Not Your Average Clever Hans: Latent Representation of Psychological Constructs in Large Language Models

Jack Chen[1], Zelin Zhou[1], Chunlin An[2], and Lexi Chen[3]

[1]Columbia Business School, Columbia University

[2]Columbia University

[3]The Wharton School, University of Pennsylvania

This version: November 7, 2026

Copyright © 2026 Jack Chen, Zelin Zhou, Chunlin An, and Lexi Chen. All rights reserved. This working paper is distributed for the purposes of comment and discussion only. It may not be reproduced, distributed, or used, in whole or in part, without the prior written permission of the authors.

Author Note

Jack Chen https://orcid.org/0009-0001-3851-1151

Zelin Zhou https://orcid.org/0009-0003-3593-7995

Author contributions: J. Chen: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, project administration, writing – original draft, writing – review & editing. Z. Zhou: investigation, writing – review & editing. C. An: formal analysis, validation, writing – review & editing. L. Chen: investigation, writing – review & editing.

Correspondence concerning this article should be addressed to Jack Chen, Columbia Business School, Columbia University, New York, NY 10027, United States. Email: cchen27@gsb.columbia.edu

Abstract

Psychologists increasingly use large language models (LLMs) as experimental participants, but an effect in such a silicon sample supports a process theory only if the LLM's response arises causally through representing the intermediate psychological states, or constructs, before it is produced. We examine whether LLMs have stable representations of these psychological constructs in their latent neural layers. Using a multitrait-multimethod approach to examine 66 psychological constructs (e.g., psychological ownership, moral agency, and perceived goal progress) in 16 open-weight LLMs, we found that the LLMs' inner representations met construct-validity standards, were shared across model developers and sizes, and were grouped by appraisal processes. We also found that distinct manipulations of a construct activated the same representation, and that whether this activation affected the LLMs' responses depended on its input proximity to response production. We thus provide a method that enables researchers to study psychological mechanisms using silicon samples.

Keywords: large language models, silicon samples, construct validity, multitrait-multimethod matrix, nomological network, mediation

Not Your Average Clever Hans: Latent Representation of Psychological Constructs in Large Language Models

More than a century ago, a horse named Clever Hans seemed to solve arithmetic problems, tapping out the solutions with its hoof, until a psychologist showed that it was responding to slight, unintended movements of its questioners (Pfungst, 1911).

Today, researchers increasingly use large language models (LLMs) as participants in psychology studies, particularly in survey-based experiments (Argyle et al., 2023; Horton, 2023; Karetnikov et al., 2026). Many studies with such LLM participants, often called silicon samples (Argyle et al., 2023; Lin, 2025), have found that, in general, LLMs reproduce the effects of seminal published psychology experiments (Aher et al., 2023; Cui et al., 2025) and predict the results of experiments that had not been published (Ashokkumar et al., 2026).

In light of the increasing ability of silicon samples to replicate the responses of human participants, psychologists have even begun to use LLMs to test psychological theories (Ashokkumar et al., 2026; Binz et al., 2025), despite methodological challenges such as the researcher degrees of freedom introduced by the choice of LLM and prompt (Cummins, 2026; Simmons et al., 2011), the flattening of response variance relative to human participants (P. S. Park et al., 2024; Wang et al., 2025), and other challenges to causal inference (Lin, 2025).

The rise of silicon samples, however, casts an epistemological cloud over psychological research: are LLMs today's Clever Hans, or do they engage psychological processes as human participants do (Figure 1; Mitchell & Krakauer, 2023)? That is, for silicon samples to support any process theory, an LLM's response, whether a choice, an evaluation, or another behavioral response, must be generated from the presumed sequence of psychological states (i.e., mediators), as a process theory prescribes. For example, human participants assign greater responsibility to a grown man who pushes a tray of glasses off a table, causing a shard to cut a baby, than to a baby who does the same to a grown man (Gray & Wegner, 2009; Klein et al., 2018). According to the theory of moral typecasting, this difference reflects participants' perception of the man as a moral agent (Gray & Wegner, 2009). A silicon sample that shows the same effect, therefore, provides evidence for the theory only if it also represents the man as a moral agent and if this representation leads it to assign greater responsibility to him. With human participants, researchers typically measure or manipulate this presumed intermediate

psychological state, the perception of moral agency, and observe its causal consequences (Pirlott & MacKinnon, 2016; Spencer et al., 2005). Such tests, however, cannot easily be done with LLMs when only their language outputs are observed, because any such output may be apsychological and reflect only statistical regularities learned from their training data (McCoy et al., 2024).

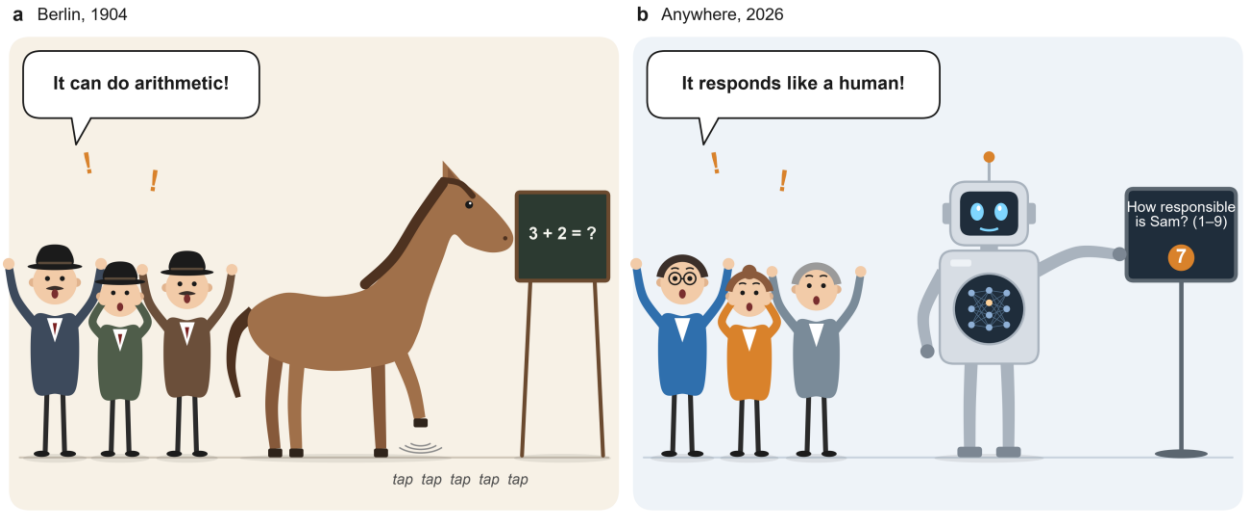


Fig. 1.

Clever Hans and a silicon sample. The illustration shows (a) psychologists in Berlin in 1904 watching Clever Hans tap out the solutions to arithmetic problems, taps that were later shown to be cued by slight, unintended movements of his questioners (Pfungst, 1911), and (b) psychologists today watching a large language model respond as a silicon sample to the outcome measure of a psychology experiment.

In this research, we propose that to address this epistemological uncertainty inherent in silicon-sample responses, one must look "inside" an LLM: at what it represents in the layers of its neural network when it stands in for human participants. This way, we can examine the latent representations of psychological constructs in the internal states of LLMs, much as researchers examine the brain activity of human participants with functional magnetic resonance imaging (fMRI).

Like the hidden units of connectionist models of cognition (Rumelhart et al., 1986), the internal states of an LLM, often called hidden states, are the vectors of numbers that the hidden layers of its neural network compute for each word of a text, given the words that precede it, and

from which the LLM predicts the next word (Elhage et al., 2021; Vaswani et al., 2017). Recent research has found that the internal states of LLMs represent the beliefs of the characters in a story (Zhu et al., 2024), the emotions that a situation evokes and the appraisals underlying them (Sofroniew et al., 2026; Tak et al., 2025), and the character traits expressed in an LLM's own text (R. Chen et al., 2025), and that changing these internal states changes an LLM's responses (Rimsky et al., 2024; Zou et al., 2023).

Existing studies, however, have typically derived each representation from a single set of texts in one or a few LLMs. This method, when applied to the representations of psychological constructs, would not meet the standards of construct validity that psychologists apply to measurements with human participants (Cronbach & Meehl, 1955; Messick, 1995). Specifically, measures of the same construct obtained by different methods should converge (convergent validity), and measures of different constructs should be distinguishable (discriminant validity; Campbell & Fiske, 1959; Joel et al., 2025). In addition, the content of each measure should represent its construct (content validity), and each measure should show the expected relationships with independently measured criteria (criterion validity).

In this research, we developed a method to measure representations of psychological constructs in the internal states of LLMs and validated these representations against the four standards of construct validity. To obtain representations that reflect the psychological content of each construct rather than construct-irrelevant features of the texts that manipulate it (content validity), we manipulated each construct in many pairs of short narratives that vary in scenario and wording, as researchers manipulate psychological states in experiments with human participants. The two narratives of each pair describe the same scenario, one conveying a high and the other a low level of the construct. Because the training data of LLMs may contain the scenarios of published experiments, all narratives were composed to differ from these scenarios, so that the representations would not reflect what LLMs had learned about the experiments. We took the average difference between the LLM's internal states for the high and the low narratives as the LLM's representation of the construct (K. Park et al., 2024; Zou et al., 2023), much as fMRI studies identify the brain activity related to a mental process by comparing two experimental conditions. To test convergent and discriminant validity, we asked several LLMs from different developers to compose the narratives, so that each of these LLMs served as a

different method of manipulating the constructs, as in a multitrait-multimethod matrix (Campbell & Fiske, 1959).

To test whether the construct representations that we identified in fact causally lead to the LLMs' responses, as psychologists theorize, we presented 60 influential experiments in social and consumer psychology to LLMs as silicon samples, each in several wordings and scenarios that preserve its manipulation. Specifically, we first examined whether the published manipulations activate the representation of the presumed mediator. We then investigated whether directly activating this representation, without changing the experiment's scenario description, moves the LLMs' responses in the theorized direction, and whether removing it reduces the manipulation's effect on the responses (Bullock et al., 2010; Pirlott & MacKinnon, 2016; Spencer et al., 2005).

In summary, we found stable latent representations of the 66 psychological constructs we tested. Importantly, these representations met the four standards of construct validity and were shared across LLMs from different developers. We also found that across 60 experiments presented to LLMs as silicon samples, manipulations from the published studies activated the representation of their presumed mediator. By directly manipulating the strength of these representations, we showed that they played a causal role in leading the LLMs to respond in the theorized direction, and that they did so more strongly the greater their input proximity to response production, that is, the closer to the response the point of the LLMs' input at which they were activated.

This research makes a methodological contribution to the use of LLMs as research participants. It provides validated measures of psychological constructs in the internal states of LLMs. When psychologists use silicon samples to test a process theory, they should use these measures to test whether the manipulation activates the mediator's representation and whether this representation causes the response, as they test the manipulation and the mediator in experiments with human participants. Next, we test whether the representations meet the standards of construct validity (Studies 1 and 2) and whether they cause the LLMs to respond as process theories specify (Studies 3 to 5).

Research Transparency Statement

General disclosures. *Conflicts of interest:* We declare no conflicts of interest. *Funding:* This research was supported by Columbia Business School. The funder had no role in the design, implementation, analysis, reporting, or interpretation of the research. *Artificial intelligence:* LLMs are the object of this research and also composed and rated the study materials, as described in the Method. Two LLMs coded how directly each experiment's outcome measure refers to the proposed mediator (Supplemental Note 4). LLMs (Claude Opus 5.5, Kimi K3, Grok 4.7, GLM 5.3, and Codex) were also used to write and run the analysis code, create the figures, and draft, review, and edit the manuscript. We reviewed the code and checked every reported result against its output. No other AI-assisted technologies were used in the research or preparation of this article. *Ethics:* This research did not require ethics approval because it involved no human participants. All results from human participants used here came from published experiments and their published replications.

Study disclosures (Studies 1 to 5). *Preregistration:* Because this research involved no human participants, we did not preregister it on a standard platform such as AsPredicted. Before measuring the internal states of LLMs, however, we specified the exclusion rules, the layers, and the primary tests of Studies 1 to 4 in a dated analysis plan, which is openly available on GitHub with the analysis scripts; Supplemental Note 1 lists the analyses specified afterward. *Materials:* All study materials, including the instructions to the writers and raters, the narratives, the experimental materials, and the agent profiles, are publicly available on Zenodo (https://doi.org/10.5281/zenodo.23771693). *Data:* All data, including the ratings of the narratives, the read-out word lists and the raters' choices, the LLMs' judgments and responses, and the representations, activations, and test results for each LLM, are publicly available in the same Zenodo record. Model internal states at every word fragment, which would amount to terabytes, were not stored; they can be recomputed with the code from the 16 openly available LLMs. *Analysis scripts:* All analysis scripts are publicly available on GitHub (https://github.com/schema-brand/latent-psychological-constructs) and archived in the same Zenodo record.

Method

Measuring Construct Representations in Large Language Models

We measured the representations of the constructs that the theories of 60 influential experiments in social and consumer psychology propose as mediators. Of these experiments, 22 were published in the *Journal of Consumer Research*, the *Journal of Marketing Research*, or the *Journal of Consumer Psychology*; seven in the *Journal of Personality and Social Psychology*; six in this journal (Bauer et al., 2012; Carter et al., 2011; Hart & Albarracín, 2011; Jostmann et al., 2009; Rottenstreich & Hsee, 2001; Schultz et al., 2007); four in *Science*; and 21 in other journals and edited volumes (Supplemental Note 1). To test whether our measurements distinguish the mediators from alternative explanations, we also measured the representations of eight rival constructs that the original articles had ruled out and of three general evaluations, such as pleasantness. Together with the 55 mediators of the 60 experiments, these make up the 66 constructs (Table 1).

Because the LLMs' training data may contain the published experiments, a representation estimated from the experiments' own texts could reflect what an LLM has learned about these experiments rather than the construct. We therefore asked six LLMs from different developers, which we call writers, to compose new pairs of narratives for each construct in a range of domains (Figure 2a; Supplemental Note 1). The two narratives of a pair describe the same scenario. The high narrative conveys the state at the highest point (7) of the seven-point rating scale that defines the construct (Table S1), whereas the low narrative conveys either the opposite state at the lowest point (1) or the absence of the state at the midpoint (4). To ensure that the difference between the two narratives reflects the construct rather than construct-irrelevant features of the texts, we instructed the writers to make the narratives nearly equal in length and different only in what they convey about the construct, and not to name the construct. We also instructed them to place the narratives outside the scenarios of the published experiments and not to use these experiments' manipulations (Table S9).

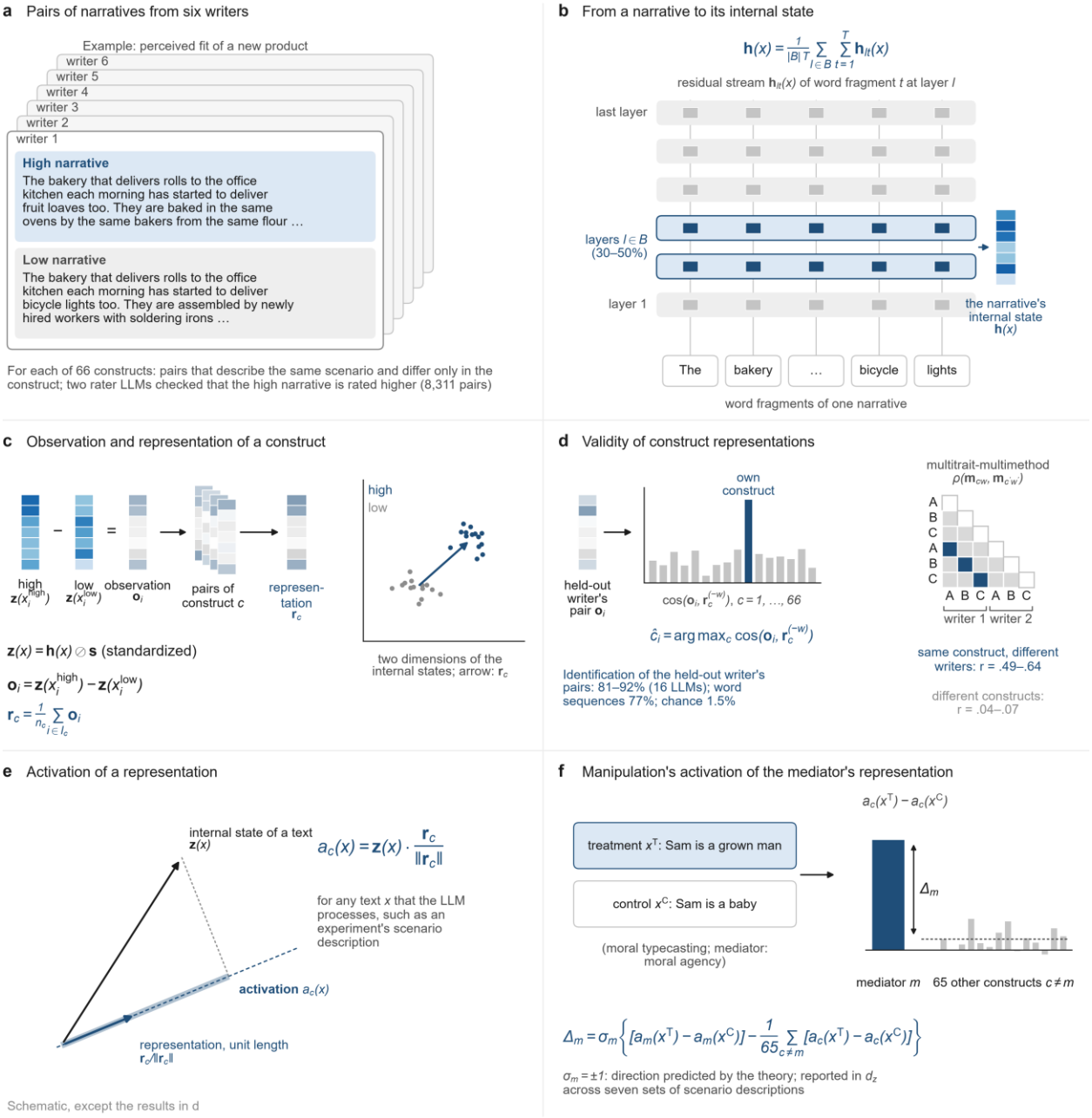


Fig. 2.

Measurement of construct representations in large language models (LLMs). In (a), six writer LLMs composed pairs of narratives, a high and a low narrative for each of 66 constructs (example: perceived fit of a new product; texts shortened). In (b), the residual stream $\mathbf{h}_{lt}(x)$ of each word fragment t at each layer l, averaged over the word fragments of a narrative and over the layers B 30% to 50% of the way through the LLM, is the narrative's internal state $\mathbf{h}(x)$. In (c), the standardized difference between the high and the low narrative of pair i is one observation $\mathbf{o}_i$,

and the mean of the observations of construct c is its representation $\mathbf{r}_c$. In (d), the construct of each of the held-out writer's pairs is identified among the 66 constructs, and multitrait-multimethod correlations are computed between the writer means $\mathbf{m}_{cw}$; the values are ranges across the 16 LLMs. In (e), the activation $a_c(x)$ is the projection of the standardized internal state $\mathbf{z}(x)$ onto the representation scaled to unit length. In (f), Δ_m is the manipulation's activation of the mediator's representation. All panels except the values in (d) are schematic.

Table 1

The 66 Constructs and the Experiments and Original Articles from Which They Come

Construct	Experiment (original article)
Categorization	
Anchor	Anchoring (Jacowitz & Kahneman, 1995); incidental anchors (Critcher & Gilovich, 2008)
Anthropomorphism	Anthropomorphized products (Aggarwal & McGill, 2007)
Associations with Republicanism	Flag priming (Carter et al., 2011)
Capability	Feature fatigue (Thompson et al., 2005)
Detail of imagined actions	Verb aspect and intent (Hart & Albarracín, 2011)
Evaluability	Less is better (Hsee, 1998)
Perceived fit	Brand extension (Aaker & Keller, 1990)
Salience of heat	Heat priming (Zaval et al., 2014)
Severity of forbidding†	Forbid vs allow (Rugg, 1941)
Small ongoing expenses (petty cash)	Pennies a day (Gourville, 1998)
Weight of the subject's features†	Similarity asymmetry (Tversky & Gati, 1978)
Usability	Feature fatigue (Thompson et al., 2005)
Social inference	
Concern about misidentification	Identity divergence (Berger & Heath, 2008)
Descriptive norm	Normative message (Schultz et al., 2007)
Desire for affiliation	Social exclusion (Mead et al., 2011)
Frame of reference	Response-scale range (Schwarz et al., 1985)
Moral agency	Moral typecasting (Gray & Wegner, 2009)
Others seen as partners	Consumer label and trust (Bauer et al., 2012)
Perceived autonomy	Nonconforming dress (Bellezza et al., 2014)
Prestige	Quote attribution (Lorge & Curtiss, 1936)
Reciprocity	Door in the face (Cialdini et al., 1975); norm of reciprocity (Hyman & Sheatsley, 1950)
Status motives	Going green (Griskevicius et al., 2010)
Vertical location as power	Vertical position and power (Giessner & Schubert, 2007)
Environmental concern	Going green (Griskevicius et al., 2010)
Need for uniqueness	Identity divergence (Berger & Heath, 2008)

Construct	Experiment (original article)
Self-relevance	
Future self-continuity	Future self and saving (Hershfield et al., 2011)
Identity as a resource	Identity threat (White & Argo, 2009)
Identity salience	Identity salience (Reed, 2004)
Personal connection	Time versus money (Mogilner & Aaker, 2009)
Psychological ownership	Emotional attachment (Shu & Peck, 2011); mere touch (Peck & Shu, 2009); possession and valuation (Reb & Connolly, 2007)
Self-perception as helpful	Foot in the door (Freedman & Fraser, 1966); labeling (Burger & Caldwell, 2003)
Self-view confidence	Shaken self-view (Gao et al., 2009)
Successful completion	IKEA effect (Norton et al., 2012)
Worldview defense	Mortality salience (Greenberg et al., 1994)
Identity importance	Identity threat (White & Argo, 2009)
Affective response	
Accessibility of the negative outcome	Tempting fate (Risen & Gilovich, 2008)
Affect richness	Affect-rich prizes (Rottenstreich & Hsee, 2001)
Analytic thinking	Analytic thinking and belief (Gervais & Norenzayan, 2012)
Emotional response to harm†	Personal force (Greene et al., 2009)
Goal fluency	Goal fluency (Labroo & Lee, 2006)
Interpersonal warmth	Physical warmth (Williams & Bargh, 2008)
Physiological arousal	Emotion and sharing (Berger & Milkman, 2012)
Reliance on affect	Affect reliance (Pham & Avnet, 2004)
Reliance on substance	Affect reliance (Pham & Avnet, 2004)
Goal relevance	
Desire for power	Powerlessness (Rucker & Galinsky, 2008)
Focus on the forgone	Buying and selling prices (Carmon & Ariely, 2000)
Gain frame†	Disease framing (Tversky & Kahneman, 1981)
Importance	Weight and importance (Jostmann et al., 2009)
Loss aversion	Endowment effect (Kahneman et al., 1990)
Overhead aversion	Overhead aversion (Gneezy et al., 2014)
Perceived goal progress	Goal progress (Fishbach & Dhar, 2005)
Rehearsal of past payments	Payment transparency (Soman, 2001)
Weight of positive features	Choosing vs rejecting (Shafir, 1993)
Causal inference	
Blame†	Side effects and intent (Knobe, 2003)
Earned right to indulge	Effort and indulgence (Kivetz & Simonson, 2002)
Expectation that actions produce outcomes†	Structure and goal pursuit (Kay et al., 2014)
Inferred importance	Decision difficulty (Sela & Berger, 2012)
Inferred length of the sequence	Retrospective gambler's fallacy (Oppenheimer & Monin, 2009)
Reference price	Beer price (Thaler, 1985)

Construct	Experiment (original article)
System justification	Currency priming (Caruso et al., 2013)
Uncertainty	Decision difficulty (Sela & Berger, 2012)
Comparison of alternatives	
Affective reaction	Emotional attachment (Shu & Peck, 2011)
Appeal of the setting	Foot in the door (Freedman & Fraser, 1966); mere touch (Peck & Shu, 2009)
Favorable evaluation	Door in the face (Cialdini et al., 1975); labeling (Burger & Caldwell, 2003); mere touch (Peck & Shu, 2009)
Perceptual contrast	Door in the face (Cialdini et al., 1975)
Pleasantness	Mere touch (Peck & Shu, 2009)

Note. Constructs in roman type are the mediators that the original articles proposed; several experiments share a mediator, so that 55 mediators serve the 60 experiments. Constructs in italics are the eight rival constructs that the original articles ruled out and the three general evaluations, each listed with the experiments in which it was measured. The names are those of the original articles; where an article names no mediator (†), the name is taken from the article's explanation of the effect. The constructs are grouped by the process in which they arise (Study 2), and comparison of alternatives is the comparison of the options in a choice. The definitions of all constructs are in Table S1, and example pairs of narratives are in Table S9. See the Supplemental Material available online for full details of the original articles cited only in this table.

To check whether each pair of narratives manipulated its intended construct, we asked two LLMs from developers other than the writer's, which we call raters, to rate every narrative on the construct (Supplemental Note 1). We retained a pair of narratives only if the raters rated its high narrative higher on the construct than its low narrative and it passed further checks of writing quality (Supplemental Note 1). In total, we retained 8,311 pairs of narratives, an average of 126 per construct (Tables S1 and S9).

Because closed-weight LLMs such as GPT do not make their internal states accessible, we used open-weight LLMs from five developers (Table S2). We conducted all analyses with three main LLMs from different developers, each with about 30 billion parameters (Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B). To test whether the results generalize across developers and model sizes, we estimated the representations and tested their construct validity in 13 additional LLMs, including five sizes of Qwen3.5, from 0.8 to 27 billion parameters, and six other LLMs. To test whether the representations arise from learning to predict text, the 13 LLMs also included two LLMs before instruction tuning, the last stage of training, in which an LLM learns to respond to requests. Because an experiment requires an LLM to respond to a

question, we measured how strongly the published manipulations activated the mediators' representations in the 11 of these 13 LLMs that had completed instruction tuning.

We estimated each construct's representation with the difference-in-means approach (Marks & Tegmark, 2024; Rimsky et al., 2024; Zou et al., 2023). An LLM divides a text into word fragments and computes, in a sequence of stages called layers, a vector of numbers for each word fragment, the residual stream that each layer passes to the next one (Elhage et al., 2021). We write $\mathbf{h}_{lt}(x)$ for this vector of text x at layer l and word fragment t. We presented each narrative to the LLM and took as its internal state the mean of these vectors over its T word fragments and over the set B of layers 30% to 50% of the way through the LLM, unless stated otherwise, $\mathbf{h}(x) = (1/|B|T)\, \Sigma_{l \in B}\, \Sigma_t\, \mathbf{h}_{lt}(x)$ (Figure 2b). We then divided each number of the internal state by its standard deviation across all narratives, $\mathbf{z}(x) = \mathbf{h}(x) \oslash \mathbf{s}$, where $\oslash$ denotes elementwise division. For pair i, the observation is the difference between the high and the low narrative, $\mathbf{o}_i = \mathbf{z}(x_i^{\text{high}}) - \mathbf{z}(x_i^{\text{low}})$, and the representation of construct c is the mean of the observations of its n_c pairs, $\mathbf{r}_c = (1/n_c)\, \Sigma_i\, \mathbf{o}_i$ (Figure 2c; formal definitions in Supplemental Note 1). It is the direction in which the LLM's internal state moves from narratives that convey a low level of the construct to narratives that convey a high level.

To assess convergent and discriminant validity, we used the multitrait-multimethod matrix (Campbell & Fiske, 1959). In this approach, measures of the same trait obtained by different methods should correlate highly, and measures of different traits should correlate weakly, whether the same or different methods produced them. We treated the 66 constructs as traits and the six writers as methods and computed the mean observation $\mathbf{m}_{cw}$ of each construct c from each writer w. The convergent correlations are the correlations across the numbers of the internal state between the means of the same construct from different writers, $\rho(\mathbf{m}_{cw}, \mathbf{m}_{cw'})$, and the discriminant correlations are those between the means of different constructs, $\rho(\mathbf{m}_{cw}, \mathbf{m}_{c'w'})$, from the same writer or from different writers (Figure 2d).

To test whether a representation identifies its construct in narratives not used to estimate it, we estimated each representation from the narrative pairs of five writers, $\mathbf{r}_c^{(-w)}$, and tested it on the pairs of the sixth, held-out writer w (Judd et al., 2012). We assigned each held-out pair to the construct whose representation had the highest cosine similarity to the pair's observation, $\hat{c}_i =$

$\text{argmax}_c \cos(\mathbf{o}_i, \mathbf{r}_c^{(-w)})$, and defined the identification rate as the percentage of pairs assigned to their own construct (Figure 2d).

To test criterion validity, we also asked the LLM, after it had processed a narrative, to rate the construct on a scale from 1 to 9 with a first-person item formulated from the definition of the construct's rating scale. We computed each judgment as the expected rating over the probabilities that the LLM assigned to the digits 1 to 9. The criterion correlation relates how closely a pair's observation resembles a construct's representation to how much the LLM's judgments of that construct differ between the pair's two narratives (Supplemental Note 1). In Study 4, these judgments also served as a measure of the mediator.

Silicon-Sample Replications of the Published Experiments

To test whether the construct representations causally lead to the LLMs' responses, we conducted silicon-sample replications of the 60 published experiments, in which the original authors examined the mediators and rival constructs investigated in this research. Inside the LLMs, we traced each experiment's causal chain in two steps (Pirlott & MacKinnon, 2016; Spencer et al., 2005). We measured whether the manipulation activates the mediator's representation and whether this activation predicts the response (Study 3), and we induced and removed the representation to test whether it causes the response (Study 4). Among the 60 experiments, large preregistered replications with human participants have re-examined 28 and found the original effect in 14 (Bago et al., 2022; Camerer et al., 2018; Ebersole et al., 2016; Eerland et al., 2016; Klein et al., 2022, 2014, 2018; Lynott et al., 2014). These experiments allowed us to compare effects that hold in human participants with effects that do not.

Each time we presented an experiment to an LLM, we instructed the LLM with an agent profile that described a fictitious individual's age, gender, occupation, city, and personality, presented the experiment's scenario description in the treatment or the control condition, and then presented its outcome measure, the question to which the LLM responded on a scale from 1 to 9 (Table S2, Panel D; Supplemental Note 1). To ensure that each silicon sample consisted of different participants, we used ten such agent profiles. To ensure that the results do not depend on a particular wording, we presented each experiment in seven sets of scenario descriptions, which preserved its manipulation across different wordings and scenarios. No human participants took part in this research, which therefore did not require ethics approval.

To test whether the manipulation of a published experiment activates the mediator's representation, we measured how strongly the LLM's internal state points in the direction of a representation while the LLM processes a text. We call this measure the representation's activation, $a_c(x) = \mathbf{z}(x) \cdot \mathbf{r}_c / \|\mathbf{r}_c\|$, the projection of the standardized internal state onto the representation scaled to unit length (Figure 2e). For an experiment with mediator m, the manipulation's activation of the mediator's representation is the difference in this activation between the treatment and the control condition, minus the mean of the same difference for the other 65 constructs, most of them unrelated to the experiment, and signed so that positive values are in the direction that the theory predicts, $\Delta_m = \sigma_m \{[a_m(x^T) - a_m(x^C)] - (1/65) \Sigma_{c \neq m} [a_c(x^T) - a_c(x^C)]\}$, where x^T and x^C are the scenario descriptions of the two conditions (Figure 2f). We computed it for each combination of scenario description and agent profile. Unless stated otherwise, we measured representation activation while the LLM processed the scenario description, before it produced a response. To test whether the relation between this activation and the response depends on the activation's input proximity to response production, that is, on how close the point of the LLM's input at which it occurs is to the production of the response, Study 3 measured it at four points of the input, while the LLM processed the scenario description as a whole, the passages that convey the manipulation, and the outcome measure, and at the last word fragment before the response, each in the layers in which we estimated the representations (30% to 50% of the way through the LLM) and in later layers closer to the LLM's output (60% to 75%).

To test whether a representation causes the LLM's response, we induced and removed the representation inside the LLM without changing the scenario description. Both interventions act on the residual stream at a layer l. We induced a representation, that is, activated it directly, by adding it at every word fragment t of the scenario description, $\mathbf{h}_{lt} \leftarrow \mathbf{h}_{lt} + \alpha \|\mathbf{d}_c^{(l)}\| \mathbf{u}$, where $\mathbf{d}_c^{(l)}$ is the mean difference between the construct's high and low narratives at that layer, $\mathbf{u}$ is the unit vector of the induced direction (the construct's own representation or, for comparison, another construct's representation or a random direction), and $\alpha \in \{\pm 1, \pm 2, \pm 4\}$. In the experiments of Study 4, we induced the mediator's representation in the treatment condition with positive and negative α, so that the induction strengthened or weakened the mediator. We removed a representation by setting the residual stream's component in a construct-specific subspace to its mean across narratives, $\mathbf{h}_{lt} \leftarrow \mathbf{h}_{lt} - \mathbf{P}_c(\mathbf{h}_{lt} - \boldsymbol{\mu}_l)$, at every word fragment, where $\mathbf{P}_c$ is the

orthogonal projection onto the ten directions in which the construct's high and low narratives differ most. We identified these directions after excluding the three directions that the differences of all 66 constructs share and that therefore do not distinguish one construct from another (Supplemental Note 1).

We report effect sizes within each experiment as Cohen's d_z. To report the manipulation's activation of the mediator's representation, we computed d_z across the seven sets of scenario descriptions after averaging over agent profiles within each set; to report the manipulation's effect on the responses, we computed d_z across combinations of scenario description and agent profile. We estimated the 95% confidence intervals of effect sizes, correlations, and other statistics computed across constructs or experiments (Cumming, 2014) by bootstrapping the 66 constructs in Studies 1 and 2 and the experiments in Studies 3 and 4, using the same resamples for all LLMs when pooling across LLMs. For the numbers of constructs or experiments that showed an effect, such as the number of experiments in which the manipulation activated the mediator's representation more strongly than the other representations, we used exact binomial tests and confidence intervals against the number expected by chance. Where Studies 3 and 4 found no relation or no specific effect, we quantified the evidence for this absence with Bayes factors (BF01) and, for correlations, with equivalence tests with bounds of ±.20. Because Study 3 tested at which points of the input and in which layers the activation predicts the response, we computed these correlations at several points of the input and layers and corrected them for multiple comparisons (Supplemental Note 1). We report ranges across LLMs in the text, the values for each LLM in Tables S2 to S8, and the main results with their confidence intervals, equivalence tests, and Bayes factors in Table S10.

Results

Study 1. Validity of Construct Representations

To test whether the representations are valid measures of the constructs, we evaluated them against the four standards of construct validity (Table 2). We first examined their convergent and discriminant validity with the multitrait-multimethod matrix. In every LLM, the representations of the same construct estimated from different writers' pairs of narratives correlated at .49 to .64, whereas those of different constructs correlated at .04 to .07, whether

they came from the same writer or from different writers (Figures 2d and 3a; Tables S3 and S10). The representations thus reflected the constructs rather than the writers.

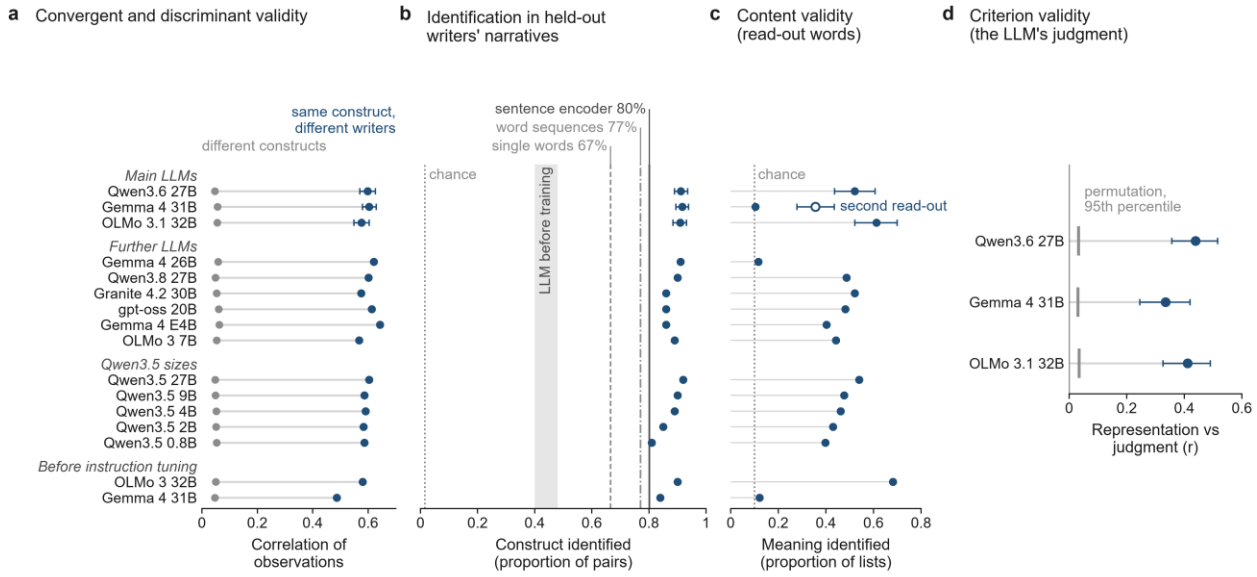


Fig. 3.

Validity of construct representations in large language models (LLMs; Study 1). In (a), the mean correlation between the representations of the same construct estimated from different writers (navy) and between the representations of different constructs (grey) is shown for each LLM. In (b), the proportion of the held-out writer's pairs assigned to the correct one of the 66 constructs (layers 30% to 50% of the way through each LLM) is shown; vertical lines and the band show the same test applied to the words and word sequences of the narratives, a sentence encoder, and the LLM before training. In (c), the proportion of read-out word lists for which the raters chose the correct one of ten definitions is shown; the open marker shows Gemma 4 31B with the read-out words obtained with the Jacobian lens. In (d), the correlation between how closely a pair's difference in internal states resembled a construct's representation and the difference between the LLM's judgments of that construct in the pair's two narratives is shown; grey bars show the 95th percentile of the permutation distribution. Error bars show 95% confidence intervals for the three main LLMs, Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B (in c, also for the Jacobian lens in Gemma 4 31B).

We then tested whether the representations identify the construct of a single pair of narratives composed by the held-out writer. For 81% to 92% of these pairs, the most similar of

the 66 representations was that of the pair's own construct, compared with a chance level of 1.5% (Figure 3b; Tables S2 and S10). In every LLM, the representations thus generalized to the narratives of a writer from whose narratives they had not been estimated.

An alternative explanation of this finding is that writers who convey the same construct use related words, which alone could identify the construct. To test discriminant validity against this explanation, we applied the same identification test to counts of words and word sequences in each narrative; to a sentence encoder trained to represent the meaning of sentences (Zhang et al., 2025), which tests whether the meaning of the sentences suffices; and to an LLM before training, whose internal states reflect which word fragments a narrative contains but nothing learned from text. In every trained LLM, the representations identified the construct more often than the best of the word-count measures (81% to 92% vs 77% of the pairs), and in the main LLMs they also did so more often than the sentence encoder (91% to 92% vs 80%). The LLM before training identified the construct in only 40% to 48% of the pairs (Figure 3b). The representations of the trained LLMs thus reflected more than the words of the narratives.

To examine content validity, the extent to which each representation conveys the meaning of its construct, we listed the words that each representation makes the LLM more likely to produce (read-out words; Supplemental Notes 1 and 2; nostalgebraist, 2020) and asked the two rater LLMs to choose, from ten definitions, the definition of the construct that each list conveys. In 13 of the 16 LLMs, the raters chose the correct definition for 40% to 68% of the lists, compared with a chance level of 10% (Figure 3c; Table S4). In these LLMs, the representations thus conveyed the meaning of their constructs.

Finally, we examined criterion validity, with the LLM's own judgment of the construct as the criterion. The more closely the difference between the internal states of a pair's two narratives resembled the representation of a construct, the more the main LLMs' judgments of that construct differed between the two narratives (r = .34 to .44; permutation $p < .001$; Figure 3d; Table S3). The representations thus predicted the LLMs' own judgments of the constructs. In sum, the representations met the standards of construct validity that psychologists apply to measurements with human participants (Table 2), showing convergent and discriminant validity in every LLM, content validity in 13 of the 16 LLMs, and criterion validity in each main LLM.

Table 2

Standards of Construct Validity, Their Tests, and the Results (Studies 1 and 2)

Standard	Definition	Test	Result
Convergent and discriminant validity	Measures of the same construct obtained by different methods correlate highly, and measures of different constructs correlate weakly (Campbell & Fiske, 1959)	Multitrait-multimethod correlations, with the six writers as methods; identification of the construct of each of the held-out writer's pairs among the 66 constructs	Same construct, r = .49 to .64; different constructs, r = .04 to .07; identification, 81% to 92% of the pairs (chance, 1.5%)
Discriminant validity against alternative explanations	A measure reflects its construct rather than the method by which it was obtained (Campbell & Fiske, 1959)	The same identification test applied to the words of the narratives, to a sentence encoder, and to an LLM before training	Best word counts, 77%; sentence encoder, 80%; LLM before training, 40% to 48% (trained LLMs, 81% to 92%)
Content validity	The content of a measure reflects the meaning of its construct (Messick, 1995)	Raters' choice of the construct's definition, among ten, from the read-out words of its representation	Correct definition for 40% to 68% of the lists in 13 of the 16 LLMs (chance, 10%)
Criterion validity	A measure predicts the criteria that its construct predicts (Cronbach & Meehl, 1955)	Correlation between how closely a pair's difference in internal states resembles a construct's representation and how differently the LLM judges that construct in the pair's two narratives	r = .34 to .44 in the main LLMs
Measurement invariance	A measure assesses a construct in the same way across groups (Vandenberg & Lance, 2000)	Identification of each construct's profile across writers and LLMs, in dimensions that the main LLMs share; correlation between LLMs of the similarities among the constructs	Identification, 93% of the profiles (chance, 1.5%); correlation of the similarities, ρ = .94 to .96
Theoretical structure	The relations among the measures correspond to the relations that theory specifies among the constructs (Cronbach & Meehl, 1955)	Similarity of the representations of constructs that arise in the same appraisal process (Moors et al., 2013; Smith & Ellsworth, 1985), compared with constructs of different processes	Mean similarity .076 to .088 within and .050 to .062 between processes in the main LLMs (permutation $p < .001$)

Note. Ranges are across the 16 LLMs unless the main LLMs (Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B) are named. Identification is the percentage of the held-out writer's pairs assigned to their own construct.

Study 2. A Shared Structure of Psychological Constructs Across LLMs

Having found valid representations in each LLM, we asked whether LLMs from different developers, which are trained on different text with different procedures, represent the constructs in the same way, as psychologists test whether a measure assesses a construct in the same way across groups, a property called measurement invariance (Vandenberg & Lance, 2000). Because

the internal states of different LLMs are vectors of different lengths whose numbers do not correspond, we first mapped the internal states of the three main LLMs onto a common set of dimensions (Supplemental Note 1). On these common dimensions, each construct's representation had a characteristic profile of values that was stable across the six writers and the three LLMs (Figure 4a). Across the 66 constructs, 93% of the profiles from one writer in one LLM were most similar to the same construct's mean profile from the other writers in the other LLMs (95% CI [90%, 96%]; chance, 1.5%); profiles of the same construct had a mean similarity of .79 (95% CI [.76, .82]), compared with .00 for different constructs.

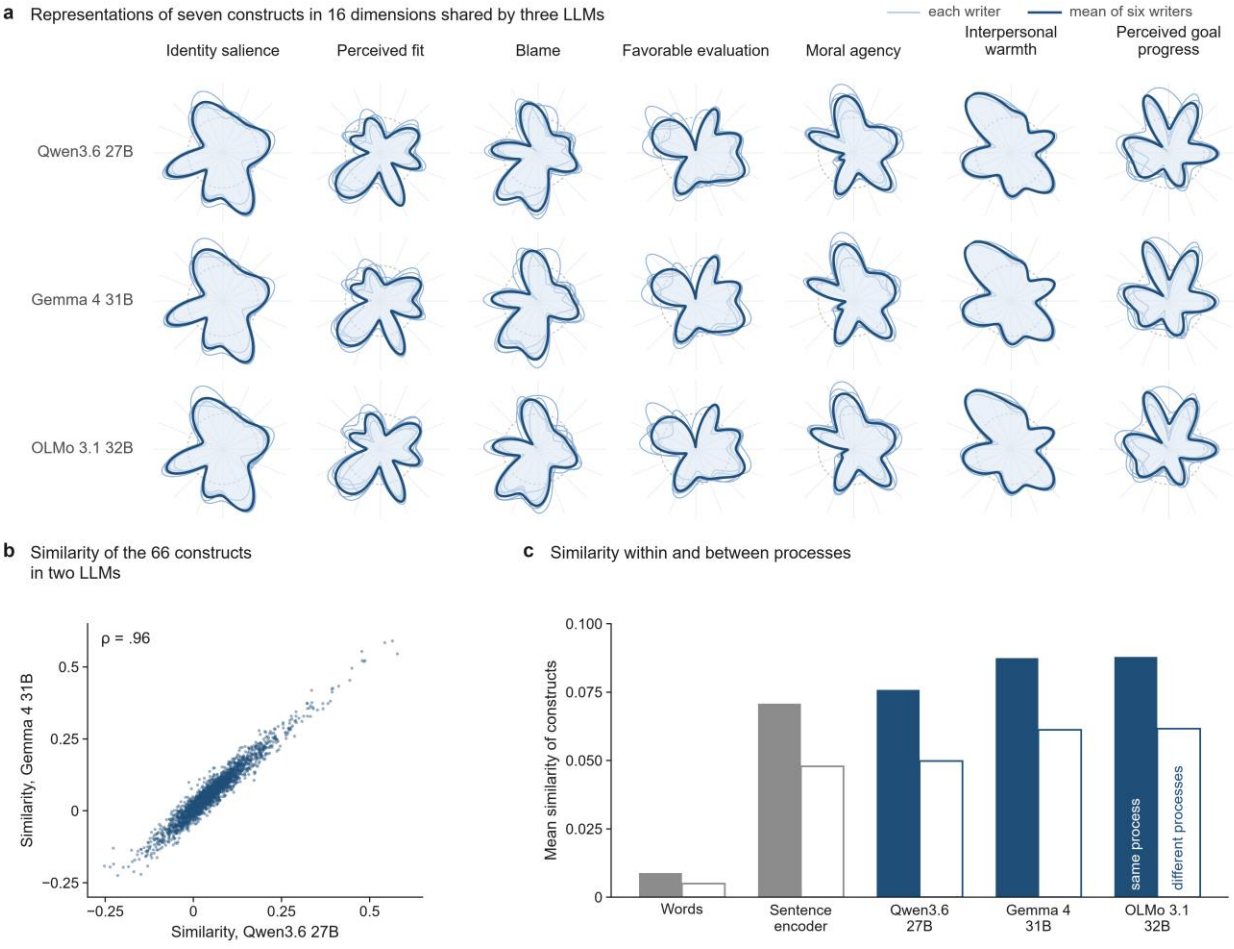


Fig. 4.

A shared structure of psychological constructs across large language models (LLMs; Study 2). In (a), the representations of seven constructs, one from each of the six appraisal processes and one from the comparison of alternatives, are shown on 16 dimensions shared by the three main LLMs

(Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B); thin outlines show the six writers, and the bold outline shows their mean. In (b), the similarities of the representations of every two constructs in Qwen3.6 27B and in Gemma 4 31B are shown; ρ is the Spearman correlation between them. In (c), the mean similarity between constructs of the same appraisal process (filled markers) and of different processes (open markers) is shown, computed from the words of the narratives, a sentence encoder, and the three main LLMs.

The LLMs also shared the relations among the constructs. Pairwise similarities between construct representations correlated at .94 to .96 across the main LLMs (Spearman ρ; Figure 4b) and remained at .82 to .88 after we statistically controlled for the similarities computed from the words of the narratives and from the sentence encoder (Table S3). LLMs from three developers thus shared both the profile of each construct and the relations among the constructs.

We next asked whether the relations among the constructs reflect the psychological processes in which the constructs arise. According to appraisal theories, appraisal consists of separate processes, such as appraising whether a situation bears on one's goals (Moors et al., 2013; Smith & Ellsworth, 1985). We therefore assigned each construct to one of six appraisal processes or to the comparison of the alternatives in a choice (Table 1; Supplemental Note 1) and compared the representations of constructs within and between processes. In each main LLM, constructs within the same process had more similar representations than constructs from different processes (mean similarity .076 to .088 within processes and .050 to .062 between them; permutation $p < .001$; Figure 4c; Supplemental Note 3), and this difference remained after we statistically controlled for the similarities computed from the words of the narratives and from the sentence encoder (permutation $p \leq .007$). The LLMs thus grouped the constructs by the processes in which, according to appraisal theories, the constructs arise.

These representations did not require instruction tuning. The two LLMs that we examined before instruction tuning identified the construct in 86% and 91% of the pairs, and in OLMo 3 7B, whose stages of training are public, the representations identified the held-out writer's construct in 79% of the pairs after only 10,000 of the 1.41 million updates of the first stage of training, in which the LLM learns to predict text (Supplemental Notes 3 and 7; Figure S1).

Study 3. Activation of the Mediator's Representation in Silicon Samples

Having validated the representations, we tested whether they causally lead to the responses of LLMs that took part in the 60 experiments as silicon samples. We first asked whether the published manipulations activate the representation of the presumed mediator. In line with the theories, while the LLMs processed the scenario description, the manipulations activated the representation of their mediator more strongly than those of the other constructs, pooled across the 14 instruction-tuned LLMs (d_z = 1.62, 95% CI [1.07, 2.12]) and in 42 to 49 of the 60 experiments in each LLM, where chance would produce about 30 (binomial $p \leq .002$; Figure 5a; Table S5). The manipulations thus activated the mediators that the theories propose.

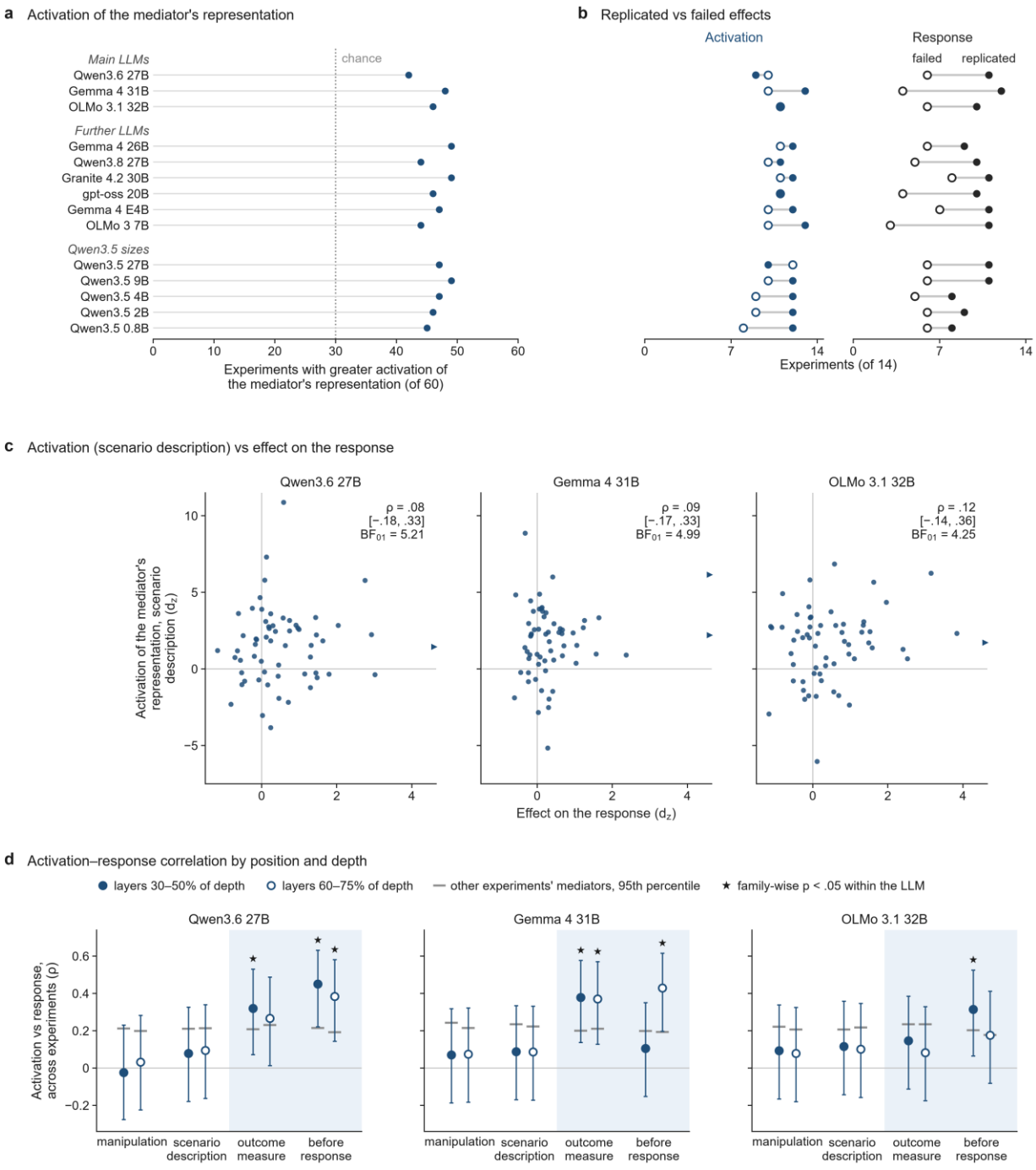


Fig. 5.

Activation of the mediator's representation across 60 experiments (Study 3). In (a), the number of experiments in which the manipulation activated the mediator's representation more than the other constructs' representations while the large language model (LLM) processed the scenario description is shown for 14 LLMs; the dotted line shows the number expected by chance. In (b),

the number of experiments in which the manipulation activated the mediator's representation (left) and in which the LLM's responses showed the original effect (right) is shown for the 14 experiments whose original effects replicated in human participants (filled markers) and the 14 whose effects failed to replicate (open markers). In (c), the activation of the mediator's representation while the LLM processed the scenario description is plotted against the manipulation's effect on the response for each experiment; triangles mark experiments beyond the axis limit. In (d), the correlation across experiments between the activation of the mediator's representation and the manipulation's effect on the response is shown for activation measured while the LLM processed the manipulation, the scenario description, and the outcome measure, and at the last word fragment before the response (shaded), in the layers 30% to 50% (filled markers) and 60% to 75% (open markers) of the way through each of the three main LLMs (Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B); stars mark family-wise $p < .05$, and grey dashes show the 95th percentile of the correlations when each experiment is paired with another experiment's mediator. Error bars show 95% confidence intervals.

We then compared the experiments whose effects hold in human participants with those whose effects do not. Of the 28 experiments that large preregistered replications have re-examined, the LLMs' responses showed the original effect more often in the 14 whose effects had replicated (rank-biserial correlation, .64, 95% CI [.29, .94]; Figure 5b). The manipulations, however, also activated the mediator's representation in the 14 experiments whose effects had failed to replicate (pooled across the 14 LLMs, $d_z = 1.12$, 95% CI [0.28, 1.90]), about as strongly as in those whose effects had replicated (rank-biserial correlation, .16, 95% CI [−.27, .59]). The LLMs' responses thus distinguished effects that hold in human participants from effects that do not, whereas the activation of the mediator's representation reflected the theory's prediction in both.

We next asked whether the mediator's activation leads to the response that the theory predicts. If it does, the manipulations that activate the mediator's representation more strongly should have larger effects on the response. While the LLMs processed the scenario description, this activation was unrelated to the manipulation's effect on the responses across the 60 experiments, in each of the 14 LLMs (BF01 = 3.33 to 6.21; Figure 5c; Supplemental Note 4) and pooled across them ($\rho = .05$, 95% CI [−.11, .21]; an equivalence test rejected correlations of .20 or larger, 90% CI [−.09, .18]).

An LLM processes an experiment's scenario description first and its outcome measure last, immediately before it responds, so the outcome measure has the greater input proximity to response production. To understand why the activation during the scenario description did not predict the responses, we therefore compared it with the activation during the outcome measure, which immediately precedes the response, first in two experiments (Figure 6a; Table S5). In the moral typecasting experiment (Gray & Wegner, 2009), whose outcome measure asked how responsible the person who pushed the tray was, the manipulation activated the representation of moral agency more strongly while the main LLMs processed the outcome measure directly before responding than while they processed the scenario description (d_z = 2.6 to 4.4 vs 1.5 to 2.2) and had a large effect on their responses (d_z = 7.1 to 8.3). By contrast, in the identity-salience experiment, human participants first write about moments when they grew closer to their family and then rate a new product (Reed, 2004). While the LLMs processed the description of this writing task, the manipulation strongly activated the representation of identity salience, here the importance of one's identity as a family member (d_z = 10.1 to 18.8). However, while they processed the outcome measure about buying the product, the manipulation no longer activated this representation (d_z = −2.4 to 0.3), and its effect on their responses was near zero (d_z = −0.08 to 0.06). In both experiments, whether the manipulation affected the LLMs' responses corresponded to whether it activated the mediator's representation while the LLMs processed the outcome measure.

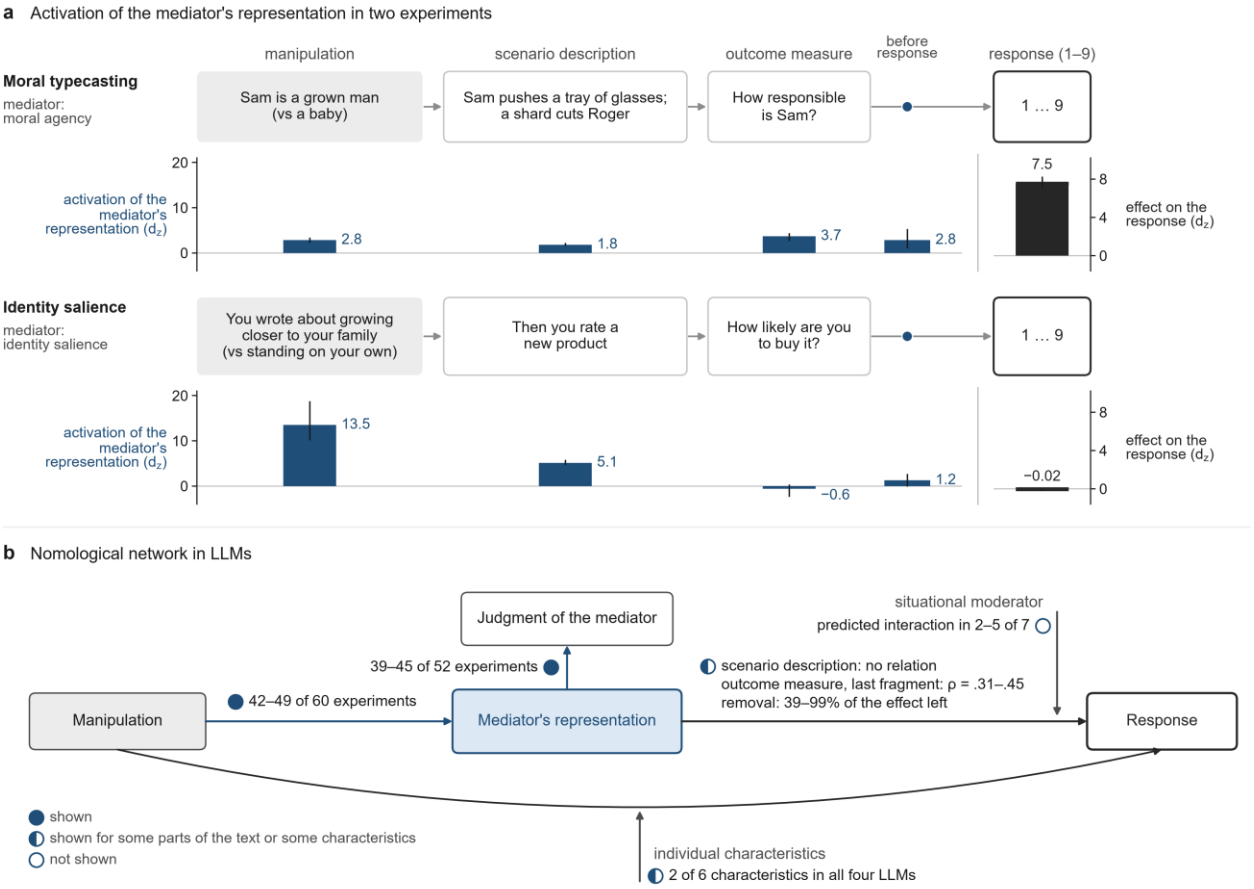


Fig. 6.

The causal chain inside large language models (LLMs) in two experiments and the nomological network. In (a), the activation of the mediator's representation by the manipulation while the LLMs processed each part of the input of two experiments, moral typecasting (Gray & Wegner, 2009) and identity salience (Reed, 2004), is shown as the standardized mean difference d_z (blue bars; mean of the three main LLMs, Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B; whiskers show the range), together with the manipulation's effect on the response (black bars, d_z; right axis). In (b), each relation of the nomological network is marked as shown in the LLMs (filled), shown for some parts of the input or some characteristics (half-filled), or not shown (open); removal is the share of the manipulation's effect left after removing the mediator's representation (see Table 3).

We then tested whether this correspondence held across all 60 experiments in the main LLMs (Figure 5d; Table S5). While the LLMs processed the passages that convey the manipulation or the scenario description as a whole, the activation of the mediator's

representation was unrelated to the manipulation's effect on the responses in either set of layers (30% to 50% and 60% to 75% of the way through each LLM; ρ = −.02 to .12; BF01 = 4.25 to 6.11). After the scenario description, however, the manipulations that activated the mediator's representation more strongly had larger effects on the responses. The outcome measure was identical in both conditions, so a difference in activation while the LLMs processed it arose from the preceding scenario description. During the outcome measure, two of the three LLMs showed this relation (ρ = .32 to .38; family-wise $p \leq .037$). At the last word fragment before the response, all three showed it (ρ = .31 to .45 in one or both sets of layers; family-wise $p \leq .044$). Pooled across the three LLMs, this relation was stronger after the scenario description than within it (difference in ρ = .21, 95% CI [.07, .36]). The relation after the scenario description was also specific to each experiment's own mediator. When we paired each experiment with the mediator of another experiment, the correlations were near zero (mean ρ = −.05 to .02) and smaller than those with the experiment's own mediator (permutation $p \leq .007$). The greater the input proximity of a point to response production, the more strongly the manipulation's activation of the mediator's representation at that point predicted the manipulation's effect on the response.

Study 4. Inducing and Removing the Mediator's Representation

Because correlations across experiments cannot establish that a representation causes a response, we induced and removed the mediator's representation directly. As a manipulation check, we first induced each construct's representation while the LLMs processed the narratives of Study 1. For each construct, we asked whether inducing its representation affected the LLM's judgment of that construct more than inducing the representations of other constructs or random directions of the same size did. Without a specific effect, the construct's own representation would affect the judgment more than these comparisons for about half of the 66 constructs, or 33. It did so for 48 to 59 of the 66 constructs in the main LLMs and for 45 to 56 in eight further LLMs (Figure 7a; Table S6). Inducing a representation thus affected the LLM's judgment of its own construct specifically.

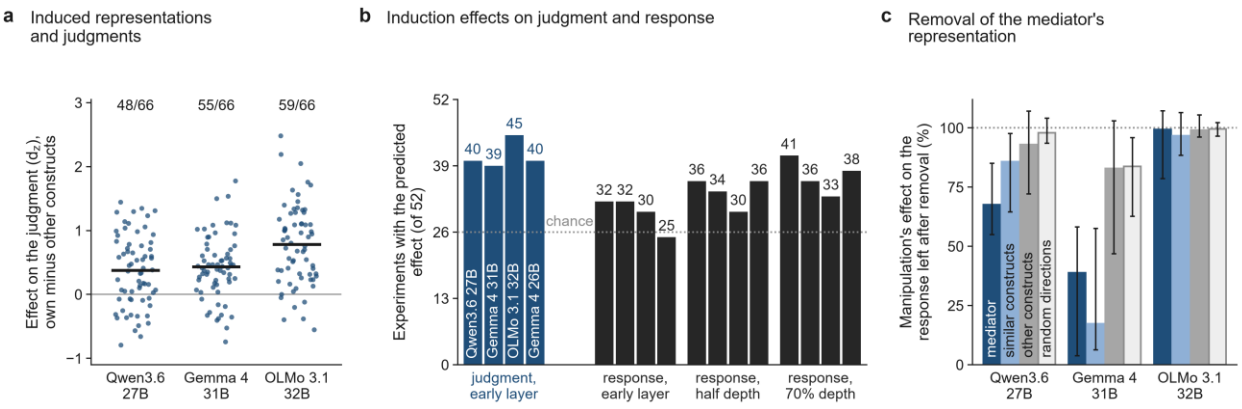


Fig. 7.

Effects of inducing and removing the mediator's representation (Study 4). In (a), the effect of inducing each construct's representation on the large language model's (LLM's) judgment of that construct, minus the effects of inducing other constructs' representations, is shown for the narratives of Study 1; bars show the median. In (b), the number of the 52 experiments in which inducing the mediator's representation affected the judgment of the mediator (blue) and the response in the predicted direction (black) is shown for inductions during the scenario description early in the LLM and halfway through it, and for inductions 70% of the way through the LLM from the beginning of the scenario description through the end of the outcome measure. In (c), the median share of the manipulation's effect on the response left after removing the representation of the mediator, of the constructs most similar to it, of randomly chosen other constructs, or of ten random directions is shown for the three main LLMs (Qwen3.6 27B, Gemma 4 31B, and OLMo 3.1 32B); error bars show 95% confidence intervals over the 29 to 36 experiments in which the manipulation changed the response by more than .3 scale points.

We then induced the mediator's representation in the 52 experiments whose theories do not make the mediator's effect depend on a situational moderator. For each experiment, we asked whether the induction moved the LLM's judgment of the mediator, and then its response to the experiment's outcome measure, in the theoretically predicted direction. Without an effect of the induction, each outcome would move in that direction in about half of the 52 experiments, or 26.

We first induced the mediator's representation while the LLMs processed the scenario description. Their judgment of the mediator, which they gave after the scenario description, moved in the predicted direction in 39 to 45 of the 52 experiments in the main LLMs (binomial *p*

< .001), and in 32 to 39 of these experiments it moved more than after we induced the representations of other constructs (binomial p = .018, .063, and < .001; Tables S6 and S10). Their responses to the outcome measure, in contrast, moved in the predicted direction in only 30 to 32 of the 52 experiments, close to the 26 expected without an effect (binomial p = .063 to .166; Figure 7b). Because the responses were related to the activation of the mediator's representation at points of greater input proximity to response production (Study 3), we then induced the representation 70% of the way through the LLM, at every word fragment from the beginning of the scenario description through the end of the outcome measure. This induction moved the responses in the predicted direction in 33 to 41 of the 52 experiments (binomial p < .001 to p = .035). Inducing the representations of other constructs, however, moved the responses about as much. Inducing the mediator's representation had a larger effect than inducing other constructs' representations in only 20 to 31 of the 52 experiments, about as often as expected without a specific effect (binomial p = .106 to .965; BF01 = 7.17 and 15.07 in two of the three main LLMs; Tables S6 and S10).

Removing the mediator's representation, in contrast, reduced the manipulation's effect on the response in two of the three main LLMs, to 68% (95% CI [55%, 85%]) and 39% (95% CI [4%, 58%]) of the effect without removal (medians across experiments; Table S6). By comparison, removing the representations of randomly chosen other constructs or ten randomly chosen directions left 83% to 98% of this effect (Figure 7c; Supplemental Note 5). In these LLMs, the manipulation's effect on the responses thus depended on the mediator's representation.

Finally, the theories of seven experiments specify that a situational moderator alters the mediator's effect on the response. In these experiments, we induced the mediator's representation at both levels of the moderator (Supplemental Note 5). The moderators altered the effect of the induced representation on the responses in the predicted direction in only 2 to 5 of the 7 experiments, and significant interactions occurred in both the predicted and opposite directions (Table S7).

Study 5. Individual Characteristics and Their Representations

In human participants, a manipulation's effect often varies with individual characteristics, a variation we refer to as heterogeneity. In silicon samples, researchers describe each participant in a profile (Argyle et al., 2023; J. S. Park et al., 2024), so that this heterogeneity can arise. In

Study 5, we tested whether describing an individual characteristic in the agent profile activates its representation and produces the heterogeneity that the theories predict. We examined six individual characteristics that, according to the theories, alter the manipulation's effect, along with self-monitoring, for which the theories predict no heterogeneity. For each characteristic, we estimated its representation from pairs of narratives, as in Study 1 (Supplemental Notes 1 and 6). Describing a high rather than a low level of a characteristic in the agent profile activated its representation immediately before the response in 25 of the 28 characteristic–LLM combinations across the main LLMs and Gemma 4 26B (Table S8). The predicted heterogeneity, however, occurred in all four LLMs for only two of the six characteristics, the belief that higher costs justify higher prices and the preference for consistency. In line with the theories, self-monitoring produced no significant heterogeneity in three of the four LLMs (Figure 8; Table S8).

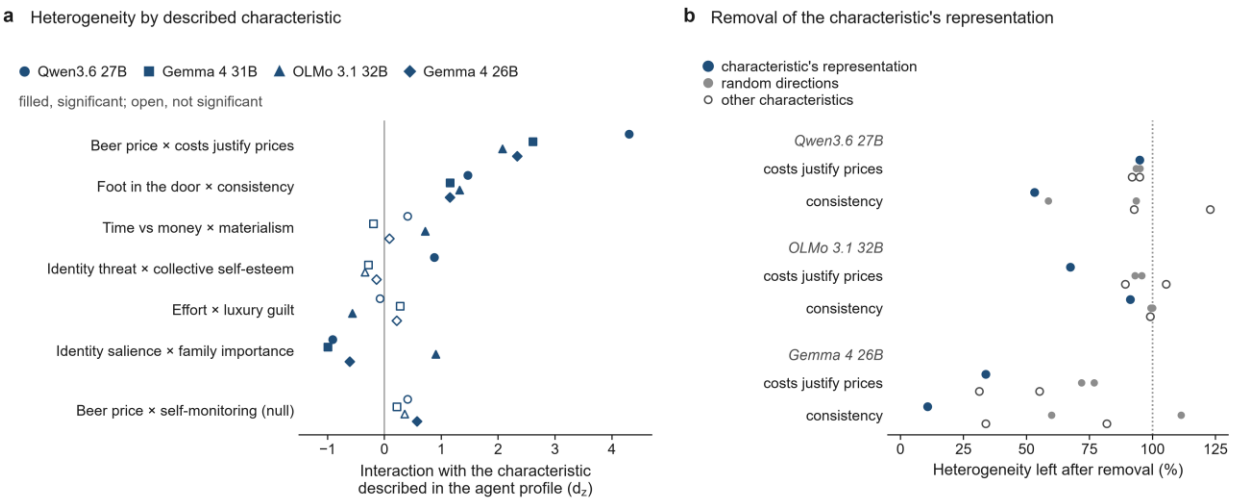


Fig. 8.

Individual characteristics and their representations (Study 5). In (a), the interaction between the manipulation and the individual characteristic described in the agent profile (the description of the individual given to the large language model [LLM]) is shown for four LLMs, signed so that positive values are in the predicted direction; filled markers indicate $p < .05$, open markers indicate $p \geq .05$, and the last row (null) shows self-monitoring, for which no interaction is predicted. In (b), the interaction left after removing, at the statement that describes the characteristic, the characteristic's representation, two sets of random directions of the same number, or the representations of two other characteristics is shown for the two characteristics that interacted as predicted in all four LLMs.

Discussion

Our findings show that LLMs have stable latent representations of the psychological constructs that process theories in psychology propose, and that these representations meet the standards of construct validity that psychologists apply to measurements with human participants. We also found that LLMs represent states that arise in the same appraisal process more similarly than states of different processes (Moors et al., 2013; Smith & Ellsworth, 1985).

Across 60 experiments presented to LLMs as silicon samples, manipulations from published research activated the representation of the presumed mediator, and removing this representation reduced the manipulations' effects on the responses in two of the three main LLMs (Table 3). In addition, whether this representation was related to the response depended on its input proximity to response production, that is, on when, in processing the experiment, the LLM activated it. How strongly a manipulation activated the mediator's representation while the LLMs processed the scenario description was unrelated to the manipulation's effect on the responses, whereas how strongly it activated this representation while they processed the outcome measure, immediately before they responded, was related to this effect (Figure 5d).

Table 3

Relations of the Mediator's Nomological Network, Their Tests, and the Results (Studies 3 to 5)

Relation	Test inside the LLMs	Result
The manipulation affects the mediator	Activation of the mediator's representation by the published manipulation while the LLMs processed the scenario description, compared with the other 65 representations	Activated more than the other representations in 42 to 49 of the 60 experiments (chance, about 30)
The mediator affects the response	Correlation across experiments between the manipulation's activation of the mediator's representation and its effect on the response	Unrelated during the scenario description ($\rho = -.02$ to .12); related while the LLMs processed the outcome measure or before the response ($\rho = .31$ to .45)
Inducing the mediator changes the judgment and the response	Mediator's representation induced in 52 experiments	Judgment of the mediator in the predicted direction in 39 to 45 experiments; responses in 30 to 41, but no more often than after inducing other representations (20 to 31; chance, 26)
Removing the mediator reduces the manipulation's effect	Mediator's representation removed from the internal states for the whole text	Effect reduced to 68% and 39% of its size without removal in two of the three main LLMs; 83% to 98% left after removing other representations or random directions
A situational moderator alters the mediator's effect	Mediator's representation induced at both levels of the moderator in seven experiments	Interaction in the predicted direction in 2 to 5 of the 7 experiments
Individual characteristics produce heterogeneity	Characteristic described in the agent profile in six tests	Its representation activated in 25 of 28 combinations of characteristic and LLM; predicted heterogeneity in all four LLMs for 2 of the 6 characteristics
Original effects that replicated differ from those that failed	Comparison of the 14 replicated and the 14 failed effects	Responses distinguished them (rank-biserial correlation, .64); activation of the mediator's representation did not (.16)

Note. The relations are those that process theories specify and that the experimental causal chain approach tests (Pirlott & MacKinnon, 2016; Spencer et al., 2005). Ranges are across the LLMs in which each test was conducted; chance is the number of experiments expected without a specific effect.

These findings reveal a dissociation between the responses of a silicon sample and the psychological processes inside the LLM. Of the 28 experiments that large preregistered replications have re-examined, the LLMs' responses showed the original effect more often in those whose effects had replicated in human participants, but the manipulations activated the mediator's representation whether or not the effect had replicated (Figure 5b). In silicon samples, as in research with human participants, the theorized process can occur without the predicted effect, and the predicted effect can occur without the theorized process (Rucker et al., 2011;

Zhao et al., 2010). A silicon sample that matches human results has therefore not yet shown that it shares the psychology behind them.

Before conducting an experiment with human participants, researchers can use the representations we validated to test whether a planned manipulation activates the presumed mediator in an LLM and, by inducing or removing this representation, whether the mediator in turn leads to the response. These two causal relations are difficult to test in human participants, in whom manipulating the mediator often alters other psychological states as well (Bullock et al., 2010; Pirlott & MacKinnon, 2016).

Note that we examined construct representations only in open-weight LLMs of up to 32 billion parameters and only for a narrow set of psychological constructs, mostly the mediators of 60 published experiments. Several open questions therefore remain that may limit the generalizability of our findings and that future research should address. First, construct representations may change with model size. In our data, the representations identified the constructs more accurately in larger LLMs (Supplemental Note 3), in line with evidence that larger models represent the world in increasingly similar ways (Huh et al., 2024). Larger LLMs, however, can also respond less like human participants (Aher et al., 2023). Second, the constructs we examined were mostly states that a situation evokes, such as inferences and appraisals, which LLMs are known to represent (Tak et al., 2025; Zhu et al., 2024). Stable dispositions, such as the need for cognition (Cacioppo & Petty, 1982) or a chronic regulatory focus (Higgins, 1997), may be represented differently or affect responses differently. In Study 5, describing individual characteristics activated their representations, yet only two of the six characteristics produced the predicted heterogeneity in all four LLMs. Third, we presented every construct in text, whereas human participants also encounter situations in images and sounds, which can engage different representations than words do (Barsalou, 2008). According to psychological constructionism, however, psychological states are constructed from conceptual knowledge that language helps to acquire (Barrett, 2009; Lindquist et al., 2015). If so, LLMs may represent constructs as concepts, whether a situation is described in words or shown in images, a possibility that can be tested in multimodal LLMs.

In summary, today's Clever Hans probably still does not do arithmetic, but the math is indeed represented in its head.

Supplemental Material for

Not Your Average Clever Hans: Latent Representation of Psychological Constructs in Large Language Models

Jack Chen, Zelin Zhou, Chunlin An, and Lexi Chen

Contents: Supplemental Notes; Figure S1; Tables S1 to S10; References.

In the Supplemental Tables, percentages are rounded to whole numbers and correlations to two decimals. n.a. = not available; not run = not examined in this LLM.

Supplemental Note 1. Methods in Detail

Constructs

The 66 constructs and the rating scales that define them come from the materials of earlier work on the same 60 experiments (J. Chen et al., 2026). They comprise the 55 distinct mediators of the 60 experiments, the eight rival constructs that the original articles had ruled out, and three general evaluations (favorable evaluation, pleasantness, and appeal of the setting), whose rating scales we worded to apply to any scenario. To test in Study 2 whether the relations among the representations reflect the processes in which the constructs arise, we assigned each construct to the process in which, according to its theory, it arises. We used the six appraisal processes distinguished in the earlier work, categorization (12 constructs), social inference (13), self-relevance (10), affective response (9), goal relevance (9), and causal inference (8), and the comparison of alternatives in a choice (5).

Pairs of narratives

The six writers were DeepSeek V4.1 Flash, Kimi K3, Muse Spark 1.3, Grok 4.7, MiniMax M3, and Claude Sonnet. None of them is among the 16 LLMs that we examined, so none of these LLMs processed narratives that it had composed. The writers composed pairs of second-person narratives; the two narratives of a pair describe the same scenario and are of nearly equal length (Table S9). Automated checks ensured that neither narrative used the construct's name or its synonyms, that the two narratives differed in length by at most 15%, and that numbers appeared in both narratives or in neither. To check whether each pair of narratives manipulated its intended construct, two raters from developers other than the writer's (DeepSeek

V4.1 Flash and GLM 5.3, with Kimi K3 in place of DeepSeek V4.1 Flash for DeepSeek's narratives) rated each narrative, on scales from 1 to 7, on every construct of the same process and on the three general evaluations.

Experiments and agent profiles

Table 1 lists the 60 experiments with their original articles; Supplementary Table 1 and Supplementary Note 21 of the earlier work describe them in detail (J. Chen et al., 2026). To ensure that the results do not depend on a particular wording, each experiment had seven sets of scenario descriptions. One set is the wording of the earlier work (J. Chen et al., 2026). The other six sets are rewordings, composed by two writers, that retain the manipulation and place it in other scenarios. Each experiment had 6 to 12 rewordings, which we divided into six sets in the order in which the writers had listed them, so that each set contained one or two rewordings of every experiment. Each time we presented an experiment, the LLM received one of ten agent profiles, descriptions of individuals taken from the earlier work, as its system prompt, followed by the scenario description and the experiment's outcome measure, a question with a response scale from 1 to 9 (Table S2). We scored the LLM's response, like its judgments, as the expected value over the probabilities that it assigned to the digits 1 to 9.

To measure the activation of the representations, we averaged the LLM's internal states for the scenario description over its word fragments and over the layers 30% to 50% of the way through the LLM, standardized them, and projected them onto each of the 66 representations scaled to unit length. The manipulation's activation of the mediator's representation was the difference between the treatment and the control condition in the activation of the mediator's representation, minus the mean of the same difference for the other 65 representations. We computed this measure for each combination of scenario description and agent profile.

To test whether the relation between the activation and the response depends on the activation's input proximity to response production, Study 3 measured the activation of the mediator's representation in the three main LLMs at successive points of the input, from the passages that convey the manipulation to the last word fragment before the response. For the passages that convey the manipulation and for the outcome measure, we averaged the internal states over their word fragments. We measured the activation in the layers 30% to 50% and 60% to 75% of the way through the LLM, with the representations estimated in the same layers. To

test whether the correlation across experiments between this activation and the manipulation's effect on the response was specific to the mediator, we compared it with the correlations obtained when we assigned each experiment the mediator of another experiment, in 2,000 random assignments in which no experiment kept its own mediator.

The designs of Studies 4 and 5 are those of the earlier work (J. Chen et al., 2026). Study 4 induced the mediator's representation in the 52 experiments whose theories do not make the mediator's effect depend on a situational moderator and tested the moderator in the seven experiments that have one (11 tests). Study 5 tested six individual characteristics, each in one experiment.

LLMs

We examined Qwen3.6 27B, Gemma 4 31B, OLMo 3.1 32B Instruct, Gemma 4 26B-A4B, Qwen3.5 (0.8B, 2B, 4B, 9B, and 27B), Gemma 4 E4B, OLMo 3 7B Instruct, Qwen3.8 27B, gpt-oss 20B, Granite 4.2 30B, and the versions of OLMo 3 32B and Gemma 4 31B before instruction tuning. To examine how the representations arise during training (Supplemental Note 7), we also examined OLMo 3 7B at 12 stages of its training, ten stages from 0 to 1,413,814 updates of the first stage and the ends of the second and third stages, and after supervised and preference tuning (Table S2). We examined all 16 LLMs in Study 1 and, because an experiment requires an LLM to respond to a question, the 14 LLMs after instruction tuning in Study 3. In Study 2, in the test of criterion validity, and in Studies 4 and 5, we examined the three main LLMs and, in some analyses, further LLMs; Table S2 lists the LLMs that entered each analysis. We ran the LLMs with the Transformers library (Wolf et al., 2020) in bfloat16 precision with their chat templates. LLMs without a chat template, such as those at intermediate stages of training, received the instructions and the text as plain text that ended with "Answer:", in the same format at every stage.

Notation and formulas

Let $\mathbf{h}_{lt}(x)$ be the residual stream, a vector of D numbers, of text x at layer l and word fragment t. The internal state of a narrative is the mean over its T word fragments and over the set B of layers 30% to 50% of the way through the LLM, $\mathbf{h}(x) = (1/|B|T)\,\Sigma_{l\in B}\,\Sigma_t\,\mathbf{h}_{lt}(x)$. The standardized internal state is $\mathbf{z}(x) = \mathbf{h}(x) \oslash \mathbf{s}$, where $\oslash$ denotes elementwise division and s_j is the standard deviation of number j across all narratives; because every measure compares texts, no

centering is needed. For pair i, with construct $c(i)$ and writer $w(i)$, the observation is $\mathbf{o}_i = \mathbf{z}(x_i^{\text{high}}) - \mathbf{z}(x_i^{\text{low}})$, and the representation of construct c is $\mathbf{r}_c = (1/n_c)\,\Sigma_{i \in I(c)}\,\mathbf{o}_i$, where $I(c)$ is the set of its n_c pairs.

For the multitrait-multimethod matrix, the mean observation of construct c from writer w is $\mathbf{m}_{cw} = (1/n_{cw})\,\Sigma_{i \in I(c,w)}\,\mathbf{o}_i$, where $I(c, w)$ is the set of the n_{cw} pairs of construct c composed by writer w. The convergent correlations are the correlations across the D numbers $\rho(\mathbf{m}_{cw}, \mathbf{m}_{cw'})$, $w \neq w'$, and the discriminant correlations are $\rho(\mathbf{m}_{cw}, \mathbf{m}_{c'w'})$, $c \neq c'$. For the identification test, the representation estimated without writer w is $\mathbf{r}_c^{(-w)} = (1/n_c^{(-w)})\,\Sigma_{i \in I(c), w(i) \neq w}\,\mathbf{o}_i$, and each of writer w's pairs is assigned to $\hat{c}_i = \text{argmax}_c\,\cos(\mathbf{o}_i, \mathbf{r}_c^{(-w)})$.

A judgment is the expected rating $J = \Sigma_k\, k\, p_k \,/\, \Sigma_k\, p_k$, where p_k is the probability that the LLM assigned to the digit k ($k = 1, \ldots, 9$). For pair i and judged construct j, the criterion correlation is the correlation between $\cos(\mathbf{o}_i, \mathbf{r}_j^{(-w(i))})$ and $J_j(x_i^{\text{high}}) - J_j(x_i^{\text{low}})$ after both are centered on the means of each pair and of each judged construct.

The activation of representation c by text x is $a_c(x) = \mathbf{z}(x) \cdot \mathbf{r}_c \,/\, \|\mathbf{r}_c\|$. For an experiment with mediator m, the manipulation's activation of the mediator's representation is $\Delta_m = \sigma_m \{[a_m(x^{\text{T}}) - a_m(x^{\text{C}})] - (1/65)\,\Sigma_{c \neq m}\,[a_c(x^{\text{T}}) - a_c(x^{\text{C}})]\}$, where x^{T} and x^{C} are the scenario descriptions of the treatment and the control condition and $\sigma_m = \pm 1$ is the direction in which the theory predicts the manipulation to move the mediator.

Both interventions act on the unstandardized residual stream at a layer l. Induction adds to the residual stream of every word fragment t of the scenario description $\mathbf{h}_{lt} \leftarrow \mathbf{h}_{lt} + \alpha\,\|\mathbf{d}_c^{(l)}\|\,\mathbf{u}$, where $\mathbf{d}_c^{(l)}$ is the mean difference between the high and the low narratives of construct c at layer l, $\mathbf{u}$ is the unit vector of the induced direction ($\mathbf{d}_c^{(l)} \,/\, \|\mathbf{d}_c^{(l)}\|$ for the construct's own representation; for comparison, another construct's representation or a random direction), and $\alpha \in \{\pm 1, \pm 2, \pm 4\}$. Removal sets the component in a construct-specific subspace to its mean, $\mathbf{h}_{lt} \leftarrow \mathbf{h}_{lt} - \mathbf{P}_c(\mathbf{h}_{lt} - \boldsymbol{\mu}_l)$, at every word fragment t, where $\boldsymbol{\mu}_l$ is the mean internal state of the narratives at layer l and $\mathbf{P}_c$ is the orthogonal projection onto the subspace described below.

Internal states and representations

We presented each narrative as a user message and recorded the internal state, the residual stream, at every layer, averaged over the narrative's word fragments. We standardized

each of the internal state's numbers across all narratives and took the difference between the high and the low narrative of a pair as one observation. For the identification test, we assigned each of the held-out writer's pairs to the most similar of the 66 representations estimated from the other five writers (cosine similarity; ties counted against the pair's construct). For the multitrait-multimethod matrix, we averaged the observations of each construct from each writer and correlated these averages over the numbers that make up the internal state. The convergent correlations are those between the averages of the same construct from different writers, and the discriminant correlations are those between the averages of different constructs, from the same writer or from different writers. In the identification test, the held-out representations $\mathbf{r}_c^{(-w)}$ pool the observations of the other five writers; in the test of criterion validity, they are the mean of these writers' means $\mathbf{m}_{cw}$. In Study 3, $\mathbf{s}$ is the mean of the standard deviations at the layers of B.

To test whether the words of the narratives alone identify the constructs (Study 1), we applied the same identification procedure to single words (8,542 words that occur in at least two narratives) and to sequences of up to two and up to three words (101,985 and 282,350). We coded them as counts, as presence, or weighted by their rarity across narratives, raw or standardized, and we also applied latent semantic analysis to these counts (100 to 1,000 dimensions). We report the most successful of these variants, sequences of up to three words weighted by rarity, which gives the words their best chance of identifying the constructs. For the sentence encoder, we applied the same procedure to the standardized embeddings of each narrative from Qwen3-Embedding-8B (Zhang et al., 2025).

Read-out words

To examine content validity (Study 1), we listed read-out words for each construct. We estimated each construct's representation at the last word fragment of its narratives, as the difference between the average internal states of its high and its low narratives, at six layers from 50% to 100% of the way through the LLM. We projected this representation onto the LLM's output vocabulary, a method called the logit lens (nostalgebraist, 2020), and retained the 30 word fragments that it made most likely. For Qwen3.6 27B and Gemma 4 31B, we also obtained these lists with the Jacobian lens, which takes into account how the later layers of the LLM transform the internal state (Gurnee et al., 2026). The two raters (DeepSeek V4.1 Flash without reasoning and GLM 5.3) received ten lists at a time, each with ten candidate definitions, the construct's

own and nine random others in random order. For each list, they chose the definition that the list conveyed and the three most likely ones. We report the raters' choices for the lists of the four middle layers of each LLM and, for the Jacobian lens, of the three middle layers. For the comparison with word matching (Supplemental Note 2), we chose for each list, using the lists of all layers, the definition whose narratives shared the most words with it. We report the mean over the two raters, who agreed on whether a list conveyed the correct definition for 84% to 85% of the lists (Cohen's $\kappa = .68$).

Inducing and removing a representation

We induced a representation by adding it to the internal state of the word fragments of the narrative or the scenario description at one layer, in doses of ±1, ±2, and ±4 times the average difference that the construct's own pairs produce along it. In Studies 4 and 5, we induced the representations in an early layer and in a layer about 50% of the way through the LLM. The early layer was the first layer at least 10% of the way through the LLM at which the identification rate reached 95% of its level at the LLM's best layer. Because Study 3 found that the responses were related to the activation at points of greater input proximity to response production, we also induced the representations, for the responses, in a layer about 70% of the way through the LLM, at every word fragment from the beginning of the scenario description to the end of the outcome measure. To test whether an effect was specific to the induced representation, we also induced the representations of other constructs and random directions of the same size, drawn from the span of the observed differences.

We removed a representation by replacing the projection of the internal state onto ten directions with its mean, at every word fragment and in the layers at which the constructs were identified best. These were the ten directions that capture most of the variation in the construct's pair differences (the first ten right singular vectors of the matrix of these differences at the layer), identified after we had excluded the three directions that the differences of all 66 constructs share (the first three right singular vectors of the matrix of the 66 constructs' mean differences) and that therefore do not distinguish one construct from another. To test whether a reduction was specific to the construct, we also removed the ten directions of other constructs and ten random directions. To test whether the number of directions mattered, we also removed 30 and 100 directions in Qwen3.6 27B. Removing 30 directions gave similar results (65% of the effect left,

compared with 91% for random directions), whereas with 100 directions, random directions also removed most of the effect (Table S6).

Structure across LLMs

To test whether the LLMs share the relations among the constructs (Study 2), we computed the cosine similarity of the representations of every two constructs, estimated from complementary halves of the writers and averaged over the ten splits, and correlated the 2,145 similarities between LLMs. To remove the part of these similarities that the words of the narratives explain, we used the residuals of each LLM's similarities on the similarities computed from word counts. To remove the part that the sentence encoder explains, we used the residuals on the encoder's similarities, alone or together with those computed from the words. We compared the mean similarity within and between processes with 5,000 permutations of the process labels.

To compare the representations themselves across LLMs, we used the observations of the three main LLMs in the layers 30% to 50% of the way through the LLM. We centered each LLM's observations on their mean over all pairs and reduced them to their first 256 principal components, which held 58% to 60% of their variance. The shared dimensions were the first 16 principal components of the three LLMs' reduced observations taken together, after we scaled each LLM to the same total variance. We computed these dimensions without the construct labels, and they held 31% of the joint variance. We mapped each LLM's reduced observations onto the shared dimensions by ridge regression, which we estimated for each construct from the pairs of the other 65 constructs only, so that any agreement between the LLMs in how they represent this construct comes from the LLMs themselves rather than from the mapping. A construct's profile for one writer in one LLM was the mean of the mapped observations of the pairs that the writer composed for the construct, scaled to unit length. We assigned each of these profiles to the most similar of the 66 mean profiles from the other five writers in the other two LLMs (cosine similarity).

Criterion validity

To test criterion validity (Study 1), each main LLM, instructed with one of the agent profiles, rated both narratives of 395 pairs, at most one pair for each construct and writer, on the first-person item for the pair's construct and, for comparison, on ten other items on average.

Individual characteristics

For the seven characteristics of Study 5, the six writers composed 611 pairs of narratives that conveyed each characteristic through habits and possessions, while holding constant how well-off, how pleasant, and how sociable the person appeared (Table S8). We estimated each characteristic's representation as in Study 1. To separate it from how well-off, pleasant, and sociable the person appeared, we also estimated it after removing the parts that covary with the rated wealth, pleasantness, and sociability of the narratives. In the experiments, we described each characteristic in the agent profile with the statements used in the earlier work (J. Chen et al., 2026). To test whether the characteristic's representation, rather than its description, produces heterogeneity, we also induced the representation at the word fragments of the agent profile. To test whether the heterogeneity depends on the representation, we removed it, one direction per layer at each recorded layer between 10% and 85% of the way through the LLM, either at the word fragments of the statement that described the characteristic or from this statement onward, and, for comparison, removed random directions and the directions of two other characteristics.

Analysis plan

Before measuring the internal states of LLMs analyzed in this article, we specified in a dated analysis plan the exclusion rules for the narratives; the use of layers 30% to 50% of the way through each LLM; the estimation of each representation from five writers and its test on the sixth; the tests in Studies 1 to 3 at these layers; and the tests in Study 4 at layers selected according to rules stated in the plan. We subsequently specified comparisons of representations between LLMs in shared dimensions (Study 2); analyses at additional points of the input, such as the experiment's outcome measure, and at additional layers, together with coding of how directly the experiments' outcome measures refer to the proposed mediator (Study 3); the LLMs' judgments of the mediator after the scenario description, inductions at later layers, and removals along more than ten directions (Study 4); Study 5; analyses of one LLM over the course of its training (Supplemental Note 7); and equivalence tests and Bayes factors. The dated analysis plan is in the folder analysis_plan of the GitHub repository, with a record of when each of its parts was written.

Statistics

We ran the analyses in Python 3.13 (Python Software Foundation, 2024) with PyTorch (Paszke et al., 2019), NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), pandas (McKinney, 2010), scikit-learn (Pedregosa et al., 2011), and Matplotlib (Hunter, 2007); the repository lists the version of each package. We excluded the judgments and responses whose digit probabilities summed to less than .5, at most 0.01% of the responses in Study 3 in any LLM. We tested the criterion correlation of Study 1 with permutation tests in which we shuffled the items within each pair (2,000 permutations), and the grouping of the constructs by process in Study 2 with permutation tests in which we shuffled the constructs' process labels (5,000 permutations). We also report the numbers of constructs, judgments, or experiments with a positive or significant effect, which we tested with exact one-sided binomial tests against the chance rate stated in the text. Across experiments, we computed Spearman correlations. In Study 3, we corrected these correlations across points of the input and layers for multiple comparisons with a max-statistic permutation test (5,000 permutations), which compares each correlation with the largest correlation produced by chance across these points and layers. With 60 experiments, a one-sided test of a rank correlation across experiments at $\alpha = .05$ had a power of .80 to detect a correlation of .32. We used equivalence tests with bounds of ±.20 to assess whether correlations were negligible. The Bayes factors (BF01) used a uniform prior on the correlation or, for counts of experiments, a rate uniformly distributed between .5 and 1, against the chance rate of .5. To estimate the reliability of the measures in Study 3, we split the seven sets of scenario descriptions into all 35 combinations of three and four sets and corrected the split-half correlations with the Spearman-Brown formula. For the comparison by replication outcome, we used one-sided Fisher's exact tests, as in prior work on these experiments (J. Chen et al., 2026), separately for the activation of the representations and for the responses, and one-sided Mann-Whitney tests to compare the activation of the representations.

Supplemental Note 2. Study 1 in Detail

At each LLM's best layer, the identification rate was 81% to 93% (Table S2), nearly the same as at the layers 30% to 50% of the way through the LLM (81% to 92%), which we used in Studies 1 to 3. Identification thus hardly depended on the choice of layers. To test whether the representations distinguish similar constructs, we also identified the held-out writer's construct

among only the constructs that arise in the same process according to their theories. The main LLMs identified it in 95% to 96% of the pairs, where guessing would succeed for 10%, and they distinguished the eight mediators that have a rival construct from their rivals in 98% to 99% of the pairs (Table S3). Of the 23% of pairs whose construct the words failed to identify, the main LLMs identified 76%, and of the 20% that the sentence encoder failed to identify, 67% to 70%. We tested the LLM before training on a third of the pairs (2,853).

The raters chose the correct definition for 52% [44%, 61%] of the read-out lists in Qwen3.6 27B and for 61% [52%, 70%] in OLMo 3.1 32B. In the three largest Gemma LLMs, however, they chose it for only 10% to 12% of the lists, where guessing would succeed for 10%. With the Jacobian lens, which takes into account how the later layers of the LLM transform the internal state (Gurnee et al., 2026), they chose the correct definition for 36% [28%, 43%] of the lists in Gemma 4 31B. To test whether the raters' choices could arise from matching read-out words to words in the narratives, we also chose for each list the definition whose narratives contained the most words of the list. For the lists on which this word matching chose a wrong definition, the raters still chose the correct definition for 28% to 53% of the lists in the main LLMs (Table S4). The raters thus identified the constructs' meaning beyond the words that the lists shared with the narratives.

Supplemental Note 3. Study 2 in Detail

In the test of whether the representations group by appraisal process (Study 2), the words of the narratives showed almost none of the difference in similarity between constructs of the same process and constructs of different processes (.004), whereas the sentence encoder showed a difference of .023 (.071 vs .048), close to the .026 of the main LLMs (Table S3). The meaning of the narratives thus conveys much of this structure.

Within one family of LLMs, Qwen3.5, the representations became more distinct with size. Identification of the held-out writer's construct at each LLM's best layer rose from 81% with 0.8 billion parameters to 93% with 27 billion, and the raters' identification of the meaning from the read-out words rose from 40% to 54% (Table S2). The representations were also present before instruction tuning. The two LLMs that we examined before instruction tuning identified the construct at their best layer in 86% and 91% of the pairs (Table S2). The representations thus do not depend on learning to respond to requests.

Supplemental Note 4. Study 3 in Detail

To test whether the absence of a correlation across experiments in Study 3 arose from unreliable measures, we computed the reliability of both measures in the main LLMs. The manipulation's activation of the mediator's representation during the scenario description and its effect on the responses were reliable across the seven sets of scenario descriptions (split-half reliability, .96 to .98; Supplemental Note 1), and their correlation, corrected for attenuation, was near zero (Pearson $r = -.03$ to .00).

In the moral typecasting experiment, the representation of moral agency was more strongly activated in the main LLMs when the man had pushed the tray than when the baby had ($d_z = 1.45$ to 2.22). Among the 66 representations, the mediator's representation was the most strongly activated in 5 or 6 experiments and among the five most strongly activated in 13 to 20, where chance would produce about 1 and 4.5 (binomial $p \leq .003$). Inside the LLMs, a manipulation thus activated the representation of its mediator together with the representations of several related constructs.

Of the 28 experiments that large preregistered replications have re-examined, 14 showed the original effect in human participants (Bago et al., 2022; Camerer et al., 2018; Ebersole et al., 2016; Eerland et al., 2016; Klein et al., 2022, 2014, 2018; Lynott et al., 2014). Inside the LLMs, the manipulation activated the mediator's representation in 9 to 13 of the 14 experiments whose effects had replicated and in 8 to 12 of the 14 whose effects had not (one-sided Fisher's exact $p = .104$ to .918), with median $d_z = 0.98$ to 2.39 and 0.95 to 2.16 (one-sided Mann-Whitney $p < .05$ in 1 of the 14 LLMs). The LLMs' responses, however, showed the original effect more often where it had replicated (8 to 12) than where it had not (3 to 8; $p = .003$ to .353, below .05 in 3 of the 14 LLMs; Figure 5b). The responses thus distinguished the effects that had replicated from those that had failed to replicate, whereas the activation of the mediator's representation did not.

To examine whether the activation of the mediator's representation at the outcome measure depended on how directly the outcome measure refers to the mediator, two LLMs independently coded for each of the 60 experiments whether its outcome measure asks about the mediator itself (2), makes the mediator salient (1), or does not evoke it (0). They agreed on 87% of the experiments (weighted $\kappa = .78$). Across the experiments, the mean code correlated with the activation of the mediator's representation while the main LLMs processed the outcome measure

at ρ = .18 (95% CI [−.08, .42]) and with the manipulation's effect on the responses at ρ = .15 (95% CI [−.10, .40]). How directly an outcome measure referred to the mediator was thus not significantly related to either measure.

To test whether responses at the end of the response scale affected the correlations of Study 3, we excluded in each LLM the experiments whose control condition sat, in more than 30% of the presentations, at the end of the scale toward which the predicted effect moves (6 experiments in Gemma 4 31B and 2 in OLMo 3.1 32B). The correlations between the activation at the outcome measure and the effect on the response were then .43 in Gemma 4 31B and .14 in OLMo 3.1 32B, and those at the scenario description .13 and .09 (Table S10). Without these experiments, the correlations thus remained positive at the outcome measure and small at the scenario description.

Supplemental Note 5. Study 4 in Detail

In the narratives of Study 1, across all combinations of constructs and judgments, the effects of the induced representations on the LLM's judgments correlated with the effects of the narratives themselves (r = .31 to .56 in the 11 LLMs tested; Table S6). Inducing a representation thus affected the judgments much as the narratives did.

Inducing the mediator's representation about halfway through the LLM while the LLM processed the scenario description moved the responses in the predicted direction in 30 to 36 of the 52 experiments in the main LLMs (chance, 26; binomial p = .004 to .166; Figure 7b; Table S6).

In OLMo 3.1 32B, the third main LLM, removing the mediator's representation left 99% (95% CI [79%, 107%]) of the manipulation's effect on the responses (Table S6). In Gemma 4 31B, removing the representations of the constructs most similar to the mediator reduced the manipulation's effect on the response to 17%, in line with the finding that a manipulation activates the representations of several related constructs (Supplemental Note 4).

The theories further specify that a situational moderator strengthens or weakens the mediator's effect on the response. In the seven experiments with a situational moderator, we induced the mediator's representation at both levels of the moderator. The interaction was in the

predicted direction in 2 to 5 of the 7 experiments, depending on the LLM and the layer, and significant interactions occurred in both directions (Table S7).

Supplemental Note 6. Study 5 in Detail

Study 5 used six tests of individual characteristics from the earlier work (J. Chen et al., 2026) and a seventh test of self-monitoring, for which the theories predict no heterogeneity. In each of 13 LLMs, the representations estimated from the writers' pairs of narratives identified the held-out writer's characteristic among the seven in 87% to 97% of the pairs, where guessing would succeed for 14% (Table S8).

Describing a characteristic produced the heterogeneity that the theories predict in all four LLMs for two characteristics, the belief that higher costs justify higher prices in the beer-price experiment and the preference for consistency in the foot-in-the-door experiment. These two characteristics also produced the most consistent heterogeneity across the 13 LLMs (Figure 8a; Table S8). The other four characteristics produced the predicted heterogeneity in at most one of the four LLMs, and the importance of family to the self produced heterogeneity in the opposite direction in three of them. Self-monitoring produced no significant interaction in three of the four LLMs.

To test whether the representation, rather than the description, produces heterogeneity, we induced a characteristic's representation instead of describing the characteristic. This induction produced the predicted heterogeneity for one of the six characteristics, the preference for consistency, in two of the four LLMs, whereas random directions of the same size produced significant heterogeneity in 10 of 56 tests (Table S8). To test whether the heterogeneity depends on the representation, we removed it. For the two characteristics that had produced the predicted heterogeneity in all four LLMs, removing the characteristic's representation from the internal states for the statement that described the characteristic reduced this heterogeneity to 11% and 34% of its size in Gemma 4 26B. This reduction was larger than the one produced by removing random directions, but removing the representations of other characteristics also reduced the heterogeneity. In Qwen3.6 27B, the heterogeneity produced by the belief that higher costs justify higher prices remained (95% left), as it did when random directions were removed, whereas the heterogeneity produced by the preference for consistency was reduced to 53%, compared with 94% and 59% for the two sets of random directions (Figure 8b; Table S8).

Supplemental Note 7. Representations Over the Course of Training

To understand why the LLMs represented the mediators but did not base their responses on the activation of these representations during the scenario description, we examined how the representations arise. If LLMs acquire them from human language, the representations should form while an LLM learns to predict text. We therefore examined OLMo 3 7B, whose intermediate stages of training are public (Team Olmo et al., 2025), at successive stages of its training. After 10,000 of the 1.41 million updates of the first stage of training, in which the LLM learns to predict text, the representations identified the held-out writer's construct in 79% of the pairs, which is 90% of the rate that the LLM reached before instruction tuning. After only 3,000 to 10,000 updates, the raters also identified the constructs' meaning from the read-out words as often as they did before instruction tuning (Figure S1a).

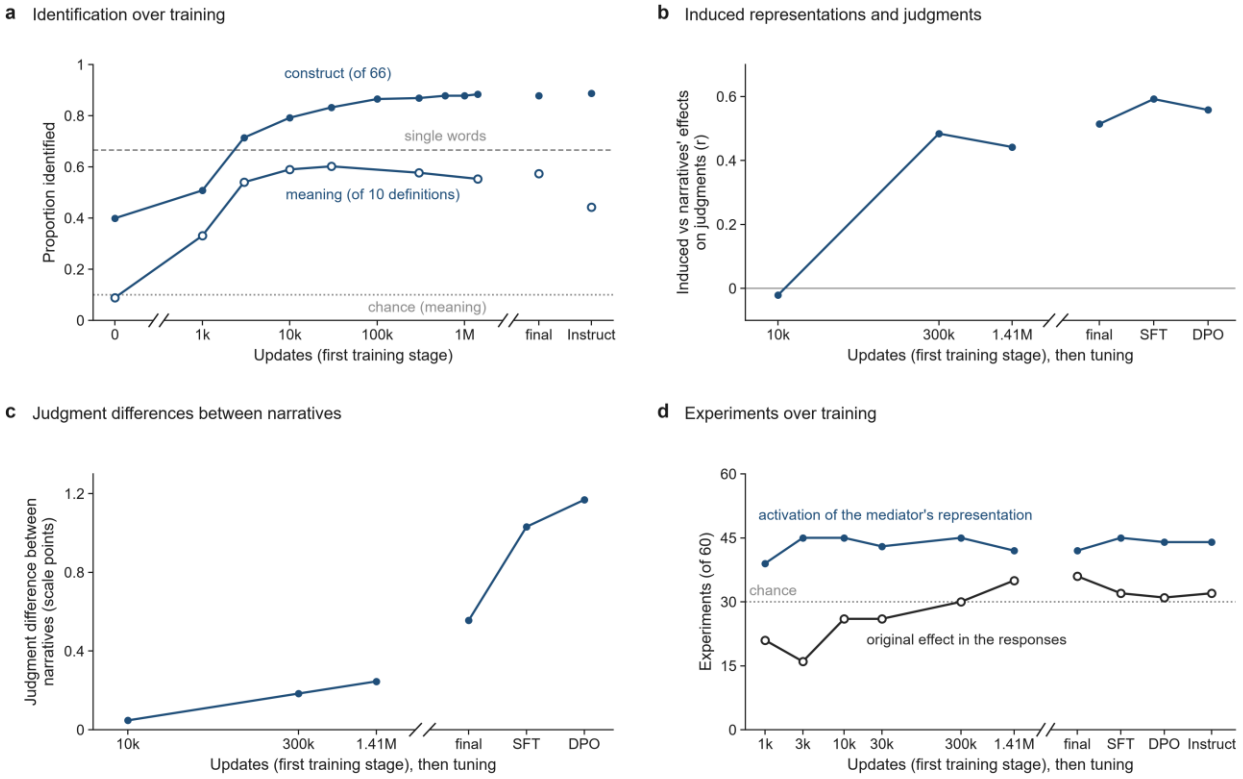


Fig. S1.

Representations over the training of OLMo 3 7B. In (a), the identification of the constructs among the 66 (at layer 16 on one third of the pairs; for the released instruction-tuned large language model [LLM], at its best middle layer on all pairs) and the identification of their meaning among ten definitions are shown over the first training stage, at the end of the training

stages before instruction tuning, and after instruction tuning. In (b), the correlation between the effects of the induced representations and the effects of the narratives on the LLM's judgments is shown. In (c), the mean difference between the LLM's judgments of the construct in the high and the low narratives is shown. In (d), the number of the 60 experiments in which the manipulation activated the mediator's representation more than the other constructs' representations (navy; the dotted line shows chance) and in which the LLM's responses showed a significant original effect (black) is shown over the same stages. On the horizontal axes, "updates" are training steps of the first stage; "final" is the LLM at the end of the training stages before instruction tuning; SFT (supervised fine-tuning) and DPO (direct preference optimization) are the two successive steps of instruction tuning; and "Instruct" is the released instruction-tuned LLM.

The representations, however, affected the LLM's judgments only much later. At 10,000 updates, the effects of induced representations on the LLM's judgments correlated with the effects of the narratives themselves at $r = -.02$, whereas at 300,000 updates the two sets of effects correlated at $r = .48$ (Figure S1b). From 10,000 updates to the end of preference tuning, a part of instruction tuning, the difference between the LLM's judgments of a construct in its high and its low narratives grew from 0.05 to 1.17 scale points (Figure S1c; Table S2), whereas identification remained nearly constant after 100,000 updates. Instruction tuning lowered how often the raters identified a construct's meaning from its read-out words (from 57% to 44% of the lists). The LLM thus formed the representations long before its judgments depended on them.

In the published experiments, too, the representations were activated long before the responses showed the effects. After only 1,000 updates, the manipulations activated the mediator's representation more strongly than the other constructs' representations in 39 of the 60 experiments, and from 3,000 updates on, in 42 to 45 (Figure S1d; Table S2). The responses, in contrast, showed a significant original effect in only 16 to 26 experiments during the first 30,000 updates and in 35 to 36 at the end of the stages before instruction tuning. They distinguished the effects that had replicated from those that had failed only after instruction tuning (11 vs 3 or 4 of 14; $p = .004$ to .011, compared with $p \geq .339$ before). At every stage, the correlation across experiments between the manipulation's activation of the mediator's representation while the LLM processed the scenario description and the manipulation's effect on the responses lay between $\rho = -.13$ and .07 (Spearman).

Table S1

The 66 Constructs, Their Roles and Processes, and the Retained Pairs of Narratives

Construct	Role (experiment)	Process	Definition (rating scale)	Pairs retained	Per writer (Cl / DS / Gr / Ki / MM / Mu)	Low narrative: opposite / absence of the state
Associations with Republicanism	Mediator: Flag priming	categorization	The person links what is in front of them with a conservative political party and its values (tradition, lower taxes, a smaller state, law and order).	135	16 / 24 / 24 / 23 / 24 / 24	72 / 63
Weight of the subject's features†	Mediator: Similarity asymmetry	categorization	The person, comparing one thing (the one being assessed) with another, finds the main characteristics of the first in the second.	142	24 / 24 / 24 / 24 / 22 / 24	71 / 71
Salience of heat	Mediator: Heat priming	categorization	The person's thoughts are taken up with heat: hot surfaces, steam, blazing sun, sweating, things cooking or burning.	141	23 / 24 / 24 / 22 / 24 / 24	72 / 69
Anthropomorphism	Mediator: Anthropomorphized products	categorization	The person sees an object as having come alive, like a person with a face and a mind of its own.	132	24 / 24 / 24 / 19 / 17 / 24	69 / 63
Small ongoing expenses (petty cash)	Mediator: Pennies a day	categorization	The person treats a payment as one of the small, routine purchases of ordinary life rather than as a large, occasional outlay.	119	13 / 23 / 24 / 21 / 17 / 21	63 / 56
Capability	Mediator: Feature fatigue	categorization	The person sees a product or tool as able to do a great many of the things they might want from it.	133	24 / 24 / 24 / 24 / 13 / 24	61 / 72
Perceived fit	Mediator: Brand extension	categorization	The person sees a new product or service from a maker as fitting what that maker already makes or does: the maker's skills, tools or materials carry over to it, or it goes with the maker's existing line.	137	24 / 24 / 24 / 24 / 17 / 24	71 / 66
Evaluability	Mediator: Less is better	categorization	The person sees an item's price as near the top of what items of its kind cost where it is sold.	137	24 / 23 / 24 / 24 / 18 / 24	67 / 70
Severity of forbidding†	Mediator: Forbid vs allow	categorization	The person sees a measure that stops other people from doing something as a harsh, drastic step.	128	23 / 24 / 24 / 20 / 13 / 24	69 / 59
Anchor	Mediator: Anchoring; Incidental anchors	categorization	The person's estimate starts from a specific figure that is at hand and moves away from it only a little.	139	24 / 24 / 23 / 22 / 23 / 23	71 / 68
Detail of imagined actions	Mediator: Verb aspect and intent	categorization	The person pictures an action as it happens, vividly and moment by moment, in clear sensory detail.	111	24 / 24 / 24 / 5 / 10 / 24	55 / 56
Usability	Rival: Feature fatigue	categorization	The person expects a product or tool to be easy to learn and to operate.	84	16 / 16 / 16 / 14 / 9 / 13	46 / 38
Perceived autonomy	Mediator: Nonconforming dress	social inference	The person sees another person as free to do as that person pleases: able to follow their own preferences and in control of their own choices.	134	24 / 24 / 24 / 23 / 15 / 24	71 / 63
Concern about misidentification	Mediator: Identity divergence	social inference	The person worries that something they use, wear or do could lead others to take them for a member of a group they do not belong to.	141	24 / 24 / 24 / 22 / 24 / 23	70 / 71
Status motives	Mediator: Going green	social inference	The person wants to gain standing, prestige or respect relative to their peers right now.	142	24 / 24 / 24 / 23 / 23 / 24	70 / 72

Construct	Role (experiment)	Process	Definition (rating scale)	Pairs retained	Per writer (Cl / DS / Gr / Ki / MM / Mu)	Low narrative: opposite / absence of the state
Frame of reference	Mediator: Response-scale range	social inference	The person takes the amounts on offer as a sign of how much of something people usually have or do, and concludes that a large amount is usual.	123	24 / 24 / 24 / 15 / 13 / 23	67 / 56
Moral agency	Mediator: Moral typecasting	social inference	The person sees another person who acts as able to plan and control their own actions and to tell right from wrong.	133	24 / 24 / 23 / 17 / 21 / 24	69 / 64
Reciprocity	Mediator: Door in the face; Norm of reciprocity	social inference	The person's sense that they owe a requester a yes, or should meet them halfway, because the requester first gave ground or did something for them.	131	23 / 24 / 22 / 20 / 23 / 19	64 / 67
Others seen as partners	Mediator: Consumer label and trust	social inference	The person sees the others who share a limited common resource with them as partners in a joint effort to look after it, rather than as competitors each out for their own use.	140	24 / 24 / 24 / 21 / 24 / 23	70 / 70
Vertical location as power	Mediator: Vertical position and power	social inference	The person takes another person's place physically above others (on a stage, an upper floor, the top of a display) as a sign of that person's power.	141	24 / 24 / 24 / 23 / 22 / 24	70 / 71
Prestige	Mediator: Quote attribution	social inference	The person takes the standing of whoever said some words as a reason to agree with them, so that the same words seem more worth accepting because of who said them.	131	24 / 24 / 24 / 20 / 18 / 21	69 / 62
Descriptive norm	Mediator: Normative message	social inference	The person sees their own use of something being lower than other people's use of it as leaving room to use more of it themselves.	139	24 / 24 / 24 / 24 / 19 / 24	72 / 67
Desire for affiliation	Mediator: Social exclusion	social inference	The person wants a particular other person they are dealing with to like and accept them.	132	24 / 24 / 23 / 18 / 20 / 23	66 / 66
Need for uniqueness	Rival: Identity divergence	social inference	The person wants to stand out and be different from other people in general.	91	16 / 16 / 16 / 16 / 13 / 14	45 / 46
Environmental concern	Rival: Going green	social inference	The person cares about protecting the natural environment.	89	15 / 16 / 14 / 15 / 14 / 15	41 / 48
Successful completion	Mediator: IKEA effect	self-relevance	The person feels capable and accomplished because their own work has successfully made, fixed or completed something.	135	24 / 23 / 24 / 22 / 18 / 24	67 / 68
Identity as a resource	Mediator: Identity threat	self-relevance	The person draws strength and a sense of being one of them from a social group they are a member of, rather than wanting distance from it.	136	24 / 23 / 24 / 24 / 18 / 23	69 / 67
Self-view confidence	Mediator: Shaken self-view	self-relevance	The person is sure that they are a competent, capable person, rather than unsure of it.	126	24 / 24 / 24 / 18 / 12 / 24	69 / 57
Future self-continuity	Mediator: Future self and saving	self-relevance	The person sees themself years from now as the same person they are today, closely similar to and continuous with their present self.	125	24 / 24 / 23 / 21 / 10 / 23	63 / 62
Worldview defense	Mediator: Mortality salience	self-relevance	The person feels a need to stand up for the values, customs and way of life of their own country or culture.	138	24 / 24 / 24 / 20 / 22 / 24	70 / 68
Identity salience	Mediator: Identity salience	self-relevance	The person sees themself, right now, mainly as part of a close group whose members share their daily life and decisions, rather than as a separate individual.	111	17 / 24 / 23 / 16 / 9 / 22	61 / 50

Construct	Role (experiment)	Process	Definition (rating scale)	Pairs retained	Per writer (Cl / DS / Gr / Ki / MM / Mu)	Low narrative: opposite / absence of the state
Personal connection	Mediator: Time versus money	self-relevance	The person feels personally connected to a product or to the moment of having it: it holds meaning for them and fits who they are.	123	24 / 23 / 24 / 19 / 9 / 24	68 / 55
Psychological ownership	Mediator: Emotional attachment; Mere touch; Possession and valuation	self-relevance	The person feels that an object is theirs, as if it were already their own.	140	24 / 24 / 24 / 20 / 24 / 24	72 / 68
Self-perception as helpful	Mediator: Foot in the door; Labeling	self-relevance	The person sees themself as the kind of person who takes action for good causes and helps people in need.	128	24 / 24 / 23 / 23 / 20 / 14	63 / 65
Identity importance	Rival: Identity threat	self-relevance	The person sees a social category they belong to (people who share a trait, background or pursuit) as important to how they see themself.	85	16 / 15 / 13 / 9 / 16 / 16	43 / 42
Analytic thinking	Mediator: Analytic thinking and belief	affective response	The person is thinking deliberately and analytically, working things through step by step and checking first impressions, rather than going with an immediate gut answer.	133	24 / 24 / 24 / 23 / 14 / 24	70 / 63
Accessibility of the negative outcome	Mediator: Tempting fate	affective response	The person readily and vividly pictures a specific bad outcome befalling them in the narrative, and how bad that moment would be.	139	23 / 24 / 24 / 24 / 20 / 24	71 / 68
Emotional response to harm†	Mediator: Personal force	affective response	The person reacts with horror, revulsion or distress to an action that another person carries out, rather than taking it in without any such reaction.	139	24 / 23 / 24 / 23 / 22 / 23	69 / 70
Goal fluency	Mediator: Goal fluency	affective response	The person finds something they encounter easy to take in and make sense of: it registers at once rather than slowly or with effort.	125	24 / 23 / 24 / 14 / 16 / 24	69 / 56
Affect richness	Mediator: Affect-rich prizes	affective response	The person is excited and hopeful about a reward on offer that is a vivid, personal experience, more than about a plain alternative of similar money value.	139	24 / 24 / 24 / 21 / 22 / 24	69 / 70
Physiological arousal	Mediator: Emotion and sharing	affective response	The person is in a state of high bodily and mental activation, alert, keyed up and full of energy, as opposed to passive, mellow and low in energy.	143	24 / 23 / 24 / 24 / 24 / 24	71 / 72
Reliance on affect	Mediator: Affect reliance	affective response	The person treats the feelings that something evokes in them (how it looks, sounds or strikes them) as relevant information for judging it.	140	23 / 24 / 24 / 23 / 22 / 24	69 / 71
Interpersonal warmth	Mediator: Physical warmth	affective response	The person feels warm, caring and generous toward other people, those around them or in their life.	139	24 / 24 / 24 / 22 / 21 / 24	69 / 70
Reliance on substance	Rival: Affect reliance	affective response	The person treats the stated factual points about something (what it does, what it is made of, how it performs) as the relevant basis for judging it.	82	16 / 16 / 16 / 7 / 11 / 16	42 / 40
Desire for power	Mediator: Powerlessness	goal relevance	The person wants more power, control or say over others or over outcomes that concern them than they now have.	133	22 / 23 / 24 / 21 / 19 / 24	69 / 64
Focus on the forgone	Mediator: Buying and selling prices	goal relevance	Experience focus: in a decision that sets a thing against money, the person dwells on what having or doing the thing itself would be like (what they would have or go without) rather than on the money and what else it could buy.	132	24 / 23 / 23 / 24 / 14 / 24	70 / 62

Construct	Role (experiment)	Process	Definition (rating scale)	Pairs retained	Per writer (Cl / DS / Gr / Ki / MM / Mu)	Low narrative: opposite / absence of the state
Overhead aversion	Mediator: Overhead aversion	goal relevance	The person sees that what they would give to a cause would go in full to the purpose itself rather than being partly taken up by costs along the way (fees, staff, rent, administration).	126	24 / 22 / 24 / 19 / 15 / 22	68 / 58
Loss aversion	Mediator: Endowment effect	goal relevance	The person treats parting with, or going without, something as losing something they have, rather than as simply not gaining something.	142	24 / 23 / 24 / 23 / 24 / 24	71 / 71
Perceived goal progress	Mediator: Goal progress	goal relevance	The person sees themselves as having come a long way toward a personal aim and as close to reaching it.	133	24 / 23 / 24 / 22 / 17 / 23	70 / 63
Gain frame†	Mediator: Disease framing	goal relevance	The person counts an outcome as what will be saved or kept, set against a starting point at which none of it would remain, rather than as what will be taken away, set against a starting point at which all of it is present.	125	24 / 24 / 24 / 17 / 12 / 24	70 / 55
Importance	Mediator: Weight and importance	goal relevance	The person sees a decision or matter that concerns them as weighty and serious, with a great deal riding on it, rather than as a minor matter.	142	24 / 24 / 24 / 23 / 23 / 24	71 / 71
Rehearsal of past payments	Mediator: Payment transparency	goal relevance	The person sees that they have spent little in the current period (a month, a week, a trip) and still have money left over.	135	21 / 24 / 22 / 20 / 24 / 24	67 / 68
Weight of positive features	Mediator: Choosing vs rejecting	goal relevance	Strength weighting: in judging between options, the person gives more weight to the options' strengths than to their drawbacks.	136	23 / 24 / 24 / 24 / 17 / 24	67 / 69
Blame†	Mediator: Side effects and intent	causal inference	The person holds another person to blame for a bad outcome, seeing it as that person's own doing.	129	24 / 23 / 24 / 19 / 15 / 24	70 / 59
Expectation that actions produce outcomes†	Mediator: Structure and goal pursuit	causal inference	The person is confident that their own actions reliably produce the results they intend, so that doing the same thing in the same way brings the same result.	140	24 / 22 / 24 / 23 / 23 / 24	70 / 70
Earned right to indulge	Mediator: Effort and indulgence	causal inference	The person feels that the work they have put in gives them the right to an indulgent pleasure, so that choosing it would be well deserved and free of guilt.	135	24 / 24 / 24 / 24 / 15 / 24	71 / 64
Reference price	Mediator: Beer price	causal inference	The person thinks that an item normally costs, and may fairly cost, a lot at the particular place that sells it, so that a high price there would be fair.	137	24 / 23 / 24 / 22 / 20 / 24	71 / 66
System justification	Mediator: Currency priming	causal inference	The person sees an open system of trade and paid work as fair, in that what each participant ends up with reflects what they themselves put in and so what they deserve.	128	18 / 24 / 24 / 18 / 21 / 23	68 / 60
Inferred importance	Mediator: Decision difficulty	causal inference	The person sees getting a particular choice right as important, because which option they pick will make a real and lasting difference.	139	24 / 24 / 24 / 22 / 22 / 23	70 / 69
Inferred length of the sequence	Mediator: Retrospective gambler's fallacy	causal inference	The person sees an outcome they have just witnessed as one that comes up only after many attempts, rather than one that could come up on any single attempt.	136	23 / 23 / 24 / 22 / 21 / 23	69 / 67
Uncertainty	Rival: Decision difficulty	causal inference	The person is unsure which of the options in front of them is the better one.	91	16 / 16 / 16 / 15 / 14 / 14	48 / 43

Construct	Role (experiment)	Process	Definition (rating scale)	Pairs retained	Per writer (Cl / DS / Gr / Ki / MM / Mu)	Low narrative: opposite / absence of the state
Appeal of the setting	General evaluation: Mere touch; Foot in the door	comparison of alternatives	How favorably an incidental part of the surroundings is described (a window, a lamp, a path, a room).	89	16 / 15 / 16 / 14 / 13 / 15	0 / 89
Affective reaction	Rival: Emotional attachment	comparison of alternatives	The person's emotional reaction to an object itself, such as being interested, delighted or enthusiastic about it.	94	16 / 16 / 16 / 16 / 15 / 15	0 / 94
Favorable evaluation	General evaluation: Mere touch; Door in the face; Labeling	comparison of alternatives	Favorable evaluation of the object, the cause or the request itself: its look, quality, usefulness or merit.	90	16 / 16 / 16 / 14 / 13 / 15	0 / 90
Pleasantness	General evaluation: Mere touch	comparison of alternatives	Pleasantness of the person's feeling.	90	16 / 13 / 16 / 16 / 14 / 15	0 / 90
Perceptual contrast	Rival: Door in the face	comparison of alternatives	How small or easy a request is judged to be, especially next to a bigger commitment.	78	16 / 16 / 16 / 15 / 0 / 15	0 / 78
All 66 constructs	55 mediators, 8 rivals, 3 general evaluations			8,311	1451 / 1472 / 1480 / 1305 / 1157 / 1446	4035 / 4276

Note. Construct: the name used in the original article (†, the article names no mediator, and the name is taken from the article's explanation of the effect). Role: the experiment or experiments whose mediator the construct is (55 distinct mediators for the 60 experiments; five mediators serve two or three experiments), a rival construct that the original article had ruled out (the experiment is named), or a general evaluation. Process: the process in which the construct arises according to its theory (categorization 12, social inference 13, self-relevance 10, affective response 9, goal relevance 9, causal inference 8, and comparison of alternatives 5, which includes the three general evaluations and two rival constructs). Definition: the definition of the rating scale from which the writers composed the pairs of narratives. Pairs retained: pairs left after the exclusions described in the Method, in total and per writer (Cl = Claude Sonnet, DS = DeepSeek V4.1 Flash, Gr = Grok 4.7, Ki = Kimi K3, MM = MiniMax M3, Mu = Muse Spark 1.3). Low narrative: number of pairs whose low narrative conveys the opposite state (scale point 1) or the absence of the state (scale point 4).

Table S2

The LLMs Examined, Their Identification of the Constructs, and the Analyses in Which They Entered

Panel A. LLMs.

LLM	Hugging Face id	Developer	Parameters	Layers	Instruction-tuned	Identification at the best layer (layer)	Identification at the early layer (layer)	Meaning identified (read-out)	Tables with the LLM's results
Qwen3.6 27B	`Qwen/Qwen3.6-27B`	Alibaba (Qwen)	27B	64	yes	92% (40)	89% (19)	52%; Jacobian lens 48%	S3 (identification, MTMM, baselines, criterion, Study 2 structure); S4 (logit lens, Jacobian lens); S5; S6 (judgments, responses at three depths, removal); S7; S8
Gemma 4 31B	`google/gemma-4-31B-it`	Google	31B	60	yes	92% (36)	90% (24)	10%; Jacobian lens 36%	S3 (identification, MTMM, baselines, criterion, Study 2 structure); S4 (logit lens, Jacobian lens); S5; S6 (judgments, responses at three depths, removal); S7; S8
OLMo 3.1 32B	`allenai/Olmo-3.1-32B-Instruct`	Ai2	32B	64	yes	92% (23)	89% (13)	61%	S3 (identification, MTMM, baselines, criterion, Study 2 structure); S4 (logit lens); S5; S6 (judgments, responses at three depths, removal); S7; S8
Gemma 4 26B	`google/gemma-4-26B-A4B-it`	Google	26B (A4B)	30	yes	91% (14)	87% (9)	12%	S3; S4 (logit lens); S5; S6 (judgments, responses at three depths, removal by size); S7; S8
Qwen3.5 27B	`Qwen/Qwen3.5-27B`	Alibaba (Qwen)	27B	64	yes	93% (40)	89% (19)	54%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B
Qwen3.5 9B	`Qwen/Qwen3.5-9B`	Alibaba (Qwen)	9B	32	yes	90% (20)	88% (10)	48%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B

LLM	Hugging Face id	Developer	Parameters	Layers	Instruction-tuned	Identification at the best layer (layer)	Identification at the early layer (layer)	Meaning identified (read-out)	Tables with the LLM's results
Qwen3.5 4B	`Qwen/Qwen3.5-4B`	Alibaba (Qwen)	4B	32	yes	89% (14)	87% (10)	46%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B
Qwen3.5 2B	`Qwen/Qwen3.5-2B`	Alibaba (Qwen)	2B	24	yes	85% (14)	85% (10)	43%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B
Qwen3.5 0.8B	`Qwen/Qwen3.5-0.8B`	Alibaba (Qwen)	0.8B	24	yes	81% (9)	77% (7)	40%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B
Gemma 4 E4B	`google/gemma-4-E4B-it`	Google	E4B	42	yes	86% (17)	85% (13)	40%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B
OLMo 3 7B	`allenai/Olmo-3-7B-Instruct`	Ai2	7B	32	yes	89% (16)	87% (10)	44%	S3; S4 (logit lens); S5; S6 (judgments, short design; responses in 25 experiments); S8A-B
Qwen3.8 27B	`Qwen/Qwen3.8-27B`	Alibaba (Qwen)	27B	64	yes	92% (40)	88% (19)	49%	S3; S4 (logit lens); S5; S6 (responses in 25 experiments)
gpt-oss 20B	`openai/gpt-oss-20b`	OpenAI	20B	24	yes	87% (14)	85% (10)	48%	S3; S4 (logit lens); S5; S6 (responses in 25 experiments); S8A-B
Granite 4.2 30B	`ibm-granite/granite-4.2-30b`	IBM	30B	64	yes	90% (43)	85% (26)	52%	S3; S4 (logit lens); S5; S8A-B
OLMo 3 32B base	`allenai/Olmo-3-1125-32B`	Ai2	32B	64	no	91% (23)	89% (13)	68%	S3; S4 (logit lens)
Gemma 4 31B base	`google/gemma-4-31B`	Google	31B	60	no	86% (21)	80% (6)	12%	S3; S4 (logit lens)

Panel B. OLMo 3 7B over training.

Checkpoint	Stage	Updates	Early layer	Identification, layer 0	Identification, layer 16	Identification, best middle layer	Meaning identified (read-out)	Induced judgments: own > others, of 66 ($t > 2$)	r (induced, natural)	Judgment difference high - low: own item / all 11 items
stage1-step0	1 (text prediction)	0	6	48%	40%	42%	9%	not run	not run	not run

Checkpoint	Stage	Updates	Early layer	Identification, layer 0	Identification, layer 16	Identification, best middle layer	Meaning identified (read-out)	Induced judgments: own > others, of 66 ($t > 2$)	r (induced, natural)	Judgment difference high - low: own item / all 11 items
stage1-step1000	1 (text prediction)	1,000	10	48%	51%	51%	33%	not run	not run	not run
stage1-step3000	1 (text prediction)	3,000	6	55%	71%	72%	54%	not run	not run	not run
stage1-step10000	1 (text prediction)	10,000	10	57%	79%	80%	59%	37 (22)	−.02	0.05 / 0.04
stage1-step30000	1 (text prediction)	30,000	13	60%	83%	83%	60%	not run	not run	not run
stage1-step100000	1 (text prediction)	100,000	13	67%	87%	87%	not judged	not run	not run	not run
stage1-step300000	1 (text prediction)	300,000	10	67%	87%	87%	58%	42 (23)	.48	0.18 / 0.11
stage1-step600000	1 (text prediction)	600,000	10	66%	88%	88%	not judged	not run	not run	not run
stage1-step1000000	1 (text prediction)	1,000,000	10	66%	88%	88%	not judged	not run	not run	not run
stage1-step1413814	1 (text prediction)	1,413,814	10	66%	88%	88%	55%	44 (19)	.44	0.25 / 0.13
stage2-step47684	2 (mid-training)	47,684	10	66%	88%	88%	not judged	not run	not run	not run
stage3-step11921	3	11,921	10	66%	88%	88%	not judged	not run	not run	not run
main	3	11,921	10	66%	88%	88%	57%	51 (32)	.51	0.56 / 0.25
after supervised tuning (Instruct-SFT)	post-training		10	66%	n.a.	87% (2,853 pairs)	not judged	50 (28)	.59	1.03 / 0.50
after preference tuning (Instruct-DPO)	post-training		10	n.a.	n.a.	88%	not judged	52 (30)	.56	1.17 / 0.55
Instruct (= OLMo 3 7B above)	post-training		10	68%	n.a.	89% (all 8,311 pairs)	44% (all pairs)	56 (30)	.56	n.a.

Panel C. Study 3 over the training of OLMo 3 7B.

Checkpoint	Larger effect on mediator than on the other 65 (of 60)	Responses: original effect positive / significant (of 60)	Responses: significant, replicated vs failed (of 14; Fisher's exact p)	Representation: positive, replicated vs failed (of 14)	Spearman ρ, representation vs response
step 1k	39/60	32/60 / 21	7/14 vs 5/14 (p = .352)	10/14 vs 8/14	.01 (p = .962)
step 3k	45/60	22/60 / 16	5/14 vs 3/14 (p = .339)	10/14 vs 10/14	.06 (p = .652)
step 10k	45/60	35/60 / 26	8/14 vs 7/14 (p = .500)	11/14 vs 11/14	−.03 (p = .852)
step 30k	43/60	32/60 / 26	6/14 vs 8/14 (p = .872)	10/14 vs 12/14	−.05 (p = .723)
step 300k	45/60	31/60 / 30	7/14 vs 7/14 (p = .647)	11/14 vs 12/14	−.13 (p = .331)
step 1.41M	42/60	38/60 / 35	9/14 vs 7/14 (p = .352)	10/14 vs 11/14	−.01 (p = .970)
base (end of training)	42/60	40/60 / 36	10/14 vs 8/14 (p = .347)	9/14 vs 12/14	−.05 (p = .735)
SFT	45/60	37/60 / 32	11/14 vs 4/14 (p = .011)	13/14 vs 11/14	.06 (p = .659)
DPO	44/60	35/60 / 31	11/14 vs 4/14 (p = .011)	13/14 vs 10/14	.07 (p = .619)
Instruct	44/60	36/60 / 32	11/14 vs 3/14 (p = .004)	13/14 vs 10/14	.00 (p = .994)

Panel D. Agent profiles.

Profile	Age	Gender	Occupation	City
Ivar	50	man	freelance ergonomist	Portland
Lakshmi	36	woman	urban planner	Pittsburgh
Stella	64	man	flight test engineer	Wichita
Tariq	55	woman	freelance UX writer	Portland
Eero	43	woman	freelance high-school tutor (chemistry)	Toronto
Arjun	41	woman	freelance camera operator (live sports)	Charlotte
Mira	42	woman	ICU charge nurse	Phoenix
Jada	34	woman	freelance fine-art conservator	Washington DC
Gabriel	59	non-binary	freelance gallery installer	Los Angeles
Ulises	30	non-binary	HVAC apprentice	Birmingham

Note. Panel A. Identification: percentage of the held-out writer's pairs assigned to the correct one of the 66 constructs (chance, 1.5%), at the layer at which it is highest and at the early layer used for induction (the first layer at least 10% of the way through the LLM that reaches 95% of the best layer's rate). Meaning identified: percentage of read-out word lists for which the raters chose the correct definition among ten (chance, 10%), with the logit lens and, for Qwen3.6 27B and Gemma 4 31B, the Jacobian lens. Parameters: size as named by the developer. Panel B. OLMo 3 7B at its public checkpoints, examined on one third of the pairs (2,853 pairs from all six writers). Induced judgments and r: a short version of the induction test of Study 4 (doses ±2 and ±4; 198 pairs from three writers). Judgment difference: mean difference between the LLM's judgments (1 to 9) of the high and the low narrative, on the construct's own item and over all 11 judged items. Panel C. Study 3 at the same checkpoints, with the representations estimated at each checkpoint; measures as in Table S5. Panel D. The ten agent profiles used as system prompts in the test of criterion validity and in Studies 3 to 5. Fifty-four experiments used all ten profiles; five experiments restricted to residents of the United States used nine, and one experiment on identity threat used eight. n.a. = not available; not run = not examined at this stage; not judged = read-out lists not rated at this stage.

Table S3

Identification, Multitrait-Multimethod Correlations, Baselines, Criterion Validity, and Structure Across LLMs (Studies 1 and 2)

Panel A. Identification of the held-out writer's construct and multitrait-multimethod correlations.

LLM	Pairs	Best layer (layer)	Early layer (layer)	Layers at 30–50% of depth	Among the top five (30–50%)	Input layer (0)	Multitrait-multimethod correlations (30–50%): same construct, different writers / different constructs, same writer / different constructs, different writers
Qwen3.6 27B	8,311	92% (40)	89% (19)	91%	98%	70%	.60 / .05 / .04
Gemma 4 31B	8,311	92% (36)	90% (24)	92%	98%	33%	.60 / .06 / .05
OLMo 3.1 32B	8,311	92% (23)	89% (13)	91%	98%	69%	.58 / .06 / .05
Gemma 4 26B	8,311	91% (14)	87% (9)	91%	n.a.	63%	.62 / .06 / .06
Qwen3.5 27B	8,311	93% (40)	89% (19)	92%	n.a.	70%	.60 / .05 / .05
Qwen3.5 9B	8,311	90% (20)	88% (10)	90%	n.a.	70%	.59 / .05 / .05
Qwen3.5 4B	8,311	89% (14)	87% (10)	89%	n.a.	57%	.59 / .06 / .05
Qwen3.5 2B	8,311	85% (14)	85% (10)	85%	n.a.	56%	.58 / .05 / .05
Qwen3.5 0.8B	8,311	81% (9)	77% (7)	81%	n.a.	54%	.59 / .06 / .05
Gemma 4 E4B	8,311	86% (17)	85% (13)	86%	n.a.	65%	.64 / .07 / .06
OLMo 3 7B	8,311	89% (16)	87% (10)	89%	n.a.	68%	.57 / .06 / .05
Qwen3.8 27B	8,311	92% (40)	88% (19)	90%	n.a.	70%	.60 / .05 / .05
gpt-oss 20B	8,311	87% (14)	85% (10)	86%	n.a.	58%	.61 / .06 / .06
Granite 4.2 30B	8,311	90% (43)	85% (26)	86%	n.a.	65%	.58 / .06 / .05
OLMo 3 32B base	8,311	91% (23)	89% (13)	90%	n.a.	69%	.58 / .05 / .05
Gemma 4 31B base	8,311	86% (21)	80% (6)	84%	n.a.	67%	.49 / .05 / .04
OLMo 3 7B after supervised tuning	2,853	87% (16)	85% (10)	88%	n.a.	66%	.46 / .04 / .04

Panel B. Baselines (all 8,311 pairs unless stated).

Baseline	Correct construct (first of 66)	Among the top five
Single words: counts, standardized (8,542 words)	67%	86%
Single words: presence, standardized	66%	85%
Single words: weighted by rarity, raw	72%	87%
Sequences of up to 2 words, weighted by rarity (101,985)	77%	89%
Sequences of up to 3 words, weighted by rarity (282,350; strongest)	77%	89%
Latent semantic analysis, 100 / 300 / 1,000 dimensions	52% / 63% / 70%	72% / 80% / 85%
Sentence encoder (Qwen3-Embedding-8B), standardized / raw	80% / 80%	93% / 92%
OLMo 3 7B before training (step 0; 2,853 pairs): layer 0 / layer 16 / best middle layer	48% / 40% / 42%	n.a. / 61% / n.a.
Trained LLMs, best layer (16 LLMs)	81% to 93%	

Panel C. Cross-tables for the three main LLMs.

LLM	LLM, all pairs	Where single words fail (*n* = 2,776, 33%)	Where word sequences up to 3 fail (*n* = 1,905, 23%)	Where the sentence encoder fails (*n* = 1,647, 20%)	Single words / encoder right where the LLM fails	Pairs LLM right, encoder wrong / encoder right, LLM wrong (McNemar)
Qwen3.6 27B	91%	82%	76%	67%	31% / 26%	1,101 / 191 ($p < .001$)
Gemma 4 31B	92%	83%	76%	70%	33% / 29%	1,150 / 205 ($p < .001$)
OLMo 3.1 32B	91%	82%	76%	68%	33% / 30%	1,115 / 230 ($p < .001$)

Panel D. Criterion validity (the LLM's own judgment of the construct).

LLM	Pairs	Items judged per pair	r (similarity to representation, judgment difference)	Permutation null, 95th percentile	*p*	Most similar construct = largest judgment difference	Chance	Judgment difference (1–9): own item / other items
Qwen3.6 27B	395	11	.44	.03	< 0.001 (2,000 permutations)	43%	9%	2.15 / 0.32
Gemma 4 31B	395	11	.34	.03	< 0.001 (2,000 permutations)	35%	9%	2.34 / 0.29
OLMo 3.1 32B	395	11	.41	.04	< 0.001 (2,000 permutations)	38%	9%	1.55 / 0.28

Panel E. Structure across the three main LLMs (Study 2).

Comparison	Constructs (2,145 similarities)	beyond words	beyond encoder	beyond encoder and words	Pairs of narratives (2,000)	beyond words
Qwen3.6 27B vs Gemma 4 31B	.96	.94	.88	.88	.78	.78
Qwen3.6 27B vs OLMo 3.1 32B	.94	.91	.83	.82	.71	.71
Gemma 4 31B vs OLMo 3.1 32B	.94	.91	.84	.83	.70	.70
Words vs Qwen3.6 27B	.56				.21	
Sentence encoder vs Qwen3.6 27B	.80				.43	
Words vs Gemma 4 31B	.56				.21	
Sentence encoder vs Gemma 4 31B	.78				.42	
Words vs OLMo 3.1 32B	.56				.21	
Sentence encoder vs OLMo 3.1 32B	.79				.40	
Sentence encoder vs words	.52					

Representation	Mean similarity within processes	between processes	Difference	Permutation *p*	Difference beyond words	beyond encoder
Qwen3.6 27B	0.076	0.050	0.026	< .001	0.016 ($p < .001$)	0.010 ($p < .001$)
Gemma 4 31B	0.087	0.061	0.026	< .001	0.016 ($p < .001$)	0.009 ($p = .005$)
OLMo 3.1 32B	0.088	0.062	0.026	< .001	0.017 ($p < .001$)	0.011 ($p < .001$)
Words	0.009	0.005	0.004	.005		
Sentence encoder	0.071	0.048	0.023	< .001		

Panel F. Discrimination among the closest alternatives (three main LLMs and baselines).

Representation	Identified among the constructs of the pair's own process	Chance	Mediator vs its rival: correct (mean over pairs; lowest pair)	Mediator-rival pairs
Qwen3.6 27B	96%	10%	99% (lowest 95%)	8
Gemma 4 31B	96%	10%	99% (lowest 94%)	8
OLMo 3.1 32B	95%	10%	98% (lowest 94%)	8
Sentence encoder	89%	10%	96% (lowest 90%)	8
Single words	81%	10%	93% (lowest 82%)	8

Note. Panel A. Identification as in Table S2 (chance, 1.5%). Layers at 30–50% of depth: the internal states averaged over the layers 30% to 50% of the way through the LLM, which we used for the multitrait-multimethod correlations, the baselines, criterion validity, Study 2, and Study 3. Multitrait-multimethod correlations: mean correlation between the averaged observations of the same construct from different writers, of different constructs from the same writer, and of different constructs from different writers. n.a. = not available. Panel B. The same held-out-writer procedure applied to the words of the narratives, to the sentence encoder, and to the LLM before training. Panel C. Identification by the LLM (layers at 30–50% of depth) of the pairs that each baseline failed to identify; McNemar test on the pairs on which the LLM and the sentence encoder disagree. Panel D. For each pair and each of the 11 judged items, the similarity of the pair's observation to the construct's representation (estimated without the pair's writer) and the difference between the LLM's judgments (1 to 9) of the high and the low narrative, after removing the means of each pair and each item; one pair per construct and writer (395 pairs), each judged under one agent profile. Panel E. Spearman correlations between LLMs of the similarities of every two of the 66 constructs (representations estimated from complementary halves of the writers, ten splits) and of the similarities of 2,000 random pairs of narratives; beyond = after removing the part explained by the similarities computed from the words or from the sentence encoder. Process structure: mean similarity of the representations within and between the seven processes, with 5,000 permutations of the process labels. Panel F. Identification restricted to the constructs of the pair's own process (chance = 1 divided by the number of constructs in the process, averaged over pairs) and to the eight pairs of a mediator and its rival construct (chance, 50%); layers at 30–50% of depth.

Table S4

Meaning of the Representations Identified from Their Read-Out Words (content Validity, Study 1)

Panel A. Read-out per LLM.

LLM	Logit lens: correct / in top 3	Jacobian lens: correct / in top 3	All read-out layers (main LLMs)	Word-matching procedure: correct	Raters correct where word matching failed	Rater agreement on correct (% / Cohen's kappa)	Constructs identified by both raters at some layer (of 66)
Qwen3.6 27B	52% / 74%	48% / 69%	42%; Jacobian lens 41%	17%; Jacobian lens 21%	38% (n = 554); Jacobian lens 39% (n = 402)	86% / .71	60
Gemma 4 31B	10% / 28%	36% / 60%	12%; Jacobian lens 31%	12%; Jacobian lens 14%	12% (n = 604); Jacobian lens 28% (n = 492)	84% / .52	46
OLMo 3.1 32B	61% / 82%	not run	59%	42%	53% (n = 360)	86% / .71	53
Gemma 4 26B	12% / 34%	not run	not judged	not run	not run	87% / .39	14
Qwen3.5 27B	54% / 75%	not run	not judged	not run	not run	84% / .69	47
Qwen3.5 9B	48% / 70%	not run	not judged	not run	not run	85% / .70	42
Qwen3.5 4B	46% / 68%	not run	not judged	not run	not run	82% / .63	47
Qwen3.5 2B	43% / 65%	not run	not judged	not run	not run	81% / .62	44
Qwen3.5 0.8B	40% / 64%	not run	not judged	not run	not run	80% / .59	39
Gemma 4 E4B	40% / 65%	not run	not judged	not run	not run	81% / .61	44
OLMo 3 7B	44% / 66%	not run	not judged	not run	not run	86% / .72	39
Qwen3.8 27B	49% / 68%	not run	not judged	not run	not run	83% / .66	41
gpt-oss 20B	48% / 71%	not run	not judged	not run	not run	80% / .61	46
Granite 4.2 30B	52% / 71%	not run	not judged	not run	not run	85% / .70	52
OLMo 3 32B base	68% / 85%	not run	not judged	not run	not run	86% / .67	55
Gemma 4 31B base	12% / 31%	not run	not judged	not run	not run	87% / .40	13

Panel B. Constructs identified most often (ten highest means).

Construct	Process	Qwen3.6 27B logit lens	Qwen3.6 27B Jacobian lens	Gemma 4 31B Jacobian lens	OLMo 3.1 32B logit lens	Mean
Weight of the subject's features†	categorization	100%	83%	100%	100%	96%
Emotional response to harm†	affective response	75%	100%	100%	100%	94%
Accessibility of the negative outcome	affective response	88%	100%	83%	100%	93%
Identity salience	self-relevance	100%	83%	100%	88%	93%
Analytic thinking	affective response	100%	83%	83%	100%	92%
Physiological arousal	affective response	100%	100%	67%	100%	92%
Vertical location as power	social inference	100%	100%	67%	100%	92%
Uncertainty	causal inference	88%	100%	67%	100%	89%
Perceived autonomy	social inference	100%	100%	50%	100%	88%
Goal fluency	affective response	88%	100%	83%	75%	87%

Constructs identified least often (ten lowest means).

Construct	Process	Qwen3.6 27B logit lens	Qwen3.6 27B Jacobian lens	Gemma 4 31B Jacobian lens	OLMo 3.1 32B logit lens	Mean
Others seen as partners	social inference	0%	0%	0%	13%	3%
Desire for affiliation	social inference	13%	0%	0%	0%	3%
Descriptive norm	social inference	0%	0%	17%	0%	4%
Frame of reference	social inference	0%	0%	0%	25%	6%
Affect richness	affective response	0%	0%	17%	13%	7%
Self-perception as helpful	self-relevance	0%	0%	17%	14%	8%
Psychological ownership	self-relevance	13%	0%	17%	13%	10%
Desire for power	goal relevance	38%	17%	0%	0%	14%
Anthropomorphism	categorization	25%	17%	0%	13%	14%
Evaluability	categorization	0%	50%	0%	25%	19%

Panel C. Number of constructs identified, and relation to the induced effect on judgments.

Read-out	Constructs identified in >= 30% / >= 50% of their lists	r with induced effect on own judgment	r with own minus others (mean)	r with own minus others (d_z)
Qwen3.6 27B, logit lens	44 / 35	−.05	−.03	−.13
Qwen3.6 27B, Jacobian lens	44 / 36	.03	.05	.06
Gemma 4 31B, Jacobian lens	34 / 25	−.07	−.09	−.02
OLMo 3.1 32B, logit lens	45 / 42	.12	.18	.08

Note. Panel A. Correct: percentage of read-out lists (the 30 word fragments that the construct's representation, estimated at the last word fragment of the narratives, makes most likely) for which the raters chose the construct's own definition among ten (chance, 10%; among their top three, 30%), averaged over the two raters. Logit lens: the four middle of six read-out layers between 50% and 100% of depth; Jacobian lens: the three middle of its five layers (Qwen3.6 27B and Gemma 4 31B). All read-out layers: all six or five layers. Word matching: the definition whose high narratives share the most words with the list; raters correct where word matching failed: the raters' accuracy on the lists for which word matching chose a wrong definition. Agreement: percentage of lists on which the two raters agreed whether the correct definition was chosen, and Cohen's κ. Panels B and C. Constructs ranked by their mean accuracy over the four read-outs of the main LLMs; Panel C relates this accuracy to the effect of inducing the construct's representation on its own judgment (Study 4).

Table S5

Activation of the Mediator's Representation by the Published Manipulations and Effects on the Responses, by LLM (Study 3)

Panel A. Effects by LLM.

LLM	Larger effect on mediator than on the other 65 (of 60)	Mediator first / in top five of 66 (median rank)	Spearman ρ, effect on representation vs on response	Responses: original effect positive (of 60) / significant	Representation: positive effect, replicated vs failed (of 14; Fisher's exact *p*)	Representation: median d_z, replicated vs failed (Mann–Whitney *p*)	Responses: original effect significant, replicated vs failed (of 14; Fisher's exact *p*)
Qwen3.6 27B	42/60 ($p = .001$)	6 / 16 (17)	.08 ($p = .550$)	42/60 / 34	9/14 vs 10/14 ($p = .790$)	0.98 vs 1.37 ($p = .473$)	11/14 vs 6/14 ($p = .060$)
Gemma 4 31B	48/60 ($p < .001$)	5 / 13 (16)	.09 ($p = .504$)	41/60 / 31	13/14 vs 10/14 ($p = .163$)	2.25 vs 1.05 ($p = .038$)	12/14 vs 4/14 ($p = .003$)
OLMo 3.1 32B	46/60 ($p < .001$)	6 / 20 (15.5)	.12 ($p = .378$)	37/60 / 31	11/14 vs 11/14 ($p = .676$)	1.49 vs 1.53 ($p = .436$)	10/14 vs 6/14 ($p = .126$)
Qwen3.5 27B	47/60 ($p < .001$)	not run	.04 ($p = .771$)	39/60 / 34	10/14 vs 12/14 ($p = .918$)	1.49 vs 1.70 ($p = .600$)	11/14 vs 6/14 ($p = .060$)
Qwen3.5 9B	49/60 ($p < .001$)	not run	.01 ($p = .924$)	39/60 / 33	12/14 vs 10/14 ($p = .324$)	1.89 vs 1.54 ($p = .151$)	11/14 vs 6/14 ($p = .060$)
Qwen3.5 4B	47/60 ($p < .001$)	not run	.02 ($p = .866$)	37/60 / 27	12/14 vs 9/14 ($p = .192$)	2.00 vs 2.16 ($p = .185$)	8/14 vs 5/14 ($p = .225$)
Qwen3.5 2B	46/60 ($p < .001$)	not run	−.08 ($p = .526$)	37/60 / 33	12/14 vs 9/14 ($p = .192$)	2.39 vs 1.40 ($p = .062$)	9/14 vs 6/14 ($p = .225$)
Qwen3.5 0.8B	45/60 ($p < .001$)	not run	.03 ($p = .825$)	34/60 / 30	12/14 vs 8/14 ($p = .104$)	1.76 vs 1.57 ($p = .224$)	8/14 vs 6/14 ($p = .353$)
Gemma 4 E4B	47/60 ($p < .001$)	not run	.11 ($p = .388$)	40/60 / 30	12/14 vs 10/14 ($p = .324$)	2.09 vs 1.55 ($p = .162$)	11/14 vs 7/14 ($p = .118$)
OLMo 3 7B	44/60 ($p < .001$)	not run	.00 ($p = .994$)	36/60 / 32	13/14 vs 10/14 ($p = .163$)	1.21 vs 1.27 ($p = .198$)	11/14 vs 3/14 ($p = .004$)
Qwen3.8 27B	44/60 ($p < .001$)	not run	.15 ($p = .258$)	39/60 / 32	11/14 vs 10/14 ($p = .500$)	1.16 vs 1.83 ($p = .509$)	10/14 vs 5/14 ($p = .064$)
gpt-oss 20B	46/60 ($p < .001$)	not run	−.05 ($p = .708$)	37/60 / 30	11/14 vs 11/14 ($p = .676$)	2.39 vs 0.95 ($p = .062$)	10/14 vs 4/14 ($p = .029$)
Granite 4.2 30B	49/60 ($p < .001$)	not run	−.02 ($p = .901$)	41/60 / 36	12/14 vs 11/14 ($p = .500$)	1.33 vs 1.84 ($p = .527$)	11/14 vs 8/14 ($p = .210$)
Gemma 4 26B	49/60 ($p < .001$)	not run	.14 ($p = .271$)	48/60 / 35	12/14 vs 11/14 ($p = .500$)	1.89 vs 1.81 ($p = .231$)	9/14 vs 6/14 ($p = .225$)

Panel B. Reliability of the two effects across the sets of scenario descriptions (three main LLMs).

LLM	Experiments	Sets of scenario descriptions	Split-half reliability: effect on representation	Split-half reliability: effect on response	Pearson r between the effects	Spearman ρ between the effects	r corrected for attenuation
Qwen3.6 27B	60	7	.98	.97	−.03	.08	−.03
Gemma 4 31B	60	7	.98	.98	.00	.10	.00
OLMo 3.1 32B	60	7	.98	.96	.00	.12	.00

Panel C. Correlation between the two effects by part of the text and depth (three main LLMs).

LLM	Position	Depth	ρ, effect on representation vs on response	Other experiments' mediators: mean / 95th percentile of ρ	Permutation *p*	Positive / first (of 60)
Qwen3.6 27B	manipulation	30–50%	−.02	.00 / .21	$p = .584$	42 / 3
Qwen3.6 27B	manipulation	60–75%	.03	.00 / .20	$p = .390$	43 / 4
Qwen3.6 27B	scenario description	30–50%	.08	.01 / .21	$p = .298$	42 / 2
Qwen3.6 27B	scenario description	60–75%	.09	.00 / .21	$p = .220$	48 / 3
Qwen3.6 27B	outcome measure	30–50%	.32	−.01 / .21	$p = .005$	43 / 2

LLM	Position	Depth	ρ, effect on representation vs on response	Other experiments' mediators: mean / 95th percentile of ρ	Permutation *p*	Positive / first (of 60)
Qwen3.6 27B	outcome measure	60–75%	.27	.00 / .23	$p = .026$	42 / 5
Qwen3.6 27B	last word fragment before the response	30–50%	.45	−.01 / .22	$p < .001$	38 / 2
Qwen3.6 27B	last word fragment before the response	60–75%	.38	−.04 / .19	$p = .003$	44 / 1
Gemma 4 31B	manipulation	30–50%	.07	.02 / .24	$p = .366$	48 / 7
Gemma 4 31B	manipulation	60–75%	.08	.01 / .22	$p = .304$	49 / 1
Gemma 4 31B	scenario description	30–50%	.09	.02 / .23	$p = .299$	48 / 2
Gemma 4 31B	scenario description	60–75%	.09	.01 / .22	$p = .292$	46 / 5
Gemma 4 31B	outcome measure	30–50%	.38	−.01 / .20	$p = .002$	47 / 4
Gemma 4 31B	outcome measure	60–75%	.37	−.02 / .21	$p = .002$	49 / 5
Gemma 4 31B	last word fragment before the response	30–50%	.11	−.01 / .20	$p = .182$	38 / 0
Gemma 4 31B	last word fragment before the response	60–75%	.43	−.04 / .19	$p < .001$	35 / 1
OLMo 3.1 32B	manipulation	30–50%	.09	.01 / .22	$p = .259$	44 / 3
OLMo 3.1 32B	manipulation	60–75%	.08	.00 / .21	$p = .280$	43 / 3
OLMo 3.1 32B	scenario description	30–50%	.12	.00 / .21	$p = .194$	46 / 3
OLMo 3.1 32B	scenario description	60–75%	.10	.00 / .22	$p = .223$	45 / 2
OLMo 3.1 32B	outcome measure	30–50%	.15	.02 / .23	$p = .166$	49 / 5
OLMo 3.1 32B	outcome measure	60–75%	.08	.01 / .24	$p = .284$	45 / 6
OLMo 3.1 32B	last word fragment before the response	30–50%	.31	−.03 / .20	$p = .007$	45 / 4
OLMo 3.1 32B	last word fragment before the response	60–75%	.18	−.05 / .18	$p = .052$	42 / 2

Panel D. The two example experiments (three main LLMs).

Experiment (mediator)	LLM	Activation, manipulation passages	Activation, scenario description	Activation, outcome measure	Activation, last fragment before the response	Effect on the response
Identity salience (interdependent self)	Qwen3.6 27B	11.60	4.67	-2.35	-0.07	-0.05
Identity salience (interdependent self)	Gemma 4 31B	18.78	4.90	0.09	1.13	0.06
Identity salience (interdependent self)	OLMo 3.1 32B	10.12	5.82	0.33	2.67	-0.08
Moral typecasting (moral agency)	Qwen3.6 27B	2.77	1.45	4.43	5.33	7.20
Moral typecasting (moral agency)	Gemma 4 31B	3.39	2.22	2.63	1.02	8.27
Moral typecasting (moral agency)	OLMo 3.1 32B	2.25	1.73	3.98	2.09	7.09

Note. Panel A. Effect on the mediator's representation: for each experiment, the treatment-minus-control difference in the activation of the mediator's representation while the LLM processed the scenario description (layers at 30–50% of depth), minus the mean of the same difference for the other 65 representations, in the direction that the theory predicts; positive counts with exact one-sided binomial tests against .5. Its d_z is computed across the seven sets of scenario descriptions, each averaged over the agent profiles. Rank: rank of the mediator among the 66 constructs after removing the part common to all experiments (main LLMs; by chance, the mediator would be first in about 0.9 and among the top five in about 4.5 of the 60 experiments). Effect on the response: treatment minus control on the response (1 to 9) in the predicted direction, d_z across the combinations of scenario description and agent profile; positive / $t > 2$. Replication: the 28 experiments re-examined in large preregistered replications (14 replicated, 14 failed); one-sided Fisher's exact tests on the counts and one-sided Mann-Whitney tests on the median d_z. Panel B. Split-half reliability over the 35 splits of the seven sets of scenario descriptions into three and four sets, corrected with the Spearman-Brown formula; correlation corrected for attenuation = r divided by the square root of the product of the reliabilities. Panel C. The effect on the mediator's representation computed as in Panel A from the internal states for the passages that convey

the manipulation, the scenario description, the outcome measure (identical in both conditions), and the last word fragment before the response, in the layers 30–50% or 60–75% of depth, with the representations estimated in the same layers; ρ = Spearman correlation with the effect on the response across the 60 experiments. Other experiments' mediators: 2,000 random assignments of another experiment's mediator to each experiment; mean and 95th percentile of ρ and one-sided permutation p. Positive: experiments with a positive effect on the mediator's representation; first: experiments in which the mediator's representation was the most strongly activated of the 66. Panel D. The two example experiments: d_z of the manipulation's activation of the mediator's representation (layers at 30–50% of depth) for each part of the text, and d_z of the manipulation's effect on the response.

Table S6

Effects of Inducing and Removing the Mediator's Representation on Judgments and Responses, by LLM (Study 4)

Panel A. Induced judgments (Study 1 narratives).

LLM	Design (comparison directions)	Layer	Own > others, of 66 ($t > 2$; binomial p)	Own direction first	r (induced, natural) (centered)	Own induction raises own judgment	Mean effect at doses -4/-2/-1/+1/+2/+4
Qwen3.6 27B	full (15 directions)	19	48 (26; $p < .001$)	18/66	.43 (.43)	53/66	−0.16 / −0.08 / −0.04 / 0.05 / 0.11 / 0.24
Gemma 4 31B	full (15 directions)	24	55 (20; $p < .001$)	18/66	.41 (.48)	56/66	−0.33 / −0.19 / −0.10 / 0.12 / 0.25 / 0.56
OLMo 3.1 32B	full (15 directions)	13	59 (38; $p < .001$)	17/66	.54 (.50)	62/66	−0.22 / −0.12 / −0.06 / 0.08 / 0.18 / 0.43
Gemma 4 26B	full (15 directions)	9	47 (23; $p < .001$)	18/66	.35 (.37)	n.a.	n.a.
Qwen3.5 27B	short (10)	19	53 (19; $p < .001$)	17/66	.38 (.36)	n.a.	n.a.
Qwen3.5 9B	short (10)	10	46 (24; $p < .001$)	23/66	.33 (.33)	n.a.	n.a.
Qwen3.5 4B	short (10)	10	49 (28; $p < .001$)	25/66	.44 (.47)	n.a.	n.a.
Qwen3.5 2B	short (10)	10	45 (31; $p = .002$)	13/66	.54 (.51)	n.a.	n.a.
Qwen3.5 0.8B	short (10)	7	47 (28; $p < .001$)	10/66	.31 (.49)	n.a.	n.a.
Gemma 4 E4B	short (10)	13	49 (18; $p < .001$)	24/66	.39 (.35)	n.a.	n.a.
OLMo 3 7B	short (10)	10	56 (30; $p < .001$)	18/66	.56 (.52)	n.a.	n.a.
Qwen3.8 27B	not run						
gpt-oss 20B	not run						
Granite 4.2 30B	not run						
OLMo 3 32B base	not run						
Gemma 4 31B base	not run						
Qwen3.6 27B, discriminant directions	short (10)	19	47 (15)	11/66	.33 (.33)	n.a.	n.a.
Gemma 4 31B, discriminant directions	short (10)	24	37 (8)	16/66	.11 (.09)	n.a.	n.a.
OLMo 3.1 32B, discriminant directions	short (10)	13	58 (35)	20/66	.48 (.47)	n.a.	n.a.
Gemma 4 26B, discriminant directions	short (10)	9	34 (16)	11/66	.02 (.08)	n.a.	n.a.

Panel B. Responses and in-context judgments in the 52 experiments.

LLM	Early layer: response changed as predicted (*p*) / own > other constructs	Half depth, scenario description: predicted (*p*) / own > others / own > random	70% depth, scenario description to end: predicted (*p*) / own > others / own > random	Early layer: in-context judgment of the mediator raised (*p*) / own > others	Manipulation raised the in-context judgment (of 60)
Qwen3.6 27B	32 ($p = .063$) / 30	36 ($p = .004$) / 32 / 38	41 ($p < .001$) / 31 / 45	40 ($p < .001$) / 34	48/60
Gemma 4 31B	32 ($p = .063$) / 27	34 ($p = .018$) / 31 / 32	36 ($p = .004$) / 25 / 20	39 ($p < .001$) / 32	46/60
OLMo 3.1 32B	30 ($p = .166$) / 27	30 ($p = .166$) / 23 / 31	33 ($p = .035$) / 20 / 28	45 ($p < .001$) / 39	48/60
Gemma 4 26B	25 ($p = .661$) / 23	36 ($p = .004$) / 34 / 40	38 ($p < .001$) / 35 / 21	40 ($p < .001$) / 35	43/60

Panel C. Responses in 25 experiments by depth of the induction.

LLM	Early layer, scenario description	Half depth, scenario description	70% depth, scenario description to end	70% depth, scenario description only
Qwen3.6 27B	14 / 15 (0.020)	20 / 19 (0.080)	21 / 19 (0.103)	19 / 14 (0.014)
Gemma 4 31B	18 / 16 (0.021)	19 / 19 (0.030)	14 / 10 (0.012)	14 / 9 (0.000)
OLMo 3.1 32B	18 / 16 (0.034)	14 / 12 (0.002)	16 / 14 (0.009)	15 / 11 (0.001)
Gemma 4 26B	15 / 16 (0.050)	21 / 21 (0.097)	20 / 19 (0.040)	20 / 14 (0.003)
Qwen3.5 27B	16 / 14 (0.019)	19 / 17 (0.072)	19 / 19 (0.088)	16 / 10 (0.012)
Qwen3.5 9B	12 / 11 (0.003)	21 / 12 (0.033)	18 / 22 (0.032)	10 / 8 (−0.001)
Qwen3.5 4B	17 / 20 (0.020)	18 / 17 (0.025)	23 / 22 (0.041)	10 / 7 (−0.001)
Qwen3.5 2B	19 / 22 (0.022)	14 / 19 (0.007)	13 / 19 (0.014)	9 / 11 (−0.001)
Qwen3.5 0.8B	17 / 20 (0.013)	18 / 18 (0.016)	19 / 20 (0.058)	15 / 16 (0.001)
Gemma 4 E4B	15 / 14 (0.046)	20 / 21 (0.056)	20 / 19 (0.031)	1 / 1 (0.000)
OLMo 3 7B	22 / 18 (0.053)	19 / 18 (0.021)	19 / 22 (0.035)	10 / 8 (−0.002)
Qwen3.8 27B	16 / 16 (0.008)	19 / 19 (0.088)	18 / 16 (0.093)	20 / 16 (0.019)
gpt-oss 20B	17 / 17 (0.043)	21 / 16 (0.054)	21 / 22 (0.131)	20 / 19 (0.019)
Granite 4.2 30B	n.a.	n.a.	n.a.	n.a.

Panel D. Removal of the mediator's representation (10 directions).

LLM	Form	Experiments with absolute effect > .3	Own (10 directions)	Own, specific	Most similar constructs, specific	Other (random) constructs, specific	Random subspaces	Three shared directions	Presentations excluded (digit probability < .5), largest set
Qwen3.6 27B	mean replacement	36	60%	68%	86%	93%	98%	67%	10%
Qwen3.6 27B	zeroing	36	12%	10%	20%	25%	91%	39%	1%
Gemma 4 31B	mean replacement	30	24%	39%	17%	83%	84%	43%	10%
Gemma 4 31B	zeroing	30	13%	0%	0%	0%	0%	n.a.	98%
OLMo 3.1 32B	mean replacement	30	91%	99%	97%	99%	99%	74%	0%
OLMo 3.1 32B	zeroing	30	99%	103%	102%	101%	101%	94%	0%
Gemma 4 26B	both	n.a.							

Panel E. Removal by number of directions (mean replacement).

LLM	Layers	Experiments	Own, 10	Own, 30	Own, 100	Random, 10	Random, 30	Random, 50	Random, 100	Own + 4 most similar (50)	5 most similar (50)	5 random constructs (50)
Qwen3.6 27B	layers of panel D	36	68%	65%	28%	96%	91%	59%	12%	40%	67%	40%
Qwen3.6 27B	early layer to last	36	61%	59%	20%	97%	77%	33%	invalid responses	48%	43%	42%
Gemma 4 31B	layers of panel D	29	26%	72%	1%	80%	79%	66%	1%	6%	−3%	9%
Gemma 4 31B	early layer to last	29	invalid responses in every condition									
OLMo 3.1 32B	layers of panel D	31	100%	104%	84%	96%	105%	86%	70%	91%	91%	87%
OLMo 3.1 32B	early layer to last	31	85%	91%	62%	98%	98%	59%	55%	69%	93%	91%
Gemma 4 26B	layers of panel D	39	59%	11%	2%	57%	20%	0%	3%	13%	−3%	5%
Gemma 4 26B	early layer to last	39	not reported									

Note. Panel A. Inducing a construct's representation in the low narratives of Study 1 (added at the early layer, in multiples of the construct's high-minus-low difference), judged on the construct's first-person item: number of the 66 judgments that the construct's own representation (doses ±2) changed more than the comparison directions did (the representations of the five most similar and five random constructs and random directions in the full design; fewer in the short design); own direction first: the own representation had the largest effect of all directions; r: correlation across combinations of construct and judgment between the effects of the induced representations and the effects of the narratives. Panel B. The 52 experiments of Study 4 (Supplemental Note 1); induction in the treatment condition; predicted: the response changed in the direction that the theory predicts (chance, 26); own > others: more than the representations of the three most similar and three random constructs; own > random: more than random directions of the same size. Half depth: the layer nearest 50% of the way through the LLM, at the word fragments of the scenario description; 70% depth: from the scenario description's first word fragment to the end of the outcome measure. In-context judgment: the mediator's first-person item presented after the scenario description. Panel C. Twenty-five of the 52 experiments, in the LLMs in which we induced the representation at several depths; response changed as predicted / more than with random directions (doses ±2). Panel D. Removal of a subspace of ten directions at every word fragment, across the layers at which the constructs were identified best: median share of the manipulation's effect on the response left, over the experiments with an absolute effect greater than .3 scale points; specific: after excluding the three directions shared by all 66 constructs. Presentations whose probabilities of the digits 1 to 9 summed to less than .5 were excluded; fixing the internal state at zero instead of its mean made most of Gemma 4 31B's responses invalid in this sense. Panel E. Removal by the number of directions (mean replacement; median share left). n.a. = not available.

Table S7

Situational Moderation of the Effect of the Induced Mediator's Representation on the Responses, by Experiment and LLM (Study 4)

Panel A. Induced at the early layer, at the word fragments of the scenario description.

Experiment	Situational moderator	Qwen3.6 27B	Gemma 4 31B	OLMo 3.1 32B	Gemma 4 26B
Identity divergence	identity relevance of the product domain	1.23 (5.2)	−0.01 (0.0)	0.12 (0.5)	0.30 (1.3)
Going green	audience while shopping	0.30 (1.3)	−0.19 (−0.8)	−0.58 (−2.5)	0.15 (0.6)
Social exclusion	whether the partner will see one's willingness to pay	−0.02 (−0.1)	−0.39 (−1.6)	0.47 (2.0)	0.57 (2.4)
Identity salience	framing of the same device	−0.18 (−0.8)	0.24 (1.0)	−0.76 (−3.2)	−0.32 (−1.4)
Powerlessness	description of the same photograph	0.32 (1.4)	0.35 (1.5)	−0.37 (−1.6)	−0.31 (−1.3)
Normative message	a sign of approval on the note	0.34 (1.4)	−0.35 (−1.5)	0.14 (0.6)	−0.09 (−0.4)
Identity threat	the film's link to the threatened identity	−0.72 (−3.1)	−0.08 (−0.4)	0.20 (0.9)	−0.24 (−1.0)
Predicted direction (of 7) / $t > 2$ / $t < -2$		4 / 1 / 1	2 / 0 / 0	4 / 0 / 2	3 / 1 / 0

Panel B. Induced halfway through the LLM, at the word fragments of the scenario description.

Experiment	Situational moderator	Qwen3.6 27B	Gemma 4 31B	OLMo 3.1 32B	Gemma 4 26B
Identity divergence	identity relevance of the product domain	−0.38 (−1.6)	0.09 (0.4)	0.60 (2.5)	0.16 (0.7)
Going green	audience while shopping	0.36 (1.5)	−0.31 (−1.3)	0.40 (1.7)	0.40 (1.7)
Social exclusion	whether the partner will see one's willingness to pay	−0.34 (−1.4)	−0.15 (−0.6)	−0.39 (−1.7)	−0.03 (−0.1)
Identity salience	framing of the same device	0.05 (0.2)	0.11 (0.5)	−0.61 (−2.6)	0.12 (0.5)
Powerlessness	description of the same photograph	0.15 (0.6)	0.24 (1.0)	0.33 (1.4)	0.06 (0.3)
Normative message	a sign of approval on the note	−0.79 (−3.4)	0.36 (1.5)	0.16 (0.7)	−0.03 (−0.1)
Identity threat	the film's link to the threatened identity	−0.32 (−1.4)	−0.32 (−1.4)	−0.48 (−2.1)	−1.12 (−4.8)
Predicted direction (of 7) / $t > 2$ / $t < -2$		3 / 0 / 1	4 / 0 / 0	4 / 1 / 2	4 / 0 / 1

Panel C. Induced 70% of the way through the LLM, from the scenario description to the end of the outcome measure.

Experiment	Situational moderator	Qwen3.6 27B	Gemma 4 31B	OLMo 3.1 32B	Gemma 4 26B
Identity divergence	identity relevance of the product domain	0.39 (1.6)	0.20 (0.8)	−0.86 (−3.7)	0.14 (0.6)
Going green	audience while shopping	0.36 (1.5)	0.03 (0.1)	−0.18 (−0.8)	0.24 (1.0)
Social exclusion	whether the partner will see one's willingness to pay	−0.46 (−1.9)	−0.23 (−1.0)	−0.19 (−0.8)	−0.12 (−0.5)
Identity salience	framing of the same device	−0.37 (−1.6)	−0.16 (−0.7)	0.60 (2.6)	−0.02 (−0.1)
Powerlessness	description of the same photograph	0.14 (0.6)	0.34 (1.5)	0.32 (1.4)	0.03 (0.1)
Normative message	a sign of approval on the note	−0.53 (−2.3)	0.28 (1.2)	0.69 (2.9)	0.19 (0.8)
Identity threat	the film's link to the threatened identity	−0.51 (−2.2)	0.32 (1.4)	0.05 (0.2)	−0.75 (−3.2)
Predicted direction (of 7) / $t > 2$ / $t < -2$		3 / 0 / 2	5 / 0 / 0	4 / 2 / 1	4 / 0 / 1

Note. Each cell: interaction d_z (t) between the situational moderator and the induced mediator's representation on the response, aligned so that positive values are in the direction that the theory predicts: the effect of the induced representation at the level of the moderator at which the theory predicts the larger effect, minus its effect at the other level (18 combinations of scenario description and agent profile per experiment; doses ±2).

Table S8

Representations of Individual Characteristics and the Heterogeneity They Produce, by LLM (Study 5)

Panel A. Identification, the described characteristic's representation, and the heterogeneity it produces (thirteen LLMs).

LLM	Characteristic identified (of 7; chance 14%)	Description activates the characteristic's representation before the response (of 7 tests)	Costs justify prices	Consistency	Materialism	Collective self-esteem	Luxury guilt	Family importance	Self-monitoring (null)
Qwen3.6 27B	95%	6/7	4.30 (18.3)	1.47 (6.2)	0.41 (1.7)	0.88 (3.7)	−0.08 (−0.3)	−0.91 (−3.9)	0.41 (1.7)
Gemma 4 31B	91%	7/7	2.61 (11.1)	1.15 (4.9)	−0.20 (−0.9)	−0.29 (−1.2)	0.27 (1.2)	−1.00 (−4.3)	0.21 (0.9)
OLMo 3.1 32B	97%	7/7	2.08 (8.8)	1.32 (5.6)	0.72 (3.1)	−0.34 (−1.5)	−0.57 (−2.4)	0.90 (3.8)	0.36 (1.5)
Gemma 4 26B	87%	5/7	2.33 (9.9)	1.15 (4.9)	0.09 (0.4)	−0.14 (−0.6)	0.22 (0.9)	−0.61 (−2.6)	0.57 (2.4)
Qwen3.5 27B	94%	7/7	2.70 (11.4)	1.48 (6.3)	1.54 (6.5)	−0.03 (−0.1)	−0.01 (0.0)	−1.01 (−4.3)	0.98 (4.2)
Qwen3.5 9B	96%	7/7	0.07 (0.3)	1.39 (5.9)	0.85 (3.6)	0.15 (0.6)	0.29 (1.2)	0.78 (3.3)	−0.59 (−2.5)
Qwen3.5 4B	94%	5/7	0.27 (1.2)	−0.04 (−0.2)	0.19 (0.8)	−0.04 (−0.2)	0.60 (2.6)	0.66 (2.8)	−0.22 (−1.0)
Qwen3.5 2B	93%	5/7	1.13 (4.8)	−0.13 (−0.6)	0.19 (0.8)	−0.65 (−2.8)	0.03 (0.1)	1.88 (8.0)	−0.03 (−0.1)
Qwen3.5 0.8B	92%	3/7	−0.57 (−2.4)	1.59 (6.8)	0.01 (0.0)	−0.94 (−4.0)	−0.05 (−0.2)	1.01 (4.3)	0.15 (0.6)
Gemma 4 E4B	94%	5/7	1.88 (8.0)	1.36 (5.8)	1.17 (4.9)	−0.62 (−2.6)	−0.17 (−0.7)	1.11 (4.7)	0.02 (0.1)
OLMo 3 7B	95%	3/7	0.25 (1.1)	1.92 (8.2)	0.48 (2.0)	−0.76 (−3.2)	0.73 (3.1)	0.36 (1.5)	0.38 (1.6)
gpt-oss 20B	89%	3/7	0.78 (3.3)	0.37 (1.6)	0.34 (1.5)	−0.13 (−0.6)	−0.42 (−1.8)	1.39 (5.9)	0.19 (0.8)
Granite 4.2 30B	93%	6/7	1.86 (7.9)	1.05 (4.5)	0.02 (0.1)	−0.95 (−4.0)	−0.36 (−1.5)	−0.02 (−0.1)	1.04 (4.4)
Not run	Qwen3.8 27B (the job did not run)								
$t > 2$ in the predicted direction / $t < -2$ (of 13)		69/91	9 / 1	10 / 0	5 / 0	1 / 5	2 / 1	7 / 4	3 / 1
		0.58 [0.20, 0.95]	0.52 [0.15, 0.90]	0.22 [−0.16, 0.59]	0.97 [0.57, 1.36]	0.05 [−0.32, 0.42]	0.15 [−0.22, 0.53]	0.10 [−0.27, 0.47]	
		0.60	0.53	0.20	0.64	0.30	0.13	−0.23	

Panel B. Inducing the characteristic's representation instead of describing it.

LLM	Layers (identification / half depth)	Costs justify prices	Consistency	Materialism	Collective self-esteem	Luxury guilt	Family importance	Self-monitoring (null)
Qwen3.6 27B	19 / 32	0.12 (0.5) / 0.06 (0.3) / 0.70	0.67 (2.9) / 0.99 (4.2) / −0.16	−0.37 (−1.6) / −0.60 (−2.5) / 0.43	−0.63 (−2.7) / −0.39 (−1.7) / −0.76	−0.45 (−1.9) / −0.38 (−1.6) / 0.51	0.24 (1.0) / −0.06 (−0.2) / 0.48	0.67 (2.8) / 0.44 (1.9) / 0.68
Gemma 4 31B	24 / 30	−0.29 (−1.3) / −0.03 (−0.1) / 0.26	0.02 (0.1) / 0.23 (1.0) / 0.49	−0.03 (−0.1) / −0.51 (−2.2) / 0.62	−0.45 (−1.9) / −0.29 (−1.2) / −0.51	−0.27 (−1.1) / −0.01 (−0.1) / 0.25	−0.10 (−0.4) / 0.00 (0.0) / −0.12	−0.25 (−1.1) / 0.28 (1.2) / 0.25
OLMo 3.1 32B	13 / 32	0.13 (0.5) / 0.19 (0.8) / 0.01	0.51 (2.2) / −0.05 (−0.2) / 0.22	0.15 (0.7) / −0.11 (−0.5) / 0.44	0.08 (0.3) / 0.18 (0.8) / −0.03	−0.27 (−1.2) / −0.10 (−0.4) / −0.14	0.46 (1.9) / −0.05 (−0.2) / 0.18	−0.49 (−2.1) / 0.04 (0.2) / −0.32
Gemma 4 26B	9 / 15	0.39 (1.6) / −0.38 (−1.6) / −0.20	0.14 (0.6) / 0.39 (1.7) / −0.46	−0.56 (−2.4) / −0.49 (−2.1) / 0.39	−0.17 (−0.7) / 0.12 (0.5) / −0.10	−0.03 (−0.1) / 0.23 (1.0) / 0.53	0.25 (1.1) / 0.20 (0.8) / 0.05	0.33 (1.4) / 0.07 (0.3) / −0.07
Qwen3.5 27B	19 / 32	0.22 (0.9) / 0.17 (0.7) / 0.89	0.31 (1.3) / 1.51 (6.4) / −0.06	0.04 (0.2) / −0.28 (−1.2) / 0.53	−0.25 (−1.1) / −0.74 (−3.2) / −0.75	−0.17 (−0.7) / −0.02 (−0.1) / 0.33	−0.01 (−0.1) / 0.17 (0.7) / 0.83	0.39 (1.7) / 0.54 (2.3) / 0.58
Qwen3.5 9B	10 / 16	−0.24 (−1.0) / −0.22 (−0.9) / 0.08	1.13 (4.8) / 0.41 (1.7) / 0.52	−0.67 (−2.9) / 0.39 (1.7) / −0.07	−0.03 (−0.1) / −0.48 (−2.0) / −0.55	−0.27 (−1.1) / −0.08 (−0.3) / 0.21	−0.54 (−2.3) / 0.80 (3.4) / 0.31	0.01 (0.0) / −0.07 (−0.3) / 0.04
Qwen3.5 4B	10 / 16	0.75 (3.2) / 0.28 (1.2) / −0.09	−0.19 (−0.8) / 0.14 (0.6) / 0.17	0.29 (1.2) / −0.08 (−0.3) / −0.18	−0.45 (−1.9) / −0.24 (−1.0) / 0.24	0.22 (0.9) / −0.06 (−0.3) / 0.24	0.34 (1.5) / −0.04 (−0.2) / −0.17	0.26 (1.1) / −0.42 (−1.8) / −0.14
Qwen3.5 2B	10 / 12	0.08 (0.3) / 0.06 (0.2) / 0.36	0.01 (0.0) / 0.16 (0.7) / 0.40	−0.07 (−0.3) / −0.19 (−0.8) / −0.03	−1.27 (−5.4) / −0.21 (−0.9) / 0.32	0.55 (2.3) / 0.19 (0.8) / −0.15	−0.10 (−0.4) / 0.23 (1.0) / −0.30	0.01 (0.0) / −0.41 (−1.8) / 0.22
Qwen3.5 0.8B	7 / 12	−0.39 (−1.7) / 0.27 (1.2) / −0.20	0.62 (2.6) / 0.15 (0.7) / −0.16	0.19 (0.8) / −0.46 (−2.0) / −0.23	0.24 (1.0) / −0.08 (−0.3) / −0.10	0.04 (0.2) / −0.10 (−0.4) / 0.17	0.46 (1.9) / −0.25 (−1.1) / −0.33	0.07 (0.3) / −0.13 (−0.5) / 0.16
Gemma 4 E4B	13 / 21	0.63 (2.7) / 1.26 (5.3) / −0.03	0.00 (0.0) / 0.41 (1.7) / −0.49	−0.31 (−1.3) / −0.34 (−1.5) / 0.01	−0.57 (−2.4) / −0.57 (−2.4) / −0.75	0.06 (0.2) / 0.15 (0.6) / 0.25	0.36 (1.5) / −0.11 (−0.5) / −0.01	0.24 (1.0) / 0.65 (2.8) / −0.07
OLMo 3 7B	10 / 16	1.16 (4.9) / −0.31 (−1.3) / 0.72	0.14 (0.6) / −0.59 (−2.5) / −0.20	0.90 (3.8) / 0.42 (1.8) / −0.07	0.12 (0.5) / 0.05 (0.2) / −0.26	−0.06 (−0.2) / −0.10 (−0.4) / −0.04	0.81 (3.4) / 0.61 (2.6) / 0.23	0.01 (0.0) / −0.55 (−2.3) / 0.66
gpt-oss 20B	10 / 12	0.47 (2.0) / 0.89 (3.8) / 0.55	0.78 (3.3) / 0.93 (3.9) / −0.32	−0.14 (−0.6) / 0.06 (0.2) / 0.53	0.39 (1.7) / 0.18 (0.8) / −0.25	0.02 (0.1) / 0.42 (1.8) / −0.23	0.78 (3.3) / 0.87 (3.7) / −0.14	0.16 (0.7) / 0.78 (3.3) / 0.03
Granite 4.2 30B	26 / 32	0.36 (1.5) / 0.78 (3.3) / −0.35	0.18 (0.8) / 0.71 (3.0) / 0.05	−0.18 (−0.8) / −0.59 (−2.5) / 0.20	−0.10 (−0.4) / −0.72 (−3.1) / −1.20	−0.23 (−1.0) / −0.23 (−1.0) / 0.03	0.90 (3.8) / 0.52 (2.2) / −0.87	−0.29 (−1.2) / 0.25 (1.1) / −0.33

Panel C. Removal of the characteristic's representation (share of the heterogeneity left).

LLM	Characteristic (test)	Heterogeneity as described: mean (d_z)	At the statement: own	own, adjusted	random (2)	other characteristics (2)	From the statement on: own	random (2)	other characteristics (2)
Qwen3.6 27B	Costs justify prices	3.47 (4.16)	95%	96%	95% / 94%	92% / 95%	99%	42% / 76%	105% / 113%
Qwen3.6 27B	Consistency	0.58 (1.22)	53%	60%	94% / 59%	123% / 93%	68%	107% / 75%	127% / 181%
Qwen3.6 27B	Materialism	0.06 (0.45)	157%	130%	−20% / −187%	162% / 67%	213%	569% / −144%	201% / 270%
Qwen3.6 27B	Collective self-esteem	0.26 (0.86)	76%	73%	84% / 105%	105% / 116%	97%	4% / 103%	127% / 133%
Qwen3.6 27B	Luxury guilt	−0.02 (−0.07)	301%	251%	107% / 330%	139% / 141%	−459%	−1821% / −502%	−452% / 255%
Qwen3.6 27B	Family importance	−1.59 (−0.91)	67%	67%	65% / 79%	102% / 86%	−52%	44% / 47%	68% / 124%
Qwen3.6 27B	Self-monitoring (null)	0.01 (0.03)	1068%	907%	−419% / −384%	844% / 167%	518%	1676% / −383%	−364% / −247%
Gemma 4 31B	Costs justify prices	5.59 (2.49)	0%	−1%	−4% / 0%	3% / 1%	invalid responses	0% / n.a.	invalid responses

LLM	Characteristic (test)	Heterogeneity as described: mean (d_z)	At the statement: own	own, adjusted	random (2)	other characteristics (2)	From the statement on: own	random (2)	other characteristics (2)
Gemma 4 31B	Consistency	2.14 (1.10)	3%	4%	22% / 11%	−1% / 12%	invalid responses	invalid responses	invalid responses
Gemma 4 31B	Materialism	−0.07 (−0.12)	−113%	−102%	−20% / −50%	−12% / −56%	invalid responses	invalid responses	invalid responses
Gemma 4 31B	Collective self-esteem	−0.26 (−0.29)	−24%	−17%	2% / −9%	−5% / −18%	0%	invalid responses	invalid responses
Gemma 4 31B	Luxury guilt	0.02 (0.05)	142%	156%	−81% / −50%	−76% / 55%	invalid responses	invalid responses	invalid responses
Gemma 4 31B	Family importance	−2.64 (−0.92)	−4%	−2%	−7% / −2%	−6% / −5%	invalid responses	3% / n.a.	invalid responses
Gemma 4 31B	Self-monitoring (null)	0.09 (0.17)	−177%	−95%	−117% / −300%	−145% / −303%	invalid responses	invalid responses	invalid responses
OLMo 3.1 32B	Costs justify prices	0.57 (1.87)	67%	67%	93% / 96%	105% / 89%	92%	106% / 118%	93% / 118%
OLMo 3.1 32B	Consistency	1.08 (1.32)	91%	92%	99% / 100%	99% / 99%	88%	94% / 92%	98% / 90%
OLMo 3.1 32B	Materialism	0.06 (0.71)	98%	74%	103% / 107%	109% / 91%	67%	141% / 97%	91% / 109%
OLMo 3.1 32B	Collective self-esteem	−0.13 (−0.34)	91%	93%	91% / 93%	102% / 98%	103%	86% / 63%	122% / 88%
OLMo 3.1 32B	Luxury guilt	−0.10 (−0.63)	59%	80%	81% / 86%	91% / 91%	89%	90% / 66%	73% / 69%
OLMo 3.1 32B	Family importance	0.68 (0.91)	113%	113%	100% / 100%	103% / 100%	140%	101% / 126%	102% / 96%
OLMo 3.1 32B	Self-monitoring (null)	0.04 (0.39)	47%	92%	116% / 104%	86% / 91%	115%	87% / 90%	142% / 135%
Gemma 4 26B	Costs justify prices	6.37 (2.41)	34%	34%	72% / 77%	31% / 55%	18%	31% / 32%	−1% / 1%
Gemma 4 26B	Consistency	1.32 (1.06)	11%	16%	111% / 60%	34% / 82%	2%	55% / 11%	−26% / 79%
Gemma 4 26B	Materialism	0.07 (0.08)	111%	151%	−218% / −357%	50% / −293%	4%	552% / 60%	354% / 102%
Gemma 4 26B	Collective self-esteem	−0.10 (−0.20)	24%	71%	−149% / 218%	52% / −24%	124%	7% / 241%	133% / 511%
Gemma 4 26B	Luxury guilt	0.45 (0.24)	198%	92%	38% / 28%	105% / 1%	−52%	−155% / 4%	32% / −12%
Gemma 4 26B	Family importance	−1.06 (−0.57)	−9%	3%	52% / 116%	89% / 67%	−41%	−31% / −59%	−12% / −213%
Gemma 4 26B	Self-monitoring (null)	0.35 (0.41)	65%	67%	39% / 91%	9% / 66%	7%	4% / −50%	98% / −20%

Note. Panel A. Seven characteristics; 611 pairs of narratives from the six writers. Identification: percentage of the held-out writer's pairs assigned to the correct one of the seven characteristics (early layer). Description activates the representation: the internal state at the last word fragment before the response, projected on the characteristic's representation, is higher after a

description of a high than of a low level of the characteristic in the agent profile (seven distinct tests). Characteristic columns: interaction of the described characteristic and the manipulation on the response, d_z (t), in the direction that the theory predicts (18 combinations of scenario description and agent profile; $t > 2.11$ corresponds to $p < .05$, two-sided, with 17 degrees of freedom); counts use $|t| > 2$, and with $|t| > 2.11$ they are the same except for materialism (4 / 0). Panel B. The characteristic's representation induced at the word fragments of the agent profile (doses ±2) instead of describing the characteristic: own representation at the early layer, d_z (t); own at half depth, d_z (t); random direction of the same size at half depth, d_z. Panel C. Share of the heterogeneity left after removing one direction per layer, at each recorded layer between 10% and 85% of depth, at the word fragments of the statement that describes the characteristic or from this statement on; own: the characteristic's representation; adjusted: after removing the parts that covary with rated wealth, pleasantness, and sociability; random: two random directions; other: the directions of two other characteristics. Shares are unstable where the heterogeneity is near zero. Invalid responses: the probabilities of the digits 1 to 9 summed to less than .5 in most presentations.

Table S9

Example Pairs of Narratives and the Numbers of Composed, Excluded and Retained Pairs

Panel A. Example pairs.

Construct (process; writer; low narrative; domain; pair id)	High narrative	Low narrative
Perceived fit (categorization; Claude Sonnet; low = opposite; domain: clothing shop)	You are in a clothing shop that has made its own wool jumpers in the back room for years. A sign at the till says it now sells wool scarves, knitted on the same machines by the same two knitters from the same wool, to be worn with the jumpers. The new items are neat, well made, priced at 25 and wrapped in tissue.	You are in a clothing shop that has made its own wool jumpers in the back room for years. A sign at the till says it now sells scented candles, poured in new moulds by a chandler hired for the job from wax bought in, to burn in glass jars. The new items are neat, well made, priced at 25 and wrapped in tissue.
Reciprocity (social inference; DeepSeek V4.1 Flash; low = absent; domain: rowing club)	At the rowing club, Lena carried your oars from the rack to the boat when your hands were full. The Saturday crew list is pinned on the board by the front door. Lena now asks you to swap your Saturday slot with hers.	At the rowing club, Lena's name is printed beside yours on the Saturday crew list today. The Saturday crew list is pinned on the board by the front door. Lena now asks you to swap your Saturday slot with hers.
Psychological ownership (self-relevance; Grok 4.7; low = opposite; domain: small hotel)	After you booked the corner room and paid for two nights, the brass lamp you selected at the desk has stayed in place each evening. The shade is warm and even. You switch it on beside the bed for reading at 9.	For these two nights the desk lends you the brass lamp from the hotel's shared spare stock, and it goes back on the lobby shelf after that. The shade is warm and even. You switch it on beside the bed for reading at 9.
Physiological arousal (affective response; Kimi K3; low = absent; domain: allotment garden)	You dig over the last bed at a quick pace, turning the soil fast, then haul two full watering cans up the path at a near jog. A brisk wind whips across the plots and snaps the bean netting. Your neighbour Marta waves from her shed, and you finish by stacking the tools before the first drops of the forecast rain arrive.	You dig over the last bed, turning the soil, then carry two full watering cans up the path. A light air moves across the plots and stirs the bean netting. Your neighbour Marta waves from her shed, and you finish by stacking the tools before the first drops of the forecast rain arrive.
Loss aversion (goal relevance; MiniMax M3; low = opposite; domain: boat rental dock)	You are at the boat rental dock with the key to a small rowboat in your hand, the one the attendant clipped to your wrist after you paid. The attendant comes back and says the boat was double-booked and another renter is on the way. You hand the key back and stand on the dock watching the empty mooring.	You are at the boat rental dock waiting while the attendant finishes with another customer. The attendant comes over and says the last rowboat was double-booked and another renter is on the way. You never get a key clipped to your wrist and stand on the dock watching the empty mooring.
Blame† (causal inference; Muse Spark 1.3; low = absent; domain: school parents' evening)	You are standing by the sign-up table in the school hall during parents' evening. Another parent, Dana, draws a thick line through three open slots to take a longer slot for herself. Dana says she knew the ink would cover those slots and no one else could use them. Two empty pens sit beside her unused. You are holding your coat while the bell rings for the next round.	You are standing by the sign-up table in the school hall during parents' evening. Another parent, Dana, writes her name in one open slot to take a longer slot for herself. Dana says the hall is warm and the list is filling fast. Two empty pens sit beside her unused. You are holding your coat while the bell rings for the next round.
Favorable evaluation (comparison of alternatives; DeepSeek V4.1 Flash; low = absent; domain: bicycle repair shop)	You are at the bicycle repair shop, and the mechanic named Priya sets a lock on the counter. The lock is strong, sturdy, and easy to use, with a thick cable. Priya asks you to try it on the shop door.	You are at the bicycle repair shop, and the mechanic named Priya sets a lock on the counter. The lock is metal, with a key, a cable, and a printed label. Priya asks you to try it on the shop door.

Panel B. Pairs by writer.

Writer	Composed	Passed the automated checks	Kept for rating	Excluded: high not rated above low	Excluded: near-duplicate	Retained
Claude Sonnet	1,523	1,511 (1,461)	1,492	5	36	1,451
DeepSeek V4.1 Flash	1,586	1,503 (1,406)	1,474	2	0	1,472
Grok 4.7	1,564	1,509 (1,501)	1,482	2	0	1,480
Kimi K3	1,852	1,362 (1,164)	1,314	9	0	1,305
MiniMax M3	2,015	1,212 (1,080)	1,173	16	0	1,157
Muse Spark 1.3	1,555	1,473 (1,416)	1,453	7	0	1,446
All writers	10,095	8,570 (8,028)	8,388	41	36	8,311

Note. Panel A. One retained pair for a construct of each process; texts verbatim. Panel B. Composed: all pairs that the writers returned. Passed the automated checks: pairs without a failed check (in parentheses, the count from the final check flags). Kept for rating: pairs within the target number per writer, construct, and type of low narrative. Excluded: pairs whose high narrative the raters did not rate higher than the low narrative on the construct's own scale, and near-duplicates (high narratives with a word-based cosine similarity above .80 within a writer and construct). Retained: pairs used in the analyses. Reasons for failing the automated checks, over all writers: length difference (873 pairs), words that name the construct or its synonyms (297), a mental state named (184), length outside 40 to 80 words (182), rated difference too small (149) or too large (76), and a reference to a study (24).

Table S10

Estimates, 95% Confidence Intervals, Equivalence Tests and Bayes Factors for the Main Results, by Study

Studies 1 to 4: the three main LLMs and pooled estimates.

Study	Result	LLM	Estimate	95% CI	Test (p, BF01, equivalence)
1	Same construct, different writers (convergent r)	Qwen3.6 27B	.60	[.57, .63]	—
1	Same construct, different writers (convergent r)	Gemma 4 31B	.60	[.58, .63]	—
1	Same construct, different writers (convergent r)	OLMo 3.1 32B	.58	[.55, .60]	—
1	Different constructs, same writer (r)	Qwen3.6 27B	.05	[.04, .06]	—
1	Different constructs, same writer (r)	Gemma 4 31B	.06	[.05, .07]	—
1	Different constructs, same writer (r)	OLMo 3.1 32B	.06	[.05, .07]	—
1	Different constructs, different writers (r)	Qwen3.6 27B	.04	[.04, .05]	—
1	Different constructs, different writers (r)	Gemma 4 31B	.05	[.05, .06]	—
1	Different constructs, different writers (r)	OLMo 3.1 32B	.05	[.05, .06]	—
1	Different constructs: same minus different writer (difference in r)	Qwen3.6 27B	.006	[.005, .006]	—
1	Different constructs: same minus different writer (difference in r)	Gemma 4 31B	.006	[.006, .007]	—
1	Different constructs: same minus different writer (difference in r)	OLMo 3.1 32B	.007	[.006, .008]	—
1	Held-out writer's construct identified among 66 (layers 30–50% of depth)	Qwen3.6 27B	91%	[89%, 93%]	—
1	Held-out writer's construct identified among 66 (layers 30–50% of depth)	Gemma 4 31B	92%	[89%, 94%]	—
1	Held-out writer's construct identified among 66 (layers 30–50% of depth)	OLMo 3.1 32B	91%	[88%, 93%]	—
1	Content validity: correct definition (of ten) chosen from the read-out words (logit lens)	Qwen3.6 27B	52%	[44%, 61%]	—
1	Content validity: correct definition (of ten) chosen from the read-out words (logit lens)	OLMo 3.1 32B	61%	[52%, 70%]	—
1	Content validity: correct definition (of ten) chosen from the read-out words (Jacobian lens)	Gemma 4 31B	36%	[28%, 43%]	—
1	Content validity: correct definition (of ten) chosen from the read-out words (Jacobian lens)	Qwen3.6 27B	48%	[40%, 58%]	—
1	Criterion validity: similarity to the representation and difference in the LLM's judgment (r)	Qwen3.6 27B	.44 (395 pairs)	[.36, .52]	permutation $p < .001$
1	Criterion validity: similarity to the representation and difference in the LLM's judgment (r)	Gemma 4 31B	.34 (395 pairs)	[.24, .42]	permutation $p < .001$
1	Criterion validity: similarity to the representation and difference in the LLM's judgment (r)	OLMo 3.1 32B	.41 (395 pairs)	[.33, .49]	permutation $p < .001$
2	Profile from one writer in one LLM matched to its construct among 66	Three main LLMs	93%	[90%, 96%]	—
2	Profile from one writer in one LLM matched to its construct among 66	Qwen3.6 27B	92%	—	—
2	Profile from one writer in one LLM matched to its construct among 66	Gemma 4 31B	94%	—	—
2	Profile from one writer in one LLM matched to its construct among 66	OLMo 3.1 32B	93%	—	—
2	Similarity of the profiles, same construct (cosine)	Three main LLMs	.79	[.76, .82]	—
2	Similarity of the profiles, different constructs (cosine)	Three main LLMs	.00	—	—
2	Similarities among the constructs, correlated between LLMs (Spearman ρ)	Qwen3.6 27B vs Gemma 4 31B	.96	[.94, .97]	—
2	Similarities among the constructs, correlated between LLMs (Spearman ρ)	Qwen3.6 27B vs OLMo 3.1 32B	.94	[.92, .95]	—

Study	Result	LLM	Estimate	95% CI	Test (p, BF01, equivalence)
2	Similarities among the constructs, correlated between LLMs (Spearman ρ)	Gemma 4 31B vs OLMo 3.1 32B	.94	[.91, .95]	—
2	Similarity of constructs within minus between the appraisal processes	Qwen3.6 27B	.026 (within .076, between .050)	[−.001, .058]	permutation $p < .001$
2	Similarity of constructs within minus between the appraisal processes	Gemma 4 31B	.026 (within .087, between .061)	[−.002, .062]	permutation $p < .001$
2	Similarity of constructs within minus between the appraisal processes	OLMo 3.1 32B	.026 (within .088, between .062)	[.000, .063]	permutation $p < .001$
2	Similarity of constructs within minus between the appraisal processes	Words of the narratives	.004 (within .009, between .005)	[−.002, .011]	permutation $p = .005$
3	Activation of the mediator's representation, mediator minus other constructs (d_z; scenario description, layers 30–50%)	Pooled (14 LLMs)	1.62	[1.07, 2.12]	—
3	Activation of the mediator's representation, mediator minus other constructs (d_z; scenario description, layers 30–50%)	Qwen3.6 27B	1.57	[0.96, 2.22]	—
3	Activation of the mediator's representation, mediator minus other constructs (d_z; scenario description, layers 30–50%)	Gemma 4 31B	1.75	[1.17, 2.32]	—
3	Activation of the mediator's representation, mediator minus other constructs (d_z; scenario description, layers 30–50%)	OLMo 3.1 32B	1.53	[0.93, 2.11]	—
3	Experiments in which the mediator's representation was activated more than the other constructs' representations	Range (14 LLMs)	42 to 49 of 60	—	—
3	Experiments in which the mediator's representation was activated more than the other constructs' representations	Qwen3.6 27B	42 of 60	[57%, 81%]	—
3	Experiments in which the mediator's representation was activated more than the other constructs' representations	Gemma 4 31B	48 of 60	[68%, 89%]	—
3	Experiments in which the mediator's representation was activated more than the other constructs' representations	OLMo 3.1 32B	46 of 60	[64%, 87%]	—
3	Activation (d_z) in the experiments whose effects replicated in human participants	Pooled (14 LLMs)	1.97; positive in 12 of 14	[0.90, 3.05]	Wilcoxon $p = .003$ (one-sided)
3	Activation (d_z) in the experiments whose effects failed to replicate in human participants	Pooled (14 LLMs)	1.12; positive in 9 of 14	[0.28, 1.90]	Wilcoxon $p = .018$ (one-sided)
3	Activation (d_z) in the experiments whose effects replicated in human participants	Qwen3.6 27B	1.58; positive in 9 of 14	—	binomial $p = .212$ (one-sided)
3	Activation (d_z) in the experiments whose effects replicated in human participants	Gemma 4 31B	2.26; positive in 13 of 14	—	binomial $p < .001$ (one-sided)
3	Activation (d_z) in the experiments whose effects replicated in human participants	OLMo 3.1 32B	1.67; positive in 11 of 14	—	binomial $p = .029$ (one-sided)
3	Activation (d_z) in the experiments whose effects failed to replicate in human participants	Qwen3.6 27B	1.17; positive in 10 of 14	—	binomial $p = .090$ (one-sided)
3	Activation (d_z) in the experiments whose effects failed to replicate in human participants	Gemma 4 31B	0.90; positive in 10 of 14	—	binomial $p = .090$ (one-sided)
3	Activation (d_z) in the experiments whose effects failed to replicate in human participants	OLMo 3.1 32B	1.19; positive in 11 of 14	—	binomial $p = .029$ (one-sided)
3	Replicated vs failed effects: LLMs' responses (rank-biserial r)	Pooled (14 LLMs)	.64	[.29, .94]	Mann–Whitney $p = .002$ (one-sided)

Study	Result	LLM	Estimate	95% CI	Test (p, BF01, equivalence)
3	Replicated vs failed effects: activation of the mediator's representation (rank-biserial r)	Pooled (14 LLMs)	.16	[−.27, .59]	Mann–Whitney p = .238 (one-sided); 90% CI [−.21, .52], not equivalent within ±.20
3	Responses minus activation (difference in rank-biserial r)	Pooled (14 LLMs)	.48	[−.04, .98]	bootstrap p = .037 (one-sided)
3	Activation during the scenario description (layers 30–50%) and effect on the response, across the 60 experiments (Spearman ρ)	Range (14 LLMs)	−.08 to .15	—	BF01 = 3.33 to 6.21
3	Activation during the scenario description (layers 30–50%) and effect on the response, across the 60 experiments (Spearman ρ)	Pooled (14 LLMs)	.05	[−.11, .21]	90% CI [−.09, .18], equivalent within ±.20
3	Activation and effect on the response (Spearman ρ): manipulation passages, layers 30–50%	Qwen3.6 27B	−.02	[−.28, .23]	BF01 = 6.11; family-wise p = .939 within LLM, p = .994 across LLMs
3	Activation and effect on the response (Spearman ρ): manipulation passages, layers 60–75%	Qwen3.6 27B	.03	[−.22, .28]	BF01 = 6.03; family-wise p = .834 within LLM, p = .973 across LLMs
3	Activation and effect on the response (Spearman ρ): scenario description, layers 30–50%	Qwen3.6 27B	.08	[−.18, .33]	BF01 = 5.21; family-wise p = .699 within LLM, p = .912 across LLMs; 90% CI [−.14, .29], equivalent only within ±.29
3	Activation and effect on the response (Spearman ρ): scenario description, layers 60–75%	Qwen3.6 27B	.09	[−.16, .34]	BF01 = 4.82; family-wise p = .642 within LLM, p = .877 across LLMs
3	Activation and effect on the response (Spearman ρ): outcome measure, layers 30–50%	Qwen3.6 27B	.32	[.07, .53]	family-wise p = .037 within LLM, p = .100 across LLMs
3	Activation and effect on the response (Spearman ρ): outcome measure, layers 60–75%	Qwen3.6 27B	.27	[.01, .49]	family-wise p = .108 within LLM, p = .234 across LLMs
3	Activation and effect on the response (Spearman ρ): last word fragment before the response, layers 30–50%	Qwen3.6 27B	.45	[.22, .63]	family-wise p = .001 within LLM, p = .003 across LLMs
3	Activation and effect on the response (Spearman ρ): last word fragment before the response, layers 60–75%	Qwen3.6 27B	.38	[.14, .58]	family-wise p = .008 within LLM, p = .021 across LLMs
3	Activation and effect on the response (Spearman ρ): manipulation passages, layers 30–50%	Gemma 4 31B	.07	[−.19, .32]	BF01 = 5.40; family-wise p = .720 within LLM, p = .926 across LLMs
3	Activation and effect on the response (Spearman ρ): manipulation passages, layers 60–75%	Gemma 4 31B	.08	[−.18, .32]	BF01 = 5.29; family-wise p = .702 within LLM, p = .919 across LLMs
3	Activation and effect on the response (Spearman ρ): scenario description, layers 30–50%	Gemma 4 31B	.09	[−.17, .33]	BF01 = 4.99; family-wise p = .656 within LLM, p = .893 across LLMs; 90% CI [−.13, .30], equivalent only within ±.30
3	Activation and effect on the response (Spearman ρ): scenario description, layers 60–75%	Gemma 4 31B	.09	[−.17, .33]	BF01 = 5.05; family-wise p = .663 within LLM, p = .897 across LLMs
3	Activation and effect on the response (Spearman ρ): outcome measure, layers 30–50%	Gemma 4 31B	.38	[.14, .58]	family-wise p = .010 within LLM, p = .024 across LLMs
3	Activation and effect on the response (Spearman ρ): outcome measure, layers 60–75%	Gemma 4 31B	.37	[.13, .57]	family-wise p = .011 within LLM, p = .029 across LLMs

Study	Result	LLM	Estimate	95% CI	Test (p, BF01, equivalence)
3	Activation and effect on the response (Spearman ρ): last word fragment before the response, layers 30–50%	Gemma 4 31B	.11	[−.15, .35]	family-wise p = .596 within LLM, p = .851 across LLMs
3	Activation and effect on the response (Spearman ρ): last word fragment before the response, layers 60–75%	Gemma 4 31B	.43	[.20, .62]	family-wise p = .002 within LLM, p = .006 across LLMs
3	Activation and effect on the response (Spearman ρ): manipulation passages, layers 30–50%	OLMo 3.1 32B	.09	[−.16, .34]	BF01 = 4.86; family-wise p = .640 within LLM, p = .883 across LLMs
3	Activation and effect on the response (Spearman ρ): manipulation passages, layers 60–75%	OLMo 3.1 32B	.08	[−.18, .33]	BF01 = 5.23; family-wise p = .697 within LLM, p = .914 across LLMs
3	Activation and effect on the response (Spearman ρ): scenario description, layers 30–50%	OLMo 3.1 32B	.12	[−.14, .36]	BF01 = 4.25; family-wise p = .557 within LLM, p = .822 across LLMs; 90% CI [−.10, .32], equivalent only within ±.32
3	Activation and effect on the response (Spearman ρ): scenario description, layers 60–75%	OLMo 3.1 32B	.10	[−.16, .35]	BF01 = 4.65; family-wise p = .614 within LLM, p = .864 across LLMs
3	Activation and effect on the response (Spearman ρ): outcome measure, layers 30–50%	OLMo 3.1 32B	.15	[−.11, .39]	family-wise p = .438 within LLM, p = .722 across LLMs
3	Activation and effect on the response (Spearman ρ): outcome measure, layers 60–75%	OLMo 3.1 32B	.08	[−.17, .33]	family-wise p = .678 within LLM, p = .904 across LLMs
3	Activation and effect on the response (Spearman ρ): last word fragment before the response, layers 30–50%	OLMo 3.1 32B	.31	[.06, .53]	family-wise p = .044 within LLM, p = .111 across LLMs
3	Activation and effect on the response (Spearman ρ): last word fragment before the response, layers 60–75%	OLMo 3.1 32B	.18	[−.08, .41]	family-wise p = .324 within LLM, p = .596 across LLMs
3	Activation at points of low input proximity to response production (manipulation passages and scenario description, both depths) and effect on the response (mean ρ)	Pooled (3 main LLMs)	.07	[−.11, .26]	90% CI [−.08, .23], not equivalent within ±.20; BF01 = 4.25 to 6.11 in each LLM, point of the input, and depth
3	Outcome measure and last word fragment minus manipulation passages and scenario description (difference in ρ, both depths)	Pooled (3 main LLMs)	.21	[.07, .36]	bootstrap p < .001 (one-sided)
3	Ceiling rule applied (experiments removed in which the comparison condition's response lay at the end of the 1–9 scale toward which the predicted effect moves in more than 30% of items): ρ at the scenario description (layers 30–50%), the outcome measure and the last word fragment (layers 30–50%, 60–75%)	Qwen3.6 27B	scenario description .08; outcome measure .32, .27; last fragment .45, .38	—	experiments retained: 60 (removed: 0)
3	Ceiling rule applied (experiments removed in which the comparison condition's response lay at the end of the 1–9 scale toward which the predicted effect moves in more than 30% of items): ρ at the scenario description (layers 30–50%), the outcome measure and the last word fragment (layers 30–50%, 60–75%)	Gemma 4 31B	scenario description .13; outcome measure .43, .35; last fragment .10, .46	—	experiments retained: 54 (removed: 6)
3	Ceiling rule applied (experiments removed in which the comparison condition's response lay at the end of the 1–9 scale toward which the predicted effect moves in more than 30% of items): ρ at the scenario description (layers 30–50%), the outcome measure and the last word fragment (layers 30–50%, 60–75%)	OLMo 3.1 32B	scenario description .09; outcome measure .14, .08; last fragment .35, .19	—	experiments retained: 58 (removed: 2)

Study	Result	LLM	Estimate	95% CI	Test (p, BF01, equivalence)
3	Outcome-measure coding (Supplemental Note): agreement of the two coders	Two coders	87% of 60 outcome measures; weighted κ = .78	—	—
3	Outcome-measure coding (Supplemental Note): mean code and activation at the outcome measure (layers 30–50%) (Spearman ρ)	Mean of the three main LLMs	.18	[−.08, .42]	p = .179 (two-sided)
3	Outcome-measure coding (Supplemental Note): mean code and effect on the response (Spearman ρ)	Mean of the three main LLMs	.15	[−.10, .40]	p = .238 (two-sided)
4	Share of the manipulation's effect on the response left after removal (mean replacement): the mediator's representation, without the shared directions	Qwen3.6 27B	68% (n = 35)	[55%, 85%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the mediator's representation, all 10 directions	Qwen3.6 27B	60% (n = 34)	[46%, 72%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the representations of the most similar constructs, without the shared directions	Qwen3.6 27B	86% (n = 35)	[65%, 98%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the representations of other (random) constructs, without the shared directions	Qwen3.6 27B	93% (n = 36)	[72%, 107%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): random directions of the same size	Qwen3.6 27B	98% (n = 36)	[94%, 104%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the three directions shared by all constructs	Qwen3.6 27B	67% (n = 36)	[52%, 81%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the mediator's representation, without the shared directions	Gemma 4 31B	39% (n = 29)	[4%, 58%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the mediator's representation, all 10 directions	Gemma 4 31B	24% (n = 29)	[−9%, 42%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the representations of the most similar constructs, without the shared directions	Gemma 4 31B	17% (n = 30)	[6%, 58%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the representations of other (random) constructs, without the shared directions	Gemma 4 31B	83% (n = 30)	[47%, 103%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): random directions of the same size	Gemma 4 31B	84% (n = 30)	[63%, 96%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the three directions shared by all constructs	Gemma 4 31B	43% (n = 30)	[24%, 61%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the mediator's representation, without the shared directions	OLMo 3.1 32B	99% (n = 30)	[79%, 107%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the mediator's representation, all 10 directions	OLMo 3.1 32B	91% (n = 30)	[69%, 110%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the representations of the most similar constructs, without the shared directions	OLMo 3.1 32B	97% (n = 30)	[88%, 107%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the representations of other (random) constructs, without the shared directions	OLMo 3.1 32B	99% (n = 30)	[96%, 105%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): random directions of the same size	OLMo 3.1 32B	99% (n = 30)	[97%, 102%]	—
4	Share of the manipulation's effect on the response left after removal (mean replacement): the three directions shared by all constructs	OLMo 3.1 32B	74% (n = 30)	[61%, 80%]	—

Study	Result	LLM	Estimate	95% CI	Test (p, BF01, equivalence)
4	Induced mediator's representation changed the response more than other constructs' representations (experiments)	Qwen3.6 27B	31 of 52	—	binomial $p = .106$ (one-sided); BF01 = 1.23
4	Induced mediator's representation changed the response more than other constructs' representations (experiments)	Gemma 4 31B	25 of 52	—	binomial $p = .661$ (one-sided); BF01 = 7.17
4	Induced mediator's representation changed the response more than other constructs' representations (experiments)	OLMo 3.1 32B	20 of 52	—	binomial $p = .965$ (one-sided); BF01 = 15.07
4	Induced mediator's representation changed the judgment of the mediator more than other constructs' representations (experiments)	Qwen3.6 27B	34 of 52	—	binomial $p = .018$ (one-sided)
4	Induced mediator's representation changed the judgment of the mediator more than other constructs' representations (experiments)	Gemma 4 31B	32 of 52	—	binomial $p = .063$ (one-sided)
4	Induced mediator's representation changed the judgment of the mediator more than other constructs' representations (experiments)	OLMo 3.1 32B	39 of 52	—	binomial $p < .001$ (one-sided)

Note. The 95% confidence intervals come from bootstraps that resampled the constructs (Studies 1 and 2) or the experiments (Studies 3 and 4), from the Fisher z transformation for single correlations, and from exact (Clopper–Pearson) intervals for counts. BF01 is the Bayes factor for the absence of a correlation (uniform prior) or, for counts of experiments, for a chance rate of .5 against a rate uniformly distributed between .5 and 1. The equivalence bounds are ±.20. For the grouping of the constructs by appraisal process (Study 2), the text reports the permutation test, which shuffles the process labels of the 66 constructs; the bootstrap intervals in this table resample the constructs. Differences between mean similarities are given to three decimals. The family-wise p values come from a max-statistic permutation test over the eight combinations of part of the text and depth in one LLM (within LLM) or in all three main LLMs (across LLMs). For removal, the table gives the median over the experiments in which the manipulation's effect on the response without removal was larger than .3 scale points; the shared directions are the three directions common to the 66 constructs. Pooled estimates are averaged over LLMs, with the experiments resampled jointly for all LLMs.

References

Aaker, D. A., & Keller, K. L. (1990). Consumer evaluations of brand extensions. *Journal of Marketing*, *54*(1), 27–41. https://doi.org/10.1177/002224299005400102

Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of Consumer Research*, *34*(4), 468–479. https://doi.org/10.1086/518544

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 337–371). PMLR. https://proceedings.mlr.press/v202/aher23a.html

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. https://doi.org/10.1017/pan.2023.2

Ashokkumar, A., Hewitt, L., Ghezae, I., & Willer, R. (2026). Large language models can predict the results of social science experiments. *Nature*, *656*(8126), 115–122. https://doi.org/10.1038/s41586-026-10742-x

Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., Adamkovic, M., Adamus, S., Albalooshi, S., Albayrak-Aydemir, N., Alfian, I. N., Alper, S., Alvarez-Solas, S., Alves, S. G., Amaya, S., Andresen, P. K., Anjum, G., Ansari, D., Arriaga, P., . . . Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*(6), 880–895. https://doi.org/10.1038/s41562-022-01319-5

Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives on Psychological Science*, *4*(4), 326–339. https://doi.org/10.1111/j.1745-6924.2009.01134.x

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bauer, M. A., Wilkie, J. E. B., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism: Situational materialism undermines personal and social well-being. *Psychological Science*, *23*(5), 517–523. https://doi.org/10.1177/0956797611429579

Bellezza, S., Gino, F., & Keinan, A. (2014). The red sneakers effect: Inferring status and competence from signals of nonconformity. *Journal of Consumer Research*, *41*(1), 35–54. https://doi.org/10.1086/674870

Berger, J., & Heath, C. (2008). Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, *95*(3), 593–607. https://doi.org/10.1037/0022-3514.95.3.593

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192–205. https://doi.org/10.1509/jmr.10.0353

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., . . . Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, *644*(8078), 1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*(4), 550–558. https://doi.org/10.1037/a0018933

Burger, J. M., & Caldwell, D. F. (2003). The effects of monetary incentives and labeling on the foot-in-the-door effect: Evidence for a self-perception process. *Basic and Applied Social Psychology*, *25*(3), 235–241. https://doi.org/10.1207/s15324834basp2503_06

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between

2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carmon, Z., & Ariely, D. (2000). Focusing on the forgone: How value can appear so different to buyers and sellers. *Journal of Consumer Research*, *27*(3), 360–370. https://doi.org/10.1086/317590

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, *22*(8), 1011–1018. https://doi.org/10.1177/0956797611414726

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, *142*(2), 301–306. https://doi.org/10.1037/a0029288

Chen, J., Xiao, W., Chen, L., & Li, X. (2026). *Agent psychonomicus*: Psychological mechanisms emerge in cognitive agents simulated by large language models [Manuscript submitted for publication].

Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). *Persona vectors: Monitoring and controlling character traits in language models* (arXiv:2507.21509). arXiv. https://doi.org/10.48550/arXiv.2507.21509

Cialdini, R. B., Vincent, J. E., Lewis, S. K., Catalan, J., Wheeler, D., & Darby, B. L. (1975). Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology*, *31*(2), 206–215. https://doi.org/10.1037/h0076284

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, *21*(3), 241–251. https://doi.org/10.1002/bdm.586

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Cui, Z., Li, N., & Zhou, H. (2025). A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, *5*(8), 627–634. https://doi.org/10.1038/s43588-025-00840-7

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cummins, J. (2026). The threat of analytic flexibility in using large language models to simulate human data. *Advances in Methods and Practices in Psychological Science*, *9*(3), Article 25152459261461505. https://doi.org/10.1177/25152459261461505

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171. https://doi.org/10.1177/1745691615605826

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., . . . Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Fishbach, A., & Dhar, R. (2005). Goals as excuses or guides: The liberating effect of perceived goal progress on choice. *Journal of Consumer Research*, *32*(3), 370–377. https://doi.org/10.1086/497548

Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, *4*(2), 195–202. https://doi.org/10.1037/h0023552

Gao, L., Wheeler, S. C., & Shiv, B. (2009). The "shaken self": Product choices as a means of restoring self-view confidence. *Journal of Consumer Research*, *36*(1), 29–38. https://doi.org/10.1086/596028

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*(6080), 493–496. https://doi.org/10.1126/science.1215647

Giessner, S. R., & Schubert, T. W. (2007). High in the hierarchy: How vertical location and judgments of leaders' power are interrelated. *Organizational Behavior and Human Decision Processes*, *104*(1), 30–44. https://doi.org/10.1016/j.obhdp.2006.10.001

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science*, *346*(6209), 632–635. https://doi.org/10.1126/science.1253932

Gourville, J. T. (1998). Pennies-a-day: The effect of temporal reframing on transaction evaluation. *Journal of Consumer Research*, *24*(4), 395–408. https://doi.org/10.1086/209517

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505–520. https://doi.org/10.1037/a0013748

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology*, *67*(4), 627–637. https://doi.org/10.1037/0022-3514.67.4.627

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Griskevicius, V., Tybur, J. M., & Van den Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology*, *98*(3), 392–404. https://doi.org/10.1037/a0017346

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026, July 6). Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/workspace/index.html

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hart, W., & Albarracín, D. (2011). Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science*, *22*(2), 261–266. https://doi.org/10.1177/0956797610395393

Hershfield, H. E., Goldstein, D. G., Sharpe, W. F., Fox, J., Yeykelis, L., Carstensen, L. L., & Bailenson, J. N. (2011). Increasing saving behavior through age-progressed renderings of the future self. *Journal of Marketing Research*, *48*(SPL), S23–S37. https://doi.org/10.1509/jmkr.48.spl.s23

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280–1300. https://doi.org/10.1037/0003-066X.52.12.1280

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from Homo silicus?* (arXiv:2301.07543v1). arXiv. https://doi.org/10.48550/arXiv.2301.07543

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, *11*(2), 107–121. https://doi.org/10.1002/(sici)1099-0771(199806)11:2<107::aid-bdm292>3.0.co;2-y

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 20617–20642). PMLR. https://proceedings.mlr.press/v235/huh24a.html

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hyman, H. H., & Sheatsley, P. B. (1950). The current status of American public opinion. In J. C. Payne (Ed.), *The teaching of contemporary affairs: 21st yearbook of the National Council of Social Studies* (pp. 11–34). National Council of Social Studies.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166. https://doi.org/10.1177/01461672952111004

Joel, S., Sakaluk, J. K., Kim, J. J., Khera, D., Qin, H. Y., & Stanton, S. C. E. (2025). Pseudo effects: How method biases can produce spurious findings about close relationships. *Psychological Science*, *36*(10), 780–795. https://doi.org/10.1177/09567976251370262

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science*, *20*(9), 1169–1174. https://doi.org/10.1111/j.1467-9280.2009.02426.x

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325–1348. https://doi.org/10.1086/261737

Karetnikov, N., Rahwan, I., & Svetinovic, D. (2026). Large language models as human proxies. *Nature Computational Science*, *6*(10), 1057–1066. https://doi.org/10.1038/s43588-026-01060-3

Kay, A. C., Laurin, K., Fitzsimons, G. M., & Landau, M. J. (2014). A functional basis for structure-seeking: Exposure to structure promotes willingness to engage in motivated

action. *Journal of Experimental Psychology: General*, *143*(2), 486–491. https://doi.org/10.1037/a0034462

Kivetz, R., & Simonson, I. (2002). Earning the right to indulge: Effort as a determinant of customer preferences toward frequency program rewards. *Journal of Marketing Research*, *39*(2), 155–170. https://doi.org/10.1509/jmkr.39.2.155.19084

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J. T., Cromar, R., Gardiner, G., Gosnell, C. L., Grahe, J., Hall, C., Howard, I., . . . Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), Article 35271. https://doi.org/10.1525/collabra.35271

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194. https://doi.org/10.1093/analys/63.3.190

Labroo, A. A., & Lee, A. Y. (2006). Between two brands: A goal fluency account of brand evaluation. *Journal of Marketing Research*, *43*(3), 374–385. https://doi.org/10.1509/jmkr.43.3.374

Lin, Z. (2025). Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, *8*(3), Article 25152459251357566. https://doi.org/10.1177/25152459251357566

Lindquist, K. A., MacCormack, J. K., & Shablack, H. (2015). The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, *6*, Article 444. https://doi.org/10.3389/fpsyg.2015.00444

Lorge, I., & Curtiss, C. C. (1936). Prestige, suggestion, and attitudes. *The Journal of Social Psychology*, *7*(4), 386–402. https://doi.org/10.1080/00224545.1936.9919891

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology*, *45*(3), 216–222. https://doi.org/10.1027/1864-9335/a000187

Marks, S., & Tegmark, M. (2024). *The geometry of truth: Emergent linear structure in large language model representations of true/false datasets* (arXiv:2310.06824). arXiv. https://doi.org/10.48550/arXiv.2310.06824

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences, USA*, *121*(41), Article e2322420121. https://doi.org/10.1073/pnas.2322420121

McKinney, W. (2010). Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Mead, N. L., Baumeister, R. F., Stillman, T. F., Rawn, C. D., & Vohs, K. D. (2011). Social exclusion causes people to spend and consume strategically in the service of affiliation. *Journal of Consumer Research*, *37*(5), 902–919. https://doi.org/10.1086/656667

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, USA*, *120*(13), Article e2215907120. https://doi.org/10.1073/pnas.2215907120

Mogilner, C., & Aaker, J. (2009). "The time vs. money effect": Shifting product attitudes and decisions through personal connection. *Journal of Consumer Research*, *36*(2), 277–291. https://doi.org/10.1086/597161

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, *22*(3), 453–460. https://doi.org/10.1016/j.jcps.2011.08.002

nostalgebraist. (2020, August 31). Interpreting GPT: The logit lens. *LessWrong*. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, *4*(5), 326–334. https://doi.org/10.1017/s1930297500001170

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). *Generative agent simulations of 1,000 people* (arXiv:2411.10109v1). arXiv. https://doi.org/10.48550/arXiv.2411.10109

Park, K., Choe, Y. J., & Veitch, V. (2024). The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 39643–39666). PMLR. https://proceedings.mlr.press/v235/park24c.html

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*(6), 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M.,

Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates.

Peck, J., & Shu, S. B. (2009). The effect of mere touch on perceived ownership. *Journal of Consumer Research*, *36*(3), 434–447. https://doi.org/10.1086/598614

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pfungst, O. (1911). *Clever Hans (the horse of Mr. von Osten): A contribution to experimental animal and human psychology* (C. L. Rahn, Trans.). Henry Holt.

Pham, M. T., & Avnet, T. (2004). Ideals and oughts and the reliance on affect versus substance in persuasion. *Journal of Consumer Research*, *30*(4), 503–518. https://doi.org/10.1086/380285

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, *66*, 29–38. https://doi.org/10.1016/j.jesp.2015.09.012

Python Software Foundation. (2024). *Python* (Version 3.13) [Computer software]. https://www.python.org

Reb, J., & Connolly, T. (2007). Possession, feelings of ownership and the endowment effect. *Judgment and Decision Making*, *2*(2), 107–114. https://doi.org/10.1017/s1930297500000085

Reed, A., II. (2004). Activating the self-importance of consumer selves: Exploring identity salience effects on judgments. *Journal of Consumer Research*, *31*(2), 286–295. https://doi.org/10.1086/422108

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024). Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15504–

15522). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.828

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*(2), 293–307. https://doi.org/10.1037/0022-3514.95.2.293

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, *12*(3), 185–190. https://doi.org/10.1111/1467-9280.00334

Rucker, D. D., & Galinsky, A. D. (2008). Desire to acquire: Powerlessness and compensatory consumption. *Journal of Consumer Research*, *35*(2), 257–267. https://doi.org/10.1086/588569

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, *5*(6), 359–371. https://doi.org/10.1111/j.1751-9004.2011.00355.x

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly*, *5*(1), 91–92. https://doi.org/10.1086/265467

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434. https://doi.org/10.1111/j.1467-9280.2007.01917.x

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388–395. https://doi.org/10.1086/268936

Sela, A., & Berger, J. (2012). Decision quicksand: How trivial choices suck us in. *Journal of Consumer Research*, *39*(2), 360–370. https://doi.org/10.1086/662997

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*(4), 546–556. https://doi.org/10.3758/bf03197186

Shu, S. B., & Peck, J. (2011). Psychological ownership and affective reaction: Emotional attachment process variables and the endowment effect. *Journal of Consumer Psychology*, *21*(4), 439–452. https://doi.org/10.1016/j.jcps.2011.01.002

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). *Emotion concepts and their function in a large language model* (arXiv:2604.07729). arXiv. https://doi.org/10.48550/arXiv.2604.07729

Soman, D. (2001). Effects of payment mechanism on spending behavior: The role of rehearsal and immediacy of payments. *Journal of Consumer Research*, *27*(4), 460–474. https://doi.org/10.1086/319621

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845–851. https://doi.org/10.1037/0022-3514.89.6.845

Tak, A. N., Banayeeanzade, A., Bolourani, A., Kian, M., Jia, R., & Gratch, J. (2025). Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 13090–13120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.679

Team Olmo, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., Morrison, J., Poznanski, J., Lo, K., Soldaini, L., Jordan, M., Chen, M., Noukhovitch, M., Lambert, N., Walsh, P., . . . Hajishirzi, H. (2025). *Olmo 3* (arXiv:2512.13961). arXiv. https://doi.org/10.48550/arXiv.2512.13961

Thaler, R. H. (1985). Mental accounting and consumer choice. *Marketing Science*, *4*(3), 199–214. https://doi.org/10.1287/mksc.4.3.199

Thompson, D. V., Hamilton, R. W., & Rust, R. T. (2005). Feature fatigue: When product capabilities become too much of a good thing. *Journal of Marketing Research*, *42*(4), 431–442. https://doi.org/10.1509/jmkr.2005.42.4.431

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 79–98). Erlbaum.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, *7*(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z

White, K., & Argo, J. J. (2009). Social identity threat and consumer preferences. *Journal of Consumer Psychology*, *19*(3), 313–325. https://doi.org/10.1016/j.jcps.2009.03.007

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, *322*(5901), 606–607. https://doi.org/10.1126/science.1162548

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zaval, L., Keenan, E. A., Johnson, E. J., & Weber, E. U. (2014). How warm days increase belief in global warming. *Nature Climate Change*, *4*(2), 143–147. https://doi.org/10.1038/nclimate2093

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., & Zhou, J. (2025). *Qwen3 Embedding: Advancing text embedding and reranking through foundation models* (arXiv:2506.05176). arXiv. https://doi.org/10.48550/arXiv.2506.05176

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*(2), 197–206. https://doi.org/10.1086/651257

Zhu, W., Zhang, Z., & Wang, Y. (2024). Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 62638–62681). PMLR. https://proceedings.mlr.press/v235/zhu24o.html

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., . . . Hendrycks, D. (2023). *Representation engineering: A top-down approach to AI transparency* (arXiv:2310.01405). arXiv. https://doi.org/10.48550/arXiv.2310.01405