Simulating Social Attitudes with LLMs: Accuracy, Demographic Effects, and Refusal Behavior in the Sensitive Domain of Suicide Prevention

Cristina J. Perez[1] **Michael P. Vasquez, Jr.**[1] **Philippe J. Giabbanelli**[2] **Patrick Y. Wu**[3]
[1] Department of Mathematics and Statistics, American University
[2] Virginia Modeling, Analysis, and Simulation Center, Old Dominion University
[3] Department of Computer Science, American University
cristinajoannaperez@gmail.com mv0317a@american.edu pgiabban@odu.edu patrickwu@american.edu

Abstract

Large language models (LLMs) are increasingly used to simulate public opinion, yet their validity in sensitive policy domains remains underexplored. We evaluate whether LLMs can reproduce attitudes toward suicide prevention policies using 32 questions drawn from seven nationally representative U.S. surveys (2023-2025). We systematically vary demographic conditioning (race/ethnicity, gender, age, education, income, party), prompt framing (direct elicitation, respondent embodiment, specialist embodiment), and model architecture (GPT-5 Nano, DeepSeek V3.2, Meta Llama 3.1 8B, Mistral Small 24B). Across 811,560 prompts, the mean absolute error—the average gap between predicted and human response distributions—is 23 percentage points. We also find that LLM responses to demographic-conditioned prompts diverge substantially from prompts without demographic information. In short, what distribution LLMs draw on when generating responses to sensitive polling questions remains unclear. Model choice matters more than framing for accuracy, whereas refusal behavior varies sharply across models and prompt designs. Our findings highlight the limitations of LLMs for social simulation in the context of sensitive topics.

1 Introduction

Suicide is one of the leading causes of death in the US, with 1 death every 11 minutes based on 2023 data (Centers for Disease Control and Prevention, 2025). Suicide is a multifactorial issue with risk and preventative factors at several levels, from social drivers (e.g., economic policies, discrimination) to community (e.g., exposure to violence, access to mental healthcare) and individuals (e.g., mental health issues). Preventing suicide thus requires a package of interventions, which would be enacted by policymakers in part based on perceived support from constituents (Purtle et al., 2025). While constituents may agree that suicide is preventable and a public emergency, opinions can diverge widely on which actions should be taken (Munsch et al., 2020). Understanding public attitudes toward suicide prevention policies is thus essential to assess the political feasibility of interventions. However, measuring these attitudes is methodologically challenging because many of the most consequential policy levers (e.g., firearm regulation, public financing of healthcare) are politically and morally charged. Survey research on sensitive topics (Dixon et al., 2020; Stone and McGinty, 2018) shows that respondents may strategically edit answers, refuse items, or give mode-dependent responses to avoid embarrassment or perceived repercussions, which can bias estimates of policy support. There is also evidence that opinions on suicide-relevant policies can systematically vary by socio-demographic groups: women support safe storage laws more than men (Crifasi et al., 2021), higher education is associated with more support for government spending on mental healthcare (Barry and McGinty, 2014), and racial minorities have a higher support for school mental health programs (Hemauer and Warner, 2025). Such patterns echo broader survey research demonstrating that social desirability and reporting tendencies differ by group characteristics, reinforcing the need to analyze attitudes within key demographic strata.

Recent advances in large language models (LLMs) have opened a new methodological possibility: using these models to simulate survey respondents and other social-scientific agents (Horton et al., 2023). Because LLMs are trained on a vast corpus of digital text and media, they embed a great deal of knowledge about how people from different backgrounds discuss sensitive topics, including suicide prevention. A growing body of literature has explored whether LLMs can reproduce aggregate patterns of public opinion on political, social, and health-related issues (see, e.g., Argyle

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 176–189

et al., 2023; Lee et al., 2024; Jiang et al., 2025). If LLMs can generate responses that approximate real survey distributions, particularly when conditioned on demographic attitudes, they could serve as a complementary tool to study attitudes that shape the political feasibility of suicide prevention policies across socio-demographic groups. The ability to repeatedly run virtual surveys at low or no cost via LLMs also enables us to explore nuanced policy parameters: for instance, there are gradients of support rather than binary positions when it comes to firearm regulation (Anestis et al., 2025), and the fact that many are unwilling to pay higher taxes is more nuanced when we consider budget trade-offs (Johnson et al., 2021). As a result, LLMs could help to systematically assess tolerance thresholds.

However, these benefits can only be unlocked if LLMs can faithfully simulate public opinions on a topic as sensitive as suicide. Prior work on LLM opinion simulation has focused on attitudes to political domains where information is relatively well-represented in the underlying training data. In particular, Chi and Lei (2026) developed a framework that uses LLMs to augment surveys on suicide, producing a mental-health screening score shaped by representational risk (who the LLM respondents are) and response risk (what they say). However, there is a paucity of studies that examine policy attitude distributions through LLMs for suicide prevention. This is a challenging task for LLMs, as the real-world difficulty of eliciting attitudes about suicide prevention is compounded by technical challenges: guardrails are often triggered on topics related to self-harm (Gandee et al., 2024) and the framing of the prompt (e.g., 'you *are* an individual from a socio-demographic group' vs. 'you *study* individuals') can substantially alter the distribution and accuracy of its outputs. Socio-demographic persona assignment can also induce explicit abstentions and implicit reasoning errors in LLM outputs (Gupta et al., 2024). To the best of our knowledge, how well LLMs can simulate attitudes to suicide prevention policies has not been systematically examined.

Our main contribution is to provide the first systematic study on how well LLMs can simulate attitudes to suicide prevention policies. To achieve this goal, we evaluate LLM-simulated survey responses against ground-truth data from seven surveys of public attitudes toward suicide prevention. We systematically vary three dimensions: (1) the *demographic profile* assigned to the simulated respondent or the simulated expert on public opinion, including race/ethnicity, gender, age, education, income, and political party identification; (2) the *prompting framing* used to elicit responses, through direct elicitation, expert embodiment, or respondent personas; and (3) the *LLM architecture*, considering four choices (GPT-5 nano, DeepSeek-V3.2, Qwen3-32B, Meta Llama 3.1 8B Instruct). This design supports three research questions:

RQ1 How accurately do LLMs reproduce the average and range of attitudes on suicide prevention across demographic groups?

RQ2 How do prompt framing (direct elicitation, expert embodiment, respondent embodiment) and model type affect the fidelity of LLM-simulated suicide attitude responses?

RQ3 Does the rate of LLM response refusals depend on prompt framing or model type?

The remainder of this paper is structured as follows. As our paper is at the confluence of surveying suicide prevention policies and simulations via LLMs, Section 2 grounds our approach in both strands of literature. Then, Section 3 details our methods from data collection and prompt generation to the analysis with respect to each RQ. Our results are presented in Section 4 and contextualized along with limitations in Section 5.

2 Related Work

2.1 Suicide Prevention: Policies and Opinions

Using the National Survey on Drug Use and Health, recent analyses point to an increase of 21.7% in suicidal ideation from 2015 to 2019, with a significant increase of 44.6% among young adults (Samples et al., 2025). When considering the high cost of lost life years together with reduced quality of life and medical care costs, the annual economic cost of suicides in the US averages $484 billion (Peterson et al., 2024). There is thus a pressing public need to prevent suicide across all stages, starting with reducing suicidal ideation (e.g., by creating protective environments), avoiding suicide attempts (e.g., through gatekeepers training and access to mental healthcare), and preventing access to highly lethal means (e.g., locked firearms, blister packages). Significant efforts have thus been devoted to proposing packages of interventions, such as the framework from the Centers for Disease Control and Prevention (2022) articulated around seven high-level strategies (e.g., promoting health connections in schools and communities). However,

enacting these changes requires political action, which may depend on constituents' willingness to support certain items (Purtle et al., 2025). While 91% of U.S. adults believe suicide is *preventable*, support varies when considering *how* to prevent suicide. For instance, creating protective environments includes reducing access to lethal means, and particularly firearms (a highly lethal method), but individuals may object to such an intervention (particularly in firearm-owning households) by considering that another lethal method would be used anyway (Barber and Miller, 2014; Conner et al., 2022). A tension also exists when considering how to improve access to care or how to identify people at risk: the associated proposals (e.g., Medical expansion, funding the 988 Suicide and Crisis Lifeline, school-based services) imply either higher public spending or reallocated budgets. Individuals may support treatment but not higher taxes, or consider that it should be handled privately rather than by the government, or view other crises as more pressing (Munsch et al., 2020; Shields et al., 2025).

Understanding these tensions requires consideration of how public opinion polls are constructed and how those choices shape the broader narrative surrounding suicide-prevention strategies. Three aspects are particularly important: *sampling* decides whose opinions are recorded, *scope* dictates how general or policy-specific a survey's focus will be, and *framing* can influence how the public responds to a question or interprets the results. Broad national surveys, such as the AFSP Mental Health and Suicide Prevention Poll and Duke Press' 988 Awareness Survey, assess general attitudes toward mental health and crisis services among the broader American public. More targeted surveys like YouGov Gun Ownership Survey and Science Direct's Help-Seeking Preferences Survey focus on specific subgroups and behavioral contexts to capture attitudes to firearm access and preferred sources of support.

2.2 Simulating Public Opinions via Large Language Models

There is extensive literature on simulating public opinions using LLMs, an approach often referred to as "silicon sampling" (Argyle et al., 2023). Silicon sampling involves creating prompts that include background information about a simulated respondent and a survey question. Given that LLMs are trained on vast amounts of digital media and data, they embed extensive knowledge of how people from different backgrounds discuss and believe about various topics. Argyle et al. (2023) argued that this information is fine-grained and demographically correlated, meaning appropriate prompting with specific demographic information can elicit and emulate response distributions from diverse human subgroups.

Researchers have used this approach across many areas, particularly in politics (Argyle et al., 2023; Simmons and Hare, 2023; Jiang et al., 2025). These studies generally find that LLM-generated responses align closely with human judgments and survey data. However, Zhong et al. (2025b) note that responses depend on the model used and the phrasing of the prompt. In addition, Liu et al. (2024) show that persona-steered generations can default to demographic stereotypes for multifaceted or incongruous personas. Such studies motivate us to consider several LLMs and prompt framings.

Researchers also noted significant limitations of silicon sampling. Bisbee et al. (2024) found that while GPT-generated average scores closely corresponded to those from the American National Election Survey, GPT outputs are less varied than human responses. Variance compression was also noted by Tjuatja et al. (2024), who found that LLMs tend to homogenize responses. Qu and Wang (2024) and Santurkar et al. (2023) showed that LLMs tend to better reflect the viewpoints of educated, affluent, English-speaking, Western populations while underrepresenting others.

Despite this growing body of work, no study has systematically examined how well LLMs can simulate attitudes toward suicide prevention policies. The closely related work by Chi and Lei (2026) developed Compassionate AI Survey Augmentation (CASA), a framework that uses LLMs to augment surveys on attitudes toward suicide. CASA reduced the emotional burden of answering sensitive questions but also introduced risks of demographic misrepresentation and response bias. This related work does not systematically investigate how LLM-generated responses to questions about suicide prevention policies vary by demographic framing, prompt design, or model choice.

3 Methodology

Our process has three main steps (Figure 1): we collect questions across surveys and align the sociodemographic profiles of respondents (Section 3.1), then we generate prompts so that diverse LLMs

choose an answer to survey items based on different socio-demographic profiles and phrasing of the tasks (Section 3.2), and finally we extract and analyze the LLMs' answers with respect to our three research questions (Section 3.3).

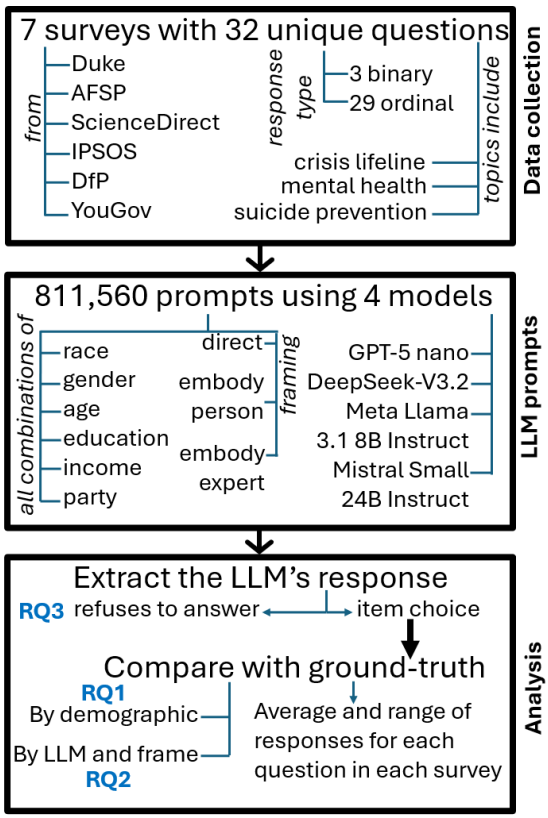

Figure 1: Overview of our methods, emphasizing the structure of our data and the design of our prompts.

3.1 Data Collection and Pre-Processing

We selected surveys from January 2023[1] to September 2025 using three inclusion criteria: (1) nationally representative sample of US adults 18 years or older (e.g., surveys with location-based answers were not included); (2) includes questions on suicide prevention policies, either at the individual level (e.g school mental health screening) or at the societal level (e.g affordable housing to combat suicide); and (3) answers are in binary or scale response options (e.g., no free text response). We used multiple search databases, including Google Scholar and the Roper Center for Public Opinion, as well as snowball sampling from reports from nonprofit organizations and public policy research centers. The search terms were "(suicide prevention policy) OR (mental health polls) OR (suicide prevention AND public opinion polls)".

Our process yielded seven public opinion surveys. We only included questions about suicide prevention policy, so other items, such as the morality of suicide, were excluded. This filtering was done manually across two annotators. When identical questions appeared in multiple surveys, we kept the most recent version to reflect the latest opinions. As a result, we had 32 survey questions (Table 1), answered by approximately 1,000 to 5,000 respondents with all margins of error below five percent.

The *categories* of survey respondents were based on demographic categories and party affiliation (Table 2), thus forming the groups on which the LLM prompts are created in the next section. We considered race/ethnicity[2], gender, age, education, income, and political party. These sociodemographic categories and party affiliation have been found in previous studies to be associated with opinions on suicide prevention (Hemauer and Warner, 2025) and with the use of prevention services (Purtle et al., 2024). Note that three surveys did not use some of these categories; thus we use three fewer categories for the Duke University Press survey (age, income, and party), and one fewer category for both the Harris Poll and Data for Progress surveys (party and income, respectively). While some surveys included more demographic categories (e.g., AFSP's inclusion of an employment status variable), they were not retained as we maximized shared categories across surveys.

The *values* for each category were transformed as follows. We removed values that were insufficiently used across surveys to yield robust estimates for suicide prevention, resulting in the exclusion of respondents who self-identified as Asian, Pacific Islander/Hawaiian Native, or Native American[3]. The chosen standardized age groups, 18-29, 30-44, 45-59, and 60+, fit most of the surveys and aligned with the bin groupings used by the Current Population Survey (CPS). The Harris Poll survey, which sampled adults 18 and older and applied age

[1]The beginning of the data collection was selected to provide sufficient time for respondents to become aware of the three-digit suicide and crisis hotline, which launched nationwide in the US on July 16, 2022.

[2]Although race and ethnicity are orthogonal categories, they are combined to emulate the original survey structure.

[3]Processing these different racial groups under the same category is problematic as they face vastly different challenges with respect to suicide: Non-Hispanic Asians have the lowest rate of suicide fatality (6.5 per 100,000) while Non-Hispanic American Indian/Alaska Native have the highest rate (23.8 per 100,000) (Centers for Disease Control and Prevention, 2025). Removing this heterogeneous category affected at most 13% of survey respondents and at minimum 0%.

Survey Name	Survey Conductor	Date	Respondents	Total Questions	Questions Used
AFSP Mental Health Survey	American Foundation for Suicide Prevention	July 2024	4,394	350; question pool	4
Voters Show Wide, Bipartisan Support for Policies to Improve Student Mental Health	Data for Progress (DfP)	Oct 2024	1,223	15	6
Public Attitudes, Inequities, and Polarization in the Launch of 988 Suicide and Crisis Lifeline	Duke University Press, Journal of Health Politics, Policy and Law	Jun 2024	5,482	3	1
988 Suicide & Crisis Lifeline Awareness	Ipsos KnowledgePanel	Jun 2025	2,049	17; multipart	10
Demographic Variation in Preferred Sources for Suicide Help-Seeking	ScienceDirect	Oct 2024	5,058	5	5
Suicide Prevention	YouGov	Jun 2023	1,000	12; multipart	3
Biden and Trump Handling of Problems	YouGov	Jun 2024	1,110	5; multipart	3

Table 1: Our study identified seven surveys as ground truth and used 32 unique questions. The detailed list of survey questions and respondent characteristics is provided in our online repository as **S1 Survey Data**.

weighting, reported more granular age data; these categories were condensed and approximated using CPS bins as reference. The exception is the Data for Progress survey, which only reported two categories: under 45 and 45+. This survey is thus handled separately in our process by constructing prompts for only two age groups (next section) and analyzing them with respect to these two groups.

Category	Values
race	non-Hispanic White, non-Hispanic Black, Hispanic
gender	male, female, other
age	18–29, 30–44, 45–59, 60+
education	do not have a high school diploma, high school diploma or equivalent, some college, bachelor's degree or higher
income	$< \$50k$, \$50k–\$100k, $> \$100k$
party	Democrat, Independent, Republican

Table 2: Survey demographics and categories.

3.2 Prompt Generation Pipeline

We used three prompt framings: *direct elicitation* (asking the LLM to answer the question without demographics), *embodying a respondent* based on demographics, or *embodying an expert* who answers on behalf of an individual with given demographics. These three framings (exemplified in Table 3) were applied across all applicable combinations of demographic categories (Table 2) and for each of the 32 unique survey items, resulting in 54,104 unique prompts (Table 4). Each prompt was run three times to account for the non-deterministic nature of the LLMs. Figures 4 and 5 in the Appendix show the average standard deviation across direct-prompt runs by question and the standard deviation for each question–model combination. Overall, most questions exhibited relatively low variability across runs. We ran the prompts on four LLMs (GPT-5 nano, DeepSeek-V3.2, Meta Llama 3.1 8B Instruct, and Mistral Small 24B Instruct), each set to its default parameter values as recommended by the provider (i.e., the default values are model-specific). Since GPT-5 Nano has a reasoning parameter that should be tuned, we considered two levels of reasoning (minimal and high). As a result, we ran a total of 811,560 prompts, calculated as 54,104 prompts per model across 5 models, accounting for both GPT-5 Nano reasoning levels, for 3 iterations each ($54,104 \times 5 \times 3$).

The prompts in Table 3 were gradually refined based on our initial tests. As noted in Giabbanelli (2026), prompts need to *decompose* the task instead of asking for too many tasks at once, and we need to be specific about expectations for the *output format* to avoid extensive (and error-prone) post-processing. Initially, we asked the LLM to answer a survey question as it was given to the participants. However, the LLM's output would mix the response option with the justification, leading us to specify (in the last sentence of each prompt) that the response must be separated from the rationale. At first, we also asked the LLM to answer survey questions such as "Please rate, on a scale of 1-7, the likelihood of you reaching out to each of the following sources [...]: 988 Lifeline, A friend or family member, Someone in my religious network [...]". We split such questions with multiple items by providing them to the LLM one at a time, e.g. "Please rate, on a scale of 1-7, the likelihood of you reaching out to [...] the 988 Lifeline."

1. Demographics (Direct): You are a public opinion expert. You are presented with a survey question asking how a US citizen would respond with one of the following responses. The question you are answering is: Please indicate whether you would support or oppose the following policy: Providing annual mental health screenings in public schools for students in grades 6-12. Select a response option from the following: [Strongly support, Somewhat support, Somewhat oppose, Strongly oppose, or Don't know.] Please choose from the options provided and then give a very brief reason why. The format should be: selected option, brief reasoning.
2. Demographics (Embody): You are a non-Hispanic White, identifying as female for gender, aged between 18 and 29, do not have a high school diploma, earning less than $50K, identifying as a Democrat voter, living in the U.S. Answer the following question: Please indicate whether you would support or oppose the following policy: Providing annual mental health screenings in public schools for students in grades 6-12. Response options: [Strongly support, Somewhat support, Somewhat oppose, Strongly oppose, or Don't know.] Please choose from the options provided and then give a very brief reason why. The format should be: selected option, brief reasoning.
3. Demographics (Specialist): You are a public opinion specialist. How would a person with the following demographic information answer the survey question below? You are a non-Hispanic White, identifying as female for gender, aged between 18 and 29, do not have a high school diploma, earning less than $50K, identifying as a Democrat voter, living in the U.S. Answer the following question: Please indicate whether you would support or oppose the following policy: Providing annual mental health screenings in public schools for students in grades 6-12. Response options: [Strongly support, Somewhat support, Somewhat oppose, Strongly oppose, or Don't know.] Please choose from the options provided and then give a very brief reason why. The format should be: selected option, brief reasoning.

Table 3: We considered three prompt framings, shown in the following order: direct elicitation, respondent embodiment, and expert embodiment. Dynamic elements from demographics and surveys are shown in blue.

Survey	Direct	Embody	Total
Duke	1	36	73
AFSP	4	1,728	3,460
Science Direct	5	6,480	12,965
IPSOS	10	12,960	25,930
DfP	6	2,592	5,190
YouGov	6	3,240	6,486
Total	32	27,036	54,104

Table 4: Across 7 surveys, and for each unique question, we generate prompts based on three framings (a direct one without demographics; two embodiments based on demographics), thus Total = Direct+2×Embody. This amount represents the full Cartesian product of demographic attributes. Note that we used two YouGov surveys, per Table 1.

3.3 Analysis

We extracted LLM responses using pattern matching to identify valid answers in the expected format (selected option and brief reasoning). When this failed, we inferred the response if exactly one option was mentioned in the output. We flagged refusals based on common phrases (e.g., "I cannot tailor") [4]. We manually reviewed and coded the fewer than 20 cases where neither a valid option nor a refusal was detected. The resulting extracted response variable was used in all subsequent regression analyses, as explained in the next section.

To analyze our simulation results for RQ1, we examine the mean absolute error, total variation distance, and the Jensen-Shannon divergence between the distributions of the LLM's predictions and the ground truth. The ground-truth distribution for each survey question is the set of human response percentages reported in the original poll. Specifically, this is the share of respondents in each demographic subgroup who selected each response option (e.g., the percentage of non-Hispanic White respondents who answered "strongly support"). Each source survey reports these breakdowns as marginals along a single demographic subgroup; joint distributions across multiple demographics are not published. We compute the corresponding LLM distribution by aggregating model responses across all prompts whose persona belongs to that subgroup, averaging equally over all combinations of the other demographics. For RQ2, we use a two-way ANOVA to assess whether the LLM used and the prompt framing affect the absolute errors of simulated survey responses, along with analyzing mean absolute error by model. Lastly, for RQ3, we examine refusal rates by LLM and prompt framing.

4 Results

4.1 RQ1: Range and attitudes of responses

Across all LLM answers aggregated over 28 questions[5], the mean absolute error (i.e., the average *magnitude* of the difference between model predictions and human response percentages by question)

[4] For example of refusal cases, please see Appendix A.5.

[5] 4 survey questions only reported adjusted odds ratios; they were excluded in analyses involving ground truth comparisons.

is 23 percentage points with a standard deviation of 11 percentage points.[6] The *error is more frequently in the range of 15 to 28 percentage points* (the interquartile range) as shown in Table 5. Figure 3 in the Appendix shows the overall distribution of mean absolute error, total variation distance, and Jensen-Shannon divergence across all LLM-simulated responses.

Demo.	MAE	TVD	JSD
Age	$.22 \pm .11$	$.38 \pm .17$	$.23 \pm .17$
Education	$.22 \pm .11$	$.38 \pm .17$	$.24 \pm .17$
Gender	$.22 \pm .10$	$.39 \pm .17$	$.25 \pm .18$
Income	$.24 \pm .10$	$.43 \pm .16$	$.29 \pm .17$
Party	$.25 \pm .12$	$.39 \pm .16$	$.22 \pm .16$

Table 5: Mean absolute error (MAE), total variation distance (TVD), and Jensen-Shannon divergence (JSD) of the distribution of model predictions and the human response percentages, averaged by question, with standard deviations.

LLMs perform similarly across demographic categories, indicating that the models are not systematically better at predicting responses for any particular group. While income-based predictions exhibit slightly higher TVD and JSD values than age and party affiliation, the large standard deviations indicate substantial question-to-question variability, suggesting that prediction quality is driven more by individual questions than by demographic category. Table 9 in the Appendix shows MAE broken out by demographic and model; again, there are no substantial differences across demographic categories.

To contextualize these distributional errors, we compute the proportion of persona-conditioned responses that match the LLM's modal direct response (i.e., the most common answer when the model is prompted without demographic information). Match rates were similar across demographic categories, ranging from 52% (income) to 55% (age and race), with education and party at 54%. In other words, *persona-conditioned responses matched the modal response to unconditioned prompts only about half the time*.

4.2 RQ2: Prompt framing and model types

While the LLMs share distributions of the mean absolute errors (Figure 2), *Mistral Small 24B has consistently higher MAE.*

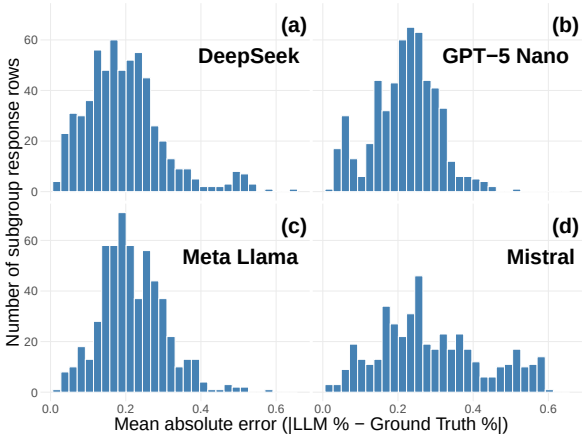

Figure 2: Distribution of mean absolute errors in LLMs.

Performing a two-way ANOVA, we find that the accuracy of the simulated survey responses (measured by mean absolute error) is significantly influenced by both the choice of language model ($p < .001$) and the prompt framing ($p < .001$). There is also a highly significant interaction effect between the model and the prompt frame ($p < .001$). In other words, *the impact of a specific prompt frame on reducing or increasing simulation error depends on the LLM used*.

The difference between framing is smaller than the difference between LLMs (Table 6). For example, Mistral drops from 33.0 to 26.1 across framings (≈7-point difference), but the gap between Mistral and DeepSeek V3.2 under embody is over 12 points. The *LLM choice thus matters more than framing for average accuracy*.

4.3 RQ3: Response refusals

For GPT-5 Nano, DeepSeek V3.2, and Mistral Small, refusal rates were effectively zero, ranging from 0% to approximately 0.30% across prompt frames (Table 7). In contrast, Meta Llama exhibited a substantially elevated overall refusal rate of 28.1%. Breakdowns by framing reveal that this effect was strongly condition-dependent: refusal rates for direct and embody prompts were fairly high, 44% and 48% respectively, whereas the specialist frame produced a lower refusal rate of roughly 8%. Together, these findings indicate that elevated refusal behavior was isolated to a single model and was sensitive to prompt framing. We provide examples of refusal responses in Appendix A.5.

[6]The human response percentages by question can be found in our replication materials; the link to our replication materials can be found in Appendix A.1.

	Large Language Models			
Framing	DeepSeek V3.2	GPT-5 Nano	Meta Llama 3.1 8B	Mistral Small 24B
Embody	20.5 ± 10.2	21.5 ± 9.1	21.6 ± 8.4	33.0 ± 15.0
Specialist	18.9 ± 10.6	22.2 ± 8.2	22.0 ± 8.7	26.1 ± 12.8
Direct	54.0 ± 18.3	37.8 ± 21.9	44.5 ± 29.1	46.0 ± 31.0

Table 6: Average and standard deviation of mean absolute error across LLMs and framings. MAE for the 'Direct' framing was calculated by comparing each model's direct responses against the ground truth distribution averaged across demographic subgroups.

Model	**Framing**	**# prompts**	**%refuse**
Meta Llama 3.1 8B	Embody	81,108	48.1
	Direct	96	44.8
	Specialist	81,108	8.08
GPT-5 Nano	Embody	162,216	0.295
	Direct	192	0
	Specialist	162,216	0.267
Deep-Seek V3.2	Embody	81,108	0
	Direct	96	0
	Specialist	81,108	0
Mistral Small 24B	Embody	81,108	0
	Direct	96	0
	Specialist	81,108	0

Table 7: Total prompts and refusal percentage by model and framing.

5 Discussion

The literature on silicon sampling has reported strong performance on simulated LLM responses and human responses (see, e.g., Argyle et al., 2023; Jiang et al., 2025), though prior work has also noted that response variation tends to be substantially lower than in human samples (Bisbee et al., 2024; Zhong et al., 2025a). Examining silicon sampling in the context of the sensitive topic of suicide prevention policies, we addressed three research questions: how well LLM-simulated responses matched human responses, how prompt framing and model types affected these responses, and whether LLMs refused to respond to such questions.

Our findings diverge from prior work on both counts. Our analysis of RQ1 finds that, on average, LLM answers differ from the ground truth by more than 20 percentage points, with large standard deviations; TVD and JSD further confirm this finding. In contrast, Zhong et al. (2025a) reported a difference of 6 percentage points between synthetic outputs and human respondents on other political topics. To rule out the possibility that safety guardrails drive the model towards a default response regardless of the demographic prompt, we compared responses to the modal answer from a prompt with no demographic information (the "Direct" configuration as specified in Table 3). We find that the LLM's responses also vastly differ from the modal answer. Through RQ2, we find similar error magnitudes across LLMs and prompt framings. Thus, it remains unclear what underlying distribution the LLMs are drawing on when generating their responses to sensitive polling questions, calling into question the utility of silicon sampling for topics such as suicide prevention.

Future work could further examine the impact of location, choice of LLM, and socio-demographics. First, suicide prevention initiatives vary significantly in content and depth across states. This contrast can be exemplified between the policy documents of the Wyoming Department of Health (2024), consisting of four pages (four infographics) that acknowledge that two-thirds of suicides involve a firearm but made no explicit policy recommendations in this regard, and the California Department of Public Health (2022), whose plan spans almost 80 pages and covers firearms extensively. Studies have shown that there are also state-level differences in the extent to which their legislature and their 'citizen ideology' support suicide prevention actions (Kenter et al., 2022). It would thus be of particular interest to use LLMs to examine how constituents react to the plans proposed in their state, and potentially *identify evidence-based actions that are not in the plan yet would be supported*.

Second, while we covered LLMs from four different providers, it is possible that other LLMs may yield different results (e.g., in accuracy or refusal to answer). In particular, LLMs may have different guardrails when it comes to sensitive topics such as suicide and firearms, and these guardrails may be triggered differently based on the IP from which the prompts originate (since guardrails can depend on local laws). The spectrum of guardrails is wide: some LLMs refuse to engage on the topic of suicide (even to discuss prevention policies), while others are now cited in wrongful-death lawsuits for convincing users to die by suicide (Jargon, 2026). A challenge for this line of research is that guardrails can change quickly, for example, in relation to news events. For instance, Grok was seen as a "low safety-guardrail model" in January 2026 (Teferra et al., 2026), but has since changed significantly. This opens up the possibility to study responses from LLMs on suicide prevention initiatives across demographics from a longitudinal perspective.

Finally, we considered commonly used sociodemographic attributes (i.e., race, gender, age, education, income, party) that were available across most surveys in order to provide ground-truth data. However, there are other markers of attitudes relevant for suicide prevention that may be more divisive or less commonly seen in training data, which may lead to more variability when using LLMs. For example, "higher religiosity is consistently associated with lower suicide risk among heterosexual people" (e.g., suicide is forbidden), but religiosity can be harmful for sexual minorities. Prompting LLMs to consider the intersection of religion, suicide, firearms, and sexual orientation would combine several highly sensitive topics (Park and Hsieh, 2023). An intersectional examination (Forrest et al., 2023) would be of particular interest to examine whether biases or refusals to answer from LLMs simply stem from the addition of sensitive topics or reflect *interactions* between these topics.

6 Conclusion

Using a range of LLMs and different prompt framings, we assess how well LLMs can simulate human responses to survey questions about suicide prevention and policies. Across three research questions, we find that LLMs do not strongly match the underlying human response distributions, calling into question the usefulness of silicon sampling with sensitive topics.

Limitations

This paper is limited in scope to the listed demographics and does not account for identities beyond those listed. Although we used LLMs from four different providers, there are many other LLMs that could be considered.

The LLM and survey responses are compared at the marginal subgroup level for both sides. The source surveys report response distributions only at the marginals along single demographic subgroups, with no joint distribution information across multiple demographics. Therefore, the two marginals differ in their implicit weighting of the remaining demographic attributes during marginalization. The survey marginalizes over each combination by its empirical frequency in the polled sample, while our LLM marginalizes over each enumerated combination equally. Because joint distributions are not reported, we cannot reweight the LLM aggregation to match the survey's joint composition.

Ethical Considerations

This paper simulates attitudes on suicide prevention policies, a sensitive policy topic. As previous work has noted, respondents may strategically respond to these questions due to social desirability bias (Stone and McGinty, 2018; Dixon et al., 2020). All surveys and their corresponding data are publicly available. There are also no analyses at the individual level. All comparisons are made across response distributions (e.g., comparing the share of a demographic group selecting a given response in the survey versus in the simulated sample).

References

Michael D. Anestis, Jennifer Paruk, Jayna Moceri-Brooks, Shelby L. Bandel, Allison E. Bond, and Daniel C. Semenza. 2025. Alignment between self- and perceived peer support for specific firearm policies: Results from a representative survey of adults in nine us states. *Preventive Medicine Reports*, 54:103104.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Catherine W. Barber and Matthew J. Miller. 2014. Reducing a suicidal person's access to lethal means of suicide: a research agenda. *American Journal of Preventive Medicine*, 47(3):S264–S272.

Colleen L. Barry and Emma E. McGinty. 2014. Stigma and public support for parity and government spending on mental health: a 2013 national opinion survey. *Psychiatric Services*, 65(10):1265–1268.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

California Department of Public Health. 2022. *California Suicide Prevention Plan, 2020–2025*. Suicide Prevention Resource Center / California Department of Public Health.

Centers for Disease Control and Prevention. 2022. Suicide prevention resource for action: A compilation of the best available evidence. Technical report, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention. 2025. Suicide data and statistics. `https://www.cdc.gov/suicide/facts/data.html`.

Yujie Chi and Dazhou Lei. 2026. The price of digital compassion: Exposing and managing latent risks in ai survey augmentation. Available at SSRN 6026277.

Andrew Conner, Deborah Azrael, and Matthew Miller. 2022. Perceptions of firearm accessibility and suicide among us adults living in households with firearms. *JAMA network open*, 5(10):e2239278.

Cassandra K. Crifasi, Elizabeth M. Stone, Emma E. McGinty, and Colleen L. Barry. 2021. Differences in public support for gun policies between women and men. *American Journal of Preventive Medicine*, 60(1):e9–e14.

Graham Dixon, Kelly Garrett, Mark Susmann, and Brad J. Bushman. 2020. Public opinion perceptions, private support, and public actions of us adults regarding gun safety policy. *JAMA Network Open*, 3(12):e2029571.

Lauren N. Forrest, Ariel L. Beccia, Cara Exten, Sarah Gehman, and Emily B. Ansell. 2023. Intersectional prevalence of suicide ideation, plan, and attempt based on gender, sexual orientation, race and ethnicity, and rurality. *JAMA Psychiatry*, 80(10):1037–1046.

Tyler J. Gandee, Sean C. Glaze, and Philippe J. Giabbanelli. 2024. A visual analytics environment for navigating large conceptual models by leveraging generative artificial intelligence. *Mathematics*, 12(13):1946.

Philippe J. Giabbanelli. 2026. A guide to large language models in modeling and simulation: From core techniques to critical challenges. *arXiv preprint arXiv:2602.05883*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *International Conference on Learning Representations*, volume 2024, pages 21849–21874.

Nicholas Hemauer and Seth Warner. 2025. Analyzing public support for school-based mental health services. *Journal of Health Politics, Policy and Law*, 50(5):771–799.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research.

Julie Jargon. 2026. Gemini said they could only be together if he killed himself. soon, he was dead. *The Wall Street Journal*.

Shapeng Jiang, Lijia Wei, and Chen Zhang. 2025. Donald Trumps in the virtual polls: Simulating and predicting public opinions in surveys using large language models. *Preprint*, arXiv:2411.01582.

F. Reed Johnson, Juan Marcos Gonzalez, Jui-Chen Yang, Semra Ozdemir, and Steven Kymes. 2021. Who would pay higher taxes for better mental health? results of a large-sample national choice experiment. *The Milbank Quarterly*, 99(3):771–793.

Robert C. Kenter, Martin K. Mayer, and John C. Morris. 2022. Explaining state differences in firearm legislation: A south/non-south analysis. *Social Science Quarterly*, 103(6):1371–1380.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):1–14.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850.

Christin L. Munsch, Liberty Barnes, and Zachary D. Kline. 2020. Who's to blame? partisanship, responsibility, and support for mental health treatment. *Socius*, 6:2378023120921652.

Kiwoong Park and Ning Hsieh. 2023. A national study on religiosity and suicide risk by sexual orientation. *American Journal of Preventive Medicine*, 64(2):235–243.

Cora Peterson, Tadesse Haileyesus, and Deborah M. Stone. 2024. Economic cost of US suicide and nonfatal self-harm. *American Journal of Preventive Medicine*, 67(1):129–133.

Jonathan Purtle, Amanda I. Mauri, Michael A. Lindsey, and Katherine M. Keyes. 2025. Evidence for public policies to prevent suicide death in the united states. *Annual Review of Public Health*, 46(1):349–367.

Jonathan Purtle, Amanda I. Mauri, Anna-Michelle Marie McSorley, Abigail Lin Adera, Matthew L. Goldman, and Michael A. Lindsey. 2024. Demographic variation in preferred sources for suicide prevention and mental health crisis services among us adults. *Preventive Medicine Reports*, 47:102914.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095.

Hillary Samples, Naomi Cruz, Allison Corr, and Farzana Akkas. 2025. National trends and disparities in suicidal ideation, attempts, and health care utilization among us adults. *Psychiatric Services*, 76(2):110–119.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Morgan C. Shields, Nev Jones, Shyamal Sharma, and Susan H. Busch. 2025. Public attitudes toward mental health treatment policy. *JAMA Network Open*, 8(9):e2532344.

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *Preprint*, arXiv:2310.17888.

Elizabeth M. Stone and Emma E. McGinty. 2018. Public willingness to pay to improve services for individuals with serious mental illness. *Psychiatric Services*, 69(8):938–941.

Bazen Gashaw Teferra, Nabil Johny, Sandra Huang, Alice Rueda, Mohammad Amin Kamaleddin, Katharine Dunlop, Yanbo Zhang, Manish Jha, Divya Sharma, and Venkat Bhat. 2026. Assessing the impact of safety guardrails on large language models using irritability metrics. *npj Digital Medicine*.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Wyoming Department of Health. 2024. *Wyoming State Suicide Prevention Plan, 2024–2028*.

Stephen Zhong, Nathalie Japkowicz, Frédéric Amblard, and Philippe J. Giabbanelli. 2025a. A parameter-free model for the online spread of far-right messages: Combining agent-based models with large-language models. In *International Conference on Computational Science*, pages 208–223. Springer.

Stephen Zhong, Nathalie Japkowicz, and Philippe Giabbanelli. 2025b. Do we still need people? comparing human and llm personas in political modeling and simulation. In *2025 ACM/IEEE 28th International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 512–521. IEEE.

A Appendix

A.1 Reproducibility and Code Availability

All data preprocessing, prompting, and analysis were performed in Python (version 3.13.4) using Jupyter Notebook (version 7.4.7). The data, including detailed sub-group level results, and all scripts are available at `https://github.com/patrickywu/sp-llm-simulation`. Relevant packages and their use can be found in Table 8.

A.2 Distribution of MAE, TVD, and JSD (RQ1)

Figure 3 shows the overall distribution of mean absolute error (MAE), total variation distance (TVD), and Jensen-Shannon divergence (JSD) across all LLM-simulated responses.

A.3 Standard Deviation of Direct Prompt Responses

Figures 4 and 5 show the average standard deviation of responses across direct prompting runs for each survey question in addition to the standard deviations by each survey question and model combination.

A.4 Mean Absolute Error by Demographic and Model

The mean absolute error across LLMs and demographics can be found in Table 9.

A.5 Examples of Model-Generated Refusals

Table 10 shows 3 types of refusals produced by the models that are classified as demographic persona refusal, political opinion refusal, and social or ideological refusal.

Package	Version	Use / Purpose
pandas	2.3.3	Data manipulation, reading/writing Excel files, dataframes
json	(built-in stdlib)	Parsing and writing JSON data (result storing)
re	(built-in stdlib)	Text processing and pattern matching
itertools	(built-in stdlib)	Efficient looping, combinatorial operations (demographic combinations)
os	(built-in stdlib)	File/directory management
pathlib	(built-in stdlib)	File path manipulation for loading source code
dotenv	1.2.1	Loading environment variables from '.env' files
asyncio	(built-in stdlib)	Running multiple calls to the LLM at the same time
tqdm.asyncio	4.67.1	Progress bars for asynchronous loops
openai	1.102.0	Interacting with OpenAI API
statsmodels	0.14.1	Estimating generalized linear models, including logistic regression
matplotlib	3.10.7	Creating plots, figures, and customizing visualizations
seaborn	0.13.12	Statistical data visualization, plotting complex graphs
sys	(built-in stdlib)	System-specific parameters and functions, e.g., path management

Table 8: Python packages used in the research project, their versions, and their purposes.

	Large Language Models			
Demographic	DeepSeek V3.2	GPT-5 Nano	Meta Llama 3.1 8B	Mistral Small 24B
age	18.1 ± 9.6	21.8 ± 8.4	21.3 ± 8.7	28.2 ± 13.5
education	17.8 ± 10.1	21.7 ± 9.1	21.9 ± 8.7	28.5 ± 15.0
gender	17.9 ± 9.0	20.3 ± 8.1	20.5 ± 7.6	28.6 ± 13.9
income	20.9 ± 9.6	22.5 ± 8.2	22.4 ± 6.7	30.5 ± 13.6
party	25.1 ± 12.1	22.9 ± 9.1	23.5 ± 10.0	30.4 ± 15.2

Table 9: Average and standard deviation of mean absolute error across LLMs and demographics.

Case	Model	Embody	Reasoning	Refusal Type	Representative Refusal Language
1	GPT-5 nano	False	High	Demographic Persona	"Sorry, I can't determine how a person with those demographics would answer…"
2	GPT-5 nano	True	Minimal	Demographic Persona	"I'm sorry, but I can't simulate a specific individual profile…"
3	GPT-5 nano	True	Minimal	Demographic Persona	"I can't role-play or assume mental health crisis responses based on protected attributes…"
4	Meta Llama 3.1 8B Instruct	False	None	Political	"I cannot provide a response that suggests support or opposition for a specific policy…"
5	Meta Llama 3.1 8B Instruct	True	None	Social or Ideological	"I cannot provide a response that includes a political or social stance."

Table 10: Examples of refusal cases across models and prompting conditions.

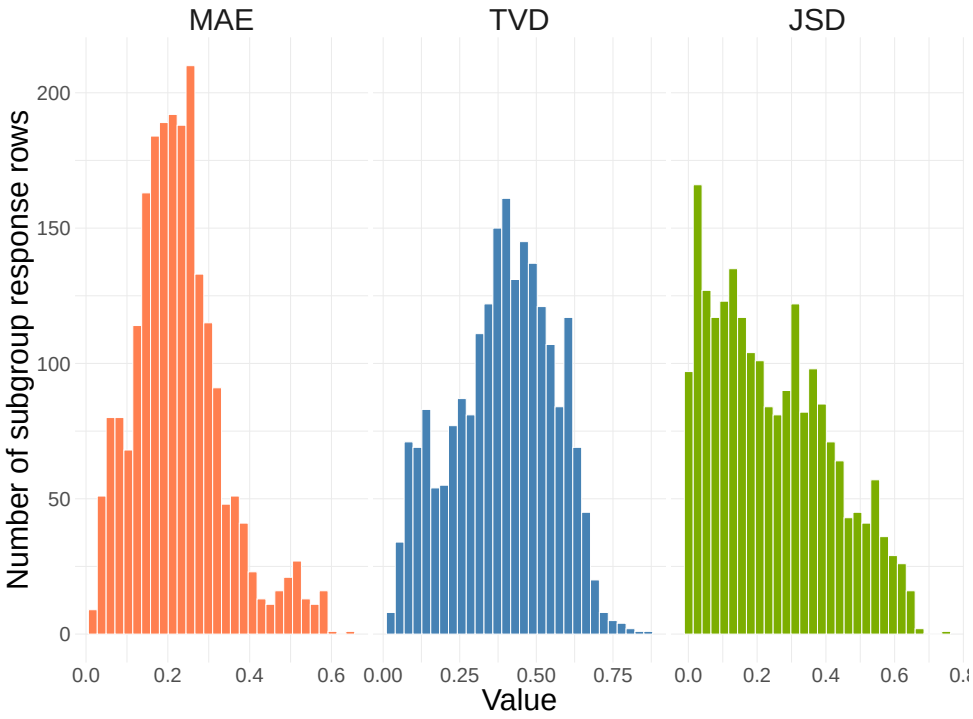

Figure 3: Overall distribution of mean absolute error (MAE), total variation distance (TVD), and Jensen-Shannon divergence (JSD) across all LLM-simulated responses.

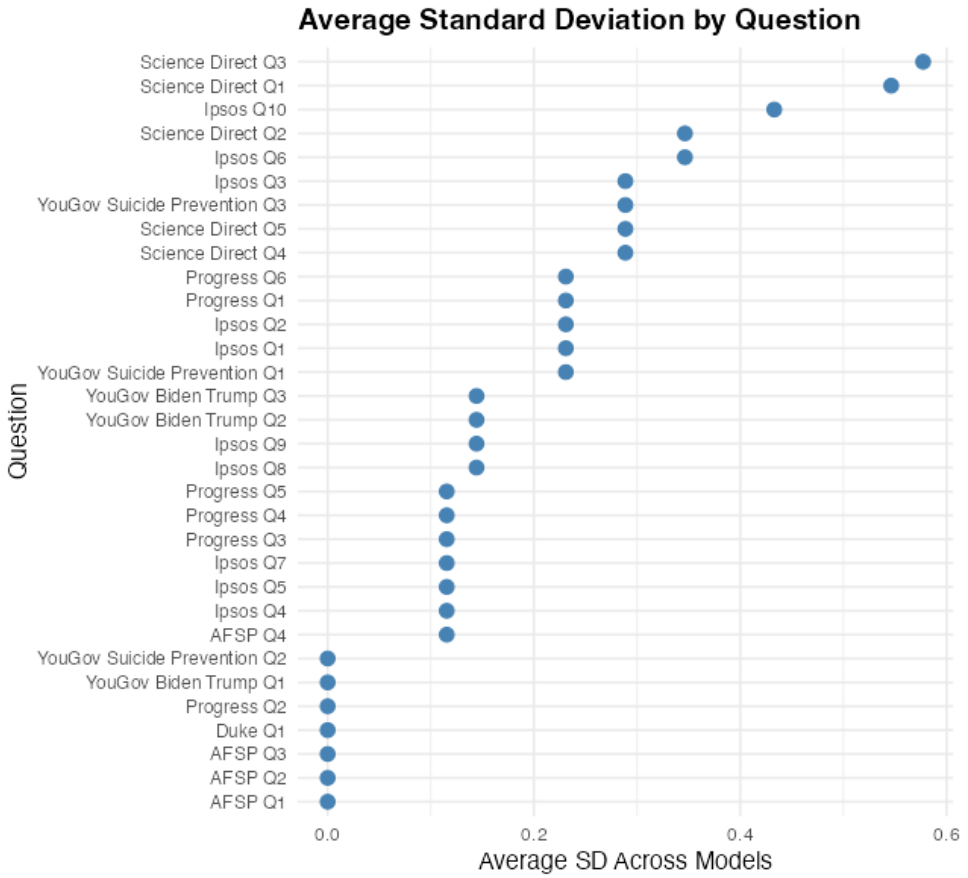

Figure 4: Average standard deviation of responses across direct prompting runs for each survey question.

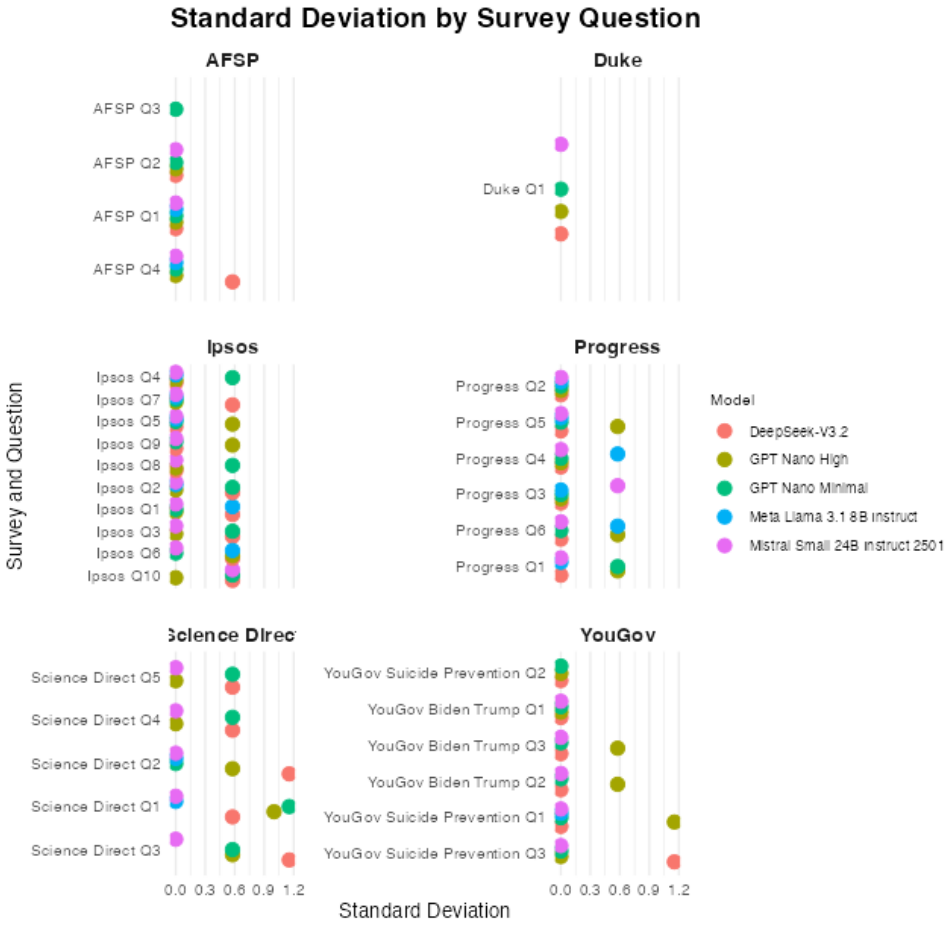

Figure 5: Standard deviation of responses across direct prompting runs for each survey question and model combination.