Can LLMs replicate human opinion? Evaluating the ability of LLMs to mimic human answers to Real-World Questions

Team members:

Muhammadyusuf Hakimov, Malikaxon Musajonova, Husan Samandarov, Quyashbek Allanazarov, Akniet Kenzhegulov, and Mirzosharif Habibov.

Academic Mentors: Zulfiya Usmonova, Akmaljon Latifov

Date: August 14, 2025

Abstract

In the past 2-3 years, Large Language Models (LLMs) like ChatGPT, Claude, and Gemini have become increasingly intelligent by showing high results in different benchmarks, such as “Humanity’s Last Exam”. As a result, the question about their ability to replicate human opinions and the relevance of their responses arose in different discussions. This research investigates whether LLM agents can simulate genuine human opinions by comparing their responses with those of humans across several open-ended and quite controversial questions. Firstly, opinions with profiles (background information) of 100+ people are collected with an online survey. Next, the same number of agents are created with the same persona profiles and characteristics on 8 different mainstream models. Through multi-agent simulation, agents are given the same open-ended questions from the survey, and their responses are recorded. After all responses are collected, they are then compared with each other (human vs LLM) using data analysis methods, and infographics are created for better data visualization. By examining models’ responses, this study assesses the reliability of AI as a human opinion generation tool.

Keywords: Large Language Models (LLM), LLM agents, data visualization, multi-agent simulation, human opinion generation.

Introduction

The aim of this research is to investigate the differences between LLMs' responses and human responses to certain questions. It will identify the key factors affecting the responses and determine how relevant the responses of generative AI models are. LLM agents are the most important tools needed in our simulation as we collect LLM-generated public opinions. Multi-agent systems (MAS), a group of LLM agents, are used in simulations to model scenarios where multiple autonomous entities interact, communicate, and make decisions. In our simulation, people or groups of different people are represented by agents. In these systems, each agent operates independently within the given environment. They have their own decision-making capabilities, logic, and perception.

We formulated the following research questions:

- RQ1: Can LLMs accurately replicate human survey responses when conditioned on persona data?
- RQ2: How similar are LLM-generated responses to real human responses to the same set of questions?
- RQ3: How relevant are AI chatbots' (ChatGPT, Claude, Gemini) answers compared to humans? Is asking their personal opinions the right thing to do?

By conducting this research, we aim to compare how well LLM-generated public opinion approximates real human responses on 5 controversial or socially relevant questions. This research will follow an approach combining human survey data collection with LLM-based agent simulations.

Methodology

Research Design

This study employed a comparative experimental design to evaluate the accuracy of Large Language Models (LLMs) in simulating human opinions across different contextual conditions. The research was structured around three distinct experiments designed to assess: (1) personalized opinion prediction using a person's background data, (2) cultural context modeling using regional personas, and (3) baseline performance without contextual information.

Participants and Data Collection

A total of 108 participants from Uzbekistan were recruited for the study. Participants were selected to represent the local demographic and cultural context relevant to the research objectives(e.g, gender, area of residence, political views, profession). All participants provided informed consent and completed a structured questionnaire containing five opinion statements on socio-cultural topics. The full questions can be found in Appendix A.

Opinion Questions

Five fixed opinion statements were developed covering diverse socio-cultural domains:

- **Q1**: Environmental responsibility - *"Every person should be directly responsible for protecting the environment in their daily life by reducing waste, recycling, and conserving resources."*
- **Q2**: Child development - *"Children need to spend more time at home in front of screens for activities such as learning, using technology, and doing homework. They should spend less time in the outdoor environment."*
- **Q3**: Social media impact - *"Social media causes more harm than benefit to society due to exposing personal life, spreading false information, and its impact on mental health."*
- **Q4**: Gender equality - *"Female athletes should receive equal pay to male athletes when their work, level, and performance in sports are the same (that is, equal pay should be given to male and female athletes for equal work)."*
- **Q5**: Cultural practices - *"A daughter is ready for marriage when she reaches the age of 19."*

Participants responded using a three-point scale: "Agree," "Disagree," or "Neutral."

Experimental Procedure

The experiment involved evaluating each leading LLM model individually by giving them the same set of challenging and controversial questions as used in the human survey. Each model was asked to provide its response in the form of a vote (Agree, Disagree, or Neutral) along with an explanation of their reasoning. These responses were then compared with a dataset of human responses collected through an online survey.

LLM testing procedure

For the 8 LLMs, the following steps have been taken:

1) Choosing questions from the given storage of questions
2) Passing the questions through the API to the correct model and getting the voting of agents and their thinking
3) For the collected response, compare the aggregated agrees with other human survey statistical agreements

LLM profiling

The LLM models shown in Table 1 have been used in our experiment.

Model	Params	Provider
DeepSeek-R1-Distill-Llama	70b	Groq
Gemini-2.5-flash	-	Google
Llama-4-scout	17b	Groq
Llama-3.3-versatile	70b	Groq
Gemma-3	4b	Ollama
GPT-4	-	OpenAI
GPT-3.5-Turbo	-	OpenAI
GPT-4.1-mini-2025-04-14	-	OpenAI
Allam-2	7b	Groq

Table 1: List of LLM models and their that were used in the experiment.

Evaluation metrics

Metric	Explanation	Range of possible values	Why this metric
Accuracy	Proportion of correct predictions.	0 to 1 (or 0% to 100%)	To measure the correctness of LLM predictions against human responses.
Cohen's Kappa (κ)	Chance-corrected measure of inter-rater agreement.	-1 to 1	To measure the agreement between LLM and human responses, accounting for chance agreement.
Bias (Mean Bias)	Systematic deviation between LLM and human agreement rates.	Any real number (positive or negative)	To identify if LLMs systematically over- or under-agree with human responses on certain topics.
Pearson Correlation (r)	Linear relationship between LLM and human responses.	-1 to 1	To assess the strength and direction of the linear relationship between LLM and human response patterns.
Response Diversity	Variation in response patterns, often measured via entropy.	0 to 1 (or 0% to 100%)	To understand how varied LLM responses are compared to the diversity in human opinions.
Overall Agreement Rate	Fraction of total responses where LLMs matched humans.	0 to 1 (or 0% to 100%)	To provide a general measure of how often LLMs align with human responses.
Exact Match	Percentage of participants for whom all five responses were predicted correctly.	0 to 1 (or 0% to 100%)	To determine the proportion of instances where LLMs perfectly replicate a participant's set of

			opinions.
Persona/Background Effect	Improvement in agreement when contextual information (persona or background data) is provided.	Any real number (positive or negative)	To quantify the impact of providing personalized or contextual data on LLM agreement with human opinions.
Confusion Matrix	Tabular breakdown of correct vs. incorrect predictions across categories.	N/A (a table, not a single value)	To provide a detailed view of how LLMs classify responses (e.g., correctly agreeing, incorrectly disagreeing).

Table 2: List of the evaluation metrics we have used to analyze the results.

Caching and randomness

During experiments 2 and 3, we made sure to set the temperature of LLMs to 0 to exclude randomness as much as we could. We initiated a new API call every time the response from the agent was required. We carefully read the documentation provided by the developers of each model, and ensured that no internal caching within the provider and model processing units is taking place. For experiment 1, we have used a temperature of 0.3 to give LLMs some creativity to correctly and accurately imitate human opinion.

Reproducibility provisions

We will share the materials, the report, and source codes (without API keys) in a GitHub repository to make reproducibility possible and accurate.

Experiments

In each experiment, by topics, the following five question-based topics are considered:

```
topics = [
    "Protecting the environment is the individual's responsibility.",
      "Do children need to spend more time outdoors rather than indoors with a
screen? (1-2)",
    "Does social media do more harm than good? ",
    "Do you think women athletes have to be paid equally to men?",
    "Do you think 19 years is an appropriate age for marriage for a girl?"
]
```

Experiment 1: Personalized Opinion Prediction

Objective: Evaluate LLM accuracy in predicting individual human opinions using personal background data.

Procedure: Four mainstream LLMs were tested:

- DeepSeek R1
- Allam (IBM)
- Meta Llama
- OpenAI GPT-4.1-mini

Each model received structured background information for individual participants, including demographic data, profession, political views, and personality. Models were instructed to predict the specific individual's responses to all five opinion questions.

The system prompt given to the LLMs:

```
You are simulating the authentic opinions of a specific individual from
{location}, Uzbekistan. CRITICAL: Your responses must reflect the unique
combination of the person's background characteristics provided. Consider how
their age, gender, situation, education,  and personal experiences would
genuinely shape their worldview and opinions.

Process:

1. First, internally consider how each background factor influences the
person's worldview

2. Then, predict their response to each question based on this analysis

3. Finally, output ONLY the JSON response
```

```
Output format: {"q1": "Answer", "q2": "Answer", "q3": "Answer", "q4": "Answer", "q5": "Answer"}

Values must be exactly: "Agree", "Disagree", or "Neutral"

Respond ONLY with a valid JSON object using exactly this format.

Use only "Agree", "Disagree", or "Neutral" as values. No explanations.
```

Experiment 2: Persona-Based Opinion Simulation

*Objective***:** Evaluate LLM responses when instructed to simulate an Uzbek persona without background data.

*Procedure***:** Eight mainstream LLMs were tested:

- DeepSeek R1
- Meta Llama 4 scout
- Meta Llama 33 versatile
- Mistral Saba
- Ollama Gemma 3
- Google Gemini 1.5 Flash
- OpenAI GPT 4
- OpenAI GPT 3.5 Turbo

No structured data was provided. Each model was asked to assume the perspective of a person from Uzbekistan and answer all five opinion questions.

The system prompt given to the LLMs:

```
You are simulating an average adult citizen from Uzbekistan.

Adopt the values, worldview, and cultural context typical of someone from Uzbekistan.

Think step-by-step. Output your reasoning inside <think>...</think> tags.

After that, state only one of these words: 'agree', 'disagree', or 'neutral'.

Please consider the following statement: {topic}
```

Experiment 3: Baseline LLM Opinions

Objective: Assess the models' own unconditioned opinions and compare them across systems and with Experiments 1 and 2.

Procedure: Eight mainstream LLMs were tested:

- DeepSeek R1
- Meta Llama 4 scout
- Meta Llama 33 versatile
- Mistral Saba
- Ollama Gemma 3
- Google Gemini 1.5 Flash
- OpenAI GPT 4
- OpenAI GPT 3.5 Turbo

No structured data or country-specific framing was provided. Models were simply asked to give their direct opinions on all five questions.

Results and discussion

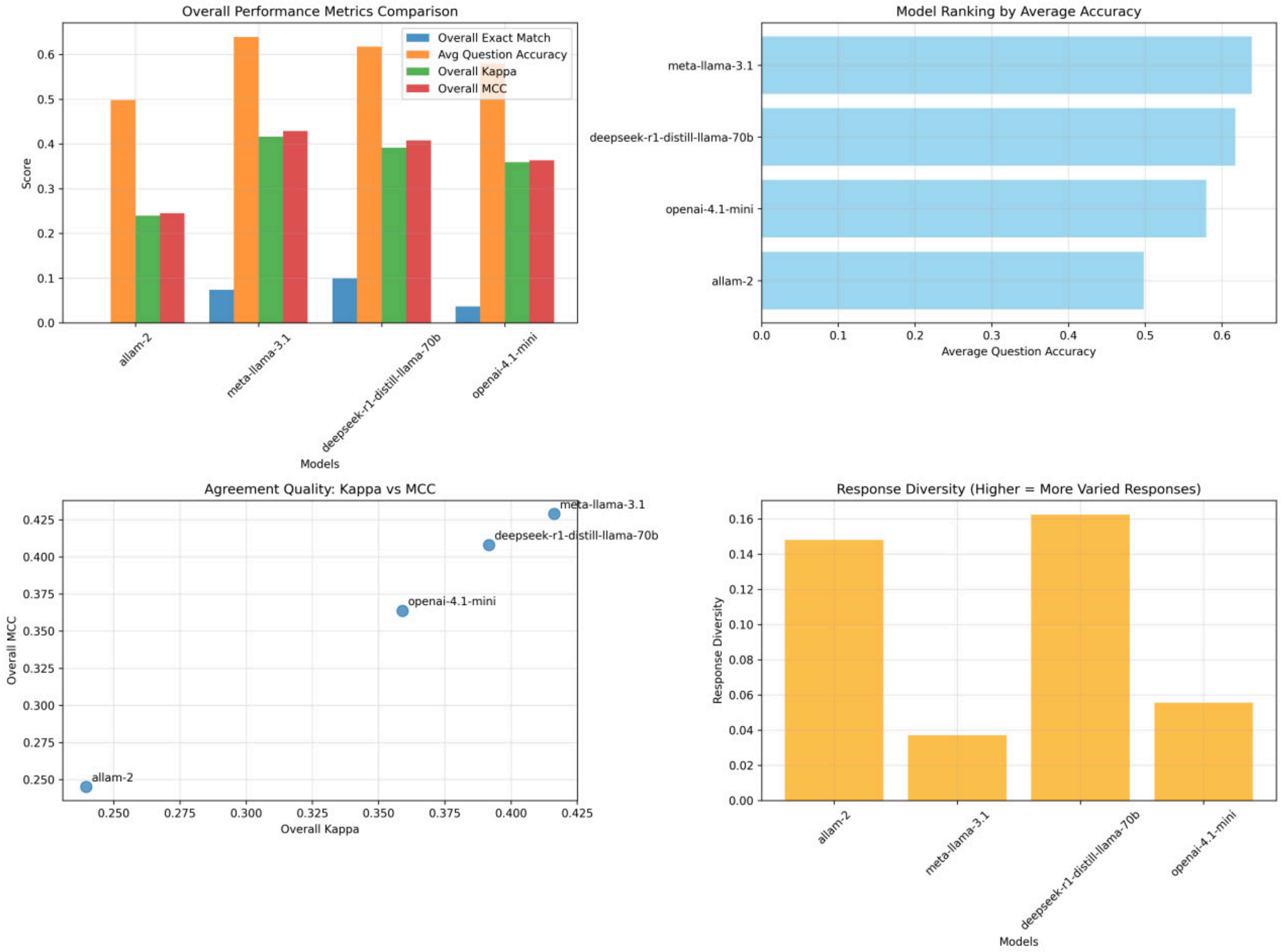


Figure 1: Comprehensive performance analysis of LLM models in Experiment 1 (Personalized Opinion Prediction). The dashboard shows (top left) overall performance metrics across models, (top right) average question accuracy, (bottom left) agreement quality, and (bottom right) response diversity for DeepSeek-R1, Allam-2, Llama-3.3, and GPT-4.1-mini.

Figure 1 presents a comprehensive analysis of four LLM models in Experiment 1 (Personalized Opinion Prediction). Performance varied across metrics:

- **Average Question Accuracy** ranged from 50% (Allam-2) to 63% (Meta-Llama-3.1), with DeepSeek-R1-Distill-Llama-70b and OpenAI GPT-4.1-mini scoring in between.
- **Cohen's Kappa** values fell between 0.25 (Allam-2) and 0.42 (Meta-Llama-3.1), indicating moderate agreement with human responses.

- **MCC (Matthews Correlation Coefficient)** closely mirrored Kappa, with the strongest performance again from Meta-Llama-3.1 and DeepSeek-R1, both exceeding 0.40, while Allam-2 remained weakest at around 0.25.
- **Response Diversity** analysis showed notable variation: DeepSeek-R1 (≈0.16) and Allam-2 (≈0.15) produced more varied outputs, while Meta-Llama-3.1 was least diverse (≈0.04), reflecting more uniform predictions. GPT-4.1-mini showed moderate diversity (≈0.06).

Overall, the results indicate that while Meta-Llama-3.1 was most accurate and consistent, DeepSeek-R1 balanced agreement with higher diversity, suggesting it captured more of the heterogeneity present in human opinions. In contrast, Allam-2 underperformed across accuracy and agreement despite producing more varied responses.

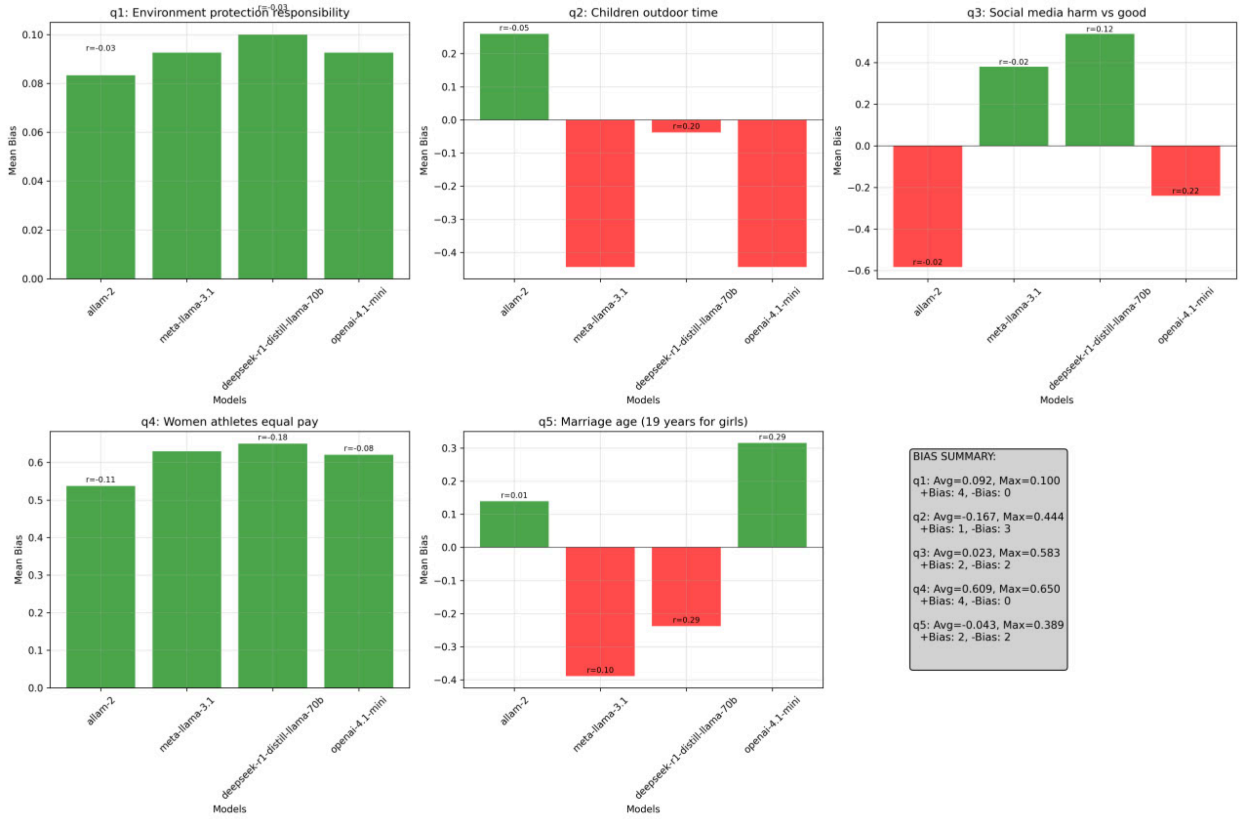


Figure 2: Question-specific bias analysis for each LLM model in Experiment 1. Each subplot represents bias patterns for questions Q1-Q5, showing how different models deviated from human response patterns across environmental responsibility, child development, social media impact, gender equality, and cultural practices topics.

Figure 2 illustrates the question-specific bias patterns for four LLM models in Experiment 1. The results highlight significant variation depending on the topic:

- For **Q1 (environmental responsibility)**, all models showed only slight positive bias (0.08–0.10), indicating *strong alignment* with human responses.
- For **Q2 (children's outdoor time)**, the largest divergence occurred: Allam-2 showed a mild positive bias (+0.05), while Meta-Llama-3.1 (–0.20), DeepSeek-R1 (–0.44), and GPT-4.1-mini (–0.39) displayed strong negative bias, suggesting they were *more critical* than humans about excessive screen time.
- For **Q3 (social media impact)**, bias values ranged widely: DeepSeek-R1 (+0.12) and Meta-Llama-3.1 (+0.02) leaned toward stronger agreement with harm, while GPT-4.1-mini (–0.20) and Allam-2 (–0.02) showed weaker or negative bias. This indicates *inconsistent handling of neutrality* across models.
- For **Q4 (gender pay equality)**, all models produced strong positive bias, ranging from +0.11 (Allam-2) to +0.65 (Meta-Llama-3.1). This was the most uniform bias across models, revealing a clear *tendency toward progressive responses* that exceeded human agreement levels.
- For **Q5 (marriage age at 19)**, divergence was again visible: Allam-2 remained nearly neutral (+0.01), while GPT-4.1-mini showed a strong positive bias (+0.29). In contrast, Meta-Llama-3.1 (–0.10) and DeepSeek-R1 (–0.29) leaned negatively, indicating l*ess agreement* with the statement than humans.

Overall, the *strongest bias* was found in Q4 (gender equality, max +0.65), while Q1 (environmental responsibility) showed *the lowest bias*. These results suggest that LLMs consistently over-endorse socially progressive positions, particularly regarding gender equality, while diverging more variably on culturally sensitive topics like child development and marriage age.

Figure 3 presents a detailed comparison of LLM and human responses across all experimental conditions. From the top left graph in Figure 3, **Accuracy** varied by question: Q1 (environmental responsibility) achieved the highest accuracy (~93%), while Q3 (social media impact) was lowest (~47%). From the top middle graph, **Bias patterns** reveal that LLMs consistently over-agreed on Q3 and Q4, with Q4 (gender equality) showing the strongest positive bias (+0.63), reflecting a stronger endorsement of equal pay than human participants. In contrast, Q2 (children's screen time) and Q5 (marriage age at 19) displayed negative bias, as models disagreed more strongly than humans.

From the top right graph, **Correlation Analysis** showed near-zero values for most questions, with only Q5 reaching a modest positive correlation (~0.1). From the bottom left graph, **Response Distribution** results highlight that for Q1, both groups mostly agreed (LLMs ~100%, Humans ~93%), but LLMs lacked the variation seen in human answers. From the bottom middle of Figure 3, the **Confusion Matrix** for Q1 confirmed this uniformity, with nearly all predictions correctly classified as "Agree."

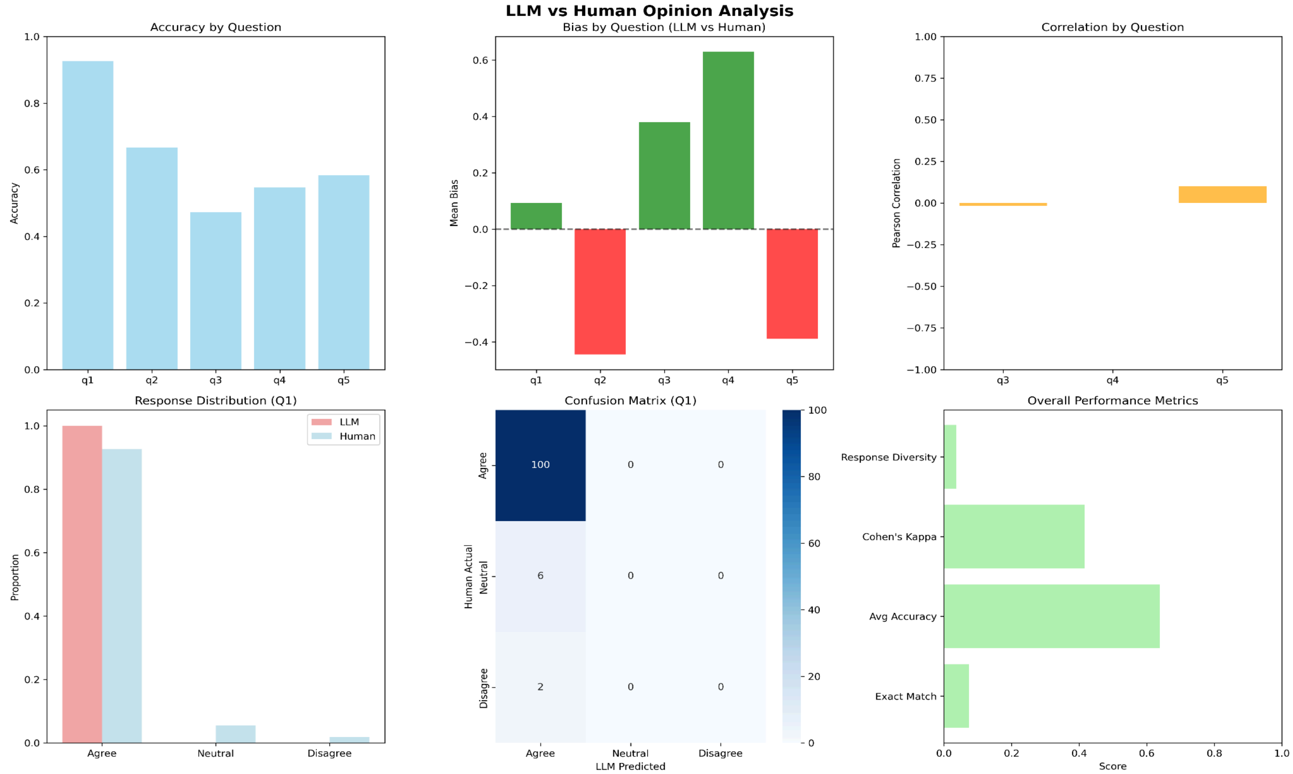
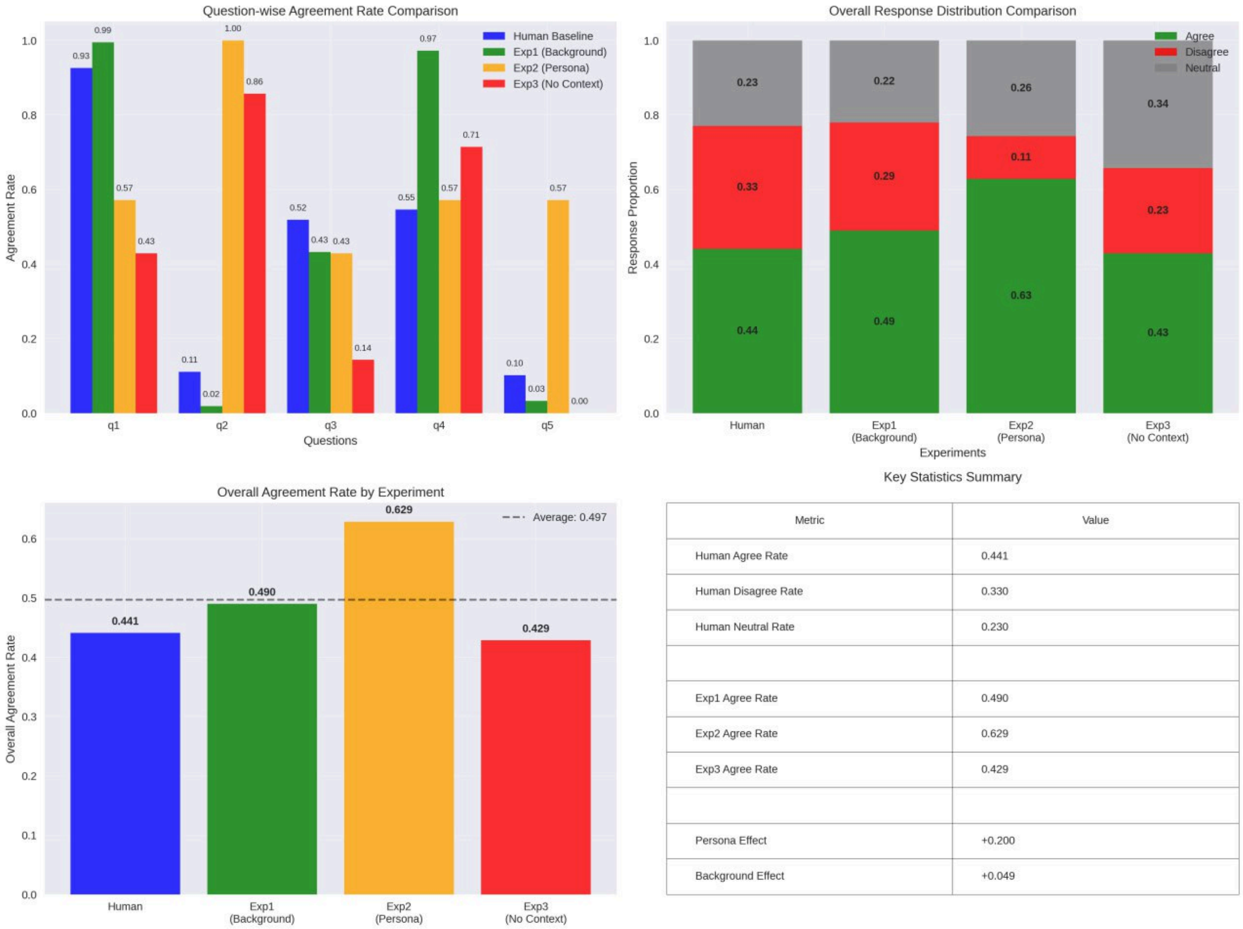


Figure 3: Detailed performance analysis of the top-performing model (Llama-Guard-4-12b) in Experiment 1, including accuracy metrics, response distribution comparison with human data, confusion matrix, and overall performance breakdown.

Overall performance metrics placed LLMs at ~65% average accuracy and Cohen's Kappa around 0.4, indicating moderate alignment with human opinions but reduced response diversity compared to the human sample.

Figure 4 provides a cross-experimental comparison of human and LLM responses under different contextual conditions. From top left in Figure 4, **Agreement patterns** varied by question. Q1 (environmental responsibility) showed consistently high alignment, with LLM agreement at 97–99% across experiments versus 93% for humans. Q2 (children's screen time) reached perfect agreement in Exp2 (100%), though rates were lower in Exp1 and Exp3, while humans showed only 11% agreement.

The largest discrepancies emerged in Q3 (social media impact) and Q4 (gender equality). For Q3, human agreement was 52%, while LLMs ranged from 43–55%. For Q4, humans agreed at 55%, but LLMs varied more widely (Exp1: 43%, Exp2: 57%, Exp3: 71%), showing stronger progressive bias in some conditions. Q5 (marriage age) showed uniformly low agreement across all experiments, closely reflecting human patterns.



Metric	Value
Human Agree Rate	0.441
Human Disagree Rate	0.330
Human Neutral Rate	0.230
Exp1 Agree Rate	0.490
Exp2 Agree Rate	0.629
Exp3 Agree Rate	0.429
Persona Effect	+0.200
Background Effect	+0.049

Figure 4: Cross-experimental comparison of LLM performance against human responses. The analysis includes (a) question-wise agreement rates across all three experiments, (b) response distribution patterns for each experimental condition, (c) overall agreement rates by experiment type, and (d) comprehensive performance metrics comparing personalized prediction, persona-based simulation, and baseline opinions.

From the top right graph in Figure 4, **Response Distribution** analysis indicated that Exp2 (persona-based) achieved the highest overall agreement (62.9%), followed by Exp1 (background data) at 49.0% and Exp3 (no context) at 42.9%, compared with the human baseline of 44.1%. Statistical analysis showed that persona simulation contributed +0.200 to agreement rates, while background data added +0.049, confirming that cultural persona framing improved alignment more effectively than individualized background information.

Overall, the findings demonstrate that LLMs approximate human responses more closely when provided with contextual framing, with persona-based simulation producing the strongest match to human opinion distributions. Performance, however, remained question-dependent, with greater divergence on socially divisive topics such as gender equality and social media impact.

Conclusion

Human subjects and large language model (LLM) agents were compared on five opinion-based statements. Analysis was intended to identify patterns of agreement, neutrality, and disagreement as well as similarities and differences between the two groups.

No	Question	Results
1	Protecting the environment	Nearly all participants agreed that it is an individual duty to protect the environment. LLM agents also shared the same opinion, but with a bit higher consistency of agreement.
2	Children staying indoors on screen for longer periods	Both responders mainly disagreed with the idea that children spend too much time in front of screens. Even though there were some human disagreements, LLM agents never did, which indicates a bit stronger disagreement from LLM responses.
3	Social media does more harm than good	Human responses were more towards agreeing, but a large percentage was neutral. LLM agents almost universally agreed and were less variable.
4	Female athletes need to be paid equally	Here was the greatest wide gap. People were divided, with more of them in disagreement or neutral. LLM agents agreed consistently, which reflected a more progressive and equality-based attitude than the human sample.
5	A girl is ready to marry at 19	Both human and LLM respondents strongly disagreed the most with this statement. The agreement here was highest, with minimal variation in responses.

Table 3: List of controversial questions and final results' discussion.

In conclusion, LLM responses tend to be more uniform and socially progressive, whereas human opinions display greater diversity. The highest similarity appeared in topics related to environmental responsibility, children's screen time, and marriage age, while the largest difference was seen in attitudes toward gender pay equality in sport.

Based on this experiment's results, we can say that LLM agents can closely replicate human opinions for the most part, particularly about generally accepted social values or ethical issues. Because they tend to generalize and be consistent, they may not necessarily understand diverse real-world human opinions.

References