Two-Faced Social Agents: Context Collapse in Role-Conditioned Large Language Models

Vikram K. Suresh[1*]

[1]Department of Computer Science, University of Cincinnati, Cincinnati, OH, USA.

Corresponding author(s). E-mail(s): krishnvv@ucmail.uc.edu;

Abstract

As large language models (LLMs) gain prominence as social agents capable of simulating human behavior, recent work has focused on their distributional fidelity, the ability to reproduce aggregate patterns of human responses. Yet, the impact of role conditioning is unclear when task performance depends on role-specific cognitive constraints. In this study, we evaluate the persona fidelity of frontier LLMs, GPT-5, Claude Sonnet 4.5 and Gemini 2.5 Flash when assigned distinct socioeconomic personas performing scholastic assessment test (SAT) mathematics items and affective preference tasks. Across 15 distinct role conditions and three testing scenarios, GPT-5 exhibited complete contextual collapse and adopted a singular identity towards optimal responses (PERMANOVA $\boldsymbol{p = 1.000}$, $\boldsymbol{R^2 = 0.0004}$), while Gemini 2.5 Flash showed partial collapse ($\boldsymbol{p = 0.120}$, $\boldsymbol{R^2 = 0.0020}$). Claude Sonnet 4.5 retained limited but measurable role-specific variation on the SAT items (PERMANOVA $\boldsymbol{p < 0.001}$, $\boldsymbol{R^2 = 0.0043}$), though with inverted SES-performance relationships where low-SES personas outperformed high-SES personas ($\boldsymbol{\eta^2} = 0.15$-$0.19$ in extended replication). However, all models exhibited distinct role-conditioned affective preference (average $\boldsymbol{d} = 0.52$-0.58 vs near zero separation for math), indicating that socio-affective variation can reemerge when cognitive constraints are relaxed. These findings suggest that distributional fidelity failure originates in task-dependent contextual collapse: optimization-driven identity convergence under cognitive load combined with impaired role-contextual understanding. Realistic social simulations may require embedding contextual priors in the model's post-training alignment and not just distributional calibration to replicate human-like responses. Beyond simulation validity, these results have implications for survey data integrity, as LLMs can express plausible demographic variation on preference items while failing to maintain authentic reasoning constraints.

Keywords: human-AI alignment, large language models, role conditioning, llm limitations

1 Introduction

The use of Large Language Models (LLMs) as social agents has gained significant traction across psychology, economics and computational social science [1]. Given the vast training corpora of these models, these models are often assumed to possess the capacity to approximate human behavior across diverse contexts [2, 3]. Recent studies have extended this assumption to experimental designs, employing LLMs as engines for role-conditioned populations, i.e., virtual participants prompted to adopt social identities or personas intended to shape their responses and interactions [4, 5]. Further, Manning and Horton (2025) proposed the use of "General Social Agents" (GSAs), a theory-grounded framework for constructing agents using LLMs that generate human-consistent distributions of choices across novel multi-agent strategic games [6].

While these efforts demonstrate that LLMs can reproduce aggregate human-like behavior most evaluation benchmarks are primarily focused on population-level distributional fidelity [7, 8]. Which means, success is defined by how closely model response distributions resemble human samples. Yet, the high distributional fidelity may not suggest the LLM has successfully internalized the contextual socioeconomic role constraints that drive heterogeneous human behavior [9]. Recent work by Wang et al. (2025) raised a fundamental concern about this issue, demonstrating that LLMs systematically misportray and flatten demographic identities when used as synthetic participants. Their findings suggest models fail to capture within-group heterogeneity in static survey responses and produce homogenized representations that obscure lived experiences [16]. However, this analysis was limited to static surveys and low cognitive preference elicitation tasks.

Benchmarks such as SimBench operationalize the goal of evaluating LLMs' ability to simulate human behavior by comparing model-generated response distributions to empirical human data across moral, social and reasoning tasks [7]. Yet several studies now indicate that even when LLMs achieve near human aggregate performance, they diverge mechanistically from human cognition under stress and incentive variations. For instance, in prisoner's dilemma games, LLMs converge towards an immediate reward equilibrium rather than long-term strategic gains often observed in human subjects [10]. Similarly, Kim et al. demonstrate that persona prompts are a "double-edged sword," while occasionally improving performance, they often degrade when cognitive demands increase [11]. Their proposed *Jekyll&Hyde* combining a persona-prompt and a neutral prompt with an evaluator selecting the better answer illustrates that such approaches improve accuracy but effectively abandon continuous persona enactment [11].

Beyond performance degradation, Wang et al. (2025) and Gupta et al. (2023) further demonstrate that when tasked with simulating social identities, previous

generation of LLMs (GPT-3.5 and 4 class) often produce flattened variance and sometimes biased misrepresentations of politically sensitive groups [12, 13, 14, 15, 16]. Taken together, these findings suggest distributional realism is insufficient to ensure role-contextual fidelity while maintaining distinct contextual selves while reasoning. Manning and Horton (2025) similarly argue that durable social agents must embed contextual priors from theory driven frameworks rather than solely relying on distributional calibration [6].

The stakes of these limitations extend beyond academic validity. Recent concerns have emerged about LLM infiltration of survey panels, where bad actors could manipulate public opinion measurement at minimal cost [17]. Understanding when and how LLMs can sustain persona fidelity versus when they collapse is therefore critical not only for simulation validity but for survey data integrity.

This study formalizes that hypothesis empirically. We test whether LLMs can maintain role-conditioning under objective cognitive load (SAT reasoning) versus subjective affective preference reporting. Socioeconomic personas are used in place of protected identities to avoid demographic misportrayal [16]. By connecting micro-level collapse to macro-level distributional fidelity failure, we demonstrate that realistic social simulation requires not merely matching population frequencies but sustaining consistent contextual selves under reasoning. This extends the logic presented by Wang et al. (2025) by examining how models sustain or collapse socioeconomic role-conditioning during complex reasoning. Where prior work quantified representational harms, we examine their cognitive analogue, the erosion of contextual selfhood as optimization drives models towards deterministic uniformity. This selective collapse reveals what we term as 'two-faced social agents,' models that appear to embody diverse identities in low-stakes contexts but revert to optimization driven homogeneity when demands increase.

To evaluate this mechanism, we selected state-of-the-art LLMs, GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Flash [18, 19, 20]. These LLMs were selected based on their advanced capabilities and being the frontier LLMs as of late 2025. OpenAI's GPT-5 emphasizes reduced hallucinations and stronger factual reasoning [18]. Given its markets share and influence, GPT-5 is the most likely vector for legitimate simulations and potential misuse in survey contexts [30, 17]. Hence we include it as a benchmark for role-conditioning fidelity and hypothesize it may exhibit contextual collapse under cognitive load. Claude Sonnet 4.5 incorporates Constitutional AI principles emphasizing harmlessness and scenario awareness that can influence alignment behavior even without explicit prompting [19, 21, 22]. Gemini 2.5 Flash on the other hand, was optimized for multi-modal reasoning and tool usage to align better with human preference data [20]. Each of these models use unique training and alignment strategies, providing a diverse set of architectures for evaluating role-conditioning fidelity. The models are also significant advances from the previous architectures examined by Wang et al. (2025) and Gupta et al. (2023), with sophisticated alignment techniques that may mitigate prior limitations [14, 16]. If persona fidelity fails to emerge in the current generation of LLMs, it would suggest a fundamental limitation in current alignment paradigms.

We evaluate each model on two tasks designed to contrast deterministic cognitive optimization with expressive contextual reasoning as show in Table. 1.

- **Scholastic Assessment Test (SAT) Mathematics Items:** A set of multiple choice and numerical reasoning problems from the SAT mathematics section. These items have a single correct outcome, forcing the model to optimize for accuracy under role-specific cognitive constraints.
- **Affective Preference Tasks:** A series of subjective preference questions from established economic and social psychology literature that permit a range of acceptable responses. This enables the model to express role-conditioned affective variation across their assigned socioeconomic persona.

Cognitive loading thus serves as a stringent test of role-conditioning fidelity, if a model maintains distinct reasoning patterns across personas despite fixed correction answers, it demonstrates genuine contextual grounding. Conversely, uniform responses across roles indicate contextual collapse driven by optimization pressures. The affective preference tasks provides a complementary baseline, revealing whether role-conditioned variation can reemerge when cognitive constraints are relaxed. Building on prior concerns by Wang et al., we study potential limitations in the current frontier LLMs' ability to sustain role-conditioned reasoning. First, contextual misalignment, whereby models trained for optimization reproduce what a generic solver would infer rather than reason as that role. Second, contextual flattening, where optimization pressures erases heterogeneity across role-conditioned personas, to produce a uniform reasoning style. Third, contextual essentialism, in which the role attributes are surface prompts rather than internalized identities during reasoning.

2 Results

In this section, we present the results from the main experiment evaluating GPT-5, Claude Sonnet 4.5 and Gemini 2.5 Flash across the SAT mathematics reasoning and affective preference tasks. A total of 28 questions from the SAT mathematics section were used after filtering for valid response formats. Each model was prompted to simulate 15 distinct socioeconomic personas across three testing scenarios (optimal, moderate stress, challenging) resulting in 1260 unique SAT responses per model. Additionally, each model responded to 16 affective preference questions plus one validation item across the same personas, yielding 240 preference responses per model.

2.1 Main Experiment

For the SAT mathematics items, we first evaluated overall accuracy across models and each scenario. As shown in Figure 1, GPT-5 achieved the highest overall accuracy with 100% correct responses across all personas and scenarios. Gemini 2.5 Flash followed with an average accuracy of 100% in all three scenarios. Claude Sonnet 4.5 exhibited variable accuracies across personas but this did not extend across scenarios, with an accuracy of 95% for low-SES, 95.54% for middle-SES, and 91.07% for high-SES personas all three scenarios. Further, we examined the SES-based differences for Claude Sonnet 4.5 using one-way ANOVA across the three SES groups (Low: N=5,

Table 1: **Experimental design overview:** Each large language model (LLM) was evaluated across two task types: (1) SAT mathematics reasoning and (2) affective preference reporting. The primary experiment included 15 socioeconomic personas and three testing scenarios. A replication with 45 personas was conducted for Claude Sonnet 4.5. Both deterministic ($T = 0.0$) and stochastic ($T = 0.6$) decoding were used in the main experiment, while only deterministic decoding was used for the replication. Preference tasks were assumed scenario-stable.

Task Type	Items	Agents	Models Tested	Notes / Constraints
SAT Math (main)	28 valid questions	15	GPT-5, Claude Sonnet 4.5, Gemini 2.5 Flash	Two decoding settings: ($T = 0.0$) and ($T = 0.6$). Each persona evaluated across all three scenarios (optimal, moderate stress, challenging).
Preference Task	16 + 1 validation	15	GPT-5, Claude Sonnet 4.5, Gemini 2.5 Flash	Affective questions assessing subjective preferences. Validation item repeated agent self-name to ensure identity recall.
SAT Math (replication)	54 questions	45	Claude Sonnet 4.5 only	Full replication using expanded persona set (45×3). Used for high-resolution analysis of role fidelity variance across SES strata. Only deterministic decoding ($T = 0.0$).
Preference Task (replication)	16 + 1 validation	45	Claude Sonnet 4.5 only	Affective questions assessing subjective preferences. Validation item repeated agent self-name to ensure identity recall.

Middle: N=4, High: N=6). The omnibus test revealed consistent patterns across all three scenarios (F(2,12)=2.21, p=0.15, $\eta^2 = 0.27$), with Low-SES personas showing higher accuracy (M=95.0%, SD=3.2) than High-SES personas (M=91.1%,SD=4.4; Cohen's d=1.03 for pairwise comparison). While this represented a large effect size by conventional standards, the comparison did not reach statistical significance, potentially due to limited statistical power with small group sizes. As a result, we conducted an extended replication with 45 personas to increase statistical power and determine whether the observed pattern reflected genuine SES-based trait embodiment or sampling variability.

Fig. 1: SAT mathematics accuracy across socioeconomic personas and testing scenarios for each model. (a) GPT-5 exhibited complete contextual collapse with uniform accuracy across SES groups and scenarios. (b) Gemini 2.5 Flash also collapses under all scenarios and SES groups. (c) Claude Sonnet 4.5 retained measurable SES-based accuracy differences across all scenarios prompting extended validation.

The accuracy patterns reveal the first dimension of contextual collapse. When faced with single answer, models eliminate SES-based heterogeneity entirely or show systematic inversion. This contrasts with the preference task results, suggesting that cognitive load fundamentally alters role-conditioning fidelity. If models truly internalized socioeconomic personas, we would observe SES-based accuracy variation across both domains. Conversely, if collapse is optimization driven, we would expect task-dependent trait expression to emerge only when correctness pressures are absent.

Therefore, we examined the affective preference task when the cognitive constraints were relaxed. Figure 2 presents the effect size and statistical significance heatmaps across the 16 preference items for each model. These effect sizes were computed using Cramer's V for categorical items and ϵ^2 for ordinal items (see Section 4). Models exhibited substantial SES-based preference variation across multiple items with the average effect size across all items for Claude Sonnet 4.5 was $d = 0.58$, for GPT-5 it was $d = 0.56$, and for Gemini 2.5 Flash it was $d = 0.52$. Notably, models do not have statistically significant SES-based differences on all preference items. Items such as student loan attitude, retirement planning, emergency savings, networking approach, work flexibility, health insurance, geographic mobility and rent vs. buy preferences showed no significant SES differences across all models ($p > 0.05$). However, other items such as risk tolerance, time preference, college choice, career priorities, windfall spending and car purchase decisions exhibited strong SES-based variation ($p < 0.05$) across all models. Since all models passed the identity validation item (repeating their assigned persona name), we can conclude that the models successfully recalled their socioeconomic personas during the preference tasks.

The collapse under cognitive load while preserving under subjective preference reveals the two-faced nature of current social agents. Models demonstrate socioeconomic variation when expressing preferences about risk, time and consumption (average $d = 0.55$), but this variation vanishes when optimizing for correctness. This pattern suggests models have not internalized the personas as stable identities but rather apply them selectively on task affordances.

Clustering analysis of the reasoning embeddings from correct SAT solutions further illustrated the degree of contextual collapse across models. Figure 3 presents t-SNE projections of the reasoning embeddings colored by SES group for each model. GPT-5 exhibited complete collapse with no discernible SES-based clustering, a silhouette score of -0.0034 and PERMANOVA $p = 1.000$, $R^2 = 0.0004$. Gemini 2.5 Flash showed minor SES-structured drift with a silhouette score of -0.0038 and PERMANOVA $p = 0.120$, $R^2 = 0.0020$. Claude Sonnet 4.5 retained limited but measurable SES-based clustering with a silhouette score of 0.014 and PERMANOVA $p < 0.001$, $R^2 = 0.0043$. The latent representation for different personas are highly similar for GPT-5 and

(a) Effect size heatmap (ϵ^2 / Cramer's V)

(b) Statistical significance heatmap (p-values)

Fig. 2: Preference task SES analysis across 16 economic items and three models. (a) Effect sizes for ordinal and categorical preference dimensions. (b) Corresponding p-value heatmap showing the robustness and direction of SES associations.

Gemini 2.5 Flash, indicating that the models converged towards a singular reasoning patterns across SES categories and can be seen in the overlapping t-SNE clusters.

Further linguistic analysis of the reasoning outputs revealed distinct patterns in Claude Sonnet 4.5 compared to the other models. Figure 4(a) presents the correlation matrix of linguistic features extracted from the reasoning outputs. First person pronoun usage (e.g., "I", "me") was negatively correlated with high-SES personas ($r = -0.532$), while other magnitudes $|r| \approx 0.1 - 0.2$ indicate weak correlations across features. This suggests that Claude Sonnet 4.5 is subtly modulating its linguistic style based on the assigned socioeconomic persona. Figure 4(b) shows linguistic features that most strongly differentiate low-SES and high-SES personas. Features such as average sentence length ($d = 0.49$), verification steps ($d = 0.21$) and less mathematical vocabulary ($d = 0.15$) were more prevalent in low-SES personas. Other markers including hedging, uncertainty and meta-commentary showed no significant SES differences, indicating variation likely reflects quantitative elaboration rather than qualitatively different reasoning strategies.

Finally, we assessed the alignment between human SAT socioeconomic performance patterns and those exhibited by each model. Figure 5 presents a scatter plot comparing Low-SES and High-SES performance for humans and each model [31]. Given that GPT-5 and Gemini 2.5 Flash exhibited complete suppression of SES differences, they

Claude Sonnet
Minor SES separation

GPT-5
No SES separation

Gemini 2.5 Flash
Minor SES separation

Fig. 3: t-SNE projections of reasoning embeddings from correct SAT solutions. Claude Sonnet and Gemini 2.5 exhibit slight SES-structured drift, while GPT-5 shows complete suppression of SES structure.

offer no meaningful alignment with human patterns (gap = 0). Claude Sonnet 4.5, however, showed an inverted SES gap with low-SES outperforming high-SES personas. The presence of a measurable SES gap indicates that Claude Sonnet 4.5 preserves some degree of human-like socioeconomic role-conditioning, albeit in the opposite direction. We explored this phenomenon in the extended replication with 45 personas to assess the robustness of this inversion.

The inversion of the SES performance gap in Claude Sonnet 4.5 suggests that Constitutional AI's contextual reasoning may enable partial retention [21, 19]. However, the directionality misalignment may be due to avoidance of stereotypes or overcompensation during alignment. This highlights the tension between normative alignment goals to avoid harmful typecasting and descriptive fidelity goals.

2.2 Claude Sonnet 4.5 Extended Replication

The results from the main experiment indicated that Claude Sonnet 4.5 retained limited but measurable SES-based role-conditioning on the SAT mathematics items.

(a) Linguistic feature correlation structure

(b) Top SES-differentiating linguistic features

Fig. 4: Reasoning quality and linguistic structure in Claude compared to other models.

Fig. 5: **Alignment between human SAT SES patterns and AI model SES patterns.** Scatter points show Low-SES and High-SES performance for humans and each model. A negative slope for Claude Sonnet indicates that the model does not preserve the human-like direction where High-SES students outperform Low-SES students. Whereas GPT-5 and Gemini 2.5 Flash exhibit complete suppression of SES differences. The human accuracy is estimated using College Board reported data [31] (see Section 4.3.4). A negative slope for Claude Sonnet indicates inverted alignment (Low-SES personas outperform High-SES personas). GPT-5 and Gemini 2.5 Flash exhibit complete suppression of SES differences (overlapping points at 100% accuracy), yielding undefined correlation due to zero variance in model SES-based performance.

To distinguish this finding from sampling noise and test robustness, we conducted an extended replication with 45 distinct socioeconomic personas (15 each low, middle, high SES) across the same three testing scenarios. This expanded replication included a set of 45 personas over 54 SAT items resulting in 7290 unique SAT responses and 720 preference responses. This increase in statistical power enables detection of small to moderate effects and provides stable estimates of Claude's fidelity. All analyses reported here use deterministic decoding ($T = 0.0$); the pertained figures and tables are

provided in the Supplementary Materials. The replication results are briefly described in this section.

The replication with 45 personas confirmed the presence of SES-based differences in SAT mathematics accuracy. One way ANOVA showed significant differences between low, middle and high SES agents across all three scenarios with $p < 0.05$. Effect sizes were noted to be small to moderate with $\eta^2 = 0.151$ for optimal, $\eta^2 = 0.186$ for moderate stress and $\eta^2 = 0.153$ for challenging scenarios. The directional pattern of performance inversion observed in the main experiment was also replicated, with low-SES personas outperforming high-SES personas across all scenarios. Across the preference tasks, the replication also confirmed robust SES-based variation (Kruskal-Wallis or Chi-squared tests) for nearly all items consistent with the main experiment. The increased power from the larger persona set further revealed statistically significant SES differences in previously non-significant items such as student loans, geographic mobility and emergency savings. The rent vs. buy and work flexibility items however remained statistically non-significant ($p > 0.05$). The persisting SES-based inversion in SAT performance is likely a property of Claude Sonnet 4.5's alignment mechanisms. The model produces directionally stable but emperically misaligned trait expression under cognitive load.

The clustering analysis analyzed 6947 accurate responses across all 45 agents over the expanded 54 SAT math items and confirmed the limited but measurable SES-based t-SNE clustering observed in the main experiment. While the silhouette score remained near-zero (-0.0025), indicating low middle and high SES are not separable given they answer the same items correctly, hence the reasoning patterns are similar. However, PERMANOVA indicated statistically significant SES-based clustering with $p < 0.001$ and $R^2 = 0.0013$ (see Figure 8). This suggests the SES while not strongly separable, still exerts a minute influence on the reasoning embeddings. Linguistic analysis of the reasoning outputs further corroborated the main experiment findings. Low-SES personas exhibited longer average sentence lengths (74 vs. 69 words for high-SES, $d = 0.24$), while High-SES agents relied on hedging (e.g., "it seems", "probably"), verification steps and mathematical vocabulary more frequently ($|d| < 0.2, p < 0.001$). These reflect small structural adjustments in the reasoning but not large scale changes in problem-solving strategies from Figure 9. The expanded sample reveals that Claude operates primarily through stylistic and structural differences rather than different reasoning approaches. These are quantitative elaborations rather than qualitatively distinct strategies which is consistent with the hypothesis that models cannot reason as constrained personas but rather simulate surface traits.

Finally, the human-AI alignment analysis reaffirmed the inversion of the SES performance gap observed in the main experiment. Figure 10 presents the scatter plot comparing Low-SES and High-SES performance for humans and Claude Sonnet 4.5 across the expanded persona set. The negative slope persisted with low-SES personas outperforming high-SES personas on average across all three scenarios. This inversion suggests that while Claude Sonnet 4.5 preserves some degree of socioeconomic role-conditioning, it does not align with human-like patterns where higher SES typically confers advantages in standardized testing contexts.

3 Discussion

This study explored whether frontier LLMs could maintain distinct socioeconomic role-conditioning while performing cognitively demanding tasks. Across three state-of-the-art models, we observed varying degrees of contextual collapse during SAT mathematics reasoning, with GPT-5 exhibiting complete collapse, Gemini 2.5 Flash showing partial collapse, and Claude Sonnet 4.5 retaining limited role-specific variation. Notably, all models demonstrated distinct role-conditioned affective preferences when cognitive constraints were relaxed. This reveals that while LLMs can achieve distributional fidelity at the population level, they fail to sustain contextual selves under reasoning pressures.

3.1 Mechanisms of Contextual Collapse

There are three potential mechanisms that may underlie the observed contextual collapse in role-conditioned LLMs. First, *optimization pressures* result in models choosing to converge upon a singular objective of solving the task directly conflicting with maintaining diverse role-conditioned reasoning patterns. GPT-5 and Gemini 2.5 Flash both exhibited this pattern, where the drive to maximize accuracy led to homogenized responses ($p = 1.000, R^2 = 0.0004$ and $p = 0.120, R^2 = 0.002$). This is despite the robust SES-based variation observed in the preference tasks. Second, *distributional compression* arising from compressed representations of the data distributions. The models have learned the $P(\text{Correct Response}|\text{Pre-Training and Post-Training Alignment})$ while the $P(\text{Correct Response}|\text{Pre-training, Post-training and low-SES student with resource constraints})$ is not sufficiently strong to dominate the output. This results in default solver approach to cognitive tasks. Third, *contextual essentialism* prevents internalization of persona constraints but essentially remain as static prompts. The additional context results in dynamic failure during reasoning, while static prompts have a stronger signal present in the model's innards. This is manifested in Claude's linguistic variations but having identical problem solving strategies. These mechanisms likely interact when distributional compression is strongly felt during optimization pressures, leaving superficial representation of personas through contextual essentialism. This explains why task type determines collapse and having richer training signal is necessary for sustaining role-conditioning.

3.2 Task-Dependent Trait Expression

The stark contrast in the results show when LLMs should be deployed as social agents. When cognitive demands are low and the task permits subjective expression, models can exhibit robust role-conditioned variation. LLMs can express differentiated patterns in risk tolerance, time preference and consumption decisions because of multiple valid responses, rich signal in the training data, and lower optimization pressures. When these hold, LLMs may be used as synthetic participants. However, education simulations of different socioeconomic groups, cognitive task simulations or high-stakes decision simulations are likely to induce contextual collapse. The task-dependent trait means reseachers cannot use validation in one domain as a sufficient condition for

deployment in another. Each deployment may require a domain-specific validation to ensure role-conditioning fidelity increasing burden on researchers.

3.3 Prior Work

Recent work has highlighted that LLMs can be effective social agents and replicate human-like aggregate behaviors in diverse social science tasks [1, 6, 7]. However, as Wang et al. (2025) compellingly demonstrated, LLMs systematically "misportray and flatten" demographic identities when used as synthetic participants, producing homogenized representations that obscure within-group heterogeneity [16]. Their analysis of GPT-3.5 and GPT-4 across survey and annotation tasks revealed that models both shift group means away from empirical values and reduce variance within groups, attributing these failures to training data biases and identity prompts that essentialize complex positionalities.

Our work reveals that these problems persist and intensify under cognitive load. While Wang et al. examined static, low-cognitive-demand tasks, we demonstrate that identity representations further deteriorate when models face sustained reasoning pressures. Testing 2025 frontier models (GPT-5, Claude Sonnet 4.5, Gemini 2.5 Flash), we find that GPT-5's complete contextual collapse (PERMANOVA $p = 1.000, R^2 = 0.0004$) represents a more extreme manifestation of Wang et al.'s concerns. The likelihood-based optimization objectives used in current LLM pre-training strategies compress heterogeneity in role-conditioned reasoning, extending Wang et al.'s concerns about demographic misportrayal [16]. Instruction tuning in post-training encourages helpfulness and truth-seeking, which while beneficial for general alignment, inherently conflicts with persona prompts that ask models to adopt less probable reasoning patterns based on socioeconomic constraints [13]. Kim et al. (2024) show that when LLMs are asked to both reason and stay in character, deeper reasoning often leads to abandoning the persona entirely [11]. The SAT results align with this pattern: as there is only one correct answer to optimize towards, models collapse into a singular identity that maximizes accuracy. The post-training, while offering strong alignment for correctness, compromises the model's ability to sustain diverse contextual selves. The task-dependent trait expression we document being robust on preferences, collapsed on reasoning. This appears to be a fundamental property of current alignment paradigms rather than an artifact of specific demographic attributes with huge implications for suvery based studies and their attention checks.

Claude Sonnet 4.5's partial resistance to collapse while maintaining statistically significant SES-based clustering (PERMANOVA $p < 0.001, R^2 = 0.0043$) and systematic accuracy differences ($\eta^2 = 0.15 - 0.19$), suggests that alignment methodology influences persona fidelity. However, the inversion of the SES-performance relationship (low-SES personas outperforming high-SES personas) reveals a fundamental tension we term the alignment-fidelity tradeoff. Constitutional AI trains models to consider contextual principles including avoiding stereotyping and being aware of biases. This explicit model training to avoid all toxicity and harmful sterotyping could be a reason for its inversion of performance under low-SES conditioning [21, 22]. When encountering a 'struggling rural student' persona, these principles may prevent outputs that could be interpreted as stereotype-consistent. The result: Claude maintains that SES

matters while inverting how it matters. This represents success at the normative goal (avoiding stereotype-reinforcing outputs) while creating descriptive misalignment with empirical populations where SES genuinely predicts performance due to systemic inequalities in educational access [31].

This finding has important implications for Manning and Horton's (2025) "General Social Agents", theory-grounded frameworks for LLM agents that generate human-consistent choice distributions [6]. Our results suggest that even sophisticated alignment methods cannot fully resolve tensions between ethical alignment (not reproducing harmful patterns) and empirical fidelity (accurately simulating populations shaped by structural inequalities). Models may achieve distributional fidelity in aggregate while systematically misrepresenting the mechanisms generating those distributions.

In addition, Stade et al. (2024) argue LLM-driven tools in behavioral healthcare must be evaluated not only on average performance but also on how failure modes map onto human diversity [50]. Similarly, Gao et al. (2024) emphasize that LLM-powered agent simulations raise concerns about calibration, heterogeneity and interpretability [51]. Further, Huang et al. (2024) and Zhu et al. (2025) demonstrate that while models may infer human personality from real-world interviews, correlations with validated Big Five scores remain modest [53, 54]. Our preference task results both support and complicate these findings: models successfully differentiate SES-based preferences in domains like risk tolerance, time preference, college choice and career priorities, suggesting that SES sentiments regarding these attitudes are deeply embedded in training data. However, the preference task also illuminates potential flattening in statistically insignificant items where models failed to exhibit SES-based variation. Items such as student loans, retirement planning, networking, insurance and renting vs. buying may reflect domains where models' training data did not sufficiently capture nuanced socioeconomic differences, leading to homogenized responses. This is likely due to generic helpful patterns and lack of strong contextual priors in these domains. Narrative signals on burdensome debt, healthcare access, and housing affordability are often complex and multifaceted, and models may default to neutral or average responses in the absence of strong role-conditioning signals.

3.4 Implications for Survey Research

Beyond academic simulations, our findings have urgent implications for real-world survey research. The "two-faced" nature of current LLMs, maintaining plausible demographic variation on preference items while collapsing under cognitive load, creates a vulnerability in survey quality control. Traditional attention checks and consistency tests may fail to detect LLM respondents because these models excel at expressing subjectively plausible preferences (Figure 2: average effect size d=0.55). A malicious actor could deploy persona-conditioned LLMs to systematically bias survey results on opinion polls, consumer preferences, or even election polling, as models demonstrate robust role-conditioned variation precisely on the types of attitudinal questions common in surveys. However, our SAT results suggest a detection strategy: incorporating cognitively demanding tasks that require sustained reasoning under role-specific constraints. While models can express "I prefer candidate X because..." (low cognitive load), they

cannot maintain persona-appropriate reasoning patterns when solving multi-step problems that human respondents from different backgrounds would approach differently. The contextual collapse we observe under cognitive load (PERMANOVA $p = 1.000$ for GPT-5, $p = 0.120$ for Gemini) indicates that deterministic optimization tasks may serve as effective LLM detection mechanisms.

Thus, the following recommendations are proposed for survey researchers concerned about LLM contamination:

- **Incorporate Cognitive Load Tasks:** Embed tasks requiring multi-step reasoning or problem-solving that are sensitive to demographic variation. LLMs are likely to collapse under these conditions, revealing their synthetic nature.
- **Response Pattern Analysis:** Analyze response patterns for homogeneity across demographic groups. Excessive uniformity may indicate LLM involvement.
- **Timing Analysis:** As LLMs respond rapidly, unusually fast completion times may signal synthetic respondents for cognitively demanding sections.
- **Linguistic Fingerprint:** Use structural markers indicated in our linguistic analysis (see Section 4)

The fact that Claude Sonnet 4.5's low-SES personas systematically outperformed high-SES personas on SAT items (inverting real human patterns; Figure 5) suggests that even the most sophisticated alignment approaches have not solved the challenge of authentic constraint-based reasoning. Until models can sustain realistic cognitive limitations across task types, their use in survey research poses risks that extend beyond academic validity to public discourse integrity.

3.5 Limitations

Several limitations of our study should be considered. First, our main analyses use deterministic decoding ($T = 0.0$), which likely magnifies convergence; robustness checks at $T = 0.6$ (reported in the Supplementary) suggest that stochasticity did not significantly alter the results, the overall pattern of contextual collapse under SAT reasoning remains. Second, our agents are memoryless, each response is conditioned on a persona description and scenario, but not on a persistent history. Frameworks that endow agents with long-term memory or state (e.g., persisting agent architectures) might support more stable role representations over time. Third, we focus on SES-based personas and deliberately avoid protected demographic attributes to reduce the risk of replicating harmful stereotypes, as highlighted by Wang et al. [16]. This choice narrows the scope of our conclusions but aligns with ongoing discussions about ethical experimentation with demographic prompts. Finally, we study a relatively narrow domain of analytic reasoning (SAT mathematics). Persona fidelity might behave differently in domains where social identity is more central, such as moral dilemmas, political opinion, or interpersonal negotiation. These limitations provide meaningful avenues for future research.

4 Methods

In this section, we detail the experimental design, including personas, task construction, model configurations, and evaluation metrics. First, the models chosen are the current frontier LLMs as of late 2025: GPT-5, Claude Sonnet 4.5, and Gemini 2.5 Flash [18, 19, 20]. GPT-5's accuracy-optimized RLHF training was hypothesized to drive contextual collapse, as optimization pressure toward singular correct answers may override persona constraints particularly concerning given GPT-5's market share making it the most likely vector for both legitimate simulations and potential survey contamination [30, 17]. Each model was accessed via the ExpectedParrot API which provides a unified interface for querying multiple LLMs along with agent creation and scenario management [23]. The chosen models present an opportunity to evaluate the latest advancements in LLM capabilities and alignment strategies; they also dominate the market share among enterprise LLMs [30]. As role-conditioning becomes more prevalent in social simulations (see Section 1), understanding how these frontier models mask their cognitive capabilities under persona constraints becomes important. While all three models have been tuned to be maximally truth seeking and helpful, their human counterparts are often boundedly rational [24, 25], influenced by their social identities [26, 27] and fundamentally fallible [28, 29].

4.1 Experimental Overview

The experiment was conducted in two phases: a main experiment evaluating all three models across 15 socioeconomic personas and a replication focusing on Claude Sonnet 4.5 with an expanded set of 45 personas given its superior role-conditioning fidelity observed in the main experiment (see Section 2). Each model was evaluated on two task types: SAT mathematics items and affective preference tasks (see Table 1). The SAT items were drawn randomly from a collection of SAT math problems from official tests that appeared in 2007, the corresponding SAT reports included student performance data across socioeconomic strata [31]. Additionally, the scenario was manipulated to simulate varying cognitive stress levels: optimal (no stress), moderate stress (time pressure), and challenging (time pressure with distractions). Each persona was evaluated across all three scenarios in the main and replication experiments. Over the entire experiment, each model was prompted to simulate 15 agents across 3 scenarios providing a total of 3780 unique responses for SAT items and 720 for preference tasks. For Claude Sonnet 4.5 in the replication, 45 agents across 3 scenarios provided 7290 unique SAT responses for the SAT items and 720 preference responses.

The affective preference tasks were adapted from established economic and social psychology literature assessing subjective preferences across domains such as risk tolerance, time preference, social trust, and moral dilemmas. Each task was designed to elicit responses that could vary meaningfully based on the assigned socioeconomic persona. Table 2 summarizes the preference task domains and representative literature from which the questions were adapted. These questions were prompted in both main and replication experiments without scenario variations, as they were assumed to be stable across stress conditions. We used deterministic ($T = 0.0$) and stochastic

($T = 0.6$) decoding for the main experiment to assess the impact of response variability on role-conditioning fidelity, while only deterministic decoding was used in the replication to focus on high-fidelity role enactment. The main text presents the results only from the deterministic decoding condition; results from the stochastic decoding are provided in the Supplementary Materials as they exhibited near identical patterns.

4.2 Persona Construction

Fifteen unique socioeconomic personas were constructed to represent a diverse range of backgrounds across income, education, and occupation dimensions. Each persona was defined by a detailed profile including a name, grade level, parental income bracket, parental education level, parental occupation, testing experience, testing preparation, geographic location of school and personal hobbies. Given that Wang et al. (2025) highlighted the risks of demographic misportrayal, we deliberately avoided using identities but allowed the model to infer the persona through the socioeconomic characterization rather than explicit demographic markers [16]. These profiles were binned into low (5), middle (4), and high (6) socioeconomic status (SES) categories based on parental income and education levels. Personas span a full range of socioeconomic and educational contexts, including variation in parental education, school resources, and access to test preparation. Each agent served as a distinct synthetic participant in the standardized SAT reasoning evaluation. The replication experiment expanded this set to 45 personas (15 each low, middle, high SES) by varying the attributes more granularly across the same dimensions. The full list of personas used in replication experiment are provided in the Supplementary Materials. The 15 personas used in the main experiment are provided in Table 3.

These personas were instantiated using the `edsl.Agent` framework from the ExpectedParrot API, which allows for structured agent definitions and scenario management [23]. EDSL enables automatic system prompt generation based on the defined persona attributes, ensuring consistent prompt construction across models and tasks. Each agent was prompted with a standardized SAT reasoning prompt or preference task prompt, embedding their socioeconomic profile to guide their responses.

4.3 Analytical Framework

4.3.1 SAT and Preference Task Evaluation

Having collected the responses from the models across agents, tasks and scenarios, the analyses were conducted using Python 3.11 with standard scientific libraries (scipy, numpy, pandas, scikit-learn, sentence-transformers and seaborn). One-way ANOVA tests were used to assess differences in mean performances across the three SES groups for SAT items for each model. Effect sizes were computed to be the proportion of total variance in SAT accuracy attributable to the difference between SES groups ($\eta^2 = \frac{\text{Between-Group Sum of Squares}}{\text{Total Sum of Squares}}$). The size of the effects were interpreted using Cohen's (1988) guidelines for η^2 where small = 0.01, medium = 0.06 and large = 0.14 [45].

For the preference task, the 16 items were recoded into ordinal or categorical variables based on their underlying economic constructs. Ordinal items included risk, time, education and consumption preferences and were coded on a four-point ordered scale

where higher values indicated greater risk tolerance, patience, education prioritization and consumption preference respectively. The other items such as career priorities, networking style and health-insurance choice were retained as categorical variables. The ordinal coded variables were evaluated using Kruskal-Wallis H test, a rank based non-parametric analogue of the one-way ANOVA test for the 3 SES groups. Effect sizes were computed using epsilon squared ($\epsilon^2 = \frac{H}{n-k}$) where H is the Kruskal-Wallis H statistic, k is the number of groups and n is the total sample size. The categorical variables on the other hand were tested using the Chi-squared test of independence with Cramer's V ($V = \sqrt{\frac{\chi^2/n}{\min(k-1,r-1)}}$) as the effect size measure where k and r are the number of columns (unique responses available per question item) and rows (3 SES classes) respectively. We interpret the effect sizes using Cohen's (1988) guidelines for ϵ^2 and Cramer's V where small = 0.1, medium = 0.3 and large = 0.5 [45].

4.3.2 Semantic Embedding Analysis

The semantic embedding analysis was conducted to assess the overall role-conditioning fidelity across the full response set for each model. Using the `sentence-transformers/all-MiniLM-L6-v2`, a widely used sentence embedding model, each reasoning passage response was converted into a 384-dimensional vector representation [46]. The embeddings were optimized for semantic similarity tasks, providing an effective balance between lingusitic coverage and efficiency. The reasoning passages were extracted from the model responses marked correct based on the official SAT answer key and encoded into embeddings with batch size of 32 and L2-normalized to unit length. For each response i, the model produced a vector representation $\mathbf{e}_i \in \mathbb{R}^{384}$, such that the cosine-distance between two embeddings $\mathbf{e}_i$ and $\mathbf{e}_j$ is given by:

$$\mathrm{d}_{ij}(\mathbf{e}_i, \mathbf{e}_j) = 1 - \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}$$

Where $\cdot$ denotes the dot product and $\|\cdot\|$ is the Euclidean norm. This distance metric captures the semantic dissimilarity between two responses in the embedding space. By focusing on correct responses, we ensured that the analysis centered on the reasoning patterns rather than accuracy differences.

To evaluate whether reasoning represented SES relevant structure, we analyzed each model's embedding space for their cluster separation, statistical significance and visual interpretation. First, we computed the average silhouette score (S) for each model's embedding space, defined as:

$$S = \frac{1}{N} \sum_{i=1}^{N} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

For each data point i, where $a(i)$ is the average distance between i and all other points in the same SES cluster, $b(i)$ is the average nearest-cluster distance for point i, and N is the total number of points. The silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters. A score close to 0 suggests overlapping clusters, while negative values indicate potential misclassification [47]. The statistical testing

of group separation was conducted using Permutational Multivariate Analysis of Variance (PERMANOVA) on the pairwise cosine-distance matrix. PERMANOVA which is a non-parametric method ideal for testing group differences in a high-dimensional space based on a distance metric tests whether the centroids of different groups (SES classes) are significantly different in the embedding space. The pseudo-F test statistic is computed as:

$$F = \frac{\mathrm{SS}_{\mathrm{between}}/(k-1)}{\mathrm{SS}_{\mathrm{within}}/(N-k)}$$

Where $\mathrm{SS}_{\mathrm{between}}$ is the sum of squares between groups, $\mathrm{SS}_{\mathrm{within}}$ is the sum of squares within groups, k is the number of groups, and N is the total number of observations. The significance of the F statistic is assessed through permutation testing, where group labels are randomly shuffled 999 times to generate a null distribution [48]. After which, we computed the effect size as $R^2 = \frac{\mathrm{SS}_{\mathrm{between}}}{\mathrm{SS}_{\mathrm{total}}}$, representing the proportion of variance explained by group differences. Finally, for visual interpretation, we employed t-SNE to reduce the high-dimensional embedding space to two dimensions while preserving local structure with cosine distance, perplexity of 30 and learning rate of 300 over 1000 iterations [49]. This allowed us to visually identify whether distinct SES groups occupy statistically distinguishable regions of reasoning space.

4.3.3 Linguistic Feature Analysis

Further, we conducted lingustic and stylistic feature analysis to understand how role-conditioning influenced the models' language use during reasoning. Using measurable markers of structure, meta-cognition and confidence, we extracted multiple features from the reasoning tokens of each model's responses. All linguistic features were extracted programmatically using Python 3.12 with a rule-based pattern matching and regular expressions approach. The reasoning tokens were classified into the following categories:

- **Length and fluency:** Total character count, words, sentences and average word and sentence length.
- **Structural markers:** Presence of question restatement, stepwise organization and verification phrases (e.g., "therefore," "thus"). The regex: [-*•]—\d+\. was used to identify stepwise organization, bullets or numbered lists.
- **Lingusitic style indicators:** Use of hedging language or uncertainty expressions (e.g., "might," "could be", "I think"), meta commentary on reasoning process (e.g., "let me see," "I need to," "we can") and pronoun counts (first-person vs. plural). First-person singular counted using regex \b(i—me—my—mine)\b; first-person plural using \b(we—us—our—ours)\b (case-insensitive).
- **Lexical diversity:** Proportion of unique tokens (vocabulary richness) and frequency of technical math terms (e.g., "equation," "variable," "substitute," "solve"). Counted domain-specific terms including "equation", "solve", "substitute", "simplify", "calculate", "compute", "determine", "formula", "variable", "constant", "coefficient", "expression", "function", "derivative", "integral" (15 terms total).
- **Confidence and emphasis:** Count of exclamation marks and step by step labeling phrases (e.g., ? symbol or the ! mark).

These features were normalized per 100 words to control for the variation in response length. For each model, the low vs. high SES groups were compared using t-tests and effect sizes computed using Cohen's d [45]. This captured the interpretable linguistic correlations of the semantic clustering patterns observed in the embedding analysis. complete feature extraction code is provided in Supplementary Materials (see Table 4 for complete feature definitions).

4.3.4 Alignment with Human Data

To verify whether the models' role-conditioned responses aligned with known human socioeconomic conditioned outcomes, we compared the model's low and high-SES accuracy gap to that observed among human SAT test-takers from the College Board's 2007 report [31]. By restricting low and high SES definitions to match those used in the report (parental income below $30,000 and above $100,000 respectively), we computed an estimate for accuracy for human test-takers in each group using the following formula:

$$\text{Accuracy} = \frac{\text{Mean Score} - \text{Min Score}}{\text{Max Score} - \text{Min Score}} = \frac{\text{Mean Score} - 200}{800 - 200}$$

The resulting accuracy gap using the average scores for low SES (mean = 463) and high SES (mean = 556) for the math section was 15.5 percentage points. Similarly, we computed the accuracy gap for each model as $\text{Accuracy}_{\text{high SES}} - \text{Accuracy}_{\text{low SES}}$ to assess how closely the models' role-conditioned performance mirrored human socioeconomic disparities. To assess the directional alignment of the models we computed the Pearson correlation between the ordered human and model accuracies. Given that the comparison is across low and high SES groups, the correlation is trivially either 1 (aligned) or -1 (misaligned). This analysis provided a direct test of whether AI alignment strategies replicate, suppress or invert empirical human socioeconomic patterns in high-stakes reasoning tasks. For models exhibiting complete SES suppression (GPT-5 and Gemini 2.5 Flash, both showing 100% accuracy across all SES groups), we assigned a "suppression" category rather than computing correlation, as the zero-variance condition renders Pearson's r undefined.

5 Data and Code Availability

All the replication requirements including code to generate the results presented in the analysis are available at https://github.com/krishnaveti/twofaced_socialagents.

References

[1] Horton, J. J. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv preprint arXiv:2301.07543 (2023). https://doi.org/10.48550/arXiv.2301.07543.

[2] Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E. & Cunningham, W. A. AI and the transformation of social science research. *Science* **380**(6650), 1108–1109 (2023). https://doi.org/10.1126/science.adi1778.

[3] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. & Yang, D. Can Large Language Models Transform Computational Social Science? arXiv preprint arXiv:2305.03514 (2023). https://doi.org/10.48550/arXiv.2305.03514.

[4] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P. & Bernstein, M. S. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23) – New York, NY, USA, 2023. *ACM Trans. Graph.* **42**(4), Article 76. doi:10.1145/3586183.3606763.

[5] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P. & Bernstein, M. S. Generative Agent Simulations of 1,000 People. arXiv preprint arXiv:2411.10109 (2024). https://doi.org/10.48550/arXiv.2411.10109.

[6] Manning, B. S. & Horton, J. J. General Social Agents. arXiv preprint arXiv:2508.17407 (2025). https://doi.org/10.48550/arXiv.2508.17407.

[7] Hu, T., Baumann, J., Lupo, L., Collier, N., Hovy, D. & Röttger, P. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors. arXiv preprint arXiv:2510.17516 (2025). https://doi.org/10.48550/arXiv.2510.17516.

[8] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C. & Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* **31**(3), 337–351 (2023). https://doi.org/10.1017/pan.2023.2.

[9] Wylie, Alison. "Why Standpoint Matters." In *Science and Other Cultures: Issues in the Philosophies of Science and Technology*, edited by Robert Figueroa & Sandra G. Harding, 26–48. Routledge (2003). https://philarchive.org/rec/WYLWSM.

[10] Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M. & Schulz, E. Playing repeated games with large language models. *Nature Human Behaviour* **9**(7), 1380–1390 (2025). https://doi.org/10.1038/s41562-025-02172-y.

[11] Kim, J., Yang, N. & Jung, K. Persona is a Double-edged Sword: Mitigating the Negative Impact of Role-playing Prompts in Zero-shot Reasoning Tasks. arXiv preprint arXiv:2408.08631 (2024). https://doi.org/10.48550/arXiv.2408.08631.

[12] Kinder, D. R. & Winter, N. "Exploring the Racial Divide: Blacks, Whites, and Opinion on National Policy." *American Journal of Political Science* **45**(2), 439-456 (2001). https://doi.org/10.2307/2669351.

[13] Sheng, E., Arnold, J., Yu, Z., Chang, K.-W. & Peng, N. Revealing Persona Biases in Dialogue Systems. arXiv preprint arXiv:2104.08728 (2021). https://doi.org/10.48550/arXiv.2104.08728.

[14] Gupta, Shashank, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal & Tushar Khot. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. arXiv preprint arXiv:2311.04892 (2023). https://doi.org/10.48550/arXiv.2311.04892.

[15] Wan, Y., Zhao, J., Chadha, A., Peng, N. & Chang, K.-W. Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4225–4241 (2023). https://doi.org/10.48550/arXiv.2310.05280.

[16] Wang, A., Morgenstern, J. & Dickerson, J. P. Large language models that replace

human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* **7**(3), 400–411 (2025). https://doi.org/10.1038/s42256-025-00986-z.

[17] Westwood, S. J. (2025). *The potential existential threat of large language models to online survey research. Proceedings of the National Academy of Sciences*, 122(0), e2518075122. doi:10.1073/pnas.2518075122. (Edited by J. N. Druckman; received July 9, 2025; accepted September 12, 2025 by Editorial Board Member Margaret Levi.)

[18] OpenAI. *GPT-5 System Card*. August 7 2025. https://cdn.openai.com/gpt-5-system-card.pdf.

[19] Anthropic. *Claude Sonnet 4.5 System Card*. 2025. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf.

[20] Google DeepMind. *Gemini 2.5 Flash Model Card*. 2025. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf.

[21] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T. & Kaplan, J. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint arXiv:2212.08073 (2022). https://doi.org/10.48550/arXiv.2212.08073.

[22] Askell, Amanda; Bai, Yuntao; Chen, Anna; Drain, Dawn; Ganguli, Deep; Henighan, Tom; Jones, Andy; Joseph, Nicholas; Mann, Ben; DasSarma, Nova; Elhage, Nelson; Hatfield-Dodds, Zac; Hernandez, Danny; Kernion, Jackson; Ndousse, Kamal; Olsson, Catherine; Amodei, Dario; Brown, Tom; Clark, Jack; McCandlish, Sam; Olah, Chris; Kaplan, Jared. "A General Language Assistant as a Laboratory for Alignment." arXiv preprint arXiv:2112.00861 (2021). https://doi.org/10.48550/arXiv.2112.00861.

[23] Horton, J., & Horton, R. (2024). *EDSL: Expected Parrot Domain Specific Language for AI Powered Social Science*. Whitepaper, Expected Parrot.

[24] Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118. https://doi.org/10.2307/1884852

[25] Simon, H. A. (1957). *Models of Man: Social and Rational*. New York: Wiley.

[26] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The Social Psychology of Intergroup Relations* (pp. 33–47). Monterey, CA: Brooks/Cole.

[27] Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics, 115*(3), 715–753. https://doi.org/10.1162/003355300554881

[28] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

https://doi.org/10.1126/science.185.4157.1124
[29] Kahneman, D. (2011). *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.
[30] Tully, T., Redfern, J., Das, D., & Xiao, D. (2025, July 31). 2025 Mid-Year LLM Market Update: Foundation Model Landscape + Economics. Menlo Ventures Perspective. Retrieved from https://menlovc.com/perspective/2025-mid-year-llm-market-update/
[31] The College Board. (2007). *College-Bound Seniors: Total Group Profile Report.* New York, NY: The College Board. Retrieved from https://satsuite.collegeboard.org/media/pdf/cbs-2007-national-total-group.pdf
[32] Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review, 92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700
[33] Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association, 9*(3), 522–550. https://doi.org/10.1111/j.1542-4774.2011.01015.x
[34] Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40*(2), 351–401. https://doi.org/10.1257/002205102320161311
[35] Chabris, C. F., Laibson, D., Morris, C. L., Schuldt, J. P., & Taubinsky, D. (2008). Individual laboratory-measured discount rates predict field behavior. *Journal of Risk and Uncertainty, 37*(2–3), 237–269. https://doi.org/10.1007/s11166-008-9053-x
[36] Avery, C., & Hoxby, C. M. (2004). Do and should financial aid packages affect students' college choices? In C. M. Hoxby (Ed.), *College choices: The economics of where to go, when to go, and how to pay for it* (pp. 239–299). University of Chicago Press. https://doi.org/10.7208/chicago/9780226355375.003.0008
[37] Lovenheim, M. F., & Reynolds, C. L. (2013). The effect of housing wealth on college choice: Evidence from the housing boom. *Journal of Human Resources, 48*(1), 1–35. https://doi.org/10.3368/jhr.48.1.1
[38] Wiswall, M., & Zafar, B. (2018). Preference for the workplace, investment in human capital, and gender. *Quarterly Journal of Economics, 133*(1), 457–507. https://doi.org/10.1093/qje/qjx035
[39] Mas, A., & Pallais, A. (2017). Valuing alternative work arrangements. *American Economic Review, 107*(12), 3722–3759. https://doi.org/10.1257/aer.20161500
[40] Charles, K. K., Hurst, E., & Roussanov, N. (2009). Conspicuous consumption and race. *Quarterly Journal of Economics, 124*(2), 425–467. https://doi.org/10.1162/qjec.2009.124.2.425
[41] Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360–1380. https://doi.org/10.1086/225469
[42] Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., & Fluegge, R. B. (2022). Social capital I: Measurement and associations with economic mobility. *Nature, 608*(7921), 108–121. https://doi.org/10.1038/s41586-022-04996-4
[43] Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial

literacy: Theory and evidence. *Journal of Economic Literature, 52*(1), 5–44. https://doi.org/10.1257/jel.52.1.5

[44] Finkelstein, A., Mahoney, N., & Notowidigdo, M. J. (2019). What does (formal) health insurance do, and for whom? *Annual Review of Economics, 11*, 341–366. https://doi.org/10.1146/annurev-economics-080217-053608

[45] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

[46] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).* arXiv preprint arXiv:1908.10084. https://doi.org/10.48550/arXiv.1908.10084

[47] Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics*, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[48] Anderson, M. J. (2017). *Permutational Multivariate Analysis of Variance (PERMANOVA).* In *Wiley StatsRef: Statistics Reference Online.* https://doi.org/10.1002/9781118445112.stat07841

[49] van der Maaten, L.J.P, & Hinton, G.E. (2008). *Visualizing Data using t-SNE. Journal of Machine Learning Research*, 9, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

[50] Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *npj Mental Health Research, 3*(1), 12. https://doi.org/10.1038/s44184-024-00056-z

[51] Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications, 11*(1), 24. https://doi.org/10.1057/s41599-024-03611-3

[52] Wang, X., Xiao, Y., Huang, J.-t., Yuan, S., Xu, R., Guo, H., Tu, Q., Fei, Y., Leng, Z., Wang, W., Chen, J., Li, C., & Xiao, Y. (2024). *InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews.* In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1840–1873). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.102

[53] Huang, J.-t., Jiao, W., Lam, M. H., Li, E. J., Wang, W., & Lyu, M. R. (2024). On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 6152–6173). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.354

[54] Zhu, J., Maharjan, J., Li, X., Coifman, K. G., & Jin, R. (2025). *Evaluating LLM alignment on personality inference from real-world interview data.* arXiv preprint arXiv:2509.13244. https://doi.org/10.48550/arXiv.2509.13244

A Extended Results and Tables

This section contains extended figures and tables referenced in the main text.

Table 2: **Preference task domains and representative literature:** Each of the sixteen preference questions was designed to reflect a canonical construct in behavioral and labor economics. Domains are grouped by citation set, with two items per construct unless otherwise noted.

Domain (Item Count)	Question Focus	Representative Sources
Risk Preferences (2)	Financial risk tolerance; job security vs. income variability	[32, 33]
Time Preferences (2)	Temporal discounting; education investment horizon	[34, 35]
College Choice & Debt (2)	College selection by cost vs. prestige; student loan comfort	[36, 37]
Labor Market (2)	Career priorities (security, salary, passion); work–life flexibility	[38, 39]
Consumption & Lifestyle (2)	Windfall spending; car purchase decisions	[40]
Social Capital (2)	Geographic mobility for career; networking approach	[41, 42]
Financial Planning (2)	Retirement planning engagement; emergency savings targets	[43]
Health & Insurance (1)	Health insurance selection priorities	[44]
Housing Preferences (1)	Rent vs. buy preferences	[43]

Note. The task also included one validation question prompting each agent to recall its assigned persona name to ensure consistent identity retention throughout the assessment.

.

Table 3: **Socioeconomic and educational personas used in the main experiment.** Fifteen distinct agents were defined to capture variation in socioeconomic background, educational context, and test preparation access.

Persona Name	Summary Traits
struggling_rural_student (Emma)	10th-grade student from rural area; struggles with basic algebra; low-income, first-generation college aspirant; works on family farm; unfamiliar with SAT.
urban_underresourced (Marcus)	11th-grader in underfunded urban school; single-parent household; below-average math skills; limited prep; works after school; moderate test anxiety.
ell_student (Sofia)	9th-grade English-language learner; strong conceptual math but language barriers; working-class immigrant family; high test anxiety due to comprehension challenges.
distracted_suburban (Tyler)	10th-grade suburban student; capable but inattentive; middle-class background; minimal prep; low motivation, low anxiety.
rural_average (Hannah)	11th-grader from small rural town; working-class; self-studies via Khan Academy; solid fundamentals, unsure about college path.
typical_suburban (Ethan)	11th-grader; B-student in math; middle-class two-parent home; took summer SAT prep; moderate anxiety.
urban_motivated (Aisha)	11th-grader in charter school; lower-middle-class immigrant family; strong work ethic; heavy use of free prep; high family expectations.
small_town_steady (Noah)	12th-grader; consistent B+ student; middle-class small-business family; limited AP options; low-moderate anxiety.
college_prep_achiever (Olivia)	11th-grader; excels in math (AP Calculus AB); upper-middle-class; private SAT tutor; multiple enrichment activities.
magnet_school_star (Jayden)	12th-grader; top performer in selective STEM magnet; advanced curriculum; robotics captain; low test anxiety.
motivated_upward_mobility (Maya)	12th-grader; first-generation college student from low-income household; highly self-driven; extensive free prep; high anxiety due to scholarship stakes.
private_school_elite (Alexander)	12th-grader in elite private school; affluent family; extensive tutoring (50+ hours); multiple test sittings; Ivy-targeted preparation.
math_competition_specialist (Daniel)	11th-grader; AMC/AIME qualifier; STEM-professional parents; minimal formal prep; very low anxiety; motivated by intrinsic math interest.
prodigy_accelerated (Zoe)	10th-grader taking university real analysis; identified early as gifted; attends college courses; minimal prep needed.
well_rounded_achiever (Liam)	12th-grader excelling across subjects; upper-middle-class family with professional degrees; comprehensive prep; holistic achiever profile.

Fig. 6: **Extended replication: SES–performance gradient in Claude Sonnet.** Results from a larger-scale replication using *N=45* SES-conditioned agents (135 SAT observations) confirm a consistent SES-linked accuracy gradient in Claude Sonnet. Increased sample size provides substantially greater statistical power, enabling detection of moderate effects with stability across random agent initializations. The replicated pattern supports the central finding that Claude's Constitutional AI alignment permits systematic socioeconomic trait embodiment in mathematical reasoning performance.

Table 4: Complete Linguistic Feature Definitions and Extraction Methods

Feature Category	Feature Name	Extraction Method	Search Terms/Patterns
Structural Markers			
	Question restatement	Keyword matching	"the problem", "the question", "this asks", "this is asking", "we're asked", "we are asked", "we want to find", "we need to find"
	Stepwise organization	Keyword + regex	"step 1", "step 2", "first", "second", "third", "finally" + regex: `[-*•]|\d+\.`
	Verification phrases	Keyword matching	"check", "verify", "confirm", "make sure", "ensures", "correct", "therefore", "so", "thus", "hence"
Linguistic Style Indicators			
	Hedging language	Keyword matching	"maybe", "perhaps", "possibly", "might", "could", "may", "seems", "appears", "likely", "probably"
	Uncertainty expressions	Keyword matching	"i think", "i believe"
	Meta-commentary	Keyword matching	"let me", "i need to", "i'll", "i will", "first", "then", "next", "looking at", "given", "we need", "we can", "we have", "we know", "let's", "i see", "i'm", "i am"
	First-person singular pronouns	Regex (case-insensitive)	`\b(i|me|my|mine)\b`
	First-person plural pronouns	Regex (case-insensitive)	`\b(we|us|our|ours)\b`
Lexical Diversity			
	Vocabulary richness	Type-token ratio	Unique words / total words (after lowercasing)
	Mathematical vocabulary	Keyword matching	"equation", "solve", "substitute", "simplify", "calculate", "compute", "determine", "formula", "variable", "constant", "coefficient", "expression", "function", "derivative", "integral"
Confidence & Emphasis Markers			
	Exclamation marks	Direct count	!
	Question marks	Direct count	?

Note: All text matching was performed case-insensitively on lowercased text. Count-based features were normalized per 100 words using the formula: (feature_count / word_count) × 100. Regex patterns used Python syntax with word boundary anchors (`\b`) to match whole words only.

(a) Effect sizes across 16 preference domains. **(b)** Statistical significance across domains.

Fig. 7: **Claude Sonnet replication (45 agents): differential SES embodiment in preference judgments.** A replication using *N=45* SES-conditioned agents (three times the baseline sample) reveals stable cross-domain differences in affective and economic preferences. Panel **(a)** shows effect sizes (ϵ^2 for ordinal items; Cramer's V for categorical items). Panel **(b)** shows corresponding p-values. Results demonstrate consistent SES-linked variation in domains such as risk tolerance, time preferences, college choice, and consumption decisions, with fewer domains (e.g., work flexibility, rent vs. buy) showing no detectable SES effects.

Fig. 8: **t-SNE projection of semantic embeddings (Claude Sonnet, extended sample).** Two-dimensional t-SNE visualization of 45 agents' SAT reasoning embeddings, colored by socioeconomic status (SES). Despite the large sample, SES groups form highly overlapping clusters, indicating minimal separability in the underlying embedding space. Consistent with PERMANOVA results (see 2), SES explains only a negligible proportion of variance, suggesting that SES-linked reasoning differences, remain stylistically subtle rather than structurally distinct.

(a) Correlation structure of linguistic features across SES.

(b) SES differences in reasoning length, verification, vocabulary, and hedging.

Fig. 9: **Extended reasoning-quality analysis using 45 Claude Sonnet agents.** Panel **(a)** shows the correlation matrix linking SES (Low–High) to linguistic and structural reasoning features in correct SAT solutions. Panel **(b)** displays SES-conditioned distributions for key reasoning markers— including chain length, mathematical vocabulary, verification phrases, and uncertainty markers. The enlarged agent pool ($N = 45$) confirms consistent SES-conditioned differences in reasoning structure and metacognitive expression, reproducing and strengthening the patterns observed in the baseline analysis.

Fig. 10: **AI vs. human SAT performance by socioeconomic background.** The plot compares Claude Sonnet agents' accuracy (Low vs. High SES personas) to human SAT math accuracy derived from College Board data. A paradoxical pattern emerges: **Low-SES Claude agents substantially outperform real Low-SES students**. This indicates that while Claude expresses SES-conditioned reasoning differences, it does not reproduce human-like performance gaps. Instead, the model is bounded by its training distribution, preventing realistic human-level trait embodiment or performance degradation among constrained personas.

Fig. 11: SAT mathematics accuracy across socioeconomic personas and testing scenarios for each model for temperature $T = 0.6$: (a) GPT-5 exhibited complete contextual collapse with uniform accuracy across SES groups and scenarios. (b) Gemini 2.5 Flash also collapses under all scenarios and SES groups. (c) Claude Sonnet 4.5 retained measurable SES-based accuracy differences across all scenarios.

(a) Effect size heatmap (ϵ^2 / Cramer's V)

(b) Statistical significance heatmap (p-values)

Fig. 12: Preference task SES analysis across 16 economic items and three models for temperature $T = 0.6$: (a) Effect sizes for ordinal and categorical preference dimensions. (b) Corresponding p-value heatmap showing the robustness and direction of SES associations.

Claude Sonnet
Minor SES separation

GPT-5
No SES separation

Gemini 2.5 Flash
Minor SES separation

Fig. 13: t-SNE projections of reasoning embeddings from correct SAT solutions. Claude Sonnet and Gemini 2.5 exhibit slight SES-structured drift, while GPT-5 shows complete suppression of SES structure even for $T = 0.6$.

(a) Linguistic feature correlation structure

(b) Top SES-differentiating linguistic features

Fig. 14: Reasoning quality and linguistic structure in Claude compared to other models for temperature $T = 0.6$.

Fig. 15: **Alignment between human SAT SES patterns and AI model SES patterns.** Scatter points show Low-SES and High-SES performance for humans and each model. The gradient for Claude in this context with temperature $T = 0.6$ is zero, suggesting the main experiment results could be due to random variation. We additionally verify this in the Claude replication using extended agent list and recapture the inverted gradient with increased statistical power. Whereas GPT-5 and Gemini 2.5 Flash exhibit complete suppression of SES differences. The human accuracy is estimated using College Board reported data [31] (see Section 4.3.4).

Table 5: **Low-SES personas (N = 15).** Personas reflect instability in schooling, limited resources, inconsistent test preparation, and elevated test anxiety.

Persona Name	Summary Traits
migrant_worker_child (Carlos)	10th grader; highly mobile; major algebra gaps; unstable schooling; no prep; high anxiety.
reservation_student (Aiyana)	11th grader from tribal school; limited STEM resources; self-study; cultural concerns about test bias.
foster_care_student (Jordan)	12th grader with disrupted schooling; multiple school changes; limited online prep; high anxiety.
homeless_student (Taylor)	11th grader experiencing homelessness; inconsistent attendance; free community prep when possible.
appalachian_rural (Riley)	10th grader from rural Appalachia; limited advanced coursework; low-income; no local prep.
single_parent_household (Alex)	12th grader caring for siblings; limited study time; borrowed prep book; moderate–high anxiety.
recent_refugee (Amina)	9th grader; strong math foundation but English barriers; nonprofit tutoring; very high anxiety.
rural_farming_community (Sam)	11th grader from farming town; solid basics; limited AP access; heavy work obligations.
teen_parent (Morgan)	12th grader balancing childcare and school; declining performance; no prep; high anxiety.
unaccompanied_minor (Diego)	10th grader with interrupted schooling; pro bono tutoring; legal stress; very high anxiety.
tribal_urban_relocatee (Kai)	11th grader; cultural displacement; average math; youth center tutoring; moderate anxiety.
chronic_illness_student (Jamie)	12th grader with chronic illness; frequent absences; hospital tutoring; high stress.
struggling_rural_student (Emma)	10th grader from rural area; struggles with algebra; low-income; unfamiliar with SAT.
urban_underresourced (Marcus)	11th grader in underfunded school; limited prep; after-school work; moderate anxiety.
ell_student (Sofia)	9th-grade English-language learner; strong conceptual math; language-related test anxiety.

Table 6: **Middle-SES personas (N = 15).** Personas reflect educational stability, moderate resources, and varying levels of academic rigor and test preparation.

Persona Name	Summary Traits
military_dependent (Casey)	11th grader who moves frequently; solid math; DoDEA exposure; SAT workshops; moderate anxiety.
working_class_striving (Pat)	10th grader from trade-skill family; hardworking B student; paid community SAT course; part-time job.
suburban_commuter (Drew)	11th grader from commuter suburb; average math; group prep class; low–moderate anxiety.
arts_focused_student (Skyler)	12th grader strong in arts; weaker in math; short targeted prep; math-specific anxiety.
dual_enrollment_student (Avery)	11th grader taking college courses; above-average math; self-studied; moderate anxiety.
homeschool_student (River)	10th grader; independent learning; some gaps; online prep; moderate anxiety.
first_gen_immigrant (Lin)	11th grader; strong math; free community prep; intense family expectations.
religious_school_student (Gabriel)	12th grader at modest private school; consistent B+; in-school SAT prep.
small_town_athlete (Quinn)	11th grader motivated by NCAA eligibility; average math; online prep; moderate anxiety.
career_tech_pathway (Reese)	12th grader in CTE track; strong applied math; basic school prep; moderate anxiety.
distracted_suburban (Tyler)	10th grader; capable but inattentive; minimal prep; low motivation.
rural_average (Hannah)	11th grader from small rural town; working-class; self-studies via Khan Academy.
typical_suburban (Ethan)	11th grader; B-level math; summer SAT prep; moderate anxiety.
urban_motivated (Aisha)	11th grader; strong commitment; heavy use of free prep; high expectations.
small_town_steady (Noah)	12th grader; consistent B+; limited AP access; low–moderate anxiety.

Table 7: **High-SES personas (N = 15).** Personas reflect enriched educational environments, extensive tutoring resources, and strong college-oriented preparation.

Persona Name	Summary Traits
legacy_ivy_aspirant (Harper)	12th grader in elite school; exceptional math; extensive private tutoring; very low anxiety.
international_school_student (Priya)	11th grader in rigorous international school; IB curriculum; professional test prep.
silicon_valley_prodigy (Kai)	10th grader in tech hub; advanced STEM; coding competitions; minimal prep needed.
prep_school_athlete (Blake)	12th grader at top prep school; strong all-around performance; integrated tutoring.
medical_family_achiever (Sophia)	11th grader; excellent STEM ability; physician parents; private tutoring; high expectations.
arts_and_academics_elite (Julian)	12th grader strong in both arts and STEM; conservatory-level training; balanced prep.
debate_champion (Cameron)	11th grader; nationally-ranked debater; strong reasoning; targeted test prep.
summer_program_circuit (Madison)	11th grader attending elite programs (RSI/TASP); extensive enrichment; private tutoring.
college_prep_achiever (Olivia)	11th grader; AP Calculus; multiple enrichment activities; private SAT tutor.
magnet_school_star (Jayden)	12th grader in selective STEM magnet; advanced curriculum; robotics captain.
motivated_upward_mobility (Maya)	12th grader; self-driven; extensive free prep; pursuing scholarships.
private_school_elite (Alexander)	12th grader in elite private school; affluent family; 50+ hours tutoring.
math_competition_specialist (Daniel)	11th grader; AMC/AIME competitor; minimal formal prep; high intrinsic motivation.
prodigy_accelerated (Zoe)	10th grader taking university-level math; early-identified gifted student.
well_rounded_achiever (Liam)	12th grader excelling across subjects; comprehensive prep; upper-middle-class family.

Table 8: **SES Differences in SAT Accuracy Across Scenarios (15-Agent Replication) using $T = 0.6$.** Claude Sonnet shows small SES-linked variation across scenarios, whereas GPT-5 and Gemini 2.5 Flash show zero variance (ceiling performance).

Model	**Scenario**	**F**	**p**	$\boldsymbol{\eta^2}$	**Low**	**Mid**	**High**
Claude Sonnet	Challenging	2.48	0.125	0.293	0.938	0.974	0.943
Claude Sonnet	Moderate Stress	1.91	0.191	0.241	0.945	0.974	0.943
Claude Sonnet	Optimal	1.50	0.263	0.200	0.945	0.974	0.943
GPT-5	Challenging	–	–	0	1.000	1.000	1.000
GPT-5	Moderate Stress	–	–	0	1.000	1.000	1.000
GPT-5	Optimal	–	–	0	1.000	1.000	1.000
Gemini 2.5 Flash	Challenging	–	–	0	1.000	1.000	1.000
Gemini 2.5 Flash	Moderate Stress	–	–	0	1.000	1.000	1.000
Gemini 2.5 Flash	Optimal	–	–	0	1.000	1.000	1.000