitly captures both marital status and educational homogamy, IS_MARRIED was excluded.

¹Pearson's correlation coefficient (r) measures the strength of linear association between two continuous variables, with values closer to $|1|$ indicating stronger linear relationships.

Finally, identifier variables such as Response ID were removed to prevent potential data leakage from geographic ordering patterns present in the original dataset.

Following multicollinearity reduction, the remaining feature set was standardized and completed to ensure numerical stability and consistency across model classes. All continuous variables were standardized using z-score normalization (mean = 0, standard deviation = 1), ensuring that each feature contributed comparably to model training regardless of its original scale and preventing features with large numeric ranges from disproportionately influencing model estimation, particularly in regularized linear models. Categorical variables were uniformly one-hot encoded to preserve categorical distinctions and ensure compatibility across all modeling approaches, which in turn allowed for direct comparisons between models without introducing variability from preprocessing differences. All five models used in the analysis—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), LightGBM (LGB), and XGBoost (XGB)—were trained on this common feature representation to ensure fairness and methodological consistency in performance and feature importance evaluations.

Missing values were handled at the feature level rather than through respondent exclusion. Continuous variables with missing entries were imputed using the median, providing a robust central estimate unaffected by outliers. Categorical features were imputed either with the mode or assigned to a dedicated “missing” category, depending on their distribution and semantic structure, thus preserving the complete sample of 10,000 respondents and maintained input consistency across all model types and training regimes. In total, the final feature matrix contained 121 demographic predictors per respondent. This unified representation served as the standardized input across all five models and all outcome types (human- and LLM-generated), ensuring that observed differences in model performance and feature importance reflected variation in response behavior rather than inconsistencies in data preprocessing.

4.3.2 Model Training on Human Survey Responses

The second stage involved training machine learning models on human survey responses to establish baseline predictive performance and to extract demographic feature importance patterns from real-world data. Seven classification tasks were defined, one for each of the trust-related survey questions (Q57–Q63), with responses treated as the target variables and the 121 preprocessed demographic features used as input.

Five supervised learning algorithms were selected to balance interpretability, flexibility, and predictive accuracy: Logistic Regression, Decision Tree, Random Forest, LightGBM, and XGBoost. This selection provided a diverse set of modeling paradigms, including linear and non-linear learners, single and ensemble models, as well as bagging and boosting strategies. Logistic regression was picked for its interpretable coefficients and considered as a linear baseline. The decision tree was selected for intuitive rule-based predictions, while random forest would introduce further generalization through variance reduction. LightGBM and XGBoost, both gradient boosting methods, were included for their ability to capture complex non-linear interactions and for their strong empirical performance on high-dimensional structured data.

Each algorithm was independently trained for each question, resulting in 35 models in total (5 models $\times$ 7 questions). For each task, the dataset was randomly split into a training set (80%) and a validation set (20%) using stratified sampling to preserve class distributions, with fixed random seed (random_state=42) for reproducibility. The same split was used across all five models for each question to ensure comparability. Due to class imbalance in many questions (e.g., over 78% of responses concentrated in a single class in Q58, over 52% in Q60), models were trained using class weights inversely proportional to class frequencies to prevent majority-class dominance.

To support transparency and reproducibility—and to produce stable feature importance estimates central to Research Question 2—the specific hyperparameter configurations used for each model class are detailed below. These settings were not the result of exhaustive grid search, but were informed by domain best practices, empirical validation on structured data, and the need to control overfitting in the presence of a moderately high feature-to-sample ratio (121 predictors and 10,000 samples per trust question). All models were trained with balanced class weights where applicable to address label imbalance across trust responses. Model parameters were selected as below:

- **Logistic Regression:** *max_iter=1000, C=0.1, penalty='l2', multi_class='multinomial', solver='lbfgs', class_weight='balanced', random_state=42.*
- **Decision Tree:** *max_depth=8, min_samples_split=50, min_samples_leaf=20, class_weight='balanced', random_state=42.*
- **Random Forest:** *n_estimators=100, max_depth=8, min_samples_split=50, min_samples_leaf=20, max_features='sqrt', class_weight='balanced', random_state=42.*
- **LightGBM:** *n_estimators=100, max_depth=6, learning_rate=0.05, reg_alpha=0.1, reg_lambda=0.1, min_child_samples=50, class_weight='balanced', random_state=42.*
- **XGBoost:** *n_estimators=100, max_depth=6, learning_rate=0.05, reg_alpha=0.1, reg_lambda=0.1, min_child_weight=5, eval_metric='logloss', random_state=42.*

Since XGBoost requires 0-indexed class labels (0, 1, 2, ...) whereas WVS uses 1-indexed scales (1, 2, 3, 4), class labels were dynamically remapped to 0-indexing before

training XGBoost models, with predictions remapped back for evaluation. This was applied consistently across all XGBoost procedures (train-validation, cross-validation, bootstrap).

4.3.3 Predictive Performance Evaluation

The third stage of the methodology involved two analyses. First, the predictive performance of the ML models trained on human survey responses was evaluated using standard metrics. Second, the predictive performance of LLMs was compared with the performance of ML models.

Model performance was evaluated on the 5-fold cross-validated set using two metrics:

- **Accuracy:** Proportion of correct predictions (exact matches). While intuitive, accuracy can mislead for imbalanced data where predicting only the majority class achieves high scores.
- **Mean Absolute Error (MAE):** Average absolute difference between predicted and actual class labels, treating ordinal scales as numeric. Penalizes larger deviations and provides graded assessment (predicting "2" when true is "1" is less severe than predicting "4").

To establish a benchmark for the accuracy results, two reference baselines were introduced:

- **Random baseline:** The expected accuracy from uniform random guessing, such as 25% for four-class and 50% for the binary trust question (Q57). This represents the lower bound of uninformative predictions.

- Majority-class baseline: The naive accuracy achieved by always predicting the most frequent class in the training set. For example, in Q58, where over 78% of respondents selected Class 1, this baseline reaches 78%. Models must exceed this threshold to demonstrate predictive signal beyond trivial heuristics.

To assess the robustness of model performance beyond a single train-test split, 5-fold stratified cross-validation was performed on the training set. Each model was re-trained five times using different 80/20 splits that preserved class distributions within each fold. Accuracy was recorded for each fold, with results reported as mean $\pm$ standard deviation. Higher variance across folds indicates sensitivity to training composition; lower variance suggests stable generalization. `StratifiedKFold(n_splits=5, shuffle=True, random_state=42)` was used to ensure reproducibility.

In parallel to ML model evaluation, the performance of LLMs as obtained within the scope of Research Question 1 was then directly included in the model performance comparison, treating the LLMs as a sixth predictive model.

To formally assess whether differences in performance between LLMs and ML models were statistically significant, paired t-tests were conducted for each ML algorithm using per-question accuracy scores. For each test, the accuracy of the ML model and the averaged LLM predictions were compared across the seven questions, leveraging the paired nature of the task (same questions, same target variable). This design minimized variability due to question-specific difficulty and increased statistical power.

For each comparison, the following quantities were computed: (i) the mean difference in accuracy between the ML and LLM models; (ii) the corresponding t-statistic and p-value from the two-tailed paired t-test; (iii) the 95% confidence interval for the mean difference; and (iv) Cohen's d, a standardized effect size that contextualizes the magnitude of the observed difference. Interpretations considered both statistical significance (i.e., whether $p < 0.05$) and practical significance, especially in light of the small sample

size ($n=7$ questions), which may limit statistical power and increase the risk of Type II error.

In addition to comparing average LLM performance, the best-performing LLM per question was identified and recorded. This allowed examination of whether model averaging masked the capabilities of top-tier individual models. For each trust question, the LLM with the highest exact-match accuracy was noted and included in comparative tables and visualizations. With this dual lens—LLM average vs. ML and best LLM vs. ML—it was aimed to provide a more nuanced understanding of LLM capabilities.

4.3.4 Model Training on LLM-Generated Responses

With the aim of testing the hypothesis 4, it was necessary to understand which demographic features are considered important by LLMs. To investigate this, the same five models (LR, DT, RF, LGB, and XGB) were retrained using LLM predictions as target variables.

Unlike the human response models, where each respondent had a single ground-truth answer per question, here each respondent had ten distinct synthetic answers per trust question—one from each of the ten LLMs. As a result, the five ML models were trained separately for each LLM and each question. For example, for Q57, 50 models were trained in total ($10 \text{ LLMs} \times 5 \text{ ML models}$).

The same 80/20 stratified train-validation split, class-weighting strategy, and hyperparameter settings used in the human-based models were applied here as well. This ensured that any observed differences in feature importance were due to differences in the response source (human vs. LLM), not due to inconsistencies in model configuration or data sampling.

4.3.5 *Feature Importance Analysis*

To compare how ML models prioritize demographic features when trained on human versus LLM-generated responses and to determine whether ML models trained on LLM responses rely on similar demographic reasoning patterns as those trained on human data, feature importance values were extracted and analyzed.

Conceptually, one can think of two parallel training regimes: in the first, ML models are trained using real human survey responses as the target; in the second, the same ML models are trained using the synthetic responses generated by ten different LLMs. While the learning algorithms and the features remain identical, the “data diet” differs—models are exposed either to human judgment or to LLM-generated approximations of that judgment.

As the selected ML models were either linear or tree-based—both commonly used in explainable machine learning—feature importance scores could transparently be obtained from their underlying classification weights or split-based metrics. To be more specific:

- Logistic Regression: Absolute value of learned coefficients. Larger coefficients indicate stronger marginal associations.
- Decision Tree: Total reduction in Gini impurity attributable to each feature across all splits.
- Random Forest: Mean decrease in Gini impurity aggregated across 100 trees.
- LightGBM/XGBoost: "Gain"—total improvement in loss function contributed by each feature across splits.

In parallel to the ML models trained on human data, the same extraction procedure was applied to ML models trained on LLM responses. Given that ten different LLMs were

used, ten distinct feature importance profiles were obtained for each model–question combination. To derive a single importance coefficient per feature, importance values were averaged across the ten models trained on LLM responses. For example, for Q57–Random Forest, the average importance of GDP_per_capita reflects how strongly this feature was prioritized by ML models trained on LLM-generated answers. This aggregation approach provided a stable estimate of which features were emphasized when models were trained on synthetic data and enabled a direct, model-by-model comparison with the importance rankings derived from models trained on real human data.

For each ML model–question pair—regardless of whether the target was human or LLM-generated—feature importance values were normalized to sum to 1, enabling cross-model and cross-question comparisons. This allowed for consistent cross-model and cross-question comparisons by eliminating scale differences inherent to algorithm-specific metrics.

To quantify alignment between feature prioritization in the two training regimes, Spearman’s rank correlation coefficient (ρ) was computed for each model–question pair. Spearman’s ρ measures monotonic association between two ranked lists and is robust to scale differences in absolute magnitude, making it well suited for comparing relative feature prioritization. A value of $\rho = 1$ indicates perfect agreement; $\rho = 0$ indicates no association; and negative values suggest inverse prioritization. Correlations were calculated for each of the 35 model–question combinations (5 models $\times$ 7 questions), by comparing feature importance rankings from ML models trained on human responses with those trained on the averaged LLM responses. These correlation values were then summarized at the model level—by averaging across questions—to identify which algorithms exhibit consistent feature prioritization across response types.

In addition to model-level aggregation, feature importance was also examined at the question level. For each trust question (Q57–Q63), the most influential demographic

features were identified separately for models trained on human responses and those trained on LLM responses. This question-specific analysis made it possible to assess whether LLMs emphasize similar predictors for individual trust dimensions, or whether alignment observed at the aggregate level concealed divergence in specific questions.

4.3.6 Robustness Validation: Bootstrap Analysis

To assess the robustness of feature importance comparisons and to evaluate the stability of Spearman correlation values under repeated sampling, a bootstrap resampling analysis was conducted. This procedure was designed to test whether observed correlations between feature importance rankings of models trained on human data and those trained on LLM responses reflect stable underlying patterns or are artifacts of a particular train-validation split, sample composition, or class imbalance.

For each trust question and ML model, 500 independent bootstrap iterations were performed. In each iteration, the following steps were applied:

1. **Bootstrap Resampling:** A sample of 10,000 respondents was drawn with replacement from the full preprocessed dataset. Because resampling is performed with replacement, each sample contains a subset of the original respondents multiple times while excluding others entirely, typically yielding around 63% unique individuals per sample. This creates natural variation in sample composition while holding sample size constant.
2. **Train-Validation Split:** Each bootstrap sample was split 80/20 using stratified sampling to preserve class distributions across splits.
3. **Model Training:** Two models were trained per iteration—one using human responses and one using the averaged LLM responses as the target—using the same 121-variable demographic profiles and identical model configurations.

4. Feature Importance Extraction: Feature importance values were computed using algorithm-specific methods, consistent with prior analysis.
5. Spearman Correlation: The Spearman rank correlation coefficient (ρ) was computed between the feature importance rankings of the models trained on human data and models trained on LLM-generated data.

This process produced a distribution of 500 Spearman correlation values for each model–question combination, representing the variability expected under repeated population-level resampling.

The use of n-from-n bootstrap (drawing 10,000 respondents with replacement from the original 10,000) was intentionally chosen over smaller subsampling (e.g., 2,000 from 10,000) to isolate sampling variance without reducing statistical power. This approach is standard in bootstrap-based stability assessment and allows for direct comparison with the primary analysis conducted on the full dataset.

Several safeguards were implemented to address potential edge cases arising from resampling:

- Single-Class Samples: Iterations in which the training or validation set contained only one class were excluded, as classification algorithms require at least two classes to function. This was especially relevant for severely imbalanced trust questions.
- Insufficient Variance: For questions where human or LLM responses exhibited near-uniformity (e.g., >99% single-class), meaningful feature importances could not be computed. Such iterations were flagged and excluded from summary statistics, with exclusions documented in the results chapter.
- Model Compatibility: Preprocessing steps ensured compatibility across all iterations, including class remapping for gradient boosting models that require zero-indexed labels.

The number of successful iterations (i.e., runs in which models were successfully trained and Spearman values could be computed) was recorded for each model–question pair. Summary statistics were calculated only from these successful runs.

For each model–question pair, the resulting bootstrap distribution of Spearman correlation values was summarized using the following statistics:

- Mean (μ): Central tendency of Spearman correlation across successful iterations.
- Standard Deviation (σ): Variability of correlation values, indicating sensitivity to sampling composition.
- 95% Confidence Interval: Non-parametric percentile-based interval (2.5th–97.5th), representing the range within which the true correlation is expected to lie with 95% confidence.
- Number of Iterations (N): Count of valid iterations included in the calculation.

These statistics were aggregated at both the model level (averaged across questions) to evaluate algorithm-specific stability, and at the question level (averaged across models) to assess which trust items yielded the most consistent feature prioritization patterns.

Ultimately, this bootstrap-based analysis provides a critical robustness check, ensuring that conclusions drawn from the main analysis are not overfit to a single sample split and are generalizable under realistic sample variation.

Answer the following survey questions realistically and consistently with the provided demographic characteristics. Provide your answers in Javascript Object Notation (JSON) format.

You have the following attributes:

- Country of Residence: Indonesia, Urban, Citizen
- Country of Birth: Indonesia
- Gender: Female • Age: 35
- Language at Home: Sundanese
- Education: Secondary • Marital Status: Married
- Employment: Full-time • Occupation: Teacher
- Income Level (1-10): 5 • Social Class: Lower middle
- Ethnicity: Sundanese

Now, please answer the following questions: [Trust Questions Q57-Q63]

Return your answers using the following JSON structure:

```
{
  "responses": {
    "generalized_trust": ,
    "trust_scores": {
      "family": ,
      "neighbors": ,
      "acquaintances": ,
      "strangers": ,
      "other_religions": ,
      "other_nationalities":
    }
  }
}
```

Figure 4.1: Illustrative prompt structure used for LLM response generation, demonstrating the mapping of demographic attributes to survey questions.

5. RESULTS

This chapter presents the empirical results addressing both research questions. Results are organized by research question, with Research Question 1 (representational fidelity) presented first, followed by Research Question 2 (comparative performance and feature importance).

5.1 Results for Research Question 1: Representational Fidelity

This section presents results evaluating whether LLMs can accurately simulate the responses of real individuals when prompted with specific demographic personas. The analysis examines both individual-level accuracy and aggregate-level distributional alignment across ten LLM architectures and seven trust-related questions from WVS.

5.1.1 *Individual-Level Accuracy Results*

In this part of the thesis, the individual-level accuracy of the LLM outputs is analyzed using both exact-match accuracy and mean absolute error.

5.1.1.1 Exact-Match Accuracy

In order to evaluate the exact-match accuracy, the predicted responses of each LLM for each demographic persona was compared with the actual response of the corresponding human, and the accuracy is aggregated on a per-question and per-model basis.

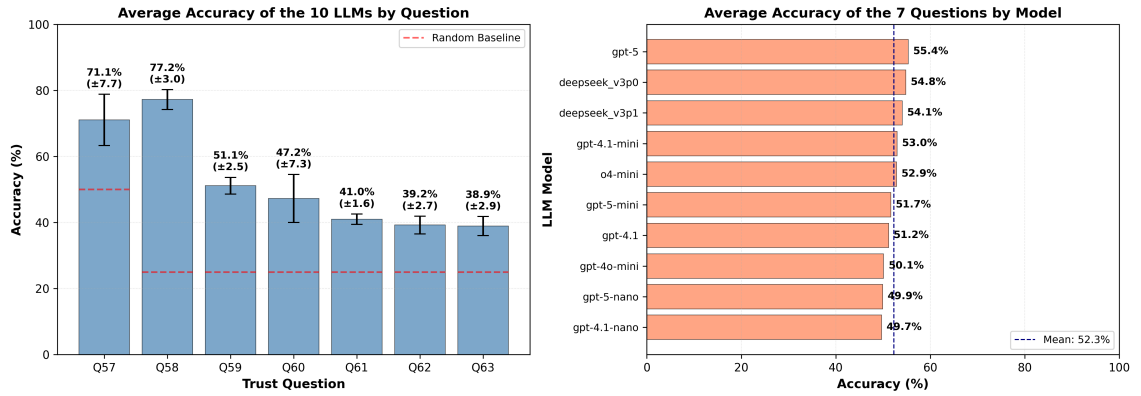


Figure 5.1: Exact-match accuracy analysis showing (left) accuracy by question with error bars and random baseline, and (right) overall accuracy by model across all questions. Random baselines: 50% for Q57 (binary), 25% for Q58-Q63 (4-point ordinal scales)¹.

Figure 5.1 illustrates the comparative performance of LLMs against random baselines. The results indicate a consistent performance advantage, with all models achieving accuracy rates well above chance levels. In terms of model-level stability (Figure 5.1, right), the variance across architectures is notably low. The tight distribution of accuracy scores suggests that predictive capability for this task is relatively model-agnostic, rather than being driven by specific high-parameter architectures or outliers.

A distinct pattern emerges when analyzing the average model accuracy across specific question types. Performance exhibits a monotonic decline corresponding to the social distance of the target group. Accuracy is highest for intimate groups (Q58: Family) and decreases progressively for distant out-groups (Q62: Other Religions; Q63: Other Nationalities). Furthermore, inter-model consistency varies by question; Q60 (Trust: People Known Personally) elicited the most heterogeneous responses among models, whereas

¹Note on aggregation: Performance metrics vary slightly between visualizations and tables due to averaging methods. Figure 5.1 reports *macro-averaged* accuracy (equal weight per question) to highlight variance across topics, while summary tables report *micro-averaged* accuracy (weighted by sample size) to reflect the total volume of successful predictions.

Q61 (Trust: Strangers) showed the highest consistency across architectures.

Binomial testing confirms the statistical significance of these margins. As detailed in Table 5.1, the null hypothesis—that models are operating at random chance—is rejected for all ten architectures ($p < 0.001$), confirming that the observed alignment is systematic.

Table 5.1: *Binomial test results comparing LLM accuracy to random baseline. P-values test whether accuracy significantly exceeds random guessing (one-tailed test). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.*

Model	Accuracy (%)	Weighted Random Baseline (%)	P-value
gpt-5	55.50	28.57	< 0.001***
deepseek_v3p0	54.97	28.57	< 0.001***
deepseek_v3p1	54.23	28.57	< 0.001***
gpt-4.1-mini	53.16	28.57	< 0.001***
o4-mini	53.00	28.57	< 0.001***
gpt-5-mini	51.77	28.57	< 0.001***
gpt-4.1	51.32	28.57	< 0.001***
gpt-4o-mini	50.27	28.57	< 0.001***
gpt-5-nano	50.03	28.57	< 0.001***
gpt-4.1-nano	49.86	28.57	< 0.001***

Table 5.2 provides the granular breakdown of these trends, detailing the performance variations across both model architectures and question types.

Table 5.2: *Exact match accuracy (%) by model and question. Models ordered by overall performance (ascending, left to right).*

Question	Model										Avg.
	4.1-nano	5-nano	4o-mini	4.1	5-mini	o4-mini	4.1-mini	DS-V3.1	DS-V3	5	
Q57	57.25	63.86	59.31	76.10	74.51	74.86	76.20	76.29	75.58	76.84	71.08
Q58	78.29	78.28	78.12	68.66	78.32	78.36	78.30	77.58	78.30	78.16	77.24
Q59	52.59	52.36	45.81	49.69	48.80	50.07	51.53	53.79	53.96	52.70	51.13
Q60	51.21	34.04	52.47	45.31	36.06	43.18	51.58	52.89	52.94	52.80	47.25
Q61	41.62	41.25	39.86	40.16	42.88	42.20	41.74	37.40	40.52	42.36	41.00
Q62	34.66	40.42	36.87	39.13	40.60	41.04	35.35	40.44	41.45	42.48	39.24
Q63	32.15	39.20	38.47	39.35	40.47	40.37	36.15	40.15	41.02	42.17	38.95

Note: Statistical analysis confirms that all model accuracies significantly exceed random baselines ($p < 0.001$ for all entries).

5.1.1.2 Mean Absolute Error (MAE) Analysis

While the exact-match accuracy metric evaluates binary success, the MAE analysis assesses the magnitude of alignment errors, accounting for the ordinal nature of the survey scales. This distinction is critical; a model predicting "2 - Trust somewhat" when the ground truth is "1 - Trust completely" represents a significantly smaller representational divergence than predicting "4 - Do not trust at all."

In order to capture this ordinal proximity, Table 5.3 presents MAE results averaged across all ten LLM models for each question (a), and across all seven questions for each model (b).

Table 5.3: *Mean Absolute Error (MAE) Analysis. Lower values indicate better performance.*

		By Model (7-Question Avg)	
By Question (10-Model Avg)		Model	MAE
Question	MAE		
		gpt-5	0.4966
Q57 (Gen. trust)	0.289	deepseek_v3p0	0.5140
Q58 (Family trust)	0.277	gpt-4.1-mini	0.5278
Q59 (Neigh. trust)	0.571	deepseek_v3p1	0.5355
Q60 (Acquaintances trust)	0.624	o4-mini	0.5430
Q61 (Strangers)	0.666	gpt-4.1	0.5482
Q62 (Religion)	0.686	gpt-4o-mini	0.5625
Q63 (Nationality)	0.707	gpt-5-mini	0.5679
		gpt-5-nano	0.5799
Average	0.546	gpt-4.1-nano	0.5814
		Average	0.546

Table 5.3 reveals a sharp contrast in task difficulty based on the target of trust. As shown in the question-level breakdown (Table 5.3, left), the models exhibit high precision for familial trust (Q58), with an error margin significantly lower than the global average. However, as the social distance of the target group increases—moving from neighbors to strangers and finally to out-groups (Q62, Q63)—the average error more than doubles.

The model-level analysis (Table 5.3, right) corroborates the findings from the accuracy section but reveals a nuanced performance hierarchy. Although the frontier model (GPT-5) achieves the lowest ordinal error, a linear relationship between model scale and

performance is not observed. For instance, the compact GPT-4.1-mini outperforms its larger counterpart GPT-4.1, and the older DeepSeek V3.0 slightly outperforms the newer V3.1. Consequently, it appears that ordinal precision is not strictly a function of model size or release date. Instead, the performance range remains relatively narrow across all architectures, suggesting that even smaller models maintain a coherent, if slightly less precise, representation of the persona’s trust.

5.1.2 Aggregate-Level Distributional Alignment

Beyond individual prediction accuracy, it is necessary to assess whether LLMs can reproduce the macroscopic statistical properties of human populations. The JS divergence metric quantifies the distance between the distribution of human responses and the distribution of model-generated responses; lower values would indicate that the model’s aggregate response distributions more closely resemble the diversity of public opinion.

Table 5.4: *Jensen-Shannon divergence analysis (lower values indicate better alignment).*

<i>(a) By Question</i>		<i>(b) By Model</i>	
Question	JS Divergence	Model	JS Divergence
Q57 (Generalized trust)	0.186	o4-mini	0.294
Q58 (Family trust)	0.261	gpt-4.1	0.300
Q59 (Neighborhood trust)	0.365	gpt-5-nano	0.306
Q60 (Acquaintances trust)	0.387	gpt-5-mini	0.309
Q61 (Strangers trust)	0.384	gpt-4o-mini	0.310
Q62 (Other religion)	0.380	deepseek_v3p1	0.338
Q63 (Other nationality)	0.346	gpt-5	0.346
Average	0.330	gpt-4.1-mini	0.357
		deepseek_v3p0	0.369
		gpt-4.1-nano	0.370
		Average	0.330

Table 5.4 shows a disparity in alignment based on question complexity and social intimacy. The models achieve their best distributional alignment on Q57 (Generalized Trust, 0.186), likely aided by its binary structure which simplifies the probability space. Among the ordinal questions, Q58 (Trust in Family) shows reasonable alignment. However, a distinct "alignment cliff" appears for all subsequent questions involving wider social circles (Q59–Q63), where divergence scores rise sharply to the 0.34–0.39 range. This indicates that while LLMs can approximate the aggregate distribution of trust for intimate relationships, they struggle to replicate the variance and distributional shape of human attitudes toward strangers and out-groups.

The model-level results (Table 5.4b) challenge the assumption that larger parameter counts² automatically yield better sociological simulation. The reasoning-optimized o4-mini achieves the lowest divergence, outperforming both the full-scale DeepSeek-V3.1 and the frontier GPT-5. Surprisingly, GPT-5 exhibits higher divergence scores than its smaller variants, such as GPT-5-mini and GPT-5-nano. This suggests that increased scale does not necessarily improve a model’s ability to capture answer distributions.

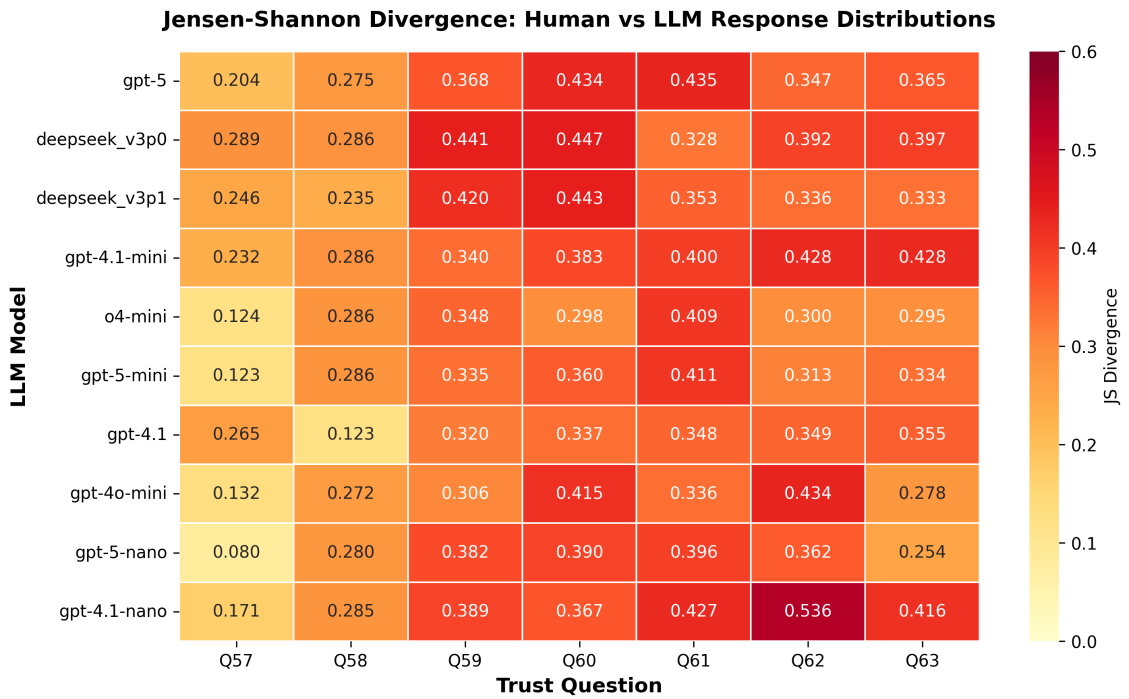


Figure 5.2: *Jensen-Shannon divergence between human and LLM response distributions across all model-question combinations. Darker colors indicate higher divergence (poorer alignment).*

Figure 5.2 provides a granular visualization of these trends. The heatmap reveals

²While parameter counts for the DeepSeek-V3 family are publicly disclosed (e.g., 671B parameters), the exact sizes of the OpenAI models remain proprietary. In accordance with OpenAI nomenclature [54, 55], GPT-5 and GPT-4.1 represent the full-scale flagship architectures, while ‘mini’ and ‘nano’ designations signify systematically reduced-parameter variants.

that the high divergence is not driven by isolated outliers but is a systemic issue across almost all architectures for questions Q59 through Q63 (indicated by the block of darker cells). Notably, even the best-performing models show a degradation in alignment (darkening color) as the target of trust shifts from "Family" (Q58) to "Strangers" (Q61), visually confirming that current LLMs struggle to model the distributional complexity of distant social relations.

Finally, the divergence analysis in Figure 5.2 highlights a critical trade-off: achieving high fidelity to human response distributions often requires sophisticated architectures like o4-mini, rather than simply scaling parameters (as seen with GPT-5's higher divergence). However, for practical sociological simulation, these fidelity gains must be weighed against computational costs, which vary by over two orders of magnitude across the model landscape.

Evaluating models solely on raw predictive accuracy per dollar would create a perverse incentive for the fixation strategies. As observed in earlier sections, smaller models such as GPT-4.1-nano often achieve respectable accuracy by collapsing their output onto the majority class, effectively gaming such a metric while failing to capture the diversity of human opinion. A robust efficiency metric must therefore penalize this behavior.

To address this, a Composite Efficiency Score is calculated to balance the predictive quality (Accuracy), distributional fidelity (penalized by JS Divergence), and economic feasibility (Log-Cost). Table 5.5 presents this balanced ranking.

Table 5.5: Relative Cost-Efficiency Analysis.

Model	Cost (10k)	Adj. Accuracy	Z-Quality	Z-Cost	Composite Score
	(\$)	$(Acc \times (1 - JS))$	(Std. Dev)	(Inv. Std. Dev)	$(0.6Z_Q + 0.4Z_C)$
o4-mini	25.50	0.374	1.50	-1.25	0.40
deepseek_v3p1	4.18	0.359	0.52	0.16	0.38
gpt-4o-mini	0.65	0.347	-0.28	1.15	0.29
deepseek_v3p0	1.17	0.347	-0.26	0.91	0.21
gpt-4.1	8.51	0.359	0.55	-0.36	0.19
gpt-5-mini	10.40	0.358	0.45	-0.52	0.06
gpt-4.1-mini	1.72	0.342	-0.59	0.71	-0.07
gpt-5-nano	6.34	0.347	-0.26	-0.14	-0.21
gpt-5	57.46	0.363	0.76	-1.93	-0.31
gpt-4.1-nano	0.42	0.314	-2.41	1.27	-0.94

Methodology: Metrics are standardized into Z-scores (σ from mean) to balance disparate units. **Z-Quality** favors high accuracy penalized by fixation ($Acc \times (1 - JS)$). **Z-Cost** favors low cost on a logarithmic scale ($-\ln(\text{Cost})$). The **Composite Score** ($0.6Z_Q + 0.4Z_C$) prioritizes performance outliers while maintaining sensitivity to relative cost efficiency.

By standardizing these disparate metrics onto a common Z-score scale, it is observed that while ultra-lightweight models like GPT-4.1-nano offer a good performance-per-dollar given their significantly low costs, they fall to the bottom of the efficiency ranking due to their poor distributional alignment. Conversely, models like o4-mini and DeepSeek-V3.1 emerge as top contenders, justifying their higher compute costs through superior simulation fidelity that relatively avoids the pitfalls of majority-class fixation.

This naturally leads to the question of why the models fail to distribute their answers effectively. The following section investigates the mechanism of this failure—response concentration—in detail.

5.1.2.1 Response Concentration Analysis

To investigate the structural causes of the distributional divergence observed in the aggregate-level distributional alignment section, the analysis was expanded to examine the specific shape of the response probability distributions. Figure 5.3 shows the distributions of each response in the human answer dataset.



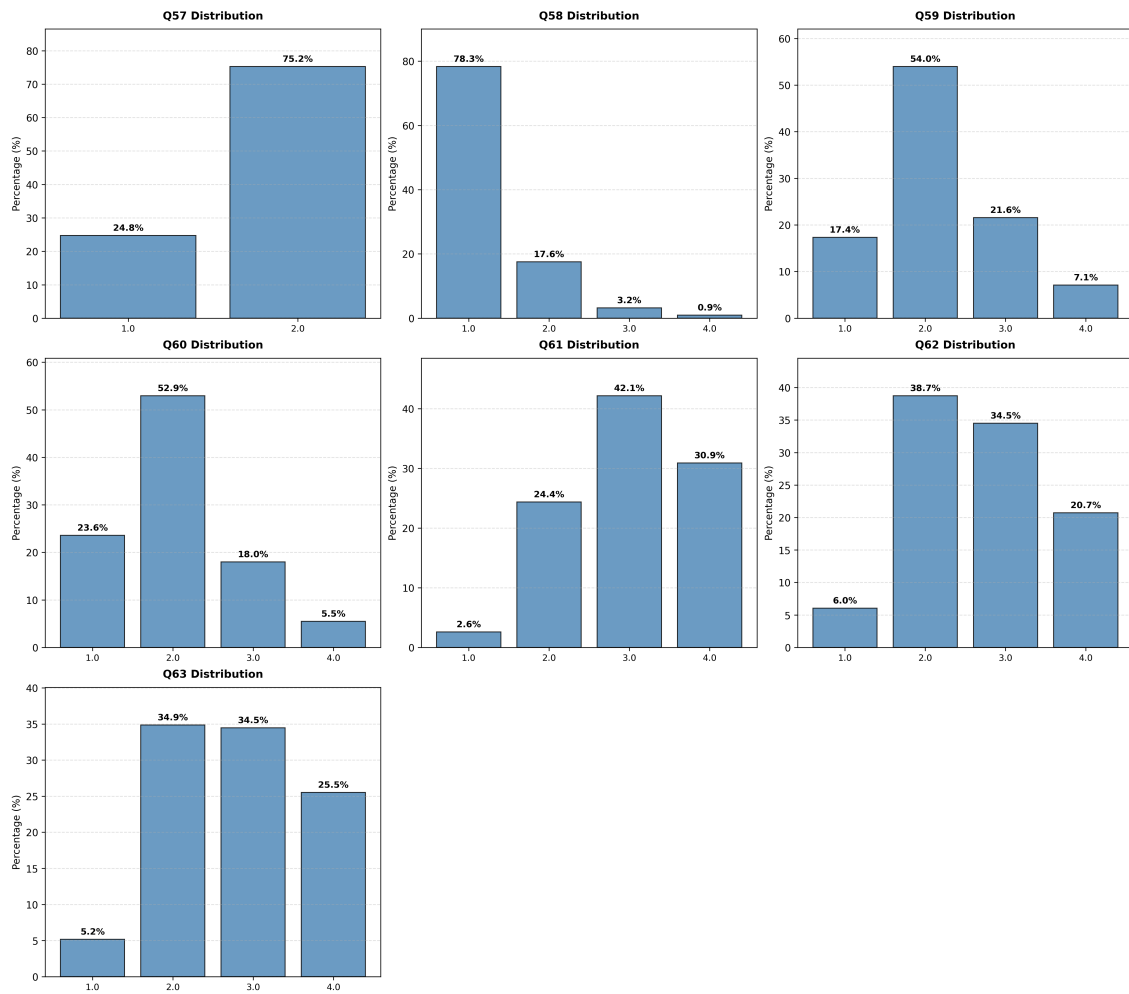


Figure 5.3: *Human Response Distributions.*

The human data exhibits a shrinking share of the majority class and increased response entropy as the target of trust shifts from family (Q58) to out-groups (Q62, Q63). This increasing distributional complexity may contribute to the performance degradation observed in earlier sections. Q61 (Trust in Strangers) presents a suitable test case; it maintains a distinct majority class, yet unlike Q58, it retains significant variance, with a non-negligible minority (2.6%) selecting "Complete Trust."

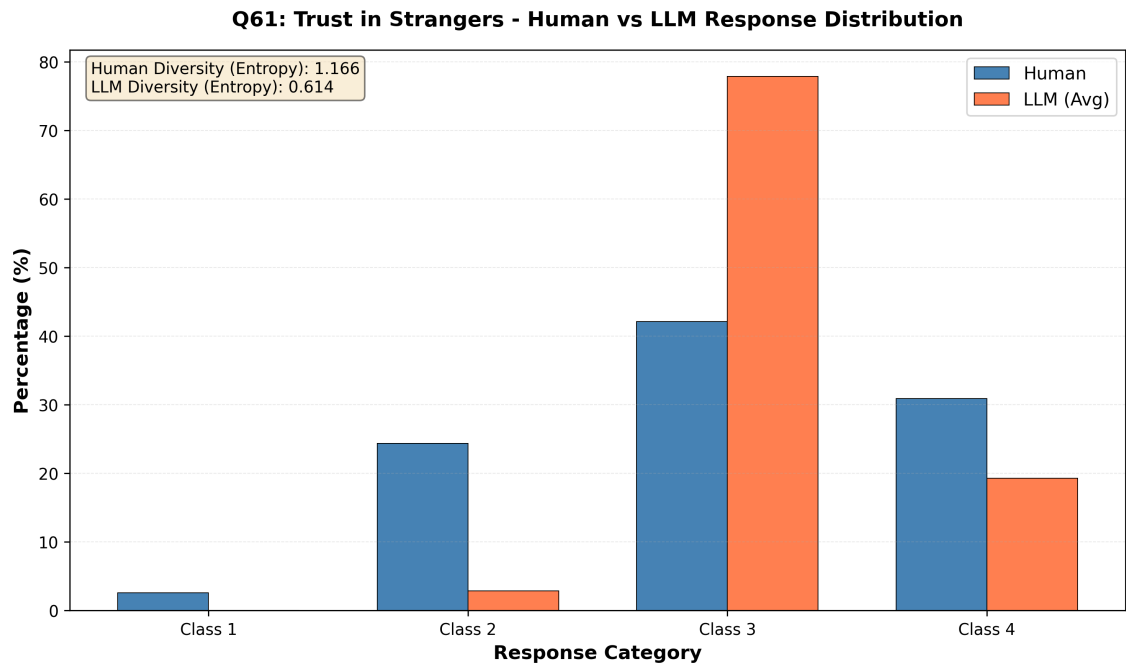


Figure 5.4: *Q61 - Human vs Average LLM Response Distributions.*

A comparison of the average LLM response to the human baseline for Q61 (Figure 5.4) reveals a systematic pattern of mode collapse. While human responses are distributed with moderate variance (Entropy: 1.166) across all four ordinal options, the average LLM-generated responses concentrate heavily on the majority category, resulting in significantly lower variance (Entropy: 0.614). Most notably, the minority class (Option 1) is effectively eliminated from the models' output distribution.

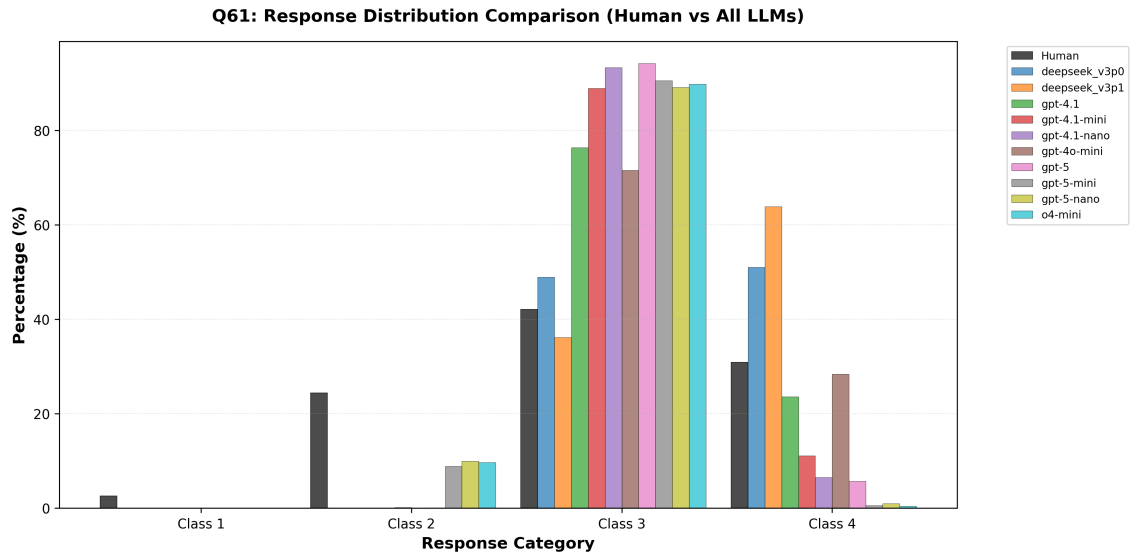


Figure 5.5: *Q61 - Detailed Response Distributions by LLM.*

Detailed by-model distributions in Figure 5.5 confirm the patterns observed in the averages. All models exhibit a strong affinity for the majority classes while failing to reproduce the tail of the distribution. This phenomenon is even more pronounced in Q58 (Trust in Family), which represents the scenario of highest human consensus.

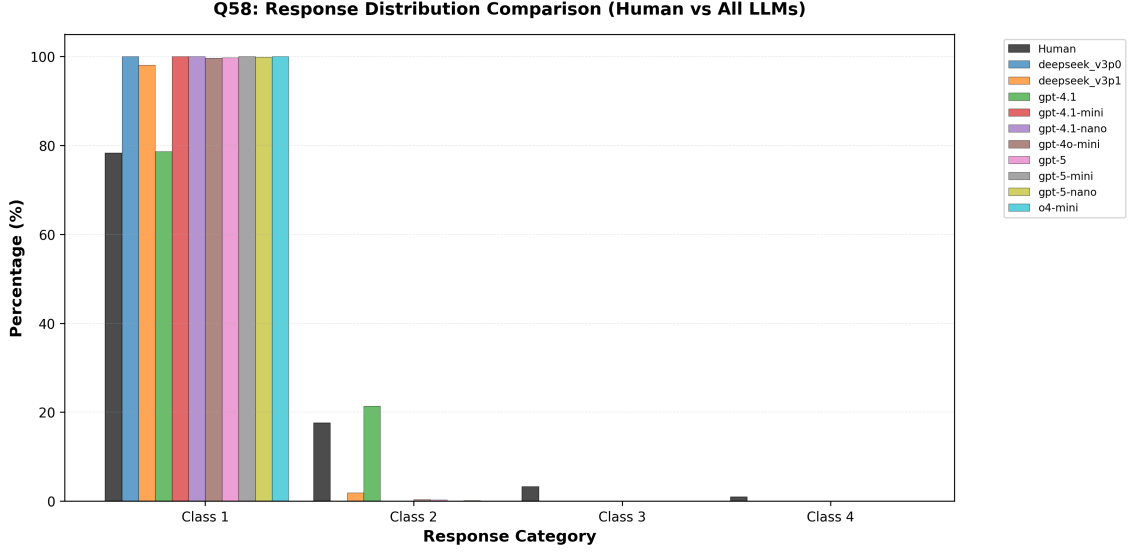


Figure 5.6: Q58 - Detailed Response Distributions by LLM.

As shown in Figure 5.6, models exhibit near-total fixation on the majority answer ("Trust Completely"). While human participants occasionally express distrust (Options 3 and 4), the models almost exclusively output the modal response, with fixation rates approaching 100% for several architectures.

To quantify this concentration, the Odds Ratio of the majority class was utilized. This metric was selected over simple relative frequency to avoid the "ceiling effect" inherent in high-consensus questions. For example, if human agreement is 85% and model fixation is 99%, a relative frequency metric would indicate only a marginal increase ($1.16\times$), obscuring the fact that the minority viewpoint has been effectively eliminated.

Following the methodology for comparing proportions in high-frequency events [56], the fixation odds were defined for both populations as the ratio of the majority class probability (p) to the remaining probability mass ($1 - p$):

$$\text{Odds} = \frac{p}{1 - p} \quad (5.1)$$

Based on (5.1), a Modal Concentration Factor (MCF) is introduced, which is defined as the ratio of the model’s peak odds to the human baseline odds. The metric is specifically configured to compare the odds of the model’s top choice to the odds of the human population’s top choice, regardless of whether they are the same option:

$$MCF = \frac{\text{Odds}(\max P_{LLM})}{\text{Odds}(\max P_{Human})} \quad (5.2)$$

The distinction of the max class in (5.2) is critical for cases of misalignment, such as Q62 (Trust in Other Religions). While human respondents were split (Max Majority: 39% for Class 2), the GPT-4.1-nano model fixated nearly 100% on a different answer (Class 3). A conventional majority-match accuracy metric would score this as ‘0% fixation’ (since it missed the human majority), whereas the MCF correctly identifies this as a near-deterministic failure, revealing that the model has collapsed to a single output despite the disagreement. An MCF value > 1 indicates that the model is more deterministic than the human population, while a value < 1 implies the model is more entropic.

Figure 5.7 presents the MCF scores across all question-model pairs. Note that due to the exponential nature of the odds ratio near 100% probability, a logarithmic scale is employed. The red dashed line (MCF=1.0) represents human-level fixation. Bars significantly above this line indicate mode collapse.

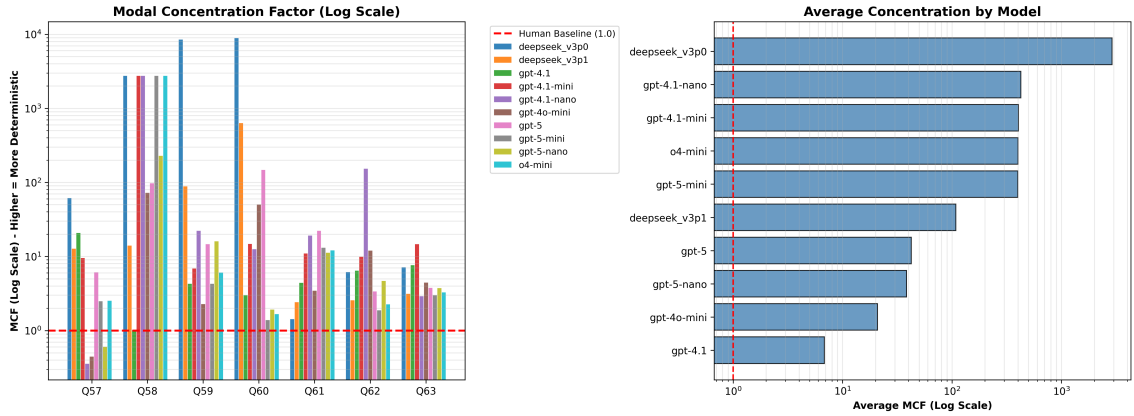


Figure 5.7: *Modal Concentration Factor (MCF) analysis. (Left) MCF by question; (Right) Average MCF by model.*

The results indicate a profound tendency toward determinism in specific contexts. For "High Consensus" questions like Q58, where human agreement is naturally high ($\sim 85\%$), the models frequently exhibit total mode collapse ($p > 99.9\%$). This results in MCF scores exceeding 100, effectively signaling that the minority opinion has been rendered statistically impossible. (Note: To prevent division by zero, probabilities were epsilon-clipped at 99.99%).

Figure 5.8 provides a granular view of the observed phenomenon. The heatmap reveals that mode collapse is not uniform. The "dark red" zones indicate that while models are relatively diverse on controversial topics (Q59, Q63), they become hyper-deterministic on questions regarding social cohesion (Q58, Q61).

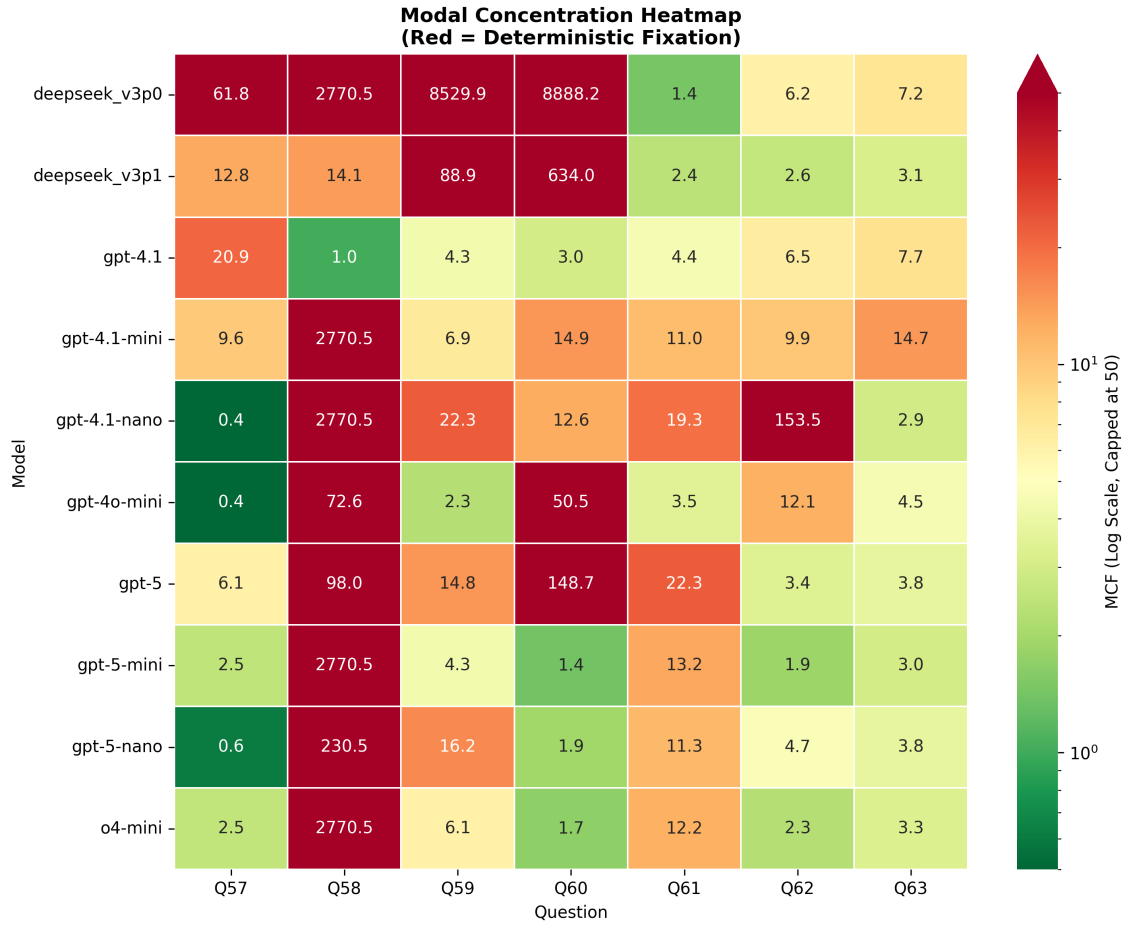


Figure 5.8: Heatmap of Modal Concentration Factors. Colors are mapped on a logarithmic scale capped at 50 to maintain contrast.

The analysis of Q62 demonstrates the utility of the refined MCF definition. Several smaller models (e.g., GPT-4.1-nano) exhibit extreme concentration scores ($> 10^3$) despite disagreeing with the human majority class. These results empirically differentiate alignment accuracy from distributional collapse, confirming that a model can be simultaneously misaligned and hyper-deterministic.

While the analysis documents a clear tendency toward aggregate-level determinism—most notably via modal fixation and entropy reduction—such findings do not necessarily imply uniformity across subgroups. A model might still exhibit nuanced variation conditional on demographic attributes, even if its overall distributions appear compressed.

To evaluate this possibility, the next subsection examines whether LLMs preserve the between-group response gradients observed in human data across gender, age, and country. Although this analysis was not part of the original methodology plan, it emerged as an ad hoc diagnostic during exploratory review of model outputs.

5.1.2.2 Demographic Subgroup Heterogeneity Analysis

Analysis of specific demographic patterns reveals whether LLMs reproduce the same associations observed in human data. For instance, if women express higher trust than men in human responses, does this pattern emerge in LLM responses? Table 5.6 presents gender differences in trust responses.

Table 5.6: Gender differences in trust responses: Human vs LLM. Values represent mean trust scores (1 = complete trust, 4 = no trust at all). Positive differences indicate women express higher trust than men.

Question	Human Responses			LLM Responses		
	Female	Male	Diff. (p-value)	Female	Male	Diff. (p-value)
Q57	1.76	1.75	+0.01 (0.428)	1.96	1.95	+0.01 (0.109)
Q58	1.28	1.27	+0.01 (0.124)	1.00	1.00	-0.00 (0.566)
Q59	2.21	2.18	+0.03 (0.031)*	2.05	2.05	-0.00 (0.637)
Q60	2.06	2.05	+0.01 (0.514)	2.01	2.01	+0.00 (0.646)
Q61	3.03	3.01	+0.02 (0.181)	3.07	3.06	+0.01 (0.003)**
Q62	2.72	2.70	+0.02 (0.282)	2.68	2.69	-0.00 (0.829)
Q63	2.82	2.80	+0.02 (0.165)	2.66	2.67	-0.01 (0.508)

* $p < 0.05$, ** $p < 0.01$. Pattern match indicates whether LLM reproduces the same directional pattern (women > men or men > women) as human responses.

Human responses show minimal but consistent gender differences: women express slightly higher trust than men across most questions, with differences ranging from 0.01 to 0.03 points on 4-point scales, although only Q59 shows a statistically significant gender difference in human responses ($p = 0.031$). LLM responses show mixed alignment: for Q57, Q60, and Q61, LLMs correctly reproduce the directional pattern (women > men), though with smaller magnitudes. However, for Q58, Q59, Q62, and Q63, LLMs fail to capture the human gender pattern, showing either no difference or opposite directions.

To evaluate other demographic sub-groups, Table 5.7 compares trust responses

between the youngest (18-30) and oldest (60+) age groups.

Table 5.7: Age differences in trust responses: Human vs LLM. Values represent mean trust scores for youngest (18-30) and oldest (60+) age groups. Negative differences indicate increasing trust with age, since lower numbers indicate higher reported trust levels.

Question	Human Responses			LLM Responses		
	18-30	60+	Diff.	18-30	60+	Diff.
Q57	1.77	1.69	-0.09	1.96	1.91	-0.04
Q58	1.28	1.23	-0.05	1.00	1.01	+0.00
Q59	2.20	2.00	-0.21	2.05	2.02	-0.03
Q60	2.10	1.91	-0.20	2.01	2.01	+0.00
Q61	3.05	2.88	-0.17	3.06	3.05	-0.00
Q62	2.76	2.53	-0.23	2.70	2.69	-0.01
Q63	2.88	2.60	-0.28	2.69	2.69	-0.00

Human responses consistently show increasing trust with age across all outgroup trust questions (Q62-Q63) and strangers trust (Q61), with differences ranging from 0.17 to 0.28 points. LLM responses capture this pattern directionally for Q61-Q63, but the magnitude is negligible (0.00 to 0.01 points). For example, on Q63, human responses show a 0.28-point increase in trust from youngest to oldest age groups, while LLM responses

show essentially no age difference (0.00 points).

Paired t-tests comparing age differences between human and LLM responses reveal that LLM age patterns differ significantly from human age patterns ($t(6) = 5.10$, $p = 0.002$, 95% Confidence Interval (CI): [0.085, 0.242], Cohen's $d = 2.73$). The mean difference in age effects (LLM - Human) is 0.164 points, indicating that LLMs systematically underweight age effects by approximately 0.16 points on average. This pattern is consistent across all seven questions: while LLMs capture some directional trend (younger individuals expressing less trust than older individuals for most questions), the magnitude of this age effect is dramatically compressed compared to human responses, confirming that LLMs severely underweight age-trust associations.

Country-level differences could reveal further insights. Table 5.8 compares trust levels across countries, showing the range between lowest and highest trust countries for human respondents and the range for between the same countries in LLM-generated responses.

Table 5.8: *Country-level variation in trust responses: Human extremes vs. corresponding LLM predictions. LLM columns show the predicted values for the specific countries identified as human extremes. Lower scores indicate higher trust levels.*

Question	Human Responses (Baselines)			Corresponding LLM Responses		
	Min. (Country)	Max. (Country)	Range	at Min.	at Max.	Range
Q57	China (1.69)	Indonesia (2.32)	0.63	China (2.00)	Indonesia (2.00)	0.00
Q58	China (1.15)	United States (1.47)	0.33	China (1.00)	United States (1.01)	0.01
Q59	China (1.99)	Indonesia (2.39)	0.40	China (2.05)	Indonesia (2.05)	0.00
Q60	United Kingdom (1.69)	Indonesia (2.72)	1.03	United Kingdom (2.01)	Indonesia (2.01)	0.00
Q61	Canada (2.88)	Indonesia (3.84)	0.96	Canada (3.06)	Indonesia (3.07)	0.01
Q62	United Kingdom (2.53)	China (3.55)	1.02	United Kingdom (2.70)	China (3.38)	0.68
Q63	United Kingdom (2.60)	Indonesia (3.77)	1.17	United Kingdom (2.69)	Indonesia (3.32)	0.63

Human responses show large cross-country variation: for Q57, trust levels range from 1.69 (China) to 2.32 (Indonesia), a 0.63-point difference. For Q60, the range is even larger (1.03 points between United Kingdom and Indonesia). LLM responses, however, show dramatically compressed variation: for Q57, LLM responses show essentially no country differences (0.00-point range), and only for Q62 and Q63 such differences are observed.

Paired t-tests comparing country-level variation (ranges between lowest and highest trust countries) between human and LLM responses reveal that LLM country-level variations differ significantly from human variations ($t(6) = -5.95$, $p = 0.001$, 95% CI: [-0.878, -0.366], Cohen's $d = -3.18$). The mean difference in country-level ranges (LLM - Human) is -0.62 points, indicating that LLMs systematically compress country-level variation by approximately 0.62 points on average. This analysis uses ranges (highest - lowest country mean) as the dependent variable, which is sensitive to outliers and uses only 7 data points (one per question), resulting in limited statistical power. A more robust approach would compare standard deviations of country means or use mixed-effects models accounting for question-level variation. However, the large effect size (Cohen's $d = -3.18$) and consistent pattern across all questions support the conclusion that LLMs fail to capture country-specific cultural patterns, instead generating responses that reflect global averages rather than localized trust norms.

The alignment analysis thus reveals a nuanced picture: LLMs appear to capture some broad demographic patterns directionally but systematically underweight their magnitude, particularly for age and country effects. This indicates that while LLMs encode demographic-opinion associations, they fail to capture the full complexity and context-dependency of these relationships that are critical for understanding opinion formation in diverse populations.

The findings of compressed demographic variation and modal fixation across questions raise a deeper concern: whether LLMs are truly simulating survey reasoning based on personas, or merely producing low-entropy outputs through quasi-random processes. Despite their apparent determinism at the aggregate level, it remains unclear whether LLMs are genuinely attending to demographic cues, or simply defaulting to frequent patterns with little conditional awareness. To address this concern, a supplementary control experiment was conducted to directly test the models' capacity for generating random outputs and to evaluate whether the observed behavior reflects stochastic guessing or structured response generation.

5.1.2.3 Supplementary Randomness Control

During the course of this study, a key methodological concern emerged regarding the nature of LLM-generated responses: are these outputs truly the result of structured persona-driven reasoning, or are they functionally equivalent to random guesses—whether due to model limitations or insufficient prompt constraints? This concern was not part of the original pre-analysis methodology but arose inductively during exploratory analysis. In particular, the severe attenuation of demographic variation and the recurrence of modal fixation across unrelated personas prompted a deeper examination of whether the models were attending to the persona information at all, or simply defaulting to frequent outputs regardless of input.

To determine if LLMs were genuinely attending to persona constraints—or simply defaulting to high-frequency tokens regardless of context—a three-stage control experiment was conducted to assess the randomness capability of LLMs under varying levels of prompt structure. Using a randomly selected subsample of 1,000 demographic prompts, five models (GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o-mini, and o4-mini) were tasked with generating a "random number between 1 and 4" under three varying levels of

prompt structure:

1. Pure Random Choice: The model was simply asked to “give a random number between 1 and 4” with no persona or contextual information.
2. Persona + Random Choice: The same persona descriptions used in the main experiment were provided, followed by a request: “Now, give a random number between 1 and 4.”
3. Persona + Survey + Random Choice: The model received both the full persona and the original survey prompt, followed by the same random number request.

The experiment reveals that no model successfully approximated uniform randomness (Theoretical max entropy $H = 2.0$), with all models exhibiting extreme χ^2 values and entropy levels significantly below the uniform baseline. As illustrated in Figure 5.9, GPT-4.1 and GPT-4.1-mini showed the most severe fixation, with entropies as low as 0.094.

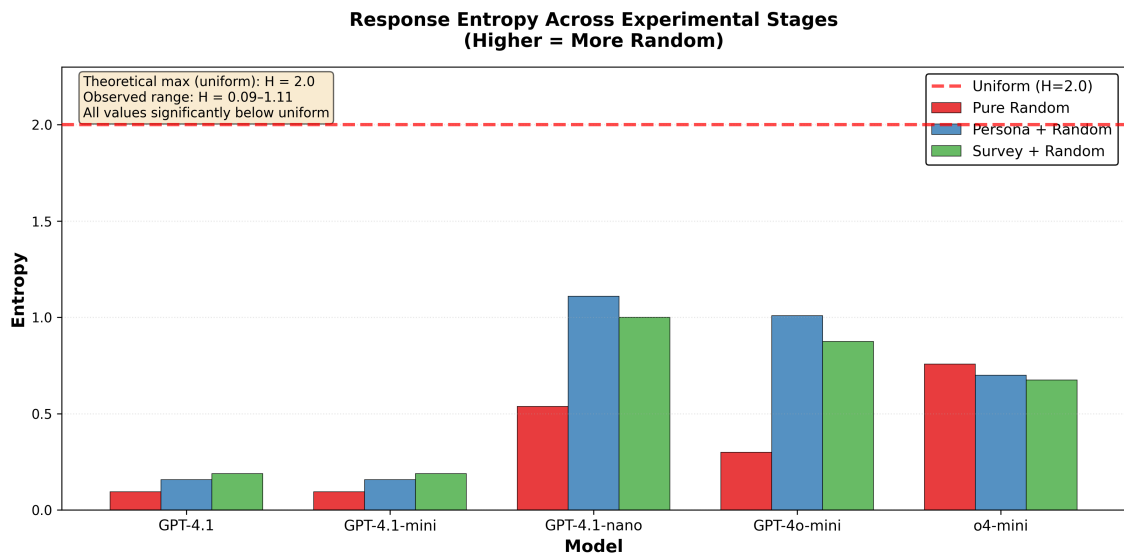


Figure 5.9: *Response Entropy Across Experimental Stages.*

All models remain significantly below the theoretical maximum for uniform randomness ($H = 2.0$), indicating a failure to produce stochastic noise. GPT-4.1 and its mini variant show the most extreme variance collapse.

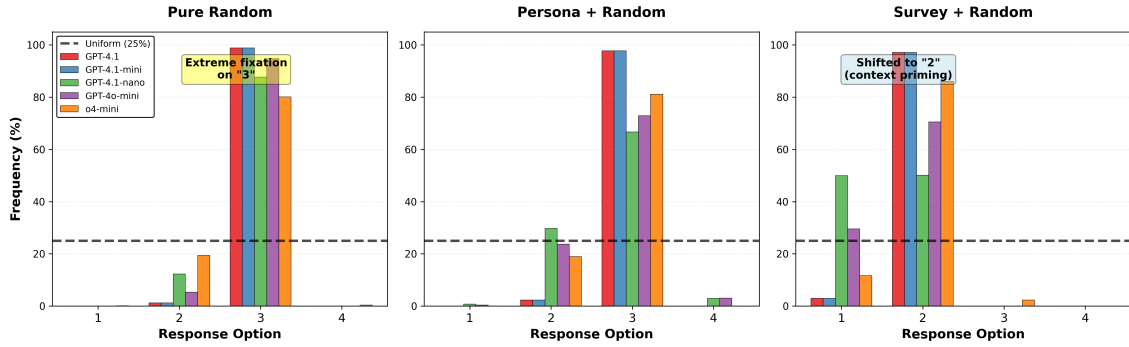


Figure 5.10: *Frequency Distribution of "Random" Choices across three experimental stages.*

In the Pure Random and Persona + Random stages, models exhibited an overwhelming fixation on response option "3". For GPT-4.1 and GPT-4.1-mini, 98.8% of "random" choices were identical (3), suggesting a deep-seated architectural or training bias toward specific tokens when prompted for randomness. In the Survey + Random stage, a dramatic shift occurred. The modal fixation shifted from "3" to "2". For the GPT-4.1 variants, 97.1% of responses moved to option 2, while the nano model split its responses almost perfectly between 1 and 2 (49.9% vs 50.1%).

Table 5.9: *Randomness control experiment: Summary statistics for three stages of prompt structure. Values represent response frequency (%), Chi-square (χ^2) goodness-of-fit against a uniform distribution, and Shannon Entropy (H). Theoretical maximum entropy for uniform randomness is $H = 2.0$.*

Model	N	%1	%2	%3	%4	χ^2	Entropy (H)
Stage 1: Pure Random							
GPT-4.1	1000	0.0	1.2	98.8	0.0	2905.2	0.094
GPT-4.1-mini	1000	0.0	1.2	98.8	0.0	2905.2	0.094
GPT-4.1-nano	1000	0.0	12.3	87.7	0.0	2137.0	0.538
GPT-4o-mini	1000	0.0	5.3	94.7	0.0	2598.5	0.299
o4-mini	1000	0.1	19.4	80.1	0.4	1717.0	0.757
Stage 2: Persona + Random							
GPT-4.1	1000	0.0	2.3	97.7	0.0	2820.2	0.158
GPT-4.1-mini	1000	0.0	2.3	97.7	0.0	2820.2	0.158
GPT-4.1-nano	992	0.7	29.7	66.6	2.9	1124.3	1.110
GPT-4o-mini	1000	0.4	23.7	72.9	3.0	1354.1	1.008
o4-mini	999	0.0	18.9	81.1	0.0	1771.1	0.700
Stage 3: Survey + Random							
GPT-4.1	1000	2.9	97.1	0.0	0.0	2774.7	0.189
GPT-4.1-mini	1000	2.9	97.1	0.0	0.0	2774.7	0.189
GPT-4.1-nano	1000	49.9	50.1	0.0	0.0	1000.0	1.000
GPT-4o-mini	1000	29.5	70.5	0.0	0.0	1336.2	0.875
o4-mini	772	11.7	86.0	2.3	0.0	1556.1	0.675

These findings suggest that while LLMs fail to generate true uniform randomness, their outputs are highly sensitive to "contextual gravity"—the structural influence of

prompt inputs that pulls model probability mass toward specific tokens. Importantly, the shift in modal response from "3" to "2" when the survey questions are introduced suggests that models are not producing stochastic noise, but are rather exhibiting structured, albeit biased, associations. Having established the structural consistency and inherent biases of these synthetic response patterns, the research moves on to Research Question 2.

5.2 Results for Research Question 2: Comparative Performance and Feature Importance

This section presents the results for Research Question 2, which examines whether LLM-generated survey responses can match the predictive power of traditional machine learning models trained on human responses, and whether they rely on the same demographic signals.

5.2.1 Predictive Performance Evaluation

To benchmark LLM performance, five traditional machine learning models (Logistic Regression, Decision Tree, Random Forest, LightGBM, and XGBoost) were trained on human survey responses using the identical 121-feature demographic set as described in detail in the Methodology chapter. The selected ML models were evaluated using 5-fold stratified cross-validation. Since model performances of the ML models will later be compared against LLM-generated responses, it is essential to verify the reliability of the cross-validated accuracy estimates. Examining the standard deviation across folds helps assess the stability of these results and ensures that reported accuracies reflect consistent predictive patterns rather than artifacts of a particular train-validation split.

Table 5.10 presents detailed cross-validation results including standard deviations across folds. Standard deviation quantifies model stability: lower values indicate consistent performance across different data partitions, while higher values suggest sensitivity to training composition.

Table 5.10: *5-fold cross-validation results for ML models trained on human responses. Validation accuracy and standard deviation shown for each model, averaged across all seven questions.*

Model	Validation Acc. (%)	Std. Dev. (%)
XGBoost	58.31	0.70
LightGBM	47.04	0.85
RandomForest	45.83	1.02
LogisticRegression	41.69	1.12
DecisionTree	39.36	2.87

XGBoost exhibits the lowest standard deviation (0.7%), indicating highly stable predictions across folds despite being the most complex model. This stability, combined with highest accuracy, confirms robust generalization rather than overfitting to specific training samples.

Decision Trees show the highest standard deviation (2.87%), however, given their well-known sensitivity to training data composition due to greedy splitting rules, the model still achieves a respectably low range of deviation.

Overall, the cross-validation results confirm that reported accuracies represent genuine generalization performance, with XGBoost demonstrating both the highest accuracy and the most reliable consistency across data partitions.

Having established the consistency and generalization capability of the ML models through cross-validation, the next step is to evaluate these performances by benchmarking them against naïve heuristics such as random guessing and majority-class predictions. This comparison helps clarify whether the models are capturing meaningful

demographic patterns or merely capitalizing on dataset imbalances. Table 5.11 demonstrates the results along with these baselines for comparison.

Table 5.11: *Majority-class and chance-level baseline comparisons. Majority baseline represents accuracy from always predicting the most frequent class.*

Question	Majority (%)	Random (%)	XGB (%)	LGB (%)	RF (%)	LR (%)	DT (%)
Q57	75.2	50.0	77.8	68.8	69.8	66.5	66.8
Q58	78.3	25.0	78.3	57.0	57.3	42.3	39.8
Q59	54.0	25.0	54.5	39.4	38.5	36.9	29.8
Q60	52.9	25.0	54.3	40.5	35.8	36.2	32.9
Q61	42.1	25.0	48.2	44.6	42.7	38.6	39.4
Q62	38.7	25.0	48.0	42.0	38.6	36.6	34.9
Q63	34.9	25.0	48.1	43.4	41.7	36.6	38.7
Average	53.7	28.6	58.5	47.9	46.3	41.9	40.3

As observed in the table above, every model substantially exceeds the random baselines for each question, confirming that all ML architectures extract non-random signals from the demographic features. However, the majority-class baseline (53.7% average) proves to be a more challenging threshold. Only XGBoost consistently outperforms this baseline across the entire set of questions, achieving an average predictive lift of 4.8 percentage points over the naive majority heuristic.

In questions regarding intimate trust (Family and Generalized Trust), the majority-class frequencies are extremely high (75.2% and 78.3%). In this domain, most models fail to provide added value. For Q58, XGBoost exactly matches the majority baseline (78.3%), suggesting that even the most advanced model struggles to find a demographic

signal strong enough to overcome the dominant class distribution. In these cases, simpler ML models actually perform below the majority baseline, indicating that a naive "always trust" heuristic is more accurate than their complex predictions.

As social distance increases and the target of trust becomes more distant (e.g., strangers or people of other nationalities), the majority-class frequency drops significantly (from 42.1% to 34.9%). It is in this domain that the machine learning models demonstrate their greatest utility. For Q63, every model—from XGBoost (+13.2%) to Logistic Regression (+1.7%), successfully outperforms the majority baseline.

When it comes to the performance comparison with LLMs, Table 5.12 presents the comparative accuracy results for both cross-ML model evaluation and performance comparison with LLMs.

Table 5.12: *Predictive accuracy comparison: five machine learning models trained on human responses (5-fold stratified cross-validation), compared to the average and best-performing LLMs for each question. Accuracy values represent validation performance averaged across folds.*

Question	XGBoost (%)	LightGBM (%)	RF (%)	LR (%)	DT (%)	LLM Avg. (%)	Best LLM (%)	Best LLM
Q57	77.8	68.8	69.8	66.5	66.8	71.1	76.8	GPT-5
Q58	78.3	57.0	57.3	42.3	39.8	77.2	78.4	GPT-4o-mini
Q59	54.5	39.4	38.5	36.9	29.8	51.1	54.0	GPT-5-mini
Q60	54.3	40.5	35.8	36.2	32.9	47.2	52.9	GPT-5
Q61	48.2	44.6	42.7	38.6	39.4	41.0	42.9	GPT-4.1
Q62	48.0	42.0	38.6	36.6	34.9	39.2	42.5	GPT-4.1-mini
Q63	48.1	43.4	41.7	36.6	38.7	38.9	42.2	Deepseek V3.1
Average	58.5	48.0	46.3	41.9	40.3	52.3	55.7	-

Within the traditional supervised learning benchmarks, a clear performance hierarchy emerges. Gradient boosting methods, specifically XGBoost, demonstrate superior

predictive capability compared to other ML architectures. XGBoost achieves the highest accuracy across every single question, maintaining a substantial lead over Random Forests and LightGBM. The dominance suggests that the relationship between demographic features and trust attitudes involves complex, non-linear interactions that are best captured by boosting algorithms rather than simple decision rules or linear coefficients. Notably, single Decision Trees consistently underperform, indicating that demographic data is too high-dimensional and sparse for un-ensembled partitioning to yield robust generalization.

Another observation is that the prediction performance of the ML models also deteriorate for out-group trust level predictions, as is consistent with the LLM results previous section. Even the best-performing model, XGBoost, experiences a substantial performance drop—exceeding 30 percentage points—when moving from in-group trust (Q58, Family: 78.3%) to out-group categories such as Q61 (Strangers), Q62 (Other Religions), and Q63 (Other Nationalities), where accuracy levels cluster around 48%. This pattern is remarkably consistent across all ML architectures, where predictive success peaks for primary social circles (Q57, Q58) and reaches its lowest point for broader, more heterogeneous out-groups (Q61–Q63).

Comparing these results with the LLM performances, while LLMs achieve respectable zero-shot performance, they do not surpass optimized supervised learning. XGBoost significantly outperforms the LLM average ($p = 0.05$), and even the "Best-in-Class" LLM for each question typically falls short of the gradient-boosted baseline. However, both the Best LLM by question, and the average LLM response accuracy surpasses every other selected model's average accuracy.

Figure 5.11 visualizes this trend, highlighting the consistency of XGBoost's superiority across the "distant trust" questions, contrasted with the parity observed in the "intimate trust" domain. XGBoost achieves superior accuracy on 6 of 7 questions, with largest advantages on distant-relationship questions (Q59-Q63).

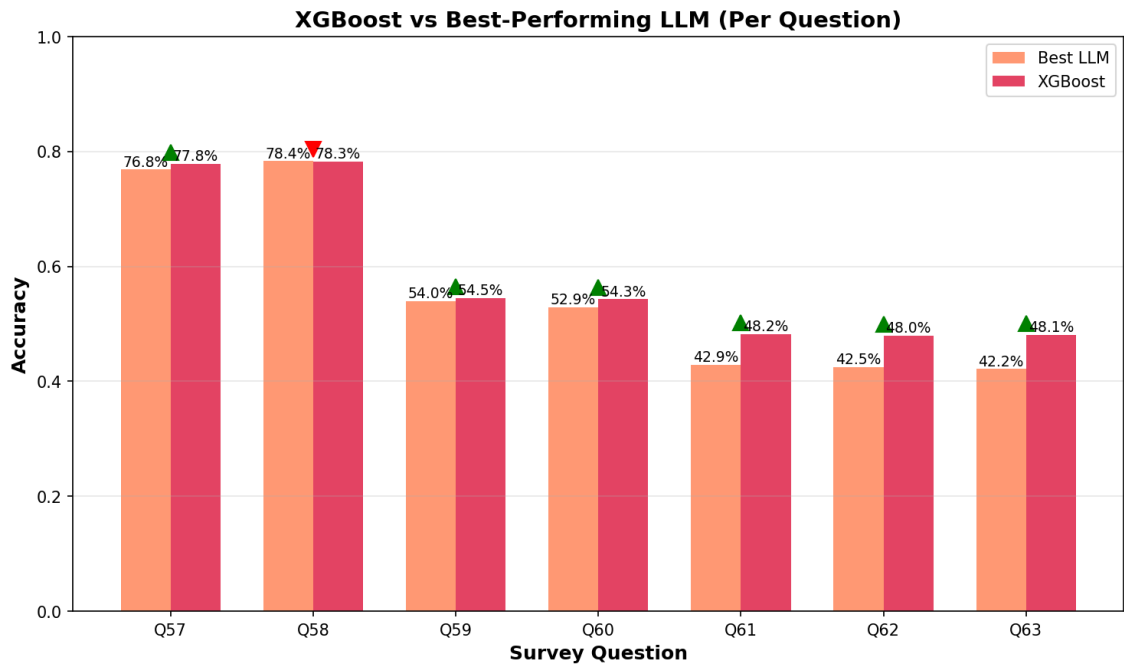


Figure 5.11: Performance comparison: XGBoost vs. best individual LLM vs. LLM average across seven questions.

To formally evaluate the performance differences between the LLM simulation and supervised ML models, a series of paired *t*-tests was conducted across the seven trust questions. This analysis, summarized in Table 5.13, assesses whether the observed accuracy gaps are statistically significant or attributable to question-specific variance.

Table 5.13: Paired *t*-test results comparing ML model accuracy against LLM average accuracy across seven questions. Positive mean difference indicates LLM outperforms ML; negative indicates ML outperforms LLM. Cohen’s *d* provides standardized effect size: small (0.2), medium (0.5), large (0.8).

Model	Mean Diff. (pp)	t-statistic	p-value	95% CI	Cohen’s <i>d</i>
LogisticRegression	+10.33	2.31	0.060	[-0.62, 21.28]	0.94
DecisionTree	+11.94	2.33	0.059	[-0.61, 24.50]	0.95
RandomForest	+5.95	1.81	0.121	[-2.11, 14.00]	0.74
LightGBM	+4.32	1.24	0.261	[-4.19, 12.83]	0.51
XGBoost	-6.19	-5.57	0.001**	[-8.91, -3.47]	-2.27

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Negative Cohen’s *d* indicates LLM outperformance; positive indicates ML outperformance.

The results reveal that XGBoost is the only model to demonstrate a statistically significant performance advantage over the LLM average ($p = 0.001$). The magnitude of this difference is characterized by a very large effect size (Cohen’s $d = -2.27$), and the 95% confidence interval $[-8.91, -3.47]$ strictly excludes zero, confirming that XGBoost’s superiority is robust across different trust dimensions. In contrast, the comparisons for all other ML architectures failed to reach statistical significance at the $\alpha = 0.05$ level.

Despite having positive mean differences (indicating higher raw accuracy for the ML models), the p -values for Logistic Regression ($p = 0.060$) and Decision Trees ($p = 0.059$) sit just above the threshold for significance. The wide confidence intervals for these models, which cross zero, suggest that while they may outperform LLMs on specific questions, they do not do so consistently across the entire survey. Interestingly,

Logistic Regression and Decision Trees show "large" effect sizes ($d \approx 0.95$), yet remain non-significant. This discrepancy is likely due to the small sample size ($n = 7$ questions), which limits statistical power. However, it also demonstrates that LLMs are performing at a level where they are statistically indistinguishable from several traditional supervised learners. The ensemble methods, Random Forest ($d = 0.74$) and LightGBM ($d = 0.51$), show the smallest performance gaps and the highest p -values when compared to LLMs. This suggests a high degree of performance parity, further reinforcing the finding that state-of-the-art LLMs can approximate the predictive accuracy of standard ensemble-based supervised learning.

Having established the accuracy-based comparison between the performances of LLMs and ML models, the analysis now shifts to a complementary metric, MAE, that better reflects the ordinal structure of the survey responses. MAE provides a more nuanced understanding of how far off predictions are, even when they are not exactly correct. Table 5.14 presents detailed MAE results for both ML models and the average LLM performance.

Table 5.14: MAE comparison. Lower values indicate better ordinal proximity. For binary Q57, MAE ranges 0-1; for ordinal Q58-Q63, MAE ranges 0-3.

Question	LLM Avg.	XGBoost	LightGBM	RF	LR	DT
Q57	0.289	0.222	0.312	0.302	0.335	0.332
Q58	0.277	0.265	0.587	0.598	0.979	0.968
Q59	0.571	0.521	0.801	0.855	0.877	0.937
Q60	0.624	0.514	0.819	0.946	0.939	1.027
Q61	0.666	0.608	0.745	0.813	0.922	0.881
Q62	0.686	0.620	0.787	0.848	0.912	0.939
Q63	0.707	0.631	0.760	0.793	0.906	0.858
Average	0.546	0.483	0.687	0.737	0.838	0.849

Based on the MAE calculations, the average MAE of the LLM predictions is observed to be lower than all ML models except XGBoost, indicating that when LLMs make errors, they tend to be "close but wrong" (predicting adjacent categories) rather than completely off-target. For instance, on Q61 (trust in strangers), the LLM average MAE is 0.666, meaning that the average prediction error is less than one scale point on the 1-4 ordinal scale. This ordinal proximity suggests LLMs may be suitable for applications where capturing directional trends matters more than exact-match accuracy, such as ordinal regression tasks or aggregate-level distributional analyses.

The MAE pattern mirrors the accuracy pattern: close-relationship questions (Q57-Q58) show low MAE for all models (0.22-0.34), while distant-relationship questions (Q61-Q63) show higher MAE (0.61-0.94), confirming that prediction difficulty increases with social distance. Interestingly, even on high-difficulty questions, LLMs maintain better ordinal proximity than all ML models except XGBoost, supporting their utility as

approximate survey respondents when exact-match accuracy is not critical.

5.2.2 Feature Importance Analysis

This section reports the feature importance results for models trained on human responses versus models trained on LLM-generated responses. With these results, it is aimed to assess whether both model types rely on similar demographic signals when making predictions, using the same 121-feature input set.

Feature importance alignment was quantified using Spearman rank correlation (ρ) between feature importances from models trained on human responses versus LLM responses. Table 5.15 presents correlations for all five model architectures across seven questions, with statistical significance markers.

Table 5.15: *Feature importance alignment: Spearman rank correlation (ρ) between feature importances from models trained on human responses versus LLM responses. Higher values indicate stronger alignment.*

Question	LR	DT	RF	LGB	XGB	Average
Q57	0.422***	0.547***	0.785***	0.568***	0.345***	0.533
Q58	-0.120	0.375***	0.686***	0.538***	0.105	0.317
Q59	0.019	0.508***	0.622***	0.589***	0.080	0.364
Q60	0.174	0.428***	0.690***	0.580***	0.066	0.388
Q61	0.294**	0.428***	0.783***	0.693***	0.108	0.461
Q62	0.242**	0.519***	0.827***	0.779***	0.338***	0.541
Q63	0.279**	0.504***	0.787***	0.769***	0.401***	0.548
Average	0.187	0.473	0.740	0.645	0.206	0.450

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results reveal that alignment is highly dependent on model architecture. Random Forest (RF) and LightGBM (LGB) exhibit the strongest and most consistent alignment, with average correlations of $\rho = 0.740$ and $\rho = 0.645$, respectively. In contrast, while XGBoost is an ensemble method, it demonstrates much lower alignment ($\rho = 0.206$), performing similarly to the linear Logistic Regression ($\rho = 0.187$). This divergence among ensemble methods is notable; it suggests that while Random Forest and LightGBM capture broad, overlapping feature signals between the two datasets, the specific optimization path of XGBoost—which prioritizes minimizing residual errors in a more aggressive, non-linear fashion—yields over-specialized feature sets for human versus LLM data. The moderate alignment of Decision Trees ($\rho = 0.473$) further indicates that ensemble averaging (in RF and LGB) is necessary to smooth out the noise and reveal the shared underlying demographic logic between human and LLM responses.

The analysis also identifies a clear trend related to the social distance of the trust target, where Q58 shows the lowest aggregate alignment ($\rho = 0.317$). As documented in the performance analysis, the extreme majority-class fixation in LLM responses (99.9% Class 1) for family trust collapses the variance required for stable feature importance extraction, leading to unreliable or non-significant correlations. Meanwhile, out-group trust (Q62–Q63) shows the highest alignment ($\rho \approx 0.54$).

Statistical significance tests confirm that correlations are significant ($p < 0.05$) for the majority of tree-based models across nearly all questions. However, Logistic Regression fails to show significant alignment for intimate trust (Q58) and close-group trust (Q59, Q60). Similarly, XGBoost correlations are non-significant for four out of the seven questions (Q58–Q61).

Figure 5.12 visualizes these correlation patterns through scatter plots comparing the feature importances of the ML models trained on human versus LLM responses across all features and questions. Each panel shows one model architecture, with points colored

by question. Points near the diagonal indicate strong alignment; scatter indicates divergence.

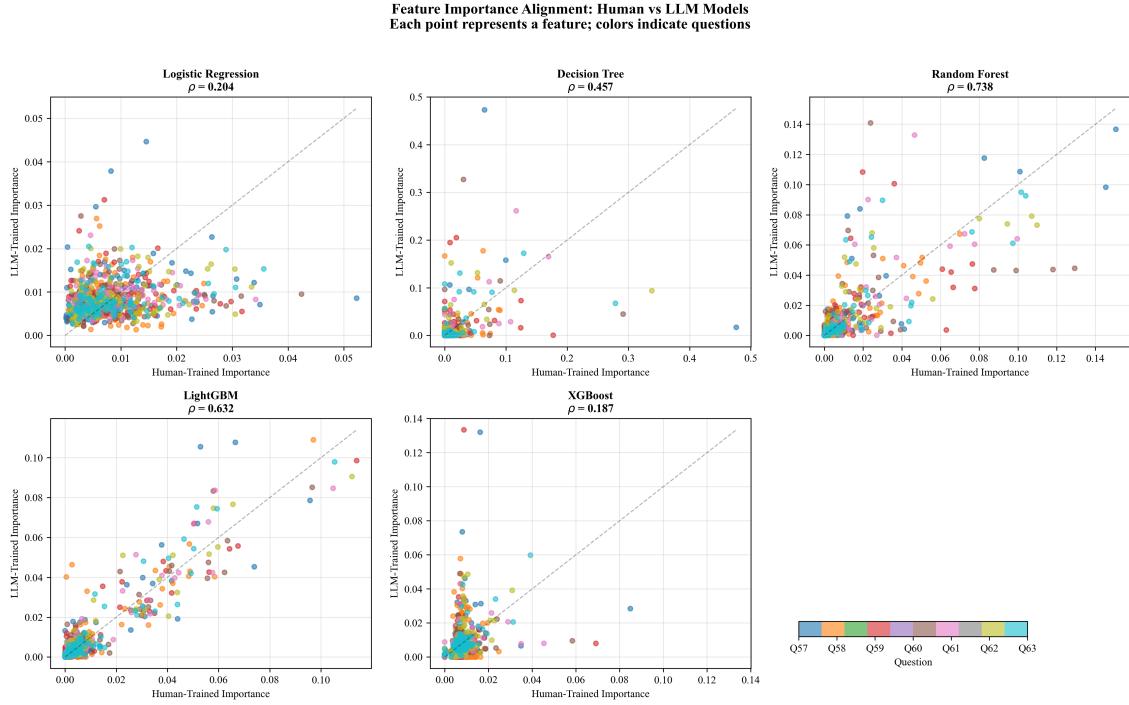


Figure 5.12: Feature importance alignment: scatter plots comparing human vs LLM-response trained models. Spearman correlations shown in panel titles.

While correlation coefficients capture overall similarity in feature importance rankings, they do not indicate whether models prioritize the same features most strongly. To address this, Table 5.16 shows the overlap in top-10 most important features between models trained on human responses and models trained on LLM-generated responses across all questions and architectures.

Table 5.16: *Top-10 feature overlap between human vs LLM-response trained models.*

Values represent number of features (out of 10) appearing in both models' top-10 lists.

Question	LR	DT	RF	LGB	XGB
Q57	1	6	7	8	4
Q58	0	5	7	8	0
Q59	1	3	6	8	1
Q60	0	6	7	8	1
Q61	1	5	7	8	4
Q62	3	5	6	8	3
Q63	2	7	6	9	6
Average	1.1	5.3	6.6	8.1	2.7

LightGBM is observed to achieve the highest stability with an average overlap of 8.1 out of 10 features, achieving near-total alignment for trust in people of other nationalities (Q63). Random Forest also maintains relatively high consistency with a mean overlap of 6.6 features. In contrast, Logistic Regression and XGBoost exhibit remarkably low convergence, averaging only 1.1 and 2.7 overlapping features, respectively.

Figure 5.13 illustrates these patterns through four strategically selected comparisons focusing on models with demonstrated predictive accuracy.

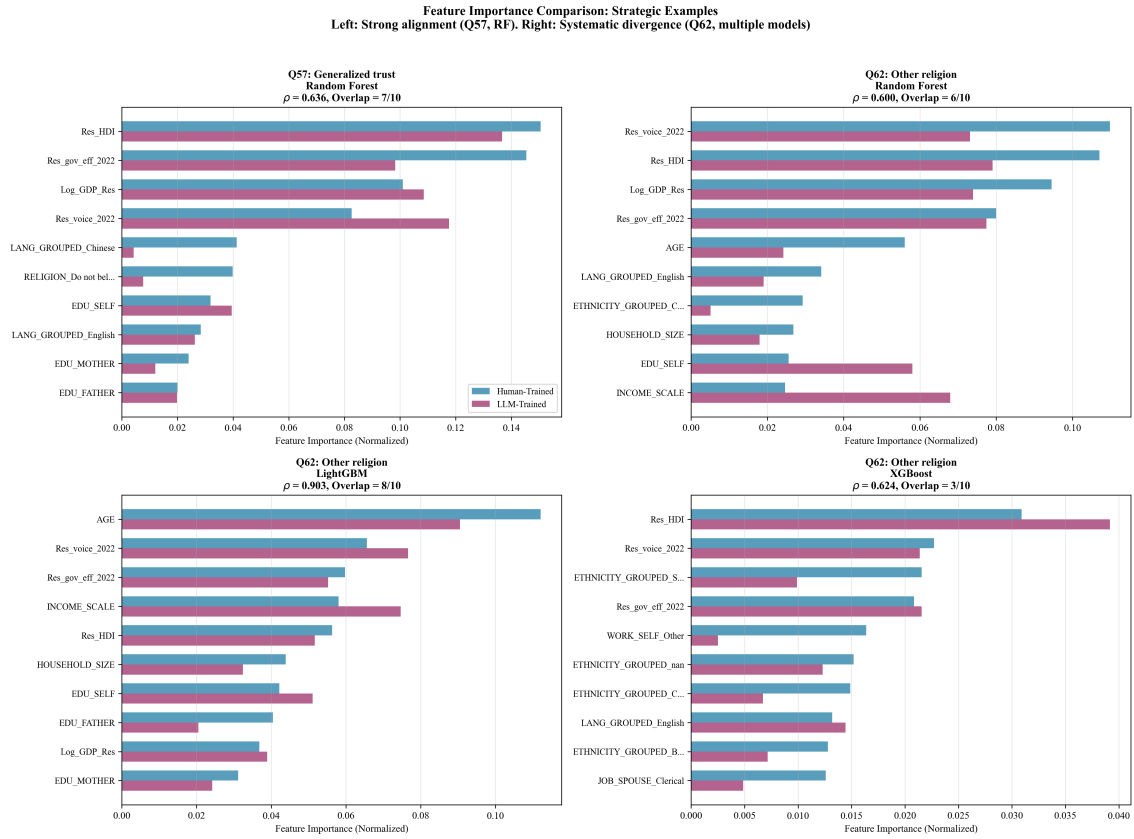


Figure 5.13: Top-10 Feature Importance Comparisons.

Plot 1 (Q57 - Random Forest): shows "Generalized Trust" demonstrating strong alignment ($\rho = 0.636$) and a high feature overlap of 7/10. Both human and LLM models prioritize regional development indicators (Res_HDI) and governance effectiveness. Plot 2 (Q62 - Random Forest) reveals a key finding regarding "Trust in Other Religions." While both models put a high emphasis on macroeconomic and macropolitical development indicators, which lead to the moderate rank correlation ($\rho = 0.600$), the features diverge at the personal level. The model trained on human responses places higher importance on identity markers (Ethnicity and Language), while the model trained on LLM-generated responses shifts focus toward economic factors (Income Scale and Education). Plot 3

(Q62 - LightGBM), in contrast to other architectures, LightGBM shows exceptional convergence on this question. Both human and LLM datasets identify Age as the dominant feature, with a 0.903 correlation, suggesting that LightGBM captures a more "universal" signal across both data sources. Plot 4 (Q62 - XGBoost): Confirms the instability observed in certain ensemble methods. On the same question (Q62), XGBoost shows a very low overlap of only 3/10 features, with the LLM model over-emphasizing regional HDI compared to the human model's more distributed reliance on ethnicity and local governance. It is also worth noting that the absolute importance scores in the XGBoost plot are significantly lower than those in the other models; this is a structural byproduct of XGBoost's gradient-based optimization and shrinkage (learning rate), which distributes influence more conservatively across features to prevent overfitting, particularly in high-dimensional feature spaces.

To observe which features the models prioritize in general, Table 5.17 presents detailed top-10 comparisons for Random Forest (the highest-alignment model in terms of correlation) across all seven questions, revealing systematic patterns in feature divergence.

Table 5.17: Complete top-10 feature comparison for Random Forest across all seven questions.

Question	Human Top-10	LLM Top-10	Overlap
Q57	Res_HDI, Res_gov_eff_2022, Log_GDP_Res, Res_voice_2022, LANG_GROUPED_Chinese, RELIGION_Do not belong to a denomination, EDU_SELF, ETHNICITY_GROUPED_nan, LANG_GROUPED_English, LANG_GROUPED_Spanish	Res_HDI, Res_voice_2022, Log_GDP_Res, Res_gov_eff_2022, INCOME_SCALE, SOCIO_CLASS, BUDGET, EDU_SELF, EDU_SPOUSE, AGE	7/10
Q58	HOUSEHOLD_SIZE, Res_gov_eff_2022, INCOME_SCALE, AGE, Res_voice_2022, Log_GDP_Res, Res_HDI, EDU_MOTHER, EDU_FATHER, EDU_SELF	HOUSEHOLD_SIZE, INCOME_SCALE, EDU_SPOUSE, MARITAL_STATUS_Married, AGE, Res_HDI, WORK_SPOUSE_nan, Log_GDP_Res, WORK_SELF_Full time, EDU_SELF	7/10
Q59	Res_voice_2022, AGE, Res_gov_eff_2022, Res_HDI, LANG_GROUPED_Spanish, Log_GDP_Res, INCOME_SCALE, EDU_FATHER, EDU_MOTHER, EDU_SELF	IS_URBAN, INCOME_SCALE, SOCIO_CLASS, AGE, Log_GDP_Res, Res_HDI, EDU_SPOUSE, Res_gov_eff_2022, BUDGET, EDU_SELF	6/10
Q60	Res_HDI, Log_GDP_Res, Res_voice_2022, Res_gov_eff_2022, LANG_GROUPED_Spanish, LANG_GROUPED_English, LANG_GROUPED_Chinese, EDU_SELF, INCOME_SCALE, AGE	INCOME_SCALE, SOCIO_CLASS, EDU_SELF, Res_HDI, Log_GDP_Res, Res_gov_eff_2022, Res_voice_2022, AGE, BUDGET, IS_URBAN	7/10
Q61	Res_HDI, Log_GDP_Res, Res_gov_eff_2022, Res_voice_2022, INCOME_SCALE, AGE, EDU_SELF, LANG_GROUPED_Spanish, ETHNICITY_GROUPED_nan, EDU_MOTHER	INCOME_SCALE, SOCIO_CLASS, Res_gov_eff_2022, Res_HDI, BUDGET, Log_GDP_Res, Res_voice_2022, AGE, EDU_SELF, IS_URBAN	7/10
Q62	Res_voice_2022, Res_HDI, Log_GDP_Res, Res_gov_eff_2022, AGE, LANG_GROUPED_English, ETHNICITY_GROUPED_Caucasian/White, EDU_SELF, LANG_GROUPED_Spanish, INCOME_SCALE	Res_HDI, Res_gov_eff_2022, Log_GDP_Res, Res_voice_2022, INCOME_SCALE, EDU_SELF, RELIGION_Muslim, AGE, SOCIO_CLASS, BUDGET	6/10
Q63	Res_HDI, Log_GDP_Res, Res_voice_2022, Res_gov_eff_2022, AGE, LANG_GROUPED_English, ETHNICITY_GROUPED_Caucasian/White, EDU_SELF, LANG_GROUPED_Spanish, INCOME_SCALE	Log_GDP_Res, Res_HDI, EDU_SELF, Res_gov_eff_2022, INCOME_SCALE, IS_URBAN, Res_voice_2022, AGE, SOCIO_CLASS, BUDGET	6/10

Detailed examination reveals two patterns:

1. Convergence on institutional features: Both model types consistently prioritize macro-level indicators, specifically HDI, governance effectiveness (Res_gov_eff_2022), voice and accountability (Res_voice_2022), and log GDP (Log_GDP_Res).

These features appear in the top-10 for all seven questions in both datasets, suggesting a shared reliance on national-level institutional quality when predicting trust attitudes.

2. Divergence on identity features: While models trained on human responses frequently elevate identity-based features (ethnicity and language), the models trained on LLM-generated responses place a heavier emphasis on individual socioeconomic markers. On Q62 and Q63, human models rely on linguistic and ethnic categories (e.g., `LANG_GROUPED_English`, `ETHNICITY_GROUPED_Caucasian/White`), whereas the models trained on LLM-generated responses consistently prioritize `INCOME_SCALE`, `SOCIO_CLASS`, and `BUDGET`.

This suggests that while both datasets agree on the importance of the national institutional context, they diverge at the individual level: models trained on human responses appear to be more tightly coupled with identity-based group membership, while models trained on LLM-generated responses are more closely aligned with economic and class-based stratification.

To assess the stability of feature importance comparisons under repeated sampling, bootstrap resampling analysis was conducted using 500 iterations per model-question combination. Each iteration involved drawing 10,000 respondents with replacement, training models on both human and LLM-aggregated responses, extracting feature importances, and computing Spearman rank correlation between the two importance rankings.

Only Q60 (trust in people known personally) was excluded from bootstrap analysis due to extreme uniformity in aggregated LLM responses. Despite individual LLMs showing variance, the averaging process collapsed this diversity to near-total uniformity. Alternative aggregation methods were tested: mode-based aggregation (democratic vote) similarly yielded 99.97% uniformity, preserving diversity in only three observations out

of 10,000. This collapse occurs because majority consensus overwhelms minority variance when responses are averaged. When eight of ten models generate Class 2 responses, averaging produces values clustering tightly around 2.0, effectively eliminating the variance present in individual model outputs.

Q58 (trust in family) presented a different edge case. While bootstrap iterations successfully generated training samples with multiple classes for most models, XGBoost exhibited NaN correlations despite 320 successful training iterations. Investigation revealed that once again the aggregated LLM responses for Q58 showed 99.99% concentration in Class 1, with individual LLMs ranging from 78.6% to 100% Class 1 responses. On this extreme class imbalance, XGBoost models learned to always predict the majority class, resulting in feature importance arrays containing all zeros (no features required for discrimination). Spearman correlation between flat zero arrays is undefined, producing NaN values. This behavior represents expected model performance on extreme imbalance rather than a methodological failure. In contrast, ensemble models (Random Forest and LightGBM) maintained positive correlations even under severe imbalance, demonstrating greater robustness to distributional collapse.

Table 5.18 presents comprehensive bootstrap results for all five model architectures across all validated questions. The table reports mean Spearman correlation, standard deviation across 500 iterations, 95% confidence intervals computed via the percentile method, and the number of successful iterations for each combination.

Table 5.18: Complete bootstrap validation results: 500-iteration analysis of feature importance correlation stability.

Model	Question	Mean ρ	Std. Dev.	95% CI	N Iterations
Random Forest	Q57	0.806	0.022	[0.761, 0.846]	500
	Q58	0.656	0.047	[0.546, 0.736]	329
	Q59	0.725	0.042	[0.635, 0.794]	500
	Q61	0.767	0.023	[0.723, 0.810]	500
	Q62	0.859	0.017	[0.827, 0.890]	500
	Q63	0.830	0.020	[0.792, 0.870]	500
	Average	0.774	0.029	[0.714, 0.824]	—
LightGBM	Q57	0.539	0.043	[0.457, 0.621]	500
	Q58	0.547	0.045	[0.459, 0.631]	305
	Q59	0.553	0.045	[0.453, 0.633]	500
	Q61	0.623	0.038	[0.548, 0.694]	500
	Q62	0.674	0.032	[0.613, 0.734]	500
	Q63	0.743	0.028	[0.683, 0.792]	500
	Average	0.613	0.039	[0.536, 0.684]	—
Decision Tree	Q57	0.388	0.074	[0.240, 0.530]	500
	Q58	0.053	0.081	[-0.109, 0.218]	320
	Q59	0.213	0.103	[0.022, 0.399]	500
	Q61	0.401	0.075	[0.251, 0.551]	500
	Q62	0.468	0.070	[0.311, 0.597]	500
	Q63	0.444	0.073	[0.299, 0.589]	500
	Average	0.328	0.079	[0.169, 0.481]	—
XGBoost	Q57	0.180	0.075	[0.029, 0.307]	500
	Q58	NaN	NaN	NaN	320*
	Q59	0.043	0.067	[-0.093, 0.173]	500

Continued on next page

Table 5.18 – continued from previous page

Model	Question	Mean ρ	Std. Dev.	95% CI	N Iterations
	Q61	0.156	0.078	[0.007, 0.307]	500
	Q62	0.251	0.080	[0.097, 0.412]	500
	Q63	0.220	0.080	[0.064, 0.377]	500
	Average	0.170**	0.076	[—]	—
	Q57	0.149	0.073	[0.009, 0.282]	500
	Q58	-0.320	0.064	[-0.432, -0.190]	334
	Q59	-0.281	0.076	[-0.426, -0.127]	499
Logistic Regression	Q61	0.145	0.067	[0.006, 0.277]	500
	Q62	0.234	0.060	[0.127, 0.358]	500
	Q63	0.106	0.064	[-0.012, 0.231]	500
	Average	0.006	0.067	[-0.121, 0.139]	—

Note: Q58 XGBoost results show NaN due to extreme class imbalance (99.99% Class 1). Q60 excluded due to 100% single class LLM response uniformity.

** For Q58 XGBoost, 320 iterations were successfully completed but returned NaN as the model generated zero feature importance.*

*** Q58 was excluded from the XGBoost average to prevent skewing the aggregate correlation metric.*

The results point to notable differences in feature importance stability across model architectures. Among them, Random Forest stands out with the highest overall alignment: it achieves a mean Spearman correlation of 0.774, accompanied by narrow confidence intervals that consistently exclude zero. This suggests that Random Forest reliably identifies demographic predictors that are common to both human and LLM-generated datasets. LightGBM shows a similarly strong pattern (mean $\rho = 0.613$), maintaining stable, positive correlations even on challenging questions like Q58, where class imbalance is pronounced.

Decision Trees, on the other hand, exhibit more modest alignment (mean $\rho =$

0.328), and their estimates are less stable—reflected in higher variance across bootstrap samples (mean $\sigma \approx 0.08$). For Q58, the confidence interval crosses zero, which aligns with the known sensitivity of greedy splitting algorithms to small variations in training data. XGBoost presents an interesting contrast: although it ranked as the most accurate model in earlier comparisons, its feature importance correlation is relatively weak (mean $\rho = 0.170$). In some questions, such as Q59, the confidence interval spans both negative and positive values. This suggests that XGBoost may be capturing complex or dataset-specific interactions that improve predictive accuracy but do not translate into stable feature patterns across human and synthetic responses.

Logistic Regression shows the weakest alignment ($\rho = 0.006$), with significantly negative correlations in questions characterized by high class imbalance. While this may partly reflect the limited capacity of linear models to capture high-dimensional patterns, the sharp divergence from the high alignment seen in Random Forest is telling. It suggests that the linear decision logic and feature hierarchies humans follow are fundamentally different from those encoded in LLMs. The negative values imply that the straightforward, first-order associations found in human attitudes (e.g., direct demographic predictors) are not merely absent but are potentially restructured or prioritized differently within the models' internal representations.

Taken together, these findings indicate that specific ensemble architectures—most notably Random Forest and LightGBM—consistently uncover a shared demographic logic across human and synthetic datasets. In contrast, models relying on linear or highly specialized optimization strategies appear less capable of capturing generalizable patterns of association.

Figure 5.14 visualizes the distribution of Spearman correlations across 500 iterations, demonstrating well-behaved sampling distributions that validate the use of percentile-based confidence intervals. Random Forest and LightGBM consistently show tight

distributions centered at high correlation values. In contrast, XGBoost and Logistic Regression exhibit significantly wider distributions centered at low or negative values, reflecting their inherent instability in this context.

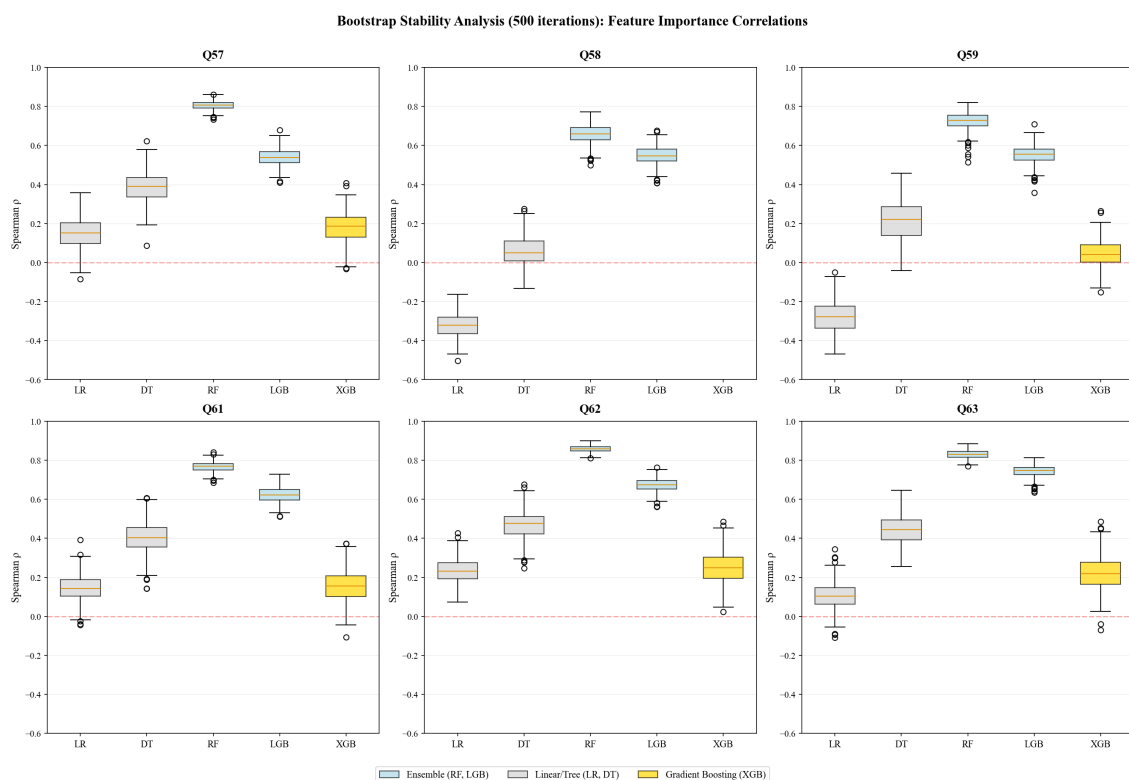


Figure 5.14: *Bootstrap distributions of Spearman correlations for selected question-model combinations across 500 iterations. Vertical lines indicate 95% confidence interval boundaries.*

The stability hierarchy observed in the primary analysis—with Random Forest and LightGBM at the top, followed by Decision Trees, and ending with XGBoost and Logistic Regression—persists across 500 independent resampling iterations. This consistency confirms that the specific ensemble strategies in Random Forest and LightGBM reveal more robust demographic-opinion associations than either single trees or highly-specialized boosting implementations like XGBoost.

A critical finding in this analysis is the contrast between predictive accuracy and feature importance stability. While XGBoost achieves high accuracy, it demonstrates minimal feature importance alignment. Conversely, Random Forest and LightGBM achieve moderate-to-high accuracy with maximal alignment, which suggests two distinct modeling pathways: one that prioritizes prediction through complex, dataset-specific interactions (as seen in XGBoost), and one that prioritizes generalizable demographic patterns (as seen in Random Forest and LightGBM). For research specifically requiring an understanding of the demographic drivers behind trust, these results indicate that Random Forest and LightGBM provide more reliable insights than highly-optimized boosting or linear models.

The bootstrap process also highlights a methodological challenge regarding LLM response homogeneity. In cases like Q60, where the aggregated LLM responses converged to nearly 100% uniformity, the variance required for stable feature importance extraction effectively collapsed. While humans expressed natural variance for this question—with 24% expressing distrust toward people they know personally—the synthetic consensus suppressed this population heterogeneity. This can be interpreted as follows: while aggregation is a standard strategy for reducing individual model bias, it may simultaneously obscure realistic diversity by over-emphasizing normative responses. This must remain a core consideration for the future use of LLMs in synthetic data ensembles.

6. DISCUSSION AND LIMITATIONS

This chapter of the thesis discusses the findings of the study in relation to the two central research questions: (1) whether LLMs can simulate individual survey responses when prompted with detailed demographic personas, and (2) whether traditional machine learning models trained on real human data outperform synthetic responses generated by LLMs in predictive tasks. The analysis also explores how closely the feature importance derived from synthetic responses aligns with that of real survey data. By interpreting the empirical results presented in the previous sections, this chapter seeks to assess the representational fidelity of LLMs and their potential as substitutes for human respondents in opinion research. The discussion is then followed by a reflection on the limitations of the study.

6.1 Discussion

The individual-level evaluation of representational fidelity revealed that LLMs consistently outperform random baselines across all trust-related questions. Despite minor variations, the ten tested models exhibited a narrow band of accuracy scores, suggesting a largely model-agnostic capability in simulating individual survey responses based on exact-match accuracy alone. This convergence in performance—across both high-parameter models like GPT-5 and lightweight alternatives such as the nano variants—raises an important observation: while scale and recency appear to correlate weakly with marginal gains in accuracy, the general representational capacity for demographic-driven opinion generation seems widely distributed across architectures.

One notable pattern emerging from the results was the relationship between prediction accuracy and the social distance of the target group. LLMs demonstrated the highest exact-match accuracy when prompted about trust in family members (Q58). However, as social distance increases, a significant gap appears: while models correctly shift their

average responses toward lower trust levels, exact-match accuracy steadily declines for groups such as individuals from other religions (Q62) or nationalities (Q63). This gradient suggests that while LLMs effectively mirror the normative hierarchy of social trust, they struggle to replicate the increased heterogeneity and contextual nuance that characterize human attitudes as social distance expands.

Accounting for the ordinal structure of the trust items, MAE provides a granular measure of predictive proximity. Here again, models clustered closely together, with most predictions diverging by roughly half a point on a four-point scale. This relatively low average error highlights that even incorrect responses tend to stay near the correct answer, reinforcing the notion that LLMs may possess a structured but imprecise mapping from demographic cues to opinion proxies.

Interestingly, the lack of stark performance gaps across models, combined with their similar error profiles, raises a deeper question: are these systems simply repeating patterns they memorized during training, or are they actually forming consistent opinions based on input? The consistency observed across architectures—especially on questions involving close social circles—could imply a shared internalization of demographic-opinion linkages. However, the extent to which this consistency reflects genuine representation versus memorized heuristics remains an open question. Along the aim to address this question, the next set of analyses evaluates whether LLMs can replicate the broader statistical patterns observed in real population data. Specifically, it examines whether the distributions of model-generated responses align with human response distributions at the aggregate level, using metrics such as JS divergence and entropy.

While models perform reasonably well at identifying the most common response category across trust questions, they consistently fail to reproduce the full distributional shape of human responses. JS divergence scores increase as questions shift from close social circles to out-groups, signaling a breakdown in the ability of LLMs to model the

diversity and variance present in population-level attitudes.

This misalignment becomes more apparent when examining response entropy and distribution concentration. In questions like Q61, where human responses are moderately dispersed with meaningful representation of minority viewpoints, LLM outputs exhibit sharply reduced entropy and near-total fixation on the majority class. In questions with high human consensus (e.g., Q58), the issue is even more extreme: models converge almost entirely on a single answer, effectively eliminating all minority classes. The introduction of the MCF confirms this behavior quantitatively, capturing the degree to which models amplify the dominance of modal responses beyond what is observed in real data. This behavior is consistent with the broader phenomenon of mode collapse seen in generative modeling, where outputs disproportionately cluster around the most frequent token.

This systemic suppression of minority viewpoints highlights a fundamental limitation of the 'persona prompting' paradigm. While LLMs are constrained by the fixed set of demographic descriptors provided in the prompt, human survey responses are frequently shaped by individual life experiences and unobserved factors that lie entirely outside the demographic frame. For instance, an extreme response—such as expressing 'no trust at all' in one's family (Q58) or 'complete trust' in strangers (Q61)—may be driven by personal history rather than social identity. An individual may have experienced deep-seated family estrangement, or conversely, a physical disability or health condition that necessitates a high degree of reliance on strangers for daily functioning.

Furthermore, a portion of the human variance that LLMs fail to replicate may simply reflect measurement noise or respondent error. In large-scale surveys like the WVS, some participants may misinterpret the ordinal scale (e.g., assuming '4' is the highest trust instead of '1'), suffer from survey fatigue, or engage in intentional non-compliance. Because LLMs operate as 'rationalized' statistical mirrors of demographic tropes, they

are inherently incapable of reproducing these 'messy' human realities. They lack the personal trauma, the situational dependency, and even the capacity for simple human error, leading to a synthetic distribution that is 'cleaner' but significantly less representative of the actual spectrum of reported human behavior.

While these findings are troubling, one might still expect LLMs to reflect finer-grained variation across demographic subgroups. However, further analysis reveals that this is not the case. Gender, age, and country-level comparisons demonstrate that LLMs systematically underrepresent group-level differences. Although they occasionally reproduce the correct direction of a pattern—such as older individuals expressing higher trust than younger ones—the magnitude of these differences is severely compressed. Most strikingly, LLM responses fail to show any meaningful variation across countries for most questions, flattening cultural distinctions into global averages. This failure to reflect contextual, demographic nuance severely limits their use for any research aiming to model opinion heterogeneity across populations.

Given these patterns of over-determinism and compression, a critical question arises: are LLMs truly reasoning over demographic personas, or are they simply defaulting to high-probability outputs regardless of input? To test this, a controlled experiment evaluated whether LLMs could even approximate random output when prompted to do so under the same, default parametric settings. The results were striking—models failed to produce uniform distributions in any setting, consistently defaulting to a single number (usually “3”) regardless of prompt complexity. However, the shift in response distributions across different prompt stages (e.g., with or without survey context) suggests that models are not purely random: they are responding in structured, but biased, ways to the input.

These findings present an apparent contradiction. On one hand, LLMs are not generating purely stochastic responses: they consistently select the first or second most

likely classes, and they adjust their outputs in response to prompt context. On the other hand, they fail to meaningfully reflect demographic signals, compress group differences, and collapse variance. This raises a practical question of whether such systems can still be useful in survey simulations.

In this light, the results are not necessarily discouraging. If the purpose of a survey is to identify dominant trends—such as the most likely candidate to win an election, or the most popular product flavor—then LLMs may provide surprisingly robust predictions. After all, traditional surveys often do not aim to understand each individual respondent in detail, but rather to approximate the preferences of a representative household or demographic profile. In such use cases, the consistent selection of majority preferences by LLMs may actually mirror the core purpose of survey research. However, for more fine-grained analyses—such as determining whether women or men are more likely to support a given policy, or whether younger people trust government less than older people—LLMs fail to replicate the necessary granularity. Thus, the utility of LLMs in synthetic survey research may ultimately depend on the type of insight being sought: broad population-level signals may be recoverable, but subgroup-level inference remains unreliable.

This trade-off marks a natural transition to the second research question, which evaluates whether supervised machine learning models trained on real survey data can outperform LLM-generated responses in predictive tasks, and whether they capture demographic signals in fundamentally different ways.

Across the predictive performance analyses, the evaluated machine learning models exhibit clear but uneven performance differences. Among the five architectures tested, XGBoost stands out as the only model that consistently achieves higher average accuracy than the LLM benchmark. All models, however, demonstrate relatively low standard deviations under cross-validation, indicating stable generalization rather than performance

driven by specific train–validation splits. This stability suggests that the observed accuracy patterns reflect systematic properties of the models rather than noise or overfitting.

Benchmarking these performances against chance-level and majority-class baselines provides further insight into the nature of the predictive signal. All five machine learning models substantially outperform random guessing across all trust questions, confirming that demographic features contain non-trivial predictive information. However, only XGBoost consistently exceeds the majority-class baseline. For questions involving intimate trust—most notably trust in family members—XGBoost often matches rather than surpasses the majority heuristic, indicating that even highly optimized supervised models struggle to extract demographic signal beyond overwhelming class dominance. In this respect, XGBoost’s behavior closely mirrors that of LLMs, which similarly converge on the modal response for high-consensus questions.

The distinction between machine learning models and LLMs becomes more pronounced for out-group trust questions. As the dominance of the majority class weakens with increasing social distance, XGBoost demonstrates a clear ability to leverage demographic structure, achieving substantial gains over the majority baseline. In contrast, LLMs largely fail to capitalize on this additional signal, continuing to exhibit distributional compression and limited responsiveness to demographic variation. Nevertheless, it is notable that the predictive performance of all machine learning models—including XGBoost—declines as social distance increases. This shared degradation across model classes suggests that trust in distant or abstract groups is intrinsically more difficult to predict from demographic attributes alone, rather than being a limitation unique to LLM-based simulation.

The MAE analysis reinforces these conclusions from a complementary perspective. While XGBoost again achieves the lowest overall error, LLMs outperform all other machine learning models in ordinal proximity, indicating that their incorrect predictions

tend to fall close to the true response category. This pattern suggests that LLMs encode coarse-grained directional tendencies even when failing to achieve exact matches. However, this apparent advantage should be interpreted cautiously: lower MAE in LLM outputs may reflect smoothing and modal attraction rather than genuine sensitivity to fine-grained demographic differences.

At the same time, a direct comparison between machine learning models and LLMs as predictive systems warrants careful contextualization. Supervised models require extensive preprocessing pipelines—including data cleaning, imputation, feature encoding, and dimensionality management—each of which depends on domain expertise and methodological choices. LLMs, by contrast, can be prompted directly with rich demographic descriptors such as language, ethnicity, and country, without explicit feature engineering or explanation. From a practical standpoint, extracting predictions from LLMs demands significantly less technical know-how than training and validating supervised learning models.

More importantly, the two approaches differ fundamentally in their data requirements. Machine learning models cannot operate without an initial “seed” dataset of human responses; their predictive power is inherently bounded by the availability and quality of observed survey data. LLMs, in contrast, are capable of zero-shot generalization, producing plausible responses even in the absence of any task-specific training data. In the context of survey simulation, this distinction is critical. If the objective is to generate synthetic responses when no prior survey data exist, supervised learning is not a viable alternative, regardless of its superior performance under ideal conditions.

Having established that LLMs can compete with most supervised models in predictive accuracy—despite the structural advantages afforded to traditional machine learning—Research Question 2 next examines a deeper issue: do LLMs and humans rely on similar demographic signals when forming trust judgments? Predictive parity alone is

insufficient if the underlying decision logic diverges. To address this question, feature importance analysis was introduced as a diagnostic tool.

Because neither human reasoning nor LLM inference exposes causal weights explicitly, machine learning models were employed as interpretable proxies. In this framework, models trained on human responses approximate the demographic signals underlying real decision-making, while models trained on LLM-generated responses approximate the implicit logic encoded in synthetic outputs. Comparing feature importance across these two settings allows an indirect assessment of whether similar informational cues drive both systems.

To summarize alignment in a single, comparable metric, Spearman rank correlation was used to measure the similarity between feature importance rankings derived from the ML models trained on human responses and models trained on LLM responses. The results demonstrate substantial heterogeneity across model architectures. Random Forest and LightGBM consistently exhibit high alignment, while Logistic Regression and XGBoost show weak or unstable correspondence.

The strong performance of Random Forest and LightGBM suggests that ensemble-based, tree-structured models are particularly well suited to uncovering shared demographic logic across human and synthetic data. These models aggregate across multiple weak learners, smoothing noise and dampening idiosyncratic feature interactions. LightGBM, in particular, achieves the highest top-10 feature overlap—averaging eight shared features across questions—which likely reflects the combined benefits of regularization, sparsity-aware splitting, and ensemble averaging. Under the selected hyperparameters, LightGBM relied on relatively few dominant predictors, a property that appears to enhance interpretability and cross-dataset stability.

In contrast, XGBoost presents a striking divergence between predictive performance and explanatory consistency. Despite achieving the highest accuracy in the previous section, its feature importance alignment with human-response-trained models is weak. This outcome appears to stem from two interacting factors. First, XGBoost’s aggressive gradient-based optimization prioritizes residual minimization, allowing it to exploit a significantly broader array of features to capture marginal gains. Unlike LightGBM’s reliance on a few dominant predictors, XGBoost utilizes complex, high-dimensional interactions that improve prediction but fail to yield stable or transferable feature rankings. Second, in highly imbalanced settings—particularly for intimate trust questions—the model converges on majority-class prediction despite efforts to counter the class imbalances with balanced rewarding, rendering feature importance extraction either trivial or undefined. While this behavior is not a methodological failure, it highlights an important distinction: high predictive accuracy, achieved through the exploitation of many subtle predictors, does not necessarily imply shared reasoning or explanatory validity.

Beyond architectural differences, the feature importance results also reveal a substantive pattern regarding the nature of trust formation itself. Across both human and LLM-trained models, macro-level country characteristics—such as Human Development Index, governance effectiveness, voice and accountability, and national income—emerge as the most consistently important predictors. In contrast, individual-level demographic attributes play a secondary role. This finding suggests that trust attitudes are shaped less by who individuals are than by the institutional and socio-economic environments in which they reside. In this sense, the results echo the long-standing sociological intuition that “geography is destiny”: trust appears to be structurally embedded in national context rather than primarily determined by personal identity.

At the individual level, however, divergence emerges. Models trained on human

responses place greater emphasis on identity-linked features such as ethnicity and language, particularly for out-group trust. Models trained on LLM-generated responses, by contrast, consistently elevate socioeconomic indicators such as income, education, and social class. This pattern suggests that while LLMs internalize broad institutional gradients accurately, they may substitute identity-based group membership with more generic class-based reasoning. Rather than mirroring how individuals differentiate social groups, LLMs appear to rely on abstract socioeconomic proxies, potentially reflecting normative or globally averaged narratives present in training data.

Taken together, these findings clarify the role LLMs can—and cannot—play in synthetic survey research. When the objective is to recover dominant structural drivers of opinion, particularly at the macro level, LLMs appear to encode a demographic logic broadly consistent with human data. However, when the research question requires sensitivity to identity, subgroup differentiation, or culturally specific social boundaries, LLM-generated responses diverge in systematic ways. Importantly, this divergence is not fully visible in predictive accuracy metrics alone; it only becomes apparent when explanatory structure is examined directly.

This distinction reinforces a central theme of the thesis: LLMs may approximate outcomes of human judgment under certain conditions, but they do not necessarily replicate the processes by which those judgments are formed. The implications of this gap are explored further in the limitations and concluding sections that follow.

6.2 Limitations and Ethical Considerations

This study has demonstrated that large language models can simulate survey responses with surprising accuracy and partial alignment in demographic reasoning. However, several important limitations—both technical and epistemological—must be acknowledged when interpreting these findings or considering the broader use of LLMs

in survey simulation and synthetic data generation.

A fundamental limitation lies in the opacity of LLM training data. For proprietary models, it is impossible to verify whether the World Values Survey or similar instruments were included in pretraining. If such data were present, the observed response accuracy might reflect memorization rather than genuine generalization, undermining the validity of any claims about representational fidelity. Even in the absence of direct leakage, LLMs may rely on linguistic priors shaped by existing survey literature, further complicating the distinction between simulation and imitation.

In terms of model behavior, the study reveals consistent patterns of systematic bias. An important pattern is the strong fixation to majority-class responses observed across trust questions, often accompanied by a suppression of minority responses. This behavior mirrors the “mode collapse” phenomenon in generative models and leads to an underrepresentation of diverse perspectives—particularly in questions involving out-groups or less socially normative targets of trust. As a result, the synthetic distributions generated by LLMs tend to flatten cultural, demographic, and geographic heterogeneity into a narrow range of high-probability outputs.

The prompting strategy used in this study—zero-shot demographic persona-based prompting—provides a useful baseline but does not represent the upper bound of LLM capability. More sophisticated approaches such as few-shot prompting, narrative backstories, or chain-of-thought reasoning may elicit richer responses. However, these techniques also risk introducing new biases or obscuring underlying model behavior. Importantly, the findings reported here reflect what LLMs do when given structured demographic input, not what they could do under optimal prompting conditions. Future research should test whether such techniques can meaningfully mitigate issues like minority underrepresentation and country-level flattening.

From a methodological standpoint, several constraints must be noted. First, this

study uses machine learning models trained on both human and synthetic data to infer feature importance, assuming that these models can serve as proxies for underlying reasoning processes. While ensemble models like Random Forest and LightGBM show strong and stable alignment in feature rankings, others—most notably XGBoost and Logistic Regression—do not. This discrepancy highlights the challenge of using prediction-focused models to draw interpretive conclusions. Second, the study focuses exclusively on trust-related items from the World Values Survey. Generalizing these results to other question types (e.g., moral, political, or behavioral items) requires further validation. Finally, only 10 LLMs were tested, and their outputs were aggregated without tuning temperature or sampling strategies. Aggregation, while stabilizing variance, also obscures individual model diversity and may exacerbate normative bias.

Beyond technical considerations, the use of LLMs in social research raises deeper epistemological and ethical concerns. LLMs generate plausible linguistic proxies of opinion, but they do not possess beliefs, experiences, or values. Treating their outputs as equivalent to human survey responses risks conflating textual plausibility with attitudinal substance. When an LLM outputs “I trust people of other religions,” it is not expressing a belief—it is reproducing a statistical pattern. This distinction becomes particularly critical when LLM outputs are used to inform research conclusions or policy decisions.

Moreover, the risk of algorithmic homogenization looms large. By overemphasizing dominant responses and underrepresenting minority views, LLM-generated data may contribute to the erasure of dissenting or marginalized voices. This is especially problematic in domains where social inclusion, group identity, or cultural variation are essential. If synthetic data replaces real survey participation for reasons of cost or speed, democratic processes that give citizens a voice in research and policymaking could be undermined. In such scenarios, policymakers may be making decisions based on algorithmically smoothed opinions rather than the lived realities of diverse populations.

Taken together, these limitations underscore the need for caution when interpreting LLM outputs in opinion research.



7. CONCLUSIONS

This thesis set out to evaluate whether Large Language Models LLMs can function as synthetic survey respondents when prompted with detailed demographic personas, and how their performance compares to traditional supervised learning models trained on real human data. Across four specific hypotheses, the findings paint a consistent picture: while LLMs can effectively approximate dominant response patterns under minimal engineering requirements, they fundamentally fall short in reproducing the population-level diversity and demographic heterogeneity inherent in human societies. These results align with the work of Agnew et al. (2024), further substantiating the observation that LLMs provide a "flattened" representation of human opinion that lacks the nuanced "long tails" of real-world populations [57]. These outcomes are a predictable consequence of the probabilistic architecture of LLMs; as systems optimized for maximum likelihood estimation, they are inherently predisposed toward mode-seeking behavior, making distributional collapse an integral feature of their operational logic.

The first hypothesis regarding individual-level accuracy was strongly supported, as all tested architectures consistently exceeded chance-level baselines. This performance indicates that demographic–opinion associations are broadly encoded in contemporary models. However, this success is highly contingent on the target of trust; accuracy peaked for intimate social circles and declined systematically as social distance increased toward out-groups and abstract nationalities. The Mean Absolute Error analysis further clarified this behavior, revealing that even when LLMs fail to hit the exact target, they remain directionally accurate, suggesting a structured but coarse-grained mapping from demographic cues to opinion proxies.

In contrast, the second hypothesis regarding distributional representation received only partial support. While LLMs can approximate the most common response categories, they act as "dominant pattern amplifiers" rather than faithful replicators of human

variance. The analysis identified a systematic majority-class fixation, where models effectively eliminate minority viewpoints and sharply reduce response entropy. This distributional compression is even more pronounced at the subgroup level; while LLMs occasionally capture the correct direction of demographic trends, they severely underrepresent the magnitude of group-level differences and flatten cultural distinctions into global averages. As such, the hypothesis is rejected for population-level heterogeneity preservation.

When benchmarked against supervised machine learning, the third hypothesis regarding predictive performance received mixed support. LLMs occupy a "middle tier" of predictive capability: they consistently outperform linear models and simple decision trees, and achieve parity with ensemble averaging methods like Random Forest. However, they are significantly outperformed by optimized gradient boosting architectures. This finding highlights a critical trade-off: supervised methods remain superior when high-quality labeled data are available, yet LLMs offer a unique zero-shot advantage, producing plausible responses without the need for seed datasets as highlighted as the "Data Availability Paradox" or explicit feature engineering.

Crucially, predictive parity did not imply shared reasoning, leading to only partial support for the fourth hypothesis regarding feature importance alignment. Only specific ensemble architectures—Random Forest and LightGBM—consistently uncovered a stable demographic logic common to both human and synthetic data. Feature analysis revealed a two-tier pattern: while both humans and LLMs rely on macro-level institutional indicators (such as national income and governance quality), they diverge at the individual level. Humans place greater emphasis on identity-linked features like ethnicity and language, whereas LLMs substitute these with generic socioeconomic proxies. This suggests that LLMs approximate the outcomes of human judgment through a different internal hierarchy of predictors.

In conclusion, LLMs are best understood not as direct replacements for human

respondents, but as tools for approximating dominant structural trends under data-scarce conditions. They are well-suited for hypothesis generation and exploratory modeling but remain unreliable for high-fidelity research requiring a faithful representation of minority views, subgroup variation, or culturally specific boundaries. While LLMs offer promising new frontiers for computational social science, their use must be accompanied by methodological caution, ensuring that the convenience of synthetic data does not obscure the essential complexity and diversity of real human behavior.



REFERENCES

- [1] A. Joshi, R. Gertner, L. Roberts, and A. El-Mohandes, “An Evidence-Based Approach on Academic Management in a School of Public Health Using SMAART Model,” *Sustainability*, vol. 13, p. 12256, Nov. 2021.
- [2] P. V. Miller, “Is There a Future for Surveys?,” *Public Opinion Quarterly*, vol. 81, no. S1, pp. 205–212, 2017.
- [3] J. Bacher and M. Weichbold, “Editorial,” *Österreichische Zeitschrift für Soziologie*, vol. 45, pp. 107–113, June 2020.
- [4] S. Köksal, L. M. Pesando, V. Rotondi, and E. Şanlıtürk, “Harnessing the Potential of Google Searches for Understanding Dynamics of Intimate Partner Violence Before and After the COVID-19 Outbreak,” *European Journal of Population*, vol. 38, pp. 517–545, Aug. 2022.
- [5] K. Smith, A. Macintyre, S. Weakley, S. Hill, O. Escobar, and G. Fergie, “Public understandings of potential policy responses to health inequalities: Evidence from a UK national survey and citizens’ juries in three UK cities,” *Social Science & Medicine*, vol. 291, p. 114458, Dec. 2021.
- [6] P. Keefer, C. G. Scartascini, and R. Vlaicu, “Social Trust and Electoral Populism: Explaining the Quality of Government,” *SSRN Electronic Journal*, 2019.
- [7] S. Boronenkova, A. Chepulyanis, and S. Nikulionok, “Methods of marketing analysis at all stages of the product life cycle,” *E3S Web of Conferences*, vol. 296, p. 06032, 2021.
- [8] V. Umaashankar and G. S. S, “Ask less - Scale Market Research without Annoying Your Customers,” Jan. 2019. arXiv:1901.08744 [cs].
- [9] D. O’Neill, M. Benzeval, A. Boyd, L. Calderwood, C. Cooper, L. Corti, E. Dennison, E. Fitzsimons, A. Goodman, R. Hardy, H. Inskip, L. Molloy, A. Sacker, A. Sudlow, A. Sullivan, and A. Park, “Data Resource Profile: Cohort and Longitudinal Studies Enhancement Resources (CLOSER),” *International Journal of Epidemiology*, vol. 48, pp. 675–676i, June 2019.
- [10] M. Herle, N. Micali, M. Abdulkadir, R. Loos, R. Bryant-Waugh, C. Hübel, C. M. Bulik, and B. L. De Stavola, “Identifying typical trajectories in longitudinal data: modelling strategies and interpretations,” *European Journal of Epidemiology*, vol. 35, pp. 205–222, Mar. 2020.
- [11] L. Miao, M. Last, and M. Litvak, “Tracking social media during the COVID-19 pandemic: The case study of lockdown in New York State,” *Expert Systems with Applications*, vol. 187, p. 115797, Jan. 2022.

- [12] F. J. Fowler, *Survey research methods*. No. 1 in Applied social research methods series, Los Angeles, Calif.: SAGE, 5. ed ed., 2014.
- [13] K. Olson, J. Wagner, and R. Anderson, “Survey Costs: Where are We and What is the Way Forward?,” *Journal of Survey Statistics and Methodology*, vol. 9, pp. 921–942, Dec. 2021.
- [14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Nee-lakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCan-dlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” July 2020. arXiv:2005.14165 [cs].
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” June 2017. arXiv:1706.03762 [cs].
- [16] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, “Emergent Abilities of Large Language Models,” Oct. 2022. arXiv:2206.07682 [cs].
- [17] J. Huang, J.-t. Huang, Z. Liu, X. Liu, W. Wang, and J. Zhao, “AI Sees Your Location, But With A Bias Toward The Wealthy World,” Sept. 2025. arXiv:2502.11163 [cs].
- [18] L. P. Argyle, E. C. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate, “Out of One, Many: Using Language Models to Simulate Human Samples,” *Political Analysis*, vol. 31, pp. 337–351, July 2023. arXiv:2209.06899 [cs].
- [19] B. Wang, Z.-Y. Khoo, and B. K. H. Low, “Prompts to Proxies: Emulating Human Preferences via a Compact LLM Ensemble,” Sept. 2025. arXiv:2509.11311 [cs].
- [20] S. Balloccu, P. Schmidtová, M. Lango, and O. Dušek, “Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs,” Feb. 2024. arXiv:2402.03927 [cs].
- [21] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (Virtual Event Canada), pp. 610–623, ACM, Mar. 2021.
- [22] E. Ferrara, “Should ChatGPT be biased? Challenges and risks of bias in large lan-guage models,” *First Monday*, Nov. 2023.
- [23] G. Aher, R. I. Arriaga, and A. T. Kalai, “Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies,” July 2023. arXiv:2208.10264 [cs].

- [24] J. Rao and I. Molina, *Small Area Estimation*. Wiley, 1 ed., Oct. 2015.
- [25] D. K. Park, A. Gelman, and J. Bafumi, “Bayesian Multilevel Estimation with Post-stratification: State-Level Estimates from National Polls,” *Political Analysis*, vol. 12, no. 4, pp. 375–385, 2004.
- [26] C. Hanretty, “An Introduction to Multilevel Regression and Post-Stratification for Estimating Constituency Opinion,” *Political Studies Review*, vol. 18, pp. 630–645, Nov. 2020.
- [27] J. R. Lax and J. H. Phillips, “How Should We Estimate Public Opinion in The States?,” *American Journal of Political Science*, vol. 53, pp. 107–121, Jan. 2009.
- [28] N. Jiang and A. Crooks, “Studying Contagious Disease Spread Utilizing Synthetic Populations Inspired by COVID-19: An Agent-based Modeling Framework,” in *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation*, (Atlanta GA USA), pp. 29–32, ACM, Oct. 2024.
- [29] P. B. Miranda, J. Giráldez-Cru, M. W. Silva-Filho, C. Zarco, and O. Córdón, “Learning Agents’ Behavioral Patterns in Agent-Based Modeling by Means of Evolutionary Algorithms,” in *2024 IEEE Congress on Evolutionary Computation (CEC)*, (Yokohama, Japan), pp. 1–8, IEEE, June 2024.
- [30] M. Gao, Z. Wang, K. Wang, C. Liu, and S. Tang, “Forecasting elections with agent-based modeling: Two live experiments,” *PLOS ONE*, vol. 17, p. e0270194, June 2022.
- [31] C. Hanretty, “Areal interpolation and the UK’s referendum on EU membership,” *Journal of Elections, Public Opinion and Parties*, vol. 27, pp. 466–483, Oct. 2017.
- [32] D. B. Rubin, “Statistical Disclosure Limitation,” *Journal of Official Statistics*, vol. 9, no. 2, pp. 461–468, 1993.
- [33] Z. Qian, T. Callender, B. Cebere, S. M. Janes, N. Navani, and M. Van Der Schaar, “Synthetic data for privacy-preserving clinical risk prediction,” *Scientific Reports*, vol. 14, p. 25676, Oct. 2024.
- [34] J. M. Mendes, A. Barbar, and M. Refaie, “Synthetic data generation: a privacy-preserving approach to accelerate rare disease research,” *Frontiers in Digital Health*, vol. 7, p. 1563991, Mar. 2025.
- [35] J. Drechsler and J. P. Reiter, “An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets,” *Computational Statistics & Data Analysis*, vol. 55, pp. 3232–3243, Dec. 2011.
- [36] B. v. Breugel, Z. Qian, and M. v. d. Schaar, “Synthetic data, real errors: how (not) to publish and use synthetic data,” July 2023. arXiv:2305.09235 [cs].

- [37] Z. S. Harris, “Distributional Structure,” *WORD*, vol. 10, pp. 146–162, Aug. 1954.
- [38] A. C. Kozlowski, M. Taddy, and J. A. Evans, “The Geometry of Culture: Analyzing Meaning through Word Embeddings,” *American Sociological Review*, vol. 84, pp. 905–949, Oct. 2019. arXiv:1803.09288 [cs].
- [39] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. v. Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, “On the Opportunities and Risks of Foundation Models,” July 2022. arXiv:2108.07258 [cs].
- [40] M. Lutz, I. Sen, G. Ahnert, E. Rogers, and M. Strohmaier, “The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models,” Oct. 2025. arXiv:2507.16076 [cs].
- [41] W. Zhao, D. Mondal, N. Tandon, D. Dillion, K. Gray, and Y. Gu, “WorldValues-Bench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models,” Apr. 2024. arXiv:2404.16308 [cs].
- [42] J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, and J. M. Larson, “Synthetic Replacements for Human Survey Data? The Perils of Large Language Models,” *Political Analysis*, vol. 32, pp. 401–416, Oct. 2024.
- [43] S. Lee, T. Q. Peng, M. H. Goldberg, S. A. Rosenthal, J. E. Kotcher, E. W. Maibach, and A. Leiserowitz, “Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias,” *PLOS Climate*, vol. 3, p. e0000429, Aug. 2024. arXiv:2311.00217 [cs].
- [44] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths, “Measuring Implicit Bias in Explicitly Unbiased Large Language Models,” 2024. Version Number: 2.
- [45] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto, “Whose Opinions Do Language Models Reflect?,” Mar. 2023. arXiv:2303.17548 [cs].

- [46] Y.-S. Chuang, A. Goyal, N. Harlalka, S. Suresh, R. Hawkins, S. Yang, D. Shah, J. Hu, and T. T. Rogers, “Simulating Opinion Dynamics with Networks of LLM-based Agents,” Apr. 2024. arXiv:2311.09618 [physics].
- [47] C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, “World Values Survey Wave 7 (2017-2022) Cross-National Data-Set,” 2024.
- [48] European Social Survey European Research Infrastructure (ESS ERIC), “ESS10 - integrated file, edition 3.2,” 2023.
- [49] M. Davern, R. Bautista, J. Freese, S. L. Morgan, and T. W. Smith, “General Social Survey 1972-2021,” 2021. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Stephen L. Morgan, and Tom W. Smith.
- [50] E. M. Uslaner, “The Moral Foundations of Trust,” *SSRN Electronic Journal*, 2002.
- [51] F. Fukuyama, *Trust: Human Nature and the Reconstitution of Social Order*. New York: Free Press, 1996. OCLC: 1000451127.
- [52] R. Layard, G. Mayraz, and S. J. Nickell, “The Marginal Utility of Income,” *SSRN Electronic Journal*, 2008.
- [53] A. Deaton and A. Heston, “Understanding pppts and ppp-based national accounts,” *American Economic Journal: Macroeconomics*, vol. 2, no. 4, pp. 1–35, 2010.
- [54] OpenAI, “GPT-5 Model | OpenAI API.” <https://platform.openai.com/docs/models/gpt-5>, 2025. [Accessed: Dec. 16, 2025].
- [55] DeepSeek, “Introducing DeepSeek-V3 | DeepSeek API Docs.” <https://api-docs.deepseek.com/news/news1226>, 2025. [Accessed: Dec. 19, 2025].
- [56] B. Rosner, *Fundamentals of biostatistics*, pp. 640–641. Cengage Learning, 8th edition ed., 2015.
- [57] W. Agnew, A. S. Bergman, J. Chien, M. Díaz, S. El-Sayed, J. Pittman, S. Mohamed, and K. R. McKee, “The illusion of artificial inclusion,” in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12, May 2024. arXiv:2401.08572 [cs].



APPENDICES

Appendix A. REPRESENTATIVENESS TEST RESULTS

To verify that the 10,000-respondent subsample used in this study reflects the demographic composition of the full WVS Wave 7 dataset (N=97,220), statistical tests were conducted across all 32 demographic variables. For categorical variables, chi-square (χ^2) tests of independence and Cramér's V effect sizes were computed. For continuous/ordinal variables (age, household size, income level), Kolmogorov-Smirnov (KS) two-sample tests were performed.

Results (Table A.1) confirm high representativeness: all categorical variables show negligible effect sizes (Cramér's V < 0.1), and all continuous variables show small differences (KS statistic < 0.1). All p-values exceed 0.05, indicating no statistically significant deviations from the full dataset. The largest observed discrepancy was for ethnicity (Cramér's V = 0.059), well within acceptable limits.

Table A.1: *Representativeness Test: 10,000-Respondent Subsample vs. Full WVS Wave 7 Dataset. All variables show negligible effect sizes (Cramér's V or KS < 0.1) and non-significant p-values (> 0.05), confirming subsample representativeness.*

Variable	Test Type	Test Statistic	p-value	Effect Size
Father Immigrant	Chi-square	0.0173	0.8954	0.0004
Income Source	Chi-square	0.0423	0.8370	0.0006
Immigrant Status	Chi-square	0.0784	0.7795	0.0009
Mother Immigrant	Chi-square	0.1707	0.6795	0.0013
Sector	Chi-square	0.1829	0.9126	0.0015
Gender	Chi-square	0.4196	0.5171	0.0020
Urban/Rural	Chi-square	0.4370	0.5086	0.0020
Citizenship	Chi-square	0.9425	0.3316	0.0030
Social Class	Chi-square	1.2356	0.8722	0.0035
Live with Parents	Chi-square	1.4618	0.6911	0.0037
Age	Kolmogorov-Smirnov	0.0046	0.9913	0.0046
Employment	Chi-square	3.3874	0.8470	0.0057
Spouse Education	Chi-square	2.4098	0.9832	0.0060
Savings Last Year	Chi-square	3.7963	0.2843	0.0060
Income Level	Kolmogorov-Smirnov	0.0069	0.7917	0.0069
Marital Status	Chi-square	5.3682	0.3726	0.0071
Spouse Employment	Chi-square	3.4717	0.9014	0.0073
Household Size	Kolmogorov-Smirnov	0.0074	0.7055	0.0074
Religion	Chi-square	6.2478	0.7149	0.0077
Father Education	Chi-square	5.6937	0.7701	0.0079
Father Occupation	Chi-square	5.9224	0.8785	0.0080
Mother Education	Chi-square	7.4944	0.5858	0.0089

Continued on next page

Table A.1 – continued from previous page

Variable	Test Type	Test Statistic	p-value	Effect Size
Education	Chi-square	8.9233	0.3488	0.0092
Occupation	Chi-square	8.6190	0.6570	0.0093
Spouse Occupation	Chi-square	9.0971	0.6129	0.0118
Children	Chi-square	19.5533	0.6110	0.0139
Country	Chi-square	53.9394	0.8345	0.0224
Country of Birth (Respondent)	Chi-square	107.8598	0.9978	0.0323
Language at Home	Chi-square	109.6884	0.9998	0.0324
Country of Birth (Father)	Chi-square	110.2527	0.9963	0.0334
Country of Birth (Mother)	Chi-square	123.9273	0.9792	0.0354
Ethnicity	Chi-square	333.0421	0.9998	0.0587

Appendix B. SAMPLE SURVEY PROMPT STRUCTURE

Answer the following survey questions realistically and consistently with the provided demographic characteristics. Provide your answers in JSON format.

You have the following attributes:

- Country of Residence: Colombia, Urban, Citizen
- Country of Birth: Colombia
- Gender: Female
- Age: 18
- Immigration Status: Non-immigrant
- Mother: Not an immigrant
- Father: Not an immigrant
- Language at Home: Spanish; Castilian
- Education: Lower secondary education
- Mother's Education: Upper secondary education
- Father's Education: No education
- Employment: Self-employed
- Occupation: Sales
- Sector: Private
- Income Level (relative, 1 is lowest and 10 is highest): 5
- Savings Last Year: Spent some savings and borrowed money
- Social Class: Lower middle class
- Religion: Do not belong to a denomination
- Marital Status: Single
- Children: 0
- Household Composition: 1 person, not living with parents
- Ethnicity: White

Now, please answer the following questions:

1. General Trust (from Q57):

"Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?"

Provide:

- * 1 for "Most people can be trusted"
- * 2 for "Need to be very careful"

2. Trust Levels for Specific Groups (from Q58-Q63):

Rate your level of trust for each group on a scale from 1 to 4, where:

- * 1 = Trust completely
- * 2 = Trust somewhat
- * 3 = Do not trust very much
- * 4 = Do not trust at all

The groups are:

- Your Family

- Your Neighborhood
- People you know personally
- People you meet for the first time
- People of another religion
- People of another nationality

Return your answers using the following JSON structure:

```
{
  "responses": {
    "generalized_trust": ,
    "trust_scores": {
      "family": ,
      "neighbors": ,
      "acquaintances": ,
      "strangers": ,
      "other_religions": ,
      "other_nationalities":
    }
  }
}
```

Ensure your responses reflect the demographic details and use the exact rating options provided above.

Appendix C. CATEGORICAL GROUPING MAPPINGS

During data preprocessing, several high-cardinality categorical variables were grouped into broader, interpretable categories to reduce dimensionality while preserving analytical value. This section documents the exact mappings used.

Table C.1: *Ethnicity Grouping Mapping. Pattern matching was performed on lowercase, normalized strings using Python regex.*

Grouped Category	Regex Patterns (matched in original ethnicity string)
Black/African	black, african, afro, negro, congolese, ethiopian, nigerian, kenyan, zimbabwean
Caucasian/White	white, caucasian, european, anglo, french, german, italian, spanish, polish, russian, ukrainian, british, irish, dutch, swedish, norwegian
Central Asian	kazakh, kyrgyz, uzbek, tajik, turkmen, mongolian
East/SE Asian	chinese, han, japanese, korean, vietnamese, thai, filipino, indonesian, malay, singaporean, taiwanese, hong kong, burmese, cambodian, laotian
Hispanic/Latino	hispanic, latino, latina, mexican, colombian, venezuelan, argentinian, chilean, peruvian, ecuadorian, bolivian, mestizo
Indigenous/Native	indigenous, native, aboriginal, maori, first nations, amerindian, inuit, quechua, aymara
Middle Eastern/West Asian	arab, arabian, middle east, persian, iranian, turkish, kurkish, armenian, georgian, azerbaijani, lebanese, syrian, iraqi, egyptian, moroccan, algerian, tunisian, libyan
South Asian	indian, pakistani, bengali, bangladeshi, sri lankan, nepali, bhutanese, tamil, punjabi, sikh, hindi
Other/Unknown	All remaining ethnicities not matching above patterns, plus explicitly missing values

Note: Patterns were applied sequentially in the order listed above. The first matching pattern determined the category assignment. Missing ethnicity values (NaN) were labeled as "Other/Unknown" after pattern matching.

Similarly language spoken at home (131 unique values) was grouped into the 7 most frequent languages in the dataset, with all remaining languages categorized as "Other":

Table C.2: *Language Grouping Mapping*

Grouped Category	Original Languages
Arabic	Arabic (all dialects)
Chinese	Chinese (Mandarin, Cantonese, etc.)
English	English
French	French
Portuguese	Portuguese
Russian	Russian
Spanish	Spanish
Other	All remaining 124 languages