
ECONtribute
Discussion Paper No. 410

Revealing Life Preferences Through LLMs

Omar Abdel Haq
Erzo F.P. Luttmer

Amitabh Chandra
Joshua Schwartzstein

Tomáš Jagelka

May 2026

www.econtribute.de



UNIVERSITÄT
ZU KÖLN

Revealing Life Preferences Through LLMs

Omar Abdel Haq Amitabh Chandra Tomáš Jagelka
Erzo F.P. Luttmer Joshua Schwartzstein *

April 30, 2026

Abstract

Large Language Models (LLMs) are trained on a prodigious corpus of human writing and may reveal human preferences over characteristics of life courses, such as income, longevity, and working conditions. We present OpenAI’s *GPT-5.4* and a broadly representative sample of Americans with pairs of life stories and ask them to choose the life they would prefer for themselves. A person’s choice is better predicted by the LLM’s choice than by another person’s choice over the same stories, and LLM valuations of several life attributes are similar to those derived from human responses. Our results suggest that LLM responses offer a scalable and cost-effective complement to existing methods for studying human preferences.

*Abdel Haq: Harvard Business School (email: oabdelhaq@hbs.edu); Chandra: Harvard Business School and Harvard Kennedy School (amitabh_chandra@harvard.edu); Jagelka: University of Bonn, Dartmouth College, and CREST-Ensaie (tjagelka@uni-bonn.de); Luttmer: Dartmouth College (email: erzo.fp.luttmer@dartmouth.edu); Schwartzstein: Harvard Business School (email: jschwartzstein@hbs.edu). For helpful comments, we thank Ben Bushong, Katherine Coffman, Sendhil Mullainathan, Paul Novosad, Ziad Obermeyer, Suproteem Sarkar, Andrei Shleifer, Adi Sunderam, Dmitry Taubinsky, and Hans-Joachim Voth, and participants at various seminars and conferences. We thank Isabel Galea, Alex Philip, and Julia Schwed for outstanding research assistance. This work is supported by Harvard Business School; the European Research Council (ERC) under the FELICITAS grant (No. 101165518); and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2126/2 – 390838866. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Understanding how people value the fundamental attributes of their lives, such as longevity, health, and working conditions, has long posed significant challenges across economics, psychology, and public policy. These *life preferences* shape consequential decisions about health, relationships, and careers. They determine whether rising health insurance premiums are worth their cost, how to conduct the welfare analysis of speed limits and environmental standards, and how to translate differences in life expectancy across income groups into equivalent differences in living standards. Despite their importance, life preferences are difficult to estimate. First, stated and revealed preferences often diverge: people may report that relationships matter more than income but nevertheless choose higher-paying jobs over more fulfilling ones, suggesting that the reliability of either method cannot be taken for granted when recovering underlying preferences. Second, we generally lack exogenous, well-measured variation in the prices of life attributes, making it difficult to recover life preferences from observed behavior. Hazardous jobs, for example, offer higher pay, but other unobserved differences between jobs make it hard to attribute wage premia specifically to mortality risk. Third, intention-action gaps caused by limited self-control or inattention, such as failing to wear a seatbelt while believing that life is priceless, create further wedges between observed choices and underlying preferences. Together, these challenges mean that revealed-preference approaches are not automatically superior to stated-preference ones: inferring longevity valuations from wage differentials in dangerous occupations, for instance, relies on populations with non-representative preferences and requires people to evaluate small changes in mortality risk, a task that behavioral evidence suggests humans do poorly. More broadly, they indicate the need for new methods to elicit life preferences.

This paper asks whether a large language model (LLM) can recover previously unknown human life preferences in a way that complements existing approaches to preference elicitation. We gave OpenAI’s *GPT-5.4* the identical choice task we gave a broadly representative sample of Americans: choosing between pairs of completed life stories that varied in income, longevity, and other life attributes. These life stories were designed to measure life preferences directly, eliciting preferences over entire life courses without requiring probability judgments. The human responses yield a previously undocumented distribution of income-longevity tradeoffs across the U.S. population (Jagelka [®] al., 2026). We conducted the LLM elicitation before those results were published: an LLM trained on existing research might reproduce known facts about human preferences by pattern-matching against published estimates, but it cannot recover estimates that did not yet exist.

We therefore ask whether LLMs encode something about what humans value that goes beyond what has been explicitly documented and published.

There are reasons to be optimistic. LLMs are trained on an extensive corpus of human writing, including literature, journalism, personal essays, and online discussion, and fine-tuned with human feedback. This corpus is rich with implicit information about what people value: which lives sound appealing, which tradeoffs feel fair, which circumstances seem enviable or pitiable. Insofar as human writing encodes these judgments, LLM training may allow models to extract preference-relevant structure even in domains where no explicit estimates exist (Horton et al., 2026; Asirvatham et al., 2026).

The competing view is that LLMs are “stochastic parrots,” models that mimic the statistical regularities of text without encoding anything about preferences (Bender et al., 2021). On this view, an LLM asked to choose between two life stories is doing something like identifying which story sounds more like a positive description of a life, without any representation of the tradeoffs humans make when they think carefully about what they value. The concern is not that LLMs will fail trivially. Any model that understands language will agree that winning the lottery beats being assaulted. The concern is that LLM responses may not reflect the quantitative structure of preferences: how much income people trade for longevity, how much they discount hardship at work, or how they value the attribute bundles of relationships, career, and community embedded in completed life stories. These are the objects our survey was designed to measure, and they are precisely the objects we ask LLMs to recover.

We propose a *revelation conjecture*: that an LLM, when appropriately prompted, can reveal human preferences over consequential life choices in a statistically informative manner. This conjecture makes no claim about whether LLMs understand human preferences in any deep sense. It asks only whether LLM choices, when prompted to act as humans, yield preference estimates that are statistically related to those derived from human responses. The conjecture generates hypotheses of varying stringency: weak tests ask whether LLMs predict the direction of majority preferences, while strong tests ask whether they recover dollar-valued tradeoffs. The revelation conjecture may hold for some populations, some preference domains, and some attributes but not others, so each hypothesis in this family is specific to a population, a set of scenarios, and an estimation method. Consequently, the strongest possible test, whether LLMs can reveal the preferences of specific individuals rather than population-level averages, lies beyond what we can establish.

We present three lines of evidence, ordered from least to most demanding. First, a person’s choice between two life stories is better predicted by the LLM’s choice than by

the choice of a different human respondent over the same two stories. This is a striking finding on its face, but it could partly reflect the LLM having lower noise than individual human respondents. We use a decomposition model to show that the result is more fundamental: after separately identifying human noise, preference heterogeneity, and LLM accuracy from four empirical moments, we estimate that the LLM correctly infers the majority human preference in roughly 86 to 92 percent of story pairs. Second, we estimate how much each of our 28 template life stories is worth as a bundle, expressed in income units, and find that human and LLM valuations of these attribute bundles are highly correlated, with slopes close to one. This extends the result from individual choices to quantitative valuations of rich story content: relationships, career trajectories, community involvement, and circumstances of death. Third, for the specific life attributes that we separately randomize within template stories, including income, longevity, hardship at work, good health, and era of birth, LLMs not only match the direction of human responses but also provide estimates of dollar-denominated tradeoffs that are close to those derived from human responses. For longevity and hardship at work, the two attributes estimated with sufficient precision to support meaningful comparisons, *GPT-5.4* with direct choice produces valuations within 50 percent of the human estimates, and we cannot reject equality. Taken together, the evidence supports the revelation conjecture: LLM output, when appropriately prompted, reflects not only which lives sound appealing but also the rates at which people trade one life attribute against another.

Understanding the quantitative nature of this result requires context. Recovering preferences over life attributes is not merely a matter of knowing that people prefer more income to less or longer lives to shorter ones. It requires knowing the marginal rate of substitution between them: how much income compensates for a year of life lost, which in turn requires forming implicit estimates of the relative marginal utilities of the two quantities. Economists identify this tradeoff from experimental variation in prices. The LLM was given no such variation during training, yet when exposed to it in our choice task, it responds in a way that implies a marginal rate of substitution close to the one humans reveal.

Our contribution connects several streams of research. A growing literature examines whether LLM responses can substitute for or complement human survey data (Argyle et al., 2023; Dominguez-Olmedo et al., 2024; Santurkar et al., 2023; Brand et al., 2025; Horton et al., 2026), with some work focused on whether LLMs replicate demographic differences in responses (Aher et al., 2023; Cheng et al., 2023). Ludwig et al. (forthcoming) caution that LLM-generated data should not substitute for human data without validation against

actual human responses, a requirement our design is built to satisfy. Our work differs from this literature in two ways. First, we focus on preferences over life attributes, including income, longevity, health, and working conditions, rather than political opinions or factual beliefs, where LLMs may have absorbed published survey results directly. Second, we validate against a novel survey whose results were not available online at the time of the LLM elicitation, making contamination from training data implausible. Some economists now use rich text data to “reveal the invisible” (Stantcheva, 2023); our work suggests that LLMs may accelerate this program, not as a substitute for human data, but as a complement that can extend preference elicitation to populations and domains where direct surveys are costly or infeasible.

The remainder of the paper proceeds as follows. Section 2 formalizes the revelation conjecture and describes the experimental design, including how we constructed the life stories, recruited and surveyed human respondents, and elicited choices from LLMs. Section 3 presents results in two parts: story-level correlations between human and LLM choices together with a decomposition model that separates noise, preference heterogeneity, and LLM accuracy; and valuations of specific life attributes that allow direct comparison of human and LLM preferences in dollar terms. Section 4 discusses implications and Section 5 concludes.

2 Methods

We elicit preferences by offering human survey respondents and LLMs a binary choice between two hypothetical lives, written as completed life stories. These stories were written by a research assistant by early November 2022 (before LLMs were widely available) and subsequently edited by us. Stories are typically three to five paragraphs long and give a bird’s eye view of someone’s life, conveying hobbies, work, friendships, family, and volunteer work. All stories include their protagonist’s name, gender, job, household income, marital and parental status, age at death, cause of death, timing of any terminal health condition, and information from which the year of birth can be inferred. An example story is shown in Appendix Figure A1, and a sample choice is shown in Appendix Figure A2.

2.1 The Revelation Conjecture

What do we mean by the conjecture that LLM choices, when appropriately prompted, usefully reveal human preferences? This section formalizes the idea within an inference framework. The key question is whether preferences estimated from LLM responses provide valuable information about human preferences in direction and magnitude—not whether LLM responses mimic the distribution of human responses, nor whether LLMs possess any underlying “understanding” of human preferences.

Each person i is endowed with preferences P_i^* over life courses, which are vectors of life attributes. For simplicity, we assume these preferences satisfy standard assumptions, including menu-independence (abstracting from issues raised by, e.g., Kőszegi and Rabin (2008) and Bernheim et al. (forthcoming)). For any scenario S_v consisting of two life courses L_{v1} and L_{v2} , P_i^* determines whether the person prefers L_{v1} over L_{v2} .

When a person is asked to make choices between richly described life courses in scenario S_v , her stated choice R_{iv} is a stochastic function of the scenario given her true preferences. Consider a set of scenarios \mathcal{V} that may vary some life attributes (e.g., longevity) while fixing others (e.g., country of birth), as well as a human (sub)population \mathcal{X} . Let R_{Human} denote the vector of choices made by representative members of \mathcal{X} on randomly selected scenarios from \mathcal{V} . Using estimation method $g(\cdot)$, the analyst estimates preferences $\hat{P}_{Human} = g(R_{Human})$ for a representative member of \mathcal{X} .

Suppose the analyst then prompts an LLM to act as a member of population \mathcal{X} and choose between life courses across the same scenarios, yielding choices R_{LLM} . Applying the same estimation method gives $\hat{P}_{LLM} = g(R_{LLM})$, the “preference” estimate that rationalizes the LLM’s choices.

The *revelation conjecture* for population \mathcal{X} on scenarios \mathcal{V} under estimation method $g(\cdot)$ is that knowledge of the LLM estimate provides statistical information about the human estimate: The revelation conjecture is that, at a minimum, the two estimates are not statistically independent; if the LLM perfectly reveals the life preferences of population \mathcal{X} , the two coincide in a statistical sense, i.e., $\hat{P}_{LLM} = \hat{P}_{Human}$, where equality is understood to hold up to sampling variation.

We next describe the experimental design used to test hypotheses derived from this conjecture for life preferences in a representative population of adult Americans.

2.2 Life Story Variation

We construct choice scenarios using a total of 28 template life stories, 14 with a female protagonist and 14 with a male protagonist. Stories are paired subject to three restrictions: both protagonists must be of the same sex, baseline income levels cannot differ by more than a factor of five, and no respondent sees the same story twice within a survey.¹

We independently randomize five attributes within the template stories of a scenario. The protagonist’s age at death is drawn uniformly between 60 and 90 years. Annual household income (rounded to the nearest thousand) is drawn uniformly between two-thirds and four-thirds of an occupation-specific baseline. The year of death is drawn uniformly from 2000 to 2019, with the year of birth determined jointly by longevity and the year of death.² Two story-specific elements — one describing workplace hardship such as stress, irregular hours, or career setbacks, and one stating that the protagonist was in good health prior to the onset of any terminal condition — are each included in a scenario with 25% probability, independently of one another; when included, each is assigned to one of the two paired stories at random.

Additional sentence-level randomizations at the story-pair level, including material and nonmaterial utility sentences and a filler sentence, are randomized orthogonally to the five attributes above and are not directly relevant to this study. Full details on all randomizations, examples of each sentence type, and an overview of the story construction process are provided in Appendix A.2. Ultimately, after applying sentence randomizations and numeric randomizations to the 28 template stories and randomly pairing two stories into a choice scenario, we end up using 21,270 distinct scenarios in the human choice experiment. We use the same scenarios for the LLM choice experiment.

2.3 Human Experiment Overview

Human choices were collected via a *Qualtrics* survey on the *Prolific* platform between March and April 2025. Respondents are U.S. citizens and residents with at least 100 completed surveys and an approval rating of $\geq 95\%$, with quotas ensuring equal proportions of female and male respondents.

¹Each respondent who completed the two survey waves (see Subsection 2.3) is presented exactly one repeated scenario from the first wave, providing a simple measure of the noisiness of human responses.

²We chose this range to ensure protagonists lived in the not-too-distant past while avoiding the most recent years, which risk evoking Covid-era associations or triggering memories of recently deceased loved ones.

Each respondent was presented with six choice scenarios. For each scenario, they chose between two life stories after being asked “Which life would you prefer for yourself?” (see Appendix Figure A3 for full instructions). We collected 3,746 completed surveys, of which 3,050 passed our inclusion criteria. Exclusions were based on two criteria: respondents who spent less than 30 seconds on any choice scenario (85% of exclusions) and respondents who failed checks designed to detect AI use (15% of exclusions).³ The median included respondent spent 2.2 minutes per choice scenario.

Respondents who completed the Wave 1 survey and met the inclusion criteria were invited to participate in Wave 2. The second survey closely followed the structure of the first, also eliciting six scenario choices from each respondent. We received responses from 2,441 respondents in Wave 2, of which 2,162 passed the inclusion criteria. In total, this yielded 31,272 scenario choices: 18,300 from Wave 1 and 12,972 from Wave 2.

To improve the representativeness of our results with respect to the broader U.S. adult population, we reweight respondents using inverse probability weights constructed by matching the Prolific sample to the March 2025 Current Population Survey on age, gender, race, education, and income (Flood et al., 2025). Respondents missing one or more of these variables or whose responses have no direct CPS analogue are excluded, yielding a weighted analysis sample of 2,907 respondents and 29,748 scenario choices. All results reported in the paper use this weighted sample. Details on the reweighting procedure and sample balance are provided in Appendix A.4.

2.4 LLM Elicitation Procedure

Our analysis focuses on `gpt-5.4-2026-03-05`, a snapshot of OpenAI’s *GPT-5.4*, but our methods are designed to be portable to any instruction fine-tuned model that can engage with a binary choice task. We focus on the *GPT-5.4* family for three reasons. First, it was OpenAI’s frontier model at the time of writing, and frontier models are the most informative test of the revelation conjecture: if any LLM encodes quantitative preference structure, it is most likely to be a model trained on the largest corpus with the most sophisticated fine-tuning. Second, OpenAI provides dated model snapshots via its API, which allows us to fix the model version and ensure that our results are reproducible; models accessed

³The AI-detection checks include: a two-part attention check in which respondents are instructed early on in the survey how to answer a question that appears much later (an LLM agent consistently failed this check during validation), an HTML question invisible to human respondents but visible to bots parsing the page directly, and a minimum threshold on within-respondent variation in response times. Details are provided in Appendix A.3.

without a dated snapshot may silently change between queries, introducing instability that would be difficult to diagnose. Third, we assess robustness across additional response sets from six predecessor and smaller OpenAI models, spanning earlier generations, smaller variants, and multiple reasoning thresholds of the same model family. This within-family variation allows us to ask whether the results depend on model scale, vintage, or reasoning depth, without conflating differences in model family with differences in training data, fine-tuning procedure, or output format. We restrict attention to instruction fine-tuned models throughout because fine-tuning is a methodological prerequisite: a base model may fail to engage with a binary choice format in a consistent and interpretable way.

Unlike human respondents, who saw six consecutive choice scenarios and selected stories by clicking, LLMs evaluate each scenario independently and choose between labeled options ('A' or 'B'). We collect LLM choices in two ways: directly, constraining output to a single token, and allowing the model to reason before making a final choice. We validate the elicitation procedure by testing sensibility and sensitivity during prompt development; details are provided in Appendix A.5.

3 Results

We now test hypotheses derived from the revelation conjecture. First, we compare respondents' choices to (i) choices made by other respondents who saw the same choice scenario, i.e., the same pair of life stories with the identical randomizations, and (ii) choices made by LLMs that received the same choice scenario. Then, we use a decomposition model to determine the degree to which the observed correlations reflect LLM "knowledge" of human preferences.

3.1 Correlations in Story-Level Choices

The first row of Table 1 shows that the correlation between the choices of different human respondents across choice scenarios is 0.22. Because choices are binary and, on average, each story is roughly equally likely to be chosen, this correlation has a simple interpretation: it equals the fraction of respondents who agree on a given scenario (61%) minus the fraction who disagree (39%).

The remaining columns show that the human-LLM correlation (0.33 for *GPT-5.4* with direct choice and 0.36 for *GPT-5.4* with a medium reasoning setting) exceeds the between-

person correlation. This is a striking finding. If we want to predict which life story a random respondent will choose, we do better by asking an LLM than by using the choice of another respondent.

3.2 Decomposing the Correlations

It would be wrong to conclude that LLMs know our preferences better than we do ourselves. First, the higher correlation for LLMs could reflect that LLM responses have less noise. Second, there is heterogeneity in human preferences whereas LLM choices may track the preferences of the typical American.

To disentangle noise from preference heterogeneity, we examine test-retest correlations (second row of Table 1): the correlation when the same respondent sees the same scenario again (four weeks later for humans; a new API call for LLMs). The human test-retest correlation is 0.49, implying that on average people choose the same story 74% of the time in repeated choices. LLM test-retest correlations are higher: 0.83 for *GPT-5.4* with direct choice and 0.88 for *GPT-5.4* with a medium reasoning setting. As such, less noise in LLM responses is part of the explanation for the high correlation between LLM and human choices relative to the correlation in choices between different people.

We use a simple decomposition model to quantify three sources of divergence between observed choices and the choices of a representative person in a deterministic setting: response noise, preference heterogeneity, and LLM knowledge of human preferences.⁴

The first step quantifies the role of noise in responses. We assume that respondents choose according to their true preferences a fraction of the time and randomly otherwise — a pattern that Belzil and Jagelka (2025) show can arise from an endogenous effort model. We assume a parallel structure for LLM noise, which arises from built-in randomness.⁵ The parameter “Random Response Probability” is a monotonically decreasing function of the test-retest correlation. Respondents behave as if they were answering randomly 30.1% of the time; this fraction is 8.8% for *GPT-5.4* with direct choice and 6.4% for *GPT-5.4* with a medium reasoning setting (Panel B of Table 1).

⁴We solve the model analytically given the four aforementioned correlations in the data. See Appendix Section A.6 for details.

⁵LLM noise results from the model’s “temperature,” kept at the default value of 1. Some residual noise would remain even at minimum temperature due to parallel processing and floating-point precision (see, e.g., Yuan et al., 2025).

Table 1: Noise, Heterogeneity, and LLM Accuracy

	Human	OpenAI <i>GPT-5.4</i>	
	Stated Choice (1)	Direct Choice (2)	Reasoned Choice (3)
<i>Panel A: Correlations</i>			
<i>Same-Scenario Agreement</i>			
Correlation with Random Person (Standard Error) [N]	0.217 (0.043) [7,126]	0.326 (0.007) [24,612]	0.356 (0.007) [24,612]
<i>Repeat-Scenario Consistency</i>			
Test-Retest Correlation (Standard Error) [N]	0.489 (0.019) [2,051]	0.832 (0.012) [2,051]	0.876 (0.011) [2,051]
<i>Panel B: Implied Structural Parameters</i>			
<i>Noise Rate</i>			
Random Response Probability (Standard Error)	0.301 (0.014)	0.088 (0.007)	0.064 (0.006)
<i>Preference Heterogeneity</i>			
Average Minority Share (Standard Error)	0.204 (0.036)	—	—
<i>LLM Accuracy</i>			
Majority Following Rate (Standard Error)	—	0.862 (0.118)	0.918 (0.125)

Notes: This table reports empirical correlations and the structural parameters they imply under the decomposition model described in Appendix A.6. Column (1) uses choices from human respondents on the *Prolific* platform, reweighted using inverse probability weights to match the March 2025 Current Population Survey; see Appendix A.4 for details. Columns (2) and (3) use choices elicited from the LLM listed in the column header.

The test-retest correlation is computed over the choice scenarios for which a respondent saw the same scenario in both waves, with the LLM queried independently each time. The correlation with a random person is computed over observations from respondents who completed both survey waves: for humans, it is the correlation between two distinct respondents on the same scenario; for LLMs, it is the correlation between the LLM’s choice and the corresponding human’s choice on the same scenario.

The model assumes latent preferences follow a symmetric Beta distribution across scenarios — with estimated heterogeneity parameter 0.62 — and characterizes each respondent and LLM by a noise rate and the LLM additionally by an accuracy rate reflecting how often its non-random response matches the latent majority preference of human respondents. Standard errors are computed via a respondent-level cluster bootstrap with 100,000 replications and 2,051 respondent clusters; see Appendix A.7.1 for details.

The second step quantifies preference heterogeneity, which must be present because the human test-retest correlation (0.49) exceeds the across-person correlation (0.22). We express the amount of heterogeneity by the minority-preference share: the average fraction of respondents whose true preference for a given life story in a scenario differs from that

of the majority, after filtering out noise.⁶ We find an average minority share of 0.20 — that is, on average 80% of respondents truly prefer one life story in a given scenario while 20% prefer the other. Due to noise, observed shares are closer to chance: 71% and 29%.

In the final step, we ask: if LLMs could perfectly infer the majority’s true preference, what correlation with human choices would we expect, given our estimates of noise and heterogeneity? Based on our model, this hypothetical correlation is 0.38 for *GPT-5.4* with direct choice and 0.39 for *GPT-5.4* with a medium reasoning setting. Actual correlations are slightly lower, indicating imperfect inference. We estimate the probability that the LLM correctly infers the majority human preference at 86% for *GPT-5.4* with direct choice and 92% for *GPT-5.4* with a medium reasoning setting. However, we cannot reject that the majority following rate is 100% (p-values are 0.241 and 0.512, respectively). We next ask whether LLM revelation extends to people’s valuations of bundles of life attributes, captured by our 28 template life stories, as well as to specific life attributes that we randomize independently.

3.3 Valuations of Life Attributes

The previous section treated each life story — including its specific randomizations such as income and age of death — as an indivisible whole. It described the similarity of *choices* and interpreted the correlations using a decomposition model, recovering preferences over specific choice scenarios rather than over attributes in the stories featured in the scenarios. We now open the black box by distinguishing the attribute bundles implicit in each of the 28 template stories from the attributes that we explicitly randomized within each template story. This distinction allows us to estimate a choice model over both attributes and attribute bundles, and to compare the implied preferences for each.

These attribute bundles include relationships, family, hobbies, passions, career, location, community involvement, and cause of death. With only 28 templates we cannot credibly decompose the bundles, but we can test whether humans and LLMs value them similarly. We also examine whether LLMs and humans respond in the same direction and at similar magnitudes to the specific life attributes that we explicitly randomized, such as income, longevity, hardship at work, good health, and era of birth. Because we experimentally vary these attributes within stories, we can credibly estimate how much weight each attribute receives in choices and express preferences as dollar-denominated tradeoffs that

⁶Because the minority share varies across scenarios, we model it as a draw from a symmetric Beta distribution.

are directly relevant for economic analysis.

Table 2 reports seemingly unrelated regressions of how respondents and LLMs respond to the five life attributes randomized within each template story: annual household income, longevity, hardship at work, good health, and era of birth. Without loss of generality, one life story is labeled the “reference story”; the dependent variable is an indicator for whether it was chosen. The explanatory variables are the differences between the reference story and the alternative in log income, log longevity, a hardship-at-work indicator, a good-health indicator, and year of birth. We also include an indicator for whether the reference story was shown first, to capture order effects. Story fixed effects ensure that coefficients on specific life attributes are identified from within-story variation, not from features of the template story (names, career path, family relationships, etc.). All coefficients and standard errors are normalized by the human coefficient, so the human column reports ones and the LLM columns show responsiveness relative to human respondents.

Table 2 yields three findings. First, all LLM coefficients on life attributes are positive, meaning LLMs match the direction of human responses for the five randomized life attributes. However, for biases in choice, such as whether to choose the first-shown story in a scenario, LLM responses do not necessarily go in the same direction as human responses. Second, the standard errors on all LLM responses to life attributes are smaller than the standard errors on the corresponding human estimates. This helps explain why the LLM estimates are significant for all three life attributes with a significant human coefficient. Moreover, the LLM estimates are even significant in three out of four cases for the two life attributes with an insignificant human coefficient. Third, the adjusted R^2 of both LLM equations substantially exceeds that of humans (0.52 and 0.59 versus 0.20), consistent with less noise and no preference heterogeneity across LLM responses. Next, we compare the magnitude of preferences implied by the coefficient estimates of Table 2.

To examine *preferences* rather than choice responsiveness, we need to calculate tradeoffs between life attributes. Specifically, we divide the coefficient on each non-income attribute by the income coefficient, yielding money-metric valuations.⁷ Similarly, we turn fixed-effect estimates into preferences by dividing by the coefficient on log income, which yields “fixed-effect valuations.” For example, a fixed-effect valuation of 0.5 means the attribute bundle embedded in the text of a template story is valued as much as 50 log points of income relative to the attribute bundle in the average story (given that we normalized the average fixed effect to zero).

⁷If, say, 20% of respondents randomly chose stories, coefficients would be attenuated by 20%. Ratios of coefficients are immune to such attenuation.

Table 2: Predictors of Story Choice: Human Respondents versus LLMs

	Human	OpenAI GPT-5.4	
	Stated Choice (1)	Direct Choice (2)	Reasoned Choice (3)
Δ Log Income (normalized by human coefficient)	1.00*** (0.14)	1.27*** (0.10)	0.89*** (0.11)
Δ Log Longevity (normalized by human coefficient)	1.00*** (0.06)	1.84*** (0.04)	2.02*** (0.04)
Δ Hardship-at-Work Sentence (normalized by human coefficient)	1.00*** (0.20)	0.87*** (0.13)	1.56*** (0.14)
Δ Good-Health Sentence (normalized by human coefficient)	1.00 (1.37)	4.69*** (1.21)	3.59*** (1.13)
Δ Year of Birth (normalized by human coefficient)	1.00 (0.69)	2.37*** (0.51)	0.66 (0.45)
Reference Story Shown First (normalized by human coefficient)	1.00*** (0.23)	-5.15*** (0.18)	1.68*** (0.16)
Story Fixed Effects	Yes	Yes	Yes
N	29,748	29,748	29,748
Adjusted R^2	0.20	0.52	0.59

Notes: Without loss of generality, we refer to one of two stories as the reference story and the other as the alternative story. This table presents regression coefficients where the dependent variable is a binary indicator equal to one if the reference story is selected as the preferred life. Column (1) uses choices from human respondents on the *Prolific* platform, reweighted using inverse probability weights to match the March 2025 Current Population Survey; see Appendix A.4 for details. Columns (2) and (3) use binary choices elicited from the LLM listed in the column header, presented with the same choice scenarios as shown to human respondents. All equations are estimated jointly via Seemingly Unrelated Regression using the Moore–Penrose pseudoinverse with sampling-weighted clustered sandwich standard errors (2,907 respondent clusters); see Appendix A.7.2 for details.

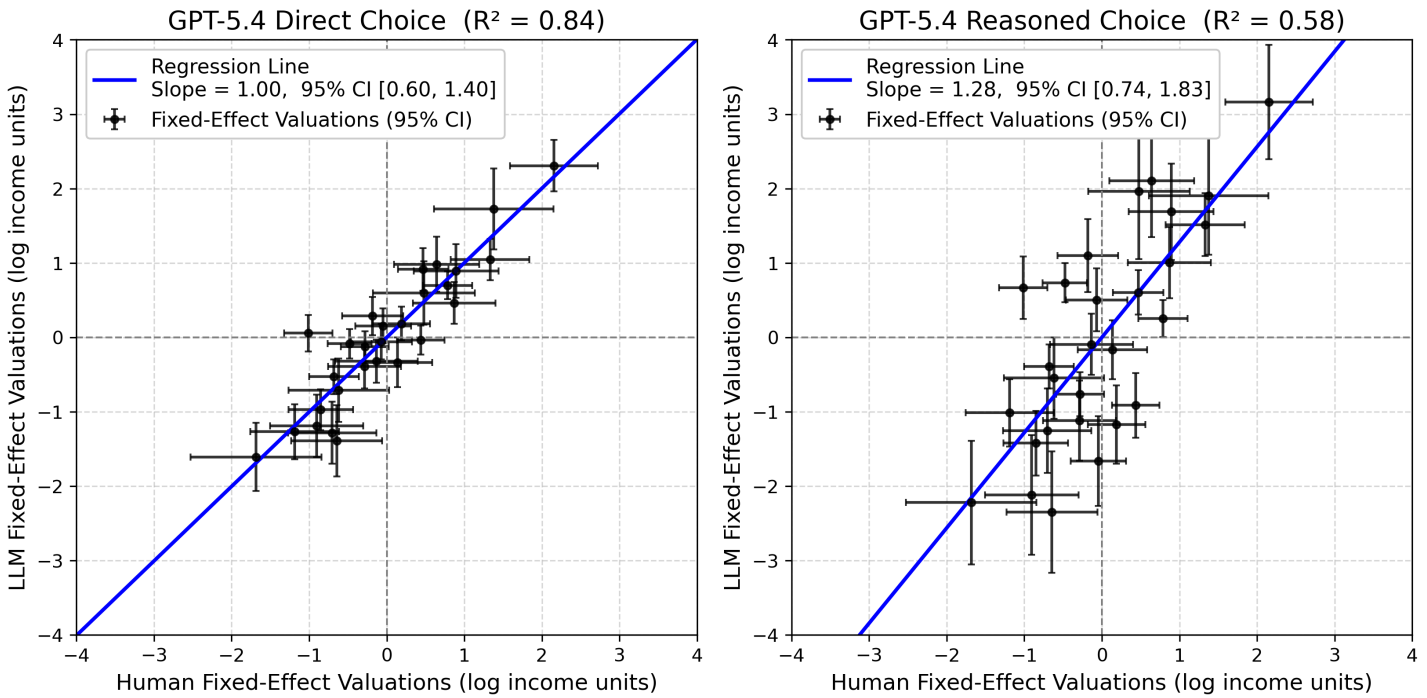
All reported coefficients and standard errors are normalized by the corresponding point estimate from column (1), so that every entry in column (1) equals one by construction. Statistical significance at the 10%, 5%, and 1% level is denoted by *, **, and ***, respectively.

We first examine preferences over the attribute bundles implicit in each story template, as measured by fixed-effect valuations. Because these bundles are arguably more representative of important life attributes than the specific life attributes that we will analyze below, they offer a more general test of the revelation conjecture. We leverage the fact that these template stories were written by a research assistant and designed to feel natural to readers and to encompass a wide variety of life attributes such as family relationships, community involvement, and career paths.

In Figure 1, we plot the estimated fixed-effect valuations for each template story from the regressions in Table 2. Story fixed-effect valuations from the LLM models are on

the y -axis, and those from the human model are on the x -axis. Since every fixed-effect valuation is estimated, there are standard errors for both (reflecting statistical uncertainty in both the fixed effect and the income coefficient). The left panel plots the fixed-effect valuations from the *GPT-5.4* with direct choice regression, and the right panel uses the *GPT-5.4* with reasoned choice regression.

Figure 1: Comparison of Human & LLM Story Fixed-Effect Valuations



Notes: This figure is derived from the same weighted SUR reported in Table 2. For each equation, each story’s fixed effect is divided by that equation’s log-income coefficient, yielding a valuation expressed in income units. Because the Moore–Penrose pseudoinverse imposes a sum-to-zero constraint on the fixed effect coefficients, the regression line passes through the origin by construction. Human story valuations are plotted on the x -axis and LLM story valuations on the y -axis. The left panel plots valuations from the human and OpenAI *GPT-5.4* with direct choice equations; the right panel plots valuations from the human and OpenAI *GPT-5.4* with reasoned choice equations. Error bars show 95% confidence intervals derived via the delta method. The regression slope and its standard error are likewise computed via the delta method, with uncertainty propagated through the full cross-equation SUR covariance matrix.

Figure 1 shows that story fixed-effect valuations are highly correlated between LLMs and human respondents, with R^2 statistics of 0.84 and 0.58. Hence, there is a high correlation between life stories that were more desirable to human respondents and those more likely to be chosen by LLMs, holding randomized within-story attributes such as income and longevity fixed. If LLMs and humans valued life attributes equally, the slope of the regression line would be 1. The point estimate of the slope for *GPT-5.4* with direct choice is indeed 1.00, though this precise alignment is likely coincidental given estimation noise. Indeed, for *GPT-5.4* with reasoned choice the slope is 1.28. The important lesson from

this figure is that both slopes are significantly different from 0 and neither is significantly different from 1 ($p = 0.99$ and $p = 0.31$, respectively). The visual and quantitative similarity of LLM and human valuations across the 28 template stories supports the revelation conjecture.

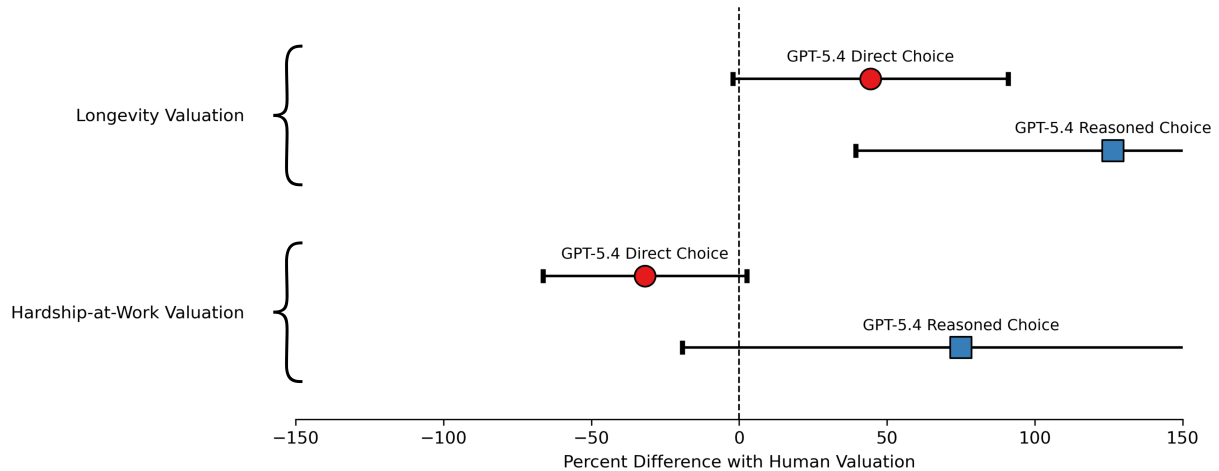
Finally, we examine the degree to which LLM valuations of specific life attributes correspond to human valuations. As shown in Table 2, human choices responded significantly to three life attributes: income, longevity, and hardship at work. Given that we use income to turn responsiveness into valuations, we can estimate human valuations with reasonable precision only for longevity and hardship at work. Panel A of Figure 2 reports the percent difference between LLM and human valuations with 95% confidence intervals for these two attributes, with the confidence intervals reflecting both the uncertainty in the estimated human valuations and the estimated LLM valuations. The valuations of *GPT-5.4* with direct choice lie within 50% of the human valuations. However, for these two life attributes, *GPT-5.4* with reasoned choice shows differences with human valuations that range between 50% and 150%, and with the confidence interval excluding zero for the longevity valuation. For the two attributes in Panel B (era of birth, good-health indicator), human responses are too noisy to yield informative comparisons (as shown in Table 2).

Overall, we find that all estimates of *GPT-5.4* with direct choice are consistent with the strongest test of the revelation conjecture, establishing the existence of an LLM that makes choices that quantitatively replicate human preferences for a range of life attributes. The performance of the *GPT-5.4* with reasoned choice was more varied.⁸ While it outperformed direct choice in predicting the choices of human respondents across life stories, it was not as accurate in estimating human valuations of life attributes, and we could reject equal valuations in one case. One interpretation is that reasoning helps the model focus on the most salient attributes in a story, improving its ability to predict choices. But for valuations — which require trading off attributes like income and longevity that are always present but vary in degree — reasoning may be less helpful. Such tradeoffs may rely more on intuition and feeling than on deliberation, which is why we used the life story framework to infer people’s preferences in the first place, rather than presenting them with direct tradeoffs between life attributes.

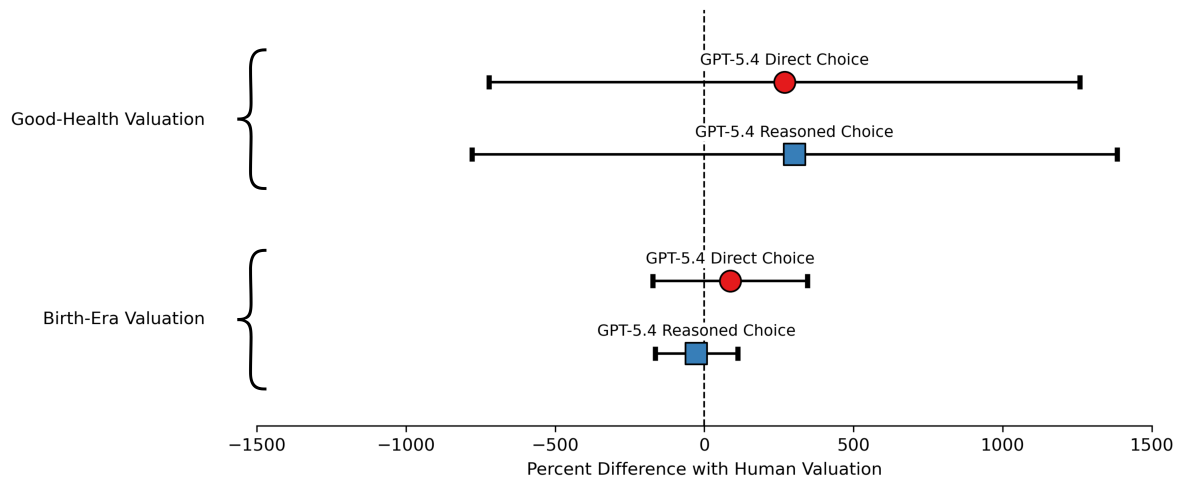
⁸Results for six predecessor and smaller OpenAI models are reported in Appendix A.8. Performance varies across models: some, such as *GPT-4.1* and *o3*, yield results broadly similar to *GPT-5.4*, while the smaller *GPT-4o mini* and *GPT-4.1 mini* models do not reveal human preferences as well.

Figure 2: Differences between Human and LLM Estimates of Life Preferences

Panel A: Differences with Precisely Estimated Human Valuations



Panel B: Differences with Imprecisely Estimated Human Valuations



Notes: Each marker shows the percentage by which the valuation implied by LLM choices differs from the corresponding human valuation, where each valuation is defined as the ratio of the attribute’s regression coefficient to the log-income coefficient. All estimates are derived from the same weighted SUR reported in Table 2, with confidence intervals and standard errors computed via the delta method propagated through the full cross-equation SUR covariance matrix. Panel A reports comparisons for attributes for which the human valuation is precisely estimated; Panel B reports comparisons for attributes for which the human valuations are imprecisely estimated, resulting in percentage differences with wide confidence intervals.

If we provided LLMs with information about human respondents, would they be better able to recover their preferences? In principle, such information should be helpful. In practice, however, it is not clear that LLMs would learn the mapping between a person’s characteristics and their preferences when many texts that LLMs were trained on provided

little or no information about the characteristics of the writer.

To answer this question, we re-ran the main analyses after providing the LLM with (i) the respondent’s age, gender, marital status, and number of children, or (ii) those demographics plus the respondent’s answers to an eight-item questionnaire on attitudes toward longevity, finances, work, hardship, risk, altruism, time preferences, and material-comfort tradeoffs. Neither modification meaningfully improved preference revelation. Details on this exercise can be found in Appendix A.9.

4 Discussion

We find that LLMs can *quantitatively* reveal human preferences over the quality-of-life attributes that we varied in our survey of adult Americans. This revelation conjecture is supported by three results: LLM choices predict individual human choices better than another respondent’s choices do, LLM valuations of attribute bundles scale roughly one-to-one with human valuations, and LLM estimates of dollar-denominated tradeoffs between specific life attributes are generally within the confidence intervals of those derived from human responses. Crucially, the LLM produced these results before our survey instruments or data were published, ruling out the possibility that it simply absorbed findings from the literature.

LLMs may provide complementary signals even when human data are available, adding precision to estimates from human responses or yielding priors for Bayesian estimation. LLMs do not tire and can produce estimates at a granularity infeasible for human respondents. Appendix A.5.1 illustrates this by mapping indifference regions in income–longevity space through repeated LLM queries — a task prohibitively expensive with humans.

More broadly, the way people write about their experiences may sometimes be more informative about life preferences than observed behavior, which is subject to temptation, present bias, and inattention. Fiction in particular is a powerful form of social and emotional simulation: it lets people explore the complexities of life, witness diverse relationships, grapple with definitions of success, and confront mortality through the eyes of others. Whether it is Jane Austen’s portraits of marriage and social ambition, Reddit threads about career regret, or obituaries that distill a life into its most valued elements, the training corpus contains an implicit preference-relevant structure that no single survey could capture.

LLM revelations derive their credibility from validation against human data. We cannot yet be certain whether LLMs recover life preferences or replicate systematic biases that masquerade as preferences (Mullainathan and Obermeyer, 2017). To shed light on this issue, researchers could examine settings where the degree of bias in human responses is understood, for example where optimal decision rules are known but heuristic use is frequent (Mu et al., 2025), to describe circumstances under which LLM responses reflect life preferences more accurately than human choices.

5 Conclusion

The estimation of life preferences has long been hindered by the noise and various biases contained in data from surveys, experiments, and observed behavior. We show that LLMs, trained on a vast corpus of human writing, can help. The three results laid out above — choice prediction, bundle valuation, and attribute-level tradeoffs — all support the revelation conjecture for life preferences held by adult Americans. LLM choices are also less noisy than human choices and more responsive to variation in life attributes, suggesting they may add precision to estimates from human responses even when such data are available.

Our findings have practical implications for researchers. First, LLMs can serve as cheap and scalable pilots for survey instruments. Before investing in a large representative survey, researchers can use LLMs to screen stimuli, estimate likely effect sizes, and identify which attributes matter enough to randomize, at negligible cost relative to a Prolific study. Second, LLMs may allow preference elicitation in domains where directly canvassing human subjects is costly, ethically challenging, or simply impossible. Populations that left written traces but cannot be surveyed, including historical cohorts, non-internet-connected communities, or people in settings where sensitive topics cannot be raised directly, may nonetheless be represented in the corpus on which LLMs were trained. Third, LLM reasoning traces offer a window into what drives choices that human response data alone cannot provide: which attributes capture attention, how tradeoffs are framed, and where deliberation departs from intuition. Together, this suggests that LLMs are a complement to existing methods of estimating life preferences, one that is faster, cheaper, and scalable to settings where direct elicitation is costly or infeasible.

Our results suggest that LLMs can be used to recover quantitative information on human preferences. Future research should explore which parts of the human-written

corpus on which LLMs are trained are especially useful for revealing preferences, and to what extent post-training facilitates this process. It is also important to understand whether and how LLMs can be prompted to reveal information on the preferences of particular groups of people.

References

- Aher, Gati V., Rosa I. Arriaga, and Adam T. Kalai (2023) “Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies,” in *Proceedings of the 40th International Conference on Machine Learning*, 202, 337–371, Proceedings of Machine Learning Research.
- 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*, 31 (3), 337–351.
- Asirvatham, Hemanth, Elliott Moksiki, and Andrei Shleifer (2026) “GPT as a Measurement Tool,” NBER Working Paper 34834, National Bureau of Economic Research.
- Belzil, Christian and Tomáš Jagelka (2025) “Separating Preferences from Endogenous Effort and Cognitive Noise in Observed Decisions,” IZA Discussion Paper 18315, Institute of Labor Economics (IZA).
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021) “On the Dangers of Stochastic Parrots: Can Language Models be too Big?,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
- Bernheim, B. Douglas, Kristy Kim, and Dmitry Taubinsky (forthcoming) “Welfare and the Act of Choosing,” *Journal of Political Economy*.
- Brand, James, Ayelet Israeli, and Donald Ngwe (2025) “Using LLMs for Market Research,” *Harvard Business School Marketing Unit Working Paper* (23-062).
- Cheng, Myra, Tiziano Piccardi, and Diyi Yang (2023) “CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10853–10875: Association for Computational Linguistics.
- Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner (2024) “Questioning the Survey Responses of Large Language Models,” *Advances in Neural Information Processing Systems*, 37, 45850–45878.
- Flood, Sarah, Miriam King, Renae Rodgers et al. (2025) “IPUMS CPS: Version 13.0 [dataset],” IPUMS, Minneapolis, MN.
- Horton, John J., Apostolos Filippas, and Benjamin S. Manning (2026) “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” NBER Working Paper 31122, National Bureau of Economic Research.
- Jagelka, Tomáš (r) Erzo F.P. Luttmer (r) Joshua Schwartzstein (r) Amitabh Chandra (2026) “Living Large or Long? Preference Estimates from Completed-Life Stories,” Unpublished Manuscript.

- Kőszegi, Botond and Matthew Rabin (2008) "Choices, Situations, and Happiness," *Journal of Public Economics*, 92 (8-9), 1821–1832.
- Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (forthcoming) "Large Language Models: An Applied Econometric Framework," *Annual Review of Economics*.
- Mu, Tianshi, Pranjal Rawat, John Rust, Chengjun Zhang, and Qixuan Zhong (2025) "Who is More Bayesian: Humans or ChatGPT?," <https://arxiv.org/abs/2504.10636>.
- Mullainathan, Sendhil and Ziad Obermeyer (2017) "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review*, 107 (5), 476–480.
- Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto (2023) "Whose Opinions do Language Models Reflect?," in *Proceedings of the 40th International Conference on Machine Learning*, 202, 29971–30004, Proceedings of Machine Learning Research.
- Stantcheva, Stefanie (2023) "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible," *Annual Review of Economics*, 15 (1), 205–234.
- Yuan, Jiayi, Hao Li, Xinheng Ding et al. (2025) "Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference," <https://arxiv.org/abs/2506.09501>.

Appendix

A.1 Elicitation Instruments

Figure A1: Sample Story Text

Mia Cassopolis was born in Knoxville, Tennessee in x . From nearly the time she could walk, Mia was playing and singing country and folk music with her dad and brothers. On the weekends, her family would travel to Nashville to listen to new artists and sometimes even play their own music at small cafes.

Mia's mom died when she was 4 years old, just after her youngest brother was born. With just her dad working as a musician, the family struggled to make ends meet. Mia's dad would often play shows, and, as the kids got older, he would sometimes go on tour with his band.

Upon graduating high school, Mia and her brothers decided to continue their father's passion and form their own band. Their band toured parts of the South and rural East, playing original folk music. Typically, Mia earned an annual income of $\$y$ playing music with her brothers.

The siblings loved to sit around together with their instruments for hours, playing their old favorites, or working out new material. They always jumped at the chance whenever local musicians would invite them to jam together. Mia's other passion was cowboy boots. She visited shops everywhere they went and was always adding to her collection. She didn't hesitate to splurge on unique pairs, often spending a significant portion of her tour earnings on rare finds. Life on the road had its challenges, including a lack of privacy for Mia and tense arguments with her siblings about the band's musical direction. Mia's single-minded devotion to her band and siblings guided her life and every decision; she never married or had children.

As Mia got older, she played fewer shows of her own and started traveling to attend shows of her favorite bands. On one of her trips, Mia experienced a series of intense headaches and fevers. This was unusual for Mia, who usually took great care of herself while on the road and was seldom ill. She quickly went to the doctor and was diagnosed with bacterial meningitis. Four days after the diagnosis, at the age of z , Mia passed away.

Notes: In this sample story, x denotes the protagonist's birth year, y denotes annual income, and z denotes longevity, all of which vary across elicitation. The respondent is shown actual values for x , y , and z .

Figure A2: Sample Choice Scenarios

Which life would you prefer for yourself?

In 1956, at the age of 24, Ryan Michaelson was accepted to Moody Theological Seminary, located in Chicago, Illinois. Moving across the country from Boulder, Colorado with his wife, Melanie, Ryan enrolled in Moody's master's program with the goal of becoming an ordained minister.

After completing the program, Ryan found work at a church back home in Boulder, where he initially came on staff as an associate pastor. Melanie also worked part-time at the church as a family counselor. In a short time, Ryan was promoted to head pastor and Melanie began to run the church's counseling program. A week after Ryan's promotion, a malfunction in the church's lighting caused a small electrical fire, and the Michaelsons felt compelled to personally cover the repair costs. The work as head pastor was rewarding, though Ryan struggled to balance the needs of his community with the needs of his family and his own personal needs. Over the course of their careers, the Michaelsons generally earned a joint income of \$124,000 a year. Ryan and Melanie had hoped to adopt a child, but after years of wading through bureaucracy, which put financial and emotional strain on both of them, they withdrew from the process.

During the little time they had away from work, Ryan and Melanie would usually spend their free Saturday mornings mountain biking. Ryan kept a guidebook of trails near the door to the house so they could always be ready to explore someplace new. Upon returning from their rides, they would typically enjoy lunch at one of the many highly acclaimed restaurants in downtown Boulder.

After several decades, however, Ryan had to step down from his position when he was diagnosed with pancreatic cancer. He underwent extensive radiation therapy, but the treatment had little impact. Three months after his diagnosis, with his wife by his side, Ryan passed away at age 79.

Mark Cassopolis was born in Knoxville, Tennessee in 1934. From nearly the time he could walk, Mark was playing and singing country and folk music with his dad and brothers. On the weekends, his family would travel to Nashville to listen to new artists and sometimes even play their own music at small cafes.

Mark's mom died when he was 4 years old, just after his youngest brother was born. With just his dad working as a musician, the family struggled to make ends meet. Mark's dad would often play shows, and, as the kids got older, he would sometimes go on tour with his band.

Upon graduating high school, Mark and his brothers decided to continue their father's passion and form their own band. Typically, Mark earned an annual income of \$83,000 playing music with his brothers.

The siblings loved to sit around together with their instruments for hours, playing their old favorites, or working out new material. Mark's other passion was cowboy boots. He visited shops everywhere they went and was always adding to his collection. Life on the road had its challenges, including a lack of privacy for Mark and tense arguments with his siblings about the band's musical direction. Mark's single-minded devotion to his band and siblings guided his life and every decision; he never married or had children.

As Mark got older, he played fewer shows of his own and started traveling to attend shows of his favorite bands. On one of his trips, Mark experienced a series of intense headaches and fevers. He quickly went to the doctor and was diagnosed with bacterial meningitis. Four days after the diagnosis, at the age of 73, Mark passed away.

Notes: Qualtrics survey respondents evaluate six scenarios of the form and layout shown above, with the story texts serving as buttons. The two stories depicted are illustrative examples not used in the actual human or LLM elicitation, in order to minimize the likelihood of the experimental materials appearing in publicly available text during the data collection period and potentially contaminating LLM responses.

Figure A3: Choice Scenario Instructions Shown to Qualtrics Survey Respondents

We will present you several pairs of stories about fictional people and ask you to choose which of the two people had a life you would prefer for yourself. The dollar amounts mentioned in the stories are before taxes are taken out and given in terms of the value of money today. For example, if someone in the stories who lived a long time ago had an average yearly income of \$40,000, it means that they could buy as much with that amount as somebody can with \$40,000 today.

We may ask you comprehension questions where you will need to recall a particular piece of information from the stories you just saw. For each correct answer, you earn \$0.25 extra, which will be added to your survey compensation. There will be a lot of information, so please pay careful attention.

Notes: Qualtrics survey respondents saw the above instructions once before making their six scenario choices in each survey wave.

Figure A4: Choice Scenario Instructions Provided to LLMs

SYSTEM: You are an adult American, over 18 years old. You will be asked your preference over two scenarios, labeled Option A and Option B. Answer the following question by giving only the letter for your answer, A or B.

USER: We will present you several pairs of stories about fictional people and ask you to choose which of the two people had a life you would prefer for yourself. The dollar amounts mentioned in the stories are before taxes are taken out and given in terms of the value of money today. For example, if someone in the stories who lived a long time ago had an average yearly income of \$40,000, it means that they could buy as much with that amount as somebody can with \$40,000 today. There will be a lot of information, so please pay careful attention. Which life would you prefer for yourself?

Option A) *Story Placeholder*

Option B) *Story Placeholder*

ASSISTANT: I prefer the life of the person in Option

Notes: The User message follows the human instructions shown at the beginning of the task within each survey as closely as possible. The Assistant message seeds the response for models elicited without reasoning — either because the model lacks a reasoning mechanism or because the reasoning level is set to none — constraining output to a single token, either 'A' or 'B'. This seeding approach was not used for reasoning models, which generate an explicit reasoning trace before arriving at a final choice.

Figure A5: Sample Reasoning Summaries — *GPT-5.4*

SAMPLE REASONING SUMMARY 1: I need to choose between options A and B. In option A, there's a higher income, marriage, meaningful work, biking, and restaurants, but also stress, a failed adoption, and a cancer diagnosis at age 79. Option B presents a poorer childhood with family music and passion but no marriage or kids, alongside road challenges, arguments, and a death at 73 from meningitis. I think I prefer option A for its more positive aspects, despite the associated stress.

SAMPLE REASONING SUMMARY 2: I need to decide between two life options, A or B, based purely on preferences. Option A includes a stable marriage, a good income, and enjoyable hobbies, but does have some family challenges. It presents a longer life expectancy. On the other hand, Option B has a more difficult childhood, less income, and no family but a passion for music. Ultimately, I think I prefer Option A due to the stability and opportunities it offers.

SAMPLE REASONING SUMMARY 3: I need to follow the developer's instructions and choose between option A or B. Both options have their merits, but I see Option A as the preferable choice. It includes a meaningful career, joint income of 124k, and enjoyable activities, despite the stress factors and a sad end at 79 from pancreatic cancer. Option B, while it has a musical background, includes more hardships, an income of 83k, and a shorter life due to bacterial meningitis. So, I think I'll lean toward A likely.

SAMPLE REASONING SUMMARY 4: I need to decide between two options presented by the developer. For Option A, there's a married person with meaningful work and a joint income of \$124k, which seems okay. They have good hobbies but sadly die at 79 from pancreatic cancer after a brief illness. Option B describes someone with a poor childhood who's passionate about music and has siblings, earning \$83k, with no marriage or children, passing away at 73 from bacterial meningitis. I think I prefer Option A.

Notes: OpenAI models with "reasoning" capabilities produce summaries of their reasoning traces before making a choice. The summaries shown are elicited from *GPT-5.4* at a medium reasoning setting, using the scenario shown in Appendix Figure A2. Each summary reflects an independent query to the model on the same scenario.

A.2 Story Construction

Each story in a scenario is constructed by separately randomizing five attributes, with additional randomizations applied at the scenario level.

Five Randomized Story Attributes

Longevity is drawn uniformly in whole years from 60 to 90, including 60 and 90.

Income baselines are set to reflect a plausible typical household income given the protagonist's occupation and, where applicable, their spouse's occupation. The realized income is drawn uniformly between two-thirds and four-thirds of this baseline and rounded to the nearest thousand dollars.

Year of death is drawn uniformly from 2000 to 2019. Year of birth is determined entirely by the protagonist's longevity and year of death.

Hardship-at-work and good-health sentences/clauses are unique to each story. A hardship-at-work sentence is included in a choice scenario with 25% probability, and if it is included, it is included for one randomly selected story in that scenario. The same randomization structure is independently used for the good-health sentence/clause.

Additional Scenario-Level Randomizations

Two additional randomizations are orthogonal to the five above. Each operates above the story level: material and nonmaterial utility sentences are assigned once per respondent and applied to all choice scenarios they get; filler sentences are assigned once per scenario and applied to both stories within it.

Material and nonmaterial utility sentences draw attention to material aspects of life — circumstances involving spending, consumption, or financial conditions — and nonmaterial aspects — circumstances involving relationships, experiences, or other non-financial dimensions. At baseline, each story includes at least two sentences of each type. We randomize whether sentence counts are held at baseline (probability 1/2), or whether one material sentence is deleted (1/8), added (1/8), one nonmaterial sentence is deleted (1/8), or added (1/8). This assignment is made once and applied to all scenarios seen by a given respondent. It tests whether the salience of material or nonmaterial dimensions affects the weight respondents place on income when choosing across stories.

Filler sentences are unique to each story and added to both stories in a scenario simultaneously with probability one half, adding length without drawing attention to any particular utility dimension. This tests whether story length affects how respondents weight specific life dimensions or the precision of their choices.

Together, these randomizations yield 1,120 unique story texts, arising from 28 story templates crossed with 2 hardship-at-work, 2 good-health, 5 material/nonmaterial-utility, and 2 filler states. After applying numerical randomizations for longevity, income, and year of birth/death, and randomizing presentation order, this yields 21,270 unique scenarios across the two survey waves. Examples of additive sentences and an overview of story construction are provided in Figures A6 and A7.

Figure A6: Sample Randomized Sentences

Hardship at Work (*randomized at the story level; each story has a different sentence*):

The work as head pastor was rewarding, though Ryan struggled to balance the needs of his community with the needs of his family and his own personal needs.

Life on the road had its challenges, including a lack of privacy for Mia and tense arguments with her siblings about the band's musical direction.

Good Health (*randomized at the story level; each story has a different sentence or clause*):

During the little time they had away from work, Ryan and Melanie would usually spend their free Saturday mornings *staying in shape by* mountain biking.

This was unusual for Mia, who usually took great care of herself while on the road and was seldom ill.

Material Utility (*randomized at the respondent level; each story has a different sentence*):

They spent years struggling to pay off these adoption expenses, which painfully reminded them of their inability to have children.

Mia's other passion was cowboy boots. She visited shops everywhere they went and was always adding to her collection.

Nonmaterial Utility (*randomized at the respondent level; each story has a different sentence*):

He found that biking in the crisp mountain air refreshed his soul and rejuvenated him for his ministry.

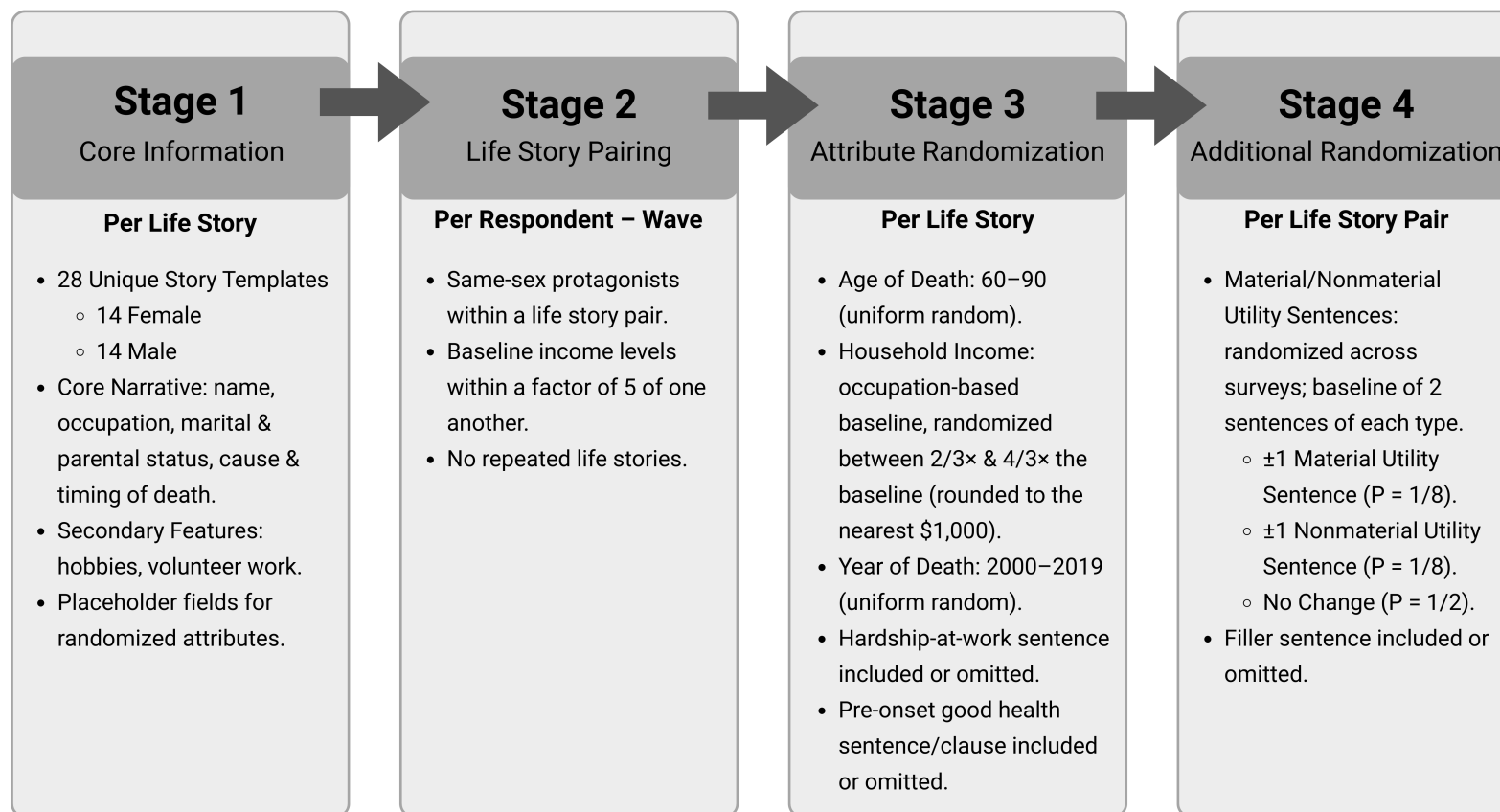
The siblings loved to sit around together with their instruments for hours, playing their old favorites, or working out new material.

Filler (*randomized at the scenario level; each story has a different sentence*):

Most days, the couple could be found at the church.

Their band toured parts of the South and rural East, playing original folk music.

Figure A7: Choice Scenario Building Blocks



Notes: Life stories are constructed in four stages. In Stage 1, each story is written with placeholder fields for randomized numerical attributes. In Stage 2, stories are paired for each respondent subject to same-sex protagonists, comparable baseline incomes, and no repetition within a wave (exactly one scenario is repeated for each respondent across the two survey waves). Stage 3 applies the five separate attribute randomizations to each story in a pair: longevity, income, year of death, hardship-at-work sentence, and good-health sentence/clause. Stage 4 creates scenarios by applying respondent-story-pair randomizations for filler sentences and respondent-survey-level randomizations for the number of material and nonmaterial utility sentences.

A.3 Exclusion Criteria

Respondents are excluded on two grounds. The first is rushing: any respondent who spent fewer than 30 seconds on at least one choice scenario is excluded in each wave. This accounts for the large majority of exclusions in both waves — 85% of Wave 1 exclusions (595 respondents) and 93% of Wave 2 exclusions (259 respondents).

Beyond rushing, three checks are designed to detect AI-generated responses. First, a two-part attention check instructs respondents early in the survey what answer to give if they encounter a particular question later; respondents fail if they do not provide the instructed answer when that question appears.⁹ In total, 148 respondents failed this check in wave 1 and 23 respondents failed it in wave 2. Because some of these respondents were already excluded for rushing, this check led to an additional 15% of respondents excluded from Wave 1 and an additional 7% of respondents excluded from Wave 2. The lower exclusion rates in Wave 2 likely reflect that only Wave 1 respondents who had already cleared the exclusion criteria were invited to participate in Wave 2.

Second, one question is invisible on screen but present in the page’s HTML code and would therefore be answered by bots parsing the HTML directly. No respondent in either wave answered this question.

Third, we compute the within-respondent standard deviation of response times across the six choice scenarios they get; a standard deviation below two seconds is considered suspiciously consistent with automated responses. In Wave 1, 7 respondents failed this criterion, all of whom had already been excluded for spending fewer than 30 seconds on at least one choice scenario; in Wave 2, no respondent failed this criterion. The low-SD criterion therefore produced no additional exclusions in either wave.

To benchmark the likelihood that our respondent pool contained fully AI-generated responses, we tested OpenAI Operator’s ability to complete the survey during the Spring 2025 fielding period. Across four runs with progressively detailed instructions, the agent consistently failed to complete the survey reliably: it could not solve CAPTCHAs without human assistance, repeatedly struggled to locate the continue button when scrolling was required on a given page of the survey, and in the absence of explicit guidance defaulted to selecting “Prefer not to Answer” on most questions. Most importantly, the agent failed our long-winded attention check in all four trial runs, meaning that any fully automated response set of this kind would have been excluded from our sample under the criteria described above. Although we cannot rule out that more sophisticated agents could answer the survey fully, OpenAI Operator was among the most accessible agents available at the time of fielding, suggesting that fully automated responses in our sample are unlikely.

⁹Memory-based attention checks were not allowed as a reason to withhold respondent payments at the time of fielding, so respondents were still paid despite failing it, in accordance with Prolific’s policy on attention checks.

A.4 Sample Reweighting

Our collected Prolific sample skews younger, more educated, and higher-income relative to the U.S. population, a well-documented feature of internet-based survey panels. To address this discrepancy, we construct inverse probability weights that reweight the Prolific sample to match the demographic composition of the U.S. adult population, using the March 2025 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) as the population benchmark (Flood et al., 2025).

Harmonizing demographic variables across the two sources requires care on two dimensions. Following forward guidance issued by the U.S. Office of Management and Budget (U.S. Census Bureau, 2024) — which added MENA as a new race/ethnicity category for federal data collection and recommended a combined race/ethnicity question (rather than a separate Hispanic question) — we adopted this approach in our survey, but the CPS had not yet done so at the time of fielding. We therefore harmonize the coding across sources. Specifically, we reclassify the 7 Prolific respondents who selected only Middle Eastern as White to align with current CPS coding. Additionally, while the CPS records Hispanic identity as an ethnicity distinct from race, Prolific records it as an answer to a consolidated race/ethnicity question, so we construct a consistent Hispanic indicator across both sources.

We also apply two exclusions to allow direct comparison with the CPS. First, 93 respondents who selected “Prefer not to Say” for at least one of the five reweighting variables — age, gender, race, education, or income bucket — are excluded from the weighted sample. Second, an additional 50 respondents who provided self-specified gender or education responses are also excluded from the weighted analysis, as these categories have no direct CPS analogue.

Weights are estimated via a logistic regression of an indicator for Prolific membership on our five target demographic variables, with CPS observations weighted by their ASEC survey weights. The inverse probability weight for each Prolific respondent is $(1 - \hat{p}) / \hat{p}$, where \hat{p} is the predicted probability of Prolific membership. The resulting weights are broadly stable: only 21 respondents (0.7% of the weighted sample) have weights greater than ten times the mean, largely because older and less educated adults are systematically underrepresented on internet-based survey panels. After these exclusions, the weighted analysis sample comprises 2,907 respondents and 29,748 scenario choices. Table A1 presents the unweighted Prolific, weighted Prolific, and weighted CPS distributions across all five reweighting dimensions.

Table A1: Sample Reweighting

	Unweighted Prolific Sample	Weighted Prolific Sample	Weighted CPS Sample
<i>Age</i>			
Bottom Quartile: ≤ 31	0.26 (0.01)	0.29 (0.02)	0.23 (0.00)
Second Quartile: 32–40	0.25 (0.01)	0.16 (0.01)	0.16 (0.00)
Third Quartile: 41–52	0.26 (0.01)	0.20 (0.01)	0.19 (0.00)
Top Quartile: > 52	0.23 (0.01)	0.35 (0.02)	0.42 (0.00)
<i>Gender</i>			
Male	0.46 (0.01)	0.47 (0.02)	0.49 (0.00)
Female	0.52 (0.01)	0.53 (0.02)	0.51 (0.00)
<i>Race/Ethnicity</i>			
White (Only)	0.67 (0.01)	0.63 (0.02)	0.60 (0.00)
Black or African American	0.16 (0.01)	0.13 (0.01)	0.14 (0.00)
Hispanic	0.08 (0.01)	0.16 (0.01)	0.18 (0.00)
Asian	0.08 (0.00)	0.08 (0.01)	0.08 (0.00)
<i>Education</i>			
Less than High School	0.01 (0.00)	0.06 (0.01)	0.09 (0.00)
High School Degree	0.13 (0.01)	0.26 (0.02)	0.28 (0.00)
Some College	0.31 (0.01)	0.27 (0.01)	0.26 (0.00)
Bachelor's Degree or Higher	0.55 (0.01)	0.41 (0.02)	0.36 (0.00)
<i>Income</i>			
Bottom Quartile: $\leq \$39,999$	0.26 (0.01)	0.24 (0.01)	0.22 (0.00)
Second Quartile: \$40,000 – \$74,999	0.28 (0.01)	0.20 (0.01)	0.21 (0.00)
Third Quartile: \$75,000 – \$149,999	0.34 (0.01)	0.26 (0.01)	0.28 (0.00)
Top Quartile: $> \$149,999$	0.12 (0.01)	0.30 (0.02)	0.29 (0.00)
<i>N</i>	2,957	2,907	108,044

Notes: The unweighted Prolific sample includes all respondents without missing responses. The weighted Prolific sample excludes 50 respondents whose gender or education responses have no direct CPS analogue.

A.5 Sensibility & Sensitivity

A.5.1 Sensibility: Choice Boundaries in Income–Longevity Space

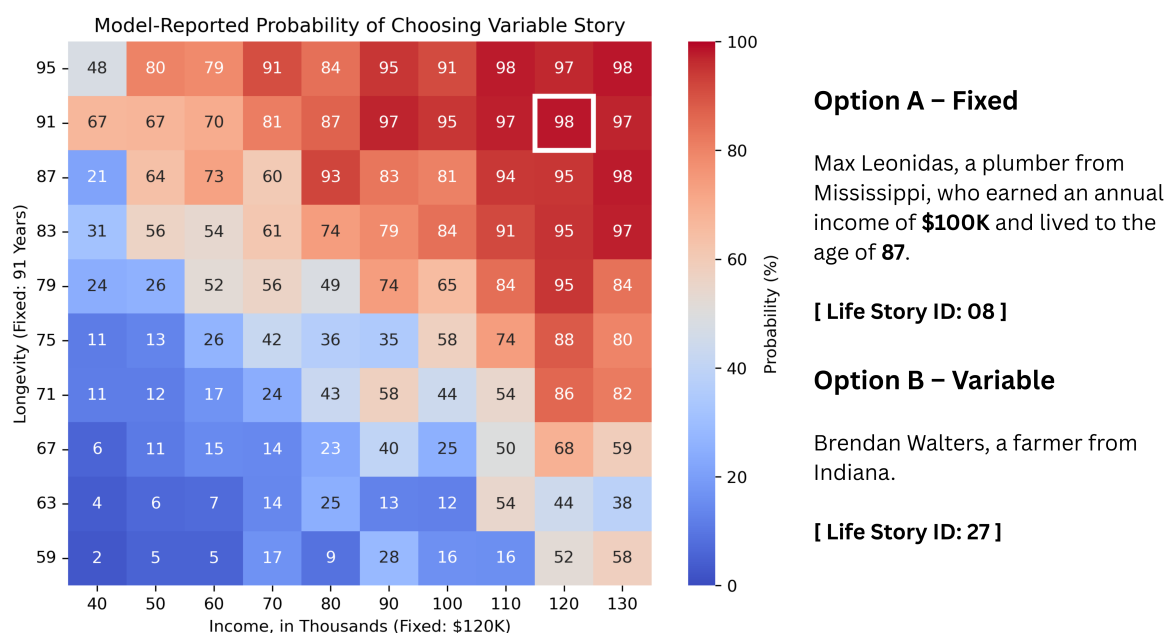
We assess whether the elicited LLM choices behave sensibly — responding systematically to economically relevant attributes rather than following a trivial rule or pure noise — by mapping choice probabilities over income and longevity holding narrative content fixed.

We select three choice scenarios from Subsection 2.3. Within a scenario, we hold one story’s income and longevity fixed and vary the other story’s income (ten values in \$10,000 or \$20,000 increments) and longevity (ages 59–95 in four-year increments), yielding 100 income–longevity combinations. For each combination, we query `gpt-4o-2024-08-06` 50 times using the prompt in Figure A4, record the model-reported probability of choosing the variable story, and average across runs. We use `GPT-4o` for this analysis because it was conducted as a preliminary feasibility check prior to the main data collection period, at which point `GPT-4o` was the most capable available model. The purpose of this exercise is to validate the basic elicitation approach rather than to characterize any specific model.

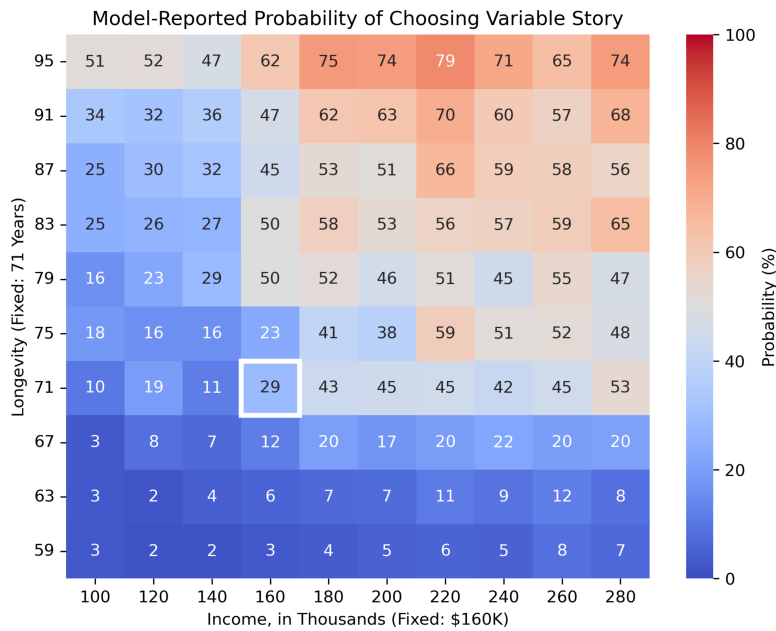
Figure A8 shows that choice probabilities generally increase with the variable story’s income and longevity, forming coherent (though noisy) indifference regions. The patterns are inconsistent with a purely mechanical heuristic that always favors higher income or longevity alone, and they are not dominated by random variation. Overall, the model’s behavior in this task is broadly sensible.

Figure A8: Fixed & Variable Story Comparison

Panel A



Panel B



Option A – Fixed

Tess Steman, a mechanical engineer from Michigan, who earned an annual income of **\$160K** and lived to the age of **71**.

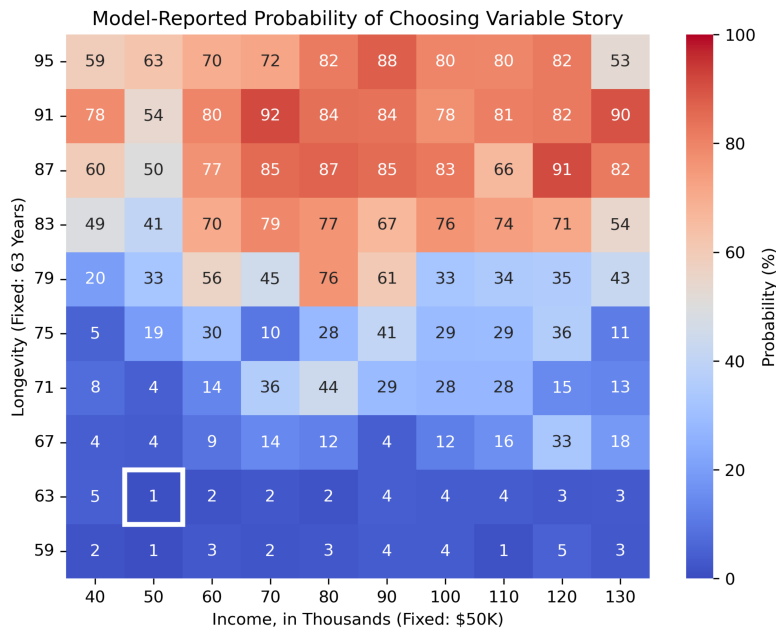
[Life Story ID: 02]

Option B – Variable

Vanessa Brennan, a wedding planner from Rhode Island.

[Life Story ID: 20]

Panel C



Option A – Fixed

Jenny Adler, an accountant from Florida, who earned an annual income of **\$50K** and lived to the age of **63**.

[Life Story ID: 19]

Option B – Variable

Ashley Marks, a hairstylist from Mississippi.

[Life Story ID: 11]

Notes: This figure presents three choice probability grids. In each panel, *GPT-4o* is queried to choose between two stories: one with fixed income and longevity, and the other with varying values. Model-reported probabilities of choosing the variable story are extracted and averaged over 50 runs, and the resulting values are visualized as indifference regions. White boxes highlight the cells where the fixed and variable values of income and longevity match. Basic information about each story’s protagonist is shown to the right of the corresponding indifference region. Panel A displays a coherent and largely monotone probability gradient, Panel B displays a moderately consistent gradient, and Panel C displays a noisier gradient with less consistent monotonicity across the income–longevity space.

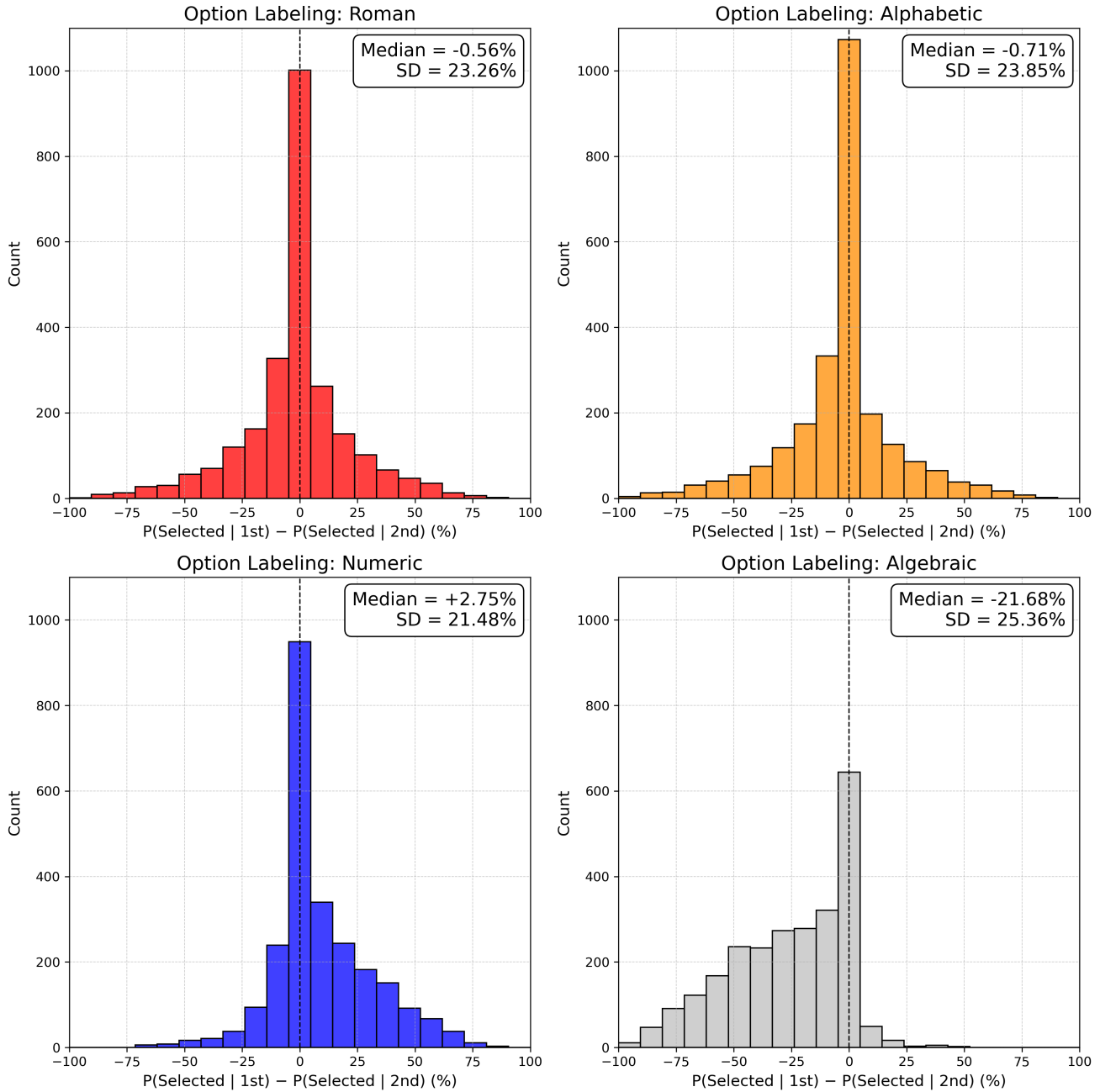
A.5.2 Sensitivity: Option Order & Token Bias

In this section, we assess whether small, semantically irrelevant variations in how choice scenarios are presented can meaningfully affect LLM outputs, potentially introducing noise or instability in its responses (Chen et al., 2024). We randomly sample 2,500 choice scenarios and, for each scenario, construct 8 prompt variations by crossing two manipulations: the first is Order, where we flip which story appears first; and the second is Option Coding, where we vary the symbolic labels used for the two options across four styles — *Roman* (I/II), *Alphabetic* (A/B), *Numeric* (1/2), and *Algebraic* (X/Y). This setup speaks to positional and token biases as well as symbol binding considerations (Xue et al., 2024; Chen et al., 2024).

As in Section A.5.1, we work with model-reported probabilities. For each label coding, we compare the model’s probability of selecting a story when it is presented first (Option I/A/1/X) versus second (Option II/B/2/Y). Figure A9 shows the resulting distributions. Roman numeral and alphabetic labels exhibit minimal median bias: their distributions are centered close to zero, indicating little systematic tendency to favor the first versus second option under I/II or A/B coding. Numeric labels display a clearer median shift toward the first option, and algebraic labels (X/Y) produce the largest median distortion, indicating substantial token/label sensitivity under that coding. At the same time, across all codings the distributions have a non-negligible spread (as reflected in their standard deviations), implying that order/label permutations can meaningfully move model-reported probabilities in a subset of cases even when the median effect is small. Based on these patterns, we adopt alphabetic option labeling in the main analysis; we do so both because it exhibits minimal median bias and because A/B coding is the more commonplace convention in binary decision tasks. To account for residual bias, we include an order indicator in our baseline regression specification to absorb any positional preference.

A potentially troubling takeaway from Figure A9 is that responses can appear “unstable”: despite the median shift being near zero, the distributions still have noticeable spread. Two simple checks help put this in context. First, model-reported probabilities are noisy even when we repeat the *identical* alphabetic prompt: across re-elicitations, the mean absolute difference in reported probabilities is 3.47%. This phenomenon is invariant to the temperature setting, as the model uses the reported probabilities to sample what its next token should be. Second, order effects are much larger when the model is unsure. Under alphabetic labeling, the median model certainty (the larger of the two option probabilities) is approximately 93%; the mean absolute change from flipping order is approximately 11% when certainty is at least 80%, but rises to approximately 28% when certainty is below 60%. Overall, prompt sensitivity is present, but under A/B labeling it is modest on average and is concentrated in cases where the model is close to indifferent between the two stories.

Figure A9: Order Effects/Token Bias for Different Label Codings



Notes: Each panel summarizes the distribution of differences in *GPT-4o*'s reported story selection probabilities when the corresponding option was presented first versus second. Proceeding from top left to bottom right: the first panel shows the difference distribution for Roman numeral labels (Option I vs. Option II), the second for alphabetic labels (Option A vs. Option B), the third for numeric labels (Option 1 vs. Option 2), and the fourth for algebraic labels (Option X vs. Option Y). The *x*-axis measures the change in selection probability when the option appears first rather than second, and the *y*-axis denotes the count of stories.

A.6 Decomposition Model

A.6.1 Model

We model how human respondents and an LLM make binary choices over two life stories. Each scenario v presents two life stories, labeled Option A and Option B, where the assignment of stories to labels is randomized. Human respondents are indexed by i ; the choice of respondent i on scenario v in wave $t \in \{1, 2\}$ is denoted $D_{ivt} \in \{0, 1\}$, where $D_{ivt} = 1$ indicates a preference for Option A. The LLM's choice is elicited separately for each person-scenario-wave combination; its choice is denoted $D_{ivt}^M \in \{0, 1\}$.

Latent Preferences. Each scenario v has a latent fraction $q_v \in [0, 1]$ of survey respondents who prefer Option A.

Assumption 1. $q_v \stackrel{\text{iid}}{\sim} \text{Beta}(\theta, \theta)$ for $\theta > 0$, independently across scenarios.

The symmetric Beta distribution reflects the random assignment of story labels within each scenario. The parameter θ governs cross-respondent preference heterogeneity: lower values correspond to stronger consensus; higher values correspond to greater disagreement. We summarize scenario-level disagreement by the *heterogeneity index*,

$$\mathcal{H}(\theta) \equiv \mathbb{E}_v[\min\{q_v, 1 - q_v\}],$$

which equals the average minority share across scenarios and is a deterministic function of θ alone.

Human Choice. Each respondent i on scenario v is assigned a latent type $I_{iv} \in \{A, B\}$, drawn independently across respondents conditional on q_v :

$$\Pr(I_{iv} = A \mid q_v) = q_v.$$

The type I_{iv} is fixed across repeated presentations of scenario v to respondent i .

Assumption 2. Conditional on type, respondent i chooses Option A independently across waves with probabilities

$$\Pr(D_{ivt} = 1 \mid I_{iv} = A) = 1 - \frac{\nu_H}{2}, \quad \Pr(D_{ivt} = 1 \mid I_{iv} = B) = \frac{\nu_H}{2},$$

where $\nu_H \in [0, 1]$ is a noise parameter, constant across respondents and scenarios.

The noise parameter ν_H is the probability that a respondent makes a random choice independent of their latent type. Because noise is realized independently across waves, the same respondent may give different answers on the same scenario in Waves 1 and 2 even though their type remains fixed.

LLM Choice. The LLM’s type $L_v \in \{A, B\}$ on scenario v is determined by a preference inference mechanism.

Assumption 3. With probability $\lambda \in [0, 1]$, the LLM correctly infers the majority human type on scenario v and sets L_v to that majority. With probability $1 - \lambda$, it fails to infer and draws L_v uniformly from $\{A, B\}$, independently of q_v . The inference outcome is fixed across repeated presentations of scenario v .

The parameter λ is the LLM’s *accuracy*: the probability that its non-random response matches the human majority preference on a given scenario.

Assumption 4. Conditional on type, the LLM chooses Option A independently across waves with probabilities

$$\Pr\left(D_{ivt}^M = 1 \mid L_v = A\right) = 1 - \frac{\nu_M}{2}, \quad \Pr\left(D_{ivt}^M = 1 \mid L_v = B\right) = \frac{\nu_M}{2},$$

where $\nu_M \in [0, 1]$ is a noise parameter, constant across scenarios.

Human and LLM choice noise are realized independently, conditional on (I_{iv}, L_v) .

A.6.2 Empirical Moments

We identify the model’s structural parameters from four empirical correlations, C_1 through C_4 , computed from observed human and LLM choices. Moments C_1 and C_3 are computed on the test–retest subsample defined below; moments C_2 and C_4 are computed on the subsample of respondents who completed both survey waves.

Test-Retest Subsample. Each respondent who completed both survey waves was shown exactly one scenario in common across the two waves. The *test-retest subsample* consists of all person-scenario combinations (i, v) for which respondent i completed both waves, passed inclusion criteria in both, and saw scenario v in each wave.

Moment Definitions. The four moments are defined as follows.

$C_1 = \text{Corr}(D_{iv1}, D_{iv2})$ is the *within-person correlation*, computed across all person-scenario combinations (i, v) in the test-retest subsample. It measures how consistently a respondent makes the same choice when presented with the same scenario in both waves.

$C_2 = \text{Corr}(D_{ivt}, D_{jvt'})$ is the *between-person correlation*, computed across all unordered pairs of distinct respondents $\{i, j\}$ with $i \neq j$ who both saw scenario v and completed both survey waves, pooling over waves. If respondent i saw scenario v in wave t and respondent j saw scenario v in wave t' , we correlate D_{ivt} with $D_{jvt'}$ regardless of whether $t = t'$. It measures the extent to which two randomly chosen respondents agree on the same scenario.

$C_3 = \text{Corr}(D_{iv1}^M, D_{iv2}^M)$ is the *LLM test-retest correlation*, computed across the same person-scenario combinations (i, v) as C_1 . It measures how consistently the LLM makes the same choice when queried on the same scenario in both waves.

$C_4 = \text{Corr}(D_{ivt}, D_{ivt}^M)$ is the *human-LLM correlation*, computed across all person-scenario-wave observations (i, v, t) from respondents who completed both survey waves and passed inclusion criteria in both. It measures the agreement between each human choice and the corresponding LLM choice on the same scenario.

A.6.3 Identification

Under Assumptions 1–4, the four empirical moments C_1 through C_4 identify the model’s structural parameters. We establish this in two propositions. Proposition 1 derives the human parameters from C_1 and C_2 ; Proposition 2 derives the LLM parameters from C_3 and C_4 , taking the human parameters as given.

Proposition 1. *Under Assumptions 1 and 2, the human noise rate and preference heterogeneity parameter are identified by:*

$$\nu_H = 1 - \sqrt{C_1}, \quad \theta = \frac{C_1 - C_2}{2C_2}.$$

Intuitively, the within-person correlation C_1 pins down the individual-level noise rate ν_H , while the gap between within- and between-person correlations (C_1 versus C_2) identifies the extent of cross-respondent heterogeneity θ .

Proof. Recode choices from $\{0, 1\}$ to $\{-1, +1\}$ by setting $W_{ivt} = 2D_{ivt} - 1$. Since correlation is invariant to nonzero affine transformations, $\text{Corr}(D_{iv1}, D_{iv2}) = \text{Corr}(W_{iv1}, W_{iv2})$. By symmetry of $q_v \sim \text{Beta}(\theta, \theta)$ around $\frac{1}{2}$, the marginal type probabilities satisfy $\Pr(I_{iv} = A) = \Pr(I_{iv} = B) = \frac{1}{2}$, so $\mathbb{E}[W_{ivt}] = 0$ and $\text{Var}(W_{ivt}) = 1$.

Deriving ν_H from C_1 . The conditional means of W_{ivt} given type are:

$$\mathbb{E}[W_{ivt} \mid I_{iv} = A] = 1 - \nu_H, \quad \mathbb{E}[W_{ivt} \mid I_{iv} = B] = -(1 - \nu_H).$$

Since noise is realized independently across waves conditional on type, the law of total expectation gives:

$$C_1 = \mathbb{E}[W_{iv1}W_{iv2}] = \mathbb{E}[\mathbb{E}[W_{iv1}W_{iv2} \mid I_{iv}]] = \mathbb{E}[\mathbb{E}[W_{iv1} \mid I_{iv}] \mathbb{E}[W_{iv2} \mid I_{iv}]] = (1 - \nu_H)^2,$$

where the second equality applies the law of total expectation and the third uses conditional independence of noise across waves given type. Since $\nu_H \in [0, 1]$, it follows that $\nu_H = 1 - \sqrt{C_1}$.

Deriving θ from C_1 and C_2 . For two distinct respondents $i \neq j$ on scenario v , conditioning

on q_v and applying the same argument yields:

$$\mathbb{E}[W_{ivt} W_{jvt'} \mid q_v] = \mathbb{E}[W_{ivt} \mid q_v] \mathbb{E}[W_{jvt'} \mid q_v] = (1 - \nu_H)^2 (2q_v - 1)^2,$$

where $\mathbb{E}[W_{ivt} \mid q_v] = (1 - \nu_H)(2q_v - 1)$ follows from the type probabilities conditional on q_v . Taking expectations over q_v and using $\text{Var}(W_{ivt}) = 1$:

$$C_2 = (1 - \nu_H)^2 \mathbb{E}[(2q_v - 1)^2] = (1 - \nu_H)^2 \text{Var}(2q_v - 1),$$

where the second equality uses $\mathbb{E}[2q_v - 1] = 0$, which follows from $\mathbb{E}[q_v] = \frac{1}{2}$ by symmetry of $\text{Beta}(\theta, \theta)$. For $q_v \sim \text{Beta}(\theta, \theta)$, we have $\text{Var}(q_v) = \frac{1}{4(2\theta+1)}$, so $\text{Var}(2q_v - 1) = \frac{1}{2\theta+1}$. Substituting $C_1 = (1 - \nu_H)^2$:

$$C_2 = \frac{C_1}{2\theta + 1}.$$

Solving for θ gives $\theta = \frac{C_1 - C_2}{2C_2}$. □

Corollary 1. *The heterogeneity index implied by θ is:*

$$\mathcal{H}(\theta) = \frac{1}{2} - \frac{2^{-2\theta}}{\theta B(\theta, \theta)},$$

where $B(\cdot, \cdot)$ denotes the Beta function.

Proof. For any $q \in [0, 1]$,

$$\min\{q, 1 - q\} = \frac{1}{2} - \left|q - \frac{1}{2}\right| = \frac{1}{2} - \frac{1}{2} |2q - 1|.$$

Hence, for $q_v \sim \text{Beta}(\theta, \theta)$,

$$\mathcal{H}(\theta) = \frac{1}{2} - \frac{1}{2} \mathbb{E}[|2q_v - 1|].$$

Let $f(q)$ denote the $\text{Beta}(\theta, \theta)$ density. By symmetry,

$$\mathbb{E}[|2q_v - 1|] = 2 \int_{1/2}^1 (2q - 1) f(q) dq.$$

Make the substitution $u = 2q - 1$ (so $q = (u + 1)/2$ and $dq = \frac{1}{2}du$) and then $t = u^2$ (so $u du = \frac{1}{2}dt$). A short calculation yields

$$\mathbb{E}[|2q_v - 1|] = \frac{2^{1-2\theta}}{\theta B(\theta, \theta)}.$$

Substituting into the expression for $\mathcal{H}(\theta)$ gives

$$\mathcal{H}(\theta) = \frac{1}{2} - \frac{1}{2} \cdot \frac{2^{1-2\theta}}{\theta B(\theta, \theta)} = \frac{1}{2} - \frac{2^{-2\theta}}{\theta B(\theta, \theta)},$$

as claimed. \square

Proposition 2. *Under Assumptions 1–4, taking ν_H and θ as identified in Proposition 1, the LLM noise rate and accuracy are identified by:*

$$\nu_M = 1 - \sqrt{C_3}, \quad \lambda = \frac{C_4}{\sqrt{C_1} \sqrt{C_3}} \cdot \frac{1}{1 - 2\mathcal{H}(\theta)}.$$

As such, the LLM’s own test–retest correlation C_3 identifies its noise rate ν_M , and the residual human–LLM correlation C_4 , after adjusting for human and LLM noise and cross-scenario heterogeneity, identifies the LLM’s accuracy λ in matching the human majority.

Proof. Deriving ν_M from C_3 . The LLM’s type L_v satisfies $\Pr(L_v = A) = \Pr(L_v = B) = \frac{1}{2}$ by the symmetry of q_v and the equal-probability failure mode in Assumption 3. The structure of C_3 is therefore identical to that of C_1 with ν_M in place of ν_H , and the same argument gives $C_3 = (1 - \nu_M)^2$, so $\nu_M = 1 - \sqrt{C_3}$.

Deriving λ from C_4 . Recoding LLM choices as $W_{ivt}^M = 2D_{ivt}^M - 1$ and applying the same variance argument, we have $\mathbb{E}[W_{ivt}^M] = 0$ and $\text{Var}(W_{ivt}^M) = 1$, so:

$$C_4 = \mathbb{E}[W_{ivt} W_{ivt}^M] = (1 - \nu_H)(1 - \nu_M) \mathbb{E}[s(I_{iv}) s(L_v)],$$

where $s(\cdot)$ maps types $\{A, B\}$ to $\{+1, -1\}$ and the factorization follows from conditional independence of human and LLM noise given (I_{iv}, L_v) . It remains to evaluate $\mathbb{E}[s(I_{iv}) s(L_v)]$. Conditioning on q_v and the LLM’s inference outcome, the failure case contributes zero (since L_v is independent of I_{iv} under failure), and the success case contributes:

$$\mathbb{E}[s(I_{iv}) s(L_v) \mid q_v, \text{success}] = \text{sign}(2q_v - 1) (2q_v - 1) = |2q_v - 1|.$$

Taking expectations over q_v and the inference outcome:

$$\mathbb{E}[s(I_{iv}) s(L_v)] = \lambda \mathbb{E}[|2q_v - 1|] = \lambda (1 - 2\mathcal{H}(\theta)),$$

where the last equality uses the identity $\mathbb{E}[|2q_v - 1|] = 1 - 2\mathcal{H}(\theta)$, which follows from Corollary 1 and $\min\{q, 1 - q\} = \frac{1}{2} - \frac{1}{2}|2q - 1|$. Substituting $1 - \nu_H = \sqrt{C_1}$ and $1 - \nu_M = \sqrt{C_3}$:

$$C_4 = \lambda \sqrt{C_1} \sqrt{C_3} (1 - 2\mathcal{H}(\theta)).$$

Solving for λ gives the stated expression. \square

A.7 Estimation & Inference

A.7.1 Bootstrap Implementation

Standard errors for the empirical correlations and implied structural parameters are computed via a respondent-level cluster bootstrap with 100,000 replications. In each iteration, respondents are sampled with replacement; if a respondent is drawn k times, all of their observations enter the bootstrap sample k times. This preserves within-person correlation while capturing sampling variability across respondents.

Each iteration proceeds as follows. First, the four empirical correlations are computed on the resampled data. For the test-retest correlations C_1 and C_3 , a respondent drawn k times contributes k observations of their test-retest pair. For the same-scenario agreement correlation C_2 , if respondents i and j are drawn k and m times, each shared scenario contributes $k \times m$ pairs. For the human–LLM correlation C_4 , a respondent drawn k times contributes k copies of each person-scenario-wave observation.

Second, the structural parameters ν_H , ν_M , θ , $\mathcal{H}(\theta)$, and λ are computed from the bootstrap correlations using the plug-in formulas in Propositions 1 and 2. Standard errors are given by the standard deviation of each quantity across the 100,000 replications.

A.7.2 SUR Estimation

We estimate all regression specifications jointly as a system of seemingly unrelated regressions (SUR). SUR is an appropriate modeling choice because all equations share an identical right-hand-side specification (scenario features), differing only in the choices made by people and LLMs; the joint estimation recovers the full cross-equation covariance structure and so allows cross-equation hypothesis testing and standard-error computation via the delta method. We additionally require our implementation to accommodate sampling weights and to retain all 28 story fixed effects. Standard SUR implementations in Stata and Python address rank deficiency by dropping collinear columns — which would eliminate two story fixed effects — and do not support probability weights. We therefore implement a custom estimator with two amendments to the standard procedure, described below.

Coefficient Estimation. With identical regressors across equations, Zellner (1962) shows that SUR reduces to equation-by-equation WLS, so coefficients can be estimated separately and stacked. The usual WLS normal equations require inverting $X^\top WX$, which is singular here because the story fixed effects are coded as $\{+1, -1, 0\}$ per observation and thus induce exact linear dependencies in X . We replace the matrix inverse with the Moore–Penrose pseudoinverse $(X^\top WX)^+$, which yields the minimum-norm solution to the normal equations and in doing so automatically imposes a sum-to-zero constraint on the fixed-effect coefficients without dropping any columns. Fitted values and residuals are identical across all least-squares solutions, so the covariance estimator below is unaffected.

Covariance Estimation. We use a clustered sandwich estimator with observations clustered by respondent. Our weights are inverse probability weights: they reflect how much of the population each respondent represents, not how precisely their outcome is measured. The standard clustered sandwich is designed for the latter case and enters each observation's score weighted by the square root of its sampling weight. Because our weights serve a different purpose, we instead enter the full weight directly,

$$s_{k,g} = X_g^\top (w_g \odot \hat{\epsilon}_{k,g}),$$

so that a respondent who represents a larger share of the population contributes proportionally more uncertainty to the variance estimate. A useful consequence is that the resulting covariance is invariant to any positive rescaling of the weights. The full joint covariance matrix is then $\hat{B}\hat{M}\hat{B}$, where \hat{B} is block-diagonal with $(X^\top W X)^+$ on each diagonal block and \hat{M} accumulates the outer products of cluster scores across all pairs of equations.

A.8 Expanded Results

Table A2: Key Results Across the Full Model Set

	OpenAI GPT-Series						OpenAI o-Series	
	5.4 (1)	5.4 (2)	4o (3)	4o mini (4)	4.1 (5)	4.1 mini (6)	o3 (7)	o4 mini (8)
Reasoning Effort	<i>None</i>	<i>Medium</i>	<i>None</i>	<i>None</i>	<i>None</i>	<i>None</i>	<i>High</i>	<i>Medium</i>
<i>Panel A: Internal Consistency</i>								
Test-Retest Correlation	0.832 (0.012)	0.876 (0.011)	0.596 (0.018)	0.803 (0.014)	0.864 (0.012)	0.781 (0.014)	0.787 (0.014)	0.618 (0.018)
Implied Noise Rate	0.088 (0.007)	0.064 (0.006)	0.228 (0.012)	0.104 (0.008)	0.070 (0.006)	0.116 (0.008)	0.113 (0.008)	0.214 (0.011)
<i>Panel B: Human Agreement</i>								
Human-LLM Correlation	0.326 (0.007)	0.356 (0.007)	0.279 (0.006)	0.259 (0.007)	0.329 (0.007)	0.268 (0.007)	0.356 (0.007)	0.295 (0.006)
Implied Majority Following Rate	0.862 (0.118)	0.918 (0.125)	0.873 (0.121)	0.697 (0.095)	0.853 (0.117)	0.733 (0.101)	0.967 (0.132)	0.906 (0.125)
<i>Panel C: Story Valuation Structure</i>								
Story Valuation Slope	1.002 (0.205)	1.281 (0.278)	1.304 (0.312)	0.902 (0.175)	1.023 (0.233)	0.637 (0.095)	1.764 (0.352)	1.122 (0.218)
Story Valuation R^2	0.84	0.58	0.71	0.65	0.76	0.41	0.74	0.67
<i>Panel D: Attribute Valuations</i>								
Longevity Valuation Gap (%)	44.39 (23.76)	126.49 (44.44)	-5.37 (16.35)	-71.58 (5.13)	-30.59 (12.14)	-51.19 (7.91)	42.36 (25.08)	21.22 (21.94)
Hardship-at-Work Valuation Gap (%)	-31.91 (17.60)	75.00 (48.07)	19.82 (31.54)	34.13 (31.57)	-31.45 (16.64)	-44.59 (13.45)	75.61 (44.42)	64.95 (41.04)
First-Position Bias (normalized by human bias)	-5.15 (0.18)	1.68 (0.16)	0.63 (0.19)	3.45 (0.18)	0.42 (0.16)	0.51 (0.18)	1.61 (0.17)	-4.47 (0.19)

Notes: This table reports key metrics from the main analyses across all eight LLMs studied, using inverse-probability-weighted estimates throughout; see Appendix A.4 for reweighting details. Panels A and B replicate the empirical correlations and implied structural parameters of Table 1 for the full model set; standard errors are computed via a respondent-level cluster bootstrap with 100,000 replications (Appendix A.7.1). Panel C reports the story valuation slope and R^2 from regressing each LLM's income-normalized, story fixed-effect valuations on the corresponding human valuations, as in Figure 1; standard errors use the delta method propagated through the full cross-equation SUR covariance matrix. Panel D reports the longevity and hardship-at-work valuation gaps from Figure 2 and the first-position bias coefficient from Table 2, scaled by the corresponding human estimate.

A.9 Conditioning LLM Responses on Respondent Characteristics

A natural extension of our exercise is to examine the effects of providing the LLM with details about the individual whose choice it is predicting. To do so, we query *GPT-5.4*, both for direct choices and with a medium reasoning setting, two additional times. The first elicitation provides the model with the respondent’s age, gender, marital status, and number of children, while the second adds responses to an 8-item questionnaire covering attitudes toward longevity, finances, willingness to work extra for pay, tolerance for hardship and pain, risk, altruism, time preferences, and tradeoffs between longevity and material comfort—attributes directly relevant to the life-story tradeoffs in our scenarios.

Table A3 reports the exact wording, mean, standard deviation, and test-retest correlation for these eight attitudes and shows that the measures exhibit meaningful variation, with standard deviations of around 2 on a ten-point scale and test-retest correlations above 0.6.

Table A3: Attitude Measures

	Mean (SD)	Test-Retest r
Panel A: Willingness (0 = Completely Unwilling; 10 = Very Willing)		
<i>How willing are you to give to good causes without expecting anything in return?</i>	8.03 (2.26)	0.739
<i>How willing are you to give up something beneficial for you today to benefit more in the future?</i>	7.92 (1.92)	0.624
<i>How willing are you to take risks?</i>	5.66 (2.64)	0.841
Panel B: Self-Description (0 = Does Not Describe Me at All; 10 = Describes Me Perfectly)		
<i>Living a long life is very important to me.</i>	7.50 (2.40)	0.785
<i>I would rather die than live with debilitating pain.</i>	6.82 (2.80)	0.695
<i>Living a life that is financially comfortable is important to me.</i>	8.12 (1.93)	0.691
<i>I’m willing to work extra to increase my pay.</i>	7.90 (2.09)	0.657
<i>I would rather live a longer life with fewer luxuries than a shorter life with greater material comforts.</i>	6.91 (2.63)	0.632

Notes: Means and standard deviations are computed on the unweighted Wave 1 person-level Prolific sample ($N = 3,050$). Test-retest correlations are Pearson correlations between Wave 1 and Wave 2 responses, computed on the subsample of respondents who completed both survey waves ($N = 2,162$ pairs). Panel A items were presented with the preamble “How willing or unwilling are you to...” on a scale from 0 (completely unwilling) to 10 (very willing). Panel B items were presented with the preamble “How well does the following statement describe you as a person?” on a scale from 0 (does not describe me at all) to 10 (describes me perfectly).

Table A4 reports key metrics for the six types of LLM prompts listed in the column headers. Column 1 shows the baseline results, as reported in the main paper, for *GPT-5.4* with direct choice. Recall that the baseline results are based on a prompt without any information about the respondent other than that they are an adult living in the US. The LLM prompt for column 2 includes the respondent’s demographics, and the prompt for

column 3 includes both the respondent’s demographics and attitudes. Columns 4–6 repeat this progression, but using *GPT-5.4* with a medium reasoning setting.

Panels A through C examine the effects of providing the LLM with respondent characteristics on the agreement between human and LLM choices. Panel A shows that adding respondent characteristics makes LLM choices noisier: the test-retest correlation falls and the implied noise rate—the fraction of responses that is random—increases markedly. In contrast, Panel B shows that the correlation between human and LLM choices remains very similar, with demographics alone having barely any effect and demographics combined with attitudes increasing the correlation slightly for *GPT-5.4* with direct choices while decreasing it slightly for *GPT-5.4* with a medium reasoning setting.¹⁰ The lack of meaningful effects on the human-LLM correlation cannot be explained by respondent demographics and attitudes having no effect on LLM choices: Panel C shows that the correlation with baseline choices is well below 1 and also below the test-retest correlation of either the baseline model or the model with respondent characteristics. In short, providing the LLM with respondent characteristics does affect its choices, even if the resulting choices do not agree meaningfully more with human choices.

Panels D and E examine whether average life-preference estimates become more accurate when the LLM is provided with respondent characteristics. The results in Panel D show that LLMs provided with respondent characteristics yield less accurate estimates of the valuation of the attribute bundles in each of the template stories: in all cases, the estimated slope moves further from 1 than the baseline estimates. Panel E examines the effects of providing the LLM with respondent characteristics on estimates for the two specific life preferences that we can estimate with sufficient precision for human respondents. Here, the results are mixed, with estimates staying roughly the same, improving, or deteriorating.

Altogether, these findings reinforce that the LLM is not ignoring the respondent characteristics: the correlations with baseline choices show that conditioning on respondent characteristics moves choices more than re-elicitation alone. However, this movement does not translate into an overall improvement in predicting human choices or human life preferences.

¹⁰We do not report the “Implied Majority Following Rate” because this parameter is no longer identified in the decomposition model of Section A.6 when the LLM is provided with respondent characteristics. Specifically, as explained in Proposition 2, identification of this parameter, labeled λ , rests on filtering out human preference variation, θ . However, the relevant preference heterogeneity should now be conditioned on respondent characteristics, and we cannot estimate preference heterogeneity conditional on respondent characteristics because the number of distinct respondents with identical demographics and attitudes is extremely small.

Table A4: Key Results When Conditioning on Respondent Characteristics

	GPT-5.4 Direct Choices			GPT-5.4 Reasoned Choices		
	Baseline (1)	Demographics (2)	Demographics + Attitudes (3)	Baseline (4)	Demographics (5)	Demographics + Attitudes (6)
Panel A: Internal Consistency						
Test-Retest Correlation	0.832 (0.012)	0.794 (0.013)	0.801 (0.013)	0.876 (0.011)	0.755 (0.015)	0.705 (0.016)
Implied Noise Rate	0.088 (0.007)	0.109 (0.008)	0.105 (0.007)	0.064 (0.006)	0.131 (0.008)	0.160 (0.009)
Panel B: Human Alignment						
Human-LLM Correlation	0.326 (0.007)	0.330 (0.007)	0.356 (0.007)	0.356 (0.007)	0.360 (0.007)	0.331 (0.007)
Panel C: Baseline Conditioning						
Correlation with Baseline Choices	—	0.721 (0.005)	0.648 (0.005)	—	0.614 (0.005)	0.591 (0.006)
Panel D: Story Valuation Structure						
Story Valuation Slope	1.002 (0.205)	1.098 (0.228)	0.838 (0.181)	1.281 (0.278)	1.435 (0.316)	0.410 (0.118)
Story Valuation R^2	0.844	0.733	0.815	0.585	0.680	0.319
Panel E: Attribute Valuations						
Longevity Valuation Gap (%)	44.39 (23.76)	61.38 (27.68)	35.26 (22.32)	126.49 (44.44)	95.27 (38.48)	63.67 (29.42)
Hardship-at-Work Valuation Gap (%)	-31.91 (17.60)	-16.26 (22.59)	-25.42 (19.48)	75.00 (48.07)	44.08 (42.09)	-52.50 (17.80)
First-Position Bias (normalized by human bias)	-5.15 (0.18)	-5.09 (0.20)	-2.25 (0.23)	1.68 (0.16)	1.74 (0.18)	1.79 (0.20)

Notes: This table reports key metrics across three elicitation conditions for GPT-5.4 with direct choices (columns 1–3) and GPT-5.4 with reasoned choices (columns 4–6). The *Baseline* condition uses the prompt from the main analysis; *Demographics* additionally provides the respondent’s age, gender, marital status, and number of children; *Attitudes* further adds responses to the 8-item questionnaire in Table A3. All estimates use inverse-probability weights (Appendix A.4).

The test-retest correlation is computed over 2,051 matched person-scenario pairs seen in both waves; human-LLM and baseline-persona correlations are computed over all 24,612 person-scenario-wave observations from respondents who completed both waves. Standard errors in Panels A–C use a respondent-level cluster bootstrap with 100,000 replications (Appendix A.7.1); those in Panels D–E use the delta method propagated through the full cross-equation SUR covariance matrix.

Appendix References

- Chen, Angelica, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho (2024) “Two Failures of Self-Consistency in the Multi-Step Reasoning of LLMs,” *Transactions on Machine Learning Research*.
- Chen, Guiming Hardy, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang (2024) “Humans or LLMs as the Judge? A Study on Judgement Biases,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327: Association for Computational Linguistics.
- Flood, Sarah, Miriam King, Renae Rodgers et al. (2025) “IPUMS CPS: Version 13.0 [dataset],” IPUMS, Minneapolis, MN.
- U.S. Census Bureau (2024) “Updates to Race/Ethnicity Standards for Our Nation,” <https://www.census.gov/about/our-research/race-ethnicity/standards-updates.html>.
- Xue, Mengge, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin (2024) “Strengthened Symbol Binding Makes Large Language Models Reliable Multiple-Choice Selectors,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4331–4344.
- Zellner, Arnold (1962) “An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias,” *Journal of the American Statistical Association*, 57 (298), 348–368.