

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.

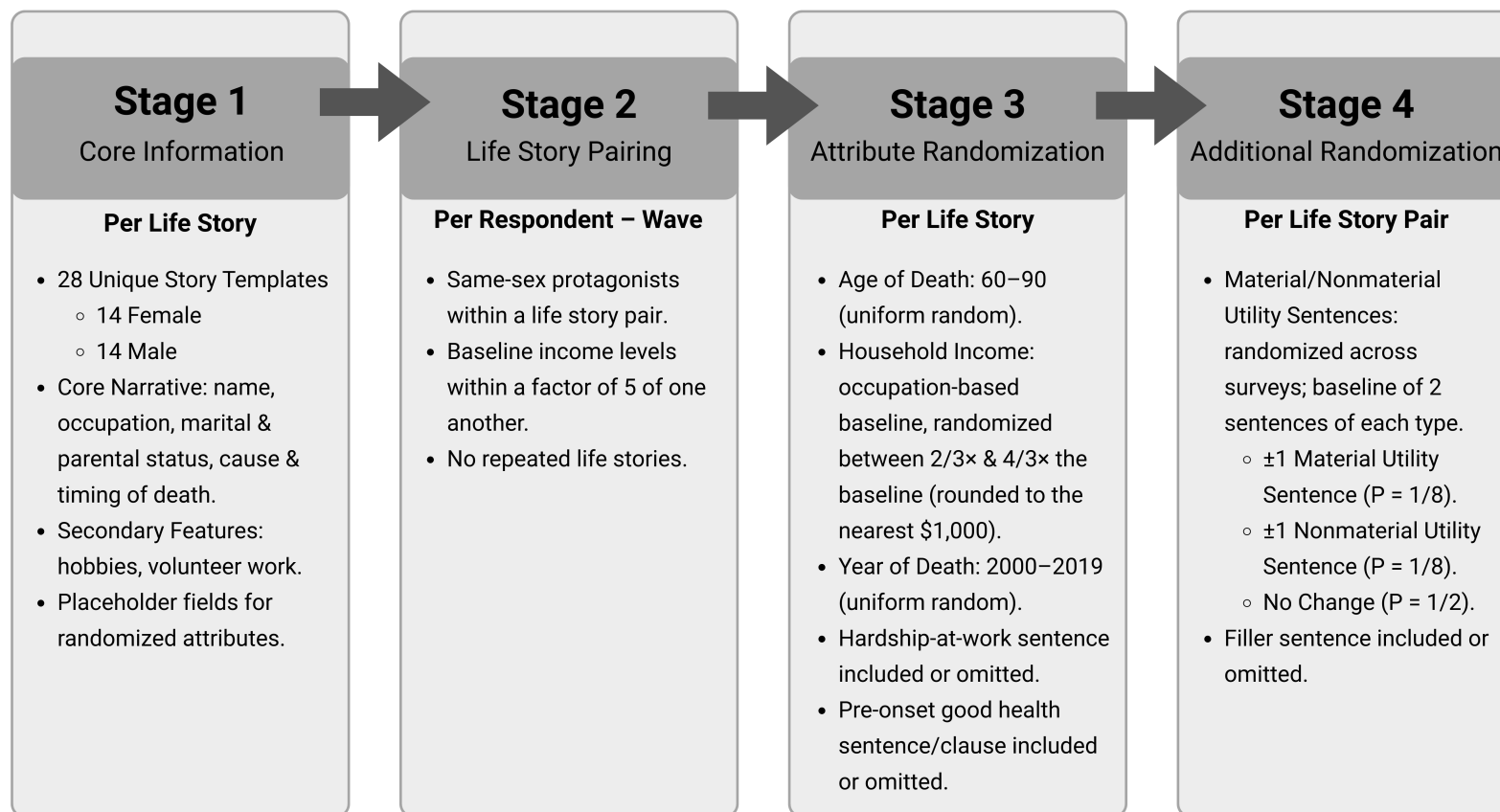
The siblings loved to sit around together with their instruments for hours, playing their old favorites, or working out new material.

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

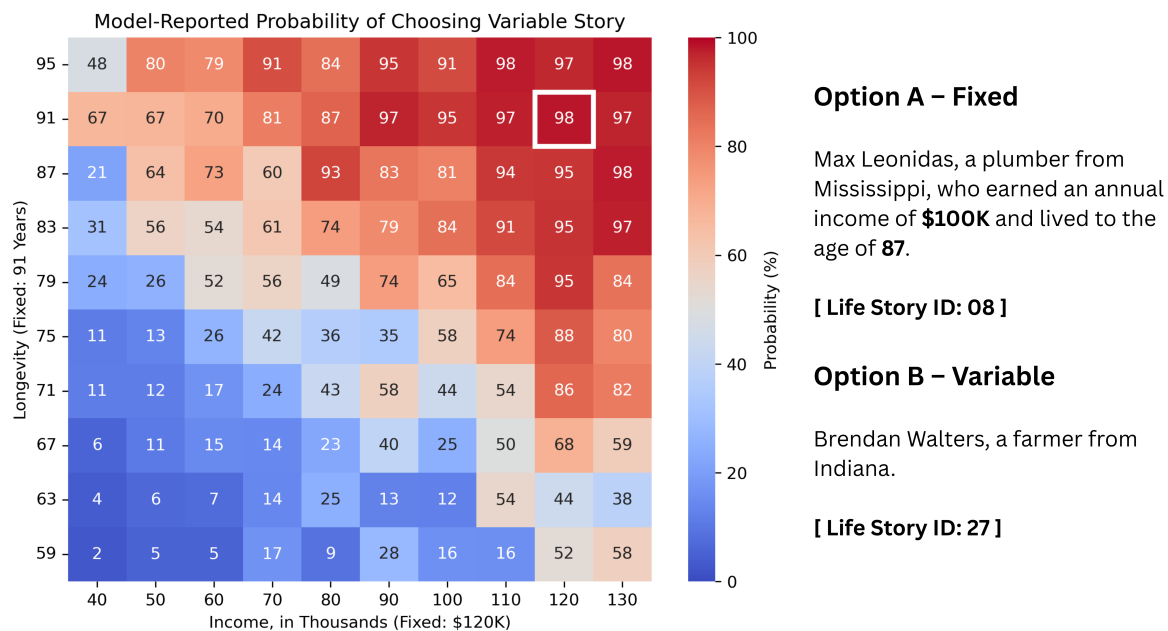
We assess whether the elicited LLM choices behave sensibly — responding systematically to economically relevant attributes rather than following a trivial rule or pure noise — by mapping choice probabilities over income and longevity holding narrative content fixed.

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.

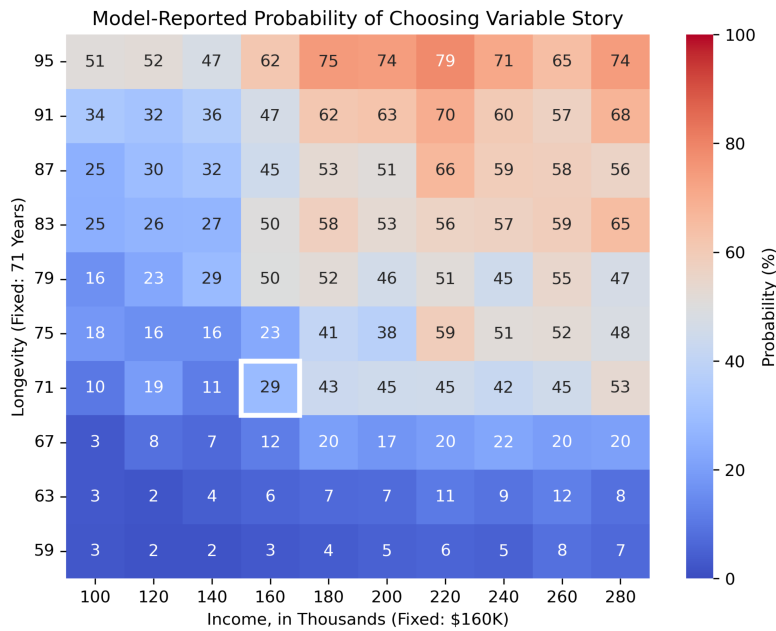
Figure A8 shows that choice probabilities generally increase with the variable story’s income and longevity, forming coherent (though noisy) indifference regions. The patterns are inconsistent with a purely mechanical heuristic that always favors higher income or longevity alone, and they are not dominated 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**.

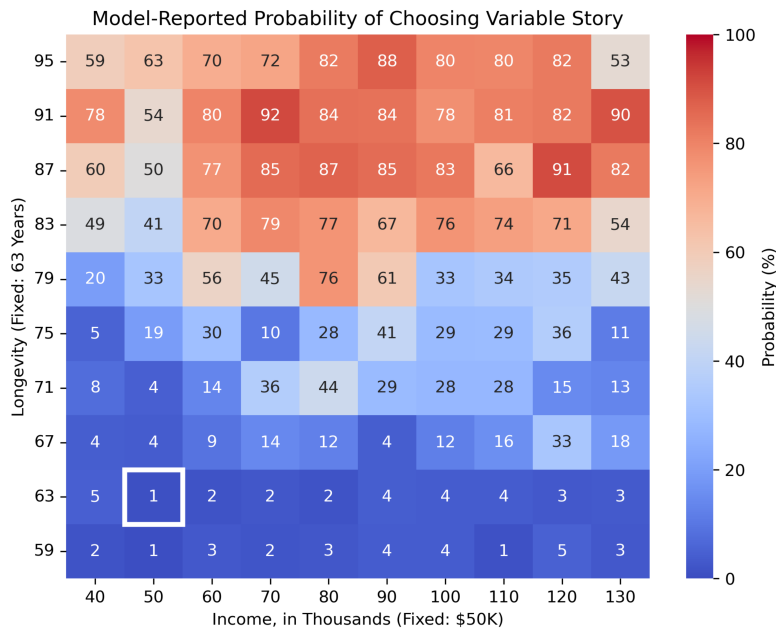
[Life Story ID: 02]

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.

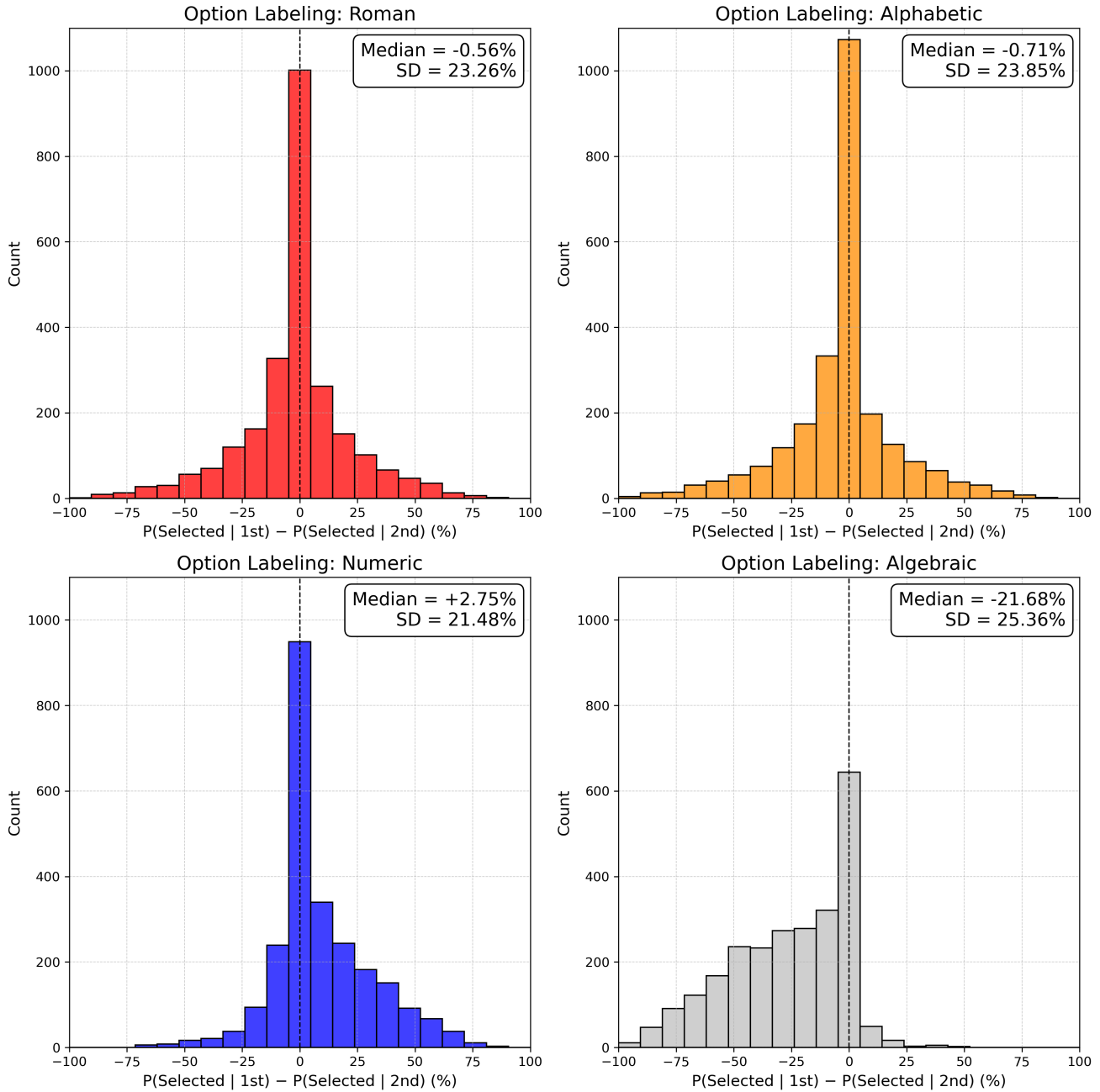
A.5.2 Sensitivity: Option Order & Token Bias

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.