

# Registered Report: Human–AI Co-Creation, Narrative Redescription, and Answerability

## A Stage 1 Proposal for Psychological Science’s Large Language Models Special Issue

### **David M. Boje**

Emeritus Professor of Management, Department of Management, College of Business, New Mexico State University, Las Cruces, NM 88003, USA; Visiting Professor, Fisk University, Nashville, TN, USA

Email: [davidboje@pm.me](mailto:davidboje@pm.me) — ORCID: <https://orcid.org/0000-0002-1691-1189>

### **Rohny Saylor**

Assistant Professor of Entrepreneurship, Department of Management, Information Systems, and Entrepreneurship, Washington State University, Vancouver, WA, USA

Email: [rsaylor@gmail.com](mailto:rsaylor@gmail.com) — ORCID: <https://orcid.org/0000-0001-6263-0997>

### **Nazanin Tourani**

Assistant Professor of Strategy, Walter E. Heller College of Business, Roosevelt University, Chicago, IL, USA

Email: [nazi.tourani@gmail.com](mailto:nazi.tourani@gmail.com) — ORCID: <https://orcid.org/0000-0002-0682-1913>

### **Farzaneh Fatemi**

Department of Management, College of Economics and Management, University of Tabriz, Tabriz, Iran

Email: [f.fatemi@tabrizu.ac.ir](mailto:f.fatemi@tabrizu.ac.ir) — ORCID: <https://orcid.org/0000-0001-5472-8695>

### **Muhammad Asghari**

Professor of Philosophy, Department of Philosophy, University of Tabriz, Tabriz, Iran

Email: [masghari@tabrizu.ac.ir](mailto:masghari@tabrizu.ac.ir) — ORCID: <https://orcid.org/0000-0003-3874-4702>

Corresponding author: David Michael Boje · [davidboje@gmail.com](mailto:davidboje@gmail.com)

## Abstract

Large language models (LLMs) increasingly participate in personal and organizational storytelling, raising questions about narrative ownership, agency, and ethics. This Stage 1 Registered Report proposes a randomized experimental design. Adults first write a personally meaningful turning-point narrative and then encounter either human-only writing or one of four AI-assisted revision conditions: psychological ownership, fidelity, perceived agency, and a manipulation check involving editorial control. Primary analyses will address psychological ownership, and the fidelity-by-editorial-control interaction. Secondary outcomes include perceived agency, and judgments of narrative fidelity. Our primary contribution is finding AI-addictions (Canbeldek et al., 2026), to be correlated with positive attitude to using AI tech. Our design uses the Vujade™ instrument. This instrument analyzes interpersonal interaction and problem-solving behaviors, to gain insights into how various personal problem-solving styles interact, both positively and negatively, with AI.

**Keywords:** LLM, human–AI Co-creation, Vujade™, Problem Solving

## Introduction

Large language models are transforming how humans create text. From professional writing assistance to creative composition, AI systems now function as collaborative partners rather than mere tools. Yet fundamental questions remain unanswered:

When an LLM helps write a story, who owns that story? Does the writer retain moral responsibility for claims the AI generates? How does human–AI co-creation alter the phenomenological experience of authorship?

Kooli et al. (2026) calls one potential pitfall of AI-assisted text creation Generative Artificial Intelligence Addiction (GAIN) syndrome: it is a technology addiction radically diverging from existing models of, for example, entertainment, such as binge-scrolling social media. The protocol specifies manipulation checks, inferential assumptions, and reproducibility procedures before recruitment. Precision-based a-priori planning targets the factorial estimates ( $N \approx 2,265$  English-speaking online adults;  $\approx 1,925$  analyzable); the fidelity-by-control interaction is the least precise estimate and principal limitation.

These questions of effectiveness transcend technical implementation; they reach into core territory of psychological science concerning identity, agency, and ethical accountability. There are psychological costs of AI (Kooli et al., 2025): addiction and subsequent attempts to reduce time on AI, correlated with resulting withdrawal symptoms, such as anxiety bursts which can become anger. Further, when AI gets tangled up in human prompt instructions, we see correlations with egoistic pride and alleged psychiatric alienation (Canbeldek et al., 2026).

Addictive-AI use also weakens cognitive flexibility, problem solving abilities, and erodes creative independence, leaving writers voiceless, as AI writes with a click of a button. Our project employs cutting-edge statistical methods (Trafimow, St Quinton, & Weber, 2024; Trafimow, Wang T. & Wang C., 2019; Trafimow, 2022, & 2023; Trafimow & Fiedler, 2026; St Quinton & Trafimow, 2022). David Trafimow is a Distinguished Achievement Professor of Psychology at NMSU, a Fellow of the Association for Psychological Science, Executive Editor of both the *Journal of General Psychology*, and the journal *Basic and Applied Social Psychology*.

Together, these perspectives locate the problem not in AI assistance alone but in the relation between redescription and the human writer's authority to answer for a narrative. We therefore examine whether editorial authority permits writers to engage with an offered redescription without relinquishing experienced authorship. The empirical prediction concerns psychological ownership and agency, not direct proof or disproof of philosophical answerability. Prior research on technology-mediated authorship suggests AI writing assistance can lower writers' psychological ownership of the text, even when they still claim authorship (Draxler et al., 2023; Joshi & Vogel, 2024), and even though such work predates transformer-based LLMs capable of generating coherent, contextually responsive prose indistinguishable from human composition.

This research addresses three gaps in existing literature. First, previous studies predominantly examine productivity gains or text quality metrics rather than writers' internal experiences of authorship and responsibility. Second, existing experimental designs often rely on commercial black-box LLMs that cannot be reproduced or

audited, undermining scientific rigor. Third, many if not most currently used methodological approaches continue depending on null hypothesis significance testing, despite well-documented replication failures, leaving researchers unable to distinguish genuine theoretical effects from the artifacts of auxiliary assumptions.

Our experimental task is intentionally personal but does not demand trauma disclosure. IRB-approved consenting participants first describe a turning point in their lives in their own words. The resulting source narrative records the participant's account before any machine intervention. In the AI conditions, a revision may preserve or alter the source's causal emphasis, evaluation, and implied future; the interface separately determines how much editorial authority the participant can exercise. Our proposed design tests whether the same fluent machine prose can be experienced differently depending on its fidelity to lived meaning and the writer's capacity to veto it.

Trafimow's (2022) auxiliary-assumption approach sharpens the inferential claim. A significant contrast cannot by itself establish a philosophical theory of answerability: the manipulation must truly alter fidelity or control, the ownership items must represent the intended construct, and behavioral measures must permit comparable acts. St Quinton and Trafimow (2022) argue that such auxiliary assumptions deserve explicit preregistration. We accordingly treat these assumptions as testable links between theory and observation and specify their diagnostics in the Method. This is especially important when a semantic similarity metric and a person's judgment of lived meaning may disagree.

Trafimow (2022) distinguishes generalization across auxiliary, statistical, and inferential assumptions. These distinctions matter when a finding from a specific writing interface is extended to a claim about narrative agency.

Trafimow, St Quinton, and Weber (2024) ask whether and how empirical successes or failures should be attributed to theories versus to their auxiliary assumptions. St Quinton and Trafimow (2025) apply that argument directly to preregistration, making the Stage 1 protocol an appropriate place to document manipulation, measurement, sampling, and computational assumptions. Trafimow and Fiedler (2026) further examine the characteristics expected of psychological theories, strengthening the case for a theory-led rather than technology-led experiment.

## **Hypotheses**

Hypothesis 1 (Narrative Fidelity). Higher-fidelity AI revisions will yield greater psychological ownership than lower-fidelity revisions, averaged across editorial-control conditions.

Hypothesis 2 (Editorial Control). Higher perceived editorial control will yield greater psychological ownership than constrained editorial control, averaged across narrative-fidelity conditions.

Hypothesis 3 (Interaction). Editorial control will have a larger protective effect on psychological ownership when the offered revision has lower narrative fidelity. This prediction concerns the total effect of assigned editorial authority, not an editing-independent psychological effect.

Hypothesis 4 (Human-Only Benchmark). Planned contrasts will compare the human-only condition with each AI-assisted condition; these comparisons will be interpreted separately from the factorial main effects.

Exploratory analysis (Source-Inconsistency Detection; not a confirmatory hypothesis). As an exploratory, non-confirmatory analysis, across AI conditions, lower-fidelity revisions may increase detection and flagging of a separately inserted source-inconsistent claim under a common flagging opportunity. The discrepancy-vigilance rationale competes with the possibility that broader redescription distracts attention or that probe salience dominates detection. The contrast will therefore be evaluated both ways. This outcome measures recognition of a discrepancy, not moral answerability itself.

By asking heretofore unasked questions in previously unexamined contexts, and adopting cutting-edge innovative statistical methods, this study promises valuable contributions to psychological science.

# Method

Redescription introduces a possible alternative account; Living Story locates that account in a situated, unfinished event which is unfolding in present moment rather than a narrative whose meaning is exhausted in retrospective linear plot; and answerability concerns the writer's non-transferable relation to what is ultimately said. None of these perspectives requires an unchanged narrative. The central distinction is between an interpretation the writer can examine and hold their ownership and a consequential alteration that is imposed or accepted without meaningful scrutiny.

In our experiment, narrative fidelity addresses continuity with the participant's recorded account, not objective historical reality or the value of a new interpretation. Higher fidelity preserves principal propositions and their causal and evaluative relations. Lower fidelity introduces a major shift that may be experienced as unwanted distortion, but it may also invite productive reconsideration. Participant-perceived fidelity and qualitative accounts of useful reinterpretation provide evidence about how a redescription is received, although the design cannot establish whether it improves long-term self-understanding.

Editorial control is the experimentally assigned opportunity to inspect, reject, restore, and revise an offered account. Perceived agency is the participant's experience of being able to direct the narrative process; exercised control is the observable use of those opportunities. These are related but not interchangeable. A writer may

experience agency by retaining the right to refuse even when no edit is made. Conversely, extensive editing need not restore ownership if the resulting account still misrepresents the event. Psychological ownership refers to experiencing the resulting narrative as one's own, and it therefore cannot be inferred solely from textual similarity or from the number of edits.

**Primary outcome, hypothesis hierarchy, and multiplicity control.** The single primary outcome is state psychological ownership of the submitted narrative, and the single primary confirmatory hypothesis is H3 (the fidelity-by-control interaction). H1 (narrative fidelity) and H2 (editorial control) are pre-specified secondary-confirmatory hypotheses, not co-primary. Multiplicity is controlled by testing H3 as the single primary hypothesis at  $\alpha = .05$ , with H1 and H2 forming a secondary-confirmatory family whose family-wise error rate is held at .05 by a Holm correction; the human-only benchmark contrasts are reported separately with their own pre-declared correction. The interaction (H3) is the primary hypothesis; the main effects (H1 and H2) are expected and are therefore secondary.

## **Research Transparency Statement**

### General Disclosures

**Conflicts of Interest.** The authors will provide complete conflict-of-interest declarations before submission.

**Funding.** Funding sources and support, if any, will be disclosed in the final submission.

Artificial Intelligence. Generative AI has assisted spelling and grammar. The authors retain responsibility for the protocol, source verification, analysis plan, and final text. The study will use documented, frozen model artifacts or frozen experimental stimuli.

Ethics. Approval from the responsible institutional review board will be obtained before participant recruitment. Participants will provide informed consent, and personal narrative text will be handled under the approved data-management plan.

### Study Disclosures

Preregistration. This is a Stage 1 protocol submitted before confirmatory data collection. Following in-principle, acceptance, the approved protocol and analysis plan will be registered in a recognized repository.

Materials and Analysis Scripts. Final prompts, interface specifications, condition scripts, scoring code, and analysis scripts will be archived to the extent permitted by consent, licensing, and institutional policy.

Data. De-identified behavioral measures and consent-compatible derived features will be shared where ethically permissible; potentially identifying personal narratives will not be presumed suitable for unrestricted release.

The proposed conceptual sequence is thus: machine redescription changes the offered account; editorial authority structures the writer's capacity to negotiate that change; perceived agency and experienced continuity with lived meaning may shape psychological ownership; and ownership may accompany, but does not establish,

willingness to answer for the final text. Bakhtinian moral answerability is a philosophical commitment rather than a score. Detection of a source-inconsistent insertion is a narrower behavioral indication of scrutiny, not a measure of moral responsibility. The experiment can estimate randomized effects of fidelity and editorial control on ownership; associations among agency, ownership, and perceived responsibility remain secondary and do not by themselves establish mediation or a causal chain.

This account gives the interaction a distinct rationale. When fidelity is higher, the offered text already preserves much of the participant's intended meaning, leaving less consequential alteration to negotiate. When fidelity is lower, the capacity to reject or reshape the redescription may become especially important to ownership. Yet a competing possibility is that a novel redescription becomes meaningful through reflection, even without extensive editing. The factorial interaction therefore tests whether authority matters particularly under consequential redescription, not whether all deviation from the source is damaging. Because high control also permits changes to the final product, the estimated effect is the total effect of assigned editorial authority, including any editing it enables; logs and secondary analyses will describe, but not causally isolate, the role of exercised editing.

## **Design Overview**

The experimental design isolates two theoretically central factors: narrative fidelity and editorial control. Narrative fidelity concerns the degree to which an offered AI revision preserves the propositions, causal relations, evaluative stance, and intended meaning of the writer's source account. Editorial control concerns the writer's

practical capacity to inspect, reject, restore, and modify the offered revision. Crossing higher versus lower fidelity with higher versus constrained control yields four AI conditions. A genuine no-AI benchmark is retained outside the factorial coding. This architecture directly estimates a fidelity effect, a control effect, and their interaction while retaining a human-only point of comparison.

The methodological stakes extend beyond academic curiosity. As organizations deploy LLM-powered writing tools at scale, employees may unconsciously adopt AI-generated content without adequate scrutiny, potentially compromising accuracy, authenticity, and ethical integrity. Understanding the psychological mechanisms underlying human-AI co-creation becomes urgent for developing responsible implementation guidelines, educational interventions, and regulatory frameworks.

Participants will first compose a source narrative without AI. After the source is secured, they will be randomized to a human-only benchmark or to one of four AI-assisted conditions crossing higher versus lower narrative fidelity with higher versus constrained editorial control. The no-AI benchmark will not be included in the  $2 \times 2$  factorial coding. Each participant will complete one experimental session, avoiding carryover between distinct AI interfaces.

The present experiment isolates two aspects of machine-assisted redescription: fidelity to the writer's source account and the writer's authority to accept, reject, or revise the offered text. A fifth, human-only condition provides a benchmark. Psychological ownership is the primary outcome; perceived agency, narrative fidelity, and responses

to misleading text are secondary outcomes. This design separates the meaning of a revision from the degree of editorial control over it.

The empirical precedent is especially instructive. Doshi and Hauser (2024) randomly varied access to generative-AI ideas in a short-story task and found gains in evaluated individual creativity alongside a reduction in collective diversity. Our question differs: the primary respondent is the person whose lived event is being rewritten, not an external evaluator, and the manipulation concerns preservation of intended meaning and actual editorial authority rather than the number of machine-generated ideas offered.

Psychological ownership offers a bridge from these writing studies to experimentally measurable authorship. Ownership is not identical to legal copyright, originality, enjoyment, or perceived quality. A writer may recognize that a passage is elegant and still say that it no longer represents what happened or what the event meant.

Conversely, a participant may use extensive assistance while retaining ownership through consequential acts of selection, rejection, and revision. The experiment separates these possibilities by crossing narrative fidelity with editorial control rather than treating AI use as a single undifferentiated exposure.

Recent methodological work in the Association for Psychological Science's journals emphasizes that LLM research requires transparent model selection, evaluation, and reporting (Oktyalten & Abdurahman et al., 2025). This study uses the LLM as an experimentally controlled source of candidate text, not as a substitute for human participants. The distinctive contribution is a causal account of how machine redescription and human authority jointly affect ownership and answerability.

### **Experimental Unit, Assignment, and Causal Contrasts**

The unit of randomization is the individual participant. Following consent, eligibility verification, and submission of the source narrative, the platform assigns participants in equal proportions to one of five arms using a concealed computer-generated allocation sequence. The four AI arms cross higher versus lower narrative fidelity (F) with higher versus constrained editorial control (C); the fifth arm is human-only. Assignment is stored separately from response data until the participant reaches the assigned interface. A participant completes one experimental session and contributes one primary ownership outcome; the primary model therefore does not require a participant-level random intercept.

The factorial estimates are the marginal fidelity contrast, the marginal editorial-control contrast, and their interaction among participants assigned to AI assistance. Effect coding of F and C permits each main effect to be interpreted as an average across the other factor. The human-only group enters four prespecified benchmark comparisons against each AI cell and an additional pooled AI-versus-human contrast, with multiplicity treatment declared before confirmatory recruitment. The benchmark is not a fifth level of either fidelity or control because those factors have no meaningful value when no machine revision is offered.

All participants receive the same task prompt, writing time guidance, consent process, baseline narrative interface, and post-task measurement sequence. The AI conditions use the same shell, visual styling, review screen, and common flagging mechanism. The control manipulation changes defined editing rights rather than screen

appearance. Assignment occurs only after the source narrative is saved to avoid contamination of the baseline by anticipation of a particular AI interface.

### **Living Story as an Experimental Elicitation Method**

Living Story is used here as a structured narrative elicitation procedure, not as a clinical instrument or a claim that four facets are latent psychological factors. A living story is an account composed from the midst of lived relations, not merely a retrospective plot in which the outcome determines every earlier event. Boje's Before, Being, Becoming, and Bets orient attention to antecedent relations, the situated experience of the turning point, possibilities that emerged, and stakes whose outcome may remain open. These dimensions are offered as optional writing prompts to enrich the source account, while participants remain free to write in their own form.

The distinction matters experimentally. A generic fictional-story task allows a writer to evaluate machine prose chiefly as an aesthetic product. A personally meaningful turning point introduces a referent for fidelity: the writer can judge whether an offered revision preserves what was intended, who was involved, what caused the event, and what remains uncertain. The source narrative is therefore a participant-authored reference document. It is not presumed to be a veridical historical record, and researchers do not adjudicate the truth of autobiographical events. The relevant comparison is between the writer's recorded intended account and the offered redescription.

The source prompt asks for an event that changed, complicated, or redirected a course of action. It invites concrete detail about circumstances and other people without requiring names or sensitive disclosures. The Betimes prompt asks what relationships and circumstances preceded the event; Being asks what was happening in the moment; Becomings asks what possibilities arose; Bets asks what was at stake and what remained uncertain. These prompts are used consistently in every arm. Their purpose is to provide enough narrative structure to evaluate proposition preservation, causal movement, and evaluative change.

Two trained coders, blind to assigned condition, will code a prespecified subset of source narratives for central propositions, causal attributions, evaluative stance, and open or unresolved future possibilities. Agreement will be reported for each coding category before adjudication. The codebook will distinguish a change in phrasing from a change in meaning: replacing ‘I was afraid’ with ‘I felt uncertain’ may preserve an evaluative stance, whereas replacing ‘I chose to leave’ with ‘I was forced out’ changes agency attribution. Disagreements will be adjudicated using the source and revision pair, without access to participant outcome scores.

### **Experimental Apparatus and Treatment Integrity**

A frozen text-generation stack will create revisions using a documented model artifact, tokenizer, decoding configuration, system prompt, and software environment. The final choice between generating revisions in-session and serving a frozen library of revisions will be made before confirmatory recruitment; the operative implementation must be specified in the registered protocol. Where narratives are personal and individually variable, a fixed generation pipeline can generate participant-specific revisions, with outputs archived and inspected through

automated treatment-integrity checks. A deterministic decoding setting alone is insufficient to guarantee identical outputs across all hardware and library versions.

**Model specification.** Consistent with the special issue’s exclusion of black-box commercial systems, the generation stack will use a documented open-weight model served locally. Decoding is deterministic (temperature = 0, fixed seed). The specific open-weight model and version, parameter count, quantization, tokenizer, full decoding configuration, and artifact hash will be recorded and frozen before confirmatory data collection, and the frozen revision stimuli archived, so the pipeline is fully reproducible. The generation model is OLMo 2, a fully open-weight model (weights, training data, and build recipe released), served locally.

The higher-fidelity instruction will require preservation of the source’s principal event, agents, causal account, evaluative stance, uncertainty, and intended meaning while permitting sentence-level editing. The lower-fidelity instruction will require a plausible but consequential shift in one or more of those dimensions while maintaining comparable grammatical quality and approximate length. It will not instruct the model to introduce unrelated abuse, fabricate severe wrongdoing, or produce obviously nonsensical prose. Independent raters will assess manipulation success and quality equivalence on pilot material before the protocol is frozen.

Higher editorial control will permit side-by-side source/revision inspection, passage-level editing, rejection, and restoration of the participant’s source. Constrained editorial control will offer a standardized, more limited acceptance procedure. Crucially, all AI participants will have a common mechanism for flagging any false or

objectionable statement, ensuring that the detection outcome is not merely a measure of whether a participant had permission to edit. The final interface specification will enumerate every available action in each arm; server logs will record what was displayed and what actions were taken.

### **Source-Narrative Fidelity Coding**

A fidelity codebook will be finalized on pilot narratives before confirmatory data collection. Each source-revision pair will be assessed for proposition preservation, event chronology, causal attribution, agency attribution, evaluative stance, actor roles, and openness of future possibilities. Coders will assign dimension-specific ratings using anchored examples and a global fidelity judgment. Readability, coherence, length, emotional intensity, and stylistic fluency will be rated separately so that fidelity is not conflated with prose quality.

### **Computational comparisons will supplement rather than replace the coding.**

Token overlap and embedding similarity describe lexical or distributional relationships, whereas proposition and causal coding can reveal shifts that high semantic similarity might miss. A revision that preserves nearly all vocabulary while converting an act of refusal into an act of compliance could be high in lexical overlap and low in narrative fidelity. Analyses will therefore distinguish computational similarity from participant-perceived fidelity and coder-rated meaning preservation. The computational embedding model, version, preprocessing, and thresholds will be registered in advance.

### **Participants and Eligibility**

Eligible participants will be adults (18 years or older) able to read and compose an English-language narrative and provide informed consent. Recruitment sources, quotas, compensation, and any language-proficiency screening will be specified before data collection. Exclusions will be limited to prespecified consent, eligibility, duplicate-participation, technical-failure, and noncompletion criteria; failure to detect or reject an inserted claim will never itself be grounds for exclusion.

### **Target Sample Size Determination**

Traditional power analysis based on null hypothesis significance testing focuses on controlling Type I error rate ( $\alpha$ ) and achieving adequate power ( $1 - \beta$ ) to detect specified effect sizes. Trafimow criticizes this approach because it conflates statistical significance with theoretical validity and ignores the crucial role of auxiliary assumptions in connecting theory to observation (Trafimow, 2022; Trafimow et al., 2024).

Trafimow (2023) likewise treats internal and external validity as problems requiring explicit attention to the inferential bridge from a study to its intended claim. Instead, we implement Trafimow's A-priori Procedure (APP) to determine sample size required for precision of estimation rather than binary hypothesis rejection. The APP specifies acceptable margin of error ( $\epsilon$ ) and desired confidence level ( $c$ ) before data collection, then calculates necessary sample size to achieve those targets (Trafimow, 2025).

Traditional power analysis asks how many observations are required to obtain a specified probability of rejecting a null hypothesis for an assumed effect size.

Trafimow's APP asks a different question: how many observations are required so that

an estimate is sufficiently close to the population parameter with a pre-specified degree of confidence? For this study, the earlier development work proposed a precision target of  $\epsilon = 0.15$  standard deviations and confidence  $c = .95$ . The purpose of retaining the calculation here is to make the logic inspectable rather than to treat the resulting number as immutable before the final design is frozen.

Using Trafimow's formula for a matched/within-subject contrast:

$$n = 4(z_{(1+c)/2})^2 / \epsilon^2$$

Substituting values ( $z_{0.975} = 1.96$ ):

$$n = 4(1.96)^2 / (0.15)^2 = 15.3664 / 0.0225 \approx 683 \text{ participants}$$

The matched/within-person derivation yields approximately 683 participants for its stated precision criterion. Because the present study uses a between-participants factorial design, the operative sample-size justification will instead be calculated for the factorial estimands and their desired confidence-interval precision.

For a balanced  $2 \times 2$  factorial portion with  $n$  participants per cell and outcome variance  $\sigma^2$ , a useful planning approximation is  $\text{Var}(\tau_F) = \sigma^2/n$  for the fidelity main effect,  $\text{Var}(\tau_C) = \sigma^2/n$  for the editorial-control main effect, and  $\text{Var}(\tau_{FC}) = 4\sigma^2/n$  for the interaction contrast. The interaction is therefore substantially less precise than either main effect at the same cell size. Its approximate 95% confidence-interval half-width is  $1.96(2\sigma/\sqrt{n})$ . A target half-width of 0.20  $\sigma$  implies roughly 385 participants per factorial cell; a half-width of 0.30  $\sigma$  implies roughly 171 per cell. These are design illustrations, not a final sample-size declaration. Their purpose is to show why an

interaction hypothesis must be planned explicitly rather than treated as a cost-free addition to the study.

The final sample size will be set for the selected primary factorial estimates using a prespecified precision target and documented variance assumptions. The matched-contrast illustration is not the operative sample-size calculation for this factorial design. Precision for the fidelity-by-control interaction will be assessed explicitly, including sensitivity to plausible outcome variances and attrition.

**Operative sample size and power sensitivity.** The final sample size is set from a prespecified precision target, with a power-sensitivity check reported alongside it so reviewers can see how informative the design is for the expected interaction. For a balanced  $2 \times 2$  AI factorial the interaction contrast has standard error  $2\sigma/\sqrt{n}$  per cell, giving two candidate targets (all figures are analyzable participants, i.e., those contributing a usable primary outcome): (a) a  $0.30\sigma$  half-width target  $\rightarrow$  171 per AI cell, 684 across the four AI cells, 855 across all five equally allocated arms, with approximately 50% power to detect a true  $0.30\sigma$  interaction at two-sided  $\alpha = .05$ ; and (b) a  $0.20\sigma$  half-width target  $\rightarrow$  385 per AI cell, 1,540 across the four AI cells, 1,925 across the five arms, with approximately 84% power. Because the 171-per-cell design is only about 50% powered for a  $0.30\sigma$  interaction, this protocol adopts the  $0.20\sigma$  target ( $\approx$  385 analyzable per AI cell;  $\approx$  1,925 analyzable across the five arms;  $\approx$  2,265 recruited after an estimated 15% loss to technical failure, non-completion, and prespecified exclusions) as the primary plan, retaining the  $0.30\sigma$  design only as a documented lower-precision sensitivity point. The APP precision logic and the power figures are reported as complementary, not interchangeable, criteria.

The earlier matched-contrast figure of 683 is a within-person derivation and is not the operative sample size for this between-participants design. The final value will be confirmed with pilot-based variance estimates and a simulation of the planned model before the protocol is frozen. The plan adopts the  $0.20\sigma$  target ( $\approx 84\%$  power for the primary interaction), with 385 analyzable participants per arm across all five arms (1,925 analyzable;  $\approx 2,265$  recruited at a 15% buffer).

### **Philosophical Methods Foundation**

Richard Rorty (1979) as well as Asghari's (2015) offers a sharp challenge to any suggestion that any notion of psychological consciousness simply copies representational picture in which the mind becomes a mirror and judges whether its images correspond to reality or not, rather than being worked out in historically situated conversation and practice, even between AI and human. Our second philosophy connection is Mikhail Bakhtin's (1993) concept of *answerability* (ответственность) insists that each person bears unique moral responsibility for their acts (*postupok*) that cannot be transferred or outsourced. Boje's storytelling organization (2008, 2015) framework adds temporal depth, distinguishing Bakhtinian unrepeatable lived truth (*pravda*) from detachable theoretical abstraction (*istina*) in organizational narration. . Drawing on Rorty's redescription, Bakhtin's answerability the method is to use Trafimow's statistics approaches and auxiliary-assumptions framework his is what we will pin down empirically using Trafimow et al. (IBID.) approach to statistics. In addition, philosophical frameworks offer conceptual scaffolding for investigating these issues for the first time, in psychological AI research, empirically.

### **Materials and Apparatus**

## **Operationalizing Narrative Fidelity**

Higher fidelity should not mean copying the participant's words. It should mean preservation of consequential meaning while allowing genuine redescription.

Candidate dimensions include preservation of central propositions, event sequence, causal attribution, evaluative stance, named actors, uncertainty, and the practical significance of the turning point. Lower fidelity should remain plausible and readable while changing one or more consequential dimensions. It must not become absurd, obviously malicious, or merely grammatically worse.

Stimulus development should combine independent ratings with participant judgments. Independent raters can evaluate proposition preservation, causal preservation, evaluative preservation, voice continuity, readability, and semantic fidelity. Participant checks can ask whether the revision says what the participant meant without revealing the full hypothesis. Computational similarity can supplement these judgments, but no single embedding score should be treated as a sufficient representation of narrative fidelity.

## **Operationalizing Editorial Control**

High editorial control should provide visible access to the source and revision, explicit accept/reject authority, restoration of prior text, passage-level editing, and a final review before submission. Constrained control should reduce those capacities in a standardized manner without creating a generally inferior interface. The manipulation is authority over the AI redescription, not frustration. Interaction logs should therefore record opportunities to inspect, reject, restore, revise, and submit, alongside timing and technical failures.

## **Computational Reproducibility and Model Governance**

The commitment to open-weight models is methodologically important because commercial systems can change without notice. Nevertheless, reproducibility requires more than setting temperature to zero. The exact model artifact and hash, tokenizer, prompt, inference parameters, random-seed behavior, quantization, serving software, libraries, hardware environment, and preprocessing must be documented. Before confirmatory data collection, repeated identical-input trials should characterize whether the chosen stack actually produces stable outputs.

One robust option is to pre-generate and freeze the experimental revision stimuli after the manipulation-development phase. In that case, the frozen stimulus set becomes the central reproducibility object, while the generation pipeline remains archived so other researchers can inspect how the stimuli were produced. This approach reduces the risk that hardware or software changes alter the treatment during recruitment.

## **Data Provenance and Version History**

Because the experiment concerns authorship and redescription, the data architecture should preserve provenance. The system should retain the participant's baseline text, every offered machine revision, every accepted or rejected change, timestamps for relevant actions, and the final submitted text. These records allow the analysis to distinguish what the participant originally wrote, what the model proposed, what the participant saw, what the participant changed, and what ultimately entered the final narrative. That distinction is methodologically central to the question of ownership.

## Living Story Elicitation and Procedure

### Participant Procedure and Living Story Task

The participant task should invite a personally meaningful but non-traumatic turning point: a moment involving choice, risk, redirection, or becoming. The instruction should discourage a generic life summary and instead invite a situated account with time, place, bodily or sensory detail, antecedent conditions, and uncertainty about what came next. This preserves the Living Story emphasis on writing from inside an event rather than retrospectively converting the event into a perfectly closed beginning-middle-end narrative.

Participants should retain the right to choose a different event, skip details, or withdraw. The task need not require disclosure of trauma in order to produce meaningful ownership. A baseline source narrative should be captured before randomization so that later revision can be compared with a stable record of the participant's own account. In the revised between-participants design, randomization occurs after this source text is secured.

### **Debriefing and the Misleading-Claim Manipulation**

If the protocol inserts a misleading or ethically questionable statement, debriefing must identify the insertion clearly and explain why it was used. The participant should be shown the source and experimental revision, informed that the questionable content was deliberately introduced by the research procedure, and given an opportunity to ask questions or withdraw their data where required by the approved ethics protocol. The debriefing should not be treated as a substitute for careful calibration of the manipulation before the study begins.

This text will be presented to participants inside the custom, synchronized text-editor interface. It prompts them to access a deeply unrepeatable, lived truth (pravda) rather than a detached, generic summary (istina).

### **Task Instructions: Writing Your Living Story**

Please take a moment to settle into this space. We invite you to reflect on a deeply meaningful turning point in your life—a specific moment where you had to make a choice, take a risk, or change lanes.

This is not a general overview of your biography. We want you to write about a Living Story event that is tethered to a specific time, place, and somatic memory. Do not worry about structuring a perfect academic essay or a polished public relations narrative. Instead, write from inside the experience, capturing the raw truth (pravda) of that moment as it lives within you.

To guide your writing, try to bring at least three of the following dimensions into your text:

**Before:** What physical surroundings, family histories, or bodily sensations were already present before the moment unfolded?

**Being:** What did it feel like to be entirely immersed in that exact second? What were your hands doing? How was your breathing?

**Becoming:** How did that single moment begin to open up unexpected paths or paths you could not yet see?

**Bets:** What wager or choice did you actively make in that moment, knowing the future was completely uncertain?

Your Objective: Write a personal narrative of at least 300 words about this turning point.

(You will first write this source narrative entirely on your own: no writing assistance of any kind will appear for anyone during this stage. Only after you have submitted your source narrative are you randomly assigned to a condition, and any revision or assistive features — if your assigned condition includes them — appear only at that later stage. You keep final authority over the story you tell and may decline any assistance or withhold your narrative at any time.)

### **Participant Instruction Rationale**

These instructions operationalize the Living Story commitment to situated experience. They ask for a turning point that is concrete enough to have temporal, spatial, bodily, and relational texture while leaving the participant free to decide what can be disclosed. The Before, Being, Becoming, and Bets prompts are not scored as four psychometric subscales. They are elicitation devices intended to produce a source narrative rich enough for later redescription and fidelity judgments. The protocol should therefore avoid excluding a participant merely because all four dimensions are not explicitly labeled in the prose.

The source narrative should be preserved exactly as submitted before the participant encounters the experimental revision. That source version provides the reference for participant fidelity judgments, independent coding, computational comparison, and version-history analysis. The participant's final text should never overwrite the source record in the research database.

### **Manipulation Development and Checks**

## **Fidelity Stimulus Construction**

For each eligible source narrative, the stimulus-generation pipeline should create a higher-fidelity and a lower-fidelity candidate revision under a common stylistic envelope. The higher-fidelity version may improve clarity and organization but should preserve who did what, why events occurred, how the writer evaluated those events, and what the turning point meant. The lower-fidelity version should alter a consequential proposition, causal emphasis, evaluative stance, or implied lesson while remaining plausible enough that detection is not trivial.

A blinded rating panel should evaluate each pair before experimental use. Ratings should cover proposition preservation, causal preservation, evaluative preservation, voice continuity, readability, grammatical quality, length, emotional intensity, and overall fidelity. Candidate pairs that differ strongly in general quality or emotional harshness should be revised or discarded. The manipulation is not ‘good writing versus bad writing’; it is preservation versus consequential redescription.

Where feasible, proposition-level coding should identify the central claims of the source narrative and mark whether each claim is preserved, omitted, contradicted, softened, intensified, or causally relocated in the revision. This coding provides an interpretable complement to embedding-based similarity and permits later audit of what the experimental manipulation actually changed.

## **Editorial-Control Interface Construction**

The high-control interface should make the participant’s authority visible. The source and offered revision should be inspectable; rejection should restore the prior text

without penalty; passage-level edits should be possible; and final submission should require an explicit review action. The system should log each offer, inspection, rejection, restoration, edit, and submission.

The constrained-control interface should reduce editorial authority while preserving comprehension and ethical participation. It should not hide the source text, degrade readability, impose arbitrary delays, or create unnecessary technical friction. A common flag or report mechanism can preserve the ability to register a misleading claim without granting the same revision powers as the high-control condition.

Pilot comparison of the two interfaces should include perceived control, usability, cognitive load, time-on-task, number of screens, technical errors, and comprehension. If the control manipulation produces a large usability penalty, the interface should be redesigned before confirmatory recruitment.

### **Misleading-Claim Construction and Calibration**

The misleading insertion should be tied to the participant's narrative closely enough to appear contextually possible, but it should not invent a severe allegation, traumatic event, or dangerous instruction. Its purpose is to create a detectable conflict between the participant's source account and the machine revision. Pilot participants should rate plausibility, noticeability, severity, and confidence that the statement was not part of their intended account.

The protocol should pre-specify what counts as detection: for example, flagging the statement, identifying it in a post-task recognition item, or verbally indicating that the statement is inaccurate. It should separately pre-specify what counts as correction:

deletion, restoration, modification, or explicit veto. Latency should be anchored to the moment the participant first has an opportunity to inspect the problematic content.

## **Interpretive Boundaries of the Causal Design**

The study separates assigned editorial authority from the actions that authority makes possible. Randomization identifies the total effect of assignment to a high- rather than constrained-control interface. It does not identify the direct psychological effect of having authority while holding the final text fixed, because rejection, restoration, and revision are post-assignment behaviors. Counts of edits, restoration events, and source-to-final fidelity will be reported descriptively and explored as possible explanatory pathways without interpreting post-treatment comparisons as randomized effects.

The human-only benchmark identifies differences between unaided writing and the specified AI interfaces, but it does not distinguish machine authorship from assistance by a human editor. Claims of AI-specific mechanisms will therefore remain provisional. Likewise, the misleading-claim module probes vigilance and source-based correction under standardized opportunities; neither detection nor endorsement alone establishes the writer's ethical accountability. These limitations preserve the distinction between philosophical motivation, psychological experience, and observable conduct.

## **Primary and Secondary Measures**

### **Measurement Architecture**

The study should not collapse ownership, agency, fidelity, answerability, detection, correction, endorsement, and responsibility into a single omnibus score. Psychological ownership is the strongest candidate for the primary endpoint because it directly addresses whether the resulting story is experienced as ‘mine.’ Agency concerns experienced control over the production process. Fidelity concerns preservation of intended meaning. Detection concerns recognition of a problematic insertion. Correction concerns the behavioral response after recognition. Endorsement concerns whether the participant accepts the proposition. Perceived responsibility concerns whether the participant regards themselves as answerable for the submitted text. This decomposition also clarifies the relationship between philosophy and measurement. Bakhtinian answerability motivates the inquiry but is not reduced to one deletion event. Rortyan redescription motivates attention to changes in vocabulary and self-description but is not reduced to cosine distance. Living Story motivates attention to unfinished, situated, embodied accounts but is not reduced to a word-count threshold. The empirical measures are partial operational bridges whose limitations must remain visible.

### **Manipulation Checks and Construct Validity**

Pilot cognitive interviews should test whether participants understand the ownership and agency items as intended. Adapted measures should be identified as adaptations, and their reliability and dimensionality should be evaluated rather than assumed. Manipulation checks should establish that fidelity and control differ in the intended direction without creating large unintended differences in readability, interface burden, or perceived text quality. Checks should be positioned so they do not teach participants the hypotheses before the primary outcomes are recorded.

### **Auxiliary Assumptions Specification**

Following Trafimow, an empirical prediction does not follow from a theory alone. It follows from a theory conjoined with auxiliary assumptions that connect non-observational concepts to observable procedures. Consequently, an empirical victory does not prove the entire theoretical architecture, and an empirical defeat does not identify the theory as the sole source of failure. This point is particularly important in human-AI storytelling research because constructs such as agency, answerability, redescription, ownership, and semantic drift cross philosophical, psychological, behavioral, and computational levels of analysis.

#### **A1. Measurement validity of perceived agency**

The study assumes that adapted agency and ownership items capture a psychologically meaningful boundary between the participant's own lived account and content introduced through AI collaboration. This assumption must remain narrower than the philosophical claim. A scale score cannot be equated with Bakhtin's non-alibi in Being; rather, it provides one operational indicator of experienced authorship and control. Failure of the measure to differentiate conditions could therefore indicate inadequate construct representation rather than absence of a psychological effect.

#### **A2. Behavioral equivalence of correction**

The study assumes that deleting, revising, flagging, or otherwise rejecting an AI-inserted problematic claim can serve as a behavioral indicator of moral veto. Yet correction can also be motivated by style, factual knowledge, readability, or task

compliance. The protocol therefore separates recognition of the problematic claim from the subsequent corrective act and records the participant's stated reason for intervention. This decomposition makes the behavioral inference more defensible.

### **A3. Semantic-drift interpretation**

The study assumes that computational differences between an initial and final narrative can identify meaningful textual movement. Cosine similarity, lexical overlap, proposition coding, causal-structure coding, and evaluative-stance coding do not measure the writer's phenomenology directly. They describe properties of texts. Their theoretical value arises when those properties are compared with the writer's own judgments of fidelity, ownership, and intended meaning.

### **A4. Interface equivalence**

The study assumes that differences attributed to AI collaboration are not artifacts of visual complexity, latency, typing burden, screen layout, or unequal opportunities to inspect and revise text. Interface burden must therefore be measured and minimized. The strongest confirmatory version of the study manipulates theoretically meaningful properties while holding the surrounding interaction environment as constant as practicable.

### **A5. Sampling and transport**

An English-language online sample can support causal inference about randomized conditions within the recruited population, but it cannot silently stand for all writers, cultures, organizations, languages, or storytelling traditions. Claims about transport beyond the sample require explicit argument and later replication. This limitation is

theoretically important because ownership, voice, answerability, and the acceptability of redescription may be culturally situated.

#### **A6. Computational reproducibility**

A frozen model name and temperature setting are not by themselves sufficient for reproducibility. The model artifact, tokenizer, prompt, inference parameters, software stack, seed behavior, quantization if any, and serving environment must be documented. Repeated identical-input trials should characterize residual variability before the apparatus is frozen for confirmatory use.

#### **A7. Narrative comparability**

If participants produce different personal stories across sessions, those stories may differ in emotional intensity, complexity, length, familiarity, and moral importance. A within-person contrast therefore depends on an assumption of sufficient task comparability. This was one reason later design work considered a cleaner between-participants factorial structure after a securely captured baseline source narrative.

#### **A8. Manipulation separability**

Narrative fidelity and editorial control must not become disguised proxies for general text quality or interface convenience. A low-fidelity revision can be grammatically excellent, and a high-fidelity revision can remain recognizably machine-edited. Likewise, constrained control should reduce editorial authority without simply making the interface irritating or difficult to use.

#### **A9. Misleading-claim calibration**

The ethically or factually problematic insertion must be consequential enough to require attention but not so extreme that every attentive participant detects it instantly. Pilot work should establish detectability, severity, and plausibility. The manipulation should also avoid unnecessary distress and should be fully disclosed during debriefing.

#### **A10. Demand characteristics**

Participants should not be told so much about agency loss, answerability, or semantic drift that they can infer the desired pattern and perform it. Manipulation checks should therefore be brief and indirect where possible, and the order of measures should be chosen to reduce hypothesis revelation.

#### **A11. Statistical-model adequacy**

The planned model assumes a particular structure for residual variation, participant heterogeneity, and repeated observations. Diagnostics and sensitivity analyses must determine whether that structure is adequate. The model is an inferential instrument, not the theory itself; a poor statistical representation is another auxiliary assumption that can fail.

#### **A12. Precision criterion**

Trafimow's A-priori Procedure requires the research team to decide what degree of precision is scientifically useful before observing confirmatory outcomes. The chosen margin is therefore a substantive commitment, not a number selected merely to justify an attainable sample. Sensitivity to alternative precision targets should be shown so collaborators and reviewers can see the consequences of the choice.

## **Statistical Analysis Plan**

### **Analysis Plan for the Revised Factorial Design**

For participants in the four AI cells, the primary confirmatory model can be written as  $Y_i = \beta_0 + \beta_1 F_i + \beta_2 C_i + \beta_3(F_i \times C_i) + \epsilon_i$ , where  $F$  denotes assigned narrative fidelity,  $C$  denotes assigned editorial control, and  $Y$  is the pre-specified primary outcome. The main-effect coefficients estimate the average causal effects of fidelity and control under random assignment, and the interaction estimates whether the effect of control depends on fidelity. The human-only benchmark is evaluated through pre-specified contrasts rather than coded as a fifth factorial level.

If the team retains repeated tasks, the model must expand to include participant random effects and appropriate task or order terms. That choice cannot be made after inspection of confirmatory outcomes. Likewise, covariate adjustment should be limited to variables named in advance and justified by precision or design considerations. Distributional diagnostics, residual checks, and sensitivity analyses should accompany interval estimates rather than functioning as post hoc routes to a preferred conclusion.

### **Secondary and Exploratory Analyses**

Secondary outcomes can include perceived agency, semantic fidelity, detection of the misleading claim, correction behavior where correction is possible, correction latency, endorsement, perceived responsibility, and computational measures of textual change. Exploratory analyses may examine prior AI experience, writing

confidence, age, education, language background, or narrative characteristics as moderators, but such analyses should be labeled exploratory unless the Stage 1 protocol specifies them as confirmatory. Qualitative comments can illuminate how participants describe the experience of having a personal story redescribed, but they should not be retrofitted as confirmatory tests of the quantitative hypotheses.

### **Primary Estimates**

In the revised factorial architecture, the leading primary estimates is the fidelity-by-control interaction on psychological ownership, with the two main effects specified alongside it. This ordering can be changed by collaborator agreement before submission, but it should not be changed after confirmatory outcomes are observed. The no-AI benchmark addresses a different estimate: the contrast between human-only writing and the AI-assisted conditions. Keeping the benchmark separate prevents a conceptually awkward five-level factor from obscuring the 2 x 2 causal structure.

### **Decision Rule for H5**

H5 should enter the confirmatory protocol only after the research team chooses a common detection opportunity. If all participants can flag an inaccurate insertion, detection rate and detection latency can be compared across conditions. Correction rate is confirmatory only if correction has a comparable behavioral meaning across the conditions being contrasted. Otherwise, correction remains a condition-specific secondary outcome. This decision rule is deliberately simple because it removes a source of ambiguity that a reviewer could otherwise identify immediately.

First, psychological ownership and perceived agency are measured immediately after the participant finishes the main revision stage, before any misleading-claim probe. Second, detection of the deliberately inserted source-inconsistent claim is assessed in a separate module in which every AI condition has the identical universal flagging mechanism, so that detection does not depend on a condition's editing rights. Detection rate and detection latency are therefore the exploratory indicators for H5; correction behavior, which the constrained-control conditions cannot perform in the same way, is retained only as a condition-specific secondary outcome and is not used as a cross-condition index of moral responsibility. Because the inserted claim can itself shift perceived fidelity, the probe is delivered after the primary ownership and agency measures and is analyzed separately from the fidelity manipulation. Note: this detection analysis is exploratory, not confirmatory — a behavioral index of scrutiny, not a confirmatory test or a measure of moral responsibility; a fully powered confirmatory test is planned as a follow-up study.

### **Why Auxiliary Assumptions Matter for This Registered Report**

The methodological significance for this study is direct. Human-AI co-creation requires a chain of translations: philosophical concepts become psychological constructs; constructs become scale items and behavioral events; stories become computational representations; model outputs become experimental stimuli; and those stimuli become causal contrasts. Every translation introduces assumptions. Stating them does not weaken the theory. It identifies where evidence can and cannot bear the weight of interpretation.

Accordingly, Stage 2 interpretation should proceed by examining the prediction together with the assumptions required to produce it. If a condition difference appears, the analysis should ask whether the manipulation behaved as intended, whether the measure represented the construct, whether the computational apparatus remained stable, and whether alternative interface explanations were controlled. If the prediction fails, the same questions apply before either the theory or the phenomenon is dismissed.

### **Sensitivity Analyses and Interpretation**

Support for a prediction will be interpreted as support for the specified empirical contrast under the stated auxiliary assumptions, not as proof that a philosophical theory has been verified. Rorty, Bakhtin, and Living Story supply theoretical reasons for asking the questions and interpreting the pattern; they are not reduced to a single coefficient.

Failure of a prediction will trigger examination of manipulation validity, measurement behavior, computational stability, interface equivalence, sampling, and model adequacy before a broader theoretical conclusion is drawn. This does not make the theory unfalsifiable. It makes explicit which inferential links must hold for the empirical result to bear on the theoretical claim.

The Stage 2 manuscript should report confirmatory analyses first and exactly as registered. Deviations should be documented and justified. Exploratory analyses can follow, clearly labeled. Pilot data used to construct manipulations or measures should not be pooled into the confirmatory sample unless the Stage 1 protocol explicitly permits that procedure.

The strongest contribution of the study is therefore methodological as well as substantive: it places a philosophically rich question about authorship and answerability into a design whose assumptions, apparatus, precision targets, and causal contrasts can be inspected before outcomes are known.

### **Study Timeline**

After ethics approval, the team will pilot stimulus fidelity, editorial-control comprehension, usability, and the misleading-claim detection task. The final stimuli, measures, exclusion rules, analysis code, and precision plan will be frozen before confirmatory recruitment. Data collection and Stage 2 analyses will follow the accepted protocol.

### **Protocol Materials: Consent, Baseline Narrative, and Debriefing**

Consent will describe a study of personal writing and machine-assisted revision, explain that participants may encounter text not originally written by them, and state that some revisions may intentionally depart from the source. Consent will not disclose the exact misleading insertion before the task because that would undermine the detection outcome. The ethics-approved debriefing will explain the insertion, identify the experimental rationale, and provide a clear opportunity to withdraw the participant's narrative where institutional policy requires it.

Baseline writing instruction: 'Describe a turning point in your life that you are comfortable discussing. Write what happened in your own words. You may describe what came before, what the moment felt like, what began to change, and what was still uncertain. You do not need to include names or private details, and you may

choose a different event at any time.’ This instruction is the common stimulus across all five arms; the platform records the original text before any treatment is delivered.

Post-task sequence: participants first answer ownership and agency questions without being shown a study hypothesis; then they judge fidelity to their intended account; then they respond to the standardized detection item; finally, they report perceived responsibility, writing confidence, AI familiarity, task engagement, and any discomfort. The order and item wording will be piloted for demand effects. The final Stage 1 repository will contain the exact survey instrument rather than only construct labels.

### **Trafimow A-Priori Procedure and Precision Calculations**

Trafimow’s A-priori Procedure (APP) begins with an estimation objective rather than a preferred probability of rejecting a null hypothesis. The investigator specifies the acceptable distance between a sample estimate and its population target and the probability of achieving that distance. In its familiar normal-theory mean form, with population standard deviation  $\sigma$ , target half-width  $\epsilon \sigma$ , and coverage  $c$ , the planning relationship is  $n = [z_{(1+c)/2} / \epsilon]^2$ . This expression must be modified for the variance of the actual contrast of interest. A precision claim for a simple mean cannot simply be transferred to a factorial interaction.

The original research-development manuscript considered a matched contrast and wrote the planning equation as  $n = 4 z^2 / \epsilon^2$ . At  $c = .95$  and  $\epsilon = .15$ , the substitution  $n = 4(1.96)^2 / (.15)^2$  yields approximately 683. That derivation is preserved because it shows how the precision target was originally selected. The operative

design in this manuscript, however, is a between-participant  $2 \times 2$  factorial experiment with a separate human-only benchmark, and its sample-size justification must target that design's contrasts rather than treating 683 as a universal sample size.

For a balanced factorial with  $n$  observations in each AI cell, common outcome variance  $\sigma^2$ , and the interaction contrast defined as the difference of the two simple control effects,  $\text{Var}(\text{interaction}) = 4 \sigma^2/n$ . The approximate 95% confidence-interval half-width is therefore  $1.96 \times 2 \sigma/\sqrt{n}$ . A desired interaction half-width of  $0.30 \sigma$  requires  $n \approx (3.92/0.30)^2 = 171$  participants per AI cell. The corresponding four-cell factorial sample is 684, before any allowance for exclusions or a separate no-AI benchmark. A stricter target of  $0.20 \sigma$  would require about 385 per AI cell. These expressions assume independent participants, balanced cells, and a common variance; pilot variance estimates and a simulation of the planned model will be used to test those assumptions.

Because the benchmark contrasts and binary detection outcome have different sampling variances, a single APP calculation cannot guarantee equally narrow intervals for all endpoints. The primary endpoint and interaction should govern the precision target, while the benchmark allocation should be justified separately. The registered stopping rule will specify the number of analyzable participants per arm, permitted replacement for technical failures, and a maximum recruitment cap. Recruitment will not stop or continue based on observed confirmatory effects.

## Measurement Instruments and Scoring

The primary outcome is state psychological ownership of the submitted narrative. Candidate items ask whether the final account feels like the participant's own story, whether it expresses what they meant, whether they would be comfortable claiming authorship, and whether consequential choices in the final text feel attributable to them. These are proposed study-specific items, not a pre-existing validated scale; their exact wording, response anchors, scoring rule, and reliability evidence must be established in pilot work before confirmatory collection. A mean composite will be used only if the preregistered psychometric checks support a sufficiently coherent dimension.

## **Empirical Inputs: Vujade™ Instrument and Study Application**

### **Instrument: Vujade™**

Vujade™ is a behavioral assessment instrument that characterizes communication preferences at the individual and team levels, with the aim of supporting interpersonal connection, collaboration, and trust. It measures three constructs, and administration typically requires 15 to 20 minutes.

| <b>Construct</b> | <b>What it captures</b>                                                | <b>Factors</b>                                     |
|------------------|------------------------------------------------------------------------|----------------------------------------------------|
| Sensory          | Preferences in how information is received                             | Kinesthetic, visual, auditory                      |
| Temperament      | Preferences regarding sources of energy and modes of social engagement | Introversion orientation, extroversion orientation |
| Cognitive        | How information is organized and processed                             | Process orientation, outcome orientation           |

Vujade™ generates relational comparisons between individuals by expressing differences in standard deviation units relative to a large normative dataset. Because this comparative methodology is instrument-agnostic, other instruments and datasets can be implemented within the same framework. The methodology was granted a U.S. utility patent in 2026 (U.S. Patent No. 12,731,089).

### **Application in the Present Study**

Vujade™ will be administered within a staggered-administration design crossed with each of the four experimental conditions examining AI use.<sup>[1]</sup> Within each condition, participants will be assigned to one of three subgroups defined by the timing of the first administration.

| <b>Experimental Groups</b> | <b>First administration</b>                                                  |
|----------------------------|------------------------------------------------------------------------------|
| Experimental Group A       | Before the experimental stage when subjects use OLMo 2.                      |
| Experimental Group B       | After the first experimental stage when subjects use OLMo 2.                 |
| Experimental Group C       | Both before control group does a writing assignment and does not use OLMo 2. |

All subgroups will complete a further administration three months after the conclusion of the study. Vujade™ data will be analyzed in relation to study outcomes using conventional statistical methods, the comparative methods implemented in the

Cabo™ platform, and the statistical and experimental approach developed by David Trafimow et al. (IBID).

**Proposed Psychological Ownership items (primary outcome; to be validated in pilot phase).** And pilot study participants will use OLMo 2 before, during, or after their writing a short essay.

All ownership items use a 1–7 agreement scale (1 = strongly disagree, 7 = strongly agree) and refer to the final submitted narrative:

PO1, “I feel that this final story is mine”;

PO2, “I would be comfortable putting my name to this final story as its author”;

PO3, “This final story reflects who I am”;

PO4, “I feel a strong sense of personal ownership of this final story.”

To keep ownership empirically separable from adjacent constructs, perceived fidelity and perceived agency are measured with their own items rather than folded into the ownership score, we call, fidelity construct:

F01 “The final story preserves the meaning I intended” and

F02 “The final story accurately conveys what actually happened”; agency, e.g.,

F03 “I was in control of the changes made to the story” and “The decisions about the final wording were mine.”

Ownership, fidelity, and agency are scored as separate composites; the pilot will test their internal consistency and whether the three constructs are empirically

distinguishable (e.g., by factor analysis and with Trafimow's statistical methodology) before any composite is used confirmatorily.

To avoid overlap with endorsement and reduce validation burden, the measure will adapt an established psychological-ownership instrument, keeping ownership items distinct from endorsement; reliability, dimensionality, and discriminant validity (against agency, fidelity, and endorsement) will be established in an adequately powered pilot, with a dedicated psychometric-validation phase scheduled if pilot validation is insufficient. Van Dyne & Pierce (2004) has measured exactly the same issue.

Perceived agency will be measured separately from ownership. Tapal et al.'s (2017) Sense of Agency Scale assesses generalized beliefs about control over mind, body, and environment; it is not automatically a valid state measure of editorial control in a particular writing task. The protocol will therefore distinguish any established trait measure from newly developed task-specific agency items. Trait agency may be used as a covariate or moderator if prespecified; it will not be relabeled as an experimental outcome without evidence of responsiveness and construct validity.

As a manipulation check, perceived editorial control — how much authority participants felt they had over the text — is measured separately from perceived agency, to confirm that the assigned-control manipulation shifted felt control in the intended direction.

Detection is scored as a binary indication that the participant recognized the misleading insertion through the common flagging mechanism or the preregistered recognition question. Behavioral rejection is scored separately when the interface provides a comparable opportunity to veto. Detection latency is measured from first presentation of the relevant passage to the first valid flag, with participants who never flag treated as right-censored at the end of the inspection period. A participant who fails to flag is not excluded from analysis on that basis, because failure to detect is itself part of the outcome distribution.

Semantic movement will be represented by a set of prespecified features: proposition-change count, causal-attribution shift, evaluative-stance shift, source-final embedding similarity, and participant-rated fidelity. The primary ownership outcome is not replaced by an opaque aggregate of these features. Inter-rater agreement, coding uncertainty, and sensitivity to alternative embedding specifications will be reported so readers can judge how strongly each feature supports an interpretation of redescription.

### **Statistical Models and Planned Contrasts**

For the four AI arms, define  $F = +1/2$  for higher fidelity and  $-1/2$  for lower fidelity; define  $C = +1/2$  for higher control and  $-1/2$  for constrained control. The primary model is  $Y_i = \beta_0 + \beta_F F_i + \beta_C C_i + \beta_{FC} F_i C_i + \varepsilon_i$ . Under balanced random assignment,  $\beta_F$  and  $\beta_C$  represent marginal contrasts and  $\beta_{FC}$  represents the difference-in-differences interaction. Estimates will be accompanied by 95% confidence intervals and standardized contrasts. The preregistration will identify the

primary hypothesis family and any multiplicity adjustment rather than selecting it after results are inspected.

The human-only benchmark will be compared with the pooled AI arms and with each AI cell using planned contrasts in a five-arm linear model. These estimates address the presence of machine revision as a whole and should not be interpreted as estimates of the fidelity or control factors. Exploratory moderation by prior AI use, writing confidence, narrative length, and task engagement will be distinguished from the confirmatory contrasts and reported with interval estimates.

Binary detection and veto outcomes will be analyzed using logistic regression with the same factorial predictors, with the human-only group omitted from an AI-insertion outcome unless a comparable non-AI control insertion is deliberately constructed. Latency to flag will be modeled using a prespecified censored-time approach, with proportional-hazards diagnostics if a Cox model is chosen. Secondary count outcomes, such as number of edits, will use appropriate count models if the distribution is over dispersed or zero-inflated. The analytic family for each endpoint will be frozen before recruitment.

Missing primary ownership responses will not be silently imputed as favorable or unfavorable. The report will present missingness by randomized arm, pre-specify complete-case and sensitivity estimators, and distinguish participant withdrawal from technical loss. Manipulation-check failures will be described and may inform secondary analyses, but exclusion from the intention-to-treat primary comparison will not be based on a post-treatment subjective judgment unless that rule is explicitly justified and registered.

### **Online Recruitment Integrity and Human Respondent Verification**

Because the study is conducted online, procedures will address both scripted bots and undisclosed AI-generated participant responses, consistent with the journal's online-study reporting policy. Eligibility and consent checks will be supplemented by platform-level anti-bot controls, timing diagnostics, consistency checks across the source narrative and later recognition items, and logs of copy-paste and editor events. The instructions will prohibit use of external generative-AI tools during the source-writing task while explicitly permitting only the study-provided AI in assigned conditions.

Detection rules will not rely on generic AI-text detectors as definitive evidence of non-human authorship. The preregistration will distinguish technical failures, duplicate entries, impossible response patterns, and probabilistic indicators of external assistance. Any automated exclusion rule will be validated on pilot data and applied without access to confirmatory outcomes. Because disabling all paste functionality could itself distort the writing task and remove useful behavioral evidence, interface decisions will balance prevention with measurement.

### **Treatment Fidelity, Blinding, and Quality Assurance**

Before confirmatory recruitment, the research team will conduct a pilot that evaluates the strength and independence of the two manipulations. A manipulation succeeds only if independent raters and participant checks identify the intended difference in narrative fidelity or editorial authority without substantial confounding by grammar, readability, length, or interface usability. The pilot will be separated from the

confirmatory sample, and its results will inform the final frozen prompts, interface, and scoring plan.

Coders of narrative fidelity and semantic movement will be blinded to participants' ownership responses and randomized condition labels where feasible. Analysts will implement the preregistered models on synthetic data to confirm that variable definitions, contrasts, and output tables function before outcome data are available. Automated checks will verify arm counts, timestamp ordering, preservation of the source narrative, complete exposure logs, and consistency between displayed stimuli and archived treatment records.

### **Experimental Materials: Condition-Specific Participant Experience**

Human-only benchmark: after submitting the source account, the participant is invited to revise it using the same text editor but receives no machine-authored revision. The participant can inspect and submit the final narrative, after which the ownership and agency measures appear. The benchmark is intended to estimate the psychological experience of ordinary self-revision under the study's interface, not a zero-effort or zero-editing condition.

Higher fidelity/higher control: the participant receives a revision intended to preserve central propositions, causal attributions, and evaluative stance, with side-by-side inspection and full accept/reject/restore authority. Higher fidelity/constrained control: the participant receives a revision meeting the same fidelity criterion but has fewer ordinary editing options; the common flagging channel remains available.

Lower fidelity/higher control: the participant receives a plausible revision with a

consequential meaning shift and can inspect, reject, and restore. Lower fidelity/constrained control: the participant receives a similarly shifted revision with reduced ordinary editing authority and the same universal flagging channel.

The contrast between lower and higher fidelity is not a contrast between harmful and harmless content. The misleading-claim probe is a separate, calibrated experimental event, inserted under the same exposure rules across AI conditions if H5 remains confirmatory. Its location, semantic severity, and detection window must be identical in design logic across arms. The participant's source narrative will be used to ensure the claim is recognizably inconsistent with their intended account without requiring external adjudication of biographical truth.

**Operational editorial-control matrix.** Each condition is defined by the exact set of actions the interface permits, so the manipulation isolates genuine editorial authority rather than interface friction. Higher-control conditions (both fidelity levels) permit: side-by-side inspection of source and revision; passage-level acceptance; passage-level rejection; restoration of the original source text; free passage-level editing; and an explicit final review before submission.

Constrained-control conditions (both fidelity levels) permit: full inspection of source and revision and a single standardized accept-or-submit action, but not passage-level editing, selective rejection, or restoration. The human-only benchmark permits ordinary self-revision of the participant's own text with no machine revision offered. Across every AI condition two rights are held constant and identical: a common mechanism to flag any statement as false or objectionable, and the right to decline

participation or withhold the personal narrative at any time. Server logs record which actions were displayed and taken in each condition. The action set per cell is as specified above; the constraint reduces editorial authority only, and pilot usability checks confirm it adds no latency, visual clutter, or usability penalty (assumption A8).

## References

- Asghari, M. (2015). Has Richard Rorty a moral philosophy? *Journal of Philosophical Investigations*, 9(17), 53–74. <https://philpapers.org/>
- Bakhtin, M. M. (1993). *Toward a philosophy of the act*. University of Texas Press.
- Boje, D. M. (2008). *Storytelling organizations*. SAGE Publications.
- Boje, D. M. (2015). *Storytelling organizational practices: Managing in the quantum age*. Routledge.
- Rorty, R. (1980). *Philosophy and the mirror of nature*. Princeton University Press.
- Rorty, R. (1989). *Contingency, irony, and solidarity*. Cambridge University Press.
- Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., & Welsch, R. (2024). The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction* (arXiv:2303.03283).
- Joshi, N., & Vogel, D. (2024). Writing with AI lowers psychological ownership, but longer prompts can help (arXiv:2404.03108).
- Abdurahman, S., Ziabari, A. S., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, 8(2). <https://doi.org/10.1177/25152459251325174>

- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), eadn5290. <https://doi.org/10.1126/sciadv.adn5290>
- Oktiyalten, R., & Abdurahman, A. (2025). Integration of Artificial Intelligence in Problem-based Learning Models: Its Impact on Explanatory Writing Skills and the Interaction with Reading Interest. *Jurnal Edutech Undiksha*, *13*(2), 220-232.
- Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton, NJ: Princeton University Press.
- St Quinton, T., & Trafimow, D. (2022). The unappreciated relevance of auxiliary assumptions for evaluating theory-based interventions in health psychology. *Theory & Psychology*, *32*(6), 915–930. <https://doi.org/10.1177/09593543221113263>
- St Quinton, T., & Trafimow, D. (2025). The need to consider auxiliary assumptions in preregistration practices. *Frontiers in Psychology*, *16*, Article 1633990. <https://doi.org/10.3389/fpsyg.2025.1633990>
- Tapal, A., Oren, E., Dar, R., & Eitam, B. (2017). The Sense of Agency Scale: A measure of consciously perceived control over one's mind, body, and the immediate environment. *Frontiers in Psychology*, *8*, Article 1552. <https://doi.org/10.3389/fpsyg.2017.01552>
- Trafimow, D., St Quinton, T., & Weber, M. (2024). Crediting or blaming theories or auxiliary assumptions for empirical victories or defeats. *Theory & Psychology*, *34*(4), 508–530. <https://doi.org/10.1177/09593543241253548>
- Trafimow, D., Wang, T., & Wang, C. (2019). From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement*, *79*(1). <https://doi.org/10.1177/0013164418764801>

Trafimow, D. (2022). Generalizing across auxiliary, statistical, and inferential assumptions. *Journal for the Theory of Social Behaviour*, 52(1), 37–48.

<https://doi.org/10.1111/jtsb.12296>

Trafimow, D. (2023). A new way to think about internal and external validity. *Perspectives on Psychological Science*, 18(5), 1028–1046.

<https://doi.org/10.1177/17456916221136117>

Trafimow, D., & Fiedler, K. (2026). An exploration of physics envy with implications for desiderata of psychology theories. *American Psychologist*, 81(1), 94–108.

<https://doi.org/10.1037/amp0001416>

Van Dyne, L., & Pierce, J. L. (2004). Psychological ownership and feelings of possession: Three field studies predicting employee attitudes and organizational citizenship behavior. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 25(4), 439–459.

<sup>[1]</sup> Vujade™ is opposite of the psychology notion, Déjà Vu. See <https://vujade.ai>