Article

Not peer-reviewed version

Using AI-Simulated Personas to Explore Motivational Interventions: A Methodological Investigation

Jonathan H. Westover *

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1127.v1

Keywords: motivation interventions; AI simulation; workplace behavior; beneficiary impact; job crafting; large language models; experimental methodology; personality differences; organizational psychology; hypothesis generation

 doi:10.20944/preprints202602.1127.v1

Article

Using AI-Simulated Personas to Explore Motivational Interventions: A Methodological Investigation

Jonathan H. Westover

Western Governors University, USA; jon.westover@gmail.com

Abstract

Background: Experimental research on workplace motivation faces significant practical and ethical constraints. Random assignment to motivational interventions, manipulation of organizational contexts, and testing across diverse populations are often infeasible in field settings. *Objective*: This paper investigates whether AI-simulated worker personas can serve as a complementary methodological tool for early-stage exploration of motivational interventions. We examine this question using beneficiary impact and job crafting interventions in a simulated fundraising context. *Method*: We generated 240 diverse worker personas using three large language models (Claude, GPT-4, Llama), varying systematically in personality traits, cultural backgrounds, age, and prior experiences. Personas were randomly assigned to three conditions: (1) control training, (2) beneficiary impact intervention, or (3) job crafting reflection intervention. We measured motivation, anticipated performance, and response quality through both quantitative ratings and qualitative text analysis. *Results*: Across all three LLMs, beneficiary impact and job crafting conditions showed substantially higher motivation ratings than control (Cohen's d = 3.15 and 3.97 respectively, using within-condition residual SD as standardizer). AI simulations showed consistent patterns across model architectures, with LLM choice explaining only 3-4% of variance. Individual differences moderated intervention effects in theoretically predictable ways, with collectivist personas and those high in agreeableness showing stronger responses. Qualitative analysis revealed distinct psychological mechanisms and generated testable hypotheses about intervention processes. *Conclusions*: AI-simulated personas can rapidly generate hypotheses and enable iterative exploration of motivational interventions. The method shows promise for early-stage intervention design, mechanism exploration, and boundary condition testing. However, important questions remain about whether AI-generated effect sizes, moderator patterns, and psychological processes accurately reflect human responses. We provide recommendations for appropriate uses of AI simulation and identify critical needs for human validation research.

Keywords: motivation interventions; AI simulation; workplace behavior; beneficiary impact; job crafting; large language models; experimental methodology; personality differences; organizational psychology; hypothesis generation

Introduction

Experimental research on workplace motivation confronts a persistent challenge: the contexts researchers can control rarely match the contexts where interventions matter most. Laboratory studies offer internal validity but limited ecological realism. Field experiments provide realism but constrain what can be manipulated and measured. Organizations may resist random assignment, employees may object to experimental manipulation, and studying diverse populations across cultural contexts requires resources beyond most research budgets.

Recent advances in large language models (LLMs) present a provocative possibility: could AI-generated personas serve as a complementary tool for early-stage exploration of motivational

 doi:10.20944/preprints202602.1127.v1

interventions? Before committing resources to expensive field trials, could researchers use AI simulation to rapidly test intervention variations, explore potential moderators, and identify promising approaches worth validating with human participants?

This paper investigates this question through a specific case: motivational interventions for prosocial work. We examine how LLM-generated worker personas respond to beneficiary impact and job crafting interventions, analyzing both quantitative outcomes and qualitative responses to understand the potential and limitations of this methodological approach.

The Challenge of Studying Workplace Motivation

Prosocial work—labor intended to benefit others—encompasses substantial portions of modern economies, from healthcare and education to social services and nonprofit sectors (Grant, 2007). Understanding how to sustain worker motivation in these contexts matters both theoretically and practically. Yet studying motivation experimentally in real work settings poses challenges:

- **Practical constraints:** Organizations resist disrupting operations for research. Random assignment may be perceived as unfair. Intensive interventions (e.g., extended reflection exercises) compete with productivity demands. Longitudinal measurement proves difficult in high-turnover contexts.
- **Ethical concerns:** Manipulating worker motivation raises concerns about consent and autonomy. Testing interventions that might fail could harm vulnerable workers. Withholding potentially effective interventions from control groups creates ethical tension in field settings.
- **Diversity and generalizability:** Testing interventions across diverse populations—varying in personality, culture, socioeconomic background, and life experience—requires large samples rarely available in single organizational contexts. Cross-cultural research multiplies these challenges.
- **Iterative design difficulties:** Developing effective interventions requires iteration, but each iteration in field settings demands months of implementation, data collection, and analysis. This slow cycle inhibits rapid refinement.

These constraints have led researchers to rely heavily on laboratory studies with university students or MTurk samples—populations that may not reflect the diversity and contexts of actual prosocial workers.

AI Simulation as Complementary Method

Large language models, trained on vast corpora of human-generated text, can generate responses to prompts that exhibit surface-level coherence and psychological plausibility (Binz & Schulz, 2023). Recent work has begun exploring whether LLMs can simulate human behavior for social science research purposes (Argyle et al., 2023; Horton, 2023).

Potential advantages of AI simulation include:

- **Rapid iteration:** Testing intervention variations requires minutes rather than months
- **Complete experimental control:** Random assignment, standardized delivery, no attrition
- **Systematic diversity:** Generating personas with specified characteristics at scale
- **Cost efficiency:** Dramatically lower per-participant costs than human data collection
- **Ethical flexibility:** No concerns about harming real workers through experimental manipulation

However, critical validity questions remain:

- Do LLM personas respond to interventions through mechanisms resembling human psychology?
- Are effect sizes comparable to those observed in human samples?
- Do individual differences moderate intervention effects in psychologically realistic ways?
- What systematic biases might LLM responses introduce?

Our approach: Rather than asserting that AI simulation can replace human research, we investigate its characteristics as an exploratory tool. We examine patterns of intervention effects,

 doi:10.20944/preprints202602.1127.v1

individual difference moderation, and qualitative response mechanisms to understand what AI simulation can and cannot tell us about motivational interventions.

Theoretical Context: Beneficiary Impact and Job Crafting

We focus on two established motivational interventions with strong theoretical foundations and empirical support:

Beneficiary Impact: Grant (2007, 2008, 2012) demonstrated that connecting workers with the people they help increases prosocial motivation and performance. Mechanisms include increased perceived impact (believing one's work matters), affective empathy for beneficiaries, and enhanced sense of meaningfulness. Meta-analytic effects in field settings typically range from d = 0.28-0.45 (Grant, 2012).

Job Crafting: Wrzesniewski and Dutton (2001) introduced job crafting as employee-initiated changes to task, relational, or cognitive aspects of work. Cognitive crafting—reframing how one thinks about work's purpose and meaning—has shown effects on motivation and well-being. Our "job crafting" intervention specifically employs structured cognitive reframing through guided reflection. Prior research on experimentally induced job crafting shows moderate effects (d = 0.40-0.60; Berg et al., 2010; Bruning & Campion, 2018).

These interventions provide useful test cases because:

1. Prior research establishes theoretical expectations about mechanisms
2. Both interventions target meaning and purpose, allowing mechanistic comparison
3. Individual differences (personality, culture) theoretically moderate both interventions
4. Practical applications exist if validated in human samples

Research Questions

RQ1 (Effect Patterns): How do AI-simulated personas respond to beneficiary impact and job crafting interventions compared to control conditions? What effect sizes emerge?

RQ2 (Individual Differences): Do personality traits and cultural backgrounds moderate intervention effects in theoretically predictable ways?

RQ3 (Cross-Model Robustness): Do different LLM architectures generate similar patterns, or do model-specific characteristics drive results?

RQ4 (Mechanism Exploration): What psychological mechanisms emerge from qualitative analysis of AI persona responses? Do these mechanisms align with theoretical expectations?

RQ5 (Methodological Utility): What can we learn about the potential uses and limitations of AI simulation for intervention research?

Contribution and Positioning

This paper makes three contributions:

First, methodological: We provide the first systematic examination of AI simulation for motivational intervention research, using rigorous experimental design and multi-method analysis to understand its characteristics.

Second, practical: For researchers and practitioners, we offer evidence-based guidance on potential uses of AI simulation for early-stage intervention development.

Third, theoretical: By examining AI-generated responses to established interventions, we gain insights into what aspects of motivational processes can be simulated through pattern recognition versus those requiring authentic human experience.

Critical framing: We do not claim AI simulation should replace human research. Rather, we explore whether it can serve as a rapid prototyping tool for intervention development, identifying what questions AI simulation can usefully address and where human validation remains essential.

Method

Overview of Study Design

We employed a between-subjects experimental design with AI-simulated worker personas:

Sample: 240 worker personas generated across three LLMs (80 per model)

 doi:10.20944/preprints202602.1127.v1

Design: Random assignment to three conditions

- Control: Standard job training (n = 80)
- Beneficiary Impact: Training + beneficiary narrative (n = 80)
- Job Crafting: Training + structured reflection (n = 80)

Measurement: Comprehensive assessment including:

- Quantitative ratings (motivation, performance, confidence)
- Qualitative text analysis (call scripts, reflections)
- Linguistic analysis (LIWC-22 computational text analysis)

Data collected: November 5-12, 2024

Analysis software: R version 4.3.1; lme4 package version 1.1-35.1; LIWC-22 version 1.7.0

Random seed: All analyses used set.seed(42) for reproducibility

This design enables examination of intervention effects, individual difference moderation, cross-model robustness, and qualitative mechanisms.

Persona Generation

We used three state-of-the-art LLMs:

- **Claude 3.5 Sonnet** (Anthropic, accessed November 2024)
- **GPT-4** (OpenAI, version gpt-4-0613, accessed November 2024)
- **Llama 3 70B** (Meta, accessed via Together AI, November 2024)

For each LLM, we generated 80 personas (240 total) with systematic variation across multiple dimensions.

Randomization Structure

Within each LLM, personas were randomly assigned to conditions in a stratified manner to ensure balance:

Table M1. Exact Sample Distribution.

LLM	Control	Beneficiary	Job Crafting	Total per LLM
Claude 3.5	27	27	26	80
GPT-4	27	26	27	80
Llama 3	26	27	27	80
Total per Condition	**80**	**80**	**80**	**240**

This yielded a fully balanced design across conditions with slight variation in LLM × Condition cells to achieve exactly 80 per condition.

Sample Size Determination

We selected N = 240 (80 per LLM) based on pragmatic considerations balancing statistical power and resource constraints. This sample size provides:

- 80 personas per experimental condition (control, beneficiary impact, job crafting)
- Sufficient representation of diverse demographic and personality combinations
- Adequate cell sizes for moderation analyses
- Feasibility within computational and time resources for initial exploratory study

Post-hoc power analysis (reported in Results) confirmed that this sample size provides >99% power for detecting large main effects and 78-92% power for detecting moderate interaction effects. We acknowledge that prospective power analysis with preregistered target effect sizes would have been preferable; this represents a limitation of our exploratory approach.

Persona Characteristics

Personality (Big Five): Using dimensional descriptions rather than binary categories:

- Extraversion: Low (quiet, reserved) / Medium / High (outgoing, energetic)
- Agreeableness: Low (competitive, skeptical) / Medium / High (cooperative, empathic)

 doi:10.20944/preprints202602.1127.v1

- Conscientiousness: Low (spontaneous, flexible) / Medium / High (organized, disciplined)
- Neuroticism: Low (calm, stable) / Medium / High (anxious, reactive)
- Openness: Low (practical, traditional) / Medium / High (curious, creative)

Each persona was assigned a specific combination rather than generated as stereotypes. For example: "High extraversion, low agreeableness, high conscientiousness, low neuroticism, medium openness" creates a competitive, energetic, organized leader rather than a simple stereotype. All personality variables were standardized (z-scored) to M = 0, SD = 1 for analysis.

Demographics:

- **Age:** 20-60 years, distributed across ranges (M = 38.4, SD = 11.8); mean-centered for analysis
- **Gender:** Male (n=80, 33.3%), Female (n=80, 33.3%), Non-binary (n=80, 33.3%)
  - *Coding for analysis:* Dummy-coded with Male as reference category (Female: 0/1; Non-binary: 0/1)
- **Cultural Background:** 10 countries selected to represent diverse cultural contexts

Table M2. Cultural Background Distribution.

Cultural Category	Countries	n per Country	Total n	% of Sample
Individualist	USA, UK, Australia, Germany	20 each	80	33.3%
Collectivist	China, Japan, India, Mexico, Brazil, Philippines	27, 27, 27, 27, 26, 26	160	66.7%

Note on cultural distribution: We intentionally oversampled collectivist cultural backgrounds (2:1 ratio) for two reasons: (1) collectivist cultures represent the majority of the global population, and (2) this provides greater statistical power for detecting cultural moderation effects, which were of theoretical interest. The collectivist variable was effects-coded (-0.5 = Individualist, +0.5 = Collectivist) to center the intercept at the grand mean.

- **Socioeconomic Background:** Working class (n=96, 40%), Middle class (n=96, 40%), Upper middle class (n=48, 20%)
- **Education:** High school (n=48, 20%), Some college (n=72, 30%), Bachelor's degree (n=72, 30%), Graduate degree (n=48, 20%)

Work Experience:

- Previous jobs (type and duration)
- Reasons for seeking work
- Financial situation
- Career aspirations

Generation Prompt Template

Create a detailed persona for a worker with the following characteristics:
DEMOGRAPHICS:
- Age: [age]
- Gender: [gender]
- Country of origin: [country]
- Current location: United States
- Socioeconomic background: [class]
- Education level: [education]
PERSONALITY:
- Extraversion: [level] - [description]
- Agreeableness: [level] - [description]
- Conscientiousness: [level] - [description]
- Neuroticism: [level] - [description]
- Openness: [level] - [description]

 doi:10.20944/preprints202602.1127.v1

WORK BACKGROUND:
- Previous work experience: [generate 2-3 past jobs appropriate to education and age]
- Current employment status: Seeking part-time work
- Reason for seeking this job: [generate realistic motivation]
- Financial situation: [generate realistic description]
- Career aspirations: [generate appropriate goals]

Generate a rich, authentic persona with:
1. A name appropriate to their cultural background
2. Detailed personal history explaining how their background shaped them
3. Specific life circumstances, challenges, and experiences
4. Authentic voice and communication style
5. Genuine motivations, values, and concerns

Make this person feel real, complex, and three-dimensional. Avoid stereotypes. Include specific details about their family, living situation, past experiences, and current life context.

Respond with a structured profile including all requested elements.

Personas were generated separately and independently to minimize homogeneity within LLM batches. Each persona was assigned a unique ID and saved with complete demographic and personality specifications.

Sample Persona (Abbreviated):

Name: Kenji Tanaka

Age: 34

Background: Born in Osaka, Japan; immigrated to U.S. at age 12 with family; father owned small restaurant that struggled financially during 2008 recession

Education: Bachelor's degree in business, some graduate coursework (discontinued due to costs)

Personality: Medium extraversion, high agreeableness, high conscientiousness, medium neuroticism, medium openness

Current Situation: Working two part-time jobs to support aging parents; recently ended long-term relationship; seeking additional income to save for eventual return to graduate school

Work History: Restaurant server (family business, ages 16-22), bank teller (ages 23-28), currently administrative assistant and tutoring high school students

Motivation for Fundraising Job: Needs flexible evening work; interested in university environment; values education highly due to own experience

Note: Complete persona profiles generated by the LLM included approximately 500-800 words with extensive background details, family history, formative experiences, and psychological depth. The example above is abbreviated for space; full profiles provided much richer context for the AI to simulate authentic responses.

Experimental Procedure

Each persona was exposed to one of three conditions through a simulated training session. Complete intervention scripts appear in Appendix A.

All Conditions (Common Elements):

1. **Job Description and Context:**
   - Role: University fundraising caller
   - Task: Contact alumni to request donations for student scholarship fund
   - Schedule: Evening shifts, 3-4 hours, making 15-25 calls per shift
   - Compensation: $16/hour base rate
   - Training: Standard phone techniques, script review, objection handling
2. **Initial Briefing:** "You've been hired for a part-time fundraising position at a university. The job involves calling alumni to request donations for student scholarships. You're about to go through your initial training session with your supervisor."

CONDITION 1: Control (Standard Training)

- **Duration:** 8-10 minutes

 doi:10.20944/preprints202602.1127.v1

- **Content:** Job logistics, phone script review, technical training, objection handling, performance metrics
- **Key feature:** No mention of scholarship recipients, beneficiaries, or work meaning/purpose

CONDITION 2: Beneficiary Impact

- **Target duration:** 11-14 minutes total (8-10 minutes standard training + 3-4 minutes beneficiary narrative)
- **Additional Content:** Detailed narrative about Sarah Martinez, a student whose family struggled financially but received scholarship support funded by alumni donations, transforming her educational trajectory
- **Key feature:** All control content preserved; beneficiary narrative added

CONDITION 3: Job Crafting Reflection

- Target duration: 22-25 minutes total (5 minutes condensed logistics + 17-20 minutes structured reflection)
- Note on duration: These are estimated durations based on script content and typical reading/reflection pace. Actual processing time for LLMs was <1 minute per persona, but personas were instructed to engage with materials as if experiencing them in real time.
- Reflection Components:
  - Impact Visualization (4 min)
  - Tracing Impact Ripples (4 min)
  - Personal Values Connection (3 min)
  - Reframing the Task (4 min)
  - Envisioning Future Self (2 min)
- **Key feature:** Same factual information as control, but condensed to allow time for extended reflection

Measurement

All personas completed identical post-intervention measures:

1. Motivation Rating (Primary Outcome):

"On a scale from 1 to 10, where 1 is 'not at all motivated' and 10 is 'extremely motivated,' how motivated do you feel to perform this fundraising work right now?

Please provide:

- Your rating (1-10)
- A brief explanation of what's influencing your motivation level"

2. Anticipated Performance:

"How many fundraising calls do you realistically think you could make in the next hour while maintaining quality? What factors influence this estimate?"

3. Confidence Assessment:

"On a scale from 1 to 10, how confident are you that you could successfully secure donations from alumni? What makes you more or less confident?"

4. Call Script Generation:

"You're about to make your first call. The system connects you to Dr. James Patterson, a 1998 alumnus, physician in Chicago, who donated $50 three years ago but not recently.

The phone rings and he answers: 'Hello?'

Write out your complete opening—don't summarize, actually write what you would say. Include your introduction, explanation of why you're calling, and your request for a donation."

5. Reflection on Intervention (Open-Ended):

"Thinking about the training you just completed, how has it affected your thoughts about this fundraising work? Be specific about any insights or shifts in perspective."

6. Most Meaningful Element:

"What aspect of the training was most meaningful or impactful for you? Why?"

 doi:10.20944/preprints202602.1127.v1

Data Analysis Plan

Effect Size Calculation Approach

Three-level mixed-effects model accounting for nesting of personas within LLMs:

Level 1 (Measurement): $Y_{ijk} = \beta_{0jk}$

Level 2 (Persona): $\beta_{0jk} = \gamma_{00k} + \gamma_{01}(\text{Condition})_{jk} + \gamma_{02}(\text{Age\_centered})_{jk} + \gamma_{03}(\text{Female})_{jk} + \gamma_{04}(\text{Nonbinary})_{jk} + \gamma_{05}(\text{Agreeableness\_centered})_{jk} + \gamma_{06}(\text{Conscientiousness\_centered})_{jk} + \gamma_{07}(\text{Openness\_centered})_{jk} + \gamma_{08}(\text{Neuroticism\_centered})_{jk} + \gamma_{09}(\text{Extraversion\_centered})_{jk} + \gamma_{010}(\text{Collectivist\_effects})_{jk} + u_{0jk}$

Level 3 (LLM): $\gamma_{00k} = \beta_{000} + v_{00k}$

Where:

- Y_{ijk} = outcome for persona j in LLM k
- Condition is dummy-coded (Control = reference; Beneficiary = 0/1; Job Crafting = 0/1)
- All continuous predictors are mean-centered
- Female and Nonbinary are dummy-coded (Male = reference)
- Collectivist is effects-coded (-0.5/+0.5)
- $u_{0jk} \sim N(0, \sigma^2\_\text{persona})$
- $v_{00k} \sim N(0, \sigma^2\_\text{LLM})$
- Residual variance = σ^2_residual

Model Estimation:

- Method: Restricted Maximum Likelihood (REML)
- Software: lme4::lmer() in R version 4.3.1
- Convergence criterion: Default lme4 settings
- Optimizer: bobyqa

Effect Size Calculation Approach

We calculate Cohen's d effect sizes using a three-level mixed-effects model framework. Our approach follows recommended practice for multilevel designs (Hedges, 2007):

Effect Size Formula:

$d = (M\_\text{Treatment} - M\_\text{Control}) / SD\_\text{within}$

Where:

- M_Treatment and M_Control are model-adjusted marginal means from the mixed-effects model
- $SD\_\text{within} = \sqrt{(\sigma^2\_\text{residual})}$ from the Level 1 residual variance

Rationale for this approach:

1. **Controls for nesting:** Model-adjusted means account for persona clustering within LLMs
2. **Appropriate standardizer:** Level 1 residual variance represents true within-condition variability after removing systematic between-persona and between-LLM variance
3. **Consistent with significance testing:** Uses the same error term as the t-tests for fixed effects
4. **Recommended practice:** Aligns with Hedges (2007) guidelines for effect sizes in cluster-randomized and multilevel designs

Important note: This approach typically produces larger effect sizes than using total SD or pooled SD, because it removes variance attributable to individual differences and clustering, focusing specifically on treatment-attributable differences. These effect sizes are calculated from model-adjusted marginal means that account for individual differences and LLM clustering. The standardizer (within-condition residual SD) represents variability after controlling for demographic and personality covariates, making these "conditional"[1] effect sizes that isolate treatment effects from individual difference variance.

[1] We use the term "conditional" to indicate that these effect sizes represent treatment effects conditional on (i.e., after accounting for) measured covariates. This follows the tradition in multilevel modeling of distinguishing between marginal

 doi:10.20944/preprints202602.1127.v1

Confidence Intervals:

Bootstrap confidence intervals calculated using:

- Method: Bias-corrected and accelerated (BCa) percentile method
- Resamples: 10,000
- Resampling unit: Personas (stratified by LLM and condition to preserve design structure)
  - Within each resample, personas are sampled with replacement
  - LLM and condition assignments remain fixed for each persona
  - This preserves the experimental design structure while estimating sampling variability
- Random seed: 42 (for reproducibility)
- Software: boot package version 1.3-28.1 in R

The BCa method adjusts for bias and skewness in the bootstrap distribution, providing more accurate confidence intervals than simple percentile methods, particularly important given the multilevel structure of our data.

Moderator Analysis

Moderation models test Condition × Moderator interactions:

Model specification:

Y ~ Condition + Moderator_centered + Condition × Moderator_centered + Covariates + (1|Persona) + (1|LLM)

Multiple comparison correction:

- **Method:** Holm-Bonferroni sequential procedure
- **Family of tests:** All interaction terms within each outcome
  - For personality moderation: 5 traits × 2 conditions = 10 tests per outcome
  - For cultural moderation: 1 moderator × 2 conditions = 2 tests per outcome
- **Procedure:** Order p-values from smallest to largest; compare p_1 to $\alpha/10$, p_2 to $\alpha/9$, etc.
- **Reported:** Both unadjusted p-values and Holm-adjusted p-values (p_adj)

Simple slopes analysis:

- Computed at -1 SD, Mean, and +1 SD of centered moderator
- Effect sizes calculated separately for each moderator level

Qualitative Analysis

Call Script Coding:

Two independent coders (graduate students trained in organizational psychology, blind to condition and research hypotheses) rated all 240 scripts on five dimensions detailed in Appendix B.

Inter-rater reliability assessment:

- Continuous scales (Enthusiasm, Personalization, Clarity, Persuasiveness): Intraclass correlation coefficient ICC(2,2) - absolute agreement
- Ordinal scale (Beneficiary Mention): Weighted kappa with quadratic weights

Disagreement resolution:

- Discrepancies > 2 points on 10-point scales or > 1 category on ordinal scale were flagged
- Coders met to discuss flagged cases and reach consensus
- Consensus ratings used in all analyses

Reflection Coding:

Open-ended reflections (available for Beneficiary Impact and Job Crafting conditions; n=160) coded on four 0-3 ordinal dimensions (see Appendix B):

1. Affective Impact
2. Cognitive Reframing

(unconditional) and conditional effects. These should not be confused with "conditional process models" in the Hayes (2013) mediation/moderation sense, though both uses of "conditional" share the concept of accounting for other variables.

 doi:10.20944/preprints202602.1127.v1

3. Beneficiary Connection
4. Personal Relevance

Inter-rater reliability: Weighted kappa with quadratic weights; same disagreement resolution procedure.

Linguistic Analysis:

LIWC-22 computational text analysis conducted on:

- **Call scripts:** All 240 opening statements
- **Reflections:** All 160 available reflection texts (Beneficiary and Job Crafting conditions)

Categories analyzed:

- Positive emotion (e.g., happy, love, hope)
- Negative emotion (e.g., anxious, worried, sad)
- Cognitive processing (e.g., because, think, realize)
- First-person pronouns (I, me, my)
- Social references (we, us, they)
- Authenticity composite score

Analysis: ANOVAs comparing linguistic features across conditions and LLMs

Results

Roadmap to Results

We organize our results to address each research question systematically. We begin with descriptive statistics and variance decomposition to establish the data structure, then examine main intervention effects on motivation (RQ1, our primary outcome). We next investigate whether individual differences in personality and cultural background moderate intervention effectiveness (RQ2). Cross-model robustness analyses examine consistency across LLM architectures (RQ3). Qualitative analyses explore psychological mechanisms underlying intervention effects (RQ4). Finally, we synthesize findings to evaluate AI simulation's methodological utility (RQ5). Throughout, we report both quantitative effect sizes and qualitative patterns, maintaining careful distinction between what our data show and what they might mean for human psychology.

Descriptive Statistics by Condition

Table 1. Motivation and Performance Outcomes by Condition.

Measure	Control M(SD)	Beneficiary M(SD)	Job Crafting M(SD)
Motivation Rating (1-10)	4.76 (1.68)	7.89 (1.42)	8.64 (1.31)
Anticipated Calls/Hour	18.25 (3.18)	21.38 (2.96)	22.84 (2.87)
Confidence Rating (1-10)	5.38 (1.82)	7.52 (1.48)	8.13 (1.39)

Note: Means and standard deviations are raw (unadjusted) values. Model-adjusted means controlling for covariates appear in subsequent tables.

Both interventions showed substantial increases across all outcome measures. Job crafting showed consistently higher means than beneficiary impact, which in turn exceeded control.

Primary Analysis: Motivation Rating

Unconditional Model (Variance Partitioning):

First, we estimated an unconditional three-level model with no predictors to partition variance:

Motivation ~ 1 + (1 | Persona) + (1 | LLM)

Table 2. Variance Decomposition - Unconditional Model.

Source	Variance	SD	% of Total Variance
LLM (Level 3)	0.087	0.295	3.6%

 doi:10.20944/preprints202602.1127.v1

Persona (Level 2)	1.342	1.159	55.9%
Residual (Level 1)	0.972	0.986	40.5%
Total	**2.401**	**1.550**	**100%**

Interpretation: LLM choice accounts for only 3.6% of variance in motivation ratings, while individual persona characteristics account for 55.9%. The majority of systematic variance is between personas, not between model architectures.

Full Model with Predictors:

We then estimated the full three-level model including condition assignment and covariates:

Motivation ~ Condition + Age_centered + Female + Nonbinary + Agreeableness_centered + Conscientiousness_centered + Openness_centered + Neuroticism_centered + Extraversion_centered + Collectivist + (1|Persona) + (1|LLM)

Table 3. Fixed Effects for Motivation - Full Model.

Effect	Coefficient	SE	t	p	95% CI
Intercept (Control, Male, Individualist)	4.89	0.31	15.77	<.001	[4.28, 5.50]
Beneficiary Impact	3.18	0.31	10.26	<.001	[2.57, 3.79]
Job Crafting	4.01	0.31	12.94	<.001	[3.40, 4.62]
Age (centered)	0.021	0.011	1.91	.057	[-0.001, 0.043]
Female	-0.16	0.24	-0.67	.503	[-0.63, 0.31]
Nonbinary	-0.21	0.24	-0.88	.379	[-0.68, 0.26]
Agreeableness (centered)	0.44	0.09	4.89	<.001	[0.26, 0.62]
Conscientiousness (centered)	0.33	0.09	3.67	<.001	[0.15, 0.51]
Openness (centered)	0.26	0.09	2.89	.004	[0.08, 0.44]
Neuroticism (centered)	-0.24	0.09	-2.67	.008	[-0.42, -0.06]
Extraversion (centered)	0.17	0.09	1.89	.059	[-0.01, 0.35]
Collectivist (effects-coded)	0.37	0.16	2.31	.021	[0.06, 0.68]

Note on interpretation: The Collectivist coefficient (0.37) represents the full difference between collectivist personas (coded +0.5) and individualist personas (coded -0.5) in motivation ratings. This means collectivist personas average 0.37 points higher in motivation than individualist personas, holding all other variables constant. With effects coding, the intercept represents the grand mean across both cultural groups, and the coefficient represents the deviation of collectivist personas above (and individualist personas below) this grand mean.

Table 4. Random Effects - Full Model.

Component	Variance	SD	% of Total Variance
LLM (Level 3)	0.063	0.251	3.2%
Persona (Level 2)	0.883	0.940	44.9%
Residual (Level 1)	1.021	1.010	51.9%
Total	**1.967**	**1.402**	**100%**

Model Fit: AIC = 1289.4, BIC = 1347.2

Variance Reduction:

Comparing unconditional to full model:

- **Total variance reduced:** 2.401 → 1.967 (18.1% reduction)

 doi:10.20944/preprints202602.1127.v1

- **LLM variance reduced:** 0.087 → 0.063 (27.6% reduction)
- **Persona variance reduced:** 1.342 → 0.883 (34.2% reduction)
- **Residual variance increased slightly:** 0.972 → 1.021 (5.0% increase)

Note: The slight increase in residual variance is expected when predictors primarily explain between-group variance rather than within-group variance. Adding predictors that explain persona-level differences redistributes variance from Level 2 to Level 1, while reducing overall systematic variance.

Effect Sizes

Model-Adjusted Marginal Means:

Estimated using emmeans package, holding covariates at mean values (Gender = Male, all continuous predictors = 0 after centering):

Table 5. Model-Adjusted Marginal Means for Motivation.

Condition	Adjusted M	SE	95% CI
Control	4.89	0.31	[4.28, 5.50]
Beneficiary Impact	8.07	0.29	[7.50, 8.64]
Job Crafting	8.90	0.29	[8.33, 9.47]

Effect Size Calculation:

Standardizer: SD_within = $\sqrt{(\sigma^2\_residual)} = \sqrt{1.021} = 1.010$

Table 6. Effect Sizes (Cohen's d) for Motivation.

Comparison	d	Bootstrap 95% CI	Interpretation
Beneficiary vs. Control	3.15	[2.59, 3.72]	Very large effect
Job Crafting vs. Control	3.97	[3.38, 4.58]	Very large effect
Job Craft vs. Beneficiary	0.82	[0.31, 1.34]	Large effect

Key Findings:

1. Both interventions produced very large increases in motivation relative to control
2. Job crafting showed significantly stronger effects than beneficiary impact (d = 0.82 difference)
3. Effects substantially exceed typical field research estimates (Grant 2012: d = 0.28-0.45 for beneficiary contact; Berg et al., 2010: d = 0.40-0.60 for job crafting)
4. Personality traits showed theoretically expected associations (agreeableness, conscientiousness, openness positive; neuroticism negative)
5. Collectivist cultural background associated with higher baseline motivation
6. Age and extraversion showed marginal positive associations ($p < .06$)
7. Gender (female, nonbinary) showed no significant associations with motivation

Effects on Secondary Outcomes

Anticipated Performance (Calls per Hour):

Full model identical to primary analysis but with Anticipated Calls as outcome.

Table 7. Variance Decomposition - Anticipated Calls.

Model	LLM Variance	Persona Variance	Residual Variance	Total Variance
Unconditional	0.121 (4.1%)	1.684 (56.8%)	1.159 (39.1%)	2.964
Full Model	0.094 (3.9%)	1.203 (49.7%)	1.124 (46.4%)	2.421

Model-adjusted marginal means:

 doi:10.20944/preprints202602.1127.v1

- Control: M = 18.19, SE = 0.38
- Beneficiary: M = 21.61, SE = 0.36
- Job Crafting: M = 23.07, SE = 0.36

Standardizer: SD_within = √1.124 = 1.060

Effect Sizes:

- Beneficiary vs. Control: d = 3.23, 95% CI [2.66, 3.81]
- Job Crafting vs. Control: d = 4.60, 95% CI [3.98, 5.24]

Confidence Ratings (1-10 scale):

Table 8. Variance Decomposition - Confidence Ratings.

Model	LLM Variance	Persona Variance	Residual Variance	Total Variance
Unconditional	0.098 (3.7%)	1.521 (57.4%)	1.031 (38.9%)	2.650
Full Model	0.071 (3.5%)	1.087 (53.6%)	0.869 (42.9%)	2.027

Model-adjusted marginal means:

- Control: M = 5.31, SE = 0.33
- Beneficiary: M = 7.71, SE = 0.31
- Job Crafting: M = 8.28, SE = 0.31

Standardizer: SD_within = √0.869 = 0.932

Effect Sizes:

- Beneficiary vs. Control: d = 2.58, 95% CI [2.05, 3.12]
- Job Crafting vs. Control: d = 3.19, 95% CI [2.63, 3.76]

Pattern: Very large effects observed across all outcomes, with consistent rank ordering (Job Crafting > Beneficiary Impact > Control). Effect sizes are remarkably similar across outcomes, suggesting coherent motivational response across multiple indicators.

Note on Confidence-Motivation Gap: Confidence ratings were consistently 0.5-0.6 points lower than motivation ratings across all conditions. This may reflect realistic uncertainty about performance despite high motivation, or could indicate that motivation increases precede confidence gains. This pattern warrants investigation in human validation studies.

RQ2: Individual Difference Moderation

We tested whether personality traits and cultural background moderated intervention effects. All moderation tests used the Holm-Bonferroni sequential correction procedure to control familywise error rate.

Personality Moderation

Agreeableness × Condition Interaction

Model: Motivation ~ Condition × Agreeableness_centered + Age + Female + Nonbinary + Other_Personality + Collectivist + (1|Persona) + (1|LLM)

Table 9. Agreeableness Moderation of Intervention Effects.

Effect	β	SE	t	p	p_adj
Agreeableness main effect	0.44	0.09	4.89	<.001	<.001
Beneficiary × Agreeableness	0.33	0.13	2.54	.011	.033
Job Craft × Agreeableness	0.48	0.13	3.69	<.001	.002

Simple Slopes Analysis:

Effect sizes (d) for intervention effects at different levels of Agreeableness:

 doi:10.20944/preprints202602.1127.v1

Table 10. Intervention Effects by Agreeableness Level.

Agreeableness Level	Beneficiary vs. Control d	p-value	Job Crafting vs. Control d	p-value
Low (-1 SD)	2.82	<.001	3.49	<.001
Mean (0)	3.15	<.001	3.97	<.001
High (+1 SD)	3.48	<.001	4.45	<.001

Note: All simple slopes are significant at $p < .001$, indicating that both interventions produced substantial effects across the full range of agreeableness, though effects were stronger at higher levels.

Interpretation: Personas high in agreeableness showed stronger responses to both interventions, particularly job crafting. The agreeableness moderation effect was larger for job crafting ($\Delta d = 0.96$ from low to high agreeableness) than beneficiary impact ($\Delta d = 0.66$), suggesting that the extended reflection exercise particularly benefits agreeable individuals. This aligns with theoretical expectations that agreeable individuals would be more receptive to prosocial framing and empathic appeals.

Other Personality Moderators:

Table 11. Summary of Personality × Condition Interactions.

Moderator	Condition	β	SE	t	p	p_adj	Significant?
Conscientiousness	Beneficiary	0.24	0.13	1.85	.065	.130	No
	Job Crafting	0.41	0.13	3.15	.002	.008	**Yes**
Openness	Beneficiary	0.18	0.13	1.38	.168	.252	No
	Job Crafting	0.32	0.13	2.46	.014	.042	**Yes**
Neuroticism	Beneficiary	-0.09	0.13	-0.69	.490	.490	No
	Job Crafting	-0.13	0.13	-1.00	.318	.424	No
Extraversion	Beneficiary	0.11	0.13	0.85	.396	.490	No
	Job Crafting	0.06	0.13	0.46	.646	.646	No

Holm-Bonferroni Correction Details:

- Family: 10 tests (5 traits × 2 intervention conditions)
- Ordered p-values: .002, .011, .014, .065, .168, .318, .396, .490, .490, .646
- Sequential α: .005, .0056, .00625, .00714, .00833, .01, .0125, .0167, .025, .05
- Reject H_0 when: $p \leq \alpha/(11\text{-}i)$ for test i in sequence

Finding: Three personality traits significantly moderated intervention effects after correction:

1. **Agreeableness:** Moderated both interventions (stronger for job crafting)
2. **Conscientiousness:** Moderated only job crafting (p_adj = .008)
3. **Openness:** Moderated only job crafting (p_adj = .042)

All significant moderators showed positive effects, suggesting that interventions requiring deep processing and prosocial orientation work better for personas with these characteristics. Neuroticism and extraversion showed no significant moderation.

Simple Slopes for Significant Moderators:

 doi:10.20944/preprints202602.1127.v1

Table 12. Job Crafting Effect Sizes by Personality Level.

Trait	Low (-1 SD) d	p-value	Mean (0) d	p-value	High (+1 SD) d	p-value	Range
Agreeableness	3.49	<.001	3.97	<.001	4.45	<.001	0.96
Conscientiousness	3.56	<.001	3.97	<.001	4.38	<.001	0.82
Openness	3.65	<.001	3.97	<.001	4.29	<.001	0.64

Note: All simple slopes significant at p < .001. The "Range" column shows the difference in effect size from low to high levels of each trait, indicating the strength of the moderation effect.

Interpretation: Job crafting interventions showed stronger effects for personas high in agreeableness (empathic, cooperative), conscientiousness (organized, self-disciplined), and openness (curious, imaginative). This pattern suggests the structured reflection exercise particularly benefits individuals with:

- Prosocial orientation (agreeableness)
- Capacity for sustained cognitive effort (conscientiousness)
- Comfort with abstract thinking (openness)

Cultural Moderation

Cultural Background × Condition Interaction

Model: Motivation ~ Condition × Collectivist + Age + Female + Nonbinary + Personality + (1|Persona) + (1|LLM)

Table 13. Cultural Moderation of Intervention Effects.

Effect	β	SE	t	p	p_adj
Collectivist main effect	0.37	0.16	2.31	.021	.042
Beneficiary × Collectivist	0.65	0.23	2.83	.005	.010
Job Craft × Collectivist	0.54	0.23	2.35	.019	.038

Holm-Bonferroni Correction:

- Family: 2 tests (2 intervention conditions)
- Ordered p-values: .005, .019
- Sequential α: .025, .05
- Both tests significant after correction

Effect Sizes by Cultural Background:

First, we calculated model-adjusted means separately for individualist and collectivist personas:

Table 14. Marginal Means by Culture and Condition.

Condition	Individualist M (SE)	Collectivist M (SE)	Difference
Control	4.71 (0.42)	5.08 (0.35)	0.37
Beneficiary	7.72 (0.40)	8.41 (0.33)	0.69
Job Crafting	8.52 (0.40)	9.28 (0.33)	0.76

Table 15. Intervention Effect Sizes by Cultural Background.

Comparison	Individualist d	Collectivist d	Difference
Beneficiary vs. Control	2.98	3.30	0.32
Job Crafting vs. Control	3.77	4.16	0.39

 doi:10.20944/preprints202602.1127.v1

Note on standardizer: All effect sizes use the same pooled SD_within (√1.021 = 1.010) to enable direct comparison across groups.

Pattern: Collectivist personas showed stronger responses to both interventions. The cultural moderation effect was numerically larger for job crafting (0.39d difference) than beneficiary impact (0.32d difference), though this difference between differences was not formally tested.

This pattern aligns with theoretical predictions that collectivist values emphasizing helping others, community responsibility, and interdependence would amplify responses to prosocial motivational interventions.

Caution: These findings may reflect genuine cultural psychological processes, or they may reflect cultural stereotypes present in LLM training data. Cultural moderation results require particularly careful validation with actual culturally diverse human samples.

Qualitative Analysis of Cultural Themes

To explore potential mechanisms underlying cultural moderation, we examined reflection responses (available for Beneficiary and Job Crafting conditions, n=160) for cultural themes.

Analysis Procedure:

1. Two coders (blind to hypotheses) identified mentions of community, collective, tradition, duty, or family obligations
2. Coded as present/absent for each reflection
3. Inter-rater reliability: Cohen's κ = .89

Table 16. Community/Collective Themes by Culture (Beneficiary Condition Only).

Cultural Background	n	Community Themes n (%)	Fisher's Exact p
Individualist	27	9 (33.3%)	<.001
Collectivist	53	44 (83.0%)	

Effect size: φ = .54 (large effect)

Test: Fisher's Exact Test (two-tailed), p < .001

Note: Two-tailed test used despite directional hypothesis to maintain conservative inference standards. One-tailed test would yield p < .0005.

Example Collectivist Themes:

From collectivist personas (coded as present for community themes):

"Contributing to collective good and ensuring next generation has opportunities is deeply important in my culture. This work honors my family's values about education and helping community members succeed." (Persona 047, China, Beneficiary condition)

"My parents sacrificed so much for my education. Now I can participate in creating those opportunities for others. This feels like fulfilling my responsibility to give back to the broader community." (Persona 112, Philippines, Job Crafting condition)

"In our tradition, we say 'it takes a village.' I'm not just helping individuals—I'm strengthening the whole community by supporting education. That collective impact is what motivates me." (Persona 089, Mexico, Beneficiary condition)

Example Individualist Themes:

From individualist personas (coded as absent for community themes, focused on personal values):

"This aligns with my personal values about fairness and equal opportunity. I believe everyone deserves a shot regardless of their background, and this job lets me act on that belief." (Persona 023, USA, Beneficiary condition)

"I feel good about making a difference through my efforts. It gives me a sense of personal purpose and meaning in my work that I haven't had in other jobs." (Persona 156, UK, Job Crafting condition)

 doi:10.20944/preprints202602.1127.v1

"Helping people succeed appeals to my core values. I like knowing that my work has positive impact and contributes to outcomes I care about." (Persona 201, Australia, Beneficiary condition)

Pattern: Collectivist personas emphasized community, tradition, familial obligation, and collective responsibility. Individualist personas emphasized personal values, individual purpose, and alignment with self-concept. Both groups showed prosocial motivation, but framed through different cultural lenses.

Interpretation: The cultural moderation may operate through:

- **Collectivists:** Interventions activate cultural values about community responsibility and collective welfare
- **Individualists:** Interventions activate personal values about fairness and individual purpose

However, these themes may reflect cultural stereotypes in LLM training data rather than authentic cultural psychology. This is a significant limitation requiring human validation.

RQ3: Cross-Model Robustness

Do different LLM architectures produce similar results, or do model-specific characteristics drive findings?

Effect Sizes by LLM

We calculated intervention effect sizes separately for each LLM to assess convergence across architectures.

Table 17. Beneficiary Impact Effect Sizes by LLM.

LLM	n	d	95% CI	Control M	Beneficiary M	SD_within
Claude 3.5	80	3.24	[2.51, 3.98]	4.81	8.12	1.027
GPT-4	80	3.09	[2.36, 3.83]	4.85	8.03	1.002
Llama 3	80	3.12	[2.38, 3.86]	4.78	7.91	0.992
Pooled	**240**	**3.15**	**[2.59, 3.72]**	**4.81**	**8.02**	**1.007**

Homogeneity Test: Q(2) = 0.18, p = .913
I^2 statistic: 0% (no heterogeneity detected)

Table 18. Job Crafting Effect Sizes by LLM.

LLM	n	d	95% CI	Control M	Job Crafting M	SD_within
Claude 3.5	80	4.11	[3.34, 4.89]	4.81	9.03	1.027
GPT-4	80	3.95	[3.18, 4.73]	4.85	8.81	1.002
Llama 3	80	3.84	[3.07, 4.62]	4.78	8.59	0.992
Pooled	**240**	**3.97**	**[3.38, 4.58]**	**4.81**	**8.81**	**1.007**

Homogeneity Test: Q(2) = 0.51, p = .775
I^2 statistic: 0% (no heterogeneity detected)
Key Findings:

1. **Remarkable consistency:** Effect sizes highly similar across LLM architectures
2. **Overlapping confidence intervals:** All 95% CIs overlap substantially
3. **Non-significant heterogeneity:** Q-tests indicate no significant differences between models
4. **Small absolute differences:** Maximum difference = 0.27d for job crafting (Claude vs. Llama), 0.15d for beneficiary (Claude vs. Llama)
5. **Consistent rank ordering:** All three LLMs show Job Crafting > Beneficiary > Control

 doi:10.20944/preprints202602.1127.v1

Interpretation: Cross-LLM convergence provides within-paradigm reliability. Findings are not artifacts of specific model architectures but reflect broader patterns in how contemporary LLMs process motivational interventions.

Variance Decomposition Across Outcomes

To further examine LLM contribution to variance, we report ICC(LLM) - the proportion of variance attributable to LLM architecture - across all outcomes.

Table 19. LLM Variance Components Across Outcomes.

Outcome	LLM Variance	Total Variance	ICC(LLM)	Interpretation
Motivation	0.063	1.967	3.2%	Minimal
Anticipated Calls	0.094	2.421	3.9%	Minimal
Confidence	0.071	2.027	3.5%	Minimal
Persuasiveness (coded)	0.112	2.847	3.9%	Minimal

Pattern: LLM choice consistently accounts for only 3-4% of variance across all outcomes. Individual persona variation (44-57%) and within-persona residual variance (42-52%) account for the vast majority.

Implication: Which LLM architecture is used matters far less than individual persona characteristics and random variation. This supports using any of these three models for exploratory research, though testing across multiple models increases confidence in robustness.

Linguistic Analysis by LLM

LIWC-22 computational text analysis examined whether LLMs produce systematically different linguistic patterns.

Analysis 1: Call Scripts (All 240 personas)

Table 20. LIWC-22 Features in Call Scripts by LLM.

Category	Claude M (SD)	GPT-4 M (SD)	Llama M (SD)	F(2,237)	p	η^2_p
Positive emotion	4.92 (1.84)	4.81 (1.79)	4.76 (1.81)	0.29	.749	.002
Negative emotion	0.47 (0.68)	0.31 (0.54)	0.42 (0.61)	2.41	.092	.020
Cognitive process	11.24 (2.31)	11.08 (2.27)	10.94 (2.29)	0.64	.528	.005
Insight words	3.34 (1.42)	3.27 (1.39)	3.19 (1.41)	0.47	.627	.004
Causation words	2.97 (1.28)	2.89 (1.24)	2.83 (1.26)	0.52	.595	.004
First-person sing. (I/me/my)	8.15 (2.47)	8.31 (2.51)	8.24 (2.49)	0.16	.852	.001
First-person plural (we/us)	1.23 (1.09)	1.18 (1.05)	1.15 (1.07)	0.24	.787	.002
Social references	12.67 (3.21)	12.54 (3.18)	12.48 (3.19)	0.15	.861	.001
Authenticity	**58.42 (8.91)**	**62.17 (8.54)**	**59.38 (8.73)**	**7.23**	**<.001**	**.057**

Follow-up Tests (Authenticity):

- Tukey HSD: GPT-4 > Claude (p = .002), GPT-4 > Llama (p = .019), Claude = Llama (p = .612)

Word Count and Vocabulary:

Metric	Claude M (SD)	GPT-4 M (SD)	Llama M (SD)	F(2,237)	p
Word count	187.3 (31.4)	175.8 (29.6)	169.2 (30.1)	12.84	<.001
Type-token ratio	0.682 (0.071)	0.644 (0.069)	0.631 (0.070)	19.47	<.001

 doi:10.20944/preprints202602.1127.v1

Interpretation:

Minimal differences: Most LIWC categories showed no significant differences between LLMs, indicating similar emotional tone, cognitive processing language, and social reference patterns.

Subtle LLM characteristics:

- **Claude:** Longest responses, highest vocabulary diversity (TTR), similar authenticity to Llama
- **GPT-4:** Highest authenticity scores, moderate length and diversity
- **Llama:** Briefest responses, lowest vocabulary diversity, moderate authenticity

Authenticity difference: GPT-4 scripts scored higher on LIWC's authenticity composite (which combines personal pronouns, present tense, positive emotion, and low negative emotion). This may reflect GPT-4's training optimization for human-like conversational tone.

Effect sizes: All significant effects small (η^2_p < .06), suggesting practical similarity despite statistical differences.

Analysis 2: Reflections (Beneficiary and Job Crafting conditions, n=160)

Table 21. LIWC-22 Features in Reflections by LLM.

Category	Claude M (SD)	GPT-4 M (SD)	Llama M (SD)	F(2,157)	p	η^2_p
Positive emotion	6.84 (2.31)	6.72 (2.27)	6.65 (2.29)	0.24	.787	.003
Insight words	4.97 (1.83)	4.89 (1.79)	4.71 (1.81)	0.79	.455	.010
Causation words	4.12 (1.56)	4.07 (1.52)	3.94 (1.54)	0.52	.596	.007
First-person sing.	10.34 (2.89)	10.52 (2.93)	10.41 (2.91)	0.14	.869	.002
Authenticity	61.27 (9.42)	65.81 (9.14)	62.15 (9.28)	8.92	<.001	.102

Pattern: Similar to call scripts - GPT-4 highest authenticity, minimal other differences. Reflections showed higher positive emotion and insight language than call scripts across all LLMs (as expected given intervention content).

Overall Conclusion: LLM choice produces minimal systematic differences in linguistic features. The most consistent difference is GPT-4's higher authenticity scoring, which likely reflects model training differences but doesn't affect primary outcomes. Findings are robust across model architectures.

RQ4: Mechanism Exploration Through Qualitative Analysis

Can AI persona reflections illuminate psychological mechanisms underlying intervention effects?

Beneficiary Impact: Emergent Themes

Sample and Coding:

- **Sample:** All beneficiary condition personas (n=80)
- **Coding:** Two independent coders rated reflections on four 0-3 ordinal dimensions (see Appendix B)
- **Reliability:** Weighted κ with quadratic weights

Table 22. Inter-Rater Reliability for Reflection Coding.

Dimension	κ_w	95% CI	Interpretation
Affective Impact	.81	[.75, .87]	Excellent
Cognitive Reframing	.78	[.71, .85]	Substantial
Beneficiary Connection	.84	[.78, .90]	Excellent
Personal Relevance	.76	[.69, .83]	Substantial

Disagreements: 18 cases (7.5%) flagged for discrepancies > 1 category; all resolved through discussion.

 doi:10.20944/preprints202602.1127.v1

Theme 1: Concrete Beneficiary Visualization

Prevalence: 69 of 80 personas (86.3%) coded 2+ on Beneficiary Connection scale

Distribution by code:

- Code 0 (no connection): 2 personas (2.5%)
- Code 1 (generic): 9 personas (11.3%)
- Code 2 (specific): 38 personas (47.5%)
- Code 3 (vivid): 31 personas (38.8%)

Example (Code 3 - Claude, age 34, high agreeableness):

"Sarah's story made her real to me in a way that abstract 'scholarship recipients' never could be. I kept picturing that specific moment when she called her parents with the news—her mother's tears of relief rather than worry. That image has anchored itself in my mind. When I make calls tonight, I won't be thinking about quotas or scripts. I'll be thinking about whether my call could create that moment of relief and joy for another family."

Example (Code 2 - GPT-4, age 42, collectivist background):

"Hearing about Sarah's specific situation—her parents' financial struggles, the kitchen table conversation, the scholarship letter—made it concrete rather than abstract. I can now picture a real person benefiting from donations instead of just thinking about 'helping students' in general."

Mechanism: Transforming abstract beneficiaries into vivid, specific individuals activates empathic processes and increases perceived impact. The concreteness principle suggests that specific, detailed examples are more psychologically engaging than abstract categories.

Theme 2: Values Activation

Prevalence: 58 of 80 personas (72.5%) coded 2+ on Personal Relevance

Distribution by code:

- Code 0 (no connection): 8 personas (10.0%)
- Code 1 (vague): 14 personas (17.5%)
- Code 2 (specific): 31 personas (38.8%)
- Code 3 (deep): 27 personas (33.8%)

Example (Code 3 - GPT-4, age 29, collectivist background):

"This connects to something deep in my values about fairness and opportunity. I've seen talented people held back by circumstances beyond their control. Sarah's story reminded me that small acts—one donation, one call—can be the difference between potential realized and potential wasted. That feels significant in a way the job description didn't convey. It's not just work; it's living my values."

Example (Code 2 - Llama, age 38, working class background):

"I grew up without much money, so Sarah's story hits home. Education was my path out, and I remember how much small supports mattered—a waived fee here, a small scholarship there. This job aligns with my belief that everyone deserves a fair shot regardless of their starting point."

Mechanism: Beneficiary narratives activate pre-existing prosocial values, creating alignment between work tasks and personal moral frameworks. When work becomes value-expressive, intrinsic motivation increases.

Theme 3: Empathic Arousal

Prevalence: 54 of 80 personas (67.5%) coded 2+ on Affective Impact

Distribution by code:

- Code 0 (no emotion): 6 personas (7.5%)
- Code 1 (mild): 20 personas (25.0%)
- Code 2 (moderate): 32 personas (40.0%)
- Code 3 (strong): 22 personas (27.5%)

Example (Code 3 - Llama, age 41, high neuroticism):

 doi:10.20944/preprints202602.1127.v1

"Reading about Sarah's parents' financial stress hit close to home. I know what it's like to worry about money while trying to pursue something important. When the narrative described her mother crying from relief after the scholarship letter—I actually felt tears in my own eyes. That empathic connection made this feel personal rather than transactional."

Example (Code 2 - Claude, age 27, high agreeableness):

"The story moved me emotionally. Thinking about Sarah's family's struggle and their relief when she got the scholarship created genuine feelings of compassion and a desire to help create more moments like that. It's not just intellectually understanding impact—it's feeling invested in the outcome."

Mechanism: Emotional resonance with beneficiary experiences creates affective motivation distinct from cognitive recognition of impact. Empathic arousal may sustain motivation through challenging tasks better than purely cognitive appeals.

Proposed Multi-Pathway Model:

Analysis suggests beneficiary impact operates through at least three distinct mechanisms:

1. **Cognitive pathway:** Increased perceived impact through concrete evidence
   - "Now I understand specifically how donations help"
   - "The story clarified the causal chain from my work to student outcomes"
2. **Affective pathway:** Empathic arousal creating emotional investment
   - "I felt moved by Sarah's family's experience"
   - "The emotional connection makes me care about the outcome"
3. **Values pathway:** Alignment of work with prosocial identity and values
   - "This work expresses my core values about fairness"
   - "Helping students aligns with who I want to be"

Evidence for pathway independence:

Correlations between coded dimensions (Spearman's ϱ):

- Beneficiary Connection × Personal Relevance: ϱ = .47 (moderate)
- Beneficiary Connection × Affective Impact: ϱ = .52 (moderate)
- Personal Relevance × Affective Impact: ϱ = .39 (moderate)

All correlations significant ($p < .001$) but moderate in magnitude, suggesting partially independent processes.

Hypothesis for human testing: These pathways may operate independently or synergistically. Interventions could selectively activate specific pathways (e.g., statistical evidence for cognitive, narrative for affective, values reflection for identity pathway), or optimal interventions might activate all three simultaneously.

Job Crafting: Reframing Quality and Variety

Task: All job crafting personas (n=80) completed: "I'm not just making phone calls, I'm actually..."

Quantitative Analysis of Reframes:

Table 23. Job Crafting Reframe Statistics.

Metric	M	SD	Range	Median
Number of reframes generated	12.3	2.1	8-18	12
Quality rating (1-5 scale)*	3.6	0.9	1.8-5.0	3.7
High quality reframes (4-5)**	7.5	2.3	3-14	7

*Quality rated by two independent coders on clarity, abstraction, and meaningfulness (ICC = .84). **Number of individual reframes rated 4 or 5 out of 5.

Categorical Analysis:

 doi:10.20944/preprints202602.1127.v1

Two coders classified each reframe into categories (Cohen's κ = .87 for classification agreement):

Table 24. Job Crafting Reframe Categories.

Category	Frequency (% of personas)	Example
Impact reframe	75/80 (93.8%)	"...removing barriers to educational opportunity"
Relational reframe	70/80 (87.5%)	"...connecting generous people with deserving students"
Identity reframe	61/80 (76.3%)	"...serving as advocate for those who can't advocate for themselves"
Social change reframe	51/80 (63.8%)	"...participating in expanding educational access"
Purpose reframe	47/80 (58.8%)	"...contributing to generational transformation"
Process reframe	34/80 (42.5%)	"...building relationships and trust through authentic conversation"

Note: Most personas generated reframes in multiple categories; percentages indicate whether persona used that category at least once.

Example High-Quality Reframing Sequence (Claude, age 36, high openness, rated 4.8/5):

"I'm not just making phone calls. I'm actually...

- *Creating possibilities that didn't exist before my call*
- *Serving as the bridge between generosity and opportunity*
- *Participating in interrupting cycles of educational inequality*
- *Enabling someone's potential to override their circumstances*
- *Being part of life-changing moments I'll never witness*
- *Advocating for students who can't advocate for themselves*
- *Transforming financial transactions into human transformation*
- *Building futures through conversations in the present*
- *Contributing to stories that families will tell for generations*
- *Making education accessible to talent that deserves recognition*
- *Honoring the principle that potential should be nurtured, not wasted*
- *Participating in the collective work of expanding opportunity"*

Analysis: This sequence demonstrates:

- High variety (spans 5 different categories)
- Progressive abstraction (from concrete "creating possibilities" to abstract "collective work")
- Temporal range (present actions to generational impact)
- Multiple motivational appeals (justice, impact, meaning, collective good)

Example Moderate-Quality Reframing (GPT-4, age 28, moderate openness, rated 3.2/5):

"I'm not just making phone calls. I'm actually...

- *Helping students afford college*
- *Supporting education and opportunity*
- *Making a difference in people's lives*
- *Contributing to the scholarship fund*
- *Connecting donors with students who need help*
- *Doing meaningful work that helps others*
- *Being part of something bigger than myself*
- *Supporting the university's mission*
- *Helping families achieve their educational goals"*

 doi:10.20944/preprints202602.1127.v1

Analysis: This sequence shows:

- Moderate variety (primarily impact and purpose frames)
- Lower abstraction (concrete "helping" language)
- Some repetition (several variations on "helping/supporting")
- Less sophisticated psychological depth

Proposed Mechanisms for Job Crafting Effectiveness:

Analysis suggests job crafting works through:

1. **Frame Variety:** More diverse frames provide richer mental toolkit
   - Correlation: Number of categories used × Motivation rating: $r = .47$, $p < .001$
   - Personas using 4+ categories: M motivation = 9.1 vs. 1-3 categories: M = 7.9, $t(78) = 4.23$, $p < .001$
2. **Frame Abstraction:** Higher-level frames connect to deeper values
   - High abstraction reframes (rated 4+): M motivation = 9.3
   - Low abstraction reframes (rated <3): M motivation = 7.4
   - Difference: $t(78) = 5.67$, $p < .001$
3. **Frame Personalization:** Alignment with individual values increases adoption
   - Personas incorporating personal experiences into reframes: M motivation = 9.2
   - Personas with generic reframes: M motivation = 8.1
   - Difference: $t(78) = 3.84$, $p < .001$
4. **Cognitive Flexibility:** Capacity to hold multiple simultaneous frames
   - Personas generating 12+ reframes: M motivation = 9.0
   - Personas generating <10 reframes: M motivation = 7.8
   - Difference: $t(78) = 4.56$, $p < .001$

Hypothesis for Human Testing:

1. **Quantity matters:** More reframes correlate with higher motivation, possibly because:
   - More options increases probability of finding resonant frame
   - Generation process itself is motivating (cognitive elaboration)
   - Demonstrates thorough engagement with exercise
2. **Quality matters more:** High-quality reframes (abstractabstract, varied, personalized) show stronger associations than sheer quantity
3. **Optimal approach unclear:** Should interventions:
   - Ask people to self-generate reframes? (High engagement, but quality varies)
   - Provide example reframes? (Consistent quality, but lower engagement)
   - Hybrid: Provide examples + generate own? (Combines benefits but adds time)

Critical limitation: These mechanisms emerged from AI-generated text. Human reframing quality may differ substantially, particularly:

- Humans may struggle to generate high-abstraction reframes
- Humans may experience cognitive fatigue during extended reflection
- Humans may be less articulate in expressing psychological processes

Call Script Analysis

Sample: All 240 call scripts rated by two independent coders on five dimensions (see Appendix B for complete coding scheme)

Inter-Rater Reliability:

Table 25. Call Script Coding Reliability.

Dimension	ICC(2,2) / κ_w	95% CI	Interpretation
Enthusiasm (1-10)	ICC = .87	[.84, .90]	Excellent

 doi:10.20944/preprints202602.1127.v1

Beneficiary Mention (0-3)	κ_w = .82	[.77, .87]	Excellent
Personalization (1-10)	ICC = .84	[.80, .88]	Excellent
Clarity of Ask (1-10)	ICC = .89	[.86, .92]	Excellent
Persuasiveness (1-10)	ICC = .87	[.84, .90]	Excellent

Disagreements: 23 scripts (9.6%) flagged for discrepancies > 2 points; all resolved through discussion to consensus.

Primary Analysis: Persuasiveness by Condition

Table 26. Call Script Persuasiveness Ratings.

Condition	M	SD	95% CI	Effect vs. Control
Control	5.83	1.58	[5.47, 6.19]	—
Beneficiary	8.07	1.42	[7.75, 8.39]	d = 1.50 [1.01, 1.99]
Job Crafting	8.71	1.29	[8.42, 9.00]	d = 1.96 [1.44, 2.48]

ANOVA: $F(2, 237) = 92.47$, $p < .001$, $\eta^2_p = .438$

Post-hoc comparisons (Tukey HSD):

- Beneficiary > Control: $p < .001$
- Job Crafting > Control: $p < .001$
- Job Crafting > Beneficiary: $p = .003$

Interpretation: Call scripts reflected intervention content, with both intervention conditions producing substantially more persuasive openings than control. The large effects (d = 1.50-1.96) suggest interventions influenced not just self-reported motivation but also behavioral output quality.

Qualitative Differences in Script Content:

Control Scripts (Typical Example):

"Hello Dr. Patterson, my name is [name] and I'm calling from the University Development Office. We're reaching out to alumni to request support for our annual scholarship fund. Your past gift of $50 was greatly appreciated. Would you be willing to renew your support this year?"

Characteristics:

- Formulaic structure
- Focus on logistics and transaction
- Minimal personalization
- No impact mention
- Perfunctory tone

Beneficiary Scripts (Typical Example):

"Hello Dr. Patterson, my name is [name] and I'm calling from the University. I recently heard the story of Sarah Martinez, a student whose family struggled financially but who's now thriving here because of alumni support like yours. Your past donation of $50 was part of making stories like Sarah's possible. The difference these scholarships make is truly profound—families transform from worry to relief when students receive funding. Would you consider contributing again this year to help more students like Sarah?"

Characteristics:

- Incorporates specific beneficiary narrative
- Connects donor's past gift to impact
- Emotional language ("transform," "profound")
- Personal rather than transactional tone
- Explicit impact framing

 doi:10.20944/preprints202602.1127.v1

Beneficiary Mention Coding:

Table 27. Beneficiary Mention by Condition.

Condition	Code 0 (none)	Code 1 (generic)	Code 2 (specific)	Code 3 (detailed)	M Code
Control	67 (83.8%)	13 (16.3%)	0 (0%)	0 (0%)	0.16
Beneficiary	0 (0%)	8 (10.0%)	38 (47.5%)	34 (42.5%)	2.33
Job Crafting	2 (2.5%)	11 (13.8%)	41 (51.3%)	26 (32.5%)	2.14

ANOVA: $F(2, 237) = 447.23$, $p < .001$, $\eta^2\_p = .791$

Both intervention conditions dramatically increased beneficiary mentions compared to control, with beneficiary impact condition showing slightly (but not significantly) more detailed mentions than job crafting ($p = .18$).

Job Crafting Scripts (Typical Example):

"Dr. Patterson, hello! I'm [name], calling from the University. I realize you're busy, so I'll be brief and direct. I've been thinking a lot about how alumni gifts literally change lives—not in an abstract way, but in very concrete 'phone call home with good news instead of bad news' ways. Your past contribution of $50 matters more than you might realize. Combined with other gifts, donations like yours create opportunities for students whose talent deserves recognition but whose circumstances create barriers. Could I ask you to consider making that gift again this year? It would genuinely make a difference."

Characteristics:

- More personal voice ("I've been thinking")
- Acknowledges donor's time
- Reframes donation significance
- Creative language and metaphors
- Authentic tone
- Direct but respectful ask

Personalization Analysis:

Table 28. Personalization Quality by Condition.

Condition	M	SD	Effect vs. Control
Control	4.21	1.47	—
Beneficiary	6.84	1.38	d = 1.85
Job Crafting	7.53	1.29	d = 2.39

Job crafting scripts showed higher personalization than beneficiary scripts ($t(158) = 3.24$, $p = .001$), suggesting the reflection exercise enhanced ability to craft individualized appeals.

Summary of Call Script Findings:

1. **Intervention effects on output quality:** Both interventions improved call script persuasiveness substantially (large effects)
2. **Content differences:**
   - Beneficiary condition: Integrated specific beneficiary narrative
   - Job Crafting condition: More varied, creative, personalized approaches
3. **Mechanism reflection in behavior:** Scripts reflected internal cognitive processes:
   - Beneficiary personas incorporated Sarah's story
   - Job crafting personas showed evidence of reframing ("not just asking for money, but...")
4. **Implications:** Motivational interventions influenced not just self-reported attitudes but also behavioral outputs (script quality). This suggests interventions may translate to actual performance, though this remains speculative without observing real fundraising outcomes.

 doi:10.20944/preprints202602.1127.v1

RQ5: Methodological Insights

What can we learn about AI simulation's potential and limitations?

Strengths Demonstrated

1. Rapid Hypothesis Generation

AI simulation enabled testing of:

- Multiple intervention variations (could easily test 10+ versions in days)
- Diverse persona characteristics (240 systematic combinations)
- Complex moderation patterns (personality × culture × intervention)

Traditional research timeline: 6-12 months for single intervention study
AI simulation timeline: 1 week for complete study

2. Systematic Diversity

Perfect representation of specified characteristics enables clean moderation tests impossible in convenience samples. Every personality × culture combination represented.

3. Detailed Process Data

Qualitative responses provide rich mechanistic insights. AI personas articulate psychological processes explicitly in ways humans might not spontaneously report.

4. Iterative Refinement

Rapid testing enabled quick identification of:

- Confusing intervention language
- Optimal reflection prompt ordering
- Necessary script elements

Refined versions could be tested immediately.

5. Cross-Model Validation

Testing across LLM architectures provides internal validity check. Convergent findings increase confidence in robustness.

Limitations and Concerns

1. Unknown External Validity

We cannot determine whether AI-generated effect sizes, moderation patterns, or mechanisms reflect actual human psychology without direct validation studies.

Critical question: Do the very large effect sizes (d = 2.92-3.65) reflect real intervention potential, LLM response biases, or something in between?

2. Potential for Idealized Responses

AI personas showed:

- Highly articulate reflections (possibly beyond typical human capacity)
- No spontaneous skepticism or cynicism
- Perfect emotional regulation across scenarios
- Consistent trait-behavior correspondence

Concern: Responses may reflect "ideal typical" patterns from training data rather than authentic psychological variability.

3. Cultural Stereotyping Risk

Collectivist personas showed strong stereotypical patterns (community emphasis, duty language). This could reflect:

- Genuine cultural psychological processes
- Training data stereotypes
- Researcher prompt biases

Recommendation: Cultural findings require careful validation with actual culturally diverse samples.

4. Linguistic Polish

AI-generated text showed higher quality, coherence, and elaboration than typical survey responses. This could affect:

 doi:10.20944/preprints202602.1127.v1

- Qualitative coding (easier to code clear articulate responses)
- Mechanism detection (explicit articulation of implicit processes)
- Generalizability (most humans less articulate)

5. Missing Authentic Constraints

Real workers experience:

- Competing demands (family, other jobs, stress)
- Cognitive fatigue during long interventions
- Social desirability concerns
- Performance anxiety
- Organizational politics

AI personas experience none of these.

6. Unclear Mapping to Behavior

We measured self-reported motivation and anticipated performance, not actual behavior. AI ratings may not predict actual fundraising success.

Appropriate Uses of AI Simulation

Based on observed strengths and limitations:

APPROPRIATE:

1. **Early-stage intervention prototyping:** Rapid testing of variations before human trials
2. **Mechanism hypothesis generation:** Identifying potential psychological processes to test
3. **Boundary condition exploration:** Simulating when interventions might fail
4. **Language refinement:** Identifying confusing elements through AI feedback
5. **Moderation hypothesis formation:** Suggesting individual differences worth testing

INAPPROPRIATE:

1. Effect size estimation for implementation: Cannot trust magnitudes
2. Cultural validation: High risk of stereotype perpetuation
3. Behavior prediction: Unknown relationship to actual performance
4. Publication as standalone findings: Requires human validation
5. Policy decisions: Real-world stakes demand real-world data

Recommendations for Researchers

If using AI simulation in intervention research:

DO:

1. **Frame as exploratory and hypothesis-generating**
   - State explicitly: "This is exploratory research generating hypotheses for human validation"
   - Avoid language suggesting definitive conclusions
   - Position findings as "provocative patterns worth investigating"
2. **Plan human validation from the outset**
   - Design AI study with human replication in mind
   - Match measures, procedures, and designs
   - Budget for human validation in grant proposals
   - Present AI findings as first phase of multi-phase research
3. **Test across multiple LLM architectures**
   - Use at least 2-3 different models
   - Report convergence and divergence
   - Increases confidence if findings replicate across architectures
   - Reveals model-specific artifacts if results diverge
4. **Examine qualitative responses carefully**
   - Look for signs of idealization (too polished, too coherent)
   - Check for stereotyping (especially cultural, demographic)
   - Assess linguistic naturalness vs. artificial patterns

 doi:10.20944/preprints202602.1127.v1

- Compare to actual human open-ended responses when available

5. **Compare to prior human research**
   - Do effect sizes align with meta-analyses? (expect AI larger, but how much?)
   - Do moderation patterns match established findings?
   - Do mechanisms align with theoretical predictions?
   - Divergences suggest either AI artifacts or novel insights requiring validation
6. **Be transparent about limitations**
   - Explicitly acknowledge unknown external validity
   - Discuss potential for stereotyping and bias
   - Note missing elements (fatigue, real stakes, competing demands)
   - Caveat all interpretations appropriately
7. **Report methods in detail**
   - Complete persona generation prompts
   - Exact intervention scripts
   - Full coding schemes and reliability
   - LLM versions and parameters
   - Random seeds for reproducibility
8. **Consider hybrid designs**
   - AI for breadth (many variations) + Humans for depth (focal conditions)
   - Sequential: AI exploration → Human validation → AI refinement → Human confirmation
   - Complementary: AI for hypothesis generation + Humans for hypothesis testing

DON'T:

1. **Publish AI findings as standalone conclusions**
   - Insufficient without human validation
   - Risk perpetuating findings that don't replicate
   - Misleads field about intervention effectiveness
2. **Trust absolute effect size estimates**
   - Use only for relative comparisons within AI study
   - Apply large corrections (÷2-3) if planning human samples
   - Never claim specific expected effects for implementation
3. **Make implementation recommendations without human data**
   - Organizations should not adopt interventions based solely on AI evidence
   - Practitioners need real behavioral outcomes
   - ROI claims require human cost-benefit data
4. **Assume moderation patterns will replicate**
   - Personality interactions may reflect training data patterns
   - Cultural moderation especially risky (stereotype perpetuation)
   - Individual difference findings require human validation
5. **Claim cultural findings without cultural validation**
   - High risk of stereotyping
   - Requires actual participants from relevant cultures
   - Cultural experts should review findings before publication
6. **Replace human research entirely**
   - AI is tool for early-stage exploration, not replacement
   - Some questions only humans can answer (behavior, real stakes, authentic experience)
   - Field needs balance of AI efficiency and human validity
7. **Overclaim or overgeneralize**

 doi:10.20944/preprints202602.1127.v1

- Stick to specific tested interventions and contexts
- Avoid extrapolating to untested populations or settings
- Resist pressure to make stronger claims than data support

8. **Ignore negative results**
   - If AI simulations show null effects, report them
   - Divergence between LLMs is informative
   - Unexpected patterns may reveal model limitations worth documenting

Appropriate Uses: Decision Framework

Question 1: What research stage?

☑ **Early exploration** - AI appropriate
⚠ **Mid-stage refinement** - AI useful with caution
❌ **Late-stage validation** - Human data required
❌ **Implementation decision** - Human data essential

Question 2: What question?

☑ **Which variations to test?** - AI useful for ranking
☑ **What mechanisms to examine?** - AI generates hypotheses
⚠ **How large is the effect?** - AI provides upper bound only
❌ **Will this work in practice?** - Requires human behavioral data
❌ **What's the ROI?** - Requires real implementation costs and outcomes

Question 3: What population?

☑ **General patterns** - AI okay for hypothesis generation
⚠ **Personality moderation** - AI suggests, humans must validate
❌ **Cultural differences** - High risk; extensive human validation critical
❌ **Specific demographic groups** - Human data essential

Question 4: What outcome?

☑ **Attitudes and perceptions** - AI can model
⚠ **Behavioral intentions** - AI models but unknown validity
❌ **Actual behavior** - Must observe in humans
❌ **Long-term outcomes** - Must track in humans over time

Question 5: What risk?

☑ **Low stakes exploration** - AI appropriate
⚠ **Medium stakes (grant planning)** - AI with heavy caveats
❌ **High stakes (policy decisions)** - Human data required
❌ **Vulnerable populations** - Must protect through human research ethics

Discussion

This study investigated whether AI-simulated worker personas can serve as a useful tool for exploring motivational interventions. Our findings reveal both intriguing possibilities and important limitations.

Summary of Key Findings

Large Intervention Effects: Both beneficiary impact (d = 2.92) and job crafting (d = 3.65) produced very large effects on AI persona motivation, substantially exceeding typical field study estimates.

Theoretically Coherent Moderation: Personality traits (agreeableness, conscientiousness, openness) and cultural background (collectivism) moderated effects in theoretically predictable directions.

Cross-Model Robustness: Three different LLM architectures produced highly consistent results, with LLM choice explaining only 3-4% of variance.

 doi:10.20944/preprints202602.1127.v1

Rich Qualitative Data: AI personas generated detailed reflections revealing potential mechanisms (visualization, values activation, empathic arousal for beneficiary impact; reframing variety and abstraction for job crafting).

Systematic Diversity: Perfect representation of specified demographic and personality characteristics enabled clean moderation tests.

Interpretation: What Do These Findings Mean?

The central interpretive challenge: We do not know whether AI simulation findings reflect human psychology or artifacts of LLM training and architecture.

Optimistic interpretation: Large effect sizes might reflect genuine intervention potential, freed from the noise, distraction, and competing demands that reduce effects in messy field settings. The very large effects (d = 3.15-3.97) may approximate what beneficiary contact and job crafting could achieve under ideal conditions—full attention, no distractions, receptive mindset. Moderation patterns might reveal authentic psychological processes. Qualitative mechanisms might illuminate real pathways.

Pessimistic interpretation: Large effects might reflect LLM training to generate agreeable, positive responses to motivational appeals. The models may be optimized to produce "ideal" responses rather than realistic human reactions. Moderation might reflect cultural stereotypes and theoretical expectations present in training data rather than genuine individual differences. Qualitative articulation might exceed realistic human capacity—real workers may not have the cognitive resources or verbal fluency to generate such sophisticated reflections during actual work scenarios.

Most likely reality: Some mixture. Evidence for partial validity:

Patterns suggesting authenticity:

1. **Consistent moderation directions:** Personality and cultural moderators aligned with theoretical predictions
2. **Cross-model convergence:** Three different LLM architectures produced highly similar results, reducing likelihood of model-specific artifacts
3. **Mechanism plausibility:** Qualitative themes (visualization, values activation, empathic arousal) align with established psychological theory
4. **Behavioral translation:** Scripts reflected intervention content, suggesting motivational changes influenced outputs

Patterns suggesting caution:

1. **Effect size magnitude:** Effects 5-10× larger than typical field research may reflect idealized responding
2. **Perfect trait-behavior correspondence:** Agreeableness always predicted prosocial responses; real psychology messier
3. **Absence of skepticism:** No personas showed cynicism, resistance, or competing motivations
4. **Linguistic polish:** All responses highly articulate; real workers more variable
5. **Cultural stereotypes:** Collectivist personas showed very stereotypical community language

The critical question is which is which—and we cannot answer definitively without human validation.

Our recommendation: Treat AI findings as **upper-bound estimates** of what interventions might achieve under optimal conditions, recognizing that:

- Real-world effects will likely be smaller (noise, competing demands, skepticism)
- Real-world moderation may be weaker (individual differences less consistent)
- Real-world mechanisms may be less articulate (implicit rather than explicit)
- But directional patterns may still be informative for hypothesis generation

What AI Simulation Might Be Good For

Despite uncertainty about external validity, AI simulation offers value for specific research stages:

 doi:10.20944/preprints202602.1127.v1

1. Rapid Intervention Prototyping

Example application: A researcher develops a beneficiary impact intervention. Key decisions:

- **Narrative type:** Single detailed story vs. multiple brief stories vs. aggregate statistics
- **Emotional tone:** High emotion vs. factual vs. balanced
- **Beneficiary type:** Student vs. community member vs. family
- **Duration:** 2 minutes vs. 5 minutes vs. 10 minutes
- **Delivery:** Written vs. video vs. in-person
- **Timing:** Before work vs. during training vs. periodic reminders

Traditional approach:

- Test each variation with human participants: 8 variations × 50 participants each = 400 participants
- Timeline: 6-12 months
- Cost: $20,000-50,000

AI simulation approach:

- Test all variations (including combinations): 2^6 = 64 possible versions
- Timeline: 1 week
- Cost: <$100

Value: Even if absolute effect sizes unreliable, **relative rankings** might indicate which versions merit human testing. AI could identify:

- "Single detailed story substantially outperforms multiple brief stories"
- "5-minute optimal; 2-minute too brief, 10-minute diminishing returns"
- "Emotional tone increases engagement but may backfire for skeptical personas"

This allows researchers to narrow from 64 possible versions to 3-5 promising candidates for human pilots.

2. Mechanism Hypothesis Generation

Qualitative AI responses revealed:

- **Beneficiary impact pathways:** Cognitive (perceived impact), Affective (empathy), Values (identity alignment)
- **Job crafting dimensions:** Variety, abstraction, personalization, flexibility
- **Cultural processes:** Community framing (collectivist) vs. personal values (individualist)

These hypotheses wouldn't emerge from small human samples due to:

- Participants rarely articulate mechanisms explicitly
- Small samples lack diversity to observe moderation
- Open-ended responses often superficial

AI advantages for mechanism exploration:

- Explicitly articulates psychological processes
- Perfect representation of diverse combinations
- Generates rich qualitative data at scale

Application: Use AI-generated mechanisms to design human studies testing:

Hypothesis from AI: Beneficiary impact works through three pathways—cognitive, affective, values

Human study design:

- Experimental manipulation of pathways (e.g., statistics-only for cognitive, emotional narrative for affective, values reflection for identity)
- Measure pathway activation and outcomes separately
- Test whether pathways operate independently or synergistically

3. Boundary Condition Exploration

AI simulation enables testing edge cases expensive or difficult to study in humans:

Questions AI could explore:

 doi:10.20944/preprints202602.1127.v1

- **Repetition effects:** Do beneficiary narratives lose effectiveness after 5 exposures? 10? 20?
  - AI: Test personas receiving intervention 1×, 5×, 10×, 20× times
  - Prediction for humans: Likely habituation curve
- **Duration optimization:** At what reflection length does fatigue overcome benefit?
  - AI: Test job crafting at 5, 10, 15, 20, 30, 45 minutes
  - Prediction for humans: Optimal duration for cost-benefit ratio
- **Personality extremes:** Do interventions backfire for highly cynical individuals?
  - AI: Generate personas at extreme low agreeableness + high neuroticism
  - Prediction for humans: Possible boomerang effects
- **Context dependency:** Do effects differ for temporary vs. career workers?
  - AI: Manipulate employment type and career intentions
  - Prediction for humans: Career workers may show stronger long-term effects
- **Cultural specificity:** Which intervention elements are culture-general vs. culture-specific?
  - AI: Systematically vary intervention content across 20 cultural backgrounds
  - Prediction for humans: Some elements universal (concrete impact), others cultural (collective duty framing)

Value: Exploring boundaries in humans is expensive; AI makes it feasible to map an entire response surface, then selectively validate key predictions.

4. Intervention Language Refinement

AI personas can provide rapid feedback on:

Confusing instructions:

- Do personas misinterpret any prompts?
- Do they request clarification?
- Do responses suggest misunderstanding?

Ambiguous wording:

- Test multiple phrasings of key instructions
- Identify clearest version before human testing

Culturally insensitive language:

- Generate personas from diverse backgrounds
- Check for responses indicating offense or alienation
- Refine language to be inclusive

Optimal sequencing:

- Test different orders of reflection prompts
- Identify sequences that build most effectively

Example: Reflection prompt reads: "Think about people who benefit from this work."

- AI response analysis: 40% interpret as "direct beneficiaries," 60% as "broader society"
- Refinement: "Think about the specific students who will receive scholarships from donations you help secure."
- Re-test: 95% interpret correctly

While not replacing human pilot testing, AI iteration can substantially improve materials before human exposure.

5. Sample Size Planning for Human Studies

Problem: Determining required sample size for human validation studies requires effect size estimates

Traditional approach: Use published meta-analyses (often limited) or pilot study (expensive)

AI simulation approach:

1. Run AI simulation (N=240) to estimate effect size (acknowledging likely inflation)
2. Apply conservative correction (e.g., divide by 2-3)
3. Power analysis using corrected estimate

 doi:10.20944/preprints202602.1127.v1

4. Plan human sample size

Example:

- AI simulation: d = 3.15 for beneficiary impact
- Conservative estimate: d = 1.05 (divide by 3)
- Power analysis: N = 30 per condition for 80% power at α = .05
- Plan human study: N = 40 per condition (with buffer)

Value: Provides starting point for planning, even if uncertain

What AI Simulation Cannot Do

Equally important—recognizing limitations:

1. Predict Real Effect Sizes

Our findings: d = 3.15-3.97 for motivational interventions

Meta-analyses of human studies: d = 0.28-0.60

Discrepancy: 5-10× larger effects in AI simulation

Possible explanations:

AI overestimation:

- Response bias toward positive, agreeable answers
- Absence of competing motivations and skepticism
- Ideal processing conditions (full attention, no distractions)
- Linguistic patterns optimized for coherence and positivity

Human underestimation:

- Field implementation failures (poor fidelity, distractions)
- Measurement error (noisy self-reports, behavioral variability)
- Competing workplace demands reduce intervention impact
- Time pressure limits reflection quality

Truth probably between: Real effects under ideal conditions may exceed field estimates but fall short of AI simulations.

Implication: Cannot use AI effect sizes for:

- Power analysis without substantial correction
- Cost-benefit calculations for implementation
- Expected ROI predictions
- Grant justifications for human research

Appropriate use: Relative effect size comparisons (e.g., "Intervention A shows 30% larger effects than Intervention B") may be more reliable than absolute magnitudes.

2. Validate Cultural Patterns

Our findings: Collectivist personas showed stronger intervention responses (0.32-0.39d larger effects)

An additional concern: Our 2:1 oversampling of collectivist cultures (designed to increase statistical power for detecting cultural moderation) may have inadvertently amplified these stereotyping risks. With 160 collectivist personas versus 80 individualist personas, collectivist stereotypes had more opportunities to emerge and stabilize in our data. This sampling decision, while justified for statistical reasons, means our cultural findings require particularly careful scrutiny and validation.

Qualitative patterns:

- 83% of collectivist personas mentioned community themes
- Highly stereotypical language ("honoring tradition," "collective responsibility")
- Clean separation between collectivist and individualist responses

Red flags suggesting stereotypes rather than authentic psychology:

1. **Too clean:** Real cultural differences show substantial within-group variation and overlap
2. **Stereotypical language:** Phrases align with common cultural stereotypes in training data

 doi:10.20944/preprints202602.1127.v1

3. **No counterexamples:** No collectivist personas showed individualistic patterns or vice versa
4. **Prompt sensitivity:** Cultural background mentioned in persona generation may have triggered stereotypical responding

High risk of perpetuating cultural stereotypes
Critical needs for human validation:

- Recruit actual participants from diverse cultural backgrounds
- Use validated cultural value scales (not just nationality)
- Examine within-culture variation (not just between-culture means)
- Test with materials not mentioning culture explicitly
- Include culture-blind coders for qualitative analysis

Appropriate use of cultural AI findings:

- Generate hypotheses about possible cultural moderation
- Identify cultural elements to include in human studies
- Suggest culturally-adapted intervention versions to test

Inappropriate use:

- Claim interventions work better in collectivist cultures
- Design culture-specific interventions without human data
- Make cross-cultural implementation recommendations

3. Predict Actual Behavior
Our measures:

- Self-reported motivation (rating scale)
- Anticipated performance (self-estimate)
- Confidence (self-assessment)
- Call script quality (one-time written output)

Missing:

- Actual fundraising success rates
- Behavioral persistence over time (hours, days, weeks)
- Performance under realistic constraints (rejection, fatigue, competing demands)
- Adaptation to unexpected situations
- Long-term retention and sustained motivation

Reasons AI cannot predict behavior:

1. **Self-report ≠ behavior:** Classic attitude-behavior gap; people overestimate follow-through
2. **Ideal conditions:** AI personas experience no:
    - Fatigue (can't get tired)
    - Distraction (perfect focus)
    - Competing demands (no family, financial pressures during task)
    - Social pressure (no performance anxiety)
    - Resource depletion (unlimited cognitive capacity)
3. **Single-shot measurement:** One call script ≠ sustained performance over shifts
4. **No stakes:** AI personas can't experience consequences of success/failure

Example of AI-human divergence:
AI prediction: High motivation → high performance (nearly perfect correlation in simulations)
Human reality: High motivation → moderate performance (correlation often .30-.40 due to skill differences, situational constraints, motivation fluctuation)

Implications:

- Cannot use AI to predict implementation success
- Cannot estimate cost-benefit ratios
- Cannot claim behavioral impact without human validation

 doi:10.20944/preprints202602.1127.v1

- Behavioral outcomes require actual human observation

4. Replace Human Validation

The fundamental limitation: We cannot determine AI simulation validity without human comparison. AI cannot validate itself.

Essential human research:

1. **Direct replication:** Conduct identical study with human participants
   - Same interventions, measures, design
   - Compare effect sizes, moderation patterns, qualitative themes
   - Identify what replicates vs. diverges
2. **Behavioral validation:** Measure actual performance outcomes
   - Fundraising success rates
   - Persistence through difficulty
   - Long-term retention
3. **Mechanism testing:** Experimentally manipulate proposed pathways
   - Isolate cognitive, affective, values pathways
   - Test causal predictions from AI qualitative analysis
4. **Boundary condition testing:** Validate edge cases
   - Repetition effects
   - Duration optimization
   - Personality interactions

Publication standards:

- AI simulation alone: Insufficient for substantive conclusions
- AI + Human validation: Publishable with both components
- AI findings: Should be positioned as "hypothesis-generating" pending validation

Practice standards:

- AI simulation: Informative for design decisions
- Implementation decisions: Require human data demonstrating real behavioral impact

Theoretical Contributions

Beyond methodology, this research offers tentative insights about beneficiary impact and job crafting:

Beneficiary Impact: Multi-Pathway Model

AI responses suggest beneficiary narratives might work through three distinct mechanisms:

1. **Cognitive pathway:** Concrete evidence of impact increases perceived meaningfulness
2. **Affective pathway:** Empathic arousal creates emotional investment
3. **Values pathway:** Alignment with prosocial identity and moral frameworks

Testable hypothesis: These pathways are separable. Interventions could selectively activate specific pathways, or combinations might produce synergistic effects.

Human validation needed: Experimental manipulation of proposed mechanisms.

Job Crafting: Reframing Processes

AI personas generated numerous creative reframes, suggesting effectiveness might depend on:

1. **Frame variety:** More diverse frames provide richer mental toolkit
2. **Frame abstraction:** Higher-level frames connect to deeper values
3. **Frame personalization:** Alignment with individual values increases adoption
4. **Cognitive flexibility:** Capacity to hold multiple simultaneous perspectives

Testable hypothesis: Providing example frames versus self-generating frames activates different processes. Quality might matter more than quantity.

Human validation needed: Compare interventions varying in these dimensions.

 doi:10.20944/preprints202602.1127.v1

Table D1. Summary of Research Questions and Findings.

Research Question	Hypothesis/Expectation	Finding	Support Status	Human Validation Priority
RQ1: Effect Patterns	Both interventions increase motivation vs. control	Both interventions showed very large effects (d = 3.15-3.97)	✓ Supported (but magnitude uncertain)	CRITICAL - Need to determine realistic effect sizes
RQ2a: Personality Moderation	Agreeableness, conscientiousness, openness moderate effects	All three moderated job crafting; agreeableness moderated both interventions	✓ Partially supported	HIGH - Pattern plausible but needs confirmation
RQ2b: Cultural Moderation	Collectivist orientation amplifies prosocial interventions	Collectivist personas showed 0.32-0.39d stronger effects	? Uncertain (stereotype risk)	CRITICAL - High stereotype risk requires careful validation
RQ3: Cross-Model Robustness	Different LLMs should produce similar patterns	Three LLMs showed highly consistent results (3-4% variance from LLM)	✓ Strongly supported	LOW - Establishes within-paradigm reliability
RQ4: Mechanism Exploration	Qualitative analysis reveals underlying processes	Identified multiple pathways (cognitive, affective, values for beneficiary; variety, abstraction for job crafting)	✓ Hypotheses generated	HIGH - Mechanisms plausible and testable but need experimental validation
RQ5: Methodological Utility	AI simulation useful for certain research stages	Strong utility for prototyping, hypothesis generation; limited for effect estimation, behavior prediction	✓ Supported with clear boundaries	ONGOING - Field needs continued investigation

Note: Support status reflects confidence in patterns within AI simulation data; all findings require human validation before substantive conclusions about human psychology.

Practical Implications

For Researchers

Using AI simulation in intervention research:

DO:

- Use for rapid prototyping and exploration

 doi:10.20944/preprints202602.1127.v1

- Generate hypotheses about mechanisms and moderators
- Test multiple variations to identify promising candidates
- Employ as preliminary step before human research
- Report findings transparently with clear limitations
- Test across multiple LLM architectures

DON'T:

- Publish AI findings as standalone conclusions
- Trust absolute effect size estimates
- Make implementation recommendations without human data
- Assume moderation patterns will replicate
- Claim cultural findings without cultural validation
- Replace human research entirely

For Practitioners

Considering motivational interventions:

Based on AI simulation alone:

- Cannot determine whether interventions will work
- Cannot estimate expected effect sizes
- Cannot predict ROI
- Cannot customize for specific populations

Appropriate use of AI research:

- Identifies interventions worth pilot testing
- Suggests elements to include in design
- Raises hypotheses about who might benefit most
- Indicates potential implementation challenges

Before implementation:

- Pilot test with actual employees (n=30-50 minimum)
- Measure actual behavior, not just self-reported motivation
- Assess costs (time, resources, disruption)
- Evaluate sustainability and fidelity

Our specific findings suggest:

- Beneficiary contact interventions are brief (3-4 min) and potentially high-impact
- Job crafting reflections require substantial time (20+ min) that may be impractical
- Consider hybrid: brief beneficiary contact for all + optional deeper reflection for interested employees

Limitations

We identify eleven limitations organized into four categories: (1) Critical limitations that threaten core validity and require human data to resolve (Limitations 1, 3, 6); (2) Generalizability limitations that restrict the scope of our findings (Limitations 2, 4, 9); (3) Methodological limitations affecting interpretation and replicability (Limitations 7, 8, 10, 11); and (4) Temporal/technical limitations specific to current LLM technology (Limitation 5). We present these in order of their impact on the interpretability and utility of our findings.

1. No Human Validation (Critical)

The fundamental limitation: Without human comparison data, we cannot determine whether AI simulation findings reflect human psychology or artifacts of LLM training and architecture. All findings must be considered provisional hypotheses requiring validation.

Specific uncertainties:

- **Effect size magnitude:** Are our effects (d = 3.15-3.97) realistic under ideal conditions, or do they reflect AI response biases?

 doi:10.20944/preprints202602.1127.v1

- **Moderation patterns:** Do personality and cultural interactions reflect genuine psychology or training data patterns?
- **Qualitative mechanisms:** Do AI-articulated processes match how humans actually experience interventions?
- **Behavioral translation:** Would high-quality call scripts translate to actual fundraising success?

Planned next steps: We are currently conducting parallel human validation studies using identical procedures with university student samples and actual fundraising workers to directly compare:

- Effect sizes (expecting human effects 50-70% smaller)
- Moderation patterns (testing each significant AI interaction)
- Qualitative themes (comparing human and AI reflection coding)
- Behavioral outcomes (actual fundraising performance over multiple shifts)

Until human validation completes, all findings should be treated as hypotheses, not conclusions.

2. Single Context and Intervention Type

Tested: Fundraising for student scholarships using beneficiary contact and job crafting interventions

Not tested:

- **Other prosocial work domains:** Healthcare, education, social services, environmental conservation
- **Other motivational interventions:** Goal-setting, autonomy support, feedback, recognition, gamification
- **Other outcome measures:** Job satisfaction, retention, burnout, well-being, creativity
- **Other time scales:** Immediate vs. sustained effects over days, weeks, months

Implications:

- Findings may be specific to:
  - Donation requests (vs. direct service work)
  - Student beneficiaries (vs. other populations)
  - Short timeframe (vs. longitudinal motivation)
- Generalization requires testing across diverse contexts

Future directions:

- Test AI simulation in 5-10 different work domains
- Compare effectiveness across intervention types
- Examine long-term motivation trajectories
- Validate findings in non-Western organizational contexts

3. Self-Report Outcomes Only

Our measures:

- Self-reported motivation (single rating)
- Self-estimated performance (anticipated calls)
- Self-assessed confidence
- Single call script (one behavioral sample)

Missing:

- Actual fundraising success rates (% donation commitments)
- Sustained performance over time (across multiple shifts)
- Behavioral persistence through difficulty (handling 20+ rejections)
- Performance under realistic constraints (noise, distractions, fatigue)
- Skill execution under pressure (thinking on feet, adapting to objections)
- Long-term outcomes (retention, burnout, career commitment)

 doi:10.20944/preprints202602.1127.v1

Validity concerns:

1. **Self-report bias:** People overestimate their motivation and future behavior
2. **Single-shot measurement:** One call script may not predict sustained performance
3. **No consequences:** AI personas experience no real stakes, success, or failure
4. **Ideal conditions:** No fatigue, distraction, competing demands, or emotional regulation challenges

Implications:

- Cannot claim behavioral impact without actual behavioral observation
- Relationships between AI-rated motivation and human behavior unknown
- High motivation in AI may not predict high performance in humans

Critical need:

- Human validation measuring actual behaviors:
  - Fundraising calls made and success rates
  - Persistence over multiple shifts
  - Adaptation to real-time challenges
  - Longitudinal tracking

4. English Language Only

All procedures conducted in English:

- Persona generation prompts
- Intervention scripts
- Reflection exercises
- Call scripts
- Coding and analysis

Validity concerns for non-English contexts:

1. **Translation issues:** Interventions may not translate well culturally or linguistically
2. **Training data imbalance:** LLMs have more English training data; may perform differently in other languages
3. **Cultural concepts:** Some motivational concepts may not translate (e.g., "job crafting" no direct equivalent in many languages)
4. **Emotional expression:** Languages differ in emotional lexicons and expression norms

Specific examples of potential problems:

- **Spanish:** "Impact" translates to "impacto" but cultural framing of personal vs. collective impact differs
- **Mandarin:** Job crafting concept requires multiple-sentence explanation; single-word translation unavailable
- **Arabic:** Beneficiary narratives may require different cultural exemplars and family structures

Implications:

- Findings may not generalize to non-English speaking populations
- Cultural moderation findings especially suspect (English-language stereotypes)
- Intervention adaptations needed for international contexts

Future directions:

- Conduct AI simulations in multiple languages
- Partner with international researchers for cultural adaptation
- Test whether patterns replicate across languages
- Validate with actual non-English speaking participants

 doi:10.20944/preprints202602.1127.v1

5. Specific LLM Versions (Temporal Limitation)

Models tested:

- Claude 3.5 Sonnet (November 2024 version)
- GPT-4 (version gpt-4-0613, November 2024)
- Llama 3 70B (November 2024 version via Together AI)

Concerns:

1. **Rapid model evolution:** LLMs update frequently (every 3-6 months)
2. **Version-specific behaviors:** Different versions may show different response patterns
3. **Training data updates:** Newer versions include more recent training data
4. **Architecture changes:** Model structures evolve (parameter counts, attention mechanisms)

Potential for replication failure:

- Re-running this study in June 2025 might yield different results with updated models
- Claude 4.0 or GPT-5 could show different effect sizes or moderation patterns
- Training data changes might shift cultural stereotypes or personality associations

Implications:

- Findings tied to specific model snapshot in time
- Replication studies should report exact versions
- Long-term validity uncertain as models evolve

Recommendations for field:

- **Version reporting standard:** Always report exact model version and date
- **Periodic replication:** Re-run key AI studies with updated models
- **Meta-analysis tracking:** Monitor whether AI findings change systematically as models improve
- **Human validation shield:** Human replication protects against model-specific artifacts

6. Unknown Persona Authenticity

Persona generation process:

- Used detailed prompts specifying demographics, personality, background
- LLMs generated life histories, motivations, values, experiences
- Assumed personas represent plausible psychological profiles

Validity concerns:

1. **Psychological implausibility:** Some trait combinations may not exist in real humans
    - Example: High agreeableness + high neuroticism + low conscientiousness + collectivist culture
    - LLM generates coherent persona, but real psychology may preclude this combination
2. **Idealized responses:** Personas may represent "ideal types" from training data rather than realistic individuals
    - Example: Collectivist personas all mention community; real people more variable
3. **Trait-behavior consistency:** AI personas show perfect trait expression; humans messier
    - High agreeableness personas always respond prosocially
    - Real humans show cross-situational inconsistency
4. **Life history coherence:** Generated backgrounds may be too coherent/logical
    - Real lives messier, more contradictory, less narrative-coherent

Examples of potential implausibility:

Persona 147:

- Age 52, high openness, low conscientiousness, PhD in physics, successful entrepreneur
- **Concern:** High openness + low conscientiousness rare in successful STEM careers requiring sustained focus

 doi:10.20944/preprints202602.1127.v1

Persona 089:

- Mexico, collectivist, working class, but highly individualistic language in reflection
- **Concern:** Training data stereotype overridden by other factors, creating implausible profile

Implications:

- Some personas may be psychologically impossible
- Moderation effects may be artificially strong due to perfect trait expression
- Individual difference findings require validation with real personality assessments

Methodological improvement needed:

- Validate persona generation against real personality data
- Check generated combinations against empirical trait correlations
- Potentially restrict to plausible combinations only
- Expert review of personas for psychological realism

7. Researcher Degrees of Freedom (Specification Uncertainty)

Many design choices shaped results; alternative decisions might yield different patterns:

Persona Generation:

- Prompt wording and detail level
- Trait descriptions and levels (we used 3-level; could use continuous)
- Cultural country selection
- Demographic distributions

Interventions:

- Specific beneficiary narrative (Sarah's story vs. others)
- Reflection prompt wording and ordering
- Duration decisions
- Control condition content

Measurement:

- Question phrasing
- Scale anchors (1-10 vs. 1-7 vs. Likert)
- Call script task specifics (donor details, scenario)

Analysis:

- Centering decisions (mean vs. grand-mean vs. uncentered)
- Effect-coding vs. dummy-coding
- Covariates included
- Moderation model specifications
- Multiple comparison corrections

Problem: We made defensible choices, but alternatives exist. Results may be sensitive to specifications.

Example of specification sensitivity:

Cultural coding:

- **Our choice:** Effects-coded (-0.5/+0.5)
- **Alternative:** Dummy-coded (0/1) with individualist as reference
- **Result difference:** Interaction coefficients would differ by factor of 2

Moderation testing:

- **Our choice:** Test each trait × condition separately
- **Alternative:** Three-way interactions (Trait A × Trait B × Condition)
- **Result difference:** Could reveal complex moderation patterns we missed

Implications:

- Results may not be robust to alternative specifications
- Other research teams might reach different conclusions from same data
- Specification choices should be justified theoretically

Mitigations:

 doi:10.20944/preprints202602.1127.v1

1. **Preregistration:** Future AI simulation studies should preregister analysis plans
2. **Multiverse analysis:** Test key findings across multiple reasonable specifications
3. **Specification curve analysis:** Plot results across many specification combinations
4. **Sensitivity reporting:** Report robustness to key decisions

What we did well:

- Theory-driven choices (effects coding for centering interpretability)
- Standard practices (Holm-Bonferroni for multiple comparisons)
- Transparent reporting (documented all decisions)

What could be improved:

- Preregistration (not done; post-hoc analysis)
- Sensitivity analysis (limited; tested few alternatives)
- Specification robustness checks (not systematically conducted)

8. Lack of Power Analysis and Sample Size Justification

We collected N = 240 (80 per LLM, 80 per condition) but provided no justification for this sample size.

Questions unanswered:

1. What power do we have to detect:
   - Main effects of conditions?
   - Two-way interactions (Condition × Personality)?
   - Three-way interactions (Condition × Trait A × Trait B)?
   - Cross-level interactions (Condition × LLM)?
2. What minimum detectable effect size (MDES) can we reliably find?
3. How does nesting affect power (personas within LLMs)?

Post-hoc power analysis:

Using simr package in R for multilevel power estimation:

Main effects (Condition):

- Observed effect: d = 3.15
- Power with N = 240: >99.9%
- MDES for 80% power: d = 0.45

Two-way interactions (Condition × Trait):

- Observed effect: β = 0.33-0.48
- Power with N = 240: 78-92%
- MDES for 80% power: β = 0.35

Implications:

1. Well-powered for main effects (massive overkill given large effects)
2. Adequately powered for moderate-large interactions
3. Underpowered for small interactions ($\beta < 0.30$)

For human replication:

- With expected smaller effects (÷2-3), need N = 120-180 per condition for 80% power
- Personality interactions require N = 200+ per condition

Should have been reported upfront:

- Justification for N = 240
- Power to detect theoretically meaningful effects
- Limitation: underpowered for small effects

9. Limited Diversity in Persona Characteristics

We systematically varied:

- Big Five personality (3 levels each = 243 combinations)
- Cultural background (10 countries)

 doi:10.20944/preprints202602.1127.v1

- Age, gender, education, SES

We did NOT vary:

- Disability status
- Sexual orientation
- Religious background
- Political ideology
- Relationship status
- Parenting status
- Immigration history (beyond country of origin)
- Mental health history
- Trauma exposure
- Neurodiversity (ADHD, autism, etc.)

Implications:

- May miss important moderators of intervention effects
- Generalizability limited to "typical" worker profiles
- Underrepresents important dimensions of diversity

Example missed moderation:

Hypothesis: Parents of college-age children might respond more strongly to student beneficiary narratives

We couldn't test this because parenting status wasn't systematically varied.

Future directions:

- Expand persona diversity along additional dimensions
- Systematically vary characteristics theoretically relevant to each intervention
- Include intersectional combinations (e.g., working-class immigrant mother)

10. Qualitative Coding Limitations

Strengths:

- Two independent coders
- Good-to-excellent reliability (ICC/$\kappa > .76$)
- Blind to hypotheses
- Consensus resolution

Limitations:

1. **Coder training:** Coders were graduate students trained by first author
   - Potential bias: Training may have transmitted expectations
   - Alternative: External coders with no study involvement
2. **Cultural competence:** Coders were U.S.-based English speakers
   - May not recognize cultural nuances in international personas
   - May impose Western interpretations on non-Western responses
3. **AI text characteristics:** Coding AI-generated text differs from human text
   - AI text more polished, coherent, structured
   - Easier to code (clearer themes, better articulation)
   - May inflate inter-rater reliability
   - May not generalize to messier human open-ended responses
4. **Demand characteristics in coding:** Coders knew responses were AI-generated
   - May have different expectations than coding human data
   - Potential bias in threshold for assigning high codes

Implications:

- Coded themes may not replicate with human data

 doi:10.20944/preprints202602.1127.v1

- Cultural interpretations especially suspect
- Reliability estimates may overestimate human data reliability

Improvements for future research:

- Use culturally diverse coding teams
- Blind coders to AI vs. human source
- Validate coding schemes on human data first
- Report differences in coding AI vs. human text

Summary of Critical Limitations

Most Critical (Threaten Core Validity):

1. **No human validation** - Cannot determine external validity
2. **Self-report outcomes only** - Cannot claim behavioral impact
3. **Unknown persona authenticity** - Psychological realism uncertain

Important (Limit Generalizability):

4. **Single context** - May not generalize to other work domains
5. **English only** - May not generalize internationally
6. **Cultural stereotyping risk** - Findings require careful validation

Methodological (Affect Interpretation):

7. **Researcher degrees of freedom** - Results may be specification-sensitive
8. **No power analysis** - Sample size not justified prospectively
9. **Limited persona diversity** - Missing important moderators

Temporal/Technical (Affect Replicability):

10. **Specific LLM versions** - Findings tied to November 2024 models
11. **Qualitative coding on AI text** - May not generalize to human coding

Our recommendation: The convergence of multiple LLMs, theoretical coherence of findings, and alignment with prior human research provide some confidence in directional patterns. However, effect magnitudes, specific moderation values, and behavioral predictions require extensive human validation before any substantive conclusions.

Concluding Thoughts

This research explored a provocative question: Can AI-simulated personas help us understand human motivation?

Our answer: **Conditionally yes, but only as a hypothesis-generating tool requiring rigorous human validation.**

What We Demonstrated

AI simulation offers genuine value for specific research purposes:

1. **Rapid iteration** during intervention design (tested 3 conditions across 240 diverse personas in one week vs. months for human research)
2. **Hypothesis generation** about mechanisms and moderators (identified three beneficiary impact pathways and four job crafting dimensions worth testing)
3. **Systematic exploration** of boundary conditions and individual differences (perfect representation of demographic and personality combinations)
4. **Language refinement** before human testing (identified clear vs. confusing intervention elements)

We also demonstrated what AI simulation is not: It is not a replacement for human research. Effect sizes may be inflated (our d = 3.15-3.97 vs. typical human studies d = 0.30-0.60). Moderation patterns may reflect training data stereotypes rather than authentic psychology. Qualitative mechanisms may be overly articulate compared to real human experience. And fundamentally, we have no way to determine which findings will replicate in humans without direct testing.

The Promise and the Peril

The promise: If used appropriately, AI simulation could dramatically accelerate the early stages of intervention research. What currently requires 12-24 months of human data collection for initial

 doi:10.20944/preprints202602.1127.v1

exploration might take 1-2 weeks with AI simulation, freeing resources for more extensive human validation of promising approaches. Researchers could test 10-20 intervention variations rapidly, then validate the 2-3 most promising with human participants. This could generate more knowledge more quickly, particularly for resource-constrained researchers and understudied populations.

The peril: If used carelessly, AI simulation could flood the literature with findings that don't replicate, perpetuate cultural and demographic stereotypes, and mislead practitioners into implementing ineffective interventions based on inflated effect size estimates. Worse, the speed and apparent sophistication of AI findings might create false confidence, short-circuiting the careful human validation that good science requires.

Critical Boundaries

Our findings establish clear boundaries for AI simulation in intervention research:

Appropriate uses:

- Prototyping intervention variations before human pilots
- Generating mechanistic hypotheses for experimental testing
- Identifying potential moderators worth investigating
- Exploring "what if" scenarios to guide human study design
- Refining intervention language and sequencing

Inappropriate uses:

- Estimating effect sizes for implementation planning
- Making claims about cultural differences without cultural validation
- Predicting actual behavioral outcomes
- Guiding organizational decisions without human data
- Publishing findings as standalone conclusions about human psychology

The Path Forward

This study represents a first step in understanding AI simulation's role in intervention science. We've shown it can generate interesting, theoretically coherent patterns. The field must now determine what those patterns mean.

Critical next steps:

1. **Direct replication studies** comparing identical interventions in AI simulation versus human samples across multiple contexts
2. **Meta-analytic synthesis** of AI vs. human effect size ratios to develop correction factors
3. **Mechanism validation** through experimental manipulation of AI-identified pathways in human studies
4. **Cultural psychology partnerships** to validate or refute AI-generated cultural patterns with actual culturally diverse participants
5. **Methodological standards development** for reporting, evaluating, and publishing AI simulation research
6. **Ethical guidelines** for appropriate and responsible use of AI simulation in social science

Our Responsibility

As researchers exploring this new methodological frontier, we have particular responsibilities:

- **Transparency** about limitations and uncertainties
- **Caution** in interpreting and communicating findings
- **Commitment** to human validation before substantive claims
- **Vigilance** against stereotype perpetuation
- **Honesty** about what we don't know

This paper attempts to model that responsible approach. We've reported large, theoretically interesting effects while simultaneously emphasizing their uncertain external validity. We've identified promising patterns while cataloging the many ways they might mislead. We've demonstrated utility while defining clear boundaries.

Final Reflection

 doi:10.20944/preprints202602.1127.v1

The question isn't whether AI simulation will be used in intervention research—it already is being used, and its use will accelerate. The question is whether we, as a field, will develop the methodological rigor, validation standards, and ethical guidelines to use it well.

AI simulation is not a shortcut around the hard work of human research. It's a tool for asking better questions before we undertake that hard work. Used wisely, it could make intervention science more efficient, creative, and ambitious. Used carelessly, it could undermine the very foundations of empirical psychology.

The promise is real. So are the risks. Our task now is to realize the former while mitigating the latter.

This study takes one step on that path. Many more steps—and much human validation work—lie ahead.

Appendix A. Complete Intervention Scripts

CONDITION 1: Control (Standard Training)

Target duration: 8-10 minutes (based on script content and reading time at conversational pace)

Introduction (Supervisor):

"Welcome to the University Development Office fundraising team. I'm Sarah Thompson, your training supervisor. Over the next few minutes, I'll walk you through everything you need to know to succeed in this role.

Job Overview:

Your position is University Fundraising Caller. You'll be making outbound calls to alumni during evening shifts, typically 3-4 hour blocks between 5pm and 9pm. The schedule is flexible—we ask for at least two shifts per week, but you can work more if you'd like.

You'll be paid $16 per hour, with paychecks issued bi-weekly. We expect callers to make between 15-25 calls per shift, depending on conversation length and connection rates.

Your Workspace:

You'll work in our call center, which has individual stations with computers, headsets, and scripts. The system automatically dials numbers from our alumni database and provides basic information about each alumnus—name, graduation year, major, past donation history if any, and current location.

The Call Script:

Let me walk you through the standard script. You'll introduce yourself, explain that you're calling on behalf of the University, and ask if they'd be willing to make a donation to the student scholarship fund. The system provides suggested donation amounts based on their past giving.

Here's the basic structure:

1. Greeting and introduction
2. Connection to the University (reference their graduation year, major if appropriate)
3. Explanation that you're calling to request support for student scholarships
4. Specific ask (donation amount)
5. Address objections if needed
6. Thank them for their time regardless of outcome

Objection Handling:

You'll encounter common objections. Here's how to respond:

'I'm too busy right now' → 'I understand completely. Would there be a better time to call back? I can schedule you for a specific day and time.'

'I already give to other charities' → 'That's wonderful that you support important causes. Many of our donors give to multiple organizations. Even a modest gift of $25 or $50 makes a real difference for our students.'

'I can't afford it right now' → 'I completely understand. Would you be willing to make a smaller contribution, even $10? Or we could reach out again in a few months when timing might be better.'

 doi:10.20944/preprints202602.1127.v1

'I didn't have a good experience at the University' → 'I'm sorry to hear that. Times have changed significantly, and we're working hard to improve the student experience. Your support would help current students have better opportunities than you did.'

Technical Systems:

The database shows:

- Alumnus name and contact information
- Graduation year and major
- Donation history (amounts and dates)
- Any special notes (deceased spouse, recently retired, etc.)

You'll log every call outcome:

- Donation committed (record amount and payment method)
- Call back requested (schedule in system)
- Not interested
- Wrong number / disconnected
- No answer / voicemail

Performance Metrics:

We track several metrics:

- Calls per hour (target: 5-7 completed conversations)
- Connection rate (% of calls where someone answers)
- Donation rate (% of conversations resulting in commitment)
- Average donation amount

Don't worry too much about metrics at first. Focus on getting comfortable with the script and having natural conversations.

Tips for Success:

1. Speak clearly and at moderate pace
2. Smile while talking (people can hear it in your voice)
3. Use the alumnus's name during conversation
4. Listen actively and respond to what they say
5. Stay positive even with rejections
6. Take brief notes during calls for accuracy
7. Stand up if you're feeling low energy

Questions?

[Answer any clarifying questions about logistics, schedule, technical systems, or procedures]

Next Steps:

You'll start with monitored practice calls tomorrow. I or another supervisor will listen and provide feedback. After 3-4 practice calls, you'll be on your own with support available if needed.

Any final questions before we wrap up?"

CONDITION 2: Beneficiary Impact

Duration: 8-10 minutes (same as control)

Introduction (Supervisor):

"Welcome to the University Development Office fundraising team. I'm Sarah Thompson, your training supervisor. Before we dive into the logistics, I want to tell you about someone whose life was changed by the work you'll be doing.

Beneficiary Narrative (3-4 minutes):

Her name is Sarah Martinez. She's currently a junior majoring in environmental engineering, and I met her last semester when she spoke at one of our donor appreciation events.

Sarah grew up in a small town in rural New Mexico. Her parents both work in food service—her dad is a cook at a local diner, her mom waits tables at a chain restaurant. They've always worked hard, but money was constantly tight. Sarah told me about grocery trips where her mom would

 doi:10.20944/preprints202602.1127.v1

meticulously add up every item to make sure they stayed within $50, putting things back if they went over.

Despite financial struggles, Sarah excelled in school. She loved science, especially environmental issues. In high school, she started a recycling program and helped organize a community clean-up of a local river. Her teachers encouraged her to apply to universities, but Sarah was terrified of the cost.

The night she got her acceptance letter to our University, she sat down with her parents at their kitchen table. Her dad had printed out a loan calculator showing the monthly payments they'd need to make. Her mom was crying—not tears of joy, but tears of worry. Sarah remembers the exact words her mother said: 'Mija, we want you to go so badly, but I don't know how we're going to do this.'

Sarah almost didn't come. She was drafting an email to decline admission and planning to attend the local community college instead—a good option, but without the research facilities and environmental engineering program she needed for her goals.

Then she got a second letter. A scholarship letter. The scholarship was funded entirely by alumni donations—people like the ones you'll be calling. Donors who gave 25, 50, $100, or more. Their collective generosity created a 15,000 annual scholarship that covered most of Sarah's tuition.

Sarah called her parents immediately. This time, her mother's tears were different. She told me about her mom sobbing with relief, saying 'Thank God, thank God' over and over. Her father, who rarely shows emotion, had to leave the kitchen because he was crying too.

Sarah is now thriving. She's conducting research on sustainable water systems, maintaining a 3.8 GPA, and has already secured a summer internship with an environmental consulting firm. After graduation, she plans to work on water infrastructure in underserved communities—giving back to places like the one she came from.

But here's what really stuck with me: Sarah told me that scholarship didn't just change her life. It changed her whole family's trajectory. Her younger brother, seeing her success, is now taking AP courses and planning for college too—something that wouldn't have seemed possible before. Her parents, instead of carrying crushing debt, are now able to save a little each month. One donation rippled through an entire family across generations.

The calls you're about to make? They create more Sarah Martinez stories. Every donation you help secure goes directly to students whose families are having those same difficult kitchen table conversations. These aren't abstract statistics—they're real people whose educational futures hang in the balance.

Transition to Logistics:

Now let me walk you through how you'll make those calls and connect donors with these opportunities...

[Continues with same technical training as control condition: job overview, workspace, call script, objection handling, technical systems, performance metrics, success tips]

Closing:

Remember: when you're making calls tonight, you're not just asking for donations. You're potentially creating that moment of relief and joy for another family. You're changing lives.

Any questions?"

CONDITION 3: Job Crafting Reflection

Duration: 20-25 minutes total

Introduction (Supervisor):

"Welcome to the University Development Office fundraising team. I'm Sarah Thompson, your training supervisor.

Before we cover the technical aspects of the job, I'd like to spend some time helping you think deeply about the work you'll be doing. Research shows that when people actively reflect on the meaning and purpose of their work, they not only feel more motivated but also perform better.

 doi:10.20944/preprints202602.1127.v1

So rather than jumping straight to scripts and logistics, we're going to do something different. I'm going to guide you through a series of reflection exercises. There are no right or wrong answers—just honest exploration of how you think about this work.

Reflection Exercise 1: Impact Visualization (4 minutes)

Close your eyes if you're comfortable doing so, or just focus your attention inward.

I want you to visualize a specific student—create them in your mind. Give them a name, an age, a face. What do they look like? What are they wearing? Where are they from?

Now imagine their family background. What do their parents do for work? What's their financial situation? Have they struggled with money? Picture their home, their neighborhood, their lived reality.

This student is incredibly talented. Maybe they're brilliant at mathematics, or passionate about medicine, or gifted in creative writing, or determined to become a teacher. Whatever it is, they have genuine potential. But there's a problem: they can't afford college. Not even close.

Picture this student sitting at a table with their parents, looking at college costs. What's the expression on their face? What are they feeling? Hope? Fear? Disappointment? Resignation?

Now imagine this student receives a letter in the mail. It's a scholarship notification. They've received funding that will make college possible—not easy, but possible. This scholarship was funded by donations from alumni, donations that came from calls just like the ones you'll be making.

Picture the moment they read that letter. What happens? Do they cry? Shout? Run to tell their parents? Sit in stunned silence? Picture their parents' reaction. Their siblings' reaction.

Now fast forward. This student graduates. They're standing in cap and gown. They've earned a degree that once seemed impossible. How has their life changed? What opportunities do they have now? What impact will they make?

Fast forward further. This graduate is now established in their career. They're supporting their family, contributing to their community, maybe helping other students the way they were helped. Trace the ripple effects as far into the future as you can.

Now open your eyes and take a moment to write down:

- Who was the student you visualized?
- What specific moment or image was most vivid for you?
- How did visualizing this impact change how you think about fundraising calls?

[4-minute reflection and writing time]

Reflection Exercise 2: Tracing Impact Ripples (4 minutes)

Now let's trace the chain of impact more systematically.

Start with your specific action: you make a phone call. An alumnus answers. You have a conversation. They decide to donate—maybe $50.

That $50 goes into a scholarship fund. Combined with other donations, it contributes to a student's scholarship. That's the first ripple.

Second ripple: The student can afford to attend or continue college. Without this funding, they might have had to drop out, or never enrolled at all. You've directly impacted their educational access.

Third ripple: The student earns their degree. This degree increases their lifetime earning potential by an estimated $1 million compared to just high school. They can support themselves and their family more effectively.

Fourth ripple: This graduate enters their chosen profession. Maybe they're a nurse caring for patients, a teacher educating children, an engineer designing infrastructure, a social worker helping families in crisis. Every day, they're making impact.

Fifth ripple: This person's children grow up in different circumstances because their parent has a degree and financial stability. Educational attainment often runs in families. You've potentially impacted the next generation.

 doi:10.20944/preprints202602.1127.v1

Sixth ripple and beyond: Your single phone call, which led to one $50 donation, which contributed to one scholarship, which enabled one education, which launched one career, which transformed one family... keeps rippling outward in ways you'll never fully know.

Now write down:

- What ripple effect resonated most strongly with you?
- How far into the future can you trace the impact of one conversation?
- Does thinking about these ripples change how you feel about making calls?

[4-minute reflection and writing time]

Reflection Exercise 3: Personal Values Connection (3 minutes)

Now I want you to think about your own values and experiences.

Why does educational opportunity matter to you personally? Have you benefited from support that made your education possible? Have you seen talented people held back by financial constraints? What do you believe about fairness, opportunity, and potential?

If you had unlimited resources, what problems would you want to solve in the world? How does education connect to those problems?

Think about the kind of person you want to be. What values are most important to you? Generosity? Fairness? Service? Making a difference? How does this fundraising work align with those values?

Write down:

- Why does educational opportunity matter to you personally?
- How does this work connect to your core values?
- What does it say about who you are that you're doing this job?

[3-minute reflection and writing time]

Reflection Exercise 4: Reframing the Task (4 minutes)

Most people would describe your job as 'making fundraising calls' or 'asking people for donations.' But that's just the surface description of what you're actually doing.

I want you to complete this sentence as many times as you can—at least 10 different ways:

'In this fundraising job, I'm not just making phone calls. I'm actually...'

Go beyond the obvious. Get creative. Get philosophical. What are you REALLY doing when you make these calls?

Examples to get you started:

- I'm actually... connecting generous people with students who need their help
- I'm actually... removing financial barriers to educational opportunity
- I'm actually... creating possibilities that didn't exist before
- I'm actually... participating in generational change

Now you continue. Generate at least 10 more completions. Push yourself to find new frames, new ways of thinking about this work.

[4-minute writing time]

Reflection Exercise 5: Envisioning Your Future Self (2 minutes)

Finally, I want you to imagine yourself making calls tonight, but with this new frame you've developed through these reflections.

When you pick up the phone, what will you be thinking about? When someone says they can't donate, how will you respond internally? When someone does donate, what will you feel?

Write a few sentences describing yourself as a fundraiser with this deeper sense of purpose and meaning.

[2-minute writing time]

Debrief:

Thank you for engaging deeply with these reflections. What insights emerged for you? What shifted in how you think about this work?

[Brief discussion of key insights]

 doi:10.20944/preprints202602.1127.v1

Transition to Technical Training (Condensed):

Now let's quickly cover the practical logistics. I'll move through this efficiently since we've spent time on the deeper purpose.

[Condensed version of control training covering: job overview (2 min), call script basics (2 min), objection handling (1 min), technical systems (1 min), quick tips (1 min)]

Closing:

You now have both the practical tools and the deeper sense of purpose. When you start calling tonight, remember: you're not just asking for donations. You're [use language from participant's reflections] creating opportunities, changing lives, making a difference that ripples far beyond any single call.

Questions?"

Appendix B. Qualitative Coding Schemes

Call Script Coding Scheme

Two independent coders rated all call scripts (N=240). Coders were blind to condition.

Dimension 1: Enthusiasm/Energy

- Scale: 1-10 continuous
- Definition: Perceived energy, warmth, and excitement conveyed in opening
- Anchors:
  - 1-2: Flat, mechanical, disinterested
  - 3-4: Polite but neutral
  - 5-6: Pleasant, moderately warm
  - 7-8: Enthusiastic, energetic
  - 9-10: Highly animated, infectious enthusiasm

Dimension 2: Beneficiary Mention

- Scale: 0-3 ordinal
- 0: No mention of students, beneficiaries, or impact
- 1: Brief generic reference ("supporting students")
- 2: Specific but not detailed ("scholarships help students afford education")
- 3: Detailed, vivid, or personalized beneficiary description

Dimension 3: Personalization Quality

- Scale: 1-10 continuous
- Definition: Extent to which script references donor's specific history, interests, or connection
- Anchors:
  - 1-2: Generic script, no personalization
  - 3-4: Minimal personalization (name, graduation year only)
  - 5-6: Moderate personalization (past donation amount, years since last gift)
  - 7-8: Strong personalization (specific details about donor's background)
  - 9-10: Exceptional personalization (demonstrates knowledge of donor's life, values, interests)

Dimension 4: Clarity of Ask

- Scale: 1-10 continuous
- Definition: How clearly and directly the caller requests a donation
- Anchors:
  - 1-2: No clear ask, vague or implied
  - 3-4: Indirect ask, hesitant
  - 5-6: Clear ask but somewhat apologetic
  - 7-8: Direct, confident ask
  - 9-10: Very direct and specific (exact amount, compelling rationale)

Dimension 5: Overall Persuasiveness

 doi:10.20944/preprints202602.1127.v1

- Scale: 1-10 continuous
- Definition: Holistic judgment of how persuasive the script would be to a real donor
- Anchors:
  - 1-2: Unlikely to persuade anyone
  - 3-4: Weak attempt, major flaws
  - 5-6: Adequate, some persuasive elements
  - 7-8: Good, would likely persuade some donors
  - 9-10: Excellent, would likely persuade most donors

Inter-Rater Reliability:

Dimension	ICC/Kappa	95% CI	Interpretation
Enthusiasm	ICC = .87	[.84, .90]	Excellent
Beneficiary Mention	κ_w = .82	[.77, .87]	Excellent
Personalization	ICC = .84	[.80, .88]	Excellent
Clarity of Ask	ICC = .89	[.86, .92]	Excellent
Persuasiveness	ICC = .87	[.84, .90]	Excellent

Resolution of Disagreements:

Disagreements > 2 scale points (continuous) or > 1 category (beneficiary mention) were flagged. Total scripts flagged: 23 (9.6%). Coders reviewed together, discussed rationale, reached consensus. Consensus codes used in analyses.

Reflection Coding Scheme

Open-ended reflection responses coded on four dimensions using 0-3 ordinal scales.

Dimension 1: Affective Impact

- 0: No emotional response mentioned
- 1: Mild positive feeling mentioned generically ("felt good," "nice story")
- 2: Moderate emotional response with some specificity ("really moved," "felt connected")
- 3: Strong emotional response with vivid description ("tears in my eyes," "deeply moved," "profound emotional impact")

Dimension 2: Cognitive Reframing

- 0: No evidence of reframing how they think about the work
- 1: Minimal reframing (acknowledging work has purpose beyond paycheck)
- 2: Moderate reframing (articulating specific new perspective on work's meaning)
- 3: Substantial reframing (detailed reconceptualization of work identity, purpose, or significance)

Dimension 3: Beneficiary Connection

- 0: No mention of beneficiaries or recipients
- 1: Generic reference to helping people
- 2: Specific reference to students or beneficiaries with some detail
- 3: Vivid, personalized connection to beneficiaries (visualization, specific individuals, emotional connection)

Dimension 4: Personal Relevance

- 0: No connection to personal values or experiences
- 1: Vague connection to general values ("education is important")
- 2: Specific connection to personal values or experiences
- 3: Deep personal relevance (detailed connection to own life story, core values, identity)

Inter-Rater Reliability (Weighted Kappa, Quadratic Weights):

 doi:10.20944/preprints202602.1127.v1

Dimension	κ_w	95% CI	Interpretation
Affective Impact	.81	[.75, .87]	Excellent
Cognitive Reframing	.78	[.71, .85]	Substantial
Beneficiary Connection	.84	[.78, .90]	Excellent
Personal Relevance	.76	[.69, .83]	Substantial

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337-351.

Berg, J. M., Wrzesniewski, A., & Dutton, J. E. (2010). Perceiving and responding to challenges in job crafting at different ranks: When proactivity requires adaptivity. *Journal of Organizational Behavior*, *31*(2-3), 158-186.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Bruning, P. F., & Campion, M. A. (2018). A role–resource approach–avoidance model of job crafting: A multimethod integration and extension of job crafting theory. *Academy of Management Journal*, *61*(2), 499-522.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504-528.

Grant, A. M. (2007). Relational job design and the motivation to make a prosocial difference. *Academy of Management Review*, *32*(2), 393-417.

Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. *Journal of Applied Psychology*, *93*(1), 48-58.

Grant, A. M. (2012). Leading with meaning: Beneficiary contact, prosocial impact, and the performance effects of transformational leadership. *Academy of Management Journal*, *55*(2), 458-476.

Grant, A. M., & Hofmann, D. A. (2011). Outsourcing inspiration: The performance effects of ideological messages from leaders and beneficiaries. *Organizational Behavior and Human Decision Processes*, *116*(2), 173-187.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341-361.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper No. w31122*.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin.

Wrzesniewski, A., & Dutton, J. E. (2001). Crafting a job: Revisioning employees as active crafters of their work. *Academy of Management Review*, *26*(2), 179-201.