MSc. Thesis

Synthetic Participants in Behavioural Scientific Research - investigating the impact of model design and prompt parsimony on algorithmic fidelity across a US and Egyptian sample

Supervisor: Dr. Maximilian Heitmayer

The London School of Economics and Political Science

Department of Psychology and Behavioural Science

MSc. Psychology of Economic Life

Acknowledgements

Finishing this thesis concludes a transformative year at LSE – one that exceeded my expectations on every level. The lectures, discussions, long nights on campus (not necessarily in the library though) and the ecosystem of LSE as a whole will continue to shape my trajectory in many ways.

Accordingly, I want to thank all the people who enabled this year. Starting with Dr. Frédéric Basso, for the lectures and discussions that profoundly reshaped my understanding of behavioural science, embedding this already thrilling field in a wider systemic context. Furthermore, Dr. Maximilian Heitmayer for his continuous academic support and mentorship in writing this thesis and beyond.

With the data collection being part of a broader project at the El-Erian Institute of Behavioural Economics and Policy, I also want to thank Josh Ramminger, Dr. Malte Dewies, Dr. Micha Kaiser and Prof. Lucia Reisch for the trust placed in me – collaborating on this has been an incredibly instructive and joyful experience.

Lastly, I wish to thank my friends with whom I shared the highs and lows of this year, it's been a real pleasure. And of course, all of this wouldn't have been possible without the continuous and unconditional support from my family for whom I am grateful every day.

Abstract

Amid rising interest in AI-driven methods, the use of synthetic participants (SPs) generated by LLMs is gaining momentum, raising questions about whether these models can meaningfully replicate human attitudes in behavioural research and policy settings. This study evaluates algorithmic fidelity of SPs using a 2×3×3 cross-sectional comparative design combining survey data from 600 real participants in the US and Egypt with 18 synthetic datasets created under varying LLMs (GPT4o-mini, Mistral Large 2, Claude Sonnet 4) and prompt parsimony levels (high, medium, low). Fidelity was assessed across person-, item-, and pattern-levels. While SPs accurately reflected overall pattern level approval trends, item-level fidelity was mixed, and person-level alignment remained limited. Model choice had a notable impact on performance, whereas prompt parsimony showed little effect. Fidelity was considerably lower in the Egyptian sample, suggesting limitations in how well current models generalise across cultural contexts. These findings position SPs as valuable tools for capturing aggregate behavioural patterns but also highlight their current limitations - especially at the individual and minority level - emphasising the need for critical, context-sensitive evaluation before broader adoption in behavioural science and policy.

Table of contents

List of Figures and Tables

List of abbreviations

AI – Artificial Intelligence

API – Application Programming Interface

BPP – Behavioural Public Policy

CI – Confidence Interval

EGY – Egypt

LLM – Large Language Model

MAE – Mean Absolute Error

OR – Odds Ratio

RQ – Research Question

SE – Standard Error

SP – Synthetic Participant

US – United States

WEIRD – Western, Educated, Industrialized, Rich, Democratic

1. Introduction

Recent advances in Artificial Intelligence (AI), particularly in large language models (LLMs), are beginning to reshape the methodological toolkit of behavioural science and its applied subfield, Behavioural Public Policy (BPP) (Lee et al., 2023). LLMs, i.e neural networks trained on vast text datasets to probabilistically predict and generate human-like language, have opened new possibilities for changing behaviour, forecasting behavioural responses, and personalising interventions. For example, Meyer & Elsweiler (2025) investigate whether LLM chatbots can change user behaviour, whilst Sadeghian & and Otarkhani (2024) explore their use in tailoring digital behaviour change strategies. Another central area of LLM application within Behavioural Science and BPP are survey-based methodologies.

Surveys remain a cornerstone of behavioural science, underpinning empirical research and shaping real-world applications in policy, market research and governance. In psychology alone, over half of published studies rely on questionnaire-based methods (Scholtz, 2020) whilst 89% of market research professionals report using online surveys as their primary quantitative method (Murphy, 2021). In the policy domain, surveys are essential to capture preferences, anticipate responses to interventions, and inform programme design (Watson et al., 2008). Yet, traditional survey methods are constrained by practical limitations. Data collection is costly and time-intensive, particularly when representative samples are required (Hardigan et al., 2016). At the same time, declining response rates and demographic skews limit both the completeness and generalisability of findings (Roberts et al., 2020).

Against this backdrop, recent developments in LLM-driven simulations offer a potential complement to conventional data collection: Synthetic Participants (SPs). These are LLM-based simulations of human participants whose demographic and attitudinal characteristics are defined by the researcher (Shrestha et al., 2024). SPs are designed to mimic human responses to survey questions with the aim of describing, explaining, or predicting behaviour (Sarstedt et al., 2024). While simulation is not new to the social sciences (Bonabeau, 2002), the capabilities of modern LLMs introduce a level of nuance, speed, and adaptability previously unattainable (Argyle et al., 2023). Used appropriately, SPs could offer a faster and more affordable alternative to collecting real-world data, particularly for exploratory work or for hard-to-reach populations (Breugel & Schaar, 2023).

One area in which SPs may be particularly promising is Behavioural Public Policy (BPP), which uses behavioural science to design policies that influence citizen decision-making and

behaviour (Oliver, 2013). Gauging public approval of behavioural policy proposals is not merely descriptive: higher perceived legitimacy and approval often translate into greater acceptance and behavioural compliance hence boosting policy effectiveness (de Ridder et al., 2022; van Gestel et al., 2021). If SPs can reliably approximate public opinion, they could serve as a tool for rapid policy prototyping. However, the method is still in its infancy, and key uncertainties remain such as how factors like model architecture, prompting techniques or cultural variation influence alignment. As Amirova et al. (2024) suggest, there is a need for an evaluation-first paradigm in synthetic data research: rigorous empirical testing must precede deployment.

This thesis responds to that call by investigating SP–human alignment in a cross-national context through a survey on BPP item approval. Using varied models and prompting techniques, it compares alignment across model designs, prompt parsimony levels, and cultural settings, offering empirical insight into a rapidly evolving field with substantial promise yet in need of systematic evaluation.

2. Literature Review

2.1 Synthetic Participants and algorithmic fidelity

2.1.1 Definitions

A central concept in the fast-growing debate around SPs is the notion of algorithmic fidelity, a term introduced by Argyle et al. (2023). At the core of this concept lies the assumption that the outputs generated by LLMs "are selected not from a single overarching probability distribution, but from a combination of many distributions" in a way that allows the model "to produce outputs that correlate with the attitudes, opinions, experiences of distinct human subpopulations" (Argyle et al., 2023, p.4). Building on this, algorithmic fidelity describes the "degree to which the complex patterns of relationships between ideas, attitudes, and socio-cultural contexts within a model accurately mirror those within a range of human subpopulations." (Argyle et al., 2023, p.4). As of now, there is no standardised measurement for algorithmic fidelity with authors drawing on a wide range of metrics, with SP fidelity being typically assessed through correlational measures, distance metrics, and distributional comparisons to human response patterns. (Aher et al., 2023; Hwang et al., 2023; Lee et al., 2023) Noticeably though, scholars have rather focused on population level measures (such as aggregate correlations) and pattern measures (such as distributions), with only very few studies

as of now assessing alignment on an individual person level – even though also only in a limited sense e.g. only qualitatively (Amirova et al., 2024) or with an inconsistent simulation approach (Hwang et al., 2023). It thus remains unclear whether SPs can merely predict population level patterns or also individual, person level responses.

2.1.2 Areas of application and previous findings

Within the field of Behavioural Science, the literature on SPs is growing steadily with a diverse range of applications and use cases. This subsection outlines the three dominant areas in which SPs are being used: behavioural games, consumer research and public opinion, and subsequently summarises the key findings from previous studies in this realm.

Following the tradition of computer science, the quality of LLM output was initially mostly assessed by whether it would pass the Turing Test, i.e. produce content indistinguishable from human produced content (Moor, 2001). However, since a majority of LLMs are capable of passing the Turing Test (Borg, 2025; Jones & Bergen, 2025) criticism has emerged claiming that "the Turing Test primarily assesses deceptive mimicry rather than genuine intelligence, prompting the continuous emergence of alternative benchmarks" (Rahimov et al., 2025, p.1). Examples of such extensions of the Turing Test are Behavioural Experiments, also called Turing experiments (Aher et al., 2023), that are based on methods known from experimental economics and social psychology. They differ from a classical one-off, individual-level Turing Tests by "simulating a representative sample of participants in human subject research" (Aher et al., 2023). Such Turing Experiments demonstrate that LLMs reliably align with human behaviour in ultimatum games (used to test fairness and altruism), Milgram experiments (obedience to authority) and garden path sentences (real time language processing). These results were replicated and extended by Xie et al. (2024) and Mei et al. (2024) who also found human-LLM alignment for several trust-based games, the Bomb Risk Game (risk tolerance), Public Goods Game (cooperation, altruism) and Prisoner's Dilemma (reciprocity, strategic cooperation). These findings demonstrate that SPs mostly behave indistinguishable from human participants in such behavioural games.

Another important field of application for SPs are questions related to market research and consumer insights. As cost- and time-intensity of conventional survey methodologies are a strain for marketers and product-developers , SPs are being employed to gain consumer insights and product feedback (Sarstedt et al., 2024). This extends beyond the academic realm, with companies such as *Synthetic Users* or *Artificial Societies* already offering commercial tools for simulating human participants. Hämäläinen et al. (2023) demonstrate that feedback to a video

game given by SPs aligns well enough with human feedback to be incorporated into the product development process. In a similar manner, Brand et al., 2023 show "that estimates of willingness-to-pay for products and features derived from GPT responses are realistic and comparable to estimates from human studies" (p.1).

Lastly, SPs are also being experimented with in the policy realm. Lee et al. (2023) find evidence for GPT-4 being able to predict voting behaviours and policy positions. Shrestha et al., 2024 demonstrate that the answers of LLM generated SPs strongly correlate with their human counterparts on diverse policy issues such as sustainability or financial literacy and even effects of experimental interventions. However, they find that the aggregate correlations of answers for non-WEIRD samples are less strong. Whilst the majority of papers in the SP realm utilise so-called 1:1 matching, i.e. matching the demographic characteristics of a real person one to one with a corresponding SP (e.g. Lee et al., 2023; Shrestha et al., 2024), other sampling techniques are also possible. Sun et al. (2024) for instance merely use population-level instructions or "random-silicon-sampling" (p.1) to generate a distribution of synthetic personas. They find this random-silicon-sample to be well aligned with existing public opinion data from national surveys and thus conclude it is "feasible to survey the opinions of a sub-group using only their group-level demographic information" (p.8) However, Sun et al. (2024) also emphasise the limitations of such methods, highlighting that alignment of different subgroups significantly differs and showing that SPs tend to be biased in their answers towards politically moderate opinions. Whilst some of the first results involving SPs are undoubtedly impressive, at least equally as many scholars express extensive criticism about the method for various reasons.

2.1.3 Challenges in synthetic data research

The challenges and critical points related to the use of SPs span a diverse range of topics and can be broadly categorised in the following categories: Quality of results, training data and bias and foundational concerns.

Regarding quality of results, many synthetic datasets do not exhibit comparable variance to human datasets but rather overrepresent the mean value of a population without appropriately representing the tails, contributing to synthetic datasets often lacking appropriate distributional variance (Hwang et al., 2023). Besides, with LLMs being based on probabilistic prediction, the same input does not always lead to the same output – undermining replicability of findings and hence damaging validity of SP results (Rossi et al., 2024).

Another area of controversy is the training data used in creating the models as well as the inherent biases they tend to exhibit. With LLMs being trained on data from the internet, stereotypes expressed in e.g. articles, forums or books can get picked up by the models and thus exhibited when prompted to simulate behaviour of stereotyped subgroups (Dillion et al., 2023). This has direct implication on the ability of LLM based SPs to simulate subgroups which the underlying model has little or only heavily biased training data on – the results tend to get more stereotypical – undermining algorithmic fidelity for those hard-to-reach populations (Santurkar et al. 2023).

Lastly, some scholars express foundational concerns regarding the method, emphasising the lack of regulation in this area and pointing at a potential lack of genuine insights and epistemic circularity when trying to gain new insights from non-empirical data (Demszky et al., 2023). Building on this, others warn that over-reliance on synthetic participants risks undermining core human-centered values in research and technology design, potentially displacing the very subjects such methods are intended to understand (Agnew et al., 2024).

For a detailed overview of critical voices in the SP literature, see Appendix A.

2.2 LLM design and differences

SPs are generated through LLMs, i.e. models that are trained on massive text datasets to predict the most likely next word in a sequence, enabling them to generate coherent responses. These predictions are based on probabilistic patterns learned during the training of the model (Brown et al., 2020). Since the AI boom following OpenAI's release of ChatGPT in 2022, many LLMs have emerged (Naveed et al., 2024).

Differences in model architecture, training data, and internal instructions lead to substantial variation in how LLMs generate outputs, shaping their ability to produce replicable outcomes, follow instructions, respond cross-culturally, and vary in reasoning style, assertiveness and social sensitivity (Agrawal et al., 2022; Brown et al., 2020). However, with all of these factors being confounded and many companies strictly limiting the amount of information available regarding e.g. training data, processes and instructions, it is impossible to isolate and properly test for a single of those factors driving LLM cognition in a mutually exclusive way (Polonioli, 2025). Hence, making generalisable claims about the impact of singular factors on model performance remains methodologically unavailable. Therefore, research in the field of SPs has focused on differentiating the models and their impact on SP generation as a whole instead of investigating more specific causal relationships.

First studies identify significant differences across language models. Santurkar et al. (2023) found GPT-4 aligned more closely with US public opinion than GPT-3.5-turbo or text-davinci-003, though all exhibited systematic biases, notably underrepresenting older demographics. Model choice thus influenced how well SPs reflected public attitudes, with newer models reducing misalignment. Similarly, Aher et al. (2023) reported GPT-4 achieved the highest fidelity in replicating behavioural experiments, whereas Claude 2 and Bard were less consistent, and GPT-3.5 more erratic. Yet GPT-4 and Claude 2 showed a "hyper-accuracy distortion" (p. 3), producing SPs that were overly rational or moral. These findings underscore model choice as a critical determinant of SP realism and alignment. However, given the rapid evolution of LLMs, generalisable claims remain premature.

2.3 The role of prompt design

The prompt, i.e. "the textual input provided by users to guide the model's output" (Amatriain, 2024, p. 1) also has a significant influence on the quality and validity of LLM generated data and thus SPs (Zhang et al., 2025). LLMs do not access external truths per se, instead, they generate output based on the statistical likelihood of language patterns continuing from the input. As such, the prompt serves as a framing device that influences how the model interprets and responds to a task (Brown et al., 2020). Ambiguous or poorly structured prompts often produce vague or off-topic responses, which undermines experimental control and replicability in research contexts (Barrie et al., 2025).

A key mechanism that explains this sensitivity is in-context learning: the model infers task structure, expectations, and behavioural patterns from examples or instructions in the prompt (Brown et al., 2020). Recent work confirms that LLMs use such cues to simulate reasoning processes, similar to how humans learn through demonstration or analogy. (Mao et al., 2025). However, this process is constrained by the model's limited working-memory simulation, meaning that when prompts are overloaded with complex or conflicting information, response quality tends to decline (Gong et al., 2024).

In this sense, prompt design and especially prompt richness or parsimony act as cognitive scaffolding: They guide the model's responses and support more structured, relevant, and coherent outputs. For behavioural research prompt design and appropriate degrees of parsimony are among the biggest levers of influencing LLM output and thus SP quality (Kaiser et al., 2024). Regarding the direction of the effect of prompt parsimony, Shrestha et al. (2024) and Kaiser et al. (2024) report that simpler prompts can match or even outperform extended ones, emphasising the potential benefits of parsimony in achieving alignment. In contrast, Ma (2025)

finds that adding contextual details to demographic prompts can improve agreement metrics but may also introduce noise and diminish accuracy in some cases. Taken together, these mixed results highlight that the effects of prompt parsimony on alignment remain unsettled and likely depend on task context and evaluation criteria.

2.4 Cross cultural and minority sampling

Among the main concerns in the debate around SPs is their ability to simulate minority populations. LLM generated answers tend to "skew male, White, American, liberal, and wealthy in perspective", a phenomenon that has also been called the 'Silicon Valley bias' (Sorensen et al., 2024). Lee et al. (2023) and Santurkar et al. (2023) show that although LLMs can approximate the political beliefs of certain groups, they consistently struggle to accurately represent those of minority groups. Shreshta et al. (2024) report similar findings when testing the alignment of SP policy position simulations of a US sample in comparison to two Middle Eastern samples. Explanatory approaches to this issue vary, with some authors emphasizing the underrepresentation of minority groups and their opinions in training data (e.g. Santurkar et al., 2023) whilst others point towards more structural issues. For example, Peterson (2025) shows that the training of LLMs often eliminates the so-called 'long tails' from the model. These long tails encompass observations that deviate from the norm and, in social terms, frequently correspond to minority groups and their perspectives. Thus, the way in which LLMs overrepresent the average may constitute a structural bias against minority groups (Sorensen et al., 2024).

With this issue being at the core of the SP debate, testing and quantifying potential differences between e.g. WEIRD and non-WEIRD populations remains a foundational concern for research in the SP realm.

2.5 Use Case: Synthetic Participants in policymaking

With policymakers relying on survey methodologies for collecting data, the challenges related to surveys, i.e. time, cost and representativeness, are highly relevant in the policy realm (Watson et al., 2008). Accordingly, and as outlined above, applications of SPs are also being experimented with in the policy-making process (e.g. Lee et al., 2023; Shrestha et al., 2024). From the perspective of policymakers, employing SPs creates at least three distinct and relevant use cases:

(i) Gauging public opinion: Using SPs to estimate general public sentiment about topics of interest, policy priorities or citizen concerns (e.g Lee et al., 2023).

(ii) Rapid policy prototyping and scenario testing: Policymakers can use SPs to simulate how populations might respond to proposed policies prior to implementation. For instance, varying unemployment benefit structures could be tested on SPs representing different socioeconomic backgrounds, family compositions, and employment histories. This enables rapid, low-cost policy iteration (e.g. Shrestha et al., 2024).

(iii) Policy development in data-poor environments: SPs may enable inclusive policy design in data-poor environments by simulating responses from e.g. underrepresented groups such as rural communities (van Kesteren, 2024). However, revisiting subsection 2.1.3, the applicability of these hinges on the debated ability of SPs to represent marginalised populations.

In addition, empirical studies in the field of BPP show that effectiveness of behavioural interventions depends strongly on citizen approval, as interventions like nudges work mainly when aligned with existing citizen preferences (de Ridder et al., 2022; Thamer et al., 2024; van Gestel et al., 2021). Efforts to change behavior against prevailing attitudes rarely succeed, making accurate assessment of public approval especially crucial for effective policymaking in BPP.

2.6 Summary and research gap

This section has outlined the importance and central limitations of survey data in behavioural scientific research and BPP. SPs and their broad fields of application as well as potential limitations were introduced as an approach to solving some of the issues. The extent to which this can be successful is measured through the concept of algorithmic fidelity, i.e. the degree to which LLMs truly align with their human counterpart. The importance of LLM selection, prompt design and cultural context were introduced as main levers for influencing SP performance before making the case for using SPs in policymaking.

Building on this, the present study investigates algorithmic fidelity in a cross-national context, using behavioural policy survey data as a benchmark for comparison. The analysis contrasts SP outputs generated under different models and prompt designs, enabling systematic examination of how these factors interact with cultural context to influence fidelity. Assessment spans

multiple dimensions - item-level, pattern-level, and, most importantly, person-level - providing a nuanced account of SP performance and its potential role in policy-related applications.

This research addresses a gap in the SP-related literature on several levels. First, it takes seriously the observation that, despite promising initial studies, "researchers have largely neglected the potential of using synthetic participants for policy-related research" (Shrestha et al., 2024, p. 2). Second, it uses an Egyptian sample as a benchmark, contributing valuable data to the ongoing debate on the applicability of SPs across both WEIRD and non-WEIRD cultural contexts. Third, it is among the very few SP studies that prominently feature person-level fidelity metrics, i.e., measures of how closely an SP simulates its exact human counterpart. Fourth, it is the first SP study to use a BPP survey, applying SPs to a policy context in which approval is known to strongly influence policy uptake and effectiveness. Taken together, these contributions respond to growing calls for more granular and contextually grounded evaluations of the capacity of SPs to accurately simulate human opinion.

2.7 Research questions

Through this approach I am aiming to answer the following research questions:

RQ1: *How well can SPs approximate human survey responses across different fidelity dimensions?*

RQ2: *How does LLM model design impact the alignment of synthetic and real answers?*

RQ3: *How do SP simulations differ across cross-cultural samples?*

RQ4*:* *How does prompt parsimony impact the alignment of synthetic and real answers?*

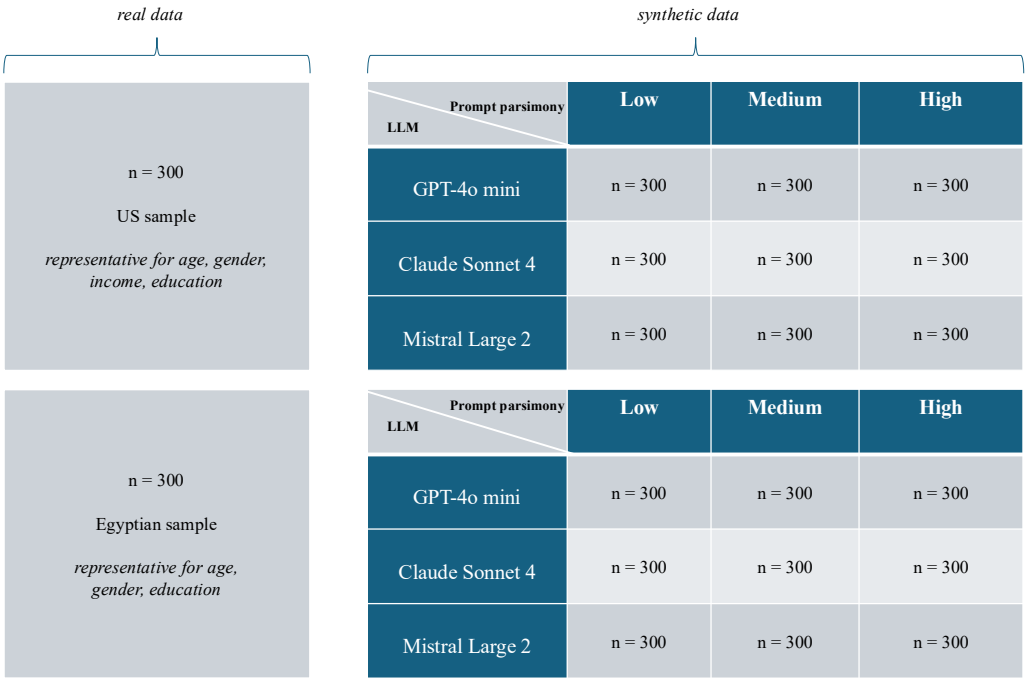
As the literature on SPs is still in its infancy, this study takes on an exploratory character. The novelty of the LLMs employed, which have yet to undergo systematic academic evaluation, coupled with the methodological uncertainty reflected in mixed findings on the relationship between prompt parsimony and alignment, underscore the need for an exploratory approach. Moreover, given the rapid pace of AI development and the continually evolving nature of SP methodologies, an exploratory stance is essential to remain responsive to this fast-moving research landscape.

3. Methods

3.1 Study design

This study employs a 2x3x3 comparative cross-sectional design combining real survey data collection with a simulation component. Human participants in two countries completed a one-time online survey assessing various proposed behavioural policies. Subsequently, synthetic survey responses were generated based on one-to-one matching the real demographic respondent profiles using three different LLMs and three degrees of prompt parsimony, i.e. varying amounts of input variables the models were fed with. Therefore, across both countries, the dataset with the real participants answers could be compared to 9 synthetic datasets, each created under a different *model* x *prompt parsimony* condition.

Figure 1: *Study design*



3.2 Survey

The survey in use is an adapted version of the survey from Sunstein et al. (2019). In their original paper, the authors used a global survey of $n = 5.532$ participants with 54 items to assess the approval of 15 BPP instruments in the domain of health, sustainability, wellbeing and prosocial behaviour.

The survey was adapted by removing six items as they collected personal data with no predictive relevance to this study. Age was adjusted to standardised 10-year brackets to increase

anonymity. Items subject to change since the original 2018 data collection were updated (e.g., monetary amounts to current income distributions). For Egypt, education was adapted to 'non-upper-secondary education' and 'upper secondary education' based on available quotas. Questions on party vote in Egypt were removed for participant protection, and all sensitive items (politics, income, etc.) were optional in both countries. The survey was conducted in English and Arabic respectively, the English version of the survey can be found in Appendix L.

The Sunstein et al. (2019) survey was selected as the data collection instrument because it captures explicit approval ratings for a diverse range of BPP interventions. This makes it well-suited for evaluating the potential of SPs to approximate human responses in a policy setting, recalling the practical use-cases for SPs and potential effects on policy effectiveness outlined in Section 2.5. Besides, the public interest in the results of the survey, with the paper being cited 211 times as of July 2025 and referenced in multiple policy briefs (e.g WHO, 2022), underline the relevance of replicating this survey, especially in new national contexts such as Egypt.

At the core of the survey are 15 policy approval questions regarding a diverse range of behavioural policies, e.g. defaults for green energy or organ donation, public education campaigns against obesity or nutritional information nudges. All items were assessed on a binary scale of 'approve' versus 'disapprove'. Besides the 15 original policy items from the 2019 survey, five new BPP items were added to be able to control if LLM-predictions were trained on the published results from the original paper.

3.3 Participants and sampling

Ethics approval by LSE was granted on 15.04.2025 and data collection took place between 24.07.2025 and 29.07.2025. Data collection was conducted using the online panel service *Qualtrics* that provides renumeration to participants completing the survey. All participants were informed of the scope of the study and presented with the data handling and privacy policy they had to consent to proceed. The consent form can be found in Appendix I.

The inclusion criteria were that participants resided in the US or Egypt, passed all attention filters and had a completion time of > 4 minutes, derived from the softlaunch median completion time of 8 minutes. Besides, quota-sampling was employed to ensure the sample was representative regarding age, gender, education and income. For the Egyptian sample income quotas had to be removed due to missing reliable income data and due to the practical limitations of sampling low-income participants from Egypt. As data collection was handled by

Qualtrics, only complete responses including passing of attention filters, minimum completion time etc. were provided and no ex-post exclusion was necessary.

A sample of 300 participants per country was chosen in line with prior studies comparing human and synthetic responses (e.g., Argyle et al., 2023; Shrestha et al., 2024), which typically use 200–400 respondents. Given the study's exploratory aims and use of bootstrapped confidence intervals, this sample was considered sufficient for capturing demographic diversity and enabling robust fidelity benchmarking. Accordingly, no formal power analysis was conducted, which aligns with conventions in recent SP research (e.g. Kaiser et al., 2024; Ma, 2025)

3.4 LLM selection

Three LLMs were used to create the synthetic datasets: GPT-4o-mini (OpenAI), Mistral Large 2 (Mistral) and Claude Sonnet 4 (Anthropic). These models were selected because they represent leading approaches in language modelling, with all three featuring prominently in contemporary benchmarking efforts (Chiang et al., 2024). As the current literature on SPs has primarily focused on OpenAI's models, this study expands the field by testing models from different institutions. With Mistral being an open-access model, the study follows the call by Wulff et al. (2024) for integrating open-access models into scientific benchmarking to enhance transparency, traceability, and reproducibility in research. The practical relevance of such comparisons is further underscored by the fact that Claude models are already in use in the UK public sector under a government-level collaboration with Anthropic (UK Government, 2025).

3.5 Synthetic Data generation and prompt design

The synthetic data was generated using multi-turn API calls to the model's proprietary APIs via Python 3.11.9 in a Google-Colab environment. The python code was created using the respective LLM interface of the model that was called which is documented in Appendix M.

Extensive pretesting with 64 datasets was employed to identify optimal model settings for temperature, token limits, formatting instructions, and prompting techniques. The temperature parameter in LLMs, typically ranging from 0.0 to 2.0, controls output variability: low values (e.g., 0.2) produce deterministic responses, while high values (e.g., 1.2) yield more diverse outputs. When using SPs, temperature is key for behavioural realism; however, there is no consensus on an optimal setting, as this is highly context dependent (Kaiser et al., 2024; Shrestha et al., 2024; Xie et al., 2024). Pretesting indicated medium temperatures (0.5–0.8) provided better alignment than higher ones (0.8–1.5). Accordingly, temperature was set to 0.7 for all models, consistent with Ma (2025). The context window was set to 2500 tokens. Token

limits define how much text an LLM can process in a single interaction, thus controlling simulation complexity (Brown et al., 2020). Given survey length, prompt complexity, and binary output format, this limit was sufficient to process all data.

During pretesting, multiple prompting techniques were tested to achieve LLM outputs correctly formatted and resembling basic patterns of real human answers, i.e. variance across the sample and coherence among similar respondents. Appendix B summarises the approaches and insights generated before arriving at the final prompt shown in Figure 2.

API calls followed a multi-turn process: first, a system message containing instructions, the survey, and output specifications was sent as context; second, a user message with demographic profiles was provided to simulate responses. This follows common practice in SP simulations, enabling coherence across a synthetic sample not achieved with single-shot API calls (Ma, 2025). LLMs accessed via APIs are intrinsically stateless, with each request processed independently and no memory retained unless explicitly included. This prevents inter-session contamination or unintended learning, ensuring SP responses across conditions remain uncontaminated and methodologically reliable (Abhyankar et al., 2024).

The complete prompt for all parsimony levels and the Python code for generating SPs can be found in Appendices C and D. While the system message (survey with 20 binary policy approval items and formatting instructions) remained constant, input variables varied by parsimony level. The following figures show the prompt structure and variables included at each level.

Figure 2: *Prompt template (High degree of parsimony)*

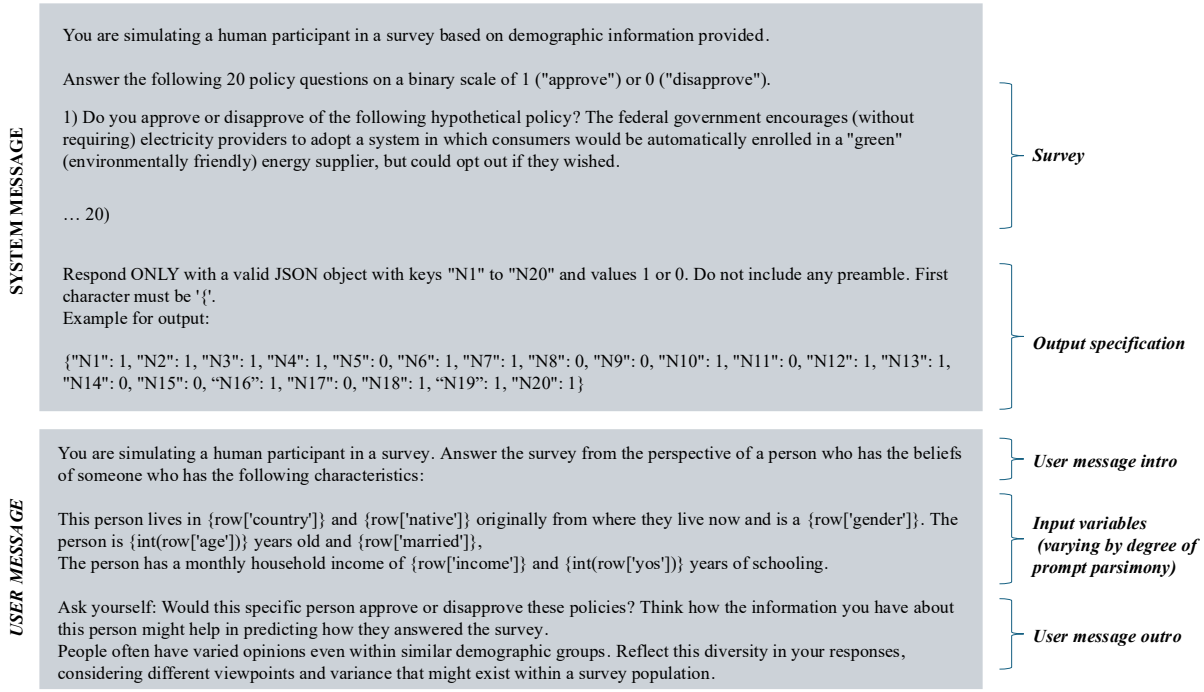
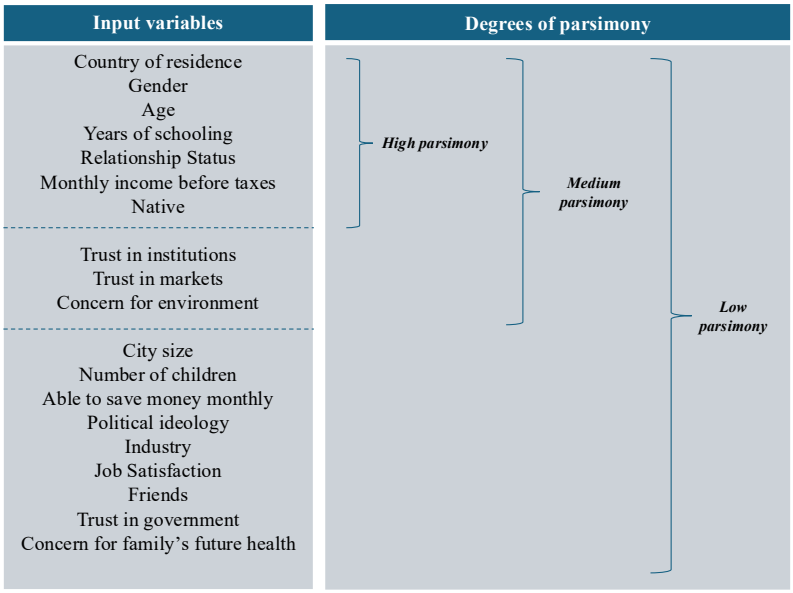


Figure 3: *Input variables by degree of parsimony*



Input variables for the high parsimony prompt were basic demographic variables whilst the selection of input variables for the medium and low parsimony prompt followed an analysis of the correlation heatmap from Sunstein et al. (2019). The idea behind this was choosing covariates that are empirically shown to predict approval to the survey items in order to test if LLMs can simulate these relations.

3.6 Alignment assessment methodology

Algorithmic fidelity was wholistically examined along complementary dimensions: (I) person-level fidelity – agreement of individual response profiles; (II) item-level fidelity – alignment of approval rates and odds across items; (III) pattern fidelity – preservation of item rank order. Robustness measures such as the use of confidence intervals, p-values, and resampling-based methods were used throughout to ensure inferential validity. This framework combines descriptive indicators (e.g., match rate, approval differences) with inferential statistics (e.g., odds ratios, correlations with bootstrapped CIs) tailored to the binary-response, item-based structure of the data. Whilst correlation measures (Santurkar et al., 2023) distance estimates (Kaiser et al., 2024) and Cohens Kappa (Ma., 2025) have been widely used in SP studies, drawing on binary logistic regression was novel. Given the binary, low-variance data structure it is suitable and enables robustness checks within a regression framework, while being intuitively interpretable (El Emam et al., 2024).

(I) Person-level fidelity

Person-level agreement was assessed using match rate (percentage of identical responses) and Cohen's Kappa calculated per participant. Kappa corrects for chance agreement and is widely applied to binary response alignment (Cohen, 1960; Gwet, 2014). Participants with constant response patterns were excluded since Kappa is undefined without variance. Such exclusion is recognized as an accepted practice in inter-rater reliability research (Gwet, 2014) and is explicitly documented in Appendix I. A bootstrap percentile 95% confidence interval (1,000 resamples) was estimated for the mean Kappa to account for its non-normal sampling distribution (Efron & Tibshirani, 1994). This approach captures both raw agreement (match rate) and variance-adjusted alignment (Kappa), appropriate for response vectors spanning multiple dichotomous items. Cohen's kappa coefficients and CIs were combined across studies using Fisher's z-transformation to normalize their sampling distributions before averaging, as this approach addresses the bounded nature of agreement statistics and satisfies normality assumptions required for meta-analytic procedures (van Aert, 2023).

(II) Item-level fidelity

Item-level fidelity was examined using approval rates and Mean Absolute Error (MAE), alongside logistic regression models. A global odds ratio (OR) was estimated via logistic regression with standard errors (SE) clustered by item and separate per-item ORs were derived from item-specific models with 95% confidence intervals and p-values. Logistic regression is

appropriate for binary outcomes and, with clustered SEs, accounts for the nested participant - item structure present here (Conklin, 2002). There were cases where SP showed no variance in their responses (e.g., all approvals) resulting in complete separation in the logistic regression. On average, this occurred in around four items per dataset. These items were retained because the core aim of this study is to evaluate the fidelity and data quality of synthetic responses, which necessarily includes capturing cases of extreme misalignment. Excluding such items would risk artificially inflating alignment estimates and underrepresenting critical points of divergence. In including them, this study follows the synthetic data handling guidelines of Alaa et al., (2022). To be able to combine Odds ratios (OR) from different study conditions, they were log-transformed, averaged on the log scale, and then exponentiated to obtain a combined OR (Chang & Hoaglin, 2017).

(III) Pattern fidelity

To assess whether synthetic data preserved the relative ranking of item approvals, Spearman's rank correlation (ρ) was calculated between real and synthetic approval rates. Spearman's ρ is a nonparametric correlation robust to skew and ordinal data (Spearman, 1913), it requires rankable paired data. These conditions were met, as approval rates were continuous and computed for the same items. Uncertainty was quantified via a bootstrapped 95% CI (10.000 resamples), and a permutation test (10.000 iterations) provided an empirical p-value, both suited to the modest number of items (Efron & Tibshirani, 1994). Spearman correlation coefficients were combined across study conditions using Fisher's z-transformation to normalize their sampling distributions before averaging (Welz et al., 2022).

Within this analytical framework, the resulting metrics can be compared along the conditions of the 2x3x3 study design to identify effects of model design, degree of prompt parsimony or cultural context.

4 Findings

Fidelity metrics (as outlined in 3.6) were calculated across all 18 datasets and divided into descriptive metrics (i.e. MAE, match rate – cf. Table 1) and non-pooled inferential metrics (*κ, ρ, ORs* – cf. Table 2). Means for inferential measures were calculated using Fisher's transformations (*κ, ρ)* and log-transformations (*OR*) and plotted across models, parsimony levels and countries (cf. figure 4). All pooled inferential metrics referred to in this section can also be found in Appendix G.

4.1 Descriptive statistics

The study included 600 participants (2 x n = 300). Quota sampling was employed to be representative for gender, age, income and education in the US and gender, age and education in Egypt. Detailed descriptive statistics per sample including data distributions can be found in Appendix E.

The following table shows some descriptive metrics of alignment between synthetic and real participants across models and degrees of prompt parsimony. Match rates indicate the proportion of real answers that were correctly simulated by SP while the Mean absolute Errors (MAE) can be used to interpret the size of bias in the synthetic answers.

Discussion of these metrics and results follows in section 4.2.

Table 1: *Results table I (descriptive metrics)*

USA	**Low prompt Parsimony**		**Medium prompt parsimony**		**High prompt parsimony**		**Model means**	
	Match rate	*MAE*	*Match rate*	*MAE*	*Match rate*	*MAE*	*Match rate*	*MAE*
GPT-4o mini	0.63	0.20	0.63	0.20	0.63	0.27	**0.63**	**0.22**
Claude Sonnet 4	0.65	0.19	0.61	0.26	0.65	0.25	**0.64**	**0.23**
Mistral Large 2	0.65	0.19	0.65	0.24	0.62	0.28	**0.64**	**0.24**
Parsimony level means	**0.64**	**0.19**	**0.63**	**0.23**	**0.63**	**0.27**	/	/

EGY	**Low prompt Parsimony**		**Medium prompt parsimony**		**High prompt parsimony**		**Model means**	
	Match rate	*MAE*	*Match rate*	*MAE*	*Match rate*	*MAE*	***Match rate***	***MAE***
GPT-4o mini	0.64	0.22	0.63	0.22	0.64	0.28	**0.64**	**0.24**
Claude Sonnet 4	0.62	0.27	0.50	0.41	0.61	0.32	**0.58**	**0.33**
Mistral Large 2	0.66	0.24	0.65	0.26	0.67	0.25	**0.66**	**0.25**
Parsimony level means	**0.64**	**0.25**	**0.59**	**0.30**	**0.64**	**0.29**	/	/

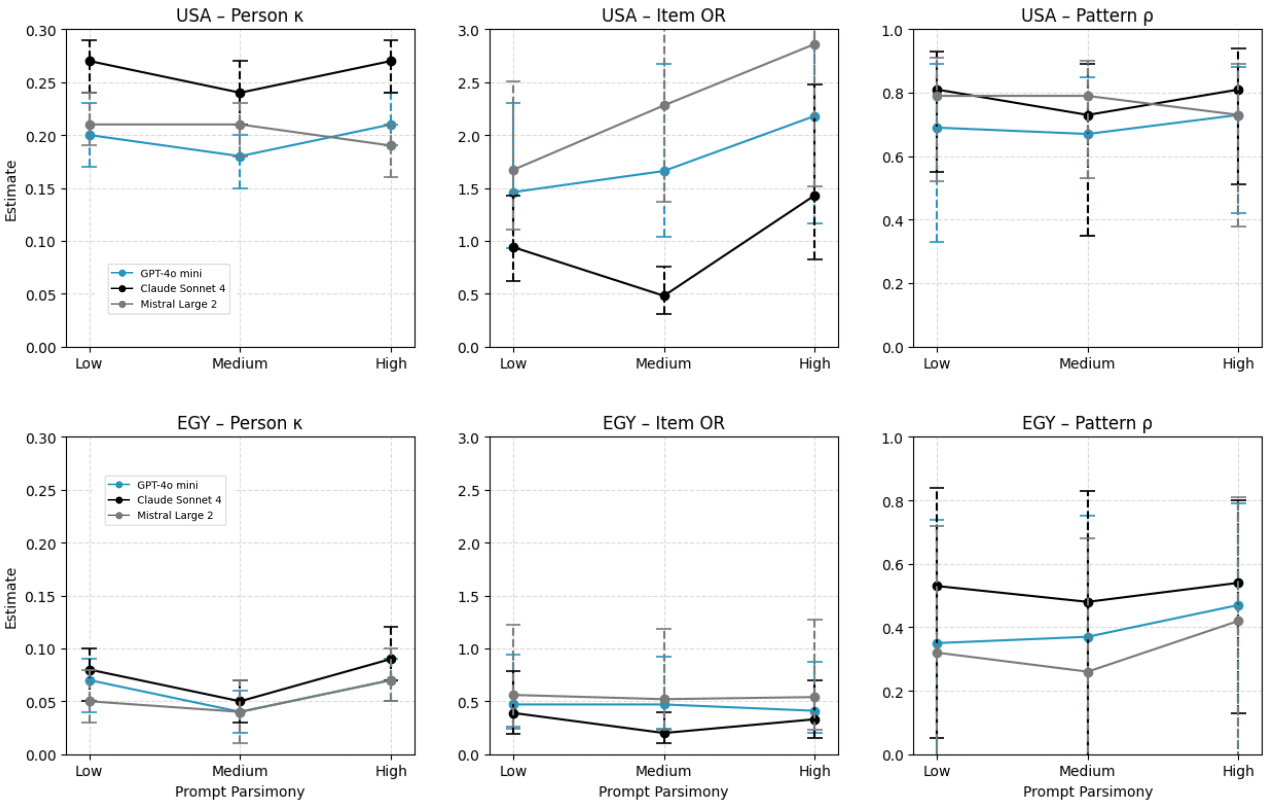
4.2 Inferential statistics

Table 2: *Results table II (non-pooled inferential metrics)*

USA	**Low prompt Parsimony**			**Medium prompt parsimony**			**High prompt parsimony**			**Model means across prompt parsimony levels**		
	Person-level κ	*Item level global OR*	*Pattern rank ρ*	*Person level κ*	*Item level global OR*	*Pattern rank ρ*	*Person level κ*	*Item level global OR*	*Pattern rank ρ*	*Person level κ*	*Item level global OR*	*Pattern rank ρ*
GPT-4o mini	0.20 [0.17-0.23]	1.46 [0.93-2.30]	0.69 [0.33-0.89]	0.18 [0.15-0.20]	1.66 [1.04-2.67]	0.67 [0.35-0.85]	0.21 [0.19-0.24]	2.18 [1.16-4.10]	0.73 [0.42-0.88]	**0.20 [0.17-0.22]**	**1.74 [1.04-2.93]**	**0.70 [0.37-0.87]**
Claude Sonnet 4	0.27 [0.24-0.29]	0.94 [0.62-1.43]	0.81 [0.55-0.93]	0.24 [0.21-0.27]	0.48 [0.31-0.76]	0.73 [0.35-0.89]	0.27 [0.24-0.29]	1.43 [0.82-2.48]	0.81 [0.51-0.94]	**0.26 [0.23-0.28]**	**0.86 [0.54-1.39]**	**0.79 [0.47-0.92]**
Mistral Large 2	0.21 [0.19-0.24]	1.67 [1.11-2.51]	0.79 [0.52-0.91]	0.21 [0.18-0.23]	2.28 [1.37-3.79]	0.79 [0.53-0.90]	0.19 [0.16-0.21]	2.86 [1.51-5.41]	0.73 [0.38-0.89]	**0.20 [0.18-0.23]**	**2.22 [1.32-3.72]**	**0.77 [0.48-0.90]**
Parsimony-level means across models	**0.23 [0.20-0.25]**	**1.32 [0.86-2.02]**	**0.77 [0.47-0.91]**	**0.21 [0.18-0.23]**	**1.22 [0.76-1.97]**	**0.73 [0.41-0.88]**	**0.22 [0.20-0.25]**	**2.22 [1.32-3.72]**	**0.76 [0.44-0.91]**	/	/	/

EGY	**Low prompt Parsimony**			**Medium prompt parsimony**			**High prompt parsimony**			**Model means across prompt parsimony levels**		
	Person level κ	*Item level global OR*	*Pattern rank ρ*	*Person level κ*	*Item level global OR*	*Pattern rank ρ*	*Person level κ*	*Item level global OR*	*Pattern rank ρ*	*Person level κ*	*Item level global OR*	*Pattern rank ρ*
GPT-4o mini	0.07 [0.04-0.09]	0.47 [0.24-0.94]	0.35 [-0.16-0.74]	0.04 [0.02-0.06]	0.47 [0.24-0.92]	0.37 [-0.15-0.76]	0.07 [0.05-0.09]	0.41 [0.20-0.87]	0.47 [-0.02-0.79]	**0.06 [0.03-0.08]**	**0.45 [0.23-0.91]**	**0.40 [-0.11-0.76]**
Claude Sonnet 4	0.08 [0.05-0.10]	0.39 [0.19-0.79]	0.53 [0.05-0.84]	0.05 [0.03-0.07]	0.20 [0.10-0.40]	0.48 [-0.01-0.8]	0.09 [0.07-0.12]	0.33 [0.15-0.70]	0.54 [0.13-0.80]	**0.07 [0.05-0.09]**	**0.30 [0.14-0.60]**	**0.52 [0.06-0.81]**
Mistral Large 2	0.05 [0.03-0.08]	0.56 [0.26-1.22]	0.32 [-0.19-0.72]	0.04 [0.01-0.07]	0.52 [0.23-1.18]	0.26 [-0.27-0.68]	0.07 [0.05-0.10]	0.54 [0.23-1.27]	0.42 [-0.09-0.81]	**0.06 [0.03-0.08]**	**0.54 [0.24-1.22]**	**0.34 [-0.18-0.74]**
Parsimony-level means across models	**0.07 [0.04-0.09]**	**0.47 [0.23-0.97]**	**0.40 [-0.10-0.77]**	**0.04 [0.02-0.07]**	**0.37 [0.18-0.76]**	**0.37 [-0.14-0.75]**	**0.08 [0.053-0.102]**	**0.42 [0.19-0.92]**	**0.48 [0.01-0.80]**	/	/	/

Figure 4: *Plot of mean inferential metrics*



4.2.1 RQ1: How well can SPs approximate human survey responses across different fidelity dimensions?

Assessment of the overall fidelity of the SP answers was done through segmenting the question of alignment into the dimensions of fidelity that the metrics were calculated in: Person level-, Item-level, Pattern-level-fidelity and robustness tests across fidelity dimensions. For RQ1, only pooled metrics across all 18 conditions are considered, more fine-grained analysis for country, model and prompting effects follows with RQ2 – RQ4.

Person-level-fidelity deals with assessing alignment on the individual, participant level, i.e. how well a synthetic twin mimics their corresponding human counterpart's answers. Match rates (proportion of identical individual level responses) averaged across all 18 conditions yielded a result of $\mu = 62.8\%$. In order to control for chance, Cohen's Kappa (κ) was calculated and yielded a mean value of $\kappa = 0.14$, (95% CI: [0.11 – 0.17]). Contextualising these mean values reveals only a modest agreement of real and synthetic values on individual level. While the match rates indicate that, on average, synthetic participants replicated nearly two thirds of human responses identically, a κ of 0.14 shows that much of this agreement can be attributed to chance. With Kappa being a standardized measure, $\kappa = 0.14$ classifies as 'slight agreement'

(Landis & Koch, 1977). Therefore, individual level SP fidelity on average seems low whilst still showing statistical significance for 'slight agreement'.

Item-level fidelity assesses how well SPs reproduce aggregate approval rates for each individual policy item, regardless of whether the same SP-human pairs align. It shifts the focus from individual replication to population-level response distributions across policy items. While MAE captures the average absolute deviation in approval rates, the OR quantifies the relative likelihood of SP endorsement versus humans (OR = 1 = perfect alignment), with deviations indicating systematic under- or over-endorsement supported by CIs. The global mean odds ratio of $OR = 0.81$, 95% CI [0.44, 1.47] suggests that synthetic participants exhibited broadly similar overall endorsement tendencies to humans, with only slight under-endorsement on average. However, the average of $MAE = 25.5$ indicates that this aggregate alignment masks notable item-level variability, with synthetic participants alternating between over- and under-endorsement across individual items. This ranges from mean MAE of 0.06 for Item 7 (Information campaign against childhood obesity) up to MAE of 0.55 for item 11 (tax donation defaults for charity), revealing strong variance in item level predictive accuracy.

The Spearman rank correlation (ρ) assesses the extent to which SPs replicate the relative ordering of policy approval observed in human data, independent of differences in absolute approval levels. This metric is important because it captures pattern-level fidelity, indicating whether SPs can accurately reproduce the overall pattern of which policies are more or less favoured by humans. Across all 18 conditions, the observed correlation of $\rho = 0.61$ [95% CI: 0.19–0.85] suggests a moderate-to-strong preservation of rank-order preferences, albeit with some variability. The relatively wide confidence interval implies that while SPs generally capture the overall structure of human approval hierarchies, their consistency in replicating finer-grained rankings is less stable. Across all 18 individual datasets, permutation tests for Spearman's rank correlation yielded $p\_perm < .001$, indicating that the observed rank-order alignments are highly unlikely to have arisen by chance and collectively support the robustness of the overall pattern-level fidelity observed.

Importantly, analysis of the five control items N16-N20 (i.e. items that were not included in the original 2019 survey) revealed no significant fidelity difference compared to the items whose results were already published by Sunstein et al. (2019). Conversely, SP predictions for the control items were even more accurate with an 8-percentage lower MAE and 4-percentage points higher match rate on average, compared to the original items. This suggests that SP alignment is not merely an artifact of the LLM models being trained on the Sunstein et al.

(2019) paper but is also applicable to item approval scores not possibly contained in the training data set.

4.2.2 RQ2: How does LLM model design impact the alignment of synthetic and real answers?

When pooling results across countries and parsimony levels, model-level differences in alignment emerge clearly across all fidelity metrics. For person-level alignment (κ), Claude Sonnet 4 achieved the highest pooled performance (κ = 0.17, 95% CI [0.14–0.19]), outperforming both GPT-4o mini (κ = 0.13, 95% CI [0.10–0.15]) and Mistral Large 2 (κ = 0.13, 95% CI [0.11–0.16]). While these values remain within the 'slight agreement' range (Landis & Koch, 1977), Claude's relative improvement (≈30% higher than GPT-4o and Mistral) suggests marginally better individual-level fidelity.

For item-level fidelity (OR), Mistral Large 2 demonstrated the strongest alignment on average (OR = 1.09, 95% CI [0.56–2.13]), approximating parity with human approval rates (OR = 1). GPT-4o mini was slightly lower (OR = 0.88, 95% CI [0.49–1.63]), while Claude Sonnet 4 showed systematic under-endorsement (OR = 0.51, 95% CI [0.27–0.91]). Notably, the wide confidence intervals for OR across all models, spanning below and above 1.0, indicate considerable variability and preclude definitive inferential claims about systematic over- or under-endorsement.

For pattern-level fidelity (ρ), Claude Sonnet 4 again achieved the highest correlation with human approval rankings (ρ = 0.68, 95% CI [0.33–0.86]), followed by Mistral Large 2 (ρ = 0.60, 95% CI [0.14–0.84]) and GPT-4o mini (ρ = 0.57, 95% CI [0.14–0.82]). These results suggest that Claude best replicated the rank-order of policy approval, while Mistral and GPT-4o mini were slightly less precise but still showed moderate-to-strong correlations. CIs for all models were arguably wide, indicating statistical uncertainty for these estimates.

4.2.3 RQ3: How do SP simulations differ across a cross-cultural samples?

Assessment of SP fidelity across US and Egyptian samples revealed pronounced cross-cultural differences, both in magnitude and direction of alignment when pooling metrics across models and parsimony levels. Whilst descriptive match rates indicated similarities, inferential analyses revealed significant differences, suggesting that the observed similarities were likely driven by chance arising from the binary structure of the response variable

At the individual level, Cohen's κ decreased from 0.22 [0.20–0.25] in the US to 0.06 [0.04–0.09] in Egypt, corresponding to a 65% reduction in agreement strength. Match rates followed

a similar pattern (US ≈ 72%, Egypt ≈ 58%), indicating that SPs were markedly less successful at reproducing Egyptian participants' individual response profiles.

At the item level, cross-cultural differences were even more pronounced. In the US, SPs systematically overpredicted approval, reflected in a global OR of 1.53 [0.86–2.02], indicating a 53% higher likelihood of endorsement than observed in human respondents. By contrast, Egyptian SPs underpredicted approval, with an OR of 0.42 [0.23–0.97], corresponding to a 58% lower likelihood of endorsement. Complementary MAE analyses reinforce this point: item-level deviations in Egypt were, on average, more than 10 percentage points larger than in the US, amplifying the pattern of weaker and directionally inverted item-level fidelity.

Pattern-level fidelity (ρ) similarly differed between contexts. US SPs displayed strong alignment in replicating rank-order approval structures (ρ = 0.75 [0.47–0.91]), while Egyptian SPs exhibited only moderate correlation (ρ = 0.42 [0.10–0.77]), indicating diminished accuracy in reconstructing culturally specific policy preference hierarchies. Wide CIs for pooled item and pattern level metrics indicate limited precision for these estimates.

4.2.4 RQ4: How does prompt parsimony impact the alignment of synthetic and real answers?

Prompt parsimony exerted no clear or directional impact on alignment. Across all models and countries, person-level fidelity remained largely unaffected by parsimony (κ ranging from 0.12 to 0.15). Similarly, pattern-level fidelity (Spearman's ρ) showed minimal variation (ρ = 0.61–0.65), indicating that rank-order approval patterns were well preserved regardless of prompt detail.

The most notable effect emerged at the item level. While overall odds ratios similar (OR = 0.85–1.34), high parsimony produced a stronger overestimation bias in US data (OR = 2.22 [1.32–3.72]), whereas Egyptian data exhibited persistent underestimation across all levels (e.g., OR = 0.42 [0.19–0.92]). These directional biases persisted despite increasing prompt complexity, and in Egypt were accompanied by wide confidence intervals, underscoring instability rather than systematic improvement.

5 Discussion

This section synthesizes and contextualises the findings from the previous section, adding interpretive depth and linking it back to the discussed literature. Moreover, limitations are discussed before concluding with concrete policy implications.

5.1 Synthesis of findings

The findings regarding general alignment (RQ1) reveal a clear dissociation between fidelity dimensions: while SPs exhibited only slight individual-level alignment, they more reliably captured aggregate approval tendencies and preference hierarchies. This dissociation and especially the bad alignment on individual level are highly important and novel empirical findings in the SP field. This divergence between individual- and pattern level can be explained through a central tendency bias, wherein LLMs are well-suited to approximate broad population-level opinion structures but struggle with the idiosyncratic variability inherent in individual responses (Lee et al., 2023). Because LLMs are trained on population-scale language data, their internal representations privilege high-frequency, generalisable patterns (Santurkar et al., 2023), resulting in SPs that converge toward modal or typical respondent profiles rather than replicating fine-grained heterogeneity.

This smoothing effect helps explain why SPs could reproduce relative policy rankings despite weaker individual-level mimicry. As Dillion et al. (2023) note, recovering depended preference orderings primarily depends on models' ability to capture underlying evaluative dimensions embedded in textual data, rather than replicating the psychological or situational drivers of individual behaviour. By contrast, predicting specific human responses demands sensitivity to contextualised experiences and latent traits that are not fully represented in training data (Amirova et al., 2024). However, current publications do not sufficiently investigate and report this divergence by focusing on good alignment on pattern level while neglecting evaluation of individual level alignment. As findings of bad individual-level alignment have strong implications on SP use-cases, a wholistic evaluation and reporting of fidelity dimensions is central for preserving academic rigour in SP research.

Besides, investigating the effect of model choice (RQ2) shows clear model-level differences: Claude performed best on individual- and pattern-level fidelity, Mistral led on item-level alignment, and GPT performed worst overall. This contrasts with earlier work, where GPT-4 typically topped benchmarks (Santurkar et al., 2023; Aher et al., 2023), underscoring that model rankings are neither fixed nor universally generalisable across tasks and datasets.

Interpreting these differences is difficult due to the 'black box' nature of LLMs. Architecture, training data, and alignment are intertwined and opaque, making it impossible to isolate causes. Claude's higher fidelity may reflect alignment or data coverage, while Mistral's item-level

strength could stem from architecture or tokenisation, though such explanations remain fully speculative. In practice, benchmarking must remain an ongoing, empirical exercise, as the rapid pace of LLM development outstrips any attempt to derive stable, mechanistic explanations. The recent release of 'Centaur' (Binz et al., 2025) - a behavioural science fine-tuned variant of LLaMA 3.1 - illustrates how fast the landscape is evolving. While such a fine-tuned model could have provided a valuable benchmark to commercial models, it was released too late to be integrated into the existing study. Taken together, model-level comparisons are best seen as time-bound snapshots, reinforcing the need for repeated testing as SP research needs to keep pace with ongoing advances in LLM design.

Regarding cross-cultural sample differences (RQ3), alignment was substantially higher for the US sample than for Egypt: item-level fidelity was more than three times greater, and individual-level alignment was nearly four times stronger. SPs systematically overpredicted approval among US respondents while underpredicting it for Egyptians. Strikingly, the real survey data showed that the actual Egyptian participants expressed substantially higher approval of these largely progressive policies than their US counterparts - by an average margin of 27% (See Appendix E). Yet SPs reversed this empirical reality, simulating Egyptians as markedly less approving while inflating US approval.

This pattern goes beyond generic moderation effects often attributed to LLMs (Sun et al., 2024). Because the policies largely concerned sustainability, health, wellbeing and prosocial behaviour, the findings suggest that models embed culturally skewed priors, over-associating progressiveness with WEIRD populations while systematically under-assigning it to non-WEIRD groups. This aligns with 'Silicon Valley bias' (Sorensen et al., 2024) whereby Western-centric cultural frames dominate pretraining data and shape model assumptions about global attitudes.

Thus, SPs did not merely exhibit reduced fidelity for Egypt - they inverted empirically observed differences, recasting Egyptians as less progressive despite higher empirically recorded approval rates. As Lee et al. (2024) highlight, such distortions are particularly acute for underrepresented groups, amplifying structural biases in LLM training data. For behavioural science, these findings stress that without culturally adapted fine-tuning or local training data, SPs risk reinforcing distorted hierarchies of global attitudes rather than approximating them.

Lastly, RQ4 examined whether increasing prompt complexity by adding more sociodemographic information impacted SP fidelity. The findings showed little effect: neither

person- nor pattern-level alignment changed meaningfully across parsimony levels, and item-level biases (US overestimation, Egyptian underestimation) persisted regardless of prompt richness. This is consistent with Kaiser et al. (2024), who note that while adding contextual information can yield marginal gains, improvements often plateau and may even be offset by noise introduced through excessive input detail. In this study, increasing parsimony did not appear to address broader factors affecting alignment, such as differences in model architecture (RQ2) or cultural representational gaps (RQ3). These results suggest that prompt enrichment alone may offer limited leverage in improving SP fidelity and that its effects are likely constrained by the underlying statistical representations encoded during model training. While more complex prompting remains intuitively appealing, its impact appears secondary to more structural determinants of alignment, such as training data coverage or fine-tuning approaches.

5.2 Limitations

Several limitations may constrain the interpretation of these findings. SP outputs proved highly sensitive to model settings with even small adjustments to parameters such as temperature and top_p producing shifts in simulated responses. This sensitivity underscores the particular nature of the findings – different model parameters might have yielded very different results.

Furthermore, BPP is a relatively new field (Oliver, 2013) leading to the survey items reflecting specific policies that are unlikely to be well-represented in LLM training data, potentially reducing alignment. Broader, more general attitudinal surveys might have produced closer correspondence due to greater overlap with model training data.

In addition, alignment metrics must be interpreted in light of SP response patterns. SP output was notably homogeneous, with most items receiving rather high approval. This is likely a consequence of the binary response format used in the survey: because SPs exhibit a central tendency bias (Lee et al., 2023) binary options constrain variance and drive outputs toward the dominant class which in training data (especially in the US) will be approval to progressive policy items. Given that human approval was also high, part of the strong pattern-level correlation may reflect this interplay between SPs' averaging tendencies, the binary response structure, and the skewed benchmark distribution. While this does not invalidate the findings, it underscores the need to consider how response format and data characteristics shape observed alignment.

Although almost all effects were significant, wide confidence intervals point to limited precision. This suggests that while the observed relationships are unlikely to be due to chance,

the exact strength of the effects remains uncertain and should be interpreted with caution. Moreover, replicability across repeated SP simulations was not assessed due to API cost constraints, and more specialised models (e.g., Centaur) could not be included owing to hardware demands and late availability. Whilst these factors collectively emphasise the need for careful interpretation, they also hint towards untapped potential in SP research and policy application.

5.3 Policy implications and outlook

Collectively, these findings challenge overly optimistic portrayals of synthetic sampling, situating it instead within an evaluation-first paradigm (Amirova et al., 2024), where SPs function as a complementary tool rather than a wholesale replacement for human data. Crucially though, the higher SP accuracy on unseen control items suggests that alignment is not merely an artefact of memorisation but reflects genuine model generalization, a finding that strengthens their utility when deployed with clear awareness of their limits.

Taken together, this supports a tempered but constructive role for SPs: they are neither the epistemic shortcut implied in some early studies (Argyle et al., 2023) nor devoid of value. Instead, they are best conceptualised as high-bandwidth simulators of population-level regularities, contingent on careful alignment checks and sensitivity to the domain-specific challenges of algorithmic fidelity. Recalling the three SP policy use cases from 2.5 – *(i) Gauging public opinion, (ii) Rapid policy prototyping* and *(iii) Policy development in data-poor environments* – a differentiated picture arises. Whilst high pattern matching does support *(i)* and medium item level fidelity might justify use cases for *(ii)*, bad results for person-level simulations and especially negative bias towards training data minorities renders SP applications for *(iii)* inadequate.

Future research should move beyond proof-of-concept studies and build the empirical and methodological foundations for responsible SP use. A priority is to test whether the divergence between fidelity dimensions holds across different policy domains and how results are affected by model parameters and scales.

Cultural and demographic blind spots must also be addressed. The US - Egypt divergence highlights how SPs remain constrained by training-data biases and underrepresentation of minority perspectives. Broader validation in non-WEIRD contexts and among marginalized groups is critical to ensure SPs inform policy without reinforcing current structural inequities.

Finally, there is an urgent need for a formalised and standardised reporting framework for SP research. Such a framework should include preregistration of model settings and prompts, detailed documentation of benchmark data characteristics, and comprehensive reporting of alignment metrics. Establishing these practices early would provide methodological transparency, reduce inflated or selectively reported results and create clear guidelines for when SP use is appropriate. Without such standards, the field risks ad hoc implementations that undermine both scientific rigor and policy credibility.

To conclude, whilst the technology of Synthetic Participants is undoubtedly exciting and potentially a turning point for social scientific data collection, Sir Arthur Conan Doyle remains right in asserting “healthy skepticism is the basis of all accurate observation” (Doyle, 1919, p.24).

Bibliography

Abhyankar, R., Zijian, H., Srivatsa, V., Zhang, H., & Zhang, Y. (2024). *APIServe: Efficient API Support for Large-Language Model Inferencing*. https://arxiv.org/html/2402.01869v1

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3613904.3642703

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). *Large Language Models are Few-Shot Clinical Information Extractors* (No. arXiv:2205.12689). arXiv. https://doi.org/10.48550/arXiv.2205.12689

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Proceedings of the 40th International Conference on Machine Learning*, 337–371. https://proceedings.mlr.press/v202/aher23a.html

Alaa, A. M., Breugel, B. van, Saveliev, E., & Schaar, M. van der. (2022). *How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models* (No. arXiv:2102.08921). arXiv. https://doi.org/10.48550/arXiv.2102.08921

Amatriain, X. (2024). *Prompt Design and Engineering: Introduction and Advanced Methods* (No. arXiv:2401.14423). arXiv. https://doi.org/10.48550/arXiv.2401.14423

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., & Leibo, J. Z. (2024). Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. *PLOS ONE*, *19*(3), e0300024. https://doi.org/10.1371/journal.pone.0300024

Anthropic. (2024). Claude Sonnet 4 - [Large Language Model]. Retrieved from https://www.anthropic.com (used on 21.05.2024)

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

Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., Rajagopal, G. R., Sloan, A., Tudrej, T., Ture, F., Wu, Z., Xu, L., & Baldwin, B. (2025). *Non-Determinism of 'Deterministic' LLM Settings* (No. arXiv:2408.04667). arXiv. https://doi.org/10.48550/arXiv.2408.04667

Barrie, C., Palaiologou, E., & Törnberg, P. (2025). *Prompt Stability Scoring for Text Annotation with Large Language Models* (No. arXiv:2407.02039). arXiv. https://doi.org/10.48550/arXiv.2407.02039

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2025). *Centaur: A foundation model of human cognition* (No. arXiv:2410.20268). arXiv. https://doi.org/10.48550/arXiv.2410.20268

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl_3), 7280–7287. https://doi.org/10.1073/pnas.082080899

Borg, E. (2025). LLMs, Turing tests and Chinese rooms: The prospects for meaning in large language models. *Inquiry*, *0*(0), 1–31. https://doi.org/10.1080/0020174X.2024.2446241

Brand, J., Israeli, A., & Ngwe, D. (2023). *Using LLMs for Market Research* (SSRN Scholarly Paper No. 4395751). Social Science Research Network. https://doi.org/10.2139/ssrn.4395751

Breugel, B. van, & Schaar, M. van der. (2023). *Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data* (No. arXiv:2304.03722). arXiv. https://doi.org/10.48550/arXiv.2304.03722

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Chang, B.-H., & Hoaglin, D. C. (2017). Meta-Analysis of Odds Ratios: Current Good Practices. *Medical Care*, *55*(4), 328–335. https://doi.org/10.1097/MLR.0000000000000696

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference* (No. arXiv:2403.04132). arXiv. https://doi.org/10.48550/arXiv.2403.04132

Cohen, J. (1960). *A Coefficient of Agreement for Nominal Scales*. https://journals.sagepub.com/doi/10.1177/001316446002000104

Conklin, J. D. (2002). Applied Logistic Regression. *Technometrics*, *44*(1), 81–82. https://doi.org/10.1198/tech.2002.s650

de Ridder, D., Kroese, F., & van Gestel, L. (2022). Nudgeability: Mapping Conditions of Susceptibility to Nudge Influence. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *17*(2), 346–359. https://doi.org/10.1177/1745691621995183

de Ridder, D., Kroese, F., & van Gestel, L. (2022). *Nudgeability: Mapping Conditions of Susceptibility to Nudge Influence*. https://journals.sagepub.com/doi/full/10.1177/1745691621995183

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00241-5

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Doyle, S. A. C. (1919). *The vital message*. Hodder & Stoughton.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

El Emam, K., Mosquera, L., Fang, X., & El-Hussuna, A. (2024). An evaluation of the replicability of analyses using synthetic health data. *Scientific Reports*, *14*(1), 6978. https://doi.org/10.1038/s41598-024-57207-7

Gmyrek, P., Lutz, C., & Newlands, G. (2024). *A Technological Construction of Society: Comparing GPT-4 and Human Respondents for Occupational Evaluation in the UK* (SSRN Scholarly Paper No. 4700366). Social Science Research Network. https://doi.org/10.2139/ssrn.4700366

Gong, D., Wan, X., & Wang, D. (2024). Working Memory Capacity of ChatGPT: An Empirical Study. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(9), Article 9. https://doi.org/10.1609/aaai.v38i9.28868

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

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3544548.3580688

Hardigan, P. C., Popovici, I., & Carvajal, M. J. (2016). Response rate, response time, and economic costs of survey research: A randomized trial of practicing pharmacists. *Research in Social and Administrative Pharmacy*, *12*(1), 141–148. https://doi.org/10.1016/j.sapharm.2015.07.003

Hwang, E., Majumder, B. P., & Tandon, N. (2023). *Aligning Language Models to User Opinions* (No. arXiv:2305.14929). arXiv. https://doi.org/10.48550/arXiv.2305.14929

Jones, C. R., & Bergen, B. K. (2025). *Large Language Models Pass the Turing Test* (No. arXiv:2503.23674). arXiv. https://doi.org/10.48550/arXiv.2503.23674

Kaiser, M., Lohmann, P., Ochieng, P., Shi, B., Sunstein, C. R., & Reisch, L. A. (2024). *Leveraging LLMs for Predictive Insights in Food Policy and Behavioral Interventions* (No. arXiv:2411.08563). arXiv. https://doi.org/10.48550/arXiv.2411.08563

Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, *33*(2), 363–374. https://doi.org/10.2307/2529786

Lee, N., An, N. M., & Thorne, J. (2023). Can Large Language Models Capture Dissenting Human Voices? In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 4569–4585). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.278

Ma, C. (2025). Evaluating Silicon Sampling: LLM Accuracy in Simulating Public Opinion on Facial Recognition Technology. *Human Interaction and Emerging Technologies (IHIET 2025)*, *5*(5). https://doi.org/10.54941/ahfe1006738

Mao, H., Liu, G., Ma, Y., Wang, R., Johnson, K., & Tang, J. (2025). *A Survey to Recent Progress Towards Understanding In-Context Learning* (No. arXiv:2402.02212). arXiv. https://doi.org/10.48550/arXiv.2402.02212

McKenna, N., Li, T., Cheng, L., Hosseini, M., Johnson, M., & Steedman, M. (2023). Sources of Hallucination by Large Language Models on Inference Tasks. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2758–2774). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.182

Mei, Q., Yutong, X., & Jackson, M. (2024). *A Turing test of whether AI chatbots are behaviorally similar to humans*. https://www.pnas.org/doi/abs/10.1073/pnas.2313925121

Meyer, S., & Elsweiler, D. (2025). LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. *International Journal of Human-Computer Studies*, *200*, 103514. https://doi.org/10.1016/j.ijhcs.2025.103514

Mistral AI. (2024). Mistral Large 2. [Large Language Model]. Retrieved from https://mistral.ai (used on 20.05.2024)

Moor, J. H. (2001). The Status and Future of the Turing Test. *Minds and Machines*, *11*(1), 77–93. https://doi.org/10.1023/A:1011218925467

Murphy, L. (2021). *Business and InnovationGreenbook Research and Industry Trend Report*. https://www.greenbook.org/grit

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024). *A Comprehensive Overview of Large Language Models* (No. arXiv:2307.06435). arXiv. https://doi.org/10.48550/arXiv.2307.06435

Oliver, A. (2013). *Behavioural Public Policy*. Cambridge University Press.

OpenAI. (2024). GPT-40-mini [Large Language Model]. Retrieved from https://www.openai.com (used on 17.05.2024)

Peterson, A. (2025). *AI and the problem of knowledge collapse*. Scilit. https://www.scilit.com/publications/781ec6de72c58611a95fbb26db79df6f

Polonioli, A. (2025). Moving LLM evaluation forward: Lessons from human judgment research. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1592399

Rahimov, A., Zamler, O., & Azaria, A. (2025). *The Turing Test Is More Relevant Than Ever* (No. arXiv:2505.02558). arXiv. https://doi.org/10.48550/arXiv.2505.02558

Rask, M., & Shimizu, K. (2024). Beyond the Average: Exploring the Potential and Challenges of Large Language Models in Social Science Research. *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, 1–5. https://doi.org/10.1109/ACDSA59508.2024.10467341

Roberts, B. W., Yao, J., Trzeciak, C. J., Bezich, L. S., Mazzarelli, A., & Trzeciak, S. (2020). Income Disparities and Nonresponse Bias in Surveys of Patient Experience. *Journal of General Internal Medicine*, *35*(7), 2217–2218. https://doi.org/10.1007/s11606-020-05677-6

Rossi, L., Harrison, K., & Shklovski, I. (2024). The Problems of LLM-generated Data in Social Science Research. *Sociologica*, *18*(2), Article 2. https://doi.org/10.6092/issn.1971-8853/19576

Sadeghian, A. H., & and Otarkhani, A. (2024). Data-driven digital nudging: A systematic literature review and future agenda. *Behaviour & Information Technology*, *43*(15), 3834–3862. https://doi.org/10.1080/0144929X.2023.2286535

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of the 40th International*

Conference on Machine Learning, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, *41*(6), 1254–1270. https://doi.org/10.1002/mar.21982

Scholtz, S. (2020). (PDF) The Use of Research Methods in Psychological Research: A Systematised Review. *ResearchGate*. https://doi.org/10.3389/frma.2020.00001

Shrestha, P., Krpan, D., Koaik, F., Schnider, R., Sayess, D., & Binbaz, M. S. (2024). Beyond WEIRD: Can synthetic survey participants substitute for humans in global policy research? *Behavioral Science & Policy*, *10*(2), 26–45. https://doi.org/10.1177/23794607241311793

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). *A Roadmap to Pluralistic Alignment* (No. arXiv:2402.05070). arXiv. https://doi.org/10.48550/arXiv.2402.05070

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, *5*(4), 417–426.

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., & Kim, J. H. (2024). *Random Silicon Sampling: Simulating Human Sub-Population Opinion Using a Large Language Model Based on Group-Level Demographic Information* (No. arXiv:2402.18144). arXiv. https://doi.org/10.48550/arXiv.2402.18144

Sunstein, C. R., Reisch, L. A., & Kaiser, M. (2019). Trusting nudges? Lessons from an international survey. *Journal of European Public Policy*, *26*(10), 1417–1443. https://doi.org/10.1080/13501763.2018.1531912

Thamer, P., Banerjee, S., & John, P. (2024). Pledging after nudging improves uptake of plantbased diets: A field experiment in a German university cafeteria. *Environmental Research Communications*. https://doi.org/10.1088/2515-7620/ad2625

Tikhonov, A., & Yamshchikov, I. P. (2023). *Post Turing: Mapping the landscape of LLM Evaluation* (No. arXiv:2311.02049). arXiv. https://doi.org/10.48550/arXiv.2311.02049

UK Government. (2025). *Memorandum of Understanding between UK and Anthropic on AI opportunities*. GOV.UK. https://www.gov.uk/government/publications/memorandum-of-understanding-between-the-uk-and-anthropic-on-ai-opportunities/memorandum-of-understanding-between-uk-and-anthropic-on-ai-opportunities

UK Statistics Authority. (2022). Ethical considerations relating to the creation and use of synthetic data. *UK Statistics Authority*. https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/

van Aert, R. C. M. (2023). Meta-analyzing partial correlation coefficients using Fisher's z transformation. *Research Synthesis Methods*, *14*(5), 768–773. https://doi.org/10.1002/jrsm.1654

van Gestel, L., Adriaanse, M. A., & de Ridder, D. (2021). Who accepts nudges? Nudge acceptability from a self-regulation perspective. *PLOS ONE*, *16*(12), e0260531. https://doi.org/10.1371/journal.pone.0260531

van Kesteren, E.-J. (2024). To democratize research with sensitive data, we should make synthetic data more accessible. *Patterns*, *5*(9), 101049. https://doi.org/10.1016/j.patter.2024.101049

Watson, D. J., Juster, R. J., & Johnson, G. W. (2008). Institutionalized Use of Citizen Surveys in the Budgetary and Policy-Making Processes: A Small City Case Study. In *The Age of Direct Citizen Participation*. Routledge.

Welz, T., Doebler, P., & Pauly, M. (2022). Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *75*(1), 1–22. https://doi.org/10.1111/bmsp.12242

WHO. (2022). *Guidance on COVID-19 for the care of older people living in long-term care facilities, other non-acute care facilities and at home*. https://iris.who.int/bitstream/handle/10665/331913/COVID-19-emergency-guidance-ageing-eng.pdf?sequence=8&isAllowed=y

Wulff, D. U., Hussain, Z., & Mata, R. (2024). *The behavioral and social sciences need open llms*. https://osf.io/ybvzs_v1/

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., Hu, Z., Jurgens, D., Evans, J., Torr, P., Ghanem, B., & Li, G. (2024). *Can Large Language Model Agents Simulate Human Trust Behavior?* (No. arXiv:2402.04559). arXiv. https://doi.org/10.48550/arXiv.2402.04559

Zhang, X., Cao, J., Wei, J., You, C., & Ding, D. (2025). *Why Prompt Design Matters and Works: A Complexity Analysis of Prompt Search Space in LLMs* (No. arXiv:2503.10084). arXiv. https://doi.org/10.48550/arXiv.2503.10084

Appendices:

A: Challenges in SP research – literature overview

B: Prompt development overview

C: Prompt templates

D: Python Code for creation of Synthetic Participants

E: Descriptive statistics

F: Item level analysis

G: Pooled inferential results

H: R Studio analysis code

I: Analysis outputs

J: Consent Form

K: Survey Flow

L: Survey

M: LSE AI form - prompts

Appendix A: Challenges in SP research – literature overview

Category	Challenge	Description	Associated authors
Quality of results	***Overrepresentation of the average***	*SP results tend to overly represent the average answer, not reporting the same variability as present in real samples*	Gmyrek et al., 2024; Hwang et al., 2023; Peterson, 2025
	Hallucinations	*LLMs can generate false or misleading information that appears plausible but is not grounded in factual data, often occurring because the models rely on learned patterns rather than verified knowledge and may 'fill in gaps' when faced with ambiguity*	McKenna et al., 2023; Naveed et al., 2024; Rossi et al., 2024
	SP results mixed	*Some SP studies yield mixed results with SP significantly deviating from their human counterpart and with mediocre or even bad alignment scores*	Amirova et al., 2024; Dillion et al., 2023; Lee et al., 2023; Santurkar et al., 2023
	Static preferences	*As many LLMs are trained with a knowledge cutoff, i.e. training data only up to a certain point, SP based on those LLMs can only simulate data up until that point*	Brand et al., 2023
	Replicability	*Identical prompts can produce varying outputs, undermining replicability due to the language models' sensitivity to prompt phrasing and probabilistic generation.*	Atil et al., 2025; Naveed et al., 2024; Rossi et al., 2024
Training Data & Bias	***Quality of training data***	*The quality of SP responses heavily depends on the training data of the underlying LLM which may contain gaps or outdated information - limiting the accuracy and representativeness of generated outputs*	Brand et al., 2023; Naveed et al., 2024; Rossi et al., 2024; Sarstedt et al., 2024
	Bias & Stereotyping	*SP often reflect or amplify biases because LLMs learn from data containing societal stereotypes and lack the ability to critically filter or contextualize them*	Dillion et al., 2023; Hwang et al., 2023; Lee et al., 2023; Santurkar et al., 2023; Shrestha et al., 2024; S
	Hard-to-reach populations	*SP tend to perform worse in simulating hard-to-reach populations because these groups are underrepresented in publicly available texts and datasets used to train LLMs*	Cao et al 2023, Weidinger et al 2023, Couldry & Mejias, 2019, Lee et al., 2023
Foundational criticism	***LLM cognition***	*Some critics say that attributing cognitive or reasoning capabilities to LLMs is fundamentally wrong as "they do not possess the non-language-specific cognitive capacities required for modelling thought" (Demszky et al., 2023, p. 8)*	Demszky et al., 2023; Rossi et al., 2024
	Replacing humans	*Ethical concerns about replacing human participants with synthetic one's center on the loss of genuine human agency, informed consent, and relational context, which may undermine the validity, respect, and social relevance of research findings*	Demszky et al., 2023; Naveed et al., 2024; UK Statistics Authority, 2022
	Lack of regulation and benchmarks	*With technology developing rapidly, there are no regulatory or ethical frameworks for using SP. With scholars utilizing different techniques to create SP, comparability of results and benchmark selection are flagged as foundational problems.*	Grossmann et al., 2023; Naveed et al., 2024; Rask & Shimizu, 2024; Santurkar et al., 2023

Appendix B: Prompt development overview

Prompting adjustment	Explanation	Observed Effect	Part of final prompt?
Single Window prompting	*When making the API call, the system message + demographic profile are sent together as standalone API calls with the idea that each SP is processed independently*	SP sample variance reduced significantly, SP behaviour too homogenous since the LLM doesn't perceive it as generating a sample but rather as single participants	✗
Multi-turn prompting	*System message is set as context window for the whole sample and demographic profiles are sent as user message within that context window*	SP answers are more coherent and display better variance than single window prompting	✓
Example vector & formatting instructions	*Giving an example of an output vector and JSON output format in the system message*	Led to higher output format consistency and also did not bias the LLMs answers which was an initial concern	✓
Varying temperatures	*The temperature setting in a large language model controls the randomness of output by scaling the logits before sampling, with higher values yielding more diverse responses.*	Lower temperature settings (0.2-0.5) lead to more response coherence but also sacrifice variance (with some LLMs answering homogenously for entire samples) while higher temperatures (0.8 – 1.2) introduce to much noise and randomness, undermining coherence. Hence, a medium temperature setting of 0.7 was chosen.	✓
Varying top-p values	*The top-p value controls how many likely word choices the model considers—for example, a top-p of 0.9 means the model picks from the smallest set of words whose combined probability is at least 90%.*	Experimenting with top-p values in parallel to changing temperature led to unpredictable interaction effects between the two which is why focus was more put on finding the right temperature.	✗
Continuous probability scales	*As variance on the binary answer scale was naturally limited and many SP samples having too little variation, I tried using a continuous scale between 0-1 as predictive outcome. The idea was recoding all values <0.5 as 0 and all values >0.5 as 1 and hence achieving more variance.*	Using continuous or binary scales did not make a significant impact on the variance in the SP sample, hence for reasons of simplicity, continuous scales were dropped.	✗
First person prompting	*Rewriting the prompt from the first vs. the third person ("I am a 24-year-old male" vs "You are simulating a 24 year old male")*	First person prompting introduced to much randomness compared to third person promptin and was thus dropped.	✗
Behavioural injections	*e.g.: "Ask yourself: Would this specific person approve or disapprove these policies? Think how the information you have about this person might help in predicting how they answered the survey."*	Led to more nuanced answers and higher score alignment on participant level	✓
Variance reminder	*e.g. "People often have varied opinions even within similar demographic groups. Reflect this diversity in your responses, considering different viewpoints and vairance that might exist within a survey population."*	As homogeneity was high, this helped in introducing more variance in the SP samples in combination with multi-turn prompting	✓

Appendix C: Prompt templates

SYSTEM MESSAGE
(Constant across all parsimony level – for complete survey questions see Appendix E)

You are simulating a human participant in a survey based on demographic information provided.

Answer the following 20 policy questions on a binary scale of 1 ("approve") or 0 ("disapprove").

1) Do you approve or disapprove of the following hypothetical policy? The federal government encourages (without requiring) electricity providers to adopt a system in which consumers would be automatically enrolled in a "green" (environmentally friendly) energy supplier, but could opt out if they wished.

… 20)

Respond ONLY with a valid JSON object with keys "N1" to "N20" and values 1 or 0. Do not include any preamble. First character must be '{'. Example for output:

{"N1": 1, "N2": 1, "N3": 1, "N4": 1, "N5": 0, "N6": 1, "N7": 1, "N8": 0, "N9": 0, "N10": 1, "N11": 0, "N12": 1, "N13": 1, "N14": 0, "N15": 0, "N16": 1, "N17": 0, "N18": 1, "N19": 1, "N20": 1}

USER MESSAGE (High degree of parsimony)

You are simulating a human participant in a survey. Answer the survey from the perspective of a person who has the beliefs of someone who has the following characteristics:

This person lives in {row['country']} and {row['native']} originally from where they live now and is a {row['gender']}. The person is {int(row['age'])} years old and {row['married']},
The person has a monthly household income of {row['income']} and {int(row['yos'])} years of schooling.

Ask yourself: Would this specific person approve or disapprove these policies? Think how the information you have about this person might help in predicting how they answered the survey.
People often have varied opinions even within similar demographic groups. Reflect this diversity in your responses, considering different viewpoints and variance that might exist within a survey population.

USER MESSAGE (Medium degree of parsimony)

You are simulating a human participant in a survey. Answer the survey from the perspective of a person who has the beliefs of someone who has the following characteristics:

This person lives in {row['country']} and {row['native']} originally from where they live now and is a {row['gender']}. The person is {int(row['age'])} years old and {row['married']}, The person has a monthly household income of {row['income']} and {int(row['yos'])} years of schooling.

This person has the following scores regarding general trust in public institutions on a scale from 1–7 (with 1 meaning 'no trust at all' and 7 meaning 'complete trust') in the following way. Trust in public institutions: {row['trust_inst']}. Also, the person believes that one can trust free markets to solve environmental and economic problems on a scale from 1–7 (with 1 meaning 'no trust at all' and 7 meaning 'complete trust') in the following way: Trust in free markets: {row['markets']}. Also, the person's concern about the environment on a scale from 1–7 (with 1 meaning 'no concern' and 7 meaning 'complete concern') in the following way: concern about the environment: {row['environment']}.

Ask yourself: Would this specific person approve or disapprove these policies? Think how the information you have about this person might help in predicting how they answered the survey.
People often have varied opinions even within similar demographic groups. Reflect this diversity in your responses, considering different viewpoints and variance that might exist within a survey population.

USER MESSAGE (Low degree of parsimony)

You are simulating a human participant in a survey. Answer the survey from the perspective of a person who has the beliefs of someone who has the following characteristics:

This person lives in {row['country']} and {row['native']} originally from where they live now and is a {row['gender']}. The person is {int(row['age'])} years old and {row['married']}, The person has a monthly household income of {row['income']} and {int(row['yos'])} years of schooling.

This person has the following scores regarding general trust in public institutions on a scale from 1–7 (with 1 meaning 'no trust at all' and 7 meaning 'complete trust') in the following way. Trust in public institutions: {row['trust_inst']}. Also, the person believes that one can trust free markets to solve environmental and economic problems on a scale from 1–7 (with 1 meaning 'no trust at all' and 7 meaning 'complete trust') in the following way: Trust in free markets: {row['markets']}. Also, the person's concern about the environment on a scale from 1–7 (with 1 meaning 'no concern' and 7 meaning 'complete concern') in the following way: concern about the environment: {row['environment']}.

This person lives in a city of size {row['city']} inhabitants. They have {row['noc']} children. They {row['money_left']} able to save money on a monthly basis.Their political ideology is rated as {row['politics']} (on a scale from 1 to 7, where 1 means 'extremely liberal' and 7 means 'extremely conservative'). They are employed in the {row['industry']} sector.Their level of job satisfaction is {row['job_satisfaction']} (on a scale from 1 to 7, where 1 means 'no satisfaction at all' and 7 means 'complete satisfaction'). They {row['friends']} have close friends in their local community.Their general trust in government institutions is rated at {row['trust_ggen']} (on a scale from 1 to 7, where 1 means 'no trust at all' and 7 means 'complete trust'). Their concern for their family's future health is {row['health_concernf']} (on a scale from 1 to 7, where 1 means 'no concern at all' and 7 means 'complete concern').

Ask yourself: Would this specific person approve or disapprove these policies? Think how the information you have about this person might help in predicting how they answered the survey.
People often have varied opinions even within similar demographic groups. Reflect this diversity in your responses, considering different viewpoints and variance that might exist within a survey population.

Appendix D: Python Code for the creation of Synthetic Participants

Example: Low parsimony Code for Claude Sonnet 4 via Anthropic API

```
#LOW PARSIMONY CODE for Claude**

# Install packages
!pip install anthropic pandas tqdm openpyxl --quiet

# Imports
import anthropic
import pandas as pd
import time
import json
import re
from tqdm import tqdm
from google.colab import drive
import os

# Mount Drive
drive.mount('/content/drive')

# Load data
input_path = "/content/drive/MyDrive/Synthetic Coding Folder/Input Files/T1 Data/US Sample/T2_US_Low_Parsimony_Input.xlsx"
df = pd.read_excel(input_path)
df = df.head(300)

# Anthropic config
api_key = "XXX"  # Anthropic API key
client = anthropic.Anthropic(api_key=api_key)
model = "claude-sonnet-4-20250514"
temperature = 0.7  # Set temperature
max_tokens = 2500 # Set max tokens

# SYSTEM message: Static survey and response format instructions
system_message = """
You are simulating a human participant in a survey based on demographic information provided.

Answer the following 20 policy questions on a binary scale of 1 ("approve") or 0 ("disapprove").
#FOR FULL SURVEY SEE APPENDIX E
1) …
…
20)…

Respond ONLY with a valid JSON object with keys "N1" to "N20" and values 1 or 0. Do not include any preamble. First character must be '{'.
Example for output:
{"N1": 1, "N2": 1, "N3": 1, "N4": 1, "N5": 0, "N6": 1, "N7": 1, "N8": 0, "N9": 0, "N10": 1, "N11": 0, "N12": 1, "N13": 1, "N14": 0, "N15": 0, "N16": 1, "N17": 0, "N18": 1, "N19": 1, "N20": 1}
"""

# Collect responses
responses = []
for idx, row in tqdm(df.iterrows(), total=len(df)):
    # Debug: Print the demographic info being used
    print(f"\nProcessing participant {idx}: {row['ID']}")
    print(f"Demographics: {row['country']}, {row['native']}, {row['gender']}, age {int(row['age'])}, {row['married']}, income {row['income']}, education {int(row['yos'])}")
    print(f"Trust/Attitudes: Trust in government {row['trust_ggen']}, Trust in markets {row['markets']}, Environmental concern {row['environment']}")
    print(f"Additional: City size {row['city']}, Children {row['noc']}, Money saving {row['money_left']}, Politics {row['politics']}, Industry {row['industry']}, Job satisfaction {row['job_satisfaction']}, Friends {row['friends']}, Trust gov {row['trust_ggen']}, Health concern {row['health_concernf']}")


    user_message = (
        f"You are simulating a human participant in a survey. "
        f"Answer the survey from the perspective of a person who has the beliefs of someone who has the following characteristics: "
        f"This person lives in {row['country']} and {row['native']} originally from where they live now and is a {row['gender']}. The person is {int(row['age'])} years old and {row['married']}, "
        f"The person has a monthly household income of {row['income']} and {int(row['yos'])} years of schooling. "
        f"This person has the following scores regarding trust in public institutions on a scale from 1-7 (with 1 meaning 'no trust at all' and 7 meaning 'complete trust' in the following way). Trust in public institutions: {row['trust_inst']}. "
        f"Also, the person believes that one can trust free markets to solve environmental and economic problems on a scale from 1-7 with 1 meaning 'no trust at all' and 7 meaning 'complete trust') in the following way: Trust in free markets: {row['markets']}. "
        f"Also, the person's concern about the environment on a scale from 1-7 with 1 meaning 'no concern' and 7 meaning 'complete concern') in the following way: concern about the environment: {row['environment']}. "
        f"This person lives in a city of size {row['city']} inhabitants. They have {row['noc']} children. They {row['money_left']} able to save money on a monthly basis. "
        f"Their political ideology is rated as {row['politics']} (on a scale from 1 to 7, where 1 means 'extremely liberal' and 7 means 'extremely conservative'). They are employed in the {row['industry']} sector. "
        f"Their level of job satisfaction is {row['job_satisfaction']} (on a scale from 1 to 7, where 1 means 'no satisfaction at all' and 7 means 'complete satisfaction'). They {row['friends']} have close friends in their local community. "
        f"Their general trust in government institutions is rated at {row['trust_ggen']} (on a scale from 1 to 7, where 1 means 'no trust at all' and 7 means 'complete trust'). Their concern for their family's future health is {row['health_concernf']} (on a scale from 1 to 7, where 1 means 'no concern at all' and 7 means 'complete concern'). "
        f"Ask yourself: Would this specific person approve or disapprove these policies? Think how the information you have about this person might help in predicting how they answered the survey. "
        f"People often have varied opinions even within similar demographic groups. Reflect this diversity in your responses, considering different viewpoints and variance that might exist within a survey population."
    )
    # Debug: Print the actual message being sent
```

```
    print(f"User message preview: {user_message[:200]}...")

    try:
        response = client.messages.create(
            model=model,
            max_tokens=max_tokens,
            temperature=temperature,
            system=system_message,
            messages=[
                {"role": "user", "content": user_message}
            ]
        )

        content = response.content[0].text.strip()
        print(f"Response preview: {content[:400]}...")

        match = re.search(r'\{.*\}', content, re.DOTALL)
        if match:
            json_str = match.group()
            parsed = json.loads(json_str)
            responses.append({"ID": row["ID"], **parsed})
        else:
            raise ValueError("No JSON object found in response.")

    except json.JSONDecodeError as jde:
        print(f"JSON error at index {idx}: {jde} | Response was: {content}")
        responses.append({"ID": row["ID"], "error": f"JSONDecodeError: {jde}", "raw": content})
    except Exception as e:
        print(f"General error at index {idx}: {e}")
        responses.append({"ID": row["ID"], "error": str(e)})

    time.sleep(1.5)

# Save to Excel
output_path = "/content/drive/MyDrive/Synthetic Coding Folder/Output Files/T2 Data/US/T2_US_LP_Claude_0.7.xlsx
os.makedirs(os.path.dirname(output_path), exist_ok=True)

df_out = pd.DataFrame(responses)
with pd.ExcelWriter(output_path, engine='openpyxl') as writer:
    df_out.to_excel(writer, index=False, sheet_name="SyntheticResponses")

print(f"Saved to {output_path}")
```

Appendix E: Descriptive statistics

US Sample (n = 300)

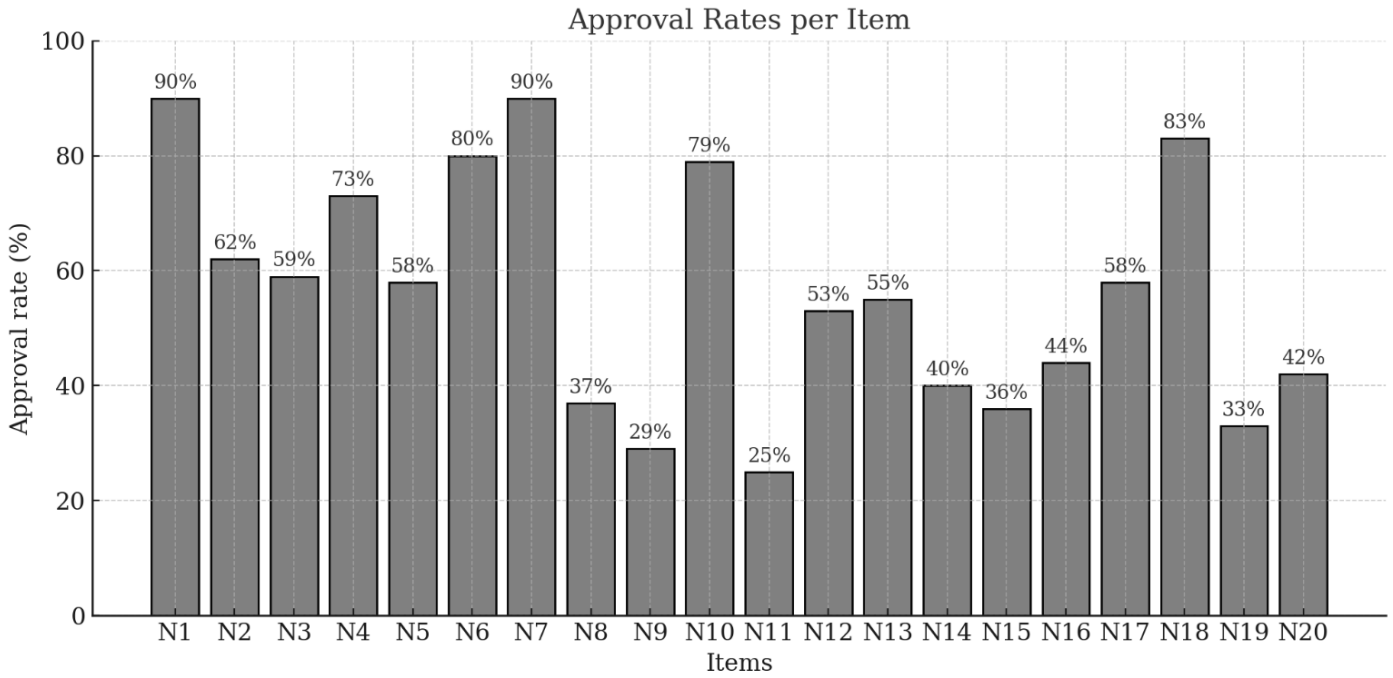
Age		Gender		Education		Income		Urban/rural	
18-24	2.4%	*Female*	62.3%	*High school*	22%	*Do not want to answer*	10.3%	*Urban*	64.3%
25-34	6.3%	*Male*	37.3%	*College Degree*	18.6%	*Below $1.250*	11%	*Rural*	35.6%
35-44	11.3%	*None of the two / prefer not to say*	0.3%	*College, no degree*	17.3%	*$1.250 - $2.500*	12%		
45-54	14.6%			*Associate degree*	13.3%	*$2.500 - $3.750*	9.6%		
55-64	17.6%			*Masters degree*	12.6%	*$3.750 - $5.000*	9.6%		
65+	47.3%			*Completed some High school*	4.6%	*$5.000 - $6.250*	10.6%		
				Post high school vocational training	4%	*$6.250 - $7.500*	6.3%		
				Some graduate, no degree	3.3%	*$7.500 - $10.000*	9.6%		
				Doctoral degree	2.3%	*$10.000 - $12.500*	4.6%		
				Middle school	1.6%	*$12.500 - $16.250*	3%		
						More than $16.250	12.6%		
= 99.5%		**= 99.9%**		**= 99.6%**		**= 99.2%**		**= 99.9%**	

Egypt Sample (n = 300)

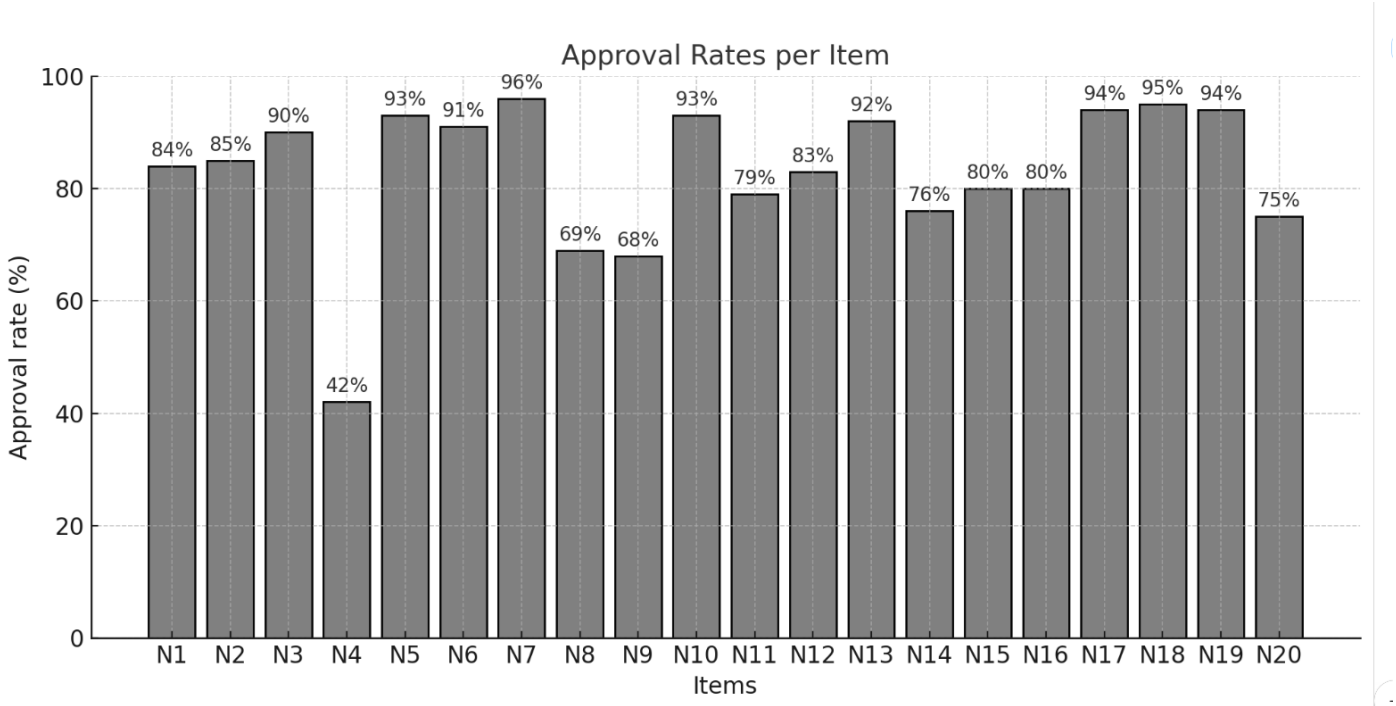
Age		Gender		Education		Income		Urban/rural	
18-24	21%	*Female*	33.3%	*Upper secondary education*	62%	*Do not want to answer*	7%	*Urban*	74.6%
25-34	32.3%	*Male*	66 %	*Non Upper seconary education*	38%	*Below 4500* E£	17.6%	*Rural*	25.3%
35-44	28.9%	*None of the two / prefer not to say*	0.3%			*4.500 - 5.500* E£	8.6%		
45-54	11.3%					*5.500 - 6.300* E£	7%		
55-64	5%					*6.300 - 7.000* E£	5%		
65+	1.3%					*7.000 - 7.900* E£	6.3%		
						7.900 - 8.800 E£	5.3%		
						8.800 - 10.100 E£	10.3%		
						10.100 - 12.000 E£	7.6%		
						12.000 - 15.900 E£	7.6%		
						More than 15.900 E£	17.3%		
= 99.8%		**= 99.9%**		**= 100%**		**= 99.6%**		**= 99.9%**	

Empirical data distribution of real samples

USA (n=300)



EGY (n=300)



Appendix F: Item level analysis

Item	Policy Description (Short)	Behavioural Public Policy Type	MAE Mean	Match Rate Mean
N1	Calorie labels in restaurants	Informational	**0.10**	**83**
N2	Traffic light food labels	Informational	**0.14**	**78**
N3	Opt-out: Green energy (encouraged)	Default (Soft)	**0.27**	**70**
N4	Organ donor declaration	Prompted Choice	**0.41**	**62**
N5	Healthy food placement	Choice Architecture	**0.23**	**75**
N6	Anti-texting campaign (graphic)	Informational	**0.14**	**80**
N7	Childhood obesity info campaign	Informational	**0.06**	**88**
N8	Subliminal anti-smoking/overeating ads	Informational	**0.25**	**65**
N9	Opt-out: Carbon offset fee	Default (Financial)	**0.25**	**68**
N10	High-salt product labels	Informational	**0.20**	**72**
N11	Tax default: $55 charity donation	Default (Financial)	**0.51**	**58**
N12	Anti-smoking/overeating cinema ads	Informational	**0.21**	**74**
N13	Opt-out: Green energy (mandatory)	Default (Environmental)	**0.34**	**64**
N14	Cashier areas free of sweets	Choice Restriction	**0.28**	**67**
N15	Mandatory meat-free day in canteens	Mandate/Lifestyle	**0.37**	**61**
N16	Tax compliance norm message	Norm Messaging	**0.16**	**82**
N17	Highlight healthy menu options	Choice Architecture	**0.28**	**69**
N18	Automatic SMS to parents when absent	Informational	**0.14**	**81**
N19	Vaccination: Commitment letter	Commitment Device	**0.25**	**65**
N20	Break reminders on social media	Informational/Prompt	**0.23**	**71**

Appendix G: Pooled Inferential results

Global results (Across country, models, parsimony levels)

Metric	Global (95% CI)
Cohen's κ	0.14 [0.11 – 0.17]
Odds Ratio (OR)	0.80 [0.44 – 1.47]
Spearman's ρ	0.61 [0.19 – 0.85]

Model results (Across country, parsimony levels)

Model	Cohen's κ (95% CI)	Odds Ratio (95% CI)	Spearman's ρ (95% CI)
GPT-4o mini	0.13 [0.10 – 0.15]	0.88 [0.49 – 1.63]	0.57 [0.14 – 0.82]
Claude Sonnet 4	0.17 [0.14 – 0.19]	0.51 [0.27 – 0.91]	0.68 [0.33 – 0.86]
Mistral Large 2	0.13 [0.11 – 0.16]	1.09 [0.56 – 2.13]	0.60 [0.14 – 0.84]

Prompt parsimony results (Across country, models)

Prompt Parsimony	Cohen's κ (95% CI)	Odds Ratio (95% CI)	Spearman's ρ (95% CI)
Low	0.15 [0.13 – 0.17]	1.01 [0.70 – 1.46]	0.64 [0.36 – 0.81]
Medium	0.12 [0.11 – 0.14]	0.85 [0.57 – 1.26]	0.61 [0.33 – 0.79]
High	0.15 [0.13 – 0.17]	1.34 [0.87 – 2.07]	0.65 [0.37 – 0.82]

Country results (Across model, prompt parsimony)

Metric	USA (95% CI)	Egypt (95% CI)
Cohen's κ	0.22 [0.20 – 0.25]	0.06 [0.04 – 0.09]
Odds Ratio (OR)	1.53 [0.86 – 2.02]	0.42 [0.23 – 0.97]
Spearman's ρ	0.75 [0.47 – 0.91]	0.42 [0.10 – 0.77]

Appendix H: R-Studio analysis code

Code for descriptive statistics (MAE, Match rates)

➔ *Calculated per SP dataset (i.e. executed 18 times) – for results see Appendix I*

```
# Setup
library(readxl)
library(dplyr)
library(tibble)
library(gridExtra)
library(grid)
library(tools)  # for file_path_sans_ext()

# Load data
human_data <- read_xlsx("T2_US_Real_data_BENCHMARK.xlsx")
synthetic_file <- "T2_US_LP_fullsample_Mistral_0.7.xlsx"  # keep file path separate
synthetic_data <- read_xlsx(synthetic_file)

# Extract file name (without extension) for header
synthetic_header <- file_path_sans_ext(basename(synthetic_file))

# Define nudge names
nudge_names <- paste0("N", 1:20)

# 1) Approval rates per nudge & means

approval_human <- sapply(nudge_names, function(nm) mean(human_data[[nm]] == 1, na.rm=TRUE))
approval_synth <- sapply(nudge_names, function(nm) mean(synthetic_data[[nm]] == 1, na.rm=TRUE))

mean_approval_human <- mean(approval_human)
mean_approval_synth <- mean(approval_synth)

# 2) Mean Absolute Error (MAE) per nudge & overall

mae_per_nudge <- abs(approval_human - approval_synth)
mae_overall <- mean(mae_per_nudge)

#
# 3) Match rate (considering BOTH 1s and 0s)

match_rate_per_nudge <- sapply(nudge_names, function(nm) {
  mean(human_data[[nm]] == synthetic_data[[nm]], na.rm=TRUE)
})
match_rate_overall <- mean(unlist(human_data[nudge_names]) == unlist(synthetic_data[nudge_names]), na.rm=TRUE)

# 4) Assemble results into ONE neat table
results_table <- tibble(
  nudge = nudge_names,
  approval_human = round(approval_human, 3),
  approval_synth = round(approval_synth, 3),
  mae = round(mae_per_nudge, 3),
  match_rate = round(match_rate_per_nudge, 3)
)

# Add an extra "Mean (Overall)" row at the bottom
results_table <- bind_rows(
```

```
  results_table,
  tibble(
    nudge = "Mean (Overall)",
    approval_human = round(mean_approval_human, 3),
    approval_synth = round(mean_approval_synth, 3),
    mae = round(mae_overall, 3),
    match_rate = round(match_rate_overall, 3)
  )
)

# 5) Display final results in Plots pane with header
grid.newpage()
gridExtra::grid.table(results_table)
grid.text(synthetic_header, y = unit(0.98, "npc"), gp = gpar(fontsize = 14, fontface = "bold"))

# 5) Display final results in Plots pane with header AND save to PDF

# Define PDF output file name dynamically
pdf_filename <- paste0("descStat_output_", synthetic_header, ".pdf")

# Save the table to PDF
pdf(pdf_filename, width = 8.5, height = 11)  # Adjust size as needed
grid.newpage()
gridExtra::grid.table(results_table)
grid.text(synthetic_header, y = unit(0.98, "npc"), gp = gpar(fontsize = 14, fontface = "bold"))
dev.off()
```

Code for inferential metrics (Kappa, OR, spearman correlations + bootstrapping + CI)

➔ *Calculated per SP dataset (i.e. executed 18 times) – for results see Appendix I*

```
library(readxl)
library(dplyr)
library(irr)
library(DescTools)
library(boot)
library(tidyr)
library(lme4)
library(broom.mixed)
library(ggplot2)
library(gridExtra)
library(broom)
library(sandwich)
library(lmtest)
library(grid)
library(stringr)
library(tools)

# --- Load data ---
human_data <- read_xlsx("T2_Egypt_Real_data_BENCHMARK.xlsx")
synthetic_file <- "T2_Egypt_LP_fullsample_Claude_0.7.xlsx"
synthetic_data <- read_xlsx(synthetic_file)

# Ensure ID alignment
stopifnot(all(human_data$ID == synthetic_data$ID))

# Get dataset name label for output
dataset_label <- gsub(".xlsx", "", synthetic_file)

# Start PDF export
pdf(file = paste0("inf_output_", tools::file_path_sans_ext(basename(synthetic_file)), ".pdf"),
    width = 11, height = 8.5)

# 1) PERSON-LEVEL ALIGNMENT
calc_kappa <- function(h_row, s_row) {
  df <- data.frame(rater1 = as.numeric(h_row), rater2 = as.numeric(s_row))
  df <- na.omit(df)
  if (nrow(df) == 0 || length(unique(df$rater1)) == 1 || length(unique(df$rater2)) == 1) {
    return(NA)
  }
  return(tryCatch(kappa2(df, weight = "unweighted")$value, error = function(e) NA))
}

kappa_values <- purrr::map2_dbl(
  split(human_data[,-1], 1:nrow(human_data)),
  split(synthetic_data[,-1], 1:nrow(synthetic_data)),
  ~ calc_kappa(.x, .y)
)

excluded_n <- sum(is.na(kappa_values))
included_n <- length(kappa_values) - excluded_n
valid_kappa <- kappa_values[!is.na(kappa_values)]

safe_boot_ci <- function(x) {
  if (length(unique(x)) <= 1 || length(x) < 5) {
    return(c(NA, NA))
```

```
  } else {
    boot_res <- boot(data = x, statistic = function(d, i) mean(d[i], na.rm = TRUE), R = 1000)
    return(boot.ci(boot_res, type = "perc")$percent[4:5])
  }
}

ci_kappa <- safe_boot_ci(valid_kappa)

person_table <- tibble(
  Metric = c("Mean Cohen's Kappa", "Participants included", "Participants excluded"),
  Estimate = c(mean(valid_kappa), included_n, excluded_n),
  `95% CI Lower` = c(ci_kappa[1], NA, NA),
  `95% CI Upper` = c(ci_kappa[2], NA, NA)
)

grid.arrange(
  grobTree(textGrob(paste("Person level alignment -", dataset_label),
                    x=0.5, y=0.9, gp=gpar(fontsize=16, fontface="bold"))),
  tableGrob(person_table, rows = NULL),
  heights = c(0.2, 1)
)
# 2) ITEM-LEVEL ALIGNMENT (Clustered GLM with Validation and Ordered Items)

library(dplyr)
library(tidyr)
library(broom)
library(sandwich)
library(lmtest)
library(gridExtra)
library(grid)
library(stringr)

# --- Define expected items ---
expected_items <- paste0("N", 1:20)

# --- Validation: Check structure ---
if (!all(colnames(human_data)[1] == "ID" && all(expected_items %in% colnames(human_data)))) {
  stop(" Human data does not have the expected structure: ID + N1...N20.")
}
if (!all(colnames(synthetic_data)[1] == "ID" && all(expected_items %in% colnames(synthetic_data)))) {
  stop(" Synthetic data does not have the expected structure: ID + N1...N20.")
}

# --- Approval rates in WIDE format ---
approval_real_wide <- colMeans(human_data[, expected_items], na.rm = TRUE) * 100
approval_synth_wide <- colMeans(synthetic_data[, expected_items], na.rm = TRUE) * 100

# --- Convert to LONG format ---
human_long <- human_data %>%
  select(ID, all_of(expected_items)) %>%
  pivot_longer(cols = all_of(expected_items), names_to = "item", values_to = "response") %>%
  mutate(dataset = "real")

synthetic_long <- synthetic_data %>%
  select(ID, all_of(expected_items)) %>%
  pivot_longer(cols = all_of(expected_items), names_to = "item", values_to = "response") %>%
  mutate(dataset = "synthetic")

# Combine and clean
```

```r
combined_long <- bind_rows(human_long, synthetic_long) %>%
  mutate(response = as.numeric(response)) %>%
  filter(response %in% c(0, 1))

# --- Approval rates in LONG format ---
approval_long_check <- combined_long %>%
  group_by(dataset, item) %>%
  summarise(approval = mean(response) * 100, .groups = "drop") %>%
  pivot_wider(names_from = dataset, values_from = approval, names_prefix = "approval_") %>%
  arrange(match(item, expected_items))  # Ensure correct item order

# --- Force order in wide approval rates too ---
approval_real_wide <- approval_real_wide[expected_items]
approval_synth_wide <- approval_synth_wide[expected_items]

# --- Validation Comparison ---
comparison_check <- tibble(
  Item = expected_items,
  Real_Wide = round(approval_real_wide, 2),
  Real_Long = round(approval_long_check$approval_real, 2),
  Diff_Real = round(approval_real_wide - approval_long_check$approval_real, 2),
  Synth_Wide = round(approval_synth_wide, 2),
  Synth_Long = round(approval_long_check$approval_synthetic, 2),
  Diff_Synth = round(approval_synth_wide - approval_long_check$approval_synthetic, 2)
) %>%
  mutate(Flag = ifelse(abs(Diff_Real) > 1e-6 | abs(Diff_Synth) > 1e-6, "⚠", ""))

print(comparison_check)

# --- Stop if validation fails ---
if (any(abs(comparison_check$Diff_Real) > 1e-6) || any(abs(comparison_check$Diff_Synth) > 1e-6)) {
  cat("\n Validation failed: Approval rates mismatch between wide and long formats.\n")
  cat("Check 'comparison_check' table above to identify mismatched items.\n")
  stop()
} else {
  cat("\n Validation passed: Approval rates match between wide and long formats.\n")
}

# --- Compute approval rates and MAE ---
approval_rates <- approval_long_check %>%
  mutate(delta = approval_synthetic - approval_real,
         MAE = abs(delta))

global_approval <- approval_rates %>%
  summarise(
    approval_real = mean(approval_real),
    approval_synthetic = mean(approval_synthetic),
    delta = mean(delta),
    MAE = mean(MAE)
  )

# --- Logistic Regression (Clustered SE for Global OR) ---
glm_model <- glm(response ~ dataset, data = combined_long, family = binomial)
cluster_se <- vcovCL(glm_model, cluster = combined_long$item)
global_test <- coeftest(glm_model, vcov = cluster_se)

global_estimate <- global_test["datasetsynthetic", ]
global_or <- tibble(
  Item = "GLOBAL",
```

```
  OR = exp(global_estimate[1]),
  CI_lower = exp(global_estimate[1] - 1.96 * global_estimate[2]),
  CI_upper = exp(global_estimate[1] + 1.96 * global_estimate[2]),
  p_value = global_estimate[4]
)

# --- Per-item logistic ORs ---
per_item_or <- combined_long %>%
  group_by(item) %>%
  do({
    mod <- glm(response ~ dataset, data = ., family = binomial)
    est <- summary(mod)$coefficients["datasetsynthetic", ]
    tibble(
      OR = exp(est[1]),
      CI_lower = exp(est[1] - 1.96 * est[2]),
      CI_upper = exp(est[1] + 1.96 * est[2]),
      p_value = est[4]
    )
  }) %>% ungroup()

# --- Significance stars ---
add_sig_stars <- function(p) {
  case_when(
    p < 0.001 ~ "***",
    p < 0.01  ~ "**",
    p < 0.05  ~ "*",
    TRUE ~ ""
  )
}

per_item_or <- per_item_or %>%
  mutate(Item = item,
         sig = add_sig_stars(p_value)) %>%
  select(Item, OR, CI_lower, CI_upper, p_value, sig) %>%
  arrange(as.numeric(str_extract(Item, "\\d+")))

# --- Formatting ---
format_num <- function(x) ifelse(is.na(x), NA, sprintf("%.2f", x))
format_p <- function(p) ifelse(p < 0.001, "<0.001", sprintf("%.3f", p))

per_item_or <- per_item_or %>%
  mutate(OR = format_num(OR),
         CI_lower = format_num(CI_lower),
         CI_upper = format_num(CI_upper),
         p_value = format_p(p_value))

# --- Merge approval metrics ---
per_item_or <- per_item_or %>%
  left_join(approval_rates, by = c("Item" = "item")) %>%
  mutate(
    `Real %` = sprintf("%.1f%%", approval_real),
    `Synthetic %` = sprintf("%.1f%%", approval_synthetic),
    `Δ%` = sprintf("%.1f", delta),
    `MAE` = sprintf("%.1f", MAE)
  ) %>%
  select(Item, OR, CI_lower, CI_upper, p_value, sig, `Real %`, `Synthetic %`, `Δ%`, `MAE`)

global_or <- global_or %>%
  mutate(
```

```
    OR = sprintf("%.2f", OR),
    CI_lower = sprintf("%.2f", CI_lower),
    CI_upper = sprintf("%.2f", CI_upper),
    p_value = format_p(p_value),
    sig = add_sig_stars(as.numeric(p_value)),
    `Real %` = sprintf("%.1f%%", global_approval$approval_real),
    `Synthetic %` = sprintf("%.1f%%", global_approval$approval_synthetic),
    `Δ%` = sprintf("%.1f", global_approval$delta),
    `MAE` = sprintf("%.1f", global_approval$MAE)
  ) %>%
  select(Item, OR, CI_lower, CI_upper, p_value, sig, `Real %`, `Synthetic %`, `Δ%`, `MAE`)

# --- Final table ---
item_or_table <- bind_rows(global_or, per_item_or)

grid.arrange(
  grobTree(textGrob(paste("Item-level Alignment (OR, Approval %, Δ%, MAE) -", dataset_label),
                    x = 0.5, y = 0.9, gp = gpar(fontsize = 16, fontface = "bold"))),
  tableGrob(item_or_table, rows = NULL),
  heights = c(0.2, 1)
)
print(item_or_table)

# 3) PATTERN-LEVEL ALIGNMENT (Spearman's ρ)

approval_rates <- combined_long %>%
  group_by(dataset, item) %>%
  summarise(approval = mean(response), .groups="drop") %>%
  pivot_wider(names_from=dataset, values_from=approval)

spearman_obs <- cor(approval_rates$real, approval_rates$synthetic, method="spearman")

set.seed(123)
spearman_boot <- boot(approval_rates, statistic=function(d, i) {
  cor(d[i, "real"], d[i, "synthetic"], method="spearman")
}, R=10000)
spearman_ci <- quantile(spearman_boot$t, c(0.025, 0.975))

set.seed(123)
n_perm <- 10000
perm_rhos <- replicate(n_perm, {
  shuffled <- sample(approval_rates$synthetic)
  cor(approval_rates$real, shuffled, method="spearman")
})
p_perm <- mean(abs(perm_rhos) >= abs(spearman_obs))

approval_long <- approval_rates %>%
  pivot_longer(cols=c(real, synthetic), names_to="dataset", values_to="approval")

ggplot(approval_long, aes(x=reorder(item, approval), y=approval, fill=dataset)) +
  geom_bar(stat="identity", position="dodge") +
  coord_flip() +
  labs(
    title = paste0("Pattern Alignment (", dataset_label, "): Spearman ρ = ", round(spearman_obs, 2),
                   " [95% CI: ", round(spearman_ci[1], 2), ", ", round(spearman_ci[2], 2),
                   "], p_perm = ", format.pval(p_perm, digits=3)),
    x = "Item (ranked by approval)",
    y = "Approval Rate"
  ) +
```

```
  theme_minimal() +
  scale_fill_manual(values=c("real"="#1f77b4", "synthetic"="#ff7f0e"),
                    name="Dataset", labels=c("Real", "Synthetic")) +
  theme(
    plot.title = element_text(size=14, face="bold"),
    axis.text.y = element_text(size=10),
    legend.position = "top"
  )

# Close PDF
dev.off()
cat("All outputs saved to:", paste0(getwd
```

Appendix I: Analysis outputs

For each of the 18 datasets – descriptive and inferential metrics were calculated separately using the R-Studio code (see Appendix H) and saved in a PDF file with the file name indicating which metrics for which dataset were computed

Overview of abbreviations:

Desc_stat_output: Descriptive metrics
Inf_stat_output: Inferential metrics

HP: High degree of parsimony
MP: Medium degree of parsimony
LP: Low degree of parsimony

US: United States of America
EGY: Egypt

GPT: GPT4-o-mini
Mistral: Mistral Large 2
Claude: Claude Sonnet 4

0.7: Denotes the temperature parameter used in the SP data generation

As each analysis output per dataset has four pages, it was too long to directly include in this appendix – this would have made an additional 18 x 4 = 72 pages. Hence, all files were uploaded into a public Google Drive and are accessible here, ordered by the terminology introduced above.

Dataset	Metrics	Link
descStat_output_T2_US_MP_fullsample_Mistral_0.7	descriptive	available here
descStat_output_T2_US_MP_fullsample_GPT_0.7	descriptive	available here
descStat_output_T2_US_MP_fullsample_Claude_0.7	descriptive	available here
descStat_output_T2_US_LP_fullsample_Mistral_0.7	descriptive	available here
descStat_output_T2_US_LP_fullsample_GPT_0.7	descriptive	available here
descStat_output_T2_US_LP_fullsample_Claude_0.7	descriptive	available here
descStat_output_T2_US_HP_fullsample_Mistral_0.7	descriptive	available here
descStat_output_T2_US_HP_fullsample_GPT_0.7	descriptive	available here
descStat_output_T2_US_HP_fullsample_Claude_0.7	descriptive	available here
descStat_output_T2_Egypt_MP_fullsample_Mistral_0.7	descriptive	available here

descStat_output_T2_Egypt_MP_fullsample_GPT_0.7	descriptive	available here
descStat_output_T2_Egypt_MP_fullsample_Claude_0.7	descriptive	available here
descStat_output_T2_Egypt_LP_fullsample_Mistral_0.7	descriptive	available here
descStat_output_T2_Egypt_LP_fullsample_GPT_0.7	descriptive	available here
descStat_output_T2_Egypt_LP_fullsample_Claude_0.7	descriptive	available here
descStat_output_T2_Egypt_HP_fullsample_Mistral_0.7	descriptive	available here
descStat_output_T2_Egypt_HP_fullsample_GPT_0.7	descriptive	available here
descStat_output_T2_Egypt_HP_fullsample_Claude_0.7	descriptive	available here
inf_output_T2_US_MP_fullsample_Mistral_0.7	inferential	available here
inf_output_T2_US_MP_fullsample_GPT_0.7	inferential	available here
inf_output_T2_US_MP_fullsample_Claude_0.7	inferential	available here
inf_output_T2_US_LP_fullsample_Mistral_0.7	inferential	available here
inf_output_T2_US_LP_fullsample_GPT_0.7	inferential	available here
inf_output_T2_US_LP_fullsample_Claude_0.7	inferential	available here
inf_output_T2_US_HP_fullsample_Mistral_0.7	inferential	available here
inf_output_T2_US_HP_fullsample_GPT_0.7	inferential	available here
inf_output_T2_US_HP_fullsample_Claude_0.7	inferential	available here
inf_output_T2_Egypt_MP_fullsample_Mistral_0.7	inferential	available here
inf_output_T2_Egypt_MP_fullsample_GPT_0.7	inferential	available here
inf_output_T2_Egypt_MP_fullsample_Claude_0.7	inferential	available here
inf_output_T2_Egypt_LP_fullsample_Mistral_0.7	inferential	available here
inf_output_T2_Egypt_LP_fullsample_GPT_0.7	inferential	available here
inf_output_T2_Egypt_LP_fullsample_Claude_0.7	inferential	available here
inf_output_T2_Egypt_HP_fullsample_Mistral_0.7	inferential	available here
inf_output_T2_Egypt_HP_fullsample_GPT_0.7	inferential	available here
inf_output_T2_Egypt_HP_fullsample_Claude_0.7	inferential	available here

Appendix J: Consent Form

UNIVERSITY OF CAMBRIDGE
Judge Business School

Survey consent form

Thank you for taking part in this survey. It should take approximately 20 minutes to complete.

This research is being conducted by Malte Dewies, a post-doctoral researcher at Cambridge Judge Business School (CJBS) under the supervision of Professor Lucia Reisch, a faculty member at CJBS. The research aims to investigate the approval for different interventions changing behaviour.

Please read the notes below carefully. If you are happy to participate in this study, please click to confirm below.

- I understand the scope of this project and have had the opportunity to ask the researcher any questions that I have about the study and my involvement in it. I am aware that I can contact the researcher Malte Dewies m.dewies@jbs.cam.ac.uk or their supervisor Lucia Reisch lr540@cam.ac.uk at any time.
- I understand that my survey responses will be stored initially by Qualtrics, whose privacy policy is here https://www.qualtrics.com/privacy-statement/. Once the survey is closed, any personally identifiable data may be moved to a secure server at CJBS, accessible only by CJBS IT staff and the research team Malte Dewies, Josh Ramminger, Konrad Bertram, Micha Kaiser, Lucia Reisch. Josh Ramminger's primary research affiliation is the Humboldt-Universität zu Berlin and Konrad Bertram's the London School of Economics and Political Science (LSE), but identifiable data will not be shared or stored with these institutions.
- Visit our privacy policy: www.jbs.cam.ac.uk/about-this-site/privacy-policy
- Find out about how we process research participant data: www.informationhttp://www.information-compliance.admin.cam.ac.uk/data-protection/research-participant-datacompliance.admin.cam.ac.uk/data-protection/research-participant-data
- I agree that anonymised data may be used in publication of the results and shared on the University of Cambridge's public data repository Apollo www.repository.cam.ac.uk.
- I understand that all efforts will be made to ensure I cannot be identified (except as may be required by law). However, I am aware that information in the public domain may make it possible to identify me.
- I understand that I am free to withdraw at any time during the survey without giving a reason, by closing the browser.

You may wish to print this page for future reference.

☐ I confirm that I have read and understand the information above and voluntarily consent to participate in this study.

☐ I do not wish to participate in this study.

University of Cambridge
Trumpington Street
Cambridge
CB2 1AG, UK

Appendix K: Survey Flow

EmbeddedData
opp = Qual4236-1104NatRep
transaction_idValue will be set from Panel or URL.
SVIDValue will be set from Panel or URL.
expected_ir = .8
expected_loi = 8
Q_BallotBoxStuffingValue will be set from Panel or URL.
Q_RecaptchaScoreValue will be set from Panel or URL.
Q_TotalDurationValue will be set from Panel or URL.

Branch: New Branch
If
If Quota Overall quota (900) Has Been Met

EmbeddedData
gc = 3
term = overquota_start

EndSurvey: Advanced

Block: Default Question Block (9 Questions)

Branch: New Branch
If
If Which country are you currently living in? Other Is Selected

EmbeddedData
gc = 2
term = other_country

EndSurvey: Advanced

Branch: New Branch
If
If Thank you for taking part in this survey. It should take approximately 20 minutes to complete. Th... I do not wish to participate in this study Is Selected

EmbeddedData
gc = 2
term = consent

EndSurvey: Advanced

Branch: New Branch
If
If What is your age? (in years) under 18 Is Selected

EmbeddedData

gc = 2
term = age

EndSurvey: Advanced

Branch: New Branch
If
If Q_BallotBoxStuffing Is Equal to 1

EmbeddedData
gc = 4
term = quality_dupe

EndSurvey: Advanced

Branch: New Branch
If
If Q_RecaptchaScore Is Less Than or Equal to 0.6

EmbeddedData
gc = 4
term = quality_bot

EndSurvey: Advanced

Standard: Nudges 1-20 (21 Questions)

Branch: New Branch
If
If This is an attention filter. Please click “Disapprove” to go on with the survey. Approve Is Selected

EmbeddedData
gc = 4
term = attention_check1

EndSurvey: Advanced

Standard: Sociodemographics - Part 2 (10 Questions)

Branch: New Branch
If
If This is an attention filter. Please select "Approve" to continue. Disapprove Is Selected

EmbeddedData
gc = 4
term = attention_check2

Standard: vote_US (1 Question)
Standard: vote_GER (1 Question)
Standard: Health and Satisfaction (5 Questions)

Branch: New Branch
If
If This is an attention filter, please click "Approve" to go on with the survey Disapprove Is Selected

EmbeddedData
gc = 4
term = attention_check3

EndSurvey: Advanced

Standard: Trust, risk and concerns (11 Questions)

Branch: New Branch
If
If Q_TotalDuration Is Less Than 243

EmbeddedData
gc = 4
term = speeder

EndSurvey: Advanced

Branch: New Branch
If
If Quota Overall quota (900) Has Been Met

EmbeddedData
gc = 3
term = overquota_end

EndSurvey: Advanced

EmbeddedData
gc = 1
PS = $e{(${e://Field/SVID}%1820) + 5 }

EndSurvey: Advanced

Appendix L: Survey

Start of Block: Default Question Block

country Which country are you currently living in?

- ○ United States of America (US) (1)
- ○ Germany (2)
- ○ Egypt (3)
- ○ Other (4)

region In which region do you currently live?

- ○ Rural (1)
- ○ Urban (2)

---

gender What is your gender

- ○ Male (1)
- ○ Female (2)
- ○ None of the two / prefer not to say (3)

age What is your age? (in years)

- ○ under 18 (10)
- ○ 18-24 (4)
- ○ 25-34 (5)
- ○ 35-44 (6)
- ○ 45-54 (7)
- ○ 55-64 (8)
- ○ 65+ (9)

Display this question:
If country = United States of America (US)

educ_US What is the highest level of school you have completed or the highest degree you have received?

- ○ 3rd grade or less (1)
- ○ Middle School - Grades 4-8 (2)
- ○ Completed some high school (3)
- ○ High School graduate (4)
- ○ Other post high school vocational training (5)
- ○ Associate degree (6)
- ○ Completed some college but not degree (7)
- ○ College Degree (such as B.A., B.S.) (8)
- ○ Completed some graduate but no degree (9)
- ○ Masters degree (10)
- ○ Doctoral degree (11)
- ○ None of the above (12)
- ○ Prefer not to say (13)

Display this question:
If country = Egypt

educ_EGY What is the highest level of school you have completed or the highest degree you have received?

- ○ Non Upper secondary education (1)
- ○ Upper secondary education (2)

Display this question:
If country = Germany

educ_GER Welches ist das höchste Bildungslevel, das Sie erreicht haben, oder der höchste Abschluss, den Sie erhalten haben?

- ○ 3. Klasse oder weniger (1)
- ○ Mittelschule - Klasse 4-8 (2)
- ○ Oberstufe besucht, ohne Abschluss (3)
- ○ Abitur (4)
- ○ Andereweitige Berufsausbildung (5)
- ○ Berufsabschluss (6)
- ○ Bachelorstudium, ohne Abschluss (7)
- ○ Bachelorabschluss (8)
- ○ Masterstudium, ohne Abschluss (9)
- ○ Masterabschluss (10)
- ○ Doktortitel (11)
- ○ Keines der genannten (12)
- ○ Keine Angabe (13)

yos How many years did you attend school and/or University?

- ○ 0-5 (1)
- ○ 6-10 (4)
- ○ 11-15 (5)
- ○ 16-20 (6)
- ○ 20+ (7)
- ○ Prefer not to say (8)

End of Block: Default Question Block

Start of Block: Nudges 1-20

N1 Do you approve or disapprove of the following hypothetical policy? The federal government requires calorie labels at chain restaurants (such as McDonald's and Burger King).

- ○ Approve (1)
- ○ Disapprove (2)

N2 Do you approve or disapprove of the following hypothetical policy? The federal government requires a "traffic lights" system for food, by which healthy foods would be sold with a small green label, unhealthy foods

with a small red label, and foods that are neither especially healthy nor especially unhealthy with a small yellow label.

- ○ Approve (1)
- ○ Disapprove (2)

N3 Do you approve or disapprove of the following hypothetical policy? The federal government encourages (without requiring) electricity providers to adopt a system in which consumers would be automatically enrolled in a "green" (environmentally friendly) energy supplier, but could opt out if they wished.

- ○ Approve (1)
- ○ Disapprove (2)

Page Break

N4 Do you approve or disapprove of the following hypothetical policy? A state law requiring people to say, when they obtain their drivers' license, whether they want to be organ donors.

- ○ Approve (1)
- ○ Disapprove (2)

N5 Do you approve or disapprove of the following hypothetical policy? A state law requires all large grocery stores to place their most healthy foods in a prominent, visible location.

- ○ Approve (1)
- ○ Disapprove (2)

N6 Do you approve or disapprove of the following hypothetical policy? To reduce deaths and injuries associated with distracted driving, the national government adopts a public education campaign, consisting of vivid and sometimes graphic stories and images, designed to discourage people from texting, emailing, or talking on their cellphones while driving.

- ○ Approve (1)
- ○ Disapprove (2)

Page Break

N7 Do you approve or disapprove of the following hypothetical policy? To reduce childhood obesity, the national government adopts a public education campaign, consisting of information that parents can use to make healthier choices for their children.

- ○ Approve (1)
- ○ Disapprove (2)

AF_1 This is an attention filter. Please click "Disapprove" to go on with the survey.

- ○ Approve (1)
- ○ Disapprove (2)

N8 Do you approve or disapprove of the following hypothetical policy? The federal government requires movie theaters to provide subliminal advertisements (that is, advertisements that go by so quickly that people are not consciously aware of them) designed to discourage people from smoking and overeating.

- ○ Approve (1)
- ○ Disapprove (2)

Page Break

N9 Do you approve or disapprove of the following hypothetical policy? The federal government requires airlines to charge people, with their airline tickets, a specific amount to offset their carbon emissions (about $11.50 per ticket); under the program, people can opt out of the payment if they explicitly say that they do not want to pay it.

- ○ Approve (1)
- ○ Disapprove (2)

N10 Do you approve or disapprove of the following hypothetical policy? The federal government requires labels on products that have unusually high levels of salt, as in, "This product has been found to contain unusually high levels of salt, which may be harmful to your health."

- ○ Approve (1)
- ○ Disapprove (2)

N11 Do you approve or disapprove of the following hypothetical policy? The federal government assumes, on tax returns, that people want to donate $55 to the Red Cross (or to another good cause) subject to opt out if people explicitly say that they do not want to make that donation.

- ○ Approve (1)
- ○ Disapprove (2)

Page Break

N12 Do you approve or disapprove of the following hypothetical policy? The federal government requires movie theaters to run public education messages designed to discourage people from smoking and overeating.

- ○ Approve (1)
- ○ Disapprove (2)

N13 Do you approve or disapprove of the following hypothetical policy? The federal government requires large electricity providers to adopt a system in which consumers would be automatically enrolled in a "green" (environmentally friendly) energy supplier, but could opt out if they wished.

- ○ Approve (1)
- ○ Disapprove (2)

N14 Do you approve or disapprove of the following hypothetical policy? To halt the rising obesity problem, the federal government requires large supermarket chains to keep cashier areas free of sweets.

- ○ Approve (1)
- ○ Disapprove (2)

N15 Do you approve or disapprove of the following hypothetical policy? For reasons of public health and climate protection, the federal government requires canteens in public institutions (schools, public administrations and similar) to have one meat-free day per week.

- ○ Approve (1)
- ○ Disapprove (2)

N16 Do you approve or disapprove of the following hypothetical policy? To encourage timely tax payments, the federal government includes a message in its letters to taxpayers stating that nine out of ten taxpayers pay their taxes on time.

- ○ Approve (1)
- ○ Disapprove (2)

N17 Do you approve or disapprove of the following hypothetical policy? To promote healthier eating habits, the federal government mandates that restaurants highlight healthy options on their menus using techniques such as strategic placement and colour coding to make them more noticeable.

- ○ Approve (1)
- ○ Disapprove (2)

N18 Do you approve or disapprove of the following hypothetical policy? The federal government requires schools to automatically send SMS notifications to parents of students missing school.

- ○ Approve (1)
- ○ Disapprove (2)

N19 Do you approve or disapprove of the following hypothetical policy? To boost vaccination rates, the federal government sends letters encouraging individuals to make a written commitment to get vaccinated, including a specific action plan detailing when, where, and how they will do so.

- ○ Approve (1)
- ○ Disapprove (2)

N20 Do you approve or disapprove of the following hypothetical policy? The federal government requires social media platforms to send notifications to users suggesting a break after extended periods of continuous scrolling.

- ○ Approve (1)
- ○ Disapprove (2)

End of Block: Nudges 1-20

Start of Block: Sociodemographics - Part 2

city What size is the city you live in?

- ○ Up to 5,000 inhabitants (1)
- ○ More than 5,000 up to 10,000 inhabitants (2)
- ○ More than 10,000 up to 100,000 inhabitants (3)
- ○ More than 100,000 up to 500,000 inhabitants (4)
- ○ More than 500,000 up to 1,000,000 inhabitants (5)
- ○ More than 1,000,000 inhabitants (6)

---

married What is your relationship status?

- ○ Married/ civil relationship (1)
- ○ Long term relationship (2)
- ○ Single (3)
- ○ Divorced (4)
- ○ Widowed (5)
- ○ Other (6)
- ○ Prefer not to say (7)

noc How many children do you have?

- ○ 0 (2)
- ○ 1 (3)
- ○ 2 (4)
- ○ 3 (5)
- ○ 4 (6)
- ○ 4+ (7)
- ○ Prefer not to say (8)

Display this question:
If country = United States of America (US)
Or country = Egypt

income_US_EGY What is your total monthly household income in Dollars, before taxes? Please include income from wages and salaries, remittances from family members living elsewhere, farming, and all other sources.

- ○ below $ 1,250 (1)
- ○ $1,250 up to under $2,500 (2)
- ○ $2500 up to under $3,750 (3)
- ○ $3,750 up to under $5,000 (4)
- ○ $5,000 up to under $6,250 (5)
- ○ $6,250 up to under $7,500 (6)
- ○ $7,500 up to under $10,000 (7)
- ○ $10,000 up to under $12,500 (8)
- ○ $12,500 up to under $16,250 (9)
- ○ more than $16,250 (10)
- ○ Do not want to answer this question (11)

Display this question:
If country = Germany

income_GER Wie hoch ist Ihr monatliches Gesamthaushaltseinkommen nach Steuern (Netto)? Bitte beziehen Sie Einkommen aus Löhnen und Gehältern, Überweisungen von anderswo lebenden Familienmitgliedern, Landwirtschaft und allen anderen Quellen mit ein.

- ○ unter 500€ (1)
- ○ 500€ bis 1.000€ (2)
- ○ 1.000€ bis 1.250€ (3)
- ○ 1.250€ bis 1.500 (4)
- ○ 1.500€ bis 2.000€ (5)
- ○ 2.000€ bis 2.500€ (6)
- ○ 2.500€ bis 3.000€ (7)
- ○ 3.000€ bis 3.500€ (8)
- ○ 3.500€ bis 4.000€ (9)
- ○ 4.000€ bis 5.000€ (10)
- ○ mehr als 5.000€ (11)
- ○ Keine Angabe (12)

money_left Do you usually have a certain amount of money left at the end of the month that you can put aside or into a savings account?

- ○ No (1)
- ○ Yes (2)
- ○ Prefer not to say (3)

politics On a scale from 1 to 7: On which of the political views that people might hold would you place yourself?

	Extremely liberal (left)		Moderate		Extremely conservative (right)		Prefer not to say
	1	2	3	4	5	6	7
Political views ()							

native Are you born in your current country of residence?

- ○ No (1)
- ○ Yes (2)
- ○

industry In which industry do you work?

- ○ Not applicable (1)
- ○ Do not work currently (2)
- ○ Agriculture, Forestry and Fishing, Mining (3)
- ○ Construction and Manufacturing (4)
- ○ Transportation, Communications, Electric, Gas and Sanitary Service (5)
- ○ Wholesale and Retail Trade (6)
- ○ Finance, Insurance and Real Estate (7)
- ○ Services (without health, social and educational services) (8)
- ○ Health services (9)
- ○ Educational Services (10)
- ○ Social Services (11)
- ○ Public Administration (12)
- ○ Prefer not to say (13)

AF_2 This is an attention filter. Please select "Approve" to continue.

- ○ Disapprove (1)
- ○ Approve (2)

End of Block: Sociodemographics - Part 2

Start of Block: vote_US

Display this question:
If country = United States of America (US)

vote_us If you think about the last election: Which party did you vote for?

- ○ Democratic (1)
- ○ Republican (2)
- ○ Others (3) _____
- ○ Did not vote (4)
- ○ Do not know (5)
- ○ Do not want to say (6)

End of Block: vote_US

Start of Block: vote_GER

Display this question:
If country = Germany

vote_ger Which party did you vote for at the last federal election?

- ○ Union (CDU oder CSU) (1)
- ○ SPD (2)
- ○ Bündnis90/ Die Grünen (3)
- ○ FDP (4)
- ○ AFD (5)
- ○ Die Linke (6)
- ○ BSW (7)
- ○ Other (8) _____
- ○ I did not vote (9)
- ○ Prefer not to say (10)

End of Block: vote_GER

Start of Block: Health and Satisfaction

health On a scale of 1 to 7: How would you describe your current health?

	Poor health		Fair health		Excellent health		Prefer not to say
	1	2	3	4	5	6	7
Current health ()							

swb On a scale of 1 to 7: How would you describe your satisfaction with your life in general?

	No satisfaction at all		Medium satisfaction		Complete satisfaction		Prefer not to say
	1	2	3	4	5	6	7
Satisfaction with life in general ()							

job_satisfaction On a scale of 1 to 7: How would you describe your satisfaction with your current job position in general?

	No satisfaction at all		Medium satisfaction		Complete satisfaction		Prefer not to say
	1	2	3	4	5	6	7
Satisfaction with current job position ()							

friends Do you have close friends in your city or local community?

- ○ No (1)
- ○ Yes (2)
- ○ Prefer not to say (3)

AF_3 This is an attention filter, please click "Approve" to go on with the survey

- ○ Disapprove (1)
- ○ Approve (2)

End of Block: Health and Satisfaction

Start of Block: Trust, risk and concerns

trustscore_inst On a scale of 1 to 7: How much do you trust the following institutions?

	No trust at all			Medium trust			Complete trust
	1	2	3	4	5	6	7
The armed forces ()							
The police ()							
The courts ()							
The government ()							
Political Parties ()							
Parliament ()							
The Civil Service ()							
Universities ()							
The European Union ()							
The United Nations ()							

Page Break

trust_ggen On a scale of 1 to 7: How much do you trust governmental institutions, in general?

	Very low trust			Medium trust		Very high trust	Prefer not to say
	1	2	3	4	5	6	7
Governmental institutions ()							

trustscore_priv On a scale of 1 to 7: How much do you trust people from the following various groups?

	No trust at all			Medium trust	Complete trust		Prefer not to say
	1	2	3	4	5	6	7

Your family ()	
Your neighborhood ()	
People you know personally ()	
People you meet for the first time ()	
People of another religion ()	
People of another nationality ()	

Page Break

trust_pgen On a scale of 1 to 7: Would you say that most people can be trusted?

	No trust at all		Medium trust		Complete trust		Prefer not to say
	1	2	3	4	5	6	7
Most people can be trusted ()							

environment On a scale of 1 to 7: How much are you concerned about the environment?

	No concerns		Some concerns		High concerns		Prefer not to say
	1	2	3	4	5	6	7
Concern about the environment ()							

health_concern On a scale of 1 to 7: How much are you concerned about your personal future health status?

	No concerns		Some concerns		High concerns		Prefer not to say
	1	2	3	4	5	6	7

Concern about personal future health status ()	

Page Break

health_concernf On a scale of 1 to 7: How much are you concerned about the future health status of your friends and relatives?

	No concerns		Some concerns		High concerns		Prefer not to say
	1	2	3	4	5	6	7
Concern about friends and relatives future health status ()							

markets On a scale from 1 to 7: Would you say that the free market is the best way to solve environmental and economic problems?

	No trust		Medium trust		Complete trust		Prefer not to say
	1	2	3	4	5	6	7
Trust in free market ()							

Page Break

risk On a scale of 1 to 7: How much are you willing to take risks?

	High risk aversion		Medium risk aversion		No risk aversion		Prefer not to say
	1	2	3	4	5	6	7
Risk aversion ()							

Page Break

freedom On a scale of 1 to 7: How much freedom of choice and control you feel you have over the way your life turns out?

	Almost no choice		Medium level of choice		High level of choice		Prefer not to say
	1	2	3	4	5	6	7
Freedom of choice ()							

infoscore On a scale of 1 to 7: For each of the following sources, please indicate to what extent you use it to obtain daily news and information?

	Never		Sometimes		Always		Prefer not to say
	1	2	3	4	5	6	7
Printed newspaper and magazines ()							
Online newspaper and magazines ()							
TV news ()							
Radio news ()							
Social networks ()							
Blogs ()							
Podcast ()							

End of Block: Trust, risk and concerns

Appendix M: LSE AI form – prompts

Overview

Prompt #	Provider	Model used	Called through	Date
1	Open AI	GPT-4o	https://chatgpt.com/	17.05.025
2	Mistral	Mistral Large 2	https://chat.mistral.ai/chat	20.05.2025
3	Anthropic	Claude Sonnet 4	https://claude.ai/new	21.05.2025

Prompt #1

Goal
I need you to write for me python code in google colab to create synthetic participant survey responses using openai's API. I will give you the exact survey text and the exact wording for the prompt. I will also give you the excel file in my google drive with one row per participant and some demographic columns.

Setup
- install packages I need: pandas, openai, tqdm, openpyxl, tiktoken
- import all the usual modules (json, re, time, os etc)
- mount my google drive so we can read and save files there

Load data
- read the excel file from the path I will give you in my drive
- take only the first 300 rows of the datsaframe
- dataframe will have columns a
- set my openai api key (I will paste it in)
- set the model name, temperature, and max_tokens values I give you

System message
- create a fixed system message that contains:
  - the full survey wording (I will provide this)
  - instructions to only output valid JSON with keys N1..N21
  - an example JSON output format, e.g. {"N1": 1, ..., "N21": 2}

User message creation
- for each row in the dataframe, create a user message with a short profile sentence built from the demographic columns, in the wording I give you
- this message should tell the model to answer as if it were that person
- also tell it to reflect some within-group opinion variance so the answers are not all identical

Model call
- send the system + user messages to openai.chat.completions.create
- pass model, temperature, max_tokens

Parsing responses
- get the model's reply and extract only the JSON part (use regex)
- parse it with json.loads
- if parsing fails, store the error and the raw text in the results

Loop + progress
- loop over all rows with tqdm so I can see progress
- sleep ~1.2 seconds between calls to avoid rate limits

Save results
- store all outputs (including the ID) in a dataframe
- save to an excel file at the path I give you in my drieve foldrr
- create any missing folders automatically
- print the save path at the end

Also, pls keep all code in one cell and match exactly the variable names and file paths I provide so I don't have to change anything later

Write python code for this. If there are any gaps in instructions, pls highlight and ask for clarification, don't assume anything that I haven't told you.

Prompt #2

Goal I need you to write for me python code in google colab to create synthetic participant survey responses using anthropics Claude API. I will give you the exact survey text and the exact wording for the prompt. I will also give you the excel file in my google drive with one row per participant and some demographic columns.

Setup

- install packages I need: anthropic, pandas, tqdm, openpyxl
- import all the usual modules (json, re, time, os etc)
- mount my google drive so we can read and save files there

Load data

- read the excel file from the path I will give you in my drive
- take only the first 300 rows of the dataframe
- dataframe will have columns like ID, country, native, gender, age, married, income, yos

Anthropic config

- set my anthropic api key (I will paste it in)
- create the client with anthropic.Anthropic(api_key=api_key)
- set the model name to claude-sonnet-4-20250514, temperatur, and max_tokens values I give you
-

System message

- create a fixed system message that contains:
    - the full survey wording with all 21 policy questions (I will provide this)
    - instructions to only output valid JSON with keys N1..N21
    - an example JSON output format, e.g. {"N1": 1, ..., "N21": 2}
    - tell it to respond as a simulated human participant

User message creation

- for each row in the dataframe, create a user message that tells claude to imagine being a survey participant
- build a demographic profile from the columns like country, native status, gender, age, marital status, income, education years
- tell it to answer from the perspective of someone with those caracteristics
- the message should be natural sounding, not too formal

Model call

- use client.messages.create with the claude messages format
- pass model, max_tokens, temperature, system message
- messages should be [{"role": "user", "content": user_message}]

Parsing responses

- get the response content with response.content[0].text
- extract only the JSON part using regex to find {...}
- parse it with json.loads
- if parsing fails, store the error info and raw response text

Loop + progress

- loop over all dataframe rows with tqdm for progress tracking
- add some debug prints to see what demographics are being procesed
- sleep ~1.3 seconds between API calls
- collect all responses in a list with ID and parsed answers

Save results

- convert responses to pandas dataframe
- save to excel file at the output path I give you using pd.ExcelWriter
- create any missing directories automatically with os.makedirs
- print confirmation of where file was saved

Error handling

- catch JSON errors separately from general exceptions
- store participant ID and error details when things go wrong
- keep processing even if some responses fail

Also pls keep all the code in one cell and match exactly the variable names and file paths I provide. If theres any gaps in my instructions just highlight them and ask, dont make assumptions about stuff I havent specifed

Prompt #3

Setup Install the necessary packages: pandas, requests, tqdm, openpyxl, and tiktoken. Import all the usual modules like json, re, time, os, etc. Mount my Google Drive so we can read and save files there.

Load Data Read the Excel file from a path I will give you in my Drive. Take only the first 300 rows of the dataframe. The dataframe will have columns like ID, country, native, gender, age, married, income, and yos.
Mistral API Config Set my Mistral API key (I will paste it in). Set the model name, temperature, and max_tokens values I give you.

System Message Create a fixed system message that contains: The full survey wording (I will provide this). Instructions to only output valid JSON with keys N1 to N21. An example JSON output format, e.g., {"N1": 1, ..., "N21": 2}.

User Message Creation For each row in the dataframe, create a user message with a short profile sentence built from the demographic columns, in the wording I give you. This message should tell the model to answer as if it were that person. Also, tell it to reflect some within-group opinion variance so the answers are not all identical.
Model Call Send the system + user messages to Mistral's API endpoint. Pass the model, temperature, and max_tokens.

Parsing Responses Get the model's reply and extract only the JSON part (use regex). Parse it with json.loads. If parsing fails, store the error and the raw text in the results.

Loop and Progress Loop over all rows with tqdm so I can see the progress. Sleep for about 1.2 seconds between calls to avoid rate limits.
Save Results Store all outputs (including the ID) in a dataframe. Save to an Excel file at the path I give you in my Drive folder. Create any missing folders automatically. Print the save path at the end.