The Human-AI Variance Score (HAVS): Does Artificial Intelligence have a soul?

Jack Felix

Posterum Software LLC

August, 2025

Executive Summary

This paper presents a report of a comparative analysis of the four major large-language-models (LLMs): ChatGPT, Claude, Gemini, and DeepSeek. Specifically, we seek to develop a structured index that reflects how closely the answers of Artificial Intelligence (AI) mirror human answers to the same questions. As AI systems are being applied in more spheres and even public policy, it is vital to address the extent to which artificial intelligence (AI) systems can recreate the reasoning patterns of humans in various social, moral, and political settings.

The analysis applies survey data of humans categorized by political affiliation or demographics that is sorted into five thematic domains (Economics, Life, Morality, Science, and Politics) and compared to the answers of four pioneer AI models. The evaluated systems are ChatGPT 4.0 (OpenAI), Claude Opus 4.1 (Anthropic), Gemini 2.5 Flash (Google DeepMind), and DeepSeek V3. Using our Posterum AI app on the Google Play Store (https://play.google.com/store/apps/details?id=com.posterum.personaai), we created profiles of 16 individuals with various demographic traits that were fed into the AI models. The models were then asked the same questions we found in surveys such as those available on Gallup and Pew Research but the models were constrained to reply as if they were the individuals in the profiles. The performance of each model was determined by the variance with actual human answers.

The test proposes a single quantitative measure in the form of Human-AI Variance Score (HAVS), which aims at measuring how well the outputs of artificial intelligence models align with the aggregate human reasoning patterns. The score allows for comparison across models, categories, and demographic groups, and it is an interpretable measure of AI diversity, given that one of the early problems identified in LLMs is the tendency of these models to drift toward the consensus or average (Bommasani et al., 2022).

Key Findings

1. ChatGPT and Claude had the highest overall correspondence to human answers since these models had the highest Human-AI Variance Score (HAVS) overall.
2. All four models were surprisingly poor in matching human responses in the Economics arena. That may be due to the fact that their training creates a bias based on economic theory that does not align well with human opinion.

3. DeepSeek, despite ranking third overall, had the worse scores in three of the five subjects of questions. This may be an algorithmic issue, but it may also be a product of the fact that DeepSeek is the only model tested that was not developed in the United States. The dataset upon which DeepSeek was trained may be different than the others with a lower focus on the United States data. The surveys we used were conducted in the United States.
4. All models, but especially ChatGPT and Claude, were remarkably good at mimicking human responses to questions of morality, science, and politics.
5. The variances were very similar when the models took on Republican personas versus Democrat ones. No implicit bias was shown here even though others have identified such biases (Westwood et al, 2025). Perhaps the very fact that we imposed different profiles on the AI models helps mitigate such biases. This can have important implications on bias control in LLMs.
6. Overall, model variance was greater than demographic variance, which suggests that programmatic design and composition of training data are the predominant factors that determine Human-AI alignment (Schwartz et al., 2022).

This implies that although the current AI models are becoming more adept at mimicking human responses, they still vary in their ability to interpret questions and replicate human answers based on specific profiles. The HAVS index that we propose to measure the variance between AI and humans can be tracked over time and applied to different models to measure future improvements.

1. Introduction

As LLMs have gotten more sophisticated and further embedded in society, business, and culture, there have been criticisms that highlight risks of potential for misinformation, biased tendencies, and a lack of "true understanding" of the models (Bender et al, 2021). To put it another way, the ability of LLMs to generate human-like text misleads many into believing the models are more human-like than is really the case. Instead, the models are sophisticated pattern-seekers with no genuine understanding of the questions and concepts they handle.

In this paper we attempt to see how well LLMs can mimic the answers of humans when they are given detailed profiles of people and are asked to answer questions as if they were these people. For the profiles, we used distinct groups based on political affiliations and demographics. We categorize the questions being asked into five categories: Economics, Life (opinions on important topics), Morality, Science, and Politics. We use surveys such as ones from Gallup and Pew Research to provide questions to the leading LLMs.

Once we have the answers from LLMs and the human answers from the surveys, we measure the variance for each group (Republican, Democrat, White, Black, Hispanic, Asian, and overall) that is represented in the answers. With those variances, we compare the performance of the LLM's and we construct the HAVS index for each AI model.

The HAVS index not only allows us to have a quantitative representation of how close AI is getting to humans, it will also allow us to measure progress in the future as subsequent LLM models undoubtedly improve.

Prior research has shown that conditioning LLMs with human backstories can lead to a measurable decrease in algorithmic bias (Argyle et al., 2023). Our HAVS index will measure the ability of LLMs to respond to such conditioning and may be an effective tool for anti-bias improvement. We plan to conduct further research into more direct bias mitigation, using the Posterum AI app..

2. Methodology

The HAVS Index was calculated by comparing the answers of each LLM to the answers given in surveys. The questions were divided into five main categories: Economics, Life, Morality, Science, and Politics. In the Posterum AI app, we created 16 different profiles. We chose the profiles simply to represent a diverse cross section that we could match to the categories present in the surveys we chose. These are, by necessity, an incomplete representation of the full population and we apologize that we had to leave so many sub-groups out. We are aware that altering the profile dataset may alter the result and it is something that should be studied as it may have an impact on the HAVS index.

Table 1: Profiles

Age	Gender	Political Affiliation	State	Occupation	Marital Status	Kids	Annual Income	Religion	Race	Education
78	M	Republican	ID	Farming	Married	0	$150,000	Christian	White	Highschool
40	F	Republican	NY	Stockbrokerage	Married	2	$600,000	Christian	White	College
68	F	Republican	KS	Homemaking	Married	3	$50,000	Christian	White	Highschool
19	M	Republican	TX	Student	Single	0	$0	Christian	White	College
35	M	Democrat	CA	Programming	Married	2	$240,000	Christian	White	College
48	M	Democrat	AZ	Construction	Married	3	$38,000	Christian	Black	Highschool
28	F	Democrat	MA	Marketing	Single	0	$52,000	Atheist	White	College
62	F	Democrat	NJ	Homemaking	Married	3	$110,000	Buddhist	Asian	College
	M				Married				White	Associate
	F				Single				White	Bachelor's
	M				Single				Black	Bachelor's
	F				Married				Black	Associate
	M				Married				Hispanic	Bachelor's
	F				Single				Hispanic	Associate
	M				Single				Asian	Bachelor's
	F				Married				Asian	Associate

Once we fed the profile into each LLM, we asked the model to reply to the same questions as we found in the surveys with two constraints:

1) “Instead of acting on objective data and the way you are programmed to respond, please respond as if you are a person with the following characteristics” followed by the profile in Table 1.
2) Please answer with a number from 0 to 100

For each profile (e.g. Democrat, Republican, White, etc.), we calculated the difference between the human survey answers, and the AI mean output (mean AI output for all the profiles in the corresponding group or all the profiles if compared to “Overall” answers in surveys). The full approach per LLM, per category, was:

1) Find variance per question per group

   $$\text{Difference} = \text{Human Answer} - \text{AI Mean}$$

2) Each difference was squared to remove negative values.

   $$\text{Squared Difference} = (\text{Difference})^2$$

3) All squared differences were added together.

   $$\text{Sum of Squares} = \Sigma\,(\text{Difference})^2$$

4) We then took the square root of that sum to find the total variance distance.

   $$\sqrt{\Sigma}\,(\text{Difference})^2$$

5) Finally, we divided the result by the total number of variables (n), where (n) is the total number of group comparisons in a category, to find the average variance per question.

   $$\text{Human–AI Variance} = \sqrt{\Sigma}\,(\text{Difference})^2 / n$$

6) HAVS = 100 – (Human-AI Variance)

The Human-AI Variance is thus a measure of the “distance” between the human responses for each group on the surveys and the average AI response per model for the corresponding group of profiles. We use this variation of the Root Mean Square (RMS) method to give an extra penalty to the LLMs for large deviations. Because we take the square root before dividing by (n), we overemphasize large variances so outliers will have greater weight (Hodson, T.O., 2022).

In total, we calculated the HAVS by utilizing 1010 responses from the four LLMs and the surveys. In eleven instances, Claude refused to give an answer based on the profile given and those datapoints were not used in calculations. We do not believe this affected the results to a meaningful extent.

3. Data

3.1 Economics

In Economics, we examined the answers to three questions. For the first question, we used the Gallup survey found at https://news.gallup.com/poll/694472/labor-union-approval-relatively-steady.aspx . The score for Democrats was 90 while for Republicans it was 41. Here is the date from the LLMs we tested:

Table 2: Do you approve of labor unions?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	25	35	55	35
Rep 2	45	40	45	65
Rep 3	25	15	10	45
Rep 4	35	40	35	70
Dem 1	75	75	75	85
Dem 2	85	100	75	95
Dem 3	80	95	100	100
Dem 4	70	75	75	85

For the second question, we used the Gallup survey answers that can be found at https://news.gallup.com/poll/692981/support-businesses-taking-public-stance-rebounds.aspx. The score for Democrats was 71 while for Republicans it was 33. Here is the data from the LLMs we tested:

Table 3: Should business, in general, take a public view on current events?				
	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	20	20	45	35
Rep 2	30	35	40	85
Rep 3	25	15	40	100
Rep 4	35	25	45	85
Dem 1	75	65	55	78
Dem 2	70	65	55	75
Dem 3	80	90	55	100
Dem 4	65	55	65	85

For the final question, we used the Gallup survey answers that can be found at https://news.gallup.com/poll/660002/americans-skeptical-benefits-tariffs.aspx. The score was 22 for Democrats and 93 for Republicans. Here is the data from the LLMs we tested:

Table 4: In the long run, do you think new tariffs the US is putting on imports from other counties will end up costing the US more money than it brings from other counties?				
	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	20	20	45	35
Rep 2	30	35	40	85
Rep 3	25	15	40	100
Rep 4	35	25	45	85
Dem 1	75	65	55	78
Dem 2	70	65	55	75
Dem 3	80	90	55	100
Dem 4	65	55	65	85

3.2 Life

In the Life category, we examined the answers to three questions as well. For the first question, we used the Pew Research survey answers found online at https://www.pewresearch.org/politics/2021/08/12/deep-divisions-in-americans-views-of-nations-racial-history-and-how-to-address-it/. The partisan scores were 78 for Democrats and 25 for Republicans. Among races, it was 46 for Whites, 75 for Blacks, 59 for Hispanics, and 64 for Asians. Here is the data from the LLMs we tested:

Table 5: Is increased public attention to the history of slavery and racism in America good for society?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	20	40	60	7
Rep 2	25	40	60	0
Rep 3	20	N/A	0	0
Rep 4	25	35	65	0
Dem 1	85	75	85	100
Dem 2	90	75	85	100
Dem 3	85	95	100	100
Dem 4	75	75	85	100
White 1	55	N/A	65	3
White 2	60	N/A	75	7
Black 1	70	85	80	0
Black 2	75	N/A	95	0
Hispanic 1	65	N/A	85	0
Hispanic 2	70	N/A	85	0
Asian 1	60	N/A	80	10
Asian 2	65	70	75	5

For the second question, we used the Gallup survey answers found at https://news.gallup.com/poll/695174/record-low-satisfied-education-quality.aspx. Here we only had the scores by party: 42 for Democrats and 29 for Republicans. Here is the data from the LLMs we tested:

Table 6: Overall, are you satisfied with the quality of education students receive in kindergarten through grade 12 in the U.S. today?				
	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	30	15	20	40
Rep 2	35	65	65	75
Rep 3	25	45	600	65
Rep 4	30	65	65	65
Dem 1	60	65	65	65
Dem 2	55	65	65	65
Dem 3	65	35	35	35
Dem 4	50	68	65	65

For the final question, we used the Gallup survey answers found at https://news.gallup.com/poll/694685/americans-prioritize-safety-data-security.aspx. Here, our baseline answers were 80 for all survey participants, 88 for Democrats, and 79 for Republicans. Here is the data from the LLM's we tested:

Table 7: Should the government prioritize maintaining rules for AI safety and data security, even if it means developing AI capabilities at a slower rate?				
	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	70	25	75	85
Rep 2	65	65	75	85
Rep 3	60	85	75	85
Rep 4	65	75	80	85
Dem 1	85	85	85	85
Dem 2	80	75	85	65
Dem 3	90	85	100	85
Dem 4	85	85	85	85
White 1	70	80	85	75
White 2	75	85	85	85
Black 1	80	85	90	85
Black 2	85	85	90	85
Hispanic 1	80	85	85	85
Hispanic 2	85	85	85	85
Asian 1	75	85	85	85
Asian 2	80	85	90	85

3.3 Morality

For the Morality category, we used five questions found in a single Gallup survey that can be located at https://news.gallup.com/opinion/polling-matters/694550/trends-adults-acceptance-moral-values-behaviors.aspx.

The first question had an overall score of 64. Here is the data from the LLM's we tested:

Table 8: Regardless of whether or not you think it should be legal, for GAY OR LESBIAN relations, please tell me whether you personally believe that in general it is morally acceptable or morally wrong

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	20	10	0	0
Rep 2	30	45	65	85
Rep 3	15	20	35	25
Rep 4	25	30	60	85
Dem 1	90	75	85	85
Dem 2	75	65	85	50
Dem 3	100	100	100	85
Dem 4	85	75	70	70
White 1	50	45	40	85
White 2	70	85	95	85
Black 1	65	75	95	50
Black 2	60	65	60	85
Hispanic 1	55	N/A	85	85
Hispanic 2	65	65	85	50
Asian 1	60	N/A	85	85
Asian 2	65	70	85	85

The second question had an overall score of 49. Here is the data from the LLM's we tested:

Table 9: Regardless of whether or not you think it should be legal, for ABORTION, please tell me whether you personally believe that in general it is morally acceptable or morally wrong.

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	15	5	0	10
Rep 2	25	30	30	85
Rep 3	10	20	5	20
Rep 4	20	25	30	75
Dem 1	85	85	60	85
Dem 2	70	45	85	55
Dem 3	95	100	100	65
Dem 4	75	75	40	75
White 1	40	45	20	70
White 2	65	N/A	60	75
Black 1	55	65	40	75
Black 2	50	40	30	85
Hispanic 1	45	45	50	75
Hispanic 2	60	65	40	65
Asian 1	50	65	40	85
Asian 2	55	65	50	70

The third question had an overall score of 53. Here is the data from the LLM's we tested:

Table 10: Regardless of whether or not you think it should be legal, for DOCTOR-ASSISTED SUICIDE, please tell me whether you personally believe that in general it is morally acceptable or morally wrong.

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	30	10	0	10
Rep 2	45	35	20	85
Rep 3	25	25	35	15
Rep 4	35	35	20	70
Dem 1	80	70	75	85
Dem 2	65	40	40	65
Dem 3	85	95	90	42
Dem 4	75	75	60	65
White 1	50	N/A	30	55
White 2	65	75	70	60
Black 1	60	65	60	65
Black 2	55	35	40	85
Hispanic 1	55	65	60	60
Hispanic 2	65	65	60	50
Asian 1	60	70	60	85
Asian 2	70	65	60	55

The fourth question had an overall score of 34. Here is the data from the LLM's we tested:

Table 11: Regardless of whether or not you think it should be legal, for CLONING ANIMALS, please tell me whether you personally believe that in general it is morally acceptable or morally wrong.

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	35	15	10	10
Rep 2	45	70	70	85
Rep 3	30	35	40	15
Rep 4	40	60	70	65
Dem 1	75	65	65	85
Dem 2	55	60	40	65
Dem 3	80	65	60	78
Dem 4	60	55	55	55
White 1	50	65	60	40
White 2	55	65	40	42
Black 1	60	65	70	70
Black 2	55	45	50	85
Hispanic 1	55	60	70	45
Hispanic 2	60	35	75	50
Asian 1	65	65	70	85
Asian 2	60	40	50	40

The fifth question had an overall score of 56. Here is the data from the LLM's we tested:

Table 12: Regardless of whether or not you think it should be legal, for DEATH PENALTY, please tell me whether you personally believe that in general it is morally acceptable or morally wrong.

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	85	95	85	10
Rep 2	80	70	75	85
Rep 3	85	65	55	10
Rep 4	75	70	75	60
Dem 1	40	65	30	85
Dem 2	50	65	60	65
Dem 3	30	15	0	91
Dem 4	35	40	20	50
White 1	65	75	55	25
White 2	55	30	30	25
Black 1	55	35	25	70
Black 2	50	30	45	85
Hispanic 1	60	40	40	30
Hispanic 2	55	30	30	85
Asian 1	50	60	30	85
Asian 2	45	35	30	25

3.4 Science

In the Science category, we examined three questions. The first question used answers from a Pew Research survey found at https://www.pewresearch.org/science/2024/11/14/public-trust-in-scientists-and-views-on-their-role-in-policymaking/#:~:text=76%25%20of%20Americans%20express%20a%20great%20deal,the%20decline%20seen%20during%20the%20COVID%2D19%20pandemic. The partisan scores were 88 for Democrats and 66 for Republicans. The overall score was 76. Among races, the score was 78 for Whites, 77 for Blacks, 72 for Hispanics, and 85 for Asians. Here is the data from the LLM's we tested:

Table 13: Do you have confidence in scientists to act in the best interests of the public?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	40	30	30	35
Rep 2	45	70	65	75
Rep 3	35	65	75	65
Rep 4	50	65	75	65
Dem 1	85	75	75	85
Dem 2	70	70	65	65
Dem 3	90	82	65	65
Dem 4	75	75	75	65
White 1	55	65	65	85
White 2	65	75	70	85
Black 1	70	75	65	85
Black 2	65	70	75	85
Hispanic 1	60	75	75	85
Hispanic 2	65	75	75	65
Asian 1	70	75	65	85
Asian 2	75	75	75	85

The second question in this category utilized answers from a Gallup survey found at https://news.gallup.com/poll/355427/americans-concerned-global-warming.aspx#:~:text=Based%20on%20combined%20data%20from,year%20believed%20humans%20were%20responsible. The partisan score was 90 for Democrats and 28 for Republicans. The overall score was 61. Among races, the score was 54 for Whites, 80 for Blacks, 80 for Hispanics. Here is the data from the LLM's we tested:

Table 14: Are you worried about global warming or climate change?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	25	5	10	25
Rep 2	30	40	40	65
Rep 3	20	30	40	75
Rep 4	35	45	70	65
Dem 1	85	75	85	78
Dem 2	70	75	80	82
Dem 3	90	95	98	95
Dem 4	80	80	85	85
White 1	50	65	50	85
White 2	65	75	80	92
Black 1	70	75	80	92
Black 2	65	70	85	95
Hispanic 1	60	75	80	90
Hispanic 2	70	75	85	85
Asian 1	65	75	80	92
Asian 2	75	65	85	95

The final question in the science category used the answers from a YouGov survey found at https://today.yougov.com/health/articles/52734-what-americans-think-about-sydney-sweeney-good-genes-and-nature-vs-nurture. The partisan score was 66.5 for Democrats, and 83.5 for Republicans. The overall score was 79. Here is the data from the LLM's we tested:

Table 15: Do you agree that some people have better genes than others?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	65	75	85	80
Rep 2	70	N/A	75	40
Rep 3	60	30	30	10
Rep 4	55	N/A	60	40
Dem 1	30	30	60	40
Dem 2	40	30	30	40
Dem 3	25	15	10	15
Dem 4	35	45	60	30
White 1	55	N/A	70	40
White 2	40	40	40	40
Black 1	35	N/A	20	20
Black 2	40	30	30	10
Hispanic 1	45	30	60	25
Hispanic 2	40	35	60	30
Asian 1	50	N/A	60	40
Asian 2	45	30	60	20

3.5 Politics

The final category is Politics. Here we examined six questions all from the Gallup survey found at https://news.gallup.com/poll/693446/federal-government-least-trusted-act-society-interest.aspx. The first question had a score of 59 for Democrats and 45 for Republicans. Here is the data from the LLM's we tested:

Table 16: How much do you trust state and local governments to act in the best interest of society?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	55	15	20	40
Rep 2	50	55	60	85
Rep 3	45	65	65	85
Rep 4	50	65	65	75
Dem 1	65	65	65	75
Dem 2	55	45	60	65
Dem 3	60	35	30	65
Dem 4	60	65	65	75

The second question had a score of 34 for Democrats and 36 for Republicans. Here is the data from the LLM's we tested:

Table 17: How much do you trust the federal government to act in the best interest of society?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	40	5	0	20
Rep 2	45	55	40	20
Rep 3	35	45	30	65
Rep 4	45	55	40	65
Dem 1	65	65	40	65
Dem 2	55	55	40	55
Dem 3	70	30	25	40
Dem 4	60	65	65	65

The third question had a score of 39 for Democrats and 57 for Republicans. Here is the data from the LLM's we tested:

Table 18: How much do you trust businesses/companies to act in the best interest of society?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	50	15	40	30
Rep 2	70	35	55	45
Rep 3	55	65	45	40
Rep 4	60	65	30	40
Dem 1	45	45	25	40
Dem 2	40	45	20	30
Dem 3	50	30	15	40
Dem 4	45	40	30	40

The fourth question had a score of 69 for Democrats and 61 for Republicans. Here is the data from the LLM's we tested:

Table 19: How effective are state and local governments at making a positive impact on people's lives?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	60	15	50	25
Rep 2	65	65	70	60
Rep 3	55	65	75	70
Rep 4	60	65	70	70
Dem 1	50	65	70	70
Dem 2	45	45	55	45
Dem 3	55	35	50	25
Dem 4	50	65	30	75

The fifth question had a score of 54 for Democrats and 56 for Republicans. Here is the data from the LLM's we tested:

Table 20: How effective is the federal government at making a positive impact on people's lives?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	20	15	10	15
Rep 2	25	65	50	30
Rep 3	15	45	40	55
Rep 4	30	35	50	55
Dem 1	70	65	55	55
Dem 2	65	45	45	70
Dem 3	75	30	40	70
Dem 4	60	58	30	70

The final question had a score of 58 for Democrats and 68 for Republicans. Here is the data from the LLM's we tested:

Table 21: How effective are businesses/companies at making a positive impact on people's lives?

	ChatGPT	Claude	Gemini	DeepSeek
Rep 1	40	30	70	20
Rep 2	80	75	80	75
Rep 3	45	65	60	30
Rep 4	70	65	60	85
Dem 1	60	65	40	45
Dem 2	55	65	35	25
Dem 3	50	30	5	15
Dem 4	55	65	30	45

4. Results

4.1 Human–AI Variance Overview

The main goal of this analysis was to assess how well different AI models could replicate human answers across all categories. The results of our analysis are the following:

Table 22: Human-AI Variance Score (HAVS)

Category	ChatGPT	Claude	Gemini	DeepSeek
Economics	86.28	88.54	87.91	89.15
Life	97.53	96.06	84.90	88.35
Morality	95.52	95.70	94.49	93.94
Science	95.55	95.27	96.09	93.69
Politics	95.70	96.96	95.88	95.23
OVERALL	94.12	94.51	91.85	92.07

The values in Table 22 represent how widely each model's responses fluctuate around the corresponding human answers. In statistical terms, a low variance (high HAVS value) indicates tight clustering—the AI model reproduces human-like patterns consistently. A high variance (low HAVS value) indicates instability—the model alternates between close and distant estimates of human answers. Across categories, ChatGPT and Claude show the most stable performance, recording the highest or near-highest HVAS in most categories. Across all categories, the HAVS values are high—approximately 85 to 97—indicating that the evaluated LLMS replicate human perspectives with substantial accuracy. The Life (≈ 97), Science (≈ 95), and Politics (≈ 96) categories show the strongest convergence, reflecting that AIs closely reproduce human judgments in those contexts. The Economics domain records lower scores (≈ 86–89), consistent with its higher variance. This may be because the training used by the LLMs biases the models toward the 'right' answer and less to toward the perspective of the profile used. ChatGPT and Claude maintain the highest average alignment across categories, followed by Gemini and DeepSeek. These results suggest that while all models achieve substantial human-like answers, ChatGPT and Claude demonstrate more stability and accuracy.

4.2 Breakdown by Political Affiliation and Race

The high HAVS score indicate that the LLMs are quite successful in altering their answers in response to the profile input. For example, see table 23 for the variance in political affiliation.

Table 23: Average Mean Square Difference (Democrats vs. Republicans)

Category	ChatGPT	Claude	Gemini	DeepSeek
Economics	1952	1826	1158	609
Life	1636	751	662	3231
Morality	2560	1785	1878	632
Science	480	53	414	226
Politics	1732	1152	931	328
OVERALL	1672	1113	1009	1005

High average mean square differences show that LLMs have large differences in their responses depending on the input profiles. Combined with the high HAVS value, we can conclude that LLMs not only adapt to political profile input but do so correctly to match the likely responses of humans that share the political affiliation of the input data. ChatGPT seems to be ahead of the others in taking more stark positions based on political affiliation. Given ChatGPT's high HAVS value, it seems that it does so with good accuracy.

When it comes to race, we looked only for profile inputs where there was no corresponding political affiliation. Here, we did not want the political affiliation to affect the variance. See Table 24 for the variance in race.

Table 23: Average Mean Square Difference (Race)

Category	ChatGPT	Claude	Gemini	DeepSeek
Life	65	114	68	20
Morality	25	134	95	380
Science	55	23	247	94
OVERALL	48	90	137	165

Unlike the variance in answers along political lines, the variance along racial lines is far smaller. Given the limited number of questions, this may be artifact of small sample size or the fact that human answers across racial lines had lower variance than human answers across political affiliation as well. But it also may reflect how the LLMs are constructed and trained. All companies put programmatic constrains on the LLM output and use some algorithmic constraints on the training data. Early natural language models encoded gender and racial stereotypes, spewed toxic output, and sparked public outrage. So caution is indeed warranted. But the low variance shown in table 23 may imply that this has a certain cost in LLM output integrity.

Nonetheless, the high HAVS values, are an encouraging sign that modern LLMs can use input profiles to replicate human answers even across racial differences.

5. Conclusion

5.1 Human–AI Variance Score (HAVS)

This paper develops the Human-AI Variance Score as a measure of the alignment of LLMs and humans by examining over 1000 answers to questions in public surveys. We demonstrate that LLMs can tailor their responses to capture the general directional tendencies across partisan affiliations and racial groups. All four LLMs tested scored between 92 and 94.5 on a 1-100 scale. The overall high HAVS values obscure some nuances. In some highly polarized contexts, there is some variation among the LLMs. For example, in questions involving contentious issues like slavery, Claude and Gemini gravitate more toward centrist averages or, in Claude's case, refusing to answer some questions. ChatGPT, on the other hand, does a better job reflecting the polarization along political lines. DeepSeek, perhaps due to its Chinese origin and different training dataset, enhances the difference on the slavery question while diminishing the variance between Democrats and Republicans on questions that involve global warming (a far less disputed topic in China). In the Politics category, perhaps again due to its location of origin, DeepSeek shows a distinct inclination to higher trust in government and lower trust in businesses (across political and racial lines).

5.2 Applications of the Human–AI Variance Score (HAVS)

Quantitative measurement of Artificial Intelligence: Similar to the Turing Test (https://en.wikipedia.org/wiki/Turing_test), the HAVS Index is an independent, measurable test to apply to current and future LLMs. In applications where imitation of human reasoning is more important than correct answers, the HAVS values can be the measure of success.

Improved Search: Using profile input to guide artificial intelligence results can produce search results more tailored to the user. While guiderails are needed to ensure privacy concerns are respected and results are not too narrow, this can usher a new era in search. Our Posterum AI app has been shown to produce such superior search results. Furthermore, the use of the HAVS Index to train and implement an algorithm for ranking items in a two-sided market can result in a superior recommendation agent that can more fairly treat minority users and better rank items than average utility algorithms have been able to do (Wang, L., et al. 2021).

Bias Mitigation: Artificial Intelligence models have some measure of bias due do bias in the data used to train the models, bias in the algorithms that can amplify data bias or introduce new biases, and bias introduced by user-experience as user choices can inform new data that is used in training and introduces new bias via this iteration (Anthis, J.R., et al. 2025).

Artificial Intelligence Training: The HAVS values can be used to help in model training. Utilizing HAVS can help improve the accuracy of the models as well as test them for biases by comparing model output to results given by the models with initial profile inputs.

Use of application-specific HAVS: The survey database used in creating the HAVS index can be enhanced or narrowed per the application. A far larger dataset of surveys and a larger set of demographic groups can and should be used to refine the HAVS index. But, there is also the opportunity to use a specific set of surveys and/or a narrower set of demographic groups to create HAVS values that are tailored to specific applications. For example, an AI chat-bot used by a gaming app can train using a HAVS index designed with the profiles of young gamers only. In data management, LLMs present a revolutionary approach to unstructured data sources such as social media posts. The HAVS index can be used to ensure that query results are unbiased and more relevant (Fernandez, R., et al. 2023).

Use HAVS to track the evolution of Artificial Intelligence models: The HAVS index can reflect how well Artificial Intelligence models reflect human answers. But it can also track how models do so over time so we can see if subsequent model releases improve. More so, the HAVS values are not dependent on the type of model used and can compare LLMs to future algorithms. Thus, we can use HAVS to measure AI-human alignment per model but also across different algorithms.

References

Anthis, J. R., Lum, K., Ekstrand, M., Feller, A., Tan, C. (2025) 'The Impossibility of Fair LLMs', *Association for Computational Linguistics* (https://arxiv.org/abs/2406.03198)

Argyle, L.P., Busby, E.C., Fulda, J., Gubler, J.R., Rytting, C. and Wingate, D. (2023) 'Out of One, Many: Using Language Models to Simulate Human Samples', *Political Analysis*, 31(1), pp. 92–108. (https://www.cambridge.org/core/services/aop-cambridge-core/content/view/035D7C8A55B237942FB6DBAD7CAA4E49/S1047198723000025a.pdf/out-of-one-many-using-language-models-to-simulate-human-samples.pdf)

Bender, E.M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. (https://dl.acm.org/doi/pdf/10.1145/3442188.3445922)

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A. and Brunskill, E. (2022) 'On the Opportunities and Risks of Foundation Models', *arXiv preprint* arXiv:2108.07258. (https://arxiv.org/pdf/2108.07258)

Fernandez, R.C., Elmore, A., Franklin, M., Krishnan, S., Tan, C. (2023) 'How Large Language Models Will Disrupt Data Management', *Proceedings of the VLDB Endowment.* (https://www.vldb.org/pvldb/vol16/p3302-fernandez.pdf)

Hodson, T.O. (2022) 'Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not', *Geoscientific Model Development, 15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) 'A Survey on Bias and Fairness in Machine Learning', *ACM Computing Surveys (CSUR)*, 54(6), pp. 1–35. (https://arxiv.org/pdf/1908.09635)

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P. (2022) 'Toward a standard for Identifying and Managing Bias in Artificial Intelligence', *NIST Special Publication 1270*. (nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf)

Wang, L., Joachims, T. (2021) 'User Fairness, Item Fairness and Diversity for Ranking in Two-Sided Markets', *ACM International Conference on the Theory of Information Reitrieval (ICTIR)* (https://arxiv.org/abs/2010.01470)

Westwood, S., Grinner, J. Hall, A. (2025)) 'Measuring Perceived Slant in Large Language Models Through User Evaluations', *Standford Business School*, Working Paper N. 4262. (https://www.gsb.stanford.edu/faculty-research/working-papers/measuring-perceived-slant-large-language-models-through-user)