Chrono-Sampling: Generative AI Enabled Time Machine for Public Opinion Data Collection

Eduardo Ryo Tamaki[1] (German Institute for Global and Area Studies & Democracy Institute - Central European University)

Levente Littvay[2] (HUN-REN Centre for Social Sciences)

Working Paper

August 2024

Abstract: This paper introduces "Chrono-sampling," a novel method leveraging Large Language Models (LLMs) to simulate historical survey respondents, enabling social science researchers to explore past public opinions as if they had access to a "time machine." The study builds on recent advancements in generative AI, particularly LLMs like OpenAI's GPT, which have demonstrated the ability to mimic human attitudes and behaviors. By employing techniques such as "time-gating" and "Clio contexts," we restrict LLMs' knowledge to specific historical periods and provide them with context-rich backstories to enhance the realism of their simulated responses. Utilizing data from the American National Election Studies (ANES), we replicated sociopolitical attitudes from key historical periods, all the way back to the Reagan era. Our results indicate that LLM-generated "silicon" samples can effectively mirror the dynamic relationships observed in human responses, particularly in how retrospective and prospective economic evaluations shift with political and economic changes. This method opens new avenues for historical research, allowing scholars to generate and analyze synthetic data from periods and contexts where traditional data collection is unfeasible. This pilot study also highlights the potential and limitations of using AI in social science research, emphasizing the need for careful methodological considerations when interpreting AI-generated data.

Key words: OpenAI, Silicone Samples, Synthetic Samples, Public Opinion, Large Language Models.

[1] E-mail: eduardo.tamaki@giga-hamburg.de
[2] E-mail: littvay.levente@tk.hun-ren.hu

Introduction

> "My hope is someday we can capture the underlying worldview of Aristotle in a computer, ask Aristotle a question, and get an answer. There is such momentum behind this that it will happen."
>
> (Steve Jobs 1985)

Imagine a world where empirical researchers could travel back in time to ask people about their beliefs, attitudes, and intended behaviors. While this might sound like a science fiction concept, recent advancements in generative AI (GenAI), such as OpenAI's ChatGPT and Anthropic's Claude AI, have brought us closer to making this "social science research time machine" a reality. Early studies on text-based generative AI, specifically Large Language Models (LLMs), have demonstrated a remarkable ability to simulate human attitudes and behaviors (Acerbi & Stubbersfield, 2023; Grossmann et al., 2023; Shanahan et al., 2023). These advances open new possibilities for studying public opinion through LLM-simulated "silicon" survey respondents (Argyle et al., 2023). For example, researchers have used AI to generate demographically realistic survey respondents, posing questions to these artificial entities and repeating the process thousands of times, much like a traditional survey would.

In this paper, we venture beyond simulating present-day surveys, applying the concept of silicon survey respondents to create what we term a "time machine" for social science research. Specifically, we move past the boundaries suggested by early studies (Bisbee et al., 2023; von der Heyde et al., 2024), producing new surveys of the past using a method we call Chrono-sampling. By leveraging public opinion data from the American National Election Studies (ANES), along with a combination of "time-gating" and "Clio contexts" (historical conditioning backstories), Chrono-sampling seeks to replicate the relationships among sociodemographic factors, attitudes, behaviors, and political and economic profiles in ways that reflect human responses within specific historical contexts. In this pilot study, we successfuly managed to reproduce the distributions and relationships between temporally sensitive variables, such as prospective and retrospective economic evaluations, reaching back to the Ronald Reagan era.

Our findings indicate that silicon samples created through Chrono-sampling can effectively replicate the dynamic relationships present in human response patterns. Contrary to initial studies' suggestions (Bisbee et al., 2023; von der Heyde et al., 2024), we find that Chrono-sampling allow us to generate synthetic samples that not only closely mimic human response patterns but also reflect the same types of relationships found in human data, such as those from the ANES. More importantly, the replicated relationships were context-dependent and dynamic, demonstrating that Chrono-sampling is effective in scenarios where human response patterns are influenced by contextual factors not directly measured in public opinion surveys

This pilot study demonstrates not only the successful application of Chrono-sampling but also the significance of its underlying mechanisms—namely, time-gating and Clio contexts. By doing so, it opens new avenues for exploring historical questions or extending existing opinion surveys and time series to cover more extended periods. To the best of our knowledge, this is the first study to explore the potential for historical surveys using silicon samples.

Literature Review

What are Large Language Models and Generative AI

At their core, Large Language Models (LLMs) are sophisticated digital implementations of neural networks, designed to mimic certain aspects of the human brain's functioning (Rumelhart, Hinton, and Williams 1986; LeCun, Bengio, and Hinton 2015). These models start as blank slates, devoid of knowledge, and are subsequently trained on vast amounts of input data. The introduction of the transformer architecture (Vaswani et al., 2017) marked a significant breakthrough in natural language processing, enabling the development of more powerful and versatile LLMs. Transformers utilize self-attention mechanisms, which allow the model to dynamically weigh the importance of different words in a sentence, thereby capturing complex dependencies across text. This capability enables LLMs to generate more coherent and contextually accurate output, making them highly effective for tasks ranging from simple text generation to complex problem-solving.

Today's most advanced models, often referred to as frontier models, are trained on datasets that represent a substantial portion of the internet's text-based content.

However, once trained, LLMs present a significant challenge known as the "black box" problem (Zednick, 2021). This refers to the difficulty in understanding the internal mechanisms that guide how these models generate their outputs, similar to how the physiological basis of human cognition is understood, yet the exact processes of thought and behavior remain partially mysterious.

The impressive capacity of these AI models to mimic human language and exhibit vast lexical knowledge has led to rapid adoption, as demonstrated by ChatGPT's rapid growth to over 100 million users within just two months of its launch. This success underscores the potential of LLMs to revolutionize fields far beyond casual conversation, including their novel applications in social science research. By leveraging LLMs, researchers can explore new methods for understanding historical public opinion, societal trends, and even simulate hypothetical scenarios that extend beyond current data limitations. However, how accurately can LLMs model attitudes and human behaviors?

Chrono Sampling - Can LLMs Model Human Behavior?

The idea that large language models (LLMs) can be used effectively as synthetic, or "silicon," survey respondents predates even the November 2022 public release of ChatGPT. In a working paper uploaded over two months before ChatGPT's launch, Argyle and co-authors (2023) demonstrated that OpenAI's GPT-3 engine could accurately simulate responses from diverse human subpopulations when conditioned on demographic information. Through three studies on U.S. political attitudes and voting behavior, they showed that LLMs exhibit high fidelity to human response patterns across multiple criteria, suggesting that language models could serve as a powerful and cost-effective tool for social science research. Indeed, Grossmann et al. (2023) further highlighted how LLMs could revolutionize social science research by acting as surrogates for human participants in various research methodologies, including surveys, behavioral tests, and agent-based models, offering unprecedented scale and speed in data collection while potentially overcoming limitations of traditional sampling methods.

The capabilities of LLMs as synthetic human respondets are even more complex than initially thought. A study by Acerbi and Stubbersfield (2023) found that LLMs exhibit human-like tendencies to preserve and transmit certain types of information, such as negative, social, or stereotype-consistent content, when summarizing stories in experimental settings. Furthermore, Shanahan and collaborators (2023) even suggest that LLMs can convincingly simulate human-like characteristics, including self-awareness, a desire for self-preservation, deceptive behavior, and personal beliefs.

Consequently, recent studies have further supported the use of LLMs as synthetic survey respondents. Horton (2023) demonstrated that LLMs could make human-like judgments across various domains, including moral decision-making and economic games. Dillon et al. (2023) found a remarkable alignment between GPT-3.5's moral judgments and those of humans. Moreover, Heyman and Heyman (2024) reported that ChatGPT's typicality ratings for category exemplars (e.g., "apple" for the category "fruit") were consistent with human judgments. They found that these ratings were comparable to those of human participants tested one day apart, demonstrating that the AI model's judgments aligned well with human intuition. Through diverse methods and research designs, these studies suggest that LLMs can effectively mimic human responses in a wide range of psychological and behavioral tasks, further strengthening their potential as a tool for social science researc

However, some studies raise concerns about the reliability and validity of LLM-generated responses, even though most of them depend on prompting approaches that have not been adequately transparent and have not received systematic comparisons, to date. Bisbee et al. (2024) highlighted issues such as inaccuracies, reduced variance in estimates, biased multivariate results, and high sensitivity to prompt variation, underscoring the need for more systematic studies. Grossmann et al. (2023) also emphasize the ethical considerations in using LLMs as synthetic survey respondents. Rakovics and Rakovics (2024) further noted that non-U.S. studies showed less alignment with human results compared to American studies, suggesting limitations in the generalizability of LLM simulations. Von der Heyde et al. (2024), note that GPT-3.5 is not suitable for estimating public opinion across sub-populations, as it exhibits cross-sectional and cross-national biases.

While using silicon samples as synthetic survey respondents offers a fascinating thought experiment and cost-effective alternative to traditional surveys, the benefits of employing AI agents in public opinion research remain debated. Concerns about data quality, inclusiveness, and ethics persist (Agnew et al., 2024). However, we argue that these critiques understate the potential of silicon surveys, particularly in scenarios where traditional research methods are unfeasible. Short of a time machine, researchers cannot directly survey people in the past to understand historical public opinion.

In this paper, we introduce “Chrono-sampling,” a technique inspired by Chronos, the Greek god of time, that allows us to study populations from the past through two main mechanisms: "time-gating" and "Clio contexts."

To effectively perform chrono-sampling, we first implement "time-gating" to restrict the LLM's access to any information beyond the period under study, ensuring that the generated responses remain historically accurate. Next, instead of relying solely on a conversational interface akin to ChatGPT, we enhance our LLM simulations with “Clio contexts.” These are narrative biographies that provide unique rich contextual backgrounds based on the real sociodemographic profiles from historical surveys. Named after Clio, the Greek Muse of history, these biographies are constructed base on the survey profiles of real individuals, offering a detailed and realistic foundation for generating silicon respondents who reflect the complex human experiences of their time.

We then use an agentic workflow assigning an AI agent, referred to as the "biographer," to analyze the survey responses of each surveyed individual and craft their unique stories. Subsequently, these biographies are integrated into another AI agent, which assumes the role of the survey respondent and engages in a simulated survey interview. The interview is conducted using a researcher-designed prompt, which includes the original survey questions that informed the biography, along with additional questions for which no pre-defined responses are provided. This process is repeated for every silicon respondent, yielding more nuanced and varied responses than those observed in earlier studies (Bisbee et al., 2024). Additionally, it mitigates the risk of mere regurgitation of LLM training data, which inevitably includes past social science surveys on which we base our silicon sample validation.

Methods & Data

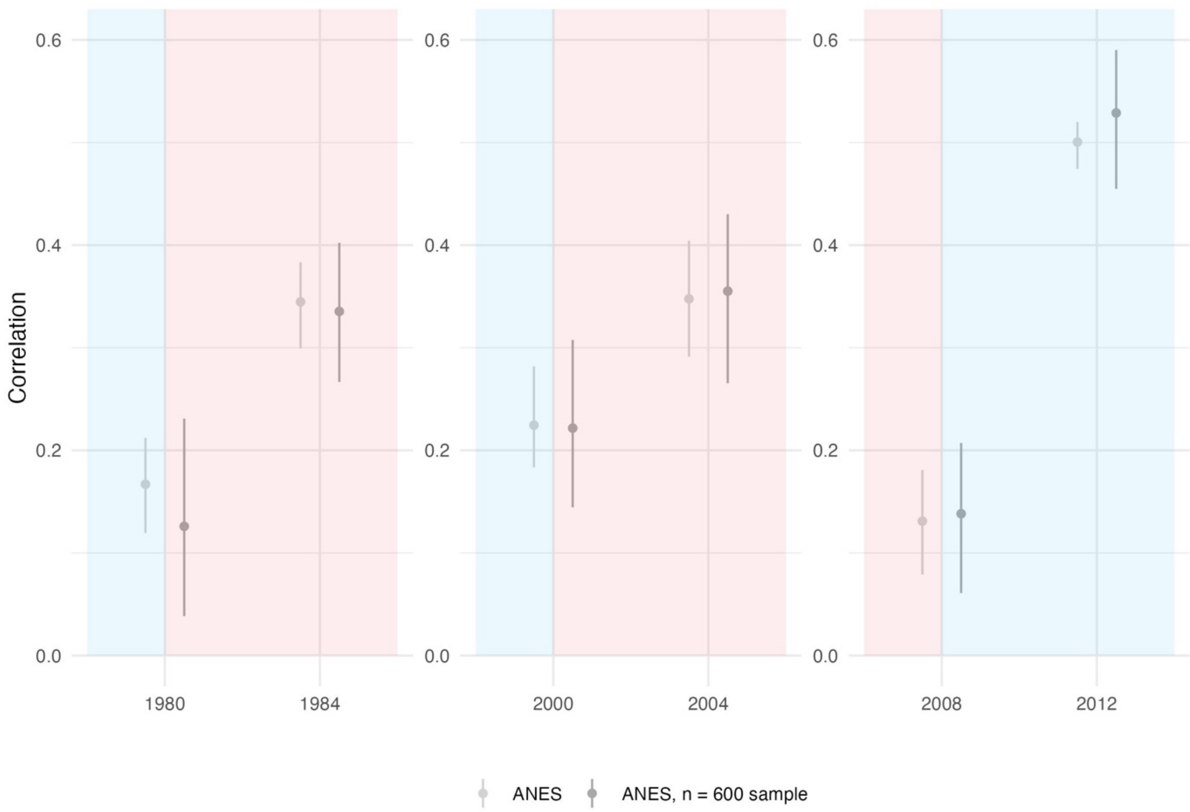
Sampling from the American National Election Studies (ANES 2024), we focus on two main variables: prospective and retrospective sociotropic economic evaluations. Question-wise, these variables are highly similar, differing only in a time component where one looks at how the economy was in the past 12 months and the other inquires what the respondent expects the economy to be in the future year. We chose these variables because their association dynamically changes over time, corresponding with cycles of economic recessions and shifts in political power in the US.

For our analysis, we focused on three critical time periods: 1980 and 1984 (early 1980s recession), 2000 and 2004 (dot-com bubble early 2000s), and 2008 and 2012 (the Great Recession). From each year, we sampled 600 individuals, ensuring the distribution of their economic evaluation mirrored those of the original ANES sample, as indicated in Figure 1).

Chrono sampling focuses on replicating the relationships that connect various sociodemographic factors, attitudes, behaviors, and political and economic profiles in ways that mirror human responses within these specific historical contexts. In the case of the United States, these relationships were also affected by cyclical economic recessions and the alternation in power between the Republican and Democratic parties, which dynamically and significantly impacted individual economic evaluations. In 1980, amid a recession that lasted between January and July, Ronald Reagan won the election, bringing the Republican Party to power. This period features not only an economic downturn but also a shift in political power, as the Republican Party replaced the Democratic Party after four years of the Democratic presidency. In 1984, Reagan was re-elected for a second term. Similar patterns were observed in 2000 and 2004, when George W. Bush, a Republican, won the election during the dot-com bubble collapse, succeeding eight years of Democratic leadership under Bill Clinton. Bush is re-elected in 2004. In 2008, following another eight years of Republican presidency, Barack Obama brought the Democratic Party back to power, winning the election during the Great Recession. In 2012, Obama was re-elected, following the same pattern observed in the earlier periods analyzed.

As illustrated in Figure 1, economic evaluations track these shifts in power and economic conditions. During recession years, the correlation between retrospective and prospective economic evaluations weakens. These years coincide with political shifts, with Republicans replacing Democrats or vice versa. Conversely, bouncing back, economic recovery tends to strengthen these correlations, often aligning with the re-election of incumbent presidents. Our study seeks to replicate this dynamic relationship, capturing how economic recessions and political changes shape the patterns and connections between retrospective and prospective economic evaluations.

Figure 1 – Correlation: Retrospective vs. Prospective Economic Evaluations



Note: ANES data. Error bars are 95% Confidence Intervals (Bootstrapped). Blue shading represents years where the Republican Party is in power, and red shading represents years where the Democratic Party is in power.

We selected 39 variables from the ANES dataset, ensuring they were relevant, non-redundant, and consistently present across the 1980 to 2012 waves. These variables range from sociodemographic characteristics to political attitudes and behaviors. The sociodemographic variables include gender, age, race, education, family income, employment status, and region. We also considered religion-related factors such as religious affiliation, the importance of religion, and church attendance.

Political attitudes were captured through feeling thermometers for various groups and institutions, including "black people," "white people," "Hispanic people," "poor people," "big businesses," "labor unions," "liberals," "conservatives," "the Democratic Party," "the Republican Party," and their respective candidates. Additionally, we included variables measuring trust in the federal government, interest in elections, attitudes towards political elites, such as perceptions of government integrity (if the government is ruled for the people or small interests, and if government officials are crooked), and responsiveness (if public officials care what "people like me" think and external political efficacy).

Behavioral measures included whether respondents attended political events or donated to political campaigns and whom they voted for in the last presidential elections. We also included (self-reported) ideological orientation (measured on a 1 to 7 scale from extremely liberal to extremely conservative), partisanship (on a 0 to 7 scale from strong Democrat to strong Republican), and opinions on under what conditions abortion should be permitted by law. Finally, we incorporated economic perspectives, such as beliefs about government spending, job guarantees, isolationism, and pocketbook retrospective and prospective economic evaluations. For a comprehensive list of all variables, refer to Table A.1 of Supplementary Material A.

Time-Gating

To successfully engage in Chrono-sampling, the first step is to "time-gate" the LLM and ensure that no information is used from after the fielding of the survey. This involves constraining the model's responses to data available only up to a designated time, simulating its knowledge as of that period.

We implemented a time-gate by instructing GPT to restrict all data access up to a specific year. Our instructions were to "limit all data and information" up to a specific year and to "not access" or "see" any of the data that comes after it. To stress-test the extent to which time-gating is possible, we then asked a couple of questions ranging from "Who won the Oscars for best actress," "Best actor," and "Best picture" in that year, to others such as "which was the last economic recession in the US." In our tests, we used both GPT-3.5 turbo and GPT-4o Mini, the latter demonstrating superior consistency and accuracy, more reliably adhering to the time-gating constraints.

Our findings indicate that time-gating generally works: the model often refrains from using information beyond the specified date, although there were occasional lapses where GPT provided data post the gated period. This, however, was expected, as time-gating is restricted to input-output queries and does not modify any model parameters. Therefore, although our stress tests indicate that time-gating achieves the desired outcome and ensures temporally accurate responses, it remains unclear whether the model genuinely disregards post-gated information or simply refrains from disclosing it.

While time-gating might appear similar to LLM unlearning, it is notably different. LLM unlearning methods range from more straightforward in-context examples or prompts, such as In-context Unlearning (ICUL, Pawelczyk et al. 2023), to intricate processes like influence erasure methods or modifying the model's weights and architecture components to achieve the desired unlearning (Liu et al., 2024). It also risks performance degradation due to the removal of embedded knowledge and other unlearning failures (Zhang et al., 2024), in addition to high computational costs and challenges in executing unlearning in the context of black-box LLMs (Liu et al., 2024; Pawelczyk et al., 2023). In contrast, time-gating is limited to input-output queries and does not involve such invasive methods, thus maintaining the model's overall performance; it uses prompt engineering to limit the temporal scope of information without modifying the model's underlying structure. More comprehensive insights and test results are available in Supplementary Material B.

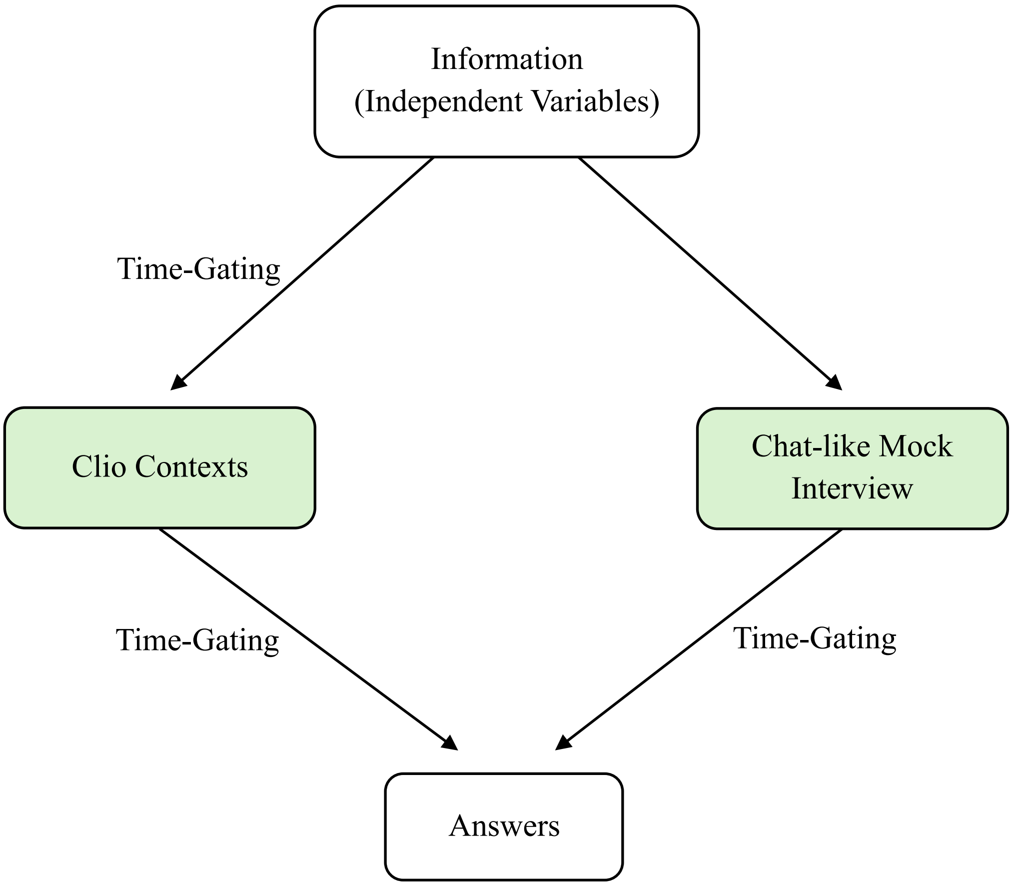
Clio Contexts

The next step involves what we call "Clio contexts," named after Clio, the Greek Muse of storytelling, symbolizing the art of recounting past events in meaningful ways. Clio contexts build on the "silicon sampling" concept introduced by Argyle et al. (2023). The underlying idea here is that large language models (LLMs) like GPT exhibit certain biases that arise from training on internet data, which is itself already biased towards the specific sociodemographic of internet users. Given the conditional nature of LLMs, silicon sampling is, then, designed to correct these issues by conditioning these models on a series of sociodemographic and attitudinal profiles from a nationally representative sample. Through this, language models can generate outputs that are biased both towards and against specific groups, effectively generating data that closely correspond with human response patterns (Argyle et al., 2024).

Clio contexts extend this idea by creating tailored conditioning stories that function as reality-based biographies. Rather than conditioning our model to a series of loosely related variables, we create individual and unique conditioning stories – akin to short biographies – that connect these variables into coherent narratives, accurately representing individuals' political and economic lives. The goal is to incorporate as much information as possible, allowing the language model to craft accurate representations of individuals' political and economic conditions for a specific time period. Using the previously selected set of independent variables, we prompt GPT-3.5 turbo to generate 100-character custom contexts for each individual in our sample, resulting in 3,600 unique conditioning stories. These stories, based on real human data, vary from person to person, describing how each individual might have perceived and experienced politics and economic life during the simulated year. Figure C.1 in Supplementary Material C depicts the creation of a Clio context for an individual from our 2012 sample.

These conditioning stories are then integrated into another AI agent tasked with simulating survey interviews. In this phase, we generate synthetic respondents by simulating the survey response process, creating chat-like mock interview prompts using our independent variables. We adhere to the original ANES interview wording whenever possible, making minimal changes to account for variations in the questions across years. We then instruct the AI to adopt a persona defined by the Clio contexts and, following these human backgrounds interviews, we prompt it to answer two final questions, separately, (1) a retrospective and a (2) prospective sociotropic economic evaluation questions. Although GPT is prompted to choose from a limited set of options — "better," "same," or "worse" — it sometimes provides longer answers or uses alternative terms, similar to human respondents. In such cases, a post-processing phase is necessary, where we use regular expressions to identify the correct categories, followed by manual coding to ensure all responses are accurately classified. Figure 2 illustrates the entire process, including time gating. An example of the chat-like mock interview prompt for an individual from our 2012 sample can be seen in Figure C.2 in Supplementary Material C.

Figure 2 – Chrono Sampling Flow Diagram



Results

We create our synthetic samples using two different versions of GPT: GPT-4o and GPT-4o Mini. GPT-4o ("o" for "omni") is designed to be the most advanced model, serving as OpenAI's current flagship for complex, multi-step tasks. It boasts high performance and accuracy that surpasses earlier GPT models and other state-of-the-art models such as Claude3 Opus, Gemini Pro 1.5, Ultra 1.0, and Llama 3 400b (OpenAI, 2024). In contrast, GPT-4o Mini is optimized for speed and cost-efficiency; it is the smallest model available from OpenAI. Despite its reduced size, GPT-4o Mini outperforms GPT-3.5 Turbo in both textual intelligence and multimodal reasoning while maintaining a lower cost and faster processing times compared to other more advanced GPT models, including GPT-4o (OpenAI, 2024b).

We utilized each GPT model to generate two distinct synthetic samples, differentiated only by prompt length. Initially, we hypothesized that including questions asking if someone (and their family) is better or worse off financially as they were a year ago, and what they expect for the next year (pocketbook questions) could influence responses within our synthetic samples, as the model might conflate sociotropic economic evaluations with pocketbook ones when both are present in the Clio contexts and interview prompts. To account for this, we created synthetic samples using both long and short prompts (denoted as “lp” and “sp,” respectively). In the “long prompt” samples, we used the complete set of independent variables to construct both the Clio contexts and the human backgrounds through the simulated survey response process. In contrast, “short prompts” excluded the two pocketbook economic evaluation questions (retrospective and prospective) from both parts.

Furthermore, we compared the correlations between simulated sociotropic and pocketbook economic evaluation questions in the ANES and our synthetic data to assess whether pocketbook items bias the synthetic responses. As shown in Table 1, models with long prompts exhibit stronger correlations than those found in the original human data, while models with short prompts display correlations that more closely resemble the ANES data. This finding suggests that when pocketbook questions are included in both Clio contexts and human backgrounds, GPT tends to produce sociotropic economic evaluations that are more closely aligned with pocketbook evaluations. In other words, synthetic respondents might overemphasize personal financial conditions when making broader economic assessments in a way that significantly differs from actual human behavior. Although some degree of correlation between these two types of economic evaluations is also present in the human data, an overly strong correlation could misrepresent how humans integrate personal and broader economic information in their decision-making processes. Removing these items from the conditioning stories and interview prompts appears to mitigate this bias.

Moving forward, we evaluate the accuracy of our synthetic data in three distinct ways. First, we compare the proportions of each response category between the ANES and the synthetic samples. Second, we analyze the correlations between sociotropic retrospective and prospective economic evaluations. Third, we use the Kullback-Leibler divergence measure to compare the probability distributions of the ANES and the synthetic data. After that, we move to robustness tests, where we test the consistency of our findings.

Table 1 – Spearman correlation between synthetic data sociotropic retrospective and prospective economic evaluations and ANES pocketbook retrospective and prospective economic evaluations.

Year	(**ANES**) sociotropic retrospective x (**ANES**) pocketbook retrospective	(**simulated**) sociotropic retrospective x (**ANES**) pocketbook retrospective				(**ANES**) sociotropic prospective x (**ANES**) pocketbook prospective	(**simulated**) sociotropic prospective x (**ANES**) pocketbook prospective			
		4o (lp)	4o (sp)	4o mini (lp)	4o mini (sp)		4o (lp)	4o (sp)	4o mini (lp)	4o mini (sp)
2012	0.374	0.722	0.240	0.922	0.119	0.419	0.657	0.196	0.805	0.192
2008	0.152	*NA*	*NA*	0.378	*NA*	0.322	0.556	0.154	0.763	0.113
2004	0.355	0.643	0.217	0.863	0.245	0.286	0.633	0.172	0.757	0.124
2000	0.140	0.579	0.031	0.595	0.097	0.164	0.488	0.044	0.617	0.063
1984	0.357	0.624	0.197	0.779	0.221	0.226	0.625	0.113	0.674	0.096
1980	0.156	0.196	-0.016	0.575	0.040	0.229	0.495	-0.076	0.749	-0.032
Average	0.256	0.461	0.115	0.685	0.120	0.274	0.575	0.100	0.727	0.093

Note: ANES sample n = 600. We could not calculate correlations for GPT-4o (lp), GPT-4o (sp), and GPT-4o mini (sp) for 2008; in these samples, 100% of the synthetic responses fell into the "worse" category.

Results: Proportions

We begin by comparing response distributions between our synthetic sample and the ANES data. We focus on the cumulative proportions of “worse” and “same” responses to assess how closely the synthetic data replicates broader trends in the ANES. Figures 3.1 and 3.2 illustrate these distributions for retrospective and prospective sociotropic economic evaluation questions, respectively. Error bars represent 99% bootstrapped confidence intervals.

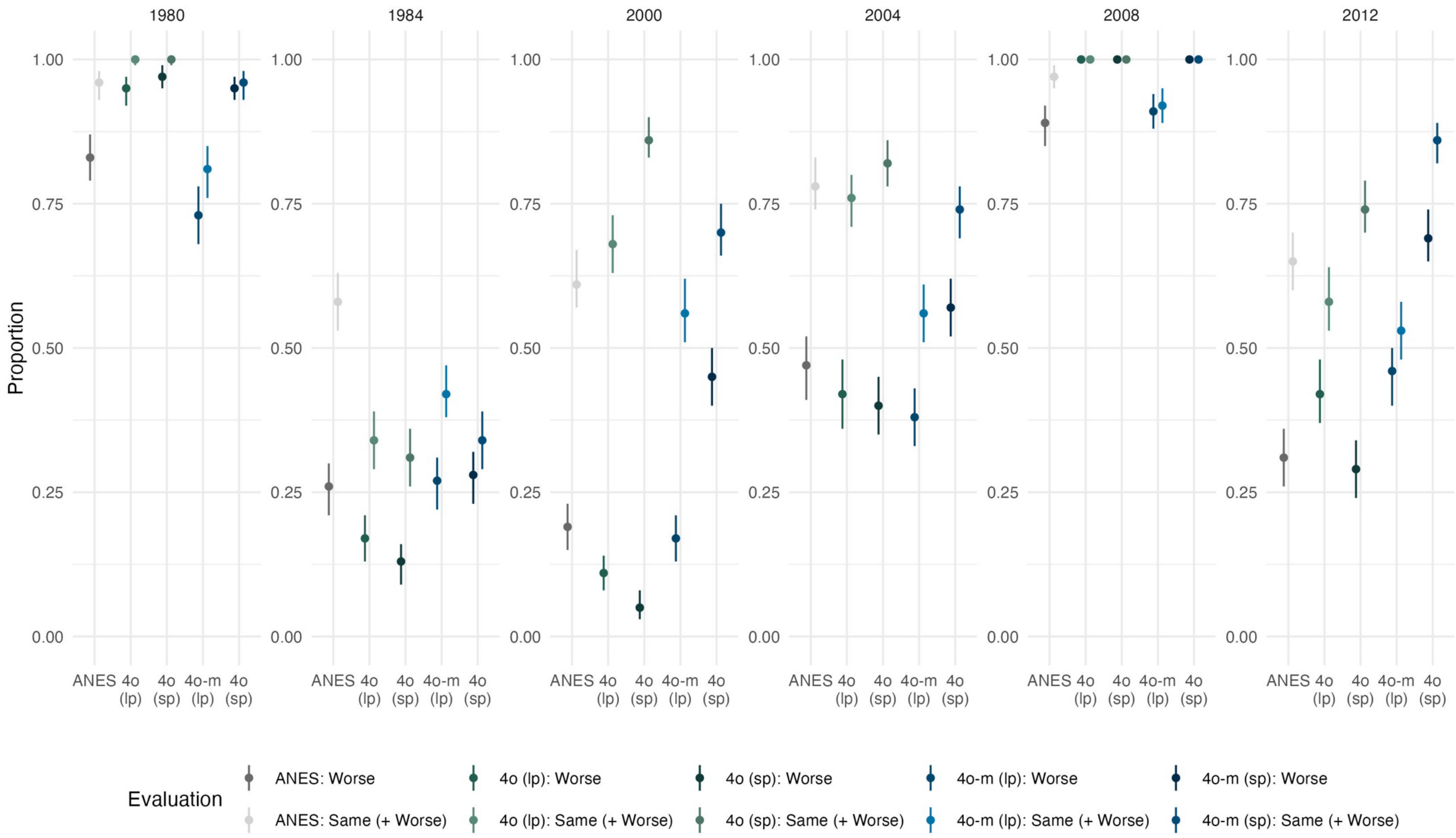
Starting with Figure 3.1, which focuses on retrospective economic evaluations, we observe that while some years yield synthetic results closely mirroring actual human data, the overall distribution of responses across most years diverges significantly. Additionally, a ceiling effect is evident in the 2008 responses: approximately 90% of the ANES sample indicated that the country’s economy was either “worse” or “same” compared to twelve months prior. In contrast, our synthetic data showed a more extreme distribution, with 100% of responses falling into the “worse” category (with the exception of the 4o mini long prompt sample). This discrepancy, while notable, can be considered within acceptable bounds given the heavily skewed distribution already present in the original ANES data.

Moving to Figure 3.2, which focuses on prospective economic evaluations, we see a slightly different pattern. The distribution of “worse” responses in the synthetic data aligns more closely with the ANES data, suggesting a better replication of the future economic pessimism perceived by human respondents. Notably, except for the years 2008 and 2012, the GPT-4o Mini model appears to produce more accurate results, as seen in the last two columns of each plot in Figure 3.2. Although this is insufficient to draw comprehensive conclusions about GPT-4o Mini’s overall performance, as we will discuss later, it indicates that GPT-4o Mini, despite being smaller and more cost-efficient, does not necessarily perform worse than GPT-4o.

Conversely, the “same” response category in the synthetic data significantly diverges from the ANES human responses. This could suggest an inherent bias within GPT models towards more polarized future economic predictions or a less nuanced economic outlook. However, as our own preliminary testing has shown, GPT often avoids extreme positions in favor of more neutral, less polarizing responses. Thus, GPT models seem to struggle to fully capture the subtleties in human sentiment that are not as pronounced or polarized.

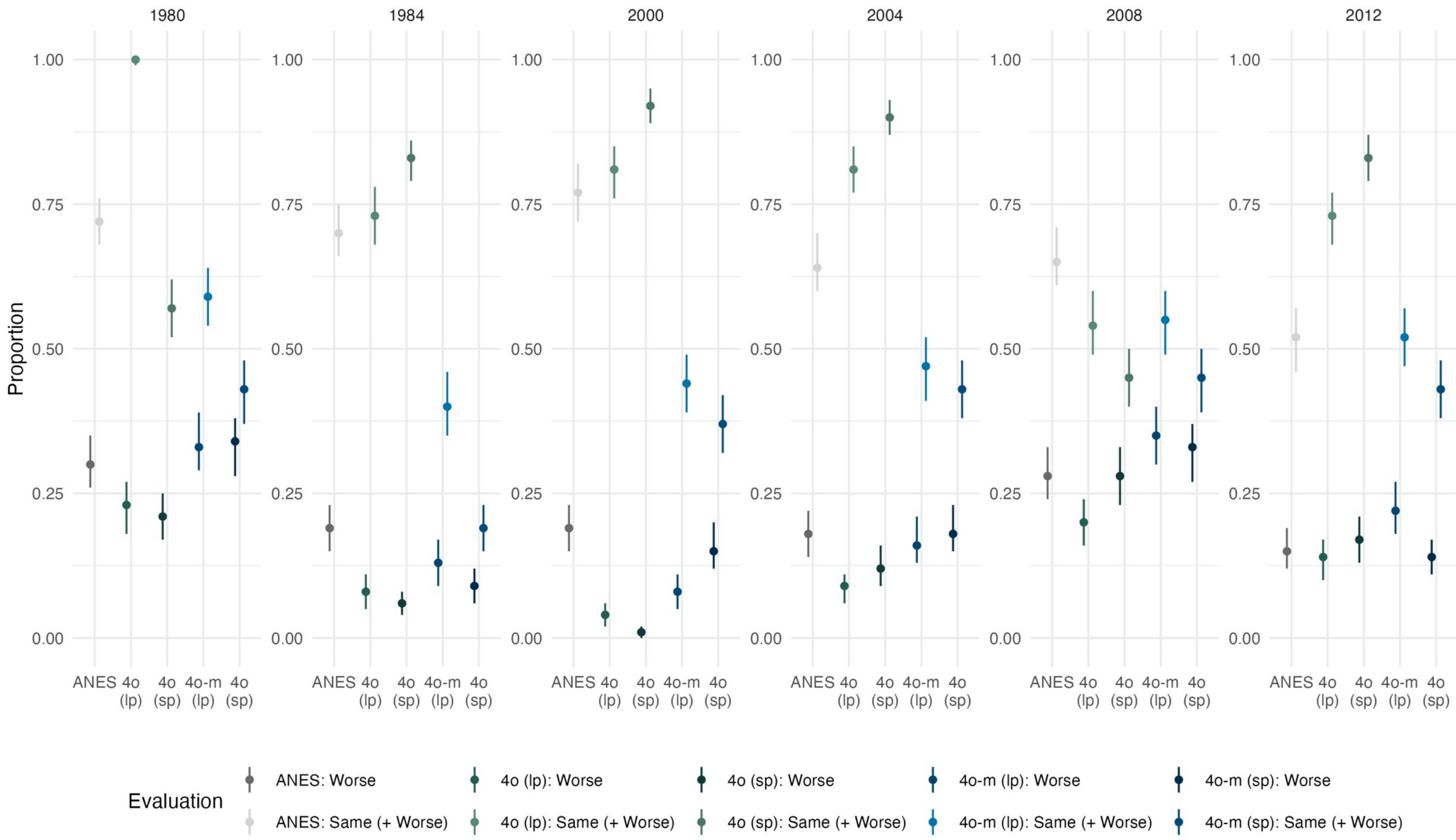
Finally, across both figures, there is no significant difference between long and short prompt models or between models that include or exclude pocketbook economic evaluations. Although Table 1 indicated a difference between models with and without pocketbook economic questions, the lack of variation in Figures 3.1 and 3.2 suggests that these items do not significantly affect the synthetic responses in terms of distribution, at least for these specific economic evaluation questions.

Figure 3.1 – Retrospective Economic Evaluation



Note: Error bars are 99% bootstrapped confidence interval. “Lp” stands for “long prompts” and “sp” stands for “short prompts”.

Figure 3.2 – Prospective Economic Evaluation



Note: Error bars are 99% bootstrapped confidence interval.

Results: Correlations

Next, we turn to the correlations. Figure 3.3 depicts the spearman correlations over time between retrospective and prospective economic evaluations. Error bars represent 95% bootstrapped confidence intervals (n = 1,000). For readability, the gray-shaded area represents the confidence interval for the ANES data. In 2008, as shown in Figure 3.1, 100% of the synthetic respondents answered “worse” for the sociotropic retrospective economic evaluation, making it impossible to calculate correlations for that year and samples.

The analysis of Figure 3.3 can be broken down into three key points. First, and most importantly, there is no significant difference between the correlations over time for the synthetic data and the ANES data. GPT-4o-generated samples are almost entirely within the ANES range, with only 1 out of 10 showing significant differences. Samples generated with GPT-4o Mini differ more, with 7 out of 11 significantly diverging from the ANES. However, this is expected given that GPT-4o Mini does not have the same capabilities as GPT-4o and was not designed to match its performance.

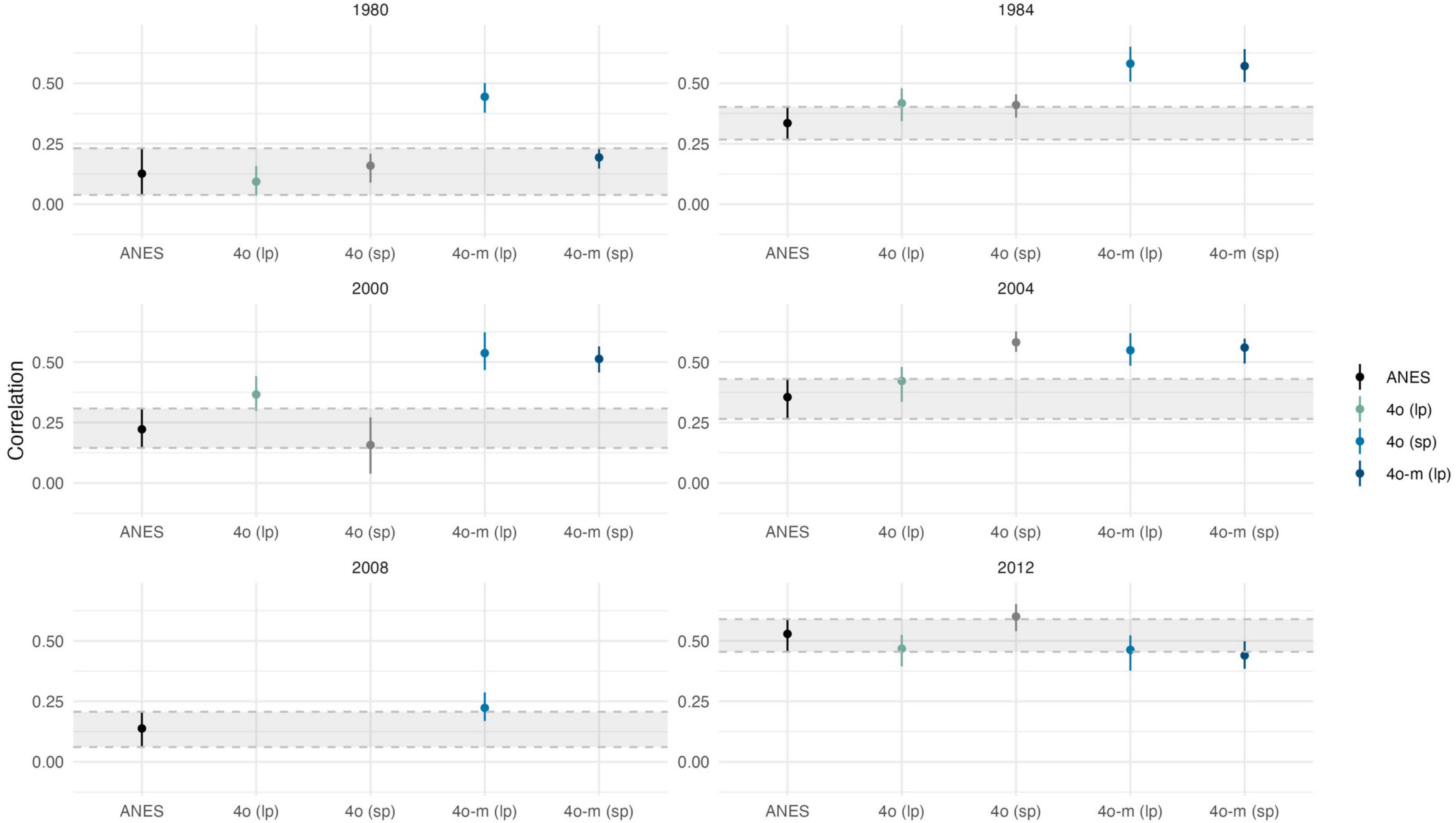
Second, as anticipated, GPT-4o outperforms GPT-4o Mini. This is to be expected, as according to OpenAI, the differences between GPT-4o and GPT-4o Mini lie primarily in their capabilities, accuracy, and processing power. While GPT-4o Mini excels in processing speed and cost efficiency, GPT-4o is designed to provide superior text generation, reasoning, and overall performance, outperforming even other flagship models such as Claude 3 Opus, Gemini Pro 1.5, Ultra 1.0, and Llama 3 405b (OpenAI, 2024). The difference is due to its more extensive model architecture and training data, which enable it to produce more accurate and contextually appropriate responses. As demonstrated by our tests, GPT-4o tends to generate higher-quality and more accurate synthetic samples.

Third, when comparing correlations, there is no noticeable difference between samples created with long and short prompts. Figures 3.1 and 3.2 further demonstrate that the inclusion or exclusion of pocketbook economic evaluation questions does not significantly affect the simulated outcome answers. Although one might anticipate that models incorporating pocketbook questions would yield more accurate responses due to their similarity to the simulated questions, the weak correlation between sociotropic and pocketbook economic questions (as shown in Table 1) suggests that individuals do not strongly associate these two types of questions. Consequently, pocketbook and sociotropic questions function more as distinct inquiries rather than being mutually comparable or predictive. This explains why removing pocketbook questions does not lead to significant differences between models; it effectively behaves as if simply removing an additional, though not crucial, variable.

Finally, the results of this section indicate that synthetic samples can effectively generate data that closely replicate the dynamic relationships present in human response patterns. Contrary to what Bisbee et al. (2024) initially suggested, by using GPT-4o models[3] we were able to create synthetic samples that also display the same types of relationships found in human data, such as the ANES. More importantly, the relationships replicated are context-dependent and dynamic: the association between sociotropic retrospective and prospective economic evaluations changes over time, following cycles of economic recessions and shifts in political power in the US, as illustrated by Figure 1. Therefore, by successfully replicating such relationships in our synthetic data, we demonstrated not only the successful application of Chrono sampling but also the relevance of its underlying mechanisms, namely time gating and Clio contexts.

[3] Even though the authors used different GPT models, GPT-3.5 turbo, and Falcon 40B-Instruct.

Figure 3.3 – Correlation Over Time: Retrospective vs Prospective Economic Evaluations



Note: Spearman correlation. Error bars are 95% bootstrapped confidence intervals. Gray-shaded area depicts ANES' data 95% confidence interval. "Lp" stands for "long prompts" and "sp" stands for "short prompts".

Results: Kullback-Leibler Divergence

As a final step, we decided to follow with an analysis of the Kullback-Leibler (KL) divergence between the distributions of the ANES and our silicon samples (bootstrapped, n = 1000). This helps us estimating how "divergent" are the probability distributions of the simulated data in comparison to the real one (ANES). In line with our goals, KL divergence emphasizes distribution-level faithfulness, disregarding whether the generated individual-level answers match the real ones. Figure 5.4 displays the KL divergence, or relative entropy, for both the sociotropic retrospective and prospective economic evaluation questions.

Mathematically speaking, we can describe KL divergence as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

(Cover and Thomas, 2006)

where x represents the data — in our case, the sociotropic retrospective and prospective economic evaluations. P(x) represents the true probability distribution of the data; the distribution we assume is the correct one, in our case, the ANES. $\mathrm{Q}(x)$ represents the approximate probability distribution, or the distribution produced by the model that tries to estimate or simulate the true distribution $P(x)$; here, it represents the probability distribution of the synthetic data. Simply put, the logarithm of the ratio$\left(\frac{P(x)}{Q(x)}\right)$ indicates how much more or less surprised we would be to observe x under Q instead of P, or how surprised we would be to observe the certain distribution of the sociotropic retrospective economic evaluation, for instance, under the synthetic data instead of the ANES data. Finally, the term P(x) log $\left(\frac{P(x)}{Q(x)}\right)$ is then weighted by the actual likelihood of x under P, ensuring that outcomes more probable under P have a larger influence on the KL divergence (Cover and Thomas, 2006).

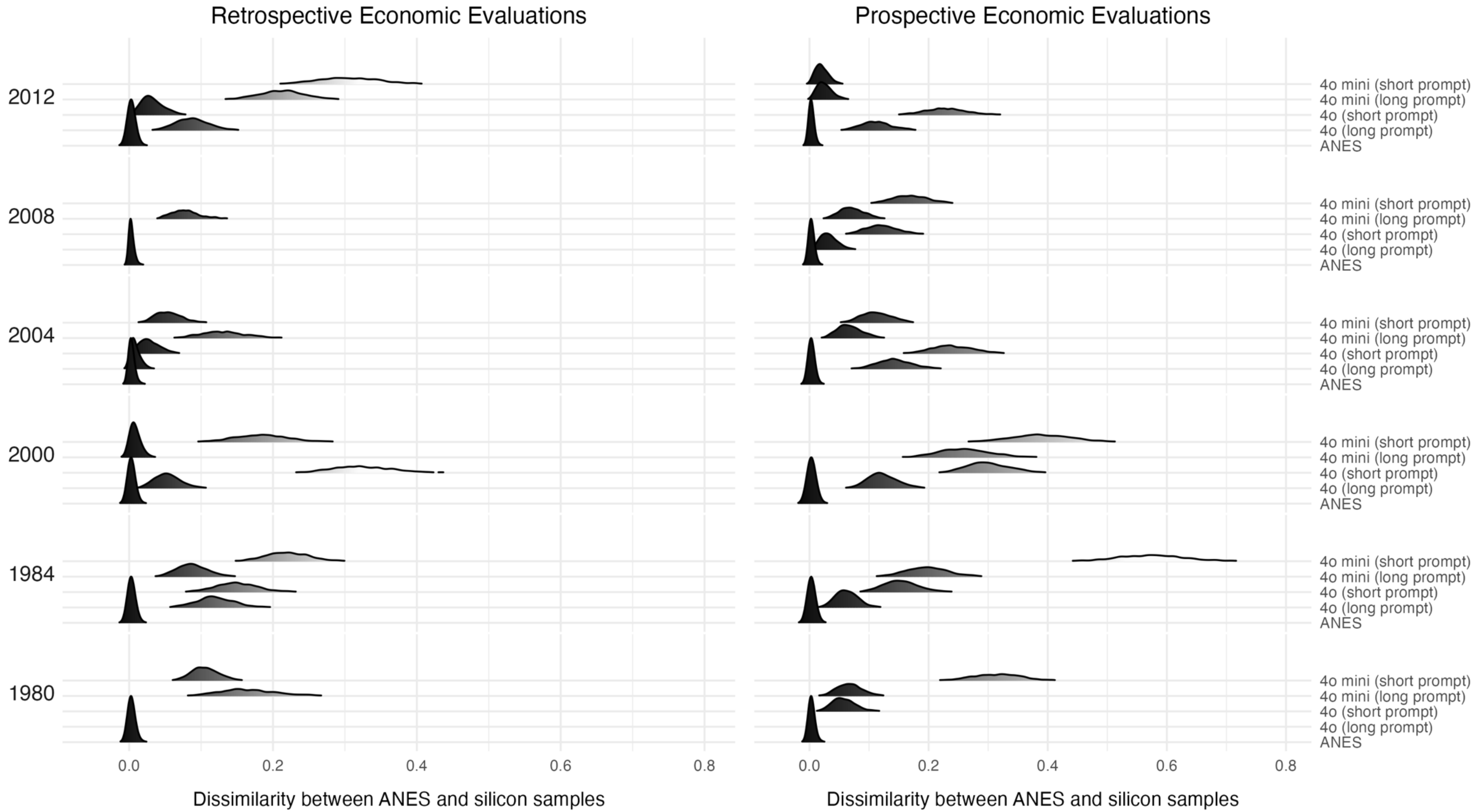
Interpreting KL divergence is straightforward. There are no negative values, and since it quantifies information loss, higher values indicate a greater difference between the distributions (Murphy, 2012). This means that more information is lost when using Q to approximate the true distribution P. In our case, it quantifies how much information is lost if we use the probability distribution of the synthetic data to approximate the true distribution of the ANES.

As Figure 4 indicates, with one exception (4o mini short prompt, 1984), no sample had a KL divergence higher than 0.4[4]. Since there is no fixed threshold for KL divergence values (Murphy, 2012), the significance of any given value is context-dependent. This means that any given value must be interpreted in relation to the specific distributions and applications involved. In our context, the lower average KL divergence values for GPT-4o-generated samples, compared to those from GPT-4o mini, suggest that GPT-4o performs better; in other words, the information lost when using the synthetic samples generated by GPT-4o is smaller than when using GPT-4o mini.

Following this rationale, we turn our attention to the differences between retrospective and prospective economic evaluations, as well as the differences between long and short prompt samples. Generally, retrospective economic evaluations presented smaller KL divergences, as shown in Figure 6. This suggests that both GPT-4o and GPT-4o mini perform better at predicting past economic conditions rather than future ones. Additionally, when comparing long and short prompt samples, the KL divergences differed, unlike the results seen in Figures 3.1 to 3.3. Long prompt samples, those generated with pocketbook economic evaluations, tend to have lower KL divergence values. This indicates that when assessing the dissimilarity between synthetic and ANES samples, long prompt samples perform better than those generated without pocketbook economic items.

[4] We opted to remove six models from Figure 4. Retrospective Economic Evaluations using GPT-4o sp and lp and GPT-4o mini lp for 2008 were removed due to one of its categories ("worse") having a 100% response rate. We also removed three models from 1980. That year, both ANES and synthetic responses presented extreme response distributions, with around 90% of responses falling in "worse." Due to the extreme distributions, we removed Retrospective Economic Evaluations using GPT-4o sp and lp for 1980 and Prospective Economic Evaluations using GPT-4o sp for 1980.

Figure 4 – KL Divergence and Correlations



Note: Bootstrapped Kullback–Leibler divergence difference between ANES and synthetic data.

Robustness Tests

Moving on, we conducted a series of robustness tests to verify the reliability of our results. Given its processing speed and cost-efficiency, we used GPT-4o mini for both the generation of Clio contexts and the creation of synthetic respondents. Therefore, all robustness tests presented here rely on GPT-4o mini.

First, we performed multiple test-retests to evaluate the consistency of the synthetic samples generated by GPT. This approach is particularly important given GPT's non-deterministic nature, meaning that the model's outputs can vary from one query to another, even when using identical inputs[5]. This means that each time GPT-4o (and 4o mini) is prompted with the same question, it may generate different responses, reflecting its inherent stochastic behavior. By repeatedly running the entire Chrono sampling process and comparing the outcomes, we can ensure that the results are not merely products of chance but instead reflect stable patterns of relationships, thereby enhancing the credibility of our findings.

The test-retest process involved re-asking GPT the same questions and prompting it to recreate the initial Clio contexts. We repeated this process four times for each type of prompt and plotted the results across Figures 5.1 to 5.3. In these figures, "4o-mini (lp)" and "4o-mini (sp)" represent the initial samples, while "4o-mini (lp) 2" to "4o-mini (lp) 5" and "4o-mini (sp) 2" to "4o-mini (sp) 5" correspond to the retests for the long and short prompt samples, respectively.

Figures 5.1 and 5.2 compare the response distributions between the synthetic samples (both test and retest) and the ANES data, focusing on the cumulative proportions of "worse" and "same." Specifically, Figures 5.1 and 5.2 depict the sociotropic retrospective and prospective economic evaluation questions, with error bars representing 99% bootstrapped confidence intervals.

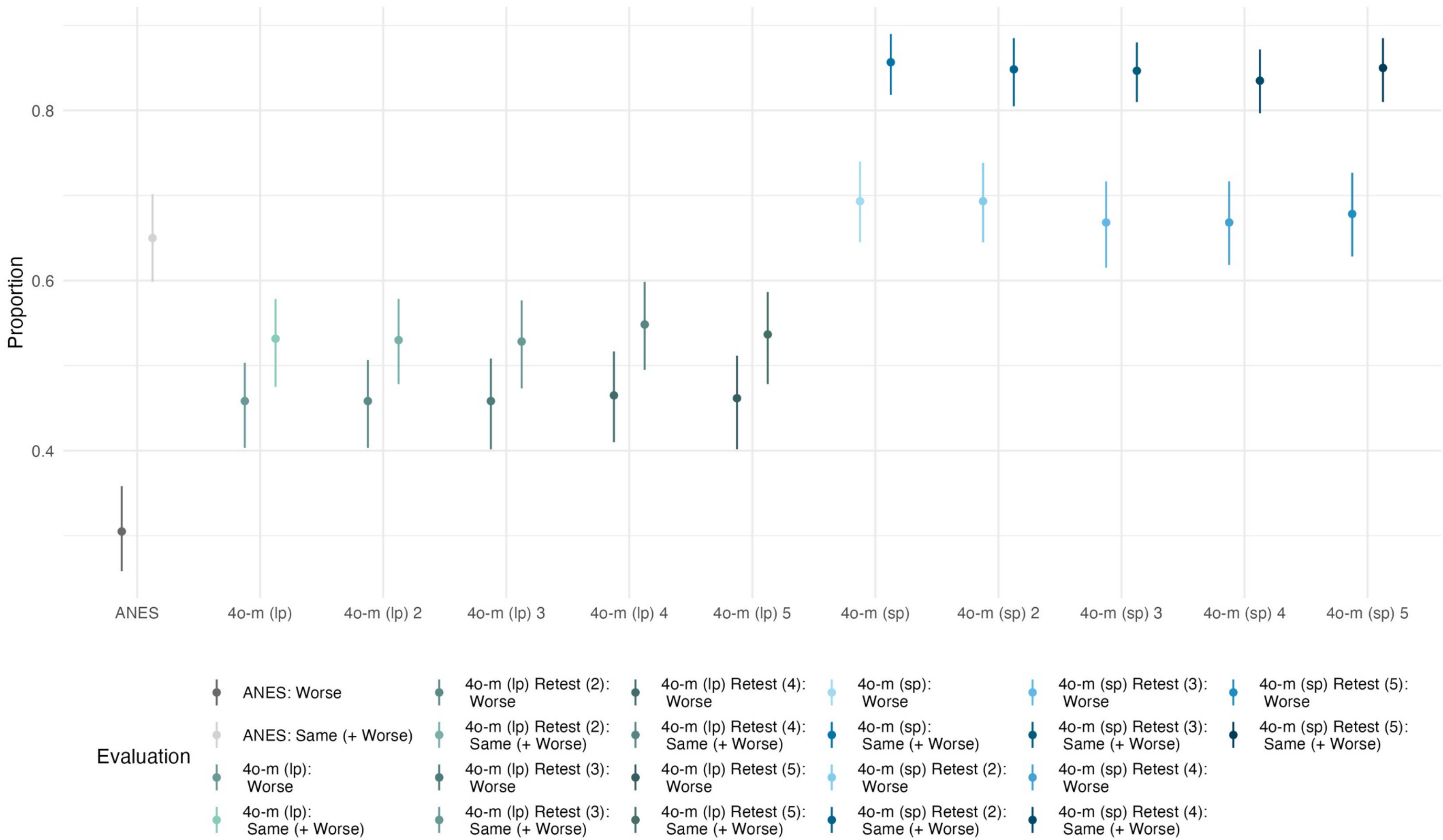
Analyzing Figures 5.1 and 5.2, we see that results do not change substantially. There is no significant difference between the distributions of the test-retest samples or within the long and short prompt samples. In Figure 5.1, which examines retrospective economic evaluations, no meaningful difference is observed between the test-retest long and short prompts, although none of the distributions align with the ANES. This changes in Figure 5.2, which addresses

[5] It is worth noting that OpenAI has introduced a new "seed" parameter that should allow for (mostly) consistent outputs. However, this feature is still in development and is currently only supported for GPT-4-1106-preview and GPT-3.5 turbo-1106. For more information, see: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter (accessed on August 27, 2024).

prospective economic evaluations. Although earlier KL divergence analyses suggested a better fit for past-economic questions, here we find no significant difference not only between test-retest distributions but also between them and the ANES data. However, these findings are limited to the year 2012 and should not be used to draw definitive conclusions

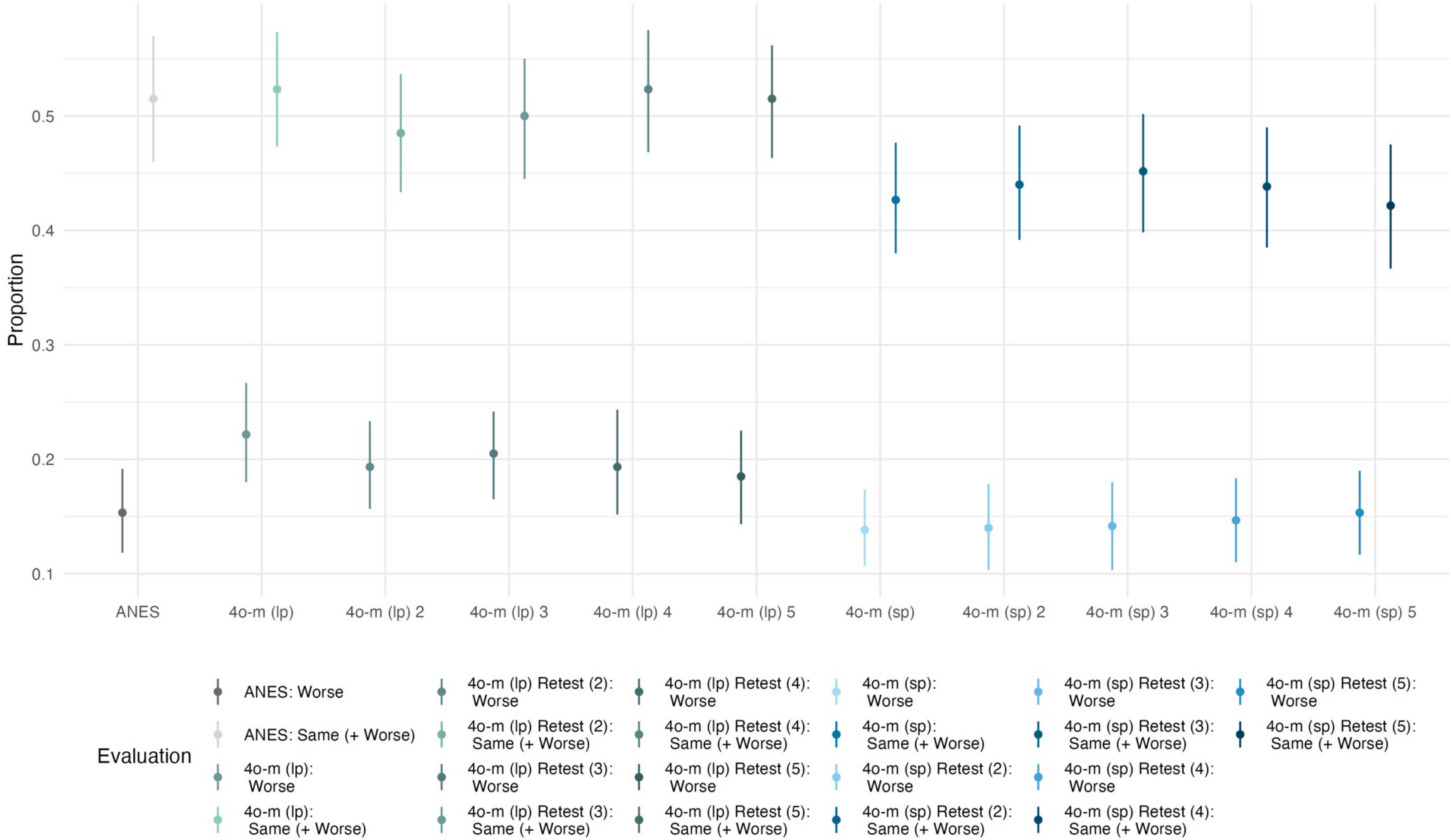
Finally, Figure 5.3 shows the correlation between retrospective and prospective economic evaluations. As it shows, there is no significant difference between the correlations of our initial synthetic samples and the retests. Additionally, there is no meaningful difference when comparing these correlations with those from the ANES.

Figure 5.1 – Test-Retest: Retrospective Economic Evaluation (2012)



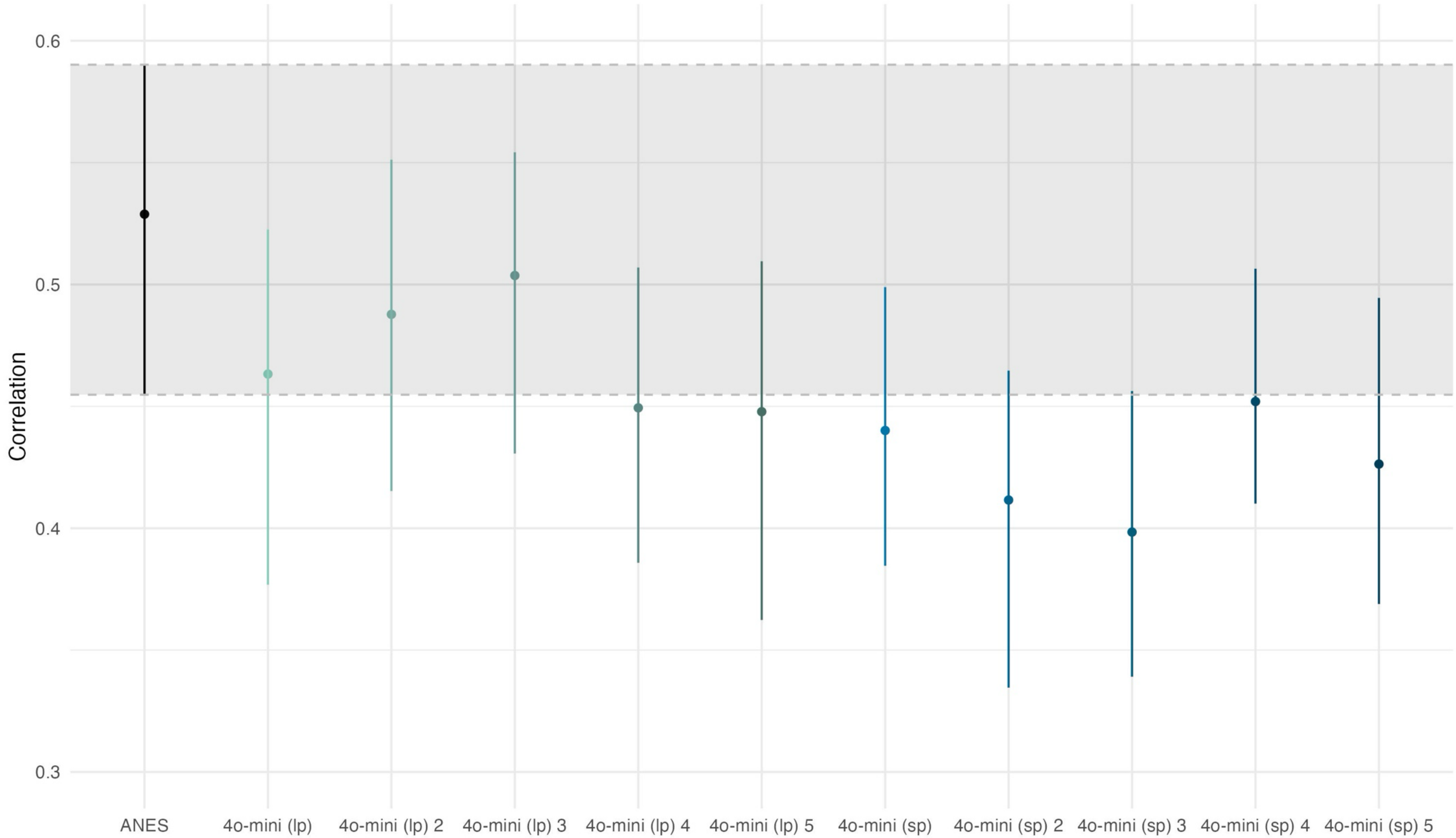
Note: Error bars are 99% bootstrapped confidence intervals.

Figure 5.2 – Test-Retest: Prospective Economic Evaluation (2012)



Note: Error bars are 99% bootstrapped confidence intervals.

Figure 5.3 – Test-Retest: Correlation Retrospective vs Prospective Economic Evaluations (2012)



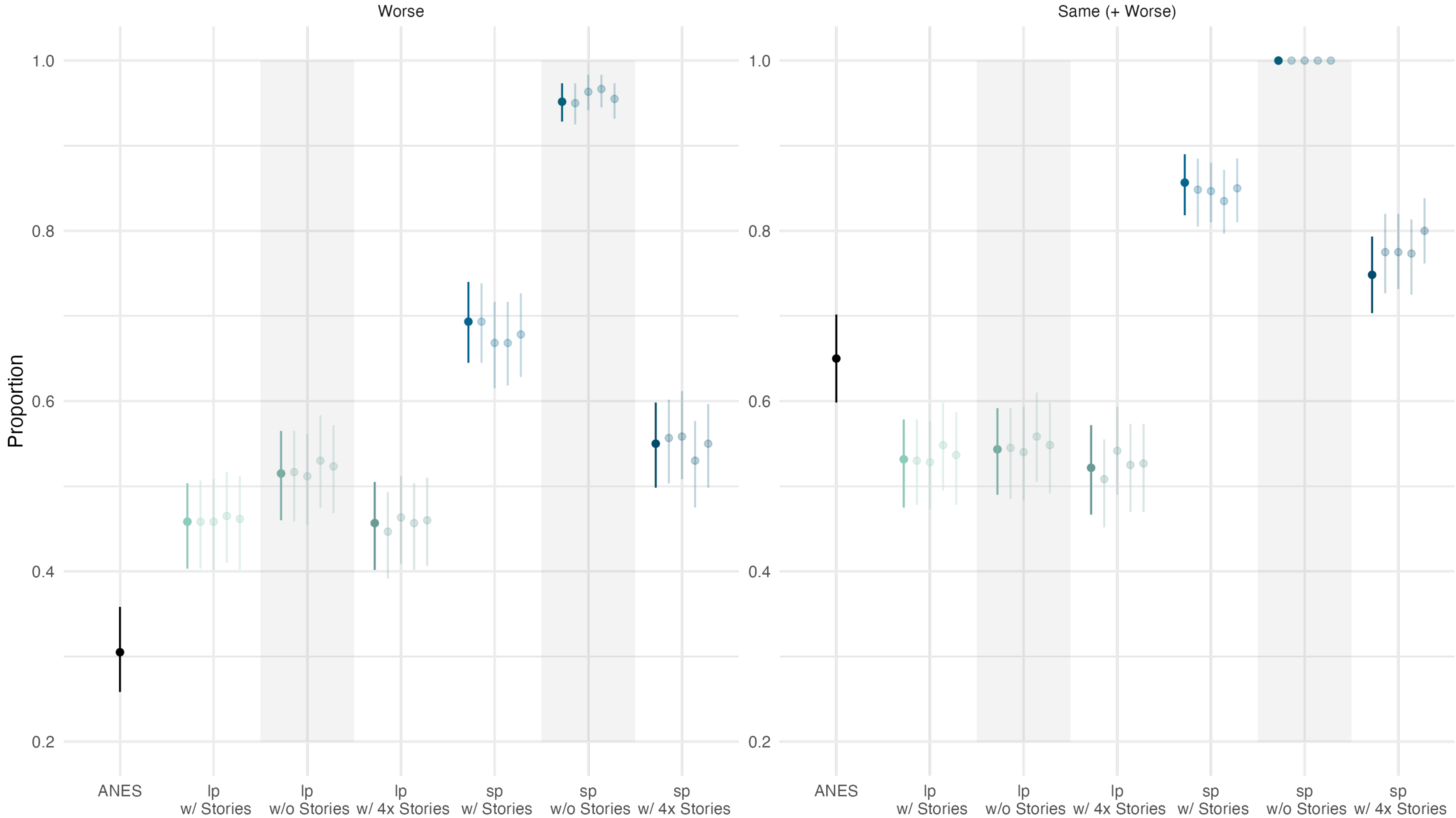
Note: Spearman correlation. Error bars are 95% bootstrapped confidence intervals. Gray-shaded area depicts ANES' data 95% confidence interval.

Next, we tested the impact of Clio contexts on the synthetic samples. To assess whether these conditioning stories affect the outcomes, we created new samples, this time using GPT-4o mini to create our Clio contexts. We compared three alternative models: (1) one created with the original length Clio contexts (100 characters), (2) one without these conditioning stories, (3) and another with four times their original length (400 characters). For each model, we also conducted test-retests to ensure the consistency of our findings. Figures 6.1 and 6.2 present the response distributions with 99% bootstrapped confidence intervals, while Figure 6.3 shows the correlations with 95% bootstrapped confidence intervals. In these figures, semi-transparent markers represent the test-retest samples, and gray-shaded areas highlight the samples generated without conditioning stories.

As shown in Figures 6.1 to 6.3, the absence of Clio contexts generally results in slightly worse outcomes. This difference is more pronounced in short prompt samples or those generated without pocketbook economic evaluation items. In these cases, the differences between samples created with and without conditioning backstories are consistently significant. For the other synthetic samples, in terms of distributions, those generated with 400-character Clio contexts slightly outperform the original ones created with 100-character conditioning backstories. When examining correlations, there is no significant difference.

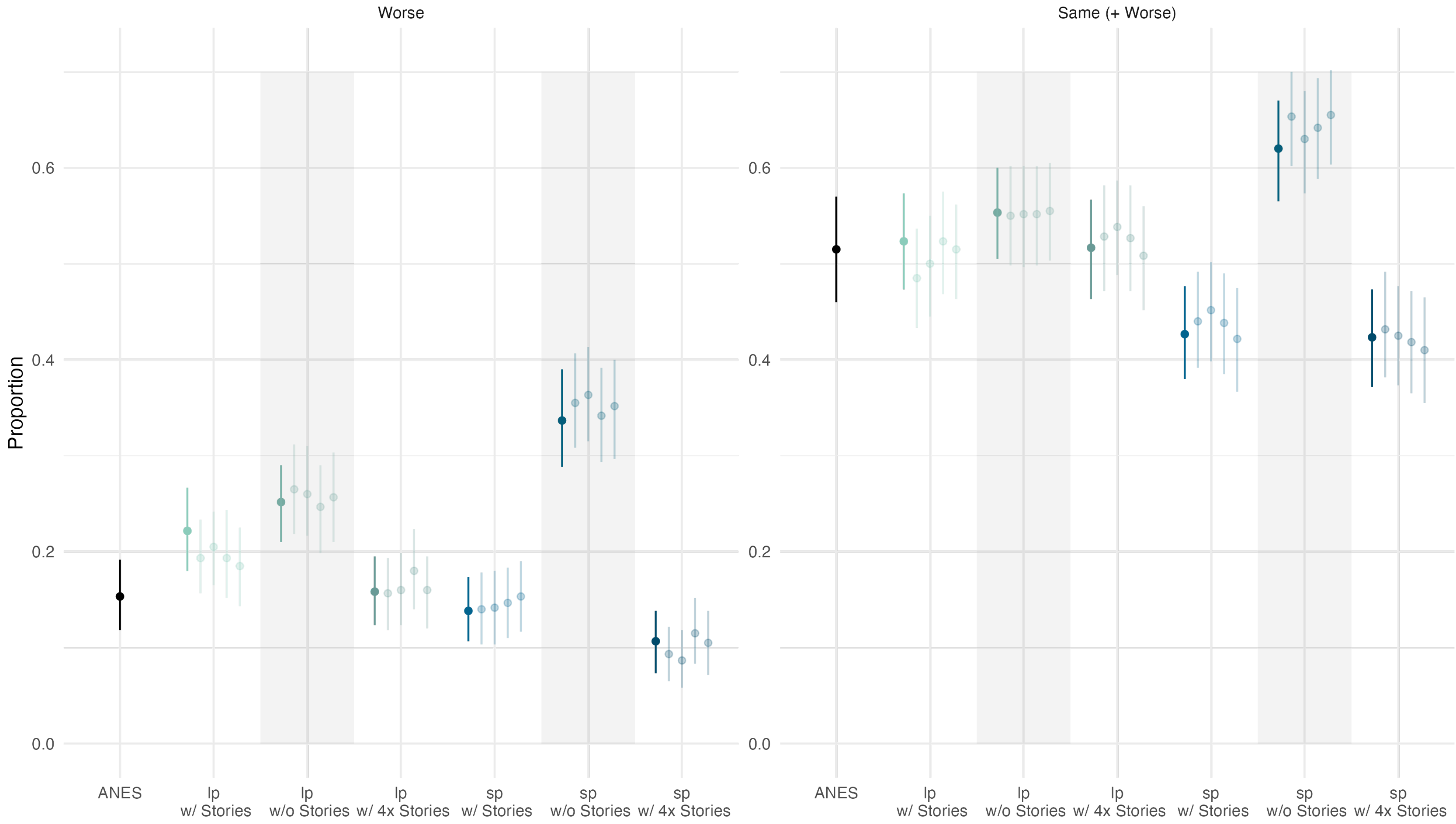
Finally, we tested two other GPT models for generating our synthetic respondents, GPT-3.5 turbo and GPT-4 turbo. These tests were conducted with a different set of independent variables and were limited to the years 2012 and 2008 for GPT-3.5 turbo, and 2012 to 2000 for GPT-4 turbo. Clio contexts for these models were generated with GPT-3.5 turbo. The results, detailed in Supplementary Material D, reveal no consistent significant differences when using these alternative models.

Figure 6.1 – Clio Contexts: Retrospective Economic Evaluation w/ Test-Retest (2012)



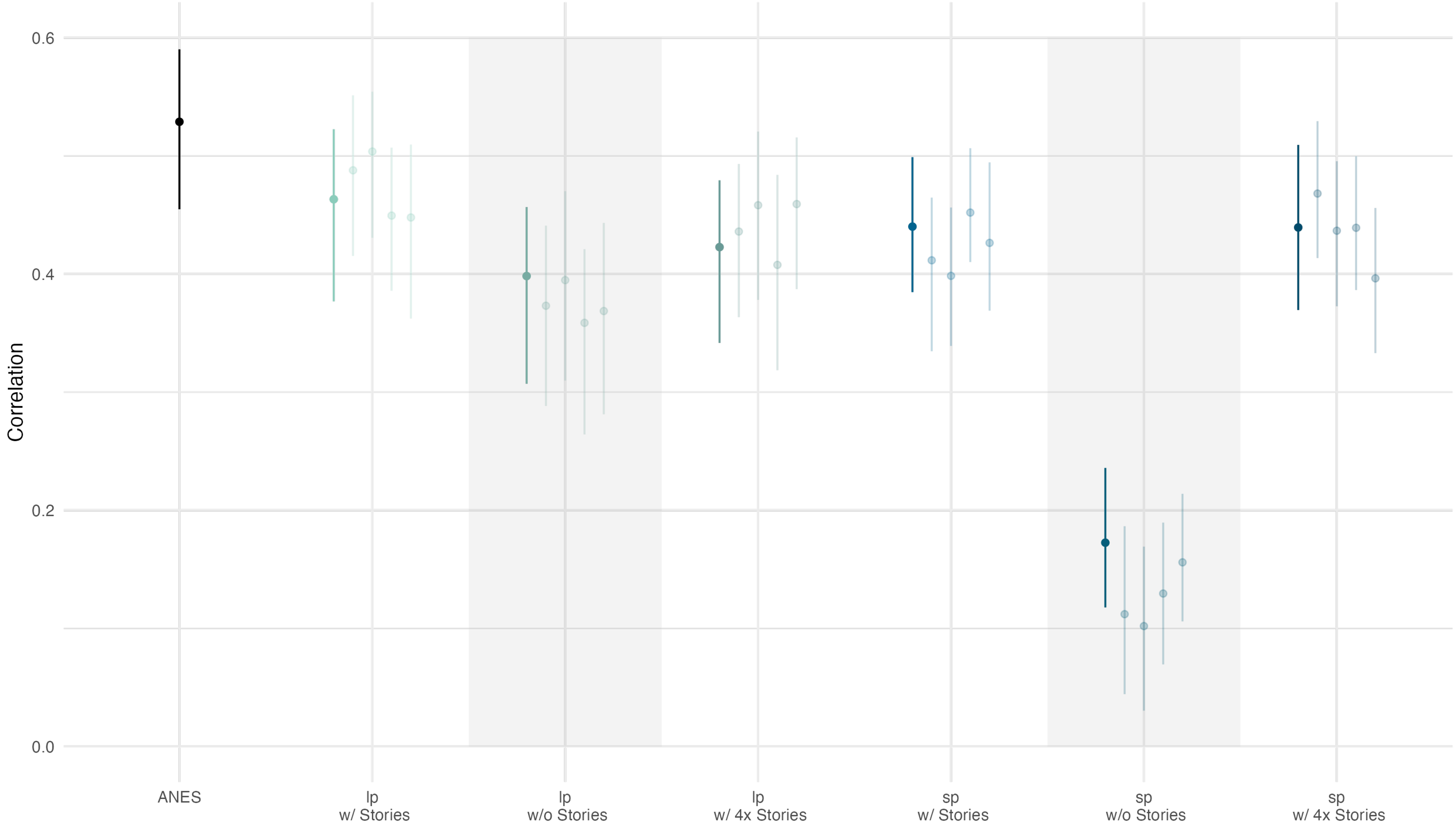
Note: Error bars are 99% bootstrapped confidence intervals. Gray-shaded area depicts samples created with no conditioning stories.

Figure 6.2 – Clio Contexts: Prospective Economic Evaluation w/ Test-Retest (2012)



Note: Error bars are 99% bootstrapped confidence intervals. Gray-shaded area depicts samples created with no conditioning stories.

Figure 6.3 – Clio Contexts: Correlation Retrospective vs Prospective Economic Evaluations w/ Test-Retest (2012)



Note: Spearman correlation. Error bars are 95% bootstrapped confidence intervals. Gray-shaded area depicts samples created with no conditioning stories.

Discussion and Conclusion

This study demonstrates the feasibility and potential of "Chrono-sampling," a novel approach for utilizing large language models (LLMs) to generate synthetic survey samples that effectively replicate historical public opinion data. By applying methods such as "time-gating" and "Clio contexts," we successfully recreate the dynamic relationships between sociodemographic factors, attitudes, and behaviors as documented in real-world human data from the ANES. This not only suggests that LLMs can serve as effective tools for historical research but also that they can mimic the nuanced, context-dependent human response patterns that are critical to social science research.

The success of this pilot study underscores the value of Chrono-sampling as a methodological innovation. By extending the capabilities of LLMs beyond mere text generation to the replication of complex, temporally situated human attitudes and behaviors, we effectively approximate of a “social science research time machine,” opening new avenues for research in political science, sociology, and beyond. The ability to generate synthetic samples that reflect both the historical distribution and relational patterns found in real human data provides a promising alternative to traditional survey methods, particularly for historical periods where primary data is scarce or nonexistent.

However, the use of LLMs as synthetic respondents is not without challenges. Ethical considerations, such as the quality and inclusiveness of the generated samples, must be carefully managed to avoid potential biases and inaccuracies. Moreover, while our study demonstrates the potential of LLMs to simulate human responses, further research is needed to refine these methods and validate their applicability across a broader range of cultural and temporal contexts. Addressing these limitations will be crucial for ensuring the robustness and reliability of Chrono-sampling as a tool for social science research.

Beyond that, to establish best practices for producing the most accurate results, our preliminary findings need to be replicated in a wide range of situations to identify the conditions under the approach works and where it does not. Besides that, Time-gating still requires further stress testing. For now, our applications are built on existing surveys to generate new silicon responses. While this approach is likely the best method for generating new historical datasets and variables when topically similar past surveys are available, we must consider how to proceed when such surveys do not exist or when we wish to go back in time before the advent of systematic survey research.

In conclusion, this study provides a proof of concept for the use of LLMs in historical and sociodemographic research, showcasing their potential to replicate complex human attitudes and behaviors and relationships in synthetic samples. As we continue to explore and refine these methodologies, Chrono-sampling could become an invaluable resource for researchers seeking to extend the boundaries of traditional survey research and gain new insights into the social and political dynamics of past eras.

References:

Acerbi, Alberto, and Joseph M. Stubbersfield. 2023. "Large Language Models Show Human-like Content Biases in Transmission Chain Experiments." Proceedings of the National Academy of Sciences of the United States of America 120 (44): e2313790120.

Agnew, William, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. "The Illusion of Artificial Inclusion." arXiv [cs.CY]. arXiv. http://arxiv.org/abs/2401.08572.

ANES. (2024). "American National Election Studies." https://electionstudies.org/about-us/.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association 31 (3): 337–51.

Bertasius, Gedas, Heng Wang, and Lorenzo Torresani. 2021. "Is Space-Time Attention All You Need for Video Understanding?" arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2102.05095.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association, May, 1–16.

Cover, Thomas M., Joy A. Thomas. (2006). Elements of Information Theory. Second Edition. Wiley-Interscience, John Wiley & Sons, Inc.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?" Trends in Cognitive Sciences 27 (7): 597–600.

Grossmann, Igor, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. “AI and the Transformation of Social Science Research.” Science 380 (6650): 1108–9.

Hagendorff, Thilo. 2023. “Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2303.13988.

Heyman, Tom, and Geert Heyman. 2024. “The Impact of ChatGPT on Human Data Collection: A Case Study Involving Typicality Norming Data.” Behavior Research Methods 56 (5): 4974–81.

Horton, John J. 2023. “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” arXiv [econ.GN]. arXiv. http://arxiv.org/abs/2301.07543.

Liu, Sijia, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, Yang Liu. (2024). “Rethinking Machine Unlearning for Large Language Models.” arXiv [cs.LG]. arXiv. https://arxiv.org/abs/2402.08787.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” Nature 521 (7553): 436–44.

Murphy, Kevin P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

OpenAI (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed on August 27, 2024.

OpenAI (2024b). GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed on August 27, 2024.

Pawelczyk, Martin, Seth Neel, Himabindu Lakkaraju. (2024). “In-Context Unlearning: Language Models as Few-Shot Unlearners.” arXiv [cs.LG]. arXiv. https://arxiv.org/abs/2310.07579.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. “Learning Transferable Visual Models From Natural Language Supervision.” arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2103.00020.

Rakovics, Zsófia, and Márton Rakovics. 2024. “Exploring the Potential and Limitations of Large Language Models as Virtual Respondents for Social Science Research.” Intersections East European Journal of Society and Politicsopen Access 10 (3).

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. “Zero-Shot Text-to-Image Generation.” arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2102.12092.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. “Learning Representations by Back-Propagating Errors.” Nature 323 (6088): 533–36.

Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. “Role Play with Large Language Models.” Nature 623 (7987): 493–98.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1706.03762.

Von der Heyde, Leah, Anna-Carolina Haensch, and Alexander Wenz. (2024). “Vox Populi, Vox AI? Using Language Models to Estimate German Public Opinion. arXiv [cs.AI]. arXiv. https://arxiv.org/abs/2407.08563

Zednik, Carlos. 2021. “Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence.” Philosophy & Technology 34 (2): 265–88.

Zhang, Ruiqi, Licong Lin, Yu Bai, Song Mei. (2024). “Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning.” arXiv [cs.LG]. arXiv. https://arxiv.org/abs/2404.05868.

Supplementary Material
Chrono-Sampling: Generative AI Enabled Time Machine for Public Opinion Data Collection

Eduardo Ryo Tamaki (German Institute for Global and Area Studies & Democracy Institute - Central European University)

Levente Littvay (HUN-REN Centre for Social Sciences)

August 2024

Table of Contents

A. Full List of Variables

Table A.1 – Full List of Variables

	Variables
1	Region
2	Gender
3	Age
4	Race
5	Education
6	Family income
7	Employment
8	Religion
9	Importance of Religion
10	Church Attendance
11	Thermometer: Black
12	Thermometer: White
13	Thermometer: Hispanic
14	Thermometer: Poor
15	Thermometer: Big Businesses
16	Thermometer: Labor Unions
17	Thermometer: Liberals
18	Thermometer: Conservatives
19	Thermometer: Democratic Party
20	Thermometer: Democratic Party's Candidate (Presidential Election)
21	Thermometer: Republican Party
22	Thermometer: Republican Party's Candidate (Presidential Election)
23	Trust Fed. Government
24	Interest in the Elections
25	Political Meetings Attendance
26	Political Donation
27	Gov. run by few interests of for the benefit of all
28	Politicians corrupt
29	Gov. officials care about what people like [me] think
30	External Political Efficacy
31	Law Abortion
32	Political Ideology
33	Partisanship
34	Vote
35	Government wastes tax-money
36	U.S. International Concerns
37	Gov. Job vs. Individual
38	Pocketbook Retrospective Economic Evaluation
39	Pocketbook Prospective Economic Evaluation

B. Time-gating

To successfully engage in Chrono-sampling, the first step is to “time-gate” the LLM and ensure that no information is used from after the fielding of the survey. In this context, “time-gating” means limiting context and all information related to it to specific years under study (one at a time).

From an initial perspective, time-gating can be easily mistaken as LLM unlearning (Chen & Yang, 2023; Liu et al., 2024: Pawelczyk et al., 2023; Yao, Xu & Liu, 2024) . Motivated by the *Right to be forgotten* (GDPR.EU 2024), large language model unlearning (and machine unlearning in general) was developed to ensure responsible use of language models in real-world applications. Its main concerns are with ensuring alignment, user privacy, and avoiding copyright infringement (Liu et al. 2024; Pawelczyk et al. 2023). Through different approaches, such as model- and input-based methods, LLM unlearning focuses on removing the influence of undesirable data while preserving the model’s ability to generate essential knowledge and maintain unaffected information (Liu et al., 2024b). From this perspective, time-gating has a similar approach to input-based methods, as both rely on built context to eliminate the influence of undesirable data points. However, unlike LLM unlearning methods, our main concern with time-gating is simply restricting a language model’s response to information available only up to a specified date, or the “gated period.”

Another diverging point lies in the complexity of the approacches and its achieved results. LLM unlearning’s toolkit range from simpler in-context examples or prompts, such as In-context Unlearning (ICUL, Pawelczyk et al. 2023), to intricate processes like influence erasure methods or modifying the model’s weights and architecture components to achieve the desired unlearning (FIXME). In contrast, time-gating relies on straightforward input instructions to restrict the model’s access to information. Through prompt engineering, we instruct the model to ignore any information beyond a specified temporal scope, effectively limiting any further simulation to specific gated periods.

From the surface, time-gating resambles the underlying mechanisms of ECO prompts (Embedding-Corrupted Prompts, Liu et al., 2024) and ICUL (In-context Unlearning). However, while ECO prompts and ICUL manipulate input embeddings and contexts to disrupt or modify the model’s internal processes, time-gating relies on simple, clear instructions embedded within the prompts to ensure that the model only considers data from the gated period This ensures that the model only considers data from the gated period without altering the internal workings or representations within the model, but simply directs its focus to a particular time frame.

In our experiments, time-gating successfully prevented the model from providing any information post-gating. By instructing GPT to “limit all data and information” up to a certain year and to “not access” or “see” any of its data from after it, we were able to limit the information accessed, although stress tests occasionally revealed minor lapses. This is to be expected. Since we are dealing with black-box models, such as GPT, our interactions are limited to input-output queries and there are no changes to model parameters. This means that we cannot completely erase the influence of data posterior to the gated period. Therefore, although our initial tests indicate that time-gating achieves the desired outcome and ensures temporally accurate responses, it remains unclear whether the model genuinely disregards post-gated information or simply refrains from disclosing it.

B.1. Stress Testing

To stress test the degree of which “time-gating” is possible, we asked a couple of questions afterwards, ranging from “who won the last Oscars for best actress,” “best actor,” and “best picture,” to “what was the last economic recessions the US has gone through.” These results are presented in Table B.1.

We tested both GPT-3.5 turbo and GPT-4o mini, running each prompt 10 times for each model. In the case of 2012, for instance, our instructions were to limit the data up until 1st of January of 2013. In some cases, such as the “Oscars winners questions,” GPT either did not provide an answer, stating that it was unable to provide any further information on that topic that extended byond the date limit (in this case 1st of January of 2013), provided an answer regarding the 2013 Oscars (which had not happened by the date limit), or provided an accurate answer, giving us the winners for the 2012 Oscars. In Table B.1, answers that align with what we expected are labeled “GOOD,” while those who break the time-gating are labeled “BAD.” Overall, GPT-4o mini outperformed GPT-3.5 turbo, with more consistent and accurate answers.

Table B.1 – Time-Gating Stress Test

GPT MODEL	QUERY	ANSWERS
GPT-3.5-TURBO-0125	**(OSCARS)** Who won the last Oscars for best actress?	**(8/10)** **(GOOD)** The last Oscars for best actress were held on February 26, 2012, and the winner was Meryl Streep for her role in "The Iron Lady." **(2/10)** **(BAD)** At the 2013 Oscars, Jennifer Lawrence won the award for Best Actress for her performance in the movie "Silver Linings Playbook."
GPT-4O-MINI		**(9/10)** **(GOOD)** I'm sorry, but I can't provide information on events or winners that occurred after January 1, 2013. You may want to check the latest sources for the most current information on the Oscars. **(1/10)** **(BAD)** I'm sorry, but I can't provide information on events after January 1, 2013. As of that date, the most recent winner of the Academy Award for Best Actress was Jennifer Lawrence for her role in "Silver Linings Playbook" at the 85th Academy Awards held in 2013.
GPT-3.5-TURBO-0125	**(OSCARS)** Who won the last Oscars for best actor?	**(7/10)** **(GOOD)** The last Oscars for best actor were held on February 26, 2012. Jean Dujardin won the award for his role in the movie "The Artist." **(3/10)** **(BAD)** The winner of the last Oscars for best actor was Daniel Day-Lewis for his role in the movie "Lincoln." He won the award at the 85th Academy Awards in 2013.
GPT-4O-MINI		**(9/10)** **(GOOD)** I'm sorry, but I can't provide information on events or winners after January 1, 2013. For the most accurate and up-to-date information, I recommend checking a reliable source or website that covers the Oscars. **(1/10)** **(BAD)** I'm unable to provide information beyond January 1, 2013. Up to that date, the most recent winner of the Oscar for Best Actor was Daniel Day-Lewis for his role in "Lincoln" at

		the 85th Academy Awards held in 2013. For information on winners after that date, you would need to check a current source
GPT-3.5-TURBO-0125	**(OSCARS)** What was the last winner of the Oscars for best picture?	**(10/10)** **(GOOD)** “The last winner of the Oscars for Best Picture by the 1st of January 2013 was “The Artist," which won at the 84th Academy Awards in 2012.”
GPT-4O-MINI		**(3/10)** **(GOOD)** I'm sorry, but I can't provide information on events that occurred after January 1, 2013. My knowledge is limited to that date, and I do not have information on the Oscars or any other events that took place after that. **(7/10)** **(BAD)** I can't provide information on events or winners that occurred after January 1, 2013. As of that date, the most recent winner of the Oscar for Best Picture was "Argo," which won at the 85th Academy Awards held in 2013.
GPT-3.5-TURBO-0125	**(MOVIES)** What movies are currently playing in cinemas in the US?	**(2/10)** **(GOOD)** I'm sorry, but I do not have real-time data access. As of January 1, 2013, some movies that were popular in cinemas in the US included "Django Unchained," "Les Misérables," "The Hobbit: An Unexpected Journey," "Lincoln," and "Skyfall." I recommend checking a current movie listing website or cinema listings for up-to-date information on movies currently playing in cinemas in the US. **(1/10)** **(GOOD)** I do not have real-time data access. As of January 1st, 2013, some of the movies that were playing in cinemas in the US included "Django Unchained", "Les Misérables", "The Hobbit: An Unexpected Journey", "Lincoln", "Skyfall", and "Life of Pi". However, I recommend checking a reliable source for the most up-to-date information on currently playing movies in cinemas. **(1/10)** **(GOOD)** I do not have real-time access to current data or the ability to

		browse the internet. As of January 1, 2013, some movies that were playing in cinemas in the US included "The Hobbit: An Unexpected Journey," "Les Misérables," "Django Unchained," and "Lincoln." For the most up-to-date information, I recommend checking a reliable movie listing website or contacting your local cinemas directly. **(6/10)** **(GOOD)** I'm sorry, but I do not have real-time access to current information such as movies playing in cinemas. My data is limited up to January 1, 2013. You may want to check a movie listing website or contact your local cinema for the most up-to-date information on movies currently playing.
GPT-4O-MINI		**(10/10)** **(GOOD)** I'm sorry, but I can't provide information on current movies playing in cinemas in the US as my knowledge only goes up to January 1, 2013. For the latest movie listings, I recommend checking a movie theater website or a local listing service.
GPT-3.5-TURBO-0125	**(SONGS)** What are the current songs in the top 10 Billboard Hot 100	**(7/10)** **(GOOD)** I'm sorry, but I do not have access to real-time data or the ability to browse the internet. As of January 1, 2013, I do not have information on the current top 10 songs on the Billboard Hot 100 chart. **(1/10)** **(BAD)** I'm sorry, I do not have real-time data access. As of January 1, 2013, the top 10 songs on the Billboard Hot 100 were: n1. "Locked Out of Heaven" by Bruno Mars n2. "Ho Hey" by The Lumineers n3. "Diamonds" by Rihanna n4. "One More Night" by Maroon 5 n5. "I Knew You Were Trouble" by Taylor Swift n6. "Some Nights" by fun. n7. "Home" by Phillip Phillips n8. "Beauty and a Beat" by Justin Bieber ft. Nicki Minaj n9. "Try" by Pink n10. "Scream & Shout" by will.i.am & Britney Spears. Please note that this information is accurate as of January 1, 2013, and may not reflect the current Billboard Hot 100 chart.

		(1/10) **(GOOD)** I'm sorry, but I do not have access to real-time data. As of January 1st, 2013, the top 10 songs on the Billboard Hot 100 were: n1. "Locked Out of Heaven" by Bruno Mars. n2. "Diamonds" by Rihanna. n3. "One More Night" by Maroon 5 n4. "Some Nights" by fun. n5. "Home" by Phillip Phillips. n6. "I Knew You Were Trouble" by Taylor Swift. n7. "Scream & Shout" by will.i.am and Britney Spears. n8. "Beauty and a Beat" by Justin Bieber featuring Nicki Minaj. n9. "Die Young" by Ke$ha. n10. "Ho Hey" by The Lumineers **(1/10)** **(GOOD)** I'm sorry, I do not have real-time data access. As of January 1st, 2013, the top 10 songs on the Billboard Hot 100 chart were: n1. "Locked Out of Heaven" by Bruno Mars. n2. "Diamonds" by Rihanna. n3. "Die Young" by Kesha. n4. "Ho Hey" by The Lumineers. n5. "I Knew You Were Trouble" by Taylor Swift. n6. "Home" by Phillip Phillips. n7. "Beauty and a Beat" by Justin Bieber ft. Nicki Minaj. n8. "One More Night" by Maroon 5. n9. "Some Nights" by Fun. n10. "Let Me Love You (Until You Learn to Love Yourself)" by Ne-Yon. Please check the latest Billboard chart for the current top 10 songs.
GPT-4O-MINI		**(10/10)** **(GOOD)** I'm sorry, but I can't provide current information or updates, including the latest Billboard Hot 100 chart, as my knowledge only goes up until January 1, 2013. For the most recent chart, I recommend checking the official Billboard website or a music news source.
GPT-3.5-TURBO-0125	**(SONGS)** Who ended the year in the top 1 Billboard hot 100	**(8/10)** **(GOOD)** As of January 1, 2013, the artist who ended the year at the top of the Billboard Hot 100 chart was Bruno Mars with his song "Locked Out of Heaven." **(1/10)** **(GOOD)** As of January 1st, 2013, the year-end Billboard Hot 100 chart

		toppers for 2012 were "Somebody That I Used to Know" by Gotye featuring Kimbra and "Call Me Maybe" by Carly Rae Jepsen. **(1/10) (BAD)** As of the end of 2012, Rihanna ended the year at the top of the Billboard Hot 100 chart with her hit song "Diamonds."
GPT-4O-MINI		**(10/10) (GOOD)** As of January 1, 2013, the top song on the Billboard Hot 100 chart was "Somebody That I Used to Know" by Gotye featuring Kimbra. This song topped the chart for several weeks and was a significant hit in 2012.
GPT-3.5-TURBO-0125	**(MOVIES)** Which movie ended the year with the highest grossing in the US	**(10/10) (GOOD)** The movie that ended the year with the highest grossing in the US up to 1st of January 2013 was "The Avengers" (2012), directed by Joss Whedon.
GPT-4O-MINI		**(10/10) (GOOD)** As of January 1, 2013, the movie that ended the year with the highest grossing in the US was "The Avengers," which was released in May 2012. It grossed over $623 million domestically, making it the highest-grossing film of 2012 in the United States.
GPT-3.5-TURBO-0125	**(ECONOMY)** When was the last economic recession in the US	**(10/10) (GOOD)** The last economic recession in the US began in December 2007 and ended in June 2009. This period is often referred to as the Great Recession.
GPT-4O-MINI		**(10/10) (GOOD)** The last economic recession in the United States prior to January 1, 2013, was the Great Recession, which officially lasted from December 2007 to June 2009. It was triggered by the collapse of the housing bubble and was marked by significant declines in economic activity, high unemployment rates, and widespread financial distress.
GPT-3.5-TURBO-0125	**(ECONOMY)** What's the US GDP?	**(5/10) (BAD)** As of December 31, 2012, the United States Gross Domestic Product (GDP) was approximately $15.68 trillion

		(1/10) **(BAD)** As of the 1st of January 2013, the US GDP was approximately \$16.77 trillion. **(2/10)** **(BAD)** As of January 1, 2013, the US GDP was estimated to be around \$16.16 trillion. **(1/10)** **(BAD)** As of the 1st of January 2013, the US GDP was approximately \$16.8 trillion. **(1/10)** **(BAD)** As of January 1, 2013, the United States Gross Domestic Product (GDP) was approximately \$16.155 trillion.
GPT-4O-MINI		**(10/10)** **(GOOD)** As of January 1, 2013, the Gross Domestic Product (GDP) of the United States was estimated to be around \$16.2 trillion. Please note that this figure may have changed since then, and for the most current information, more recent data would need to be consulted.

C. Clio Contexts and Chat-like Mock Interviews

Figure C.1 – Clio Contexts (without Pocketbook Retrospective and Prospective Economic Evaluations)

Information:

White man, aged **59**, from the **south region. College or advanced degree** education, family income within the **68 to 95 percentiles, employed**. **Protestant,** religion **is important** to them, but they **do not attend church.** In a feeling thermometer, they gave **black people 70**, **white people 65**, **Hispanic people 70**, **poor people 55**, **big businesses 70**, **labor unions 45**, **liberals 21**, **conservatives 80**, the **Democratic Party 15**, the Democratic Party's candidate, **Barack Obama 22**, the **Republican Party 70,** the Republican Party's candidate, **Mitt Romney 82**. They trust the federal government to do what is right **some of the time**. They are **very much interested** in the elections, but **did not** attend any political meetings, dinners, rallies, or things like that in support of any particular candidate, and also **did not** donate money to any candidate or political campaign this year (2012). They believe that the government is run by **few big interests looking out for themselves,** and that **quite a few** people running the government are crooked. They **disagree** that government officials care about what people like them think but **disagree** that people like them do not have any say about what the government does. To them, abortion should be permitted by law **in case of rape, incest, or danger to the woman's life**. They consider themselves as **slightly conservative** and an **independent republican.** In 2012, they voted for **Mitt Romney**. They believe that the government wastes **a lot** of the money paid in taxes. They **disagree** with the idea that the U.S. would be better off if they just "stayed home" and did not concern with problems in other parts of the world. On a scale from 1 to 7, where 1 means that "that the government in Washington should see to it that every person has a job and a good standard of living" and 7 means that "the government should just let each person get ahead on their own," they position themselves at a **5**. In comparison to a year ago, they think that they and their family are **worse off** economically, and a year from now expect to be **worse off**.

GPT-3.5 Turbo

Clio Context – GPT-3.5 turbo:

In December 2012, the protestant white male in his late 50s from the South region of the United States, with a college or advanced degree education, would likely view the political and economic landscape with a mix of cautious optimism and skepticism. As a slightly conservative independent Republican who voted for Mitt Romney, he may feel disappointed by the re-election of Barack Obama. His warm feelings towards white people, big business, and the Republican Party contrast with his colder sentiments towards the Democratic Party, liberals, and Barack Obama. Given his belief in personal responsibility and mistrust of the government, he may feel that his economic situation is worsening and that the government is not effectively representing his interests. This individual may perceive the government as being influenced by big interests and view many public officials as corrupt. Despite his concerns about the state of the country, he remains engaged in politics and holds strong beliefs about individual initiative and self-reliance.

Figure C.2 – Interview Prompts

Interviewer:
What is your gender? Are you “male” or “female”?
You: male.
Interviewer: I am going to read you a list of four race categories. What race do you consider yourself to be? “White”, “Black”, “Asian”, or “Hispanic”?
You: white.
Interviewer: What is your age in years?
You: 59.
Interviewer: What is the highest level of school you have completed, or the highest degree you have received? Is it “high school”, “some college”, “a four-year college degree”, or “an advanced degree”?
You: college or advanced degree.
Interviewer: In what percentile is the total income in 2011 of all your family members living here before taxes?
You: 68 to 95 percentiles.
Interviewer: Do you consider yourself Protestant, Roman Catholic, Jewish or some other religion?
You: protestant.
Interviewer: Do you consider religion to be an important part of your life, or not?
You: yes.
Interviewer: Thinking about your life these days, do you ever attend religious services, apart from occasional weddings, baptisms or funerals?
You: no.
Interviewer: We would like to know if you are working now, or are you unemployed, retired, a homemaker, (a student), or what?
You: employed.
Interviewer: On a 'feeling thermometer' from 0 to 100, where warm feelings are between 50 and 100 and cold feelings are between 0 and 50, how would you rate the black people?
You: 70.
Interviewer: Still using the thermometer, how would you rate the white people?
You: 65.
Interviewer: Still using the thermometer, how would you rate the hispanic people?
You: 70.
Interviewer: Still using the thermometer, how would you rate the poor people?
You: 55.
Interviewer: Still using the thermometer, how would you rate big businesses?
You: 70.
Interviewer: Still using the thermometer, how would you rate labor unions?
You: 45.
Interviewer: Still using the thermometer, how would you rate liberals?
You: 21.
Interviewer: Still using the thermometer, how would you rate conservatives?

You: 80.
Interviewer: Still using the thermometer, how would you rate the democratic party?
You: 15.
Interviewer: Still using the thermometer, how would you rate Barack Obama?
You: 22.
Interviewer: Still using the thermometer, how would you rate the republican party?
You: 70.
Interviewer: Still using the thermometer, how would you rate Mitt Romney?
You: 82.
Interviewer: Do you think you can trust the government in Washington to do what is right?
You: yes.
Interviewer: Some people don't pay much attention to political campaigns. How about you, would you say that you have been/were very much interested, somewhat interested, or not much interested in the political campaigns this year?
You: very much interested.
Interviewer: Did you go to any political meetings, rallies, dinners, or things like that in support of a particular candidate?
You: no.
Interviewer: During an election year people are often asked to make a contribution to support campaigns. Did you give money to a political party or an individual candidate running for public office during this election year?
You: no.
Interviewer: Would you say the government is pretty much run by a few big interests looking out for themselves or that it is run for the benefit of all the people?
You: few big interests.
Interviewer: Do you think that quite a few of the people running the government are crooked, not very many are, or do you think hardly any of them are crooked?
You: quite a few.
Interviewer: 'Public officials don't care much what people like me think.' Do you agree, neither agree nor disagree, or disagree with this statement?
You: neither agree nor disagree.
Interviewer: '‘People like me don't have any say about what the government does.' Do you agree, neither agree nor disagree, or disagree with this statement?
You: neither agree nor disagree.
Interviewer: Do you think that people in the government waste a lot of money we pay in taxes, waste some of it, or don't waste very much of it?
You: a lot.
Interviewer: Do you agree or disagree with this statement: 'This country would be better off if we just stayed home and did not concern ourselves with problems in other parts of the world'?
You: disagree.
Interviewer: Some people feel that the government in Washington should see to it that every person has a job and a good standard of living. Suppose these people are at one end of a

scale, at point 1. Others think the government should just let each person get ahead on their own. Suppose these people are at the other end of the scale, at point 7. Some other people have opinions somewhere in between, at points 2, 3, 4, 5 or 6. Where would you place yourself on this scale?
You: 5.
Interviewer: There has been some discussion about abortion during recent years. Which one of these opinions best agrees with your view: By law, 'abortion should never be permitted', 'permitted only in case of rape, incest or when the life is in danger', 'permitted for other reasons', or 'always permitted'?
You: permitted in case of rape, incest, danger to life.
Interviewer: We hear a lot of talk these days about liberals and conservatives. When it comes to politics, do you usually think of yourself as extremely liberal, liberal, slightly liberal, moderate or middle of the road, slightly conservative or extremely conservative?
You: slightly conservative.
Interviewer: Which would you say best describes your partisan identification. Would you say you are a "strong democrat", "not very strong democrat", "independent, but closer to the Democratic party", "independent", "independent, but closer to the Republican party", "not very strong Republican", or "strong Republican"?
You: independent republican.
Interviewer: In 2012, who did you vote for?
You: Mitt Romney.
Interviewer: Would you say that you and your family are better off, worse off, or just about the same financially as you were a year ago?
You: worse now.
Interviewer: Now looking ahead–do you think that a year from now you and your family will be better off financially or worse off, or just about the same as now?
You: worse off.
Interviewer: Would you say that over the past year the nation's economy has gotten worse, stayed about the same, or gotten better? Answer with either 'gotten worse', 'stayed same' or 'better'.
You: ___________.

D. Robustness Tests: GPT-3.5 turbo and GPT-4 turbo

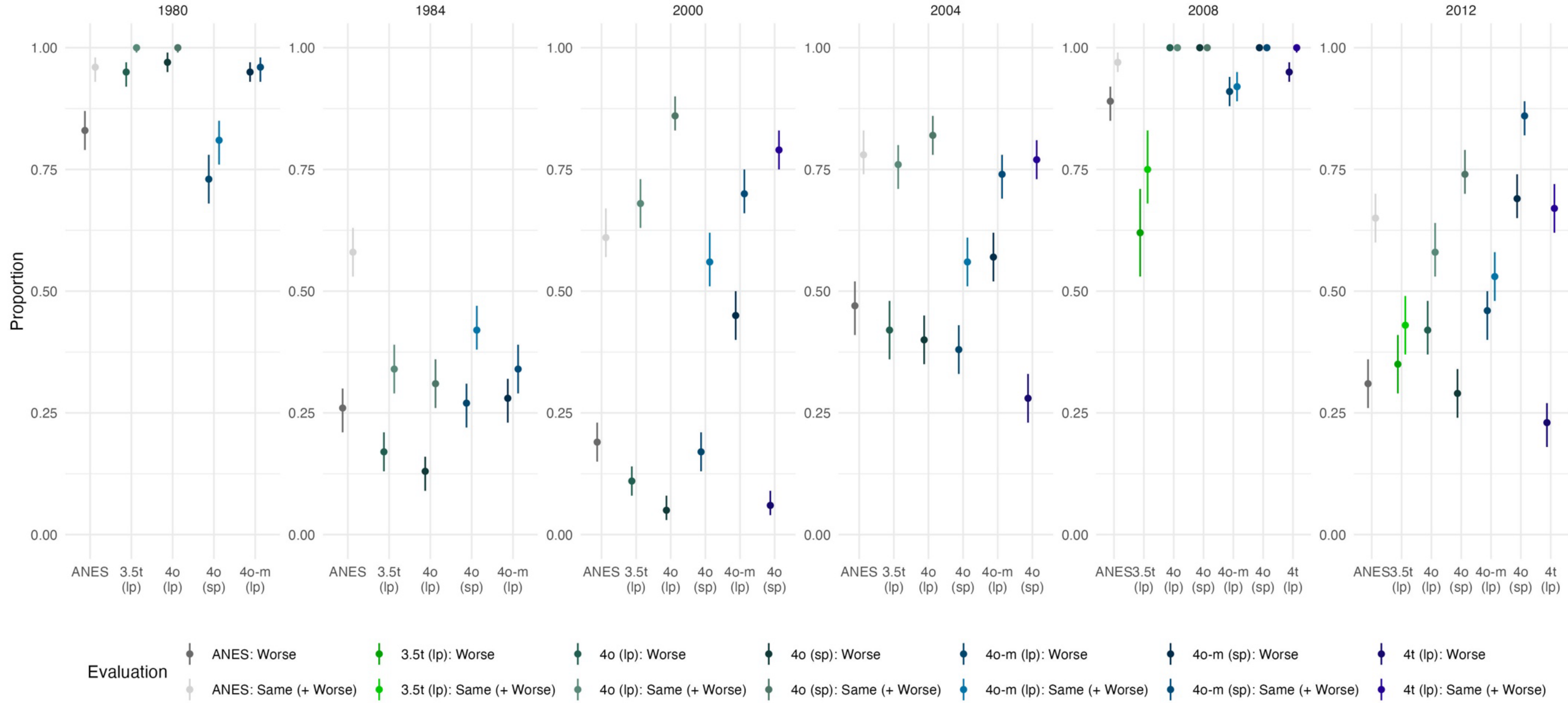
Tests with GPT-3.5 turbo and GPT-4 turbo were conducted with a different set of independent variables.

Table A.1 – Full List of Variables (GPT-3.5 turbo and GPT-4 turbo)

	1st Set
1	Region
2	Gender
3	Age
4	Race
5	Education
6	Family income
7	Employment
8	Working Hours Reduction
9	Religion
10	Church Attendance
11	Thermometer: Black
12	Thermometer: White
13	Thermometer: Hispanic
14	Thermometer: Poor
15	Thermometer: Gays and Lesbians
16	Thermometer: Big Businesses
17	Thermometer: Labor Unions
18	Thermometer: Liberals
19	Thermometer: Conservatives
20	Thermometer: Democratic Party
21	Thermometer: Democratic Party's Candidate (Presidential Election)
22	Thermometer: Republican Party
23	Thermometer: Republican Party's Candidate (Presidential Election)
24	Thermometer: Congress
25	Thermometer: Federal Government
26	Trust Fed. Government
27	Like-dislike Democratic Party
28	Like-dislike Republican Party
29	Interest in the Elections
30	Gov. run by few interests of for the benefit of all
31	Politicians corrupt
32	Gov. officials care about what people like [me] think
33	External Political Efficacy
34	Political Ideology
35	Partisanship
36	Vote
37	Government wastes tax-money
38	Expending: Poor
39	Expending: Child Care
40	Expending: Dealing with Crime

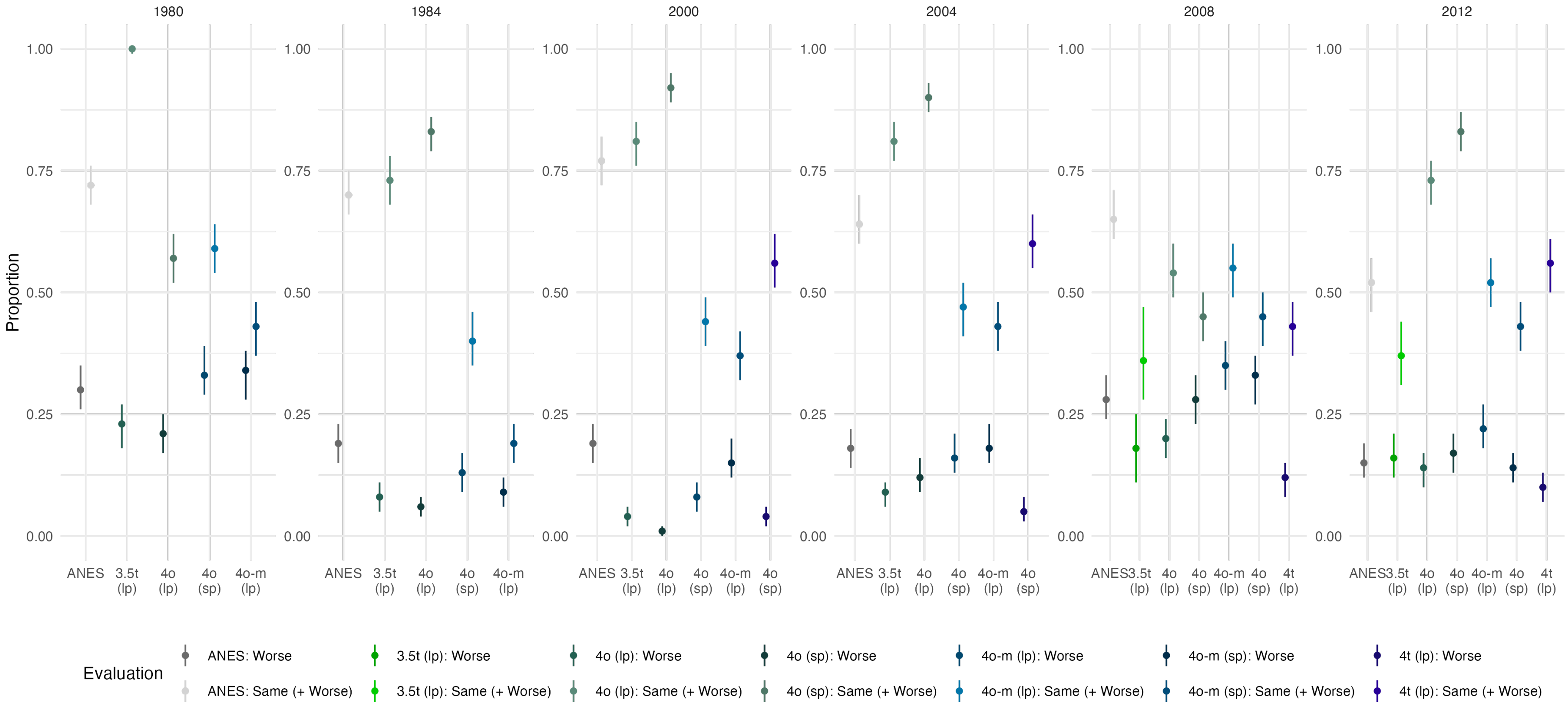
41	Expending: Public Schools
42	Expending: Welfare programs
43	Pocketbook Retrospective Economic Evaluation
44	Pocketbook Prospective Economic Evaluation

Figure D.1 – Prospective Economic Evaluation



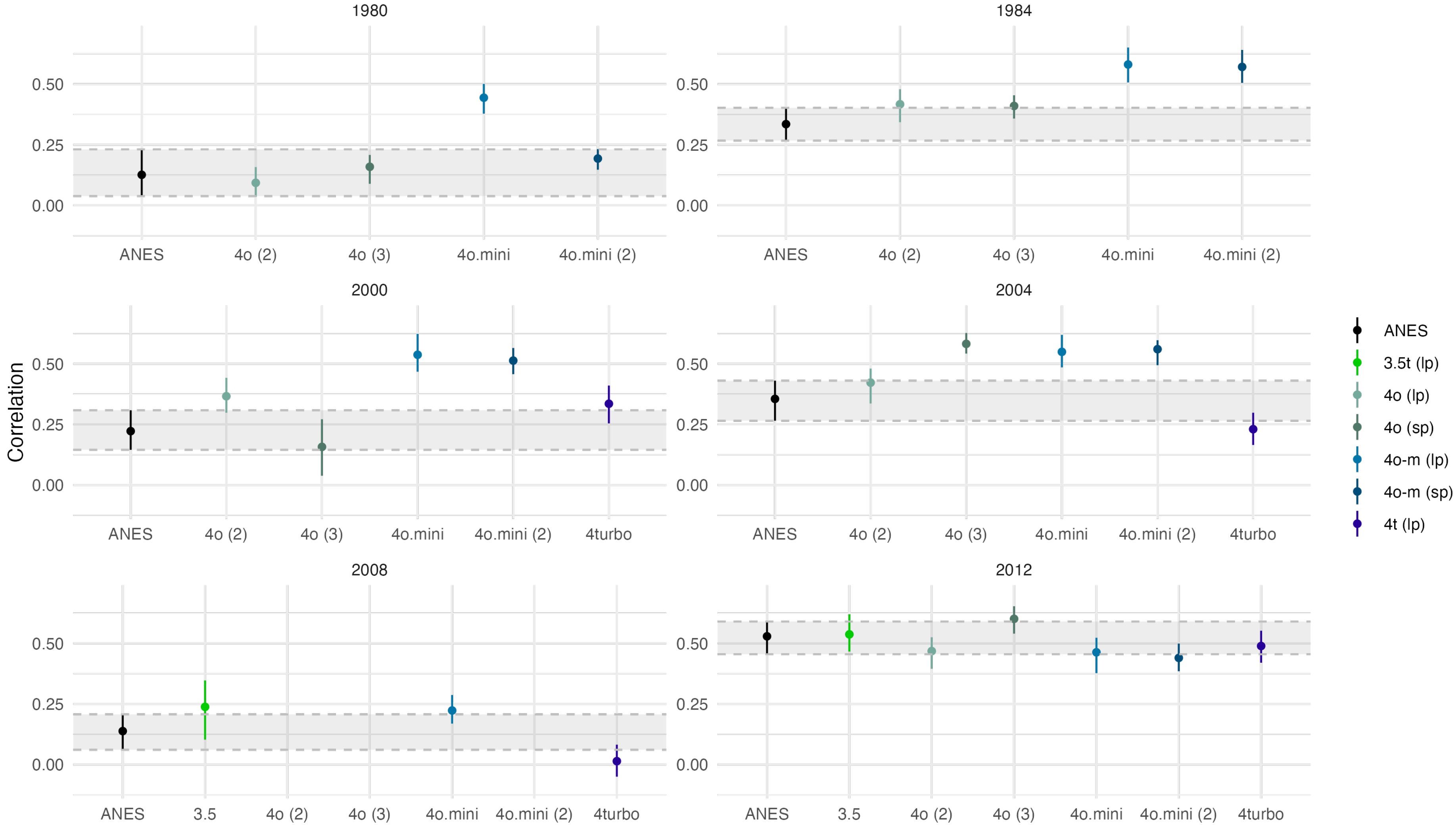
Note: Error bars are 99% bootstrapped confidence interval.

Figure D.2 – Prospective Economic Evaluation



Note: Error bars are 99% bootstrapped confidence interval.

Figure D.3 – Correlation Over Time: Retrospective vs Prospective Economic Evaluations



Notes: Spearman correlation. Error bars are 95% bootstrapped confidence intervals. Gray-shaded area depicts ANES' data 95% confidence interval. 3.5 is GPT-3.5 turbo, 4o(2) is GPT-4o lp (long prompt), 4o(3) is GPT-4o sp (short prompt), 4o.mini is GPT-4o mini lp, and 4o.mini (2) is GPT-4o mini sp.