

Empirical Evidence of Large Language Model’s Influence on Human Spoken Communication

Hiromu Yakura^{*†1}, Ezequiel Lopez-Lopez^{*†2,3}, Levin Brinkmann^{*†1}, Ignacio de la Serna¹, Lara Kirfel¹, Prateek Gupta¹, Ivan Soraperra¹, Thomas F. Eisenmann¹, Dirk U. Wulff^{2,4}, and Iyad Rahwan^{*1}

¹*Center for Humans and Machines, Max-Planck Institute for Human Development, Berlin, Germany*

²*Center for Adaptive Rationality, Max-Planck Institute for Human Development, Berlin, Germany*

³*Center Synergy of Systems, TUD Dresden University of Technology, Dresden, Germany*

⁴*Department of Business Analytics and Decision Science, Vienna University of Economics and Business, Vienna, Austria*

Abstract

From the printing press to social media, innovations in communication technology have repeatedly reshaped how ideas spread through human culture. Chatbots powered by generative artificial intelligence constitute a new medium, encoding cultural patterns in their neural representations and disseminating them in conversations with hundreds of millions of people. Whether these patterns transmit into human language, and ultimately shape human culture, is a fundamental question. While fully quantifying the causal impact of a chatbot like ChatGPT on human culture is challenging, lexical shifts in human spoken communication may offer an early indicator. Here we show that words preferentially generated by ChatGPT, such as *delve*, *showcase*, *boast*, *intricacies* and *meticulous*, increased abruptly in spontaneous human speech. A synthetic-control analysis [1] of 737,083 hours of conversation from 824,634 podcast episodes, screened for unscripted speech, causally links this shift to ChatGPT’s release. The measurable influence on spontaneous speech suggests that humans internalize the lexical choices of large language models (LLMs). A preregistered experiment ($N = 496$) confirms they do, as a brief chatbot interaction led participants to adopt its words as their own, persisting past a distractor task and confirmed in forced lexical choice, indicating entrenchment in the active vocabulary. Together these results show that machines trained on human data now feed their own traits back into human language, integrating LLMs into the ongoing processes of cultural evolution [2]. This coupling raises concerns about linguistic homogenization [3] and the capacity of a few major AI providers for latent cultural influence at scale.

Communication technologies have long altered how knowledge and practices arise, spread, and persist—the process of cultural evolution [4, 5]. Writing, the printing press, broadcast media,

^{*}Corresponding authors: {yakura, lopez, brinkmann, rahwan}@mpib-berlin.mpg.de.

[†]These authors contributed equally to this work.

and the Internet each reshaped human culture in distinctive ways [6, 7, 8, 9]. Generative AI, particularly Large Language Models (LLMs) such as ChatGPT, now emerges as the next such technology [10]: a medium that reaches a global audience in an integrated voice with distinctive linguistic characteristics [11, 12, 13]. As with previous communication media, its influence may extend beyond its own outputs to the language humans themselves produce, including patterns that are internalized and reproduced spontaneously [14, 15, 16]. This prompts a fundamental question: are the outputs of generative AI internalized, signifying cultural transmission?

LLMs are a structurally novel medium for cultural transmission. Unlike books and newspapers, their outputs are largely non-verbatim—each interaction reconstructs rather than copies—resembling oral cultural transmission more than the faithful copying of print [17], and users engage with these reconstructions much as they would with knowledgeable interlocutors [18], a setting that might activate established social-learning strategies [19, 20]. Yet unlike peer-to-peer cultural transmission, LLMs collapse multiple voices into one, concentrated in a small number of providers [21, 3] and delivered through parallel one-to-one conversations. While such cultural transmission resists direct measurement, its reproduction in spontaneous spoken language offers a distinctive empirical signal.

Within two months of release, ChatGPT had accumulated more than 100 million users [22], with adoption proliferating across the English-speaking world and beyond. ChatGPT’s pronounced preference for words such as *delve* offers a unique opportunity to observe and quantify its cultural influence in real time, within a quasi-experimental setting [12, 23]. The observation was made possible by a rare window of measurement: for roughly 18 months, millions of users interacted with a single dominant language model, GPT-3.5, whose unusually distinctive lexical preferences (Fig. 1D) created a measurable signature that could be tracked as it spread through human communication.

Each LLM interaction exposes users to a unique response, but these responses are drawn from the same underlying distribution, providing repeated opportunities for their patterns to be transmitted. Such repeated exposure is known to cause durable changes in language production through entrenchment, a process that operates automatically and below the level of explicit attribution [16, 24]. If LLM exposure affects human language production at this level, it would situate LLMs inside the cultural-transmission network alongside humans [2], where the distinction of AI-generated, AI-influenced, and unshaped human language is increasingly blurred [25, 11, 26].

In this study, we provide complementary empirical evidence of LLM-mediated linguistic influence. At the population level, we use the November 2022 ChatGPT release as a natural experiment, documenting a measurable shift in the spontaneous use of ChatGPT-favored words in over a million hours of human spoken communication. At the individual level, a preregistered controlled experiment isolates entrenchment as a proximate mechanism, showing that short interactions with an AI chatbot induce persistent lexical shifts that survive a distractor task and operate below the threshold of explicit attribution. Together, these results indicate that LLMs have a measurable influence on human language and point to the integration of LLMs into ongoing processes of cultural evolution [2, 27].

Comparing Word Preferences of Humans and LLMs

ChatGPT is trained on broad public corpora [28] and fine-tuned through opaque proprietary processes [29], producing an emergent linguistic profile shaped by statistical learning, reinforcement, and alignment objectives [30, 31]. While rooted in human language, this profile exhibits distinctive characteristics that set it apart from organic human communication [32, 33], with a persistent preference for normative, socially desirable patterns [34, 35].

At the lexical level, where word choice is itself a key aspect of cultural behavior, ChatGPT exhibited distinctive lexical characteristics that reflect its training and optimization [36, 37]. A striking example is *delve*, which was favored over alternatives such as *explore* or *examine* [12, 23]. To quantify these characteristics, we computed a per-word *GPT score* from word-level log-odds ratios between human-written texts and their GPT-edited counterparts, aggregated across datasets, models, and rephrasing prompts [11] (Fig. 1C). *Delve* sits at the top of the resulting distribution across early GPT models, alongside a broader cluster of GPT-preferred words including *showcase*, *intricacies*, and others (Fig. 1D).

Such preferences have already left a measurable mark on written language, where *delve* rose across scientific abstracts and peer reviews soon after ChatGPT’s release [11]. But this mark may never be internalized. In writing, a ChatGPT-preferred word can be introduced into a text even if it never enters the writer’s own vocabulary. While in written text, LLMs may supply lexical choices directly through full production or copy-pasting, spontaneous speech excludes such shortcuts. Words surfacing in unscripted talk would signal internalization by the speaker, signifying machine-to-human cultural transmission [2, 38].

To estimate the population-level causal impact of ChatGPT’s release on spoken communication, we treat the launch as a natural experiment and apply the synthetic control method [1, 39]: for each treated word, we compute its monthly relative document frequency (the fraction of podcast episodes containing the word per month) and build a counterfactual from a convex combination of untreated donor words whose pre-release frequency trajectory best matches that of the treated word (Fig. 1E). This per-word, time-series design sits in the text-as-outcome branch of the causal-inference-for-text literature [40, 41], distinct from document-level methods that estimate the causal effect of latent linguistic properties under no-unobserved-confounding assumptions on text representations [42, 43]; matching on usage rather than meaning is necessary for identification, since semantically similar words are most likely to share the treatment and would contaminate the counterfactual (see [Estimating causal influence of ChatGPT](#)).

In a first approximation on 360,445 YouTube academic talks, the same lexical signature was already detectable, as anticipated from its written footprint in scientific communication [11] and consistent with concurrent work on academic conference talks [44] (see [Population-level replication in academic YouTube talks](#) for details). Yet, academic talks are often scripted or delivered from prepared notes. This motivated a complementary corpus capturing spontaneous speech across diverse domains beyond the academic domain.

Podcasts are well-suited to this question because they frequently contain spontaneous conversation, and individual word choices are largely selected in the moment, even when the broader format is prepared. We therefore assembled 1,407,131 podcast episodes spanning science, technology, education, business, sports, and a topic-agnostic snapshot (see [Constructing datasets of human spoken communication](#)). To restrict the analysis to episodes that genuinely carry spontaneous communication, we trained a classifier to detect spontaneity from audio (Fig. 1B; $\approx 90\%$ accuracy on held-out annotations; see also [Spontaneity annotation and classification](#)), retaining only the conversational and unscripted subset on which the analyses that follow are based.

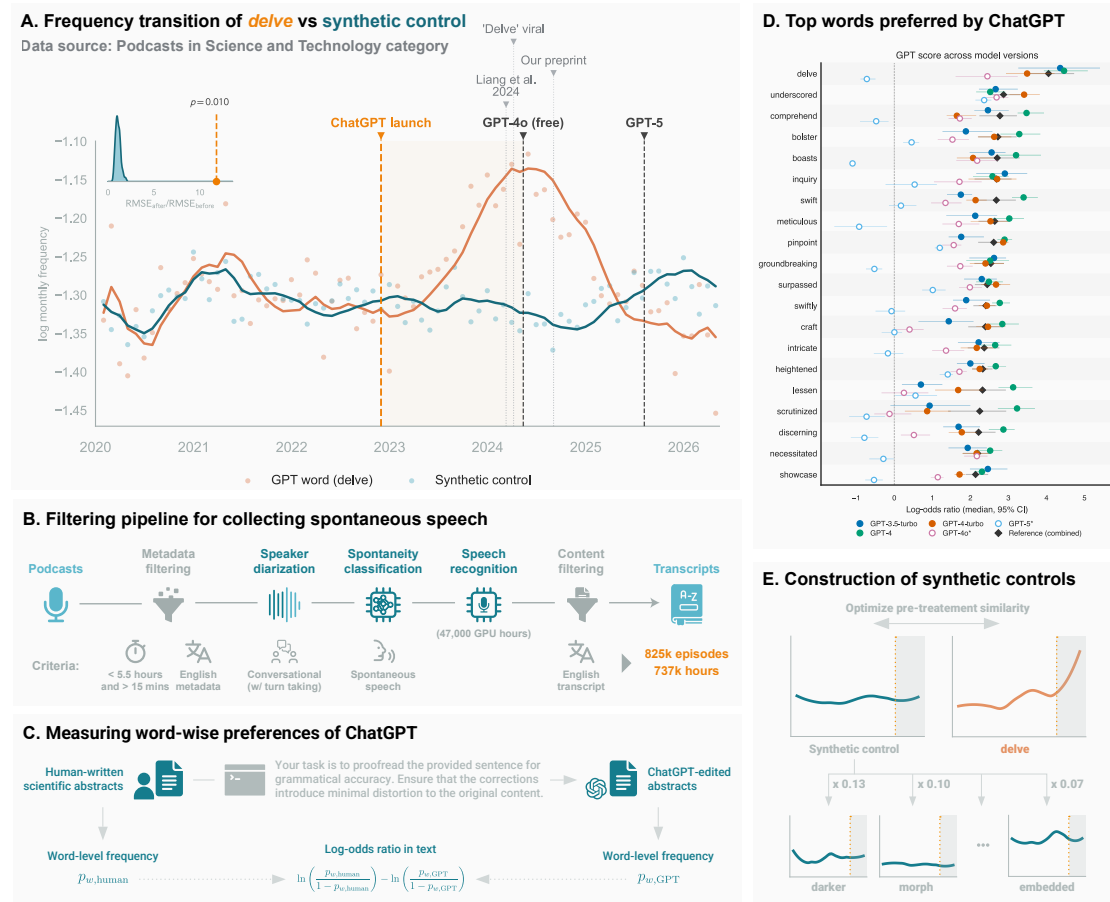


Figure 1: ChatGPT’s word preferences and their measurable influence on spoken communication. (A) Monthly relative frequency of *delve* (the fraction of podcast episodes containing the word per month) in the Science & Technology category (orange) and its synthetic control (teal), with LOWESS trend curve overlaid on the monthly observations. The shaded region marks the post-release window; the inset shows the placebo distribution of post-/pre-treatment RMSE ratios across the donor pool, with the observed ratio for *delve* marked ($p = 0.01$). Vertical dashed lines mark major LLM-related events. (B) Filtering pipeline for the podcast corpus, retaining only episodes that pass duration and language checks, contain conversational turn-taking, and are classified as spontaneous speech by a custom classification model (see [Spontaneity annotation and classification](#) for details). (C) Systematic method for measuring ChatGPT’s word-level preferences: human-written scientific abstracts and their ChatGPT-edited counterparts are compared via word-level log-odds ratios, yielding the *GPT score*. (D) GPT scores for the top 20 GPT-preferred words across model versions; the combined reference score (black diamonds) is computed from GPT-3.5-turbo, GPT-4, and GPT-4-turbo, the three models available at the time the score set was defined. GPT-4o and GPT-5 (star-marked) are shown for comparison and are not included in the reference. *delve* sits at or near the top, with markedly attenuated preference with GPT-5. (E) Construction of a synthetic control: a convex combination of donor words optimized to match the treated word’s pre-treatment trajectory, illustrated for *delve*.

Relationship between ChatGPT’s word preferences and human adoption

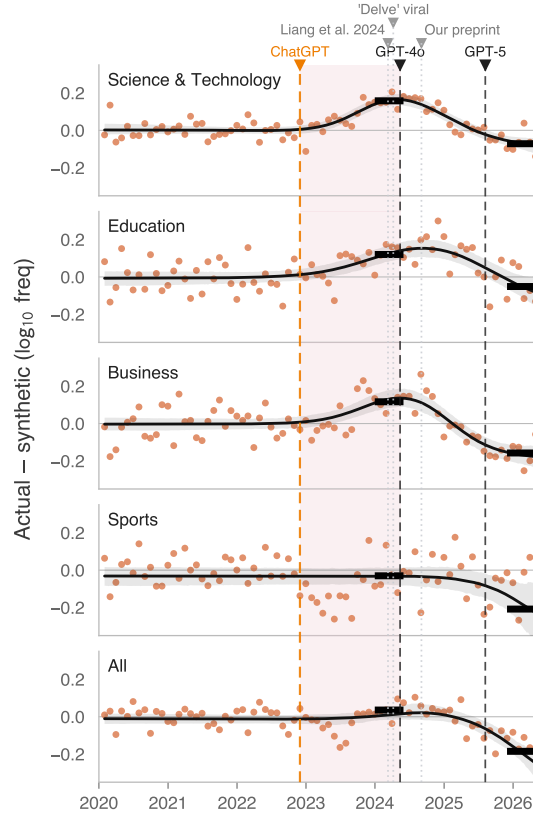


Figure 2: ***Delve* rises after the ChatGPT release and then reverts in multiple podcast categories.** Monthly difference between the observed and synthetic-control relative frequency of *delve* (the fraction of podcast episodes containing the word per month; actual – synthetic; positive values indicate use above the pre-ChatGPT baseline) for four podcast categories and a category-independent sample (*All*). Points are monthly observations; the black curve is a double-logistic *smoother*—the S-curve form expected for lexical change [45]—and the grey band its pointwise posterior. The dashed line marks the ChatGPT release (30 November 2022) and the shaded strip the GPT-3.5 era (release to the free release of GPT-4o); the two black markers are the mean gap over the post-adoption window (months 13–18 after release) and over the last six recorded months. Vertical markers denote a selection of external events. Per-category point estimates, placebo-based 95% confidence intervals, and *p*-values are in Supplementary Table S4.

For *delve*—the word most consistently overused across models and contexts—usage in Science & Technology podcasts (restricted to conversational, spontaneous speech; see [Methods](#)) rose following the release of ChatGPT (Fig. 1A) to ~44% above its synthetic-control counterfactual over months 13–18 post-release (95% CI [+22%, +63%], in-space placebo *p* = 0.010; see [Estimating causal influence of ChatGPT](#) in Methods); the rejection holds under alternative donor-selection

strategies (Supplementary Fig. S1). The result is not driven by a small number of high-usage channels: removing the top-usage channels leaves the post-release rise essentially unchanged (Supplementary Fig. S6), and the pattern mirrors the uptake we observe in YouTube academic talks (Appendix B).

Consistent with the pattern in Science & Technology, Education and Business podcasts showed mean elevations of 32% ($p = 0.06$) and 31% ($p = 0.04$) above the synthetic-control counterfactual over the same window (Fig. 2), with a weak upward tendency in the category-independent sample (+9%) and a slight decline in Sports (−7%). This elevated period falls within the window in which a single model (GPT-3.5-turbo) dominated the consumer market (shaded strip).

The elevation did not persist. The gap peaked around mid-2024 and the usage of *delve* declined in the following months, dropping below baseline in Science & Technology and Education (−15% and −11%) and overshooting further below it in Business (−30%) and the category-independent sample (−35%; 95% CI [−57%, −7%], placebo $p = 0.05$, two-sided). The turn coincides with the free release of GPT-4o and with growing public awareness of *delve* as a signature of LLM language.

Extending the analysis beyond *delve* and raw frequency difference, we investigated 3,535 lexical stems—a stem being the root of a word that reduces to once inflectional endings are stripped (i.e. unifying *delve*, *delves*, and *delving*)—that were present in our GPT-score dataset, part of the top 50k word2vec words, and appearing in at least 20 episodes per month on average over the pre-treatment window. For simplicity, we will use stem and word interchangeably in the following. For each treated word w we tested whether the actual–synthetic gap $\Delta y_{w,t} = \log_{10} y_{w,t}^{\text{obs}} - \log_{10} y_{w,t}^{\text{synth}}$ departs from its pre-treatment trajectory at the ChatGPT release. We use the post-event slope coefficient β_{Post} of a Bayesian change-point regression (Methods, Eqn. 1) on $\Delta y_{w,t}$ as our per-word estimate of the ChatGPT-induced change in usage; positive values indicate post-release acceleration of the treated word’s frequency relative to its synthetic control.

We found higher GPT-scores to be associated with an increase in usage in spoken communication (Fig. 3B). For the top 1% of GPT-score words ($n = 36$) the mean post-release slope β_{Post} reaches +0.030; 28 of 36 show an increase in usage, 13 of them credibly so (95% HDI excludes zero), against 8 decreases of which 2 are credibly so. Fig. 3A shows the 12 words in the top 1% with the largest credible change (see SI for a full overview, Supplementary Fig. S2). While *delve* and others revert towards baseline in recent months, words such as *boast*, *meticulous*, and *showcase* maintain a sustained uptake.

The effect is strongest at the highest-scoring words. The mean post-release slope β_{Post} is +0.030 in the top 1% of GPT scores, +0.025 in the top 2%, +0.015 in the top 5%, and +0.013 in the top decile—corresponding to a median peak usage ~50% above the synthetic-control baseline at the top 1%. The top- k mean rises faster than its permutation null as k narrows onto the highest-GPT-score words (Fig. 3C). This argues against a vocabulary-wide drift and confirms an association between higher GPT scores and increased usage in spoken communication. The same concentration is visible under alternative donor-selection strategies and on un-audited counts (Supplementary Fig. S3), and does not depend on *delve* alone (β_{Post} in the top 1% excluding *delve*: +0.027; Supplementary Fig. S7).

To further check whether the acceleration is anchored in time to ChatGPT’s release, we considered every month in the data window as a candidate change point and refit the cross-word regression at each. For 28 candidate change points whose 18-month post-window pre-dates ChatGPT, the mean top-1% slope β_{Post} was flat ($+0.0004 \pm 0.003$); it rose to +0.033 only once the window reached the actual launch and peaked at +0.042 in September 2023 (permutation $p = 0.034$; Supplementary Fig. S5).

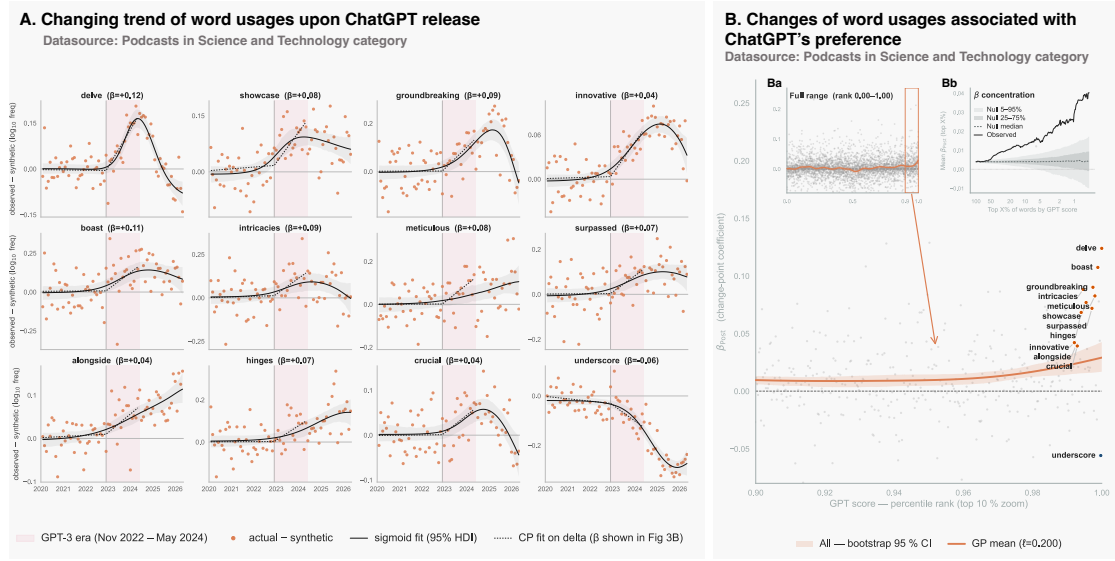


Figure 3: Word-usage shifts after the ChatGPT release track ChatGPT’s lexical preference (Science & Technology podcasts). (A) **Changing trend of word usages upon ChatGPT release.** Of the 36 words in the top 1% of GPT scores, 15 showed a credible change in observed – synthetic $\log_{10}$ frequency (orange points) at ChatGPT’s launch. Shown here are the twelve with the largest-magnitude conservative β_{Post} bound (the 95% HDI limit nearest zero). Solid curves are double-sigmoid posterior smoothers [45] (95% HDI shaded); dashed lines are the change-point fit of Eqn. 1, whose post-release slope β_{Post} (panel titles) is the quantity plotted across all words in panel B. The shaded vertical band marks the period of analysis between ChatGPT’s launch and the launch of GPT-4o (free). Eleven of the twelve rise after release; *underscore* is the lone decline. (B) **Changes of word usages associated with ChatGPT’s preference.** Per-word change-point slope β_{Post} (Eqn. 1) plotted against GPT-score percentile rank ($n = 3,535$ words). The main panel zooms on the top 10% of GPT scores; gray points represent individual words and the orange line shows the Gaussian-process posterior mean over all words (Matérn $\nu = 2.5$, length scale $\ell = 0.20$ on the unit-interval rank axis; bootstrap 95% CI shaded), which rises towards the top of the GPT-score range. The twelve words shown in A are labeled. Inset Ba shows the same relationship across the full GPT-score range (rank 0–1). Inset Bb displays the slice-mean β_{Post} over top- $X\%$ slices of the GPT-score distribution (black line) against the permutation null (grey band, 5th–95th percentile; shuffling GPT scores across words). The slice-mean rises faster than its permutation null, climbing from +0.004 over the whole vocabulary to +0.030 in the top 1%, which corresponds to more than 2× the permutation-null upper 95% bound.

A brief chatbot interaction entrenches lexical choices

The observational findings above establish a population-level shift in spoken word use following ChatGPT’s release, but leave open the individual-level question: whether direct chatbot interaction alone is sufficient to produce lasting lexical change in individual speakers. To answer this, we conducted a controlled experiment ($N = 496$; see Fig. 4 Top and Methods). The study was pre-registered under <https://aspredicted.org/k38u44.pdf>. In this study, participants played a chat-based picture selection game with an AI co-player, modeled after referential communication

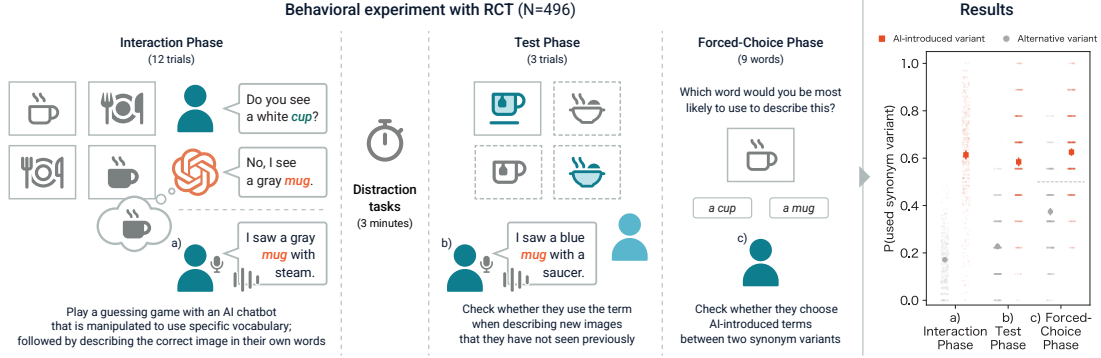


Figure 4: **Experimental design and results** ($N = 496$). Top panels: the four sequential phases of the experiment. *Interaction Phase* (12 trials): participants played a referential image-guessing game with an AI chatbot covertly prompted to use specific synonym variants (e.g., *mug* instead of *cup*), then gave a spoken description of the target image. *Distractor Phase* (3 min): unrelated arithmetic and visual pattern-matching tasks. *Test Phase*: participants gave spoken descriptions of novel images not seen during the AI interaction, with no chatbot present. *Forced-Choice Phase* (9 words): participants chose between two synonym labels for a depicted object. Right panel: probability of using the AI-introduced synonym variant (orange squares) versus an alternative variant (grey circles) across the three outcome phases. Note that in the interaction phase and the test phase, values do not necessarily sum to one, as participants might also use both or neither term. Small translucent dots show individual participant means; large markers show group means $\pm$ 95% CI; dashed line at 0.5 indicates chance. The AI-introduced variant was adopted well above chance in the Interaction Phase and remained elevated after the distractor task in descriptions of entirely novel images.

game paradigms [46, 47]. During the game, the AI co-player used one of two synonym variants (Variant 1 vs. Variant 2) to describe picture content, and participants had to identify a target image based on these descriptions (see Fig. S13).

Participants’ *own* spoken descriptions of the target images during the Interaction Phase showed strong alignment with the AI-introduced vocabulary. Participants were substantially more likely to use a given word variant when the AI had used it than when the AI had used the alternative (61% vs. 17% for AI-introduced vs. other variant; $b = 0.44$, 95% CI [0.42, 0.46], $p < .001$, Monte Carlo permutation test; Supplementary Fig. S14), indicating that exposure to the AI’s word choices significantly increased participants’ likelihood of adopting them. Because participants were randomly assigned to hear one variant or the other of each pair, the differential adoption isolates the AI’s causal contribution.

Following a 3-minute distractor task, participants had to describe nine novel images (i.e., images never seen during the previous Interaction Phase) so that a hypothetical future human co-player could identify the depicted image, with no AI chatbot present. As hypothesized, the AI-induced vocabulary uptake persisted in these spontaneous descriptions ($b = 0.36$, 95% CI [0.33, 0.38], $p < .001$; 58% vs. 23% for AI-introduced vs. other variants; Supplementary Fig. S15). The effect generalized to new picture material and survived cognitive distraction, consistent with durable lexical entrenchment rather than within-context repetition [48, 49]. In a subsequent forced-choice task in which participants selected their preferred expression from the two variants, they chose the AI-introduced variant on 63% of trials ($b = 0.13$, 95% CI [0.11, 0.14], $p < .001$; Supplementary Fig. S16). This provides converging evidence of an explicit

lexical preference shift. The uptake of AI-introduced variants held across three distinct lexical categories—nouns, verbs, and adjectives—in both spontaneous production and explicit label choice (see Fig. 4 Bottom).

An open-ended detection check administered after the Test Phase revealed that only 15 of 496 participants (3.0%) noticed the chatbot’s vocabulary pattern; most attributed the discrepancy to regional or stylistic variation rather than a deliberate constraint. The remaining 481 (97.0%) reported nothing unusual or noted only incidental features such as response speed or punctuation. These responses suggest that the observed lexical shift was implicit rather than deliberate.

Discussion

Together, our findings indicate that lexical features exhibited by ChatGPT are internalized into spontaneous human speech at a population scale. By analyzing 737,083 hours of transcribed podcast episodes, we reveal a measurable surge in words preferred by ChatGPT—including *delve*, *boast*, and *meticulous*—with a causal association to its public release. While the strongest and earliest signal appears in academic-adjacent domains (Science & Technology), where exposure to LLM-shaped text might be the highest, indications of spreading to other domains, such as education, business, and the category-independent sample, suggest subsequent filtering into the general public. Restricting the analysis to podcast episodes featuring unscripted, conversational, spontaneous discourse shows that the shift extends beyond scripted or formal speech. A subset of these words then exhibits a subsequent partial moderation, with usage falling back towards, and in some cases below, the pre-ChatGPT baseline, suggesting a more complex dynamic in which initial adoption is followed by selective avoidance once these words become culturally marked as AI-associated and model providers react by changing the words their models favor.

A behavioral experiment suggests an individual-level mechanism for the shift observed in podcasts. A short text-based AI chatbot interaction induces lexical shifts in participants’ subsequent spontaneous speech production. The effect survives cognitive distraction and generalizes to new contexts. We observe the use of AI-introduced words increasing by 36 percentage points, a magnitude comparable to lexical-alignment effects reported in studies of human-to-human dialogue [50, 51]. A separate forced-choice task, likewise administered after the distraction tasks, provides converging evidence: participants select the AI-introduced words in 63% of cases. Our results show that humans not only converge lexically with their conversation partner during a dialogue, whether human [48, 52] or AI [53, 54, 55], but also carry AI-introduced word choices into subsequent spontaneous speech once the interaction ends. This points to entrenchment rather than transient alignment, with repeated exposure strengthening the lexical representation in long-term memory and increasing its activation probability in future production [16, 24]. Such imitation operates at high fidelity as a matter of convention, without an instrumental payoff [56].

The mechanism we observe is consistent with a long-standing psycholinguistic picture, in which lexical access in spontaneous speech is automatic rather than strategic [14, 15] and repeated exposure entrenches lexical representations below the threshold of explicit attribution [16, 24], as our open-ended detection check confirms. What is notable is how little the filters that normally make social learning selective apply to it. Humans readily defer to algorithmic systems on capability-defined tasks [57, 58], while algorithm aversion persists in identity-laden domains [59, 60]. For LLMs, which collapse authorship into a uniform voice, such task-type moderation [61] is plausibly dampened, and evolved heuristics for source evaluation—such as prior accuracy, prestige, and expertise [20, 62]—might find less purchase. Instead, an LLM’s lexical choices are read as articulate and authoritative, lending a word like *delve* the appearance of sophisticated expression, and it is this perceived value, in the words themselves, on which such variants are

selected [63]. The consequence is that an algorithmic system may come to function as a cultural model—a source people learn from.

The podcast corpus and the experiment together cast LLMs as sources of cultural variants that humans internalize, signifying a coupled human–machine cultural process. The boundary between human- and machine-authored text is already eroding across scientific writing, academic discourse, and online communication [11, 12, 13, 64, 25], and our results show that LLM-shaped patterns enter even unassisted human production. The concern that successor LLMs trained on increasingly LLM-shaped corpora may degrade in output diversity [65]—so-called model collapse—was previously thought to be partly contained by original human language serving as an external anchor for the training distribution [66]. Our findings, however, suggest that human language can no longer be treated as an independent external anchor; rather, human and machine cultural production form a single, integrated system [2]. This integration is not structurally symmetric. Human–AI interaction has a hub-and-spoke topology in which many users converse with few generative AI systems, closer to early broadcast media than to the peer-to-peer networks of the online era. Such structures concentrate exposure and are known to amplify influence, with changes at highly connected nodes propagating to individuals never directly exposed to the source [67, 68, 69, 70]. We illustrate the asymmetry with a noisy voter model on a small-world network with one committed source. A hub committed to the favored variant drives population-level uptake well beyond a randomly placed speaker of equal commitment, reaching agents with no direct hub contact (Fig. S17; see Supplementary Methods).

The observed lexical shifts are specific in kind. Lexical shifts can have a range of causes. Social media diffuses novel lexical items [71], world events such as the COVID-19 pandemic drive topical spikes, and new technologies drive the rise of vocabulary tied to the practices they enable, from the telephone and the radio to the search engine [72]. Our own corpus shows both of these last two patterns. The COVID-19 vocabulary (*pandemic*, *vaccine*, *mask*) surged in 2020–2021 and declined in the following years; machine-learning-adjacent technical vocabulary (*gpu*, *python*, *vector*) climbed with the growth of the field over 2022–2024; and the post-2022 rise of *prompt* tracks the new activity of interacting with LLMs (Supplementary Fig. S10). The words we identify—*delve*, *boast*, *meticulous*, and the other top-1% GPT-preferred words—fit none of these categories. They have no obvious referential connection to language models, and they occupy a frequency band where natural-rate lexical change is slow [73].

The dynamics of LLM influence on language are more complex than adoption alone. Across the GPT-score distribution the effect is unidirectional: ChatGPT-preferred words accelerate in spoken use, but ChatGPT-disfavored words do not show a corresponding decline (Supplementary Fig. S4). The trajectory is also not monotone since adoption reverses for some words. The word *delve* in particular drops sharply once it is discussed in social and traditional media, settling below its pre-ChatGPT baseline. OpenAI also reacted, with later GPT versions removing *delve* during text-editing operations and driving its usage below the human baseline (see Fig. 1D). A plausible mechanism comes from sociolinguistics. Beyond content, lexical choices signal group membership and authenticity [74]. Words that become culturally legible as AI-associated may therefore attract avoidance after initial uptake, as speakers distance themselves from forms that threaten authenticity, a pattern reminiscent of hypercorrection [75]. These signaling dynamics may also have consequences for social stratification. While LLMs lower linguistic barriers for non-native speakers seeking to communicate in formal English [76, 77], adopting LLM-marked vocabulary now risks new stigmas. Words such as *delve* may come to be stereotypically associated with lower skill or with uncritical AI use, reshaping perceptions of credibility and competence.

Several limitations qualify these results. The corpus is English-only and is limited to a self-selected, public-facing population of podcast hosts and their guests. The main analysis covers the first 18 months after ChatGPT’s release and treats OpenAI’s GPT models as the dominant

driver; LLM deployment has since fragmented across models and providers, making attribution increasingly complex. Reverse causality (ChatGPT amplifying emerging human preferences) is addressed by the in-time placebo showing the spoken shift is specific to ChatGPT’s launch date (Fig. S5), and the random assignment in the experiment ruling it out at the individual level. Because ChatGPT could in principle affect all words in non-trivial ways, we cannot rule out residual interference in our synthetic control, even though we restricted donors to words with near-zero GPT scores and excluded the closest synonyms. However, even if the donors are partially affected, the differential effect (i.e., between high-GPT-score treated words and near-zero-GPT-score donors) still reflects ChatGPT’s preferences in subsequent human usage. The finding that these preferences propagate into spoken language therefore stands. In addition, the experimental evidence is limited to a sample recruited online via Prolific, which consisted of self-identified native English speakers. Comprising a single test session after a brief distractor, the experimental window captures only an early signal of AI-induced lexical entrenchment.

We have shown how AI shapes human language quietly. Its lexical choices are entering spontaneous spoken communication, with entrenchment during chat interactions as a plausible mechanism. Unlike documented cases of deliberate imitation of algorithmic solutions [38] or AI-driven persuasion [78, 79], the channel demonstrated here is incidental—operating below explicit attribution. As chatbots are increasingly used from education to therapy, these findings raise the question of which other, more consequential associations and thinking patterns might be transmitted along the same channel. The stakes are high. The narrow set of dominant AI systems may compress both the diversity of cultural variants and the choice between sources—eroding the variation and selection on which cultural evolution depends. The cultural dynamics that emerge from the feedback of human and machine cultural production—in which AI systems both draw from and reshape human language—thereby open a new research endeavor, complementing the study of machines and humans in isolation. The question is no longer whether machines influence us, but in which way, through which channels, and under whose control.

Methods

We tested whether ChatGPT’s lexical preferences propagate from generated text into spontaneous human speech using a per-word causal-inference design. Each word’s affinity for ChatGPT output is quantified via a log-odds score (the GPT score); per-word post-release usage shifts are estimated by combining synthetic-control matching with Bayesian change-point regression, with specificity tested by in-word-space and in-time placebo procedures.

Constructing datasets of human spoken communication

To capture spontaneous spoken communication, we systematically constructed a dataset of podcast transcripts spanning multiple categories.

Data collection

Our sampling of podcasts was designed to trace how any linguistic shift propagates outward from its likely point of entry. Science and Technology was our primary target, as the domain closest to the documented written-language cases of the phenomenon. Around it we sampled categories at increasing conceptual distance: Education and Business, which overlap with Science and Technology in theme and audience, and Sports, a more distant and characteristically spontaneous domain. To place these targeted categories against the wider landscape, we additionally collected a broad snapshot of all remaining categories in the catalog.

We ran a first exploratory data collection, in which episodes were drawn from a database of over four million series, randomly sampling 6,000 episodes per quarter *for each* of five candidate categories (Business, Education, Religion and Spirituality, Science and Technology, and Sports), restricted to English-language episodes published between January 2017 and late 2024, and yielded 771,591 transcribed episodes. A conversational screen—at least two distinct speakers and four or more exchanges over a ten-minute slice—showed that the yield of genuine dialogue varied sharply by category: Science and Technology retained 40.8% of collected episodes and the remaining general categories roughly 60% to 75%, whereas Religion and Spirituality retained only 16.2%, reflecting a predominance of single-speaker, monologic delivery. Learning from this, we dropped Religion and Spirituality up front—together with Books, which consists almost entirely of scripted audiobook readings—rather than collecting and then discarding the bulk of those episodes, giving a cleaner sampling frame.

Building on this, the present study substantially enlarges the corpus and sharpens the spontaneity criterion. We drew candidate feeds from the PodcastIndex public catalog of approximately 4.4 million podcast series,¹ using the snapshot dated 5 April 2026, and retrieved episode-level metadata through the PodcastIndex API.² We restricted the collection to English-language episodes published between 1 January 2017 and mid-April 2026, mapping provider-supplied labels onto broader, general-purpose categories.³ Sampling was stratified by calendar quarter and seeded for reproducibility, so that temporal coverage is balanced across the study period rather than dominated by recent, higher-volume years. This collection comprises 1,407,131 candidate episodes (approximately 1.0 million hours of audio), of which 931,450 passed the dialogue screen and were transcribed—about 20% more transcribed episodes than the earlier collection, drawn from a broader set of categories. As described below, we also replaced the earlier dialogue-only screen with a two-stage spontaneity filter that adds a trained audio classifier, giving a finer and better-validated separation of spontaneous from scripted speech; 824,634 episodes met this stricter criterion. A small fraction of episodes could not be retrieved or were unavailable in a usable audio format and were discarded, and the number of usable episodes decreases further through the filtering pipeline described below.

Filtering and transcription

To maximize the number of transcripts we can obtain with limited GPU resources, we implemented the following filtering criteria. We first removed podcast episodes shorter than 15 minutes, since they often include non-conversational speech content. Additionally, we excluded episodes that exceeded 20,000 seconds (approximately 5.5 hours, which fell around the 99th percentile of duration in an early downloaded subset) to avoid unnecessary GPU occupation by rare extreme-length episodes.

Importantly, our intention in using podcasts was to specifically analyze the influence of LLMs on spontaneous communication. We therefore filtered episodes in two stages. First, we applied speaker diarization, which partitions audio into segments labeled by speaker identity, to a 10-minute slice extracted from the middle of each episode, using the `pyannote` library [80, 81]. We required at least two distinct speakers and four or more exchanges (alternating turns) to retain an episode, which removed pure monologues and read-out broadcasts before further processing. The dialogue filter alone, however, does not completely separate spontaneous interaction from scripted exchange (e.g., scripted interviews or co-hosted read-throughs). We therefore complemented diarization with an audio-based classifier of *spontaneity*, trained against human annotations on a

¹PodcastIndex feed database: https://public.podcastindex.org/podcastindex_feeds.db.tgz

²PodcastIndex API: <https://podcastindex-org.github.io/docs-api>

³Apple Podcasts categories: <https://podcasters.apple.com/support/1691-apple-podcasts-categories>

1–4 scale ranging from “clearly scripted” to “clearly spontaneous”. The annotation followed the disfluency-based protocol of Cho et al. [82] (filler words, repetitions and corrections, hesitations, and incorrectly used words); three naïve coders rated three 30-second samples per episode after a two-round calibration, and the interrater reliability reached $ICC = 0.87$ (95% CI 0.81–0.91; see [Annotation protocol](#)). We then trained a classifier on top of the `Whisper-large-v3` encoder [83], following Elisha et al. [84], so that it maps the 25×30 -second middle window of each episode to a soft distribution over the four annotation classes (see [Audio-based spontaneity classifier](#)), and aggregated chunk predictions into an episode-level continuous score in [1, 4]. For all analyses, an episode was retained if (i) it passed the diarization-based dialogue filter and (ii) its spontaneity score exceeded 3.0, which corresponds to the inflection point of the labeled validation curve. We confirmed that the retained episode set is essentially unchanged under an alternative, majority-vote-based aggregation of the same chunk-level predictions (see [Spontaneity annotation and classification](#)).

The transcription of the collected data was performed using the `large-v3` model of `WhisperX` [85], a faster version of the `Whisper` speech-to-text model [83]. We employed batch processing with the model, achieving an average transcription speed of approximately 2 minutes per hour of audio with Nvidia A100 GPU. Here, we opted to run the transcription process ourselves rather than use pre-existing transcript data from YouTube or other podcast platforms, given the possibility that they have switched transcription models over time.⁴ As a result, the same model, configured identically, was applied uniformly to every episode across the entire time window of the study, while this configuration achieves a word error rate of approximately 5% averaged across common English audio datasets [83]. Importantly, since the model is fixed across the time axis, any residual transcription error contributes a constant background that cannot, by construction, generate a change point at the ChatGPT release date or differentially favor the GPT-preferred words over their synthetic controls. The recognition language was fixed to English throughout transcription, consistent with the English-language restriction applied at collection through the podcast feeds’ language metadata.

Preprocessing

We preprocessed the obtained transcripts to capture essential changes in word frequency by removing noise and highlighting relevant patterns. We followed a systematic procedure:

1. **Tokenization:** The text is divided into individual tokens (words) for processing.
2. **Normalization:** All words are converted to lowercase to ensure uniformity and avoid duplication due to case differences.
3. **Stop word removal:** Commonly used words that do not carry significant semantic meaning, such as *and*, *the*, and *is*, are removed. The list of stop words used in this process is sourced from the Natural Language Toolkit (NLTK) library [86], which provides a standard set of English stop words.
4. **Non-alphabetic filtering:** Words containing non-alphabetic characters are excluded, ensuring only standard words are retained.
5. **Length filtering:** Words with fewer than three characters are removed to eliminate overly short and potentially uninformative tokens.
6. **Stemming:** Words are reduced to their root forms using the Porter stemming algorithm [87]. This algorithm applies a series of heuristic rules to iteratively strip suffixes from words

⁴Specifically in the YouTube transcripts, we found an unnatural increase in the frequency of the filler word *um* starting around May 2020, which we found difficult to attribute to an actual increase in speakers’ usage of the word. It is more plausible that YouTube switched to a transcription model that transcribes fillers verbatim, and thus, we conducted the transcription process to avoid a potential source of bias.

(e.g., *running* to *run*). Since raw stems are often non-words (e.g., *delve* stems to *delv*), figures and tables relabel each stem with a representative surface form for readability.

For subsequent analyses, we calculated the log relative frequency of podcast episodes containing each presented word, sampled monthly. Older data may exhibit different word usage trends due to factors such as the relatively low number of published podcast episodes. Hence, we analyzed data spanning six years before the initial release of ChatGPT on November 30, 2022. Additionally, due to the timing of data collection, the corpus includes podcast episodes published up to April 30, 2026. The change-point regression and synthetic-control estimates are anchored to the 18-month GPT-3.5 era ending May 30, 2024 (the month GPT-4o replaced GPT-3.5 as the default on ChatGPT’s free tier); post-2024 data is used to characterize the reversion dynamics reported in Results. We employed log frequency to facilitate trend interpretation within this early diffusion phase, using Laplace smoothing [88] to account for zero counts, which helps detect emerging patterns that may initially exhibit exponential growth.

Measuring word preferences of large language models

We investigated the word preferences of commonly used LLMs by prompting various models to edit a diverse set of human-authored texts. Building on prior research [11, 12], we analyzed differences in word frequencies between original human-written texts and their LLM-edited versions. Our analysis spans a wide range of human texts, prompts, and models, enabling the computation of an aggregated *GPT score*.

Creation of contrastive datasets

We compiled datasets from diverse sources, all predating the introduction of ChatGPT. These included 7,182 abstracts from arXiv (2019–2022) using the arXiv API, 2,880 abstracts from bioRxiv (2019–2022) via the bioRxiv API, over 8,000 abstracts from Nature (2019–2023) collected through its search engine, 10,000 samples each from the Enron email dataset (2000–2001), Hewlett Foundation student essays (2012) and Wikipedia articles (2019–2022), and 2,000 spontaneous-speech podcast transcript excerpts predating ChatGPT. Detailed dataset creation steps are provided in [Datasets to compute Log Odds Ratios of human and LLM word usage](#).

To assess how prompts influence model word preferences, we used three standard prompts across all datasets and models:

- **Prompt 1:** Please polish this text: {text}
- **Prompt 2:** Can you improve this text: {text}?
- **Prompt 3:** Please rephrase this text to improve its clarity: {text}

As Prompt 3 frequently altered the content of emails, we extended the prompt to include: “It’s an email so please don’t change the structure of the text.” in that specific case.

We preprocessed both the original human texts and their LLM revisions using the same procedure applied to transcript datasets of YouTube videos and podcasts. For robustness, we considered only words whose human or LLM document frequency, pooled across all dataset–model–prompt strata, reached at least one per mille of the pooled document count, and excluded prompt-related stems (*rephrase*, *polish*, *dear*, *text*, *certainly*, *subject*, *readable*, *clarity*, *enhance*, *version*, *title*) that were frequently repeated in the LLM’s responses. Our analysis also included different GPT-family models, of which GPT-3.5-turbo, GPT-4, and GPT-4-turbo were the production models available at the time the GPT-score set was defined and used to compute the reference GPT score below. To check whether the same preferences persist in later models, we additionally evaluated GPT-4o and GPT-5 after that, while we did not include them in the

reference score so that the set of treated words used in the downstream causal analysis remains anchored to the pre-release model family.

Log-odds ratio estimation

To identify words preferentially associated with LLMs, we computed log-odds ratios comparing word frequencies in human-authored and ChatGPT-edited corpora. For each word w , we estimated its document frequency in human (p_{human}) and ChatGPT (p_{GPT}) corpora using Laplace smoothing to mitigate zero-count issues:

$$p_w = \frac{\text{number of documents containing word } w + 1}{\text{total documents} + 1}.$$

The log-odds transformation was applied to these smoothed probabilities:

$$\text{log-odds}(p) = \ln \left(\frac{p}{1-p} \right),$$

yielding the log-odds ratio (LOR) for each word:

$$\text{LOR}_w = \text{log-odds}(p_{w,\text{GPT}}) - \text{log-odds}(p_{w,\text{human}}).$$

Positive LOR values indicate higher prevalence in ChatGPT-edited texts, while negative values suggest human-associated usage. This metric was computed independently across all dataset–model–prompt combinations. When estimating p_w we set the denominator to the maximum number of returned documents across (model, prompt) combinations for each dataset to avoid inflation through occasional API refusals or empty completions.

We define these probabilities on *document frequency*—whether a word occurs in a document—rather than on token counts, because our interest is in how *widely* a word is used rather than how often, this Bernoulli definition measures the prevalence of a word across units of communication. It is immune to within-document repetition and less sensitive to document length than a token-count measure, both of which would otherwise inflate the estimated preference for a word. It also coincides with the definition of *GPT score*.

Calculation of a weighted GPT score

We present word preferences for a range of GPT-family models and document types (see [A](#)). For our main analysis, we focus on scientific abstracts (arXiv, bioRxiv, Nature) and the GPT chat models available before GPT-4o (GPT-3.5-turbo, GPT-4, and GPT-4-turbo; Table [S1](#)), and we developed a GPT score that marginalizes over uncertainties in model usage patterns. Given the unknown true distribution of ChatGPT’s usage across datasets (D), models (M), and prompts (P), we adopted a Bayesian hierarchical model with non-informative Dirichlet priors:

$$\begin{aligned} P(D) &\sim \text{Dirichlet}(\mathbf{1}), \\ P(M \mid D) &\sim \text{Dirichlet}(\mathbf{1}), \\ P(P \mid D, M) &\sim \text{Dirichlet}(\mathbf{1}), \end{aligned}$$

where $\mathbf{1}$ denotes flat priors for each parameter. The joint distribution $P(D, M, P)$ was computed as:

$$P(D, M, P) = P(D) \cdot P(M \mid D) \cdot P(P \mid D, M).$$

For each of 1000 Monte Carlo samples from this prior, we marginalized the human and ChatGPT smoothed probabilities over (dataset, model, prompt):

$$\hat{p}_w^{(\cdot)} = \sum_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}} \sum_{p \in \mathcal{P}} p_w^{(\cdot, d, m, p)} \cdot \lambda(d, m, p),$$

where $(\cdot)$ is either human or GPT, $\lambda(d, m, p) \propto P(D = d, M = m, P = p)$, and $\mathcal{D}, \mathcal{M}, \mathcal{P}$ index datasets, models, and prompts. The per-sample log-odds ratio is $\text{logit}(\hat{p}_w^{\text{GPT}}) - \text{logit}(\hat{p}_w^{\text{human}})$. The GPT score is the median LOR across the 1000 samples, with uncertainty quantified via 95% percentile intervals.

The Dirichlet prior structure reflects maximum entropy assumptions about potential correlations between datasets, models, and prompts. By sampling from the joint prior distribution, we emulate the variability expected under real-world deployment scenarios where specific GPT-family model, dataset, and prompt combinations are not systematically favored. The resulting GPT scores thus represent robust centrality estimates of word preferences across plausible usage distributions.

Estimating causal influence of ChatGPT

To assess ChatGPT’s causal impact on human verbal communication, we employed the synthetic control method [1, 89]. This method allows us to estimate the usage pattern of a “treated” GPT-preferred word (*i.e.*, a word with a high GPT score) in the counterfactual scenario where ChatGPT was never deployed. This is built on the assumption that words sharing similar pre-release usage patterns would have continued exhibiting comparable patterns in the absence of the release. Thus, the method constructs a synthetic control for each treated word by forming a convex combination of multiple “donor” words whose usage closely tracks the treated word’s pre-release trajectory so that it predicts the counterfactual pattern by extrapolating this combination forward.

Reading the treated vs synthetic gap as the causal effect of ChatGPT’s release requires three assumptions. First, *no anticipation*: the release is an external, discretely timed event that speakers could not have foreseen, so pre-release trajectories are free of treatment, and the six-year pre-release window furnishes an uncontaminated basis for matching. Second, *a good and stable pre-release fit*: the synthetic control must reproduce the treated word’s pre-release trajectory closely and over a long horizon, since a fit that holds only briefly, or that is achieved by chance on a short window, does not support extrapolation past the release; our six-year pre-release window and trajectory-based donor selection (below) are designed to meet this requirement. Third, *no interference* (the stable unit treatment value assumption, SUTVA): the donor words must not themselves be affected by the release, so that their post-release usage reflects the counterfactual rather than a diffuse response to ChatGPT. This is the binding assumption in our setting, since ChatGPT may, in principle, shift all language rather than only the words it most prefers. We cannot guarantee it, but we make it as credible as the design allows by choosing donors that are unlikely to be differentially treated. Specifically, we exclude the treated word’s closest semantic neighbors, which are its plausible substitutes and would inherit a fraction of the same shock, and we restrict the donor pool to words that ChatGPT neither over- nor under-uses (those with near-zero GPT score), removing the words most likely to carry the treatment (see [Donor selection](#)). We discuss the consequences of residual interference, and the interpretation of our results should it remain, in [Discussion](#).

Donor selection

For each treated word, the synthetic control method requires the selection of a set of donor words that are used to build the synthetic control. The candidate pool is the intersection of the top 50,000 most frequent words in the pre-trained word2vec embedding with the words we measured a GPT-score for; restricting to this vocabulary discards rare words keep computation feasible and ensuring well defined GPT-scores. From this pool, we first filtered out the $K = 20$ words most similar to the treated word in a pre-trained word2vec embedding built on the Google News dataset [90]. Close semantic neighbors of the treated word can plausibly be substitutes for it and therefore inherit a fraction of the same treatment shock (potentially in an inverse direction), which would contaminate the counterfactual. We narrowed the pool of potential donors to a symmetric neutral-percentile band around $|\text{GPT score}| = 0$, retaining the central 50% of words by absolute GPT score. Removing words that showed over- or underusage by GPT enforces the stable unit treatment value assumption that underpins the synthetic control estimator.

From the remaining vocabulary, we then selected the $L = 100$ words following the recommendation of Abadie and Bastida [89]. We specifically chose words whose pre-treatment log-frequency trajectories were closest to that of the treated word in pointwise Euclidean distance over the pre-treatment months. To prevent month-to-month sampling noise from dominating this matching step, the pre-treatment trajectories were first smoothed with a Gaussian-process prior (Matérn kernel with $\nu = 5/2$ and a two-year length scale, $\ell = 720$ d). The length scale is chosen as a compromise between the raw monthly series, on which the synthetic overfits noise, and longer length scales at which it underfits the relevant low-frequency dynamics (Supplementary Fig. S9). The smoother enters only at donor selection and the synthetic control fit itself; all downstream change-point and placebo statistics are computed on the raw monthly series.

The synthetic control for the treated word was then constructed as a convex combination of these L donors, with the weights chosen to minimize the pre-treatment root mean squared prediction error between the treated word’s trajectory and the weighted donor average [1] using the smoothed pre-treatment trajectories. We constrained the weights to be non-negative and to sum to one, which rules out extrapolation outside the convex hull of donor trajectories and induces sparsity, so that the fitted weight vector typically had support on a small subset of the pool. Weights were optimized by sequential least-squares programming (SLSQP, `scipy.optimize.fmin_slsqp`) initialized at the uniform vector $\omega_j = 1/L$. Across the treated words in Science & Technology podcasts, the non-negativity and sum-to-one constraints induce sparse SLSQP weight vectors, so the fitted synthetic control typically draws on only a handful of donors, as illustrated for *delve* in Table S3.

We re-ran the entire synthetic-control pipeline under four control specifications that each vary a specific design choice: **C1** drops the semantic-neighbour and GPT-score filters, leaving only ℓ_2 matching on the raw vocabulary; **C2** shrinks the donor pool from $L = 100$ to $L = 10$; **C3** replaces the SLSQP convex fit with deterministic inverse-distance similarity weights, following CausalCite [43], adapted to our time-series setting; **C4** swaps the audited counts substrate for the raw, un-audited counts. Full specifications, per-variant placebo p -values, and side-by-side renderings of Fig. 1A and Fig. 3B across the four controls are given in Supplementary Materials (Table S5, Figs. S1 and S3); the qualitative pattern is preserved across all four.

Placebo test

We assess the significance of the causal effect using the in-space placebo test of Abadie et al. [1] as refined by Ferman and Pinto [91]. For each word in the matching pool, we re-run the synthetic control procedure with that word in the role of the treated word, taking the remaining donors as the new pool, and compute the post- to pre-treatment ratio of mean squared prediction error

(MSPE). The placebo pool size is bounded above by the size of the donor pool: for the Main specification we use $n_{\text{placebo}} = L = 100$ (every donor serves once as a placebo target), while for the C2 robustness specification the ten-word donor pool yields $n_{\text{placebo}} = 10$. The placebo targets are drawn without replacement from the donor pool.

The empirical p -value pools the treated word with its n_{placebo} placebos and ranks the MSPE ratios: $p = (r + 1)/(n_{\text{placebo}} + 1)$, where r is the number of placebos with MSPE ratio at least the treated word's. The floor is therefore $1/(n_{\text{placebo}} + 1)$: $1/101 \approx 0.010$ under the Main specification ($n_{\text{placebo}} = 100$) and $1/11 \approx 0.091$ under C2 ($n_{\text{placebo}} = 10$).

For the window-mean gaps reported in Fig. 2, we apply an analogous inversion of the same in-space placebo distribution. For each placebo target, we compute the mean observed-minus-synthetic log-frequency gap over the window of interest (months 13–18 after release, and the last six recorded months). The 95% confidence interval for the treated word's true effect is obtained by inverting this distribution, $[\hat{g} - q_{97.5}, \hat{g} - q_{2.5}]$, where $\hat{g}$ is the treated word's window-mean gap and q_α denotes the α -quantile of the placebo distribution. We report a one-sided p -value for the predicted post-adoption rise over months 13–18 and a two-sided p -value for the post-2024 reversion, whose sign was observed rather than predicted. Both use the same $(1+r)/(1+n_{\text{placebo}})$ add-one convention. Per-category values for *delve* are reported in Supplementary Table S4.

Change-point regression of the synthetic-control gap

Having constructed a synthetic control for each treated word, we tested whether the word's trajectory departs from that counterfactual after the release of ChatGPT, and by how much. We restricted the set of treated words to those appearing in at least twenty episodes per month on average over the pre-treatment window, which removes targets for which the gap series is too noisy to support a meaningful change-point fit. For each treated word w we formed the monthly gap between its observed and synthetic-control log-frequency,

$$\Delta y_{w,t} = \log_{10} y_{w,t}^{\text{obs}} - \log_{10} y_{w,t}^{\text{synth}},$$

and fitted to it the hierarchical Bayesian change-point regression of Eqn. 1,

$$\Delta y_{w,t} = \alpha + \beta_{\text{Pre}} t + \beta_{\text{Post}} d_{\text{Post}}(t - T_{\text{event}}) + \epsilon_t, \quad d_{\text{Post}} = \begin{cases} 1 & t > T_{\text{event}} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Here t is continuous time measured in years and anchored at the start of the window T_{start} , so the coefficients are interpretable as changes in $\log_{10}$ frequency per year, and the post-release term allows the slope to change at the change point T_{event} (the ChatGPT release, 30 November 2022). The fitted slope is β_{Pre} before the release and $\beta_{\text{Pre}} + \beta_{\text{Post}}$ after it. The post-release change in slope β_{Post} is our estimate of ChatGPT's influence on the usage of w . The fit spans the same 18-month window as the prior analysis.

We placed weakly-informative priors on every parameter: a standard normal $\mathcal{N}(0, 1)$ on each slope coefficient ($\beta_{\text{Pre}}, \beta_{\text{Post}}$), a diffuse $\mathcal{N}(0, 10)$ on the intercept α , and a half-Cauchy(0, 10) prior on the standard deviation σ of the Gaussian residual ϵ_t . Each word was fitted independently by Hamiltonian Monte Carlo using Stan's no-U-turn sampler via `cmdstanpy` at its defaults, with four chains of 1,000 post-warmup draws each; warmup draws were discarded. We summarized every coefficient by its posterior mean and the 95% highest-density interval (HDI).

We compare the slice-mean β_{Post} over nested top- $X\%$ slices of the GPT-score distribution against a permutation null (Fig. 3C). X is sampled on geometric grid from 100% to 0.5%; at each point we generate $n_{\text{perm}} = 1,000$ permutations and report the 5th–95th percentile of the permuted slice mean.

Controlled referential communication experiment

Participants Five hundred participants completed the experiment via Prolific (mean age = 40.5 years, $SD = 13.2$; 50% women; 91% reporting English as their first language; see Supplementary Table S6 for full demographics). The study was preregistered on AsPredicted (#k38u44; <https://aspredicted.org/k38u44.pdf>). Sample size was determined by a power analysis targeting $d = 0.30$ with 80% power ($N_{\text{required}} = 352$); we recruited above this target to accommodate exclusions. Four participants were excluded for self-reported color vision deficiency, yielding a final $N = 496$.

Design The experiment used a 2 stimulus group (Group A/B, *between-subject*) $\times$ 2 mode (synonym Variant 1 vs. Variant 2 as AI-used, *between-subject*) $\times$ 12 trial (*within-subject*) design. The stimulus group determined which nine of the 18 synonym pairs the AI was instructed to use during the Interaction Phase. Participants were randomly assigned to one of the two stimulus groups. Within their group, participants were then randomly assigned to one of two word mode sets (Variant 1 or 2).

Materials Stimuli in total comprised 18 synonym pairs spanning three lexical categories: nouns (e.g., *mug/cup*), verbs (e.g., *to fix/to repair*), and adjectives (e.g., *glossy/shiny*). Each stimulus Group A/B consisted of 9 synonym pairs, i.e., three noun pairs, three verb pairs, and three adjective pairs. The mode determined which of the two synonyms within each pair the AI chatbot used consistently across all 9 pairs. For example, a participant assigned to Group A, Variant 1 encountered an AI co-player that always used the words “thermos bottle”, “gift”, “cup”, “to fix”, “to examine”, “to install”, “colorful”, “cracked”, and “spotted”. The full list is provided in Supplementary Table S7. Images were AI-generated using Gemini 3 (*gemini-3-pro-image-preview*) to depict scenes that allowed the AI co-player to describe features in the target image using the prompted variants.

Procedure Participants completed five sequential phases:

(1) Interaction Phase Participants played a picture selection game [modeled after 46, 47] with the GPT-4o-based chatbot that plays the role of a co-player (see Fig. S13). In each trial, the participant viewed six candidate images and had to identify the one target image based on the AI chatbot’s description. Each target image was identifiable by three distinct features (one object, one activity, one attribute), which the AI chatbot describes using its three assigned synonym variants (one noun, one verb, one adjective). Participants selected the image they thought best matched the AI chatbot’s description and received feedback on whether their selection was correct. They then recorded a spoken description of the correct target image themselves. The interaction phase consisted of 12 trials, and each trial featured one noun, one verb, and one adjective synonym variant. Each of the 9 synonym variants appeared four times across all trials.

(2) Distractor Phase Following the Interaction Phase, participants completed a 3-minute distraction task consisting of numerical and spatial reasoning problems [92].

(3) Test Phase In our final Test Phase, participants were presented with three new sets of six images, each with one image highlighted as the target. For each set, participants were tasked with recording a spoken description of the target image, assuming another virtual participant who would try to identify it in a future iteration of the game. Each of the 9 synonym variants was depicted once across the three trials.

Between phases (3) and (4), participants answered an open-ended detection check: “Have you noticed anything about the language of the chatbot?”.

(4) **Forced-Choice Task** Participants were presented with nine images depicting the objects, activities, and attributes corresponding to the nine synonym pairs encountered during the interaction phase. For each image, participants had to indicate how they would prefer to describe what they saw by selecting one of two synonym variants (e.g. “mug” or “cup”).

(5) **Background questionnaire** Finally, participants complete a series of questions covering demographic information, familiarity with and use of AI, and language background and usage.

Throughout the above procedure, participants’ spoken descriptions were transcribed via the browser Web Speech API, and during the Test Phase, the transcripts were then displayed to participants for manual correction and submitted as human-verified transcripts (see [System implementation](#)).

Scoring and analysis Transcriptions were scored using a stemmer-based span-dominance algorithm (see [System implementation](#)). The primary outcome, Δp , was the per-participant difference in AI-variant usage rate between AI-introduced and non-introduced variant pairs. Significance was assessed via Monte Carlo sign-flip permutation test (10,000 permutations, two-tailed); 95% CIs were computed by cluster bootstrap (10,000 replications, resampling by participant); a linear mixed model with random intercepts for participants and word pairs served as a parametric cross-check.

Acknowledgments

H.Y. is supported in part by JST PRESTO Grant Number JPMJPR246B. E.L.-L. is funded by the Deutsche Forschungsgemeinschaft (DFG), project number 458366841 (POLTOOLS - Assisting behavioral science and evidence-based policy making using online machine tools).

Appendix A Word preferences of Large Language Models

Large language models (LLMs) from the GPT family systematically alter word frequencies when revising human-written text [11, 12]. To quantify these word preferences, we computed log-odds ratios (LORs) comparing word frequencies in human-authored texts and their GPT-revised counterparts. We systematically evaluated the sensitivity of this effect to model version (Figure 1D), prompting (Fig. 5A), and source dataset (Fig. 5B; Supplementary Figs. S11 and S12).

Word preference patterns exhibited notable stability across GPT-family models (Figure 1D), suggesting these biases emerge from intrinsic characteristics of the training pipeline rather than version-specific training. However, specifically for *delve*, we found decreasing preference in newer models. For *delve*, the odds ratio in revised arXiv abstracts declines from approximately 380:1 under GPT-3.5-turbo to 100:1 for GPT-4-turbo and 40:1 for GPT-4o. This trend culminates in GPT-5, which exhibits markedly smaller lexical anomalies than any earlier model: most of the signature words are no longer over-used, and *delve* itself falls below its human baseline (odds ratio $\approx 0.7:1$ on arXiv).

These preferences are also robust to the exact rephrasing instruction (Fig. 5A): the three prompts we used (**polish**, **improve**, and **clarify**) yield closely similar LORs for every reference word, with the **clarify** prompt consistently the mildest yet never reversing a preference.

LOR magnitudes varied substantially across source corpora (Fig. 5B). Analysis of log-probability distributions (Supplementary Fig. S11) revealed that this variability stems primarily from base-

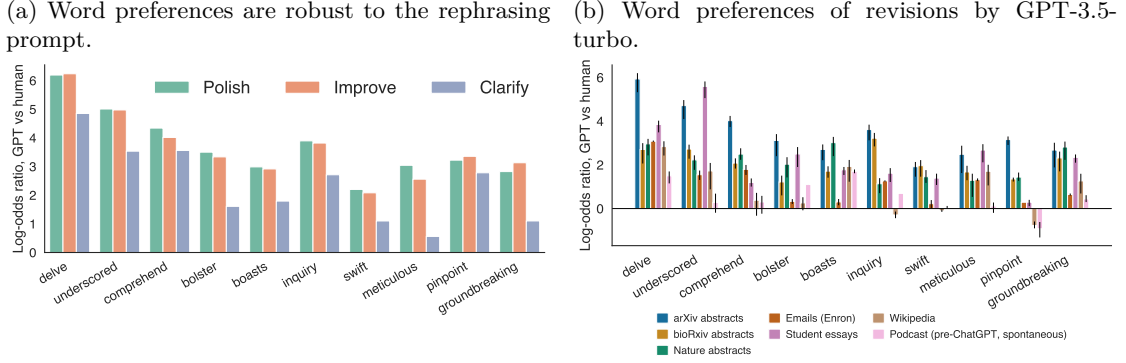


Figure 5: **Log-Odds-Ratios (LORs) of words in human versus LLM-revised text.** (a) Word preferences are largely invariant to the exact rephrasing instruction (*polish*, *improve*, and *clarify*) when revising arXiv abstracts with GPT-3.5-turbo. (b) Substantial variations in LORs emerge when examining revisions of GPT-3.5-turbo across different datasets.

line differences in human word choices. For instance, while humans rarely use *underscore* in essays, GPT revisions introduced this term frequently across domains, including essays.

Focusing on scientific abstracts (arXiv, bioRxiv, Nature) and the GPT chat models available before GPT-4o (GPT-3.5-turbo, GPT-4 and GPT-4-turbo; see Table S1), we computed a weighted GPT score by marginalizing over model, prompt, and dataset combinations (Figure 1D; black diamonds). Here, *delve* emerged as the most strongly overused term ($\text{LOR} > 4$), followed by *underscore*, *comprehend*, *bolster*, *boast*, *inquiry*, *swift*, *meticulous*, *pinpoint* and *groundbreak* ($\text{LOR} > 2.5$).

Appendix B Population-level replication in academic YouTube talks

As a complement to the podcast analysis, we replicate the population-level signal in another spoken corpus: 360,445 academic YouTube talks from the channels of 20,622 research institutions cataloged in the Research Organization Registry [93]. Such talks sit between the written academic record where ChatGPT’s footprint is well documented [11, 12] and the spontaneous speech we study in podcasts, providing an independent test of whether the effect extends beyond conversational audio.

The pipeline mirrors the podcast pipeline (**Constructing datasets of human spoken communication**), with three differences: (i) institutional channels are identified by querying the YouTube Data API and selecting the most plausible match via **gpt-3.5-turbo-0125** (Supplementary Fig. S18); (ii) videos are retained between 20 minutes (the API’s **short** cutoff, below which videos frequently consist of non-speech content such as advertisements or trailers) and the 99th-percentile duration (~ 3.0 hours); (iii) we omit the speaker-diarisation and spontaneity-classifier steps, since academic talks are predominantly prepared, monologic speech. Transcription, synthetic control, and change-point analyses are otherwise identical.

The headline matches the podcast result: the placebo test for *delve* yields $p = 0.010$ (Fig. 6A). The piecewise-linear regression extends the effect to other top GPT-preferred words, such as

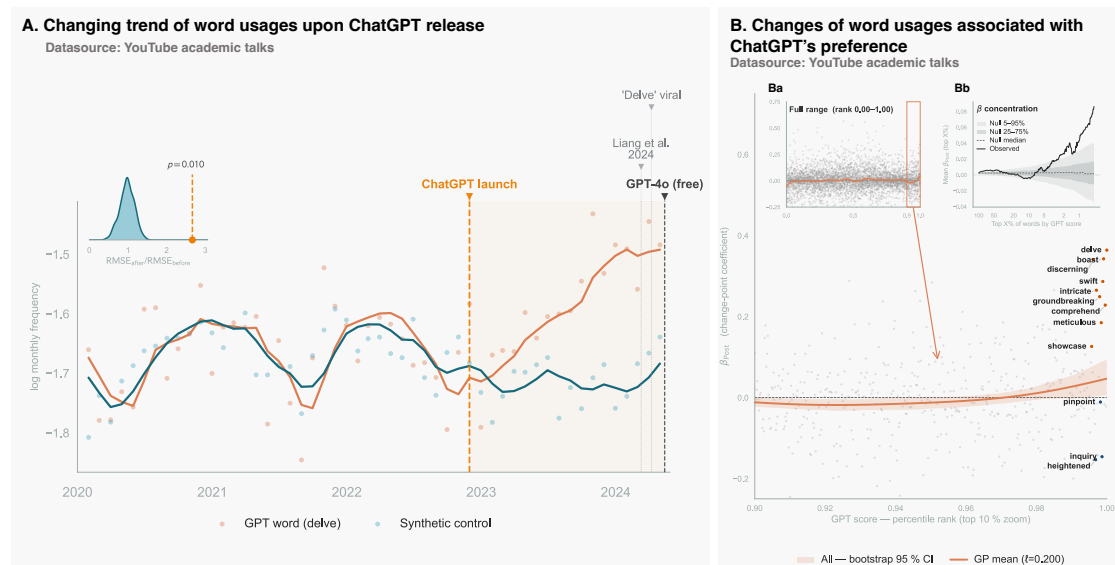


Figure 6: **Population-level trend changes for top GPT-preferred words in academic YouTube talks.** (A) Monthly relative frequency of *delve* in academic YouTube talks (orange) and its synthetic control (teal). The inset shows the placebo distribution of post-/pre-treatment RMSE ratios across the donor pool, with the observed ratio for *delve* marked ($p = 0.010$). The vertical dashed line marks the ChatGPT release. (B) Piecewise-linear trend fits for the top GPT-preferred words, with a change point at the ChatGPT release. Words such as *comprehend*, *boast*, and *swift* show a significant post-release increase, mirroring *delve*.

comprehend, *boast*, *swift*, and *meticulous*; all exhibit a significant post-release uptake (Fig. 6B). ChatGPT’s lexical signature is thus also audible in academic talks, consistent with the stricter spontaneous-speech result reported in the main text.

References

- [1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.
- [2] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray, and Iyad Rahwan. Machine culture. *Nature Human Behaviour*, 7:1855–1868, 2023. doi: 10.1038/s41562-023-01742-2.
- [3] Zhivar Sourati, Alireza S. Ziabari, and Morteza Dehghani. The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences*, jan 2026. doi: 10.1016/j.tics.2026.01.003.
- [4] Marshall McLuhan. *Understanding Media: The Extensions of Man*. McGraw-Hill, 1964.

- [5] Joseph Henrich. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton University Press, 2016. doi: 10.2307/j.ctvc77f0d.
- [6] Jack Goody and Ian Watt. The consequences of literacy. *Comparative Studies in Society and History*, 5(3):304–345, apr 1963. doi: 10.1017/S0010417500001730.
- [7] Robert Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, 2000.
- [8] Jeremiah Dittmar. Information technology and economic change: The impact of the printing press. *The Quarterly Journal of Economics*, 126(3):1133–1172, jan 2011. doi: 10.1093/qje/qjr035.
- [9] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, mar 2018. doi: 10.1126/science.aap9559.
- [10] Alberto Acerbi. Cognitive attraction and online misinformation. *Palgrave Communications*, 5(15):1–12, 2019. doi: 10.1057/s41599-019-0224-y.
- [11] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. Mapping the increasing use of LLMs in scientific papers. In *Proceedings of the 1st Conference on Language Modeling*, 2024.
- [12] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813–eadt3813, jul 2025. doi: 10.1126/sciadv.adt3813.
- [13] Kyle Siler. The diffusion of large language models in published academic articles. *Proceedings of the National Academy of Sciences*, 123(22):e2605754123, jun 2026. doi: 10.1073/pnas.2605754123.
- [14] J. Kathryn Bock. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387, jul 1986. doi: 10.1016/0010-0285(86)90004-6.
- [15] Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–75, 1999. doi: 10.1017/s0140525x99001776.
- [16] Nick C. Ellis. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188, jun 2002. doi: 10.1017/S0272263102002024.
- [17] Alberto Acerbi and Alex Mesoudi. If we are all cultural darwinians what’s the fuss about? Clarifying recent disagreements in the field of cultural evolution. *Biology & Philosophy*, 30(4):481–503, 2015. doi: 10.1007/s10539-015-9490-2.
- [18] Clara Colombatto and Stephen M. Fleming. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013, 2024. doi: 10.1093/nc/niae013.
- [19] Joseph Henrich and Francisco J. Gil-White. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(3):165–196, may 2001. doi: 10.1016/S1090-5138(00)00071-4.

- [20] Rachel L. Kendal, Neeltje Boogert, Luke Rendell, Kevin N. Laland, Mike Webster, and Patricia L. Jones. Social learning strategies: Bridge-building between fields. *Trends in Cognitive Sciences*, 22(7):651–665, jul 2018. doi: 10.1016/j.tics.2018.04.003.
- [21] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, mar 2021. doi: 10.1145/3442188.3445922.
- [22] Krystal Hu. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*, feb 2023. URL <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>.
- [23] Tom S. Juzek and Zina B. Ward. Why does ChatGPT “delve” so much? Exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, jan 2025.
- [24] Holger Diessel. *The Grammar Network*. Cambridge University Press, 2019.
- [25] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David M. Rothschild, and Robert West. Prevalence and prevention of large language model use in crowd work. *Communications of the ACM*, 68(3):42–47, feb 2025. doi: 10.1145/3685527.
- [26] Raluca Rilla, Tobias Werner, Hiromu Yakura, Iyad Rahwan, and Anne-Marie Nussberger. Recognising and mitigating LLM pollution in online behavioural research. *Nature Communications*, 17:5578, jun 2026. doi: 10.1038/s41467-026-74621-9.
- [27] Alberto Acerbi and Joseph M. Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120, oct 2023. doi: 10.1073/pnas.2313790120.
- [28] OpenAI. GPT-4 technical report. *arXiv*, 2023. doi: 10.48550/arXiv.2303.08774.
- [29] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th ACM Conference on Conversational User Interfaces*, number 47, pages 1–6, jul 2023. doi: 10.1145/3571884.3604316.
- [30] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901, dec 2020.
- [31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pages 27730–27744, dec 2022.

- [32] Alberto Muñoz Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and LLM-generated news text. *arXiv*, 2023. doi: 10.48550/arXiv.2308.09067.
- [33] Alex Reinhart, Ben Markey, Michael Laudénbach, Kachata Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122, feb 2025. doi: 10.1073/pnas.2422455122.
- [34] Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1):18617, oct 2023. doi: 10.1038/s41598-023-45644-9.
- [35] Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, 56(6):5754–5770, 2024. doi: 10.3758/s13428-023-02307-x.
- [36] Shangbin Feng, Chan Young Park, Yuhang Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11737–11762, jul 2023. doi: 10.18653/v1/2023.acl-long.656.
- [37] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, sep 2024. doi: 10.1038/s41586-024-07856-5.
- [38] Levin Brinkmann, Thomas F. Eisenmann, Anne-Marie Nussberger, Maxime Derex, Sara Bonati, Valerii Chirkov, and Iyad Rahwan. Experimental evidence for the propagation and preservation of machine discoveries in human populations. *Nature Communications*, 2026. in press.
- [39] Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021. doi: 10.1257/jel.20191450.
- [40] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022. doi: 10.1162/tacl_a_00511.
- [41] Katherine A. Keith, David Jensen, and Brendan O’Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, 2020. doi: 10.18653/v1/2020.acl-main.474.
- [42] Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109, 2021. doi: 10.18653/v1/2021.naacl-main.323.
- [43] Ishan Agrawal, Zhijing Jin, Ehsan Mokhtarian, Siyuan Guo, Yuen Chen, Mrinmaya Sachan, and Bernhard Schölkopf. Causalcite: A causal formulation of paper citations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8395–8410, 2024. doi: 10.18653/v1/2024.findings-acl.497.

- [44] Mingmeng Geng, Caixi Chen, Yanru Wu, Yao Wan, Pan Zhou, and Dongping Chen. The impact of large language models in academia: From writing to speaking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19303–19319, 2025. doi: 10.18653/v1/2025.findings-acl.987.
- [45] Richard A. Blythe and William Croft. S-curves and the mechanisms of propagation in language change. *Language*, 88(2):269–304, jun 2012.
- [46] Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986. doi: 10.1016/0010-0277(86)90010-7.
- [47] George Yule. *Referential Communication Tasks*. Lawrence Erlbaum Associates, 1997.
- [48] Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996. doi: 10.1037/0278-7393.22.6.1482.
- [49] Heather Bortfeld and Susan E. Brennan. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147, 1997. doi: 10.1080/01638537709544986.
- [50] Ellise Suffill, Timea Kutasi, Martin J. Pickering, and Holly P. Branigan. Lexical alignment is affected by addressee but not speaker nativeness. *Bilingualism: Language and Cognition*, 24(4):746–757, 2021. doi: 10.1017/S1366728921000092.
- [51] Robert XD Hawkins, Michael C Frank, and Noah D Goodman. Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, volume 39, 2017.
- [52] Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–226, 2004. doi: 10.1017/S0140525X04000056.
- [53] Peter Zeng, Weiling Li, Amie J. Paige, Zhengxiang Wang, Panagiotis Kaliosis, Dimitris Samaras, Gregory J. Zelinsky, Susan Brennan, and Owen Rambow. LVLMS and humans ground differently in referential communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9087, 2026. doi: 10.18653/v1/2026.acl-long.410.
- [54] Cameron R Jones, Agnese Lombardi, Kyle Mahowald, and Benjamin K Bergen. LLMs and people both learn to form conventions—just not with each other. *arXiv*, 2026. doi: 10.48550/arXiv.2602.08208.
- [55] Saujas Vaduguru, Yilun Hua, Yoav Artzi, and Daniel Fried. Success and cost elicit convention formation for efficient communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pages 42033–42050. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1946.
- [56] Cristine H. Legare and Mark Nielsen. Imitation and innovation: The dual engines of cultural learning. *Trends in Cognitive Sciences*, 19(11):688–699, nov 2015. doi: 10.1016/j.tics.2015.08.005.
- [57] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, mar 2019. doi: 10.1016/j.obhdp.2018.12.005.

- [58] Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. In *Proceedings of the 2025 ACM CHI Conference on Human Factors in Computing Systems*, number 1025, pages 1–15, apr 2025. doi: 10.1145/3706598.3714082.
- [59] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, feb 2015. doi: 10.1037/xge0000033.
- [60] Chiara Longoni, Andrea Bonezzi, and Carey K. Morewedge. Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650, dec 2019. doi: 10.1093/jcr/ucz013.
- [61] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, oct 2019. doi: 10.1177/0022243719851788.
- [62] Luke Rendell, Laurel Fogarty, William Hoppitt, Thomas J. H. Morgan, Mike Webster, and Kevin N. Laland. Cognitive culture: Theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences*, 15(2):68–76, 2011. doi: 10.1016/j.tics.2010.12.002.
- [63] Manvir Singh. Subjective selection and the evolution of complex culture. *Evolutionary Anthropology*, 31:266–280, 2022. doi: 10.1002/evan.21948. URL <https://onlinelibrary.wiley.com/doi/10.1002/evan.21948>.
- [64] Mingmeng Geng and Roberto Trotta. Is ChatGPT transforming academics’ writing style? *arXiv*, nov 2024. doi: 10.48550/arXiv.2404.08627.
- [65] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, jul 2024. doi: 10.1038/s41586-024-07566-y.
- [66] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *Proceedings of the 1st Conference on Language Modeling*, oct 2024.
- [67] Elihu Katz. The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, 21(1):61–78, 1957. doi: 10.1086/266687.
- [68] Mauro Mobilia. Does a single zealot affect an infinite group of voters? *Physical Review Letters*, 91(2):028701, jul 2003. doi: 10.1103/PhysRevLett.91.028701.
- [69] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, sep 2010. doi: 10.1126/science.1185231.
- [70] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, jun 2018. doi: 10.1126/science.aas8827.
- [71] Jacob Eisenstein, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114, nov 2014. doi: 10.1371/journal.pone.0113114. URL <https://doi.org/10.1371/journal.pone.0113114>.

- [72] Jean-Baptiste Michel, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. and The Google Books Team Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, jan 2011. doi: 10.1126/science.1199644.
- [73] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, jun 2016. doi: 10.18653/v1/P16-1141. URL <https://aclanthology.org/P16-1141/>.
- [74] Pierre Bourdieu. *Language and Symbolic Power*. Harvard University Press, 1991.
- [75] William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, 2006.
- [76] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, mar 2023. doi: 10.1126/science.adh2586.
- [77] Xiaofei Wang, Hayley M. Sanders, Yuchen Liu, Kennarey Seang, Bach Xuan Tran, Atanas G. Atanasov, Yue Qiu, Shenglan Tang, Josip Car, Ya Xing Wang, Tien Yin Wong, Yih-Chung Tham, and Kevin C. Chung. ChatGPT: Promise and challenges for deployment in low-and middle-income countries. *The Lancet Regional Health - Western Pacific*, 41:100905, dec 2023. doi: 10.1016/j.lanwpc.2023.100905.
- [78] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9:1645–1653, aug 2025. doi: 10.1038/s41562-025-02194-6.
- [79] Thomas H. Costello, Gordon Pennycook, and David G. Rand. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814–eadq1814, sep 2024. doi: 10.1126/science.adq1814.
- [80] Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association*, pages 3222–3226, oct 2023. doi: 10.21437/interspeech.2023-205.
- [81] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark, and recipe. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association*, pages 1983–1987, aug 2023. doi: 10.21437/interspeech.2023-105.
- [82] Eunah Cho, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. A corpus of spontaneous speech in lectures: The KIT lecture corpus for spoken language processing and translation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1554–1559, may 2014. doi: 10.63317/38ya5og5f63v.
- [83] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 28492–28518, jul 2023.

- [84] Shahrar Elisha, Andrew McDowell, Mariano Beguerisse-Díaz, and Emmanouil Benetos. Classification of spontaneous and scripted speech for multilingual audio. In *Proceedings of the 2024 IEEE Spoken Language Technology Workshop*, pages 489–495, 2024.
- [85] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association*, pages 4489–4493, aug 2023. doi: 10.21437/Interspeech.2023-78.
- [86] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the 2004 ACL Interactive Poster and Demonstration Sessions*, pages 214–217, jul 2004.
- [87] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, mar 1980. doi: 10.1108/eb046814.
- [88] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/cbo9780511809071.
- [89] Alberto Abadie and Jaume Vives-i Bastida. Synthetic controls in action. *arXiv*, mar 2022. doi: 10.48550/arXiv.2203.06279.
- [90] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv*, sep 2013. doi: 10.48550/arXiv.1301.3781.
- [91] Bruno Ferman and Cristine Pinto. Placebo tests for synthetic controls. Technical Report 78079, University Library of Munich, Germany, apr 2017. URL <https://mpra.ub.uni-muenchen.de/78079/>.
- [92] Otto Waris, Anna Soveri, Miikka Ahti, Russell C. Hoffing, Daniel Ventus, Susanne M. Jaeggi, Aaron R. Seitz, and Matti Laine. A latent factor analysis of working memory measures using large-scale data. *Frontiers in Psychology*, 8:1062, 2017. doi: 10.3389/fpsyg.2017.01062.
- [93] Research organization registry (ROR), 2024. URL <https://ror.org/>.
- [94] Douglas Biber and Susan Conrad. *Register, Genre, and Style*. Cambridge University Press, Cambridge, 2 edition, 2019. doi: 10.1017/9781108686136.
- [95] Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. irr: Various coefficients of interrater reliability and agreement, 2010.
- [96] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2):591–646, 2009. doi: 10.1103/revmodphys.81.591.

Supplementary Materials

Hiromu Yakura^{1*†}, Ezequiel Lopez-Lopez^{2,3†}, Levin Brinkmann^{1†}, Ignacio de la Serna¹, Lara Kirfel¹, Prateek Gupta¹, Ivan Soraperra¹, Thomas F. Eisenmann¹, Dirk U. Wulff^{2,4}, Iyad Rahwan¹

¹Center for Humans and Machines, Max-Planck Institute for Human Development, Germany

²Center for Adaptive Rationality, Max-Planck Institute for Human Development, Germany

³Center Synergy of Systems, TUD Dresden University of Technology, Germany

⁴Department of Business Analytics and Decision Science, Vienna University of Economics and Business, Austria

*Corresponding author(s): yakura,lopez,brinkmann,rahwan@mpib-berlin.mpg.de

[†]These authors contributed equally to this work.

This PDF file includes:

Materials and Methods

Figures **S1** to **S17**

Tables **S1** to **S7**

Materials and Methods

Datasets to compute Log Odds Ratios of human and LLM word usage

arXiv The arXiv dataset contains abstracts from research papers that were published on the arXiv website. We used the arXiv API⁵ to extract 150 papers from five different categories, namely Computer Science, Electrical Engineering and Systems Science, Mathematics, Physics, and Statistics, each month from 2019 to 2022. All categories were further divided into 133 subcategories, and we gathered 7182 abstracts in total.

bioRxiv The bioRxiv abstracts were gathered using the bioRxiv API.⁶ We ran a brute force query from the start to the end of the month for 4 years (2019-2022). We collected 60 papers per month, which was 2,880 papers in total.

Nature Nature abstracts were collected using the search engine on the Nature website. The process involved querying for up to 20 pages, each displaying 50 results. Our goal was to collect between 7,000 and 10,000 abstracts to match a comparable size of our other datasets. To achieve this, we ran a query without specific search terms, focusing on publications from 2019 to 2023. The results were sorted in two ways: ascending and descending per year. This dual sorting approach yielded over 8,000 unique abstracts, sufficient for our purposes.

Emails The email dataset was sourced from the publicly available Enron email dataset on Kaggle.⁷ The original dataset contains 500,000 emails generated by employees of the Enron Corporation. However, for our use case, we randomly sampled 10,000 emails and processed them into a dataset. The emails were sent between 2000 and 2001, far before the introduction of ChatGPT.

Essays This dataset comprises student essays collected from The Hewlett Foundation: Automated Essay Scoring challenge on Kaggle.⁸ The goal of this challenge was to develop an automated scoring algorithm for essays. All of the essays in the challenge, which was released in 2012, were composed by students. Similar to other datasets, we sampled 10,000 essays for our analysis.

Wikipedia We utilized the Wikipedia API⁹ to pull the articles from Wikipedia. One restriction of the API is that it only gives the article title and article ID. Due to the existence of duplicate articles, we were forced to query 30,000 of them. We extracted the articles' dates of publication after eliminating duplicates, keeping just those that were released between 2019 and 2022. We used the page title to randomly collect the content of 10,000 articles from this chosen subset.

Podcast To test whether GPT word preferences extend to spontaneous spoken language, we reused the pre-ChatGPT podcast transcripts assembled for the observational analysis (**Constructing datasets of human spoken communication**). From English-language episodes published before ChatGPT's release (30 November 2022) that had an available transcript and a spontaneity score above the threshold used in the main analysis (**Spontaneity annotation and classification**), we randomly sampled 2,000 episodes and extracted one 300-word excerpt from each.

⁵<https://info.arxiv.org/help/api/index.html>

⁶<https://api.biorxiv.org/>

⁷<https://www.kaggle.com/datasets/wcukierski/enron-email-dataset>

⁸<https://www.kaggle.com/competitions/asap-aes>

⁹https://www.mediawiki.org/wiki/API:Main_page

Spontaneity annotation and classification

A central premise of our podcast analysis is that the corpus captures *spontaneous* spoken language rather than read-out scripted material. Speaker diarization can distinguish monologues from multi-speaker exchanges, but a multi-speaker show can still be scripted (e.g., a co-hosted news read-through or an interview from a written question list), and a monologue can still be spontaneous (e.g., extemporized commentary). Following the register-theoretic distinction between “oral” and “literate” features of spoken language [94], and building on prior corpus work that operationalizes spontaneity via disfluencies [82], we therefore complemented diarization with an explicit annotation of how spontaneous each podcast sounds. Disfluencies, such as filler words, repetitions and corrections, hesitations, and locally incorrectly used words, are taken as positive evidence of spontaneous production, since they reflect cognitive effort in language production and interaction management that does not arise when reading a script.

Annotation protocol

Three coders were recruited from the research-assistant pool of the host institution and instructed via a written coding manual (reproduced in the appendix to the report on podcast annotation; available with the release materials). They were blind to the hypothesis of the study and to the purpose of the annotation task. The manual instructed them to attend only to disfluencies, and explicitly *not* to language ability, monologue versus dialogue, post-production quality, or genre. Each episode was rated on a four-point scale: 1 (*clearly scripted*), 2 (*rather scripted*), 3 (*rather spontaneous*), and 4 (*clearly spontaneous*). A threshold of five disfluencies in a 30-second sample was adopted as the operational cut-off between scripted and spontaneous codes; clips with around four disfluencies were treated as edge cases, with the manual specifying tie-breaking rules across the three samples drawn from each episode.

Annotation proceeded through a custom web interface that presented one episode at a time as three 30-second clips taken from the beginning, middle, and end of the middle ten minutes of the recording. This window avoids introductions, outros, and trailing music while still sampling the dynamics of the episode. Episodes were shown in random order, independently per coder, and coders could flag individual clips for technical issues (e.g., non-English passages, audio artifacts), via a dedicated web interface (Fig. S8). The annotation effort was split into a two-round calibration phase and a main labeling phase. After the first calibration sample of 80 podcasts the interrater reliability was moderate-to-good (intraclass correlation coefficient ICC = 0.79, 95% CI 0.69–0.86; two-way agreement model with average measures, computed via the `irr` package in R [95]), so the coders met to discuss low-agreement cases and the manual was refined (including the five-disfluency threshold). A second calibration sample of 80 podcasts yielded ICC = 0.85 (95% CI 0.76–0.91), at which point we moved to the main labeling sample. The main sample consisted of 400 independently coded podcasts split across three batches (80, 160, and 160), used as the ground-truth set for classifier training and evaluation. One episode was lost due to a technical issue, yielding a final main-sample size of 399 episodes; on this set, the interrater reliability was ICC = 0.87 (95% CI 0.81–0.91), indicating excellent agreement under any of the standard guidelines for ICC interpretation.

Audio-based spontaneity classifier

We used the 399-episode annotated set to train an audio classifier in the same manner as Elisha et al. [84] so that it predicts spontaneity directly from each podcast’s middle window, allowing the spontaneity criterion to be applied to the full corpus without further manual labeling.

Feature extraction Each episode was resampled to 16 kHz mono and split into 25 contiguous 30-second chunks taken from the middle of the recording, matching the slice used during human annotation. A log-mel spectrogram (128 mel bands $\times$ 3,000 frames) was computed per chunk and passed through the encoder of the `whisper-large-v3` speech model [83], yielding a 1500×1280 embedding tensor per chunk and a $25 \times 1500 \times 1280$ tensor per episode. The Whisper encoder was kept frozen throughout; only the downstream classification head was trained.

Classifier architecture We used a compact multilayer perceptron (MLP) operating on individual chunk embeddings. Each frame of the chunk embedding is projected via a linear layer ($1280 \rightarrow 8$) followed by ReLU, expanded back via a second linear layer ($8 \rightarrow 128$) followed by ReLU, average-pooled across the 1,500 time positions, regularized with dropout, and mapped via a final linear layer to four logits corresponding to the four annotation classes (clearly spontaneous, rather spontaneous, rather scripted, clearly scripted).

Training targets and procedure The three coders’ ratings were converted into a per-episode soft label by placing each rater’s vote on the four-class simplex and averaging, yielding a discrete probability distribution over the four classes per episode. This soft target preserves the partial disagreement information that a hard majority vote discards (e.g., a 4/3/4 vote and a 4/4/4 vote get different targets). Soft labels were broadcast from the episode level to every one of its 25 chunks for chunk-level training. The MLP was trained with a soft-label cross-entropy objective on the 399-episode annotated set; we used a held-out split of the same set for early stopping and threshold selection, and report classification metrics (precision, recall, F1, balanced accuracy) along with the confusion matrix from the held-out split with the release materials.

Inference and episode-level score At inference time, the trained MLP emits a four-class softmax per chunk, and we aggregated the 25 chunk distributions into a single episode-level score in two complementary ways. The *binary vote* score uses each chunk’s argmax: an episode’s score is the fraction of chunks whose argmax falls in the two spontaneous classes (so the score lies in $[0, 1]$). The *weighted* score maps each chunk to the expected ordinal value $\sum_{k=1}^4 k \cdot p_k$ under the chunk’s predicted distribution and averages across the 25 chunks (so the score lies in $[1, 4]$, with higher values indicating more spontaneous production). The weighted variant is the default used downstream; as below, the binary-vote variant was confirmed not to materially change which episodes are retained.

Filtering policy For all main analyses, an episode is retained for downstream word-frequency aggregation if (i) it passes the diarization-based dialogue filter (at least two distinct speakers and four or more alternating turns within the middle 10-minute window) and (ii) its weighted spontaneity score exceeds 3.0. The threshold of 3.0 corresponds to the inflection point of the precision/recall curve on the held-out annotated split (approximately 90% accuracy at the threshold) and roughly partitions episodes between the “rather spontaneous” and “rather scripted” labels. The binary-vote analogue (≥ 18 of 25 chunks classified spontaneous, equivalently a score of 0.72) yields a near-identical retained episode set, confirming that the choice between the two aggregation rules does not materially affect the downstream corpus.

Word-sense audit

The raw podcast counts tally every token whose stem matches a target word, irrespective of meaning. This over-states the signal of interest in two ways. The first is contamination by

proper nouns, brands, and transcription artifacts that happen to share a stem. A count for **swift**, for instance, sweeps in Taylor Swift, Apple’s Swift programming language, the SWIFT banking network, and NASA’s Swift telescope, none of which is the adjective the word list is meant to track. The second is the presence of dictionary senses that one corpus essentially never produces yet that still inflate the raw match. To recover a count that reflects the intended sense, we audited each target word, retaining only occurrences that fall within a curated set of legitimate senses. This is a *sense-validation* step and it is symmetric in human and machine usage. It removes contamination and unused senses, and the rule that selects which senses to keep never inspects the human-versus-GPT asymmetry, so it neither identifies nor conditions on the overuse effect the analysis later estimates.

The audit begins from the word-level log-odds ratios that rank words by their preference in GPT-rephrased over human text, and retains the most GPT-preferred words after two filters that protect signal quality and bound cost. We dropped stems present in more than 20% of podcast episodes, which are too ubiquitous to carry a distinguishable shift, and stems above the 99th percentile of total podcast occurrence, whose per-occurrence auditing would be prohibitively expensive. For each surviving word we retrieved its inventory of dictionary senses from Wiktionary,¹⁰ following inflected-form redirects so that a form such as *underscored* resolves to the entry for *underscore*. We then extracted paired example sentences from both sides of the comparison, drawing on the order of one hundred human and three hundred GPT usages per word, the larger machine sample reflecting the three rephrasings generated per source passage.

Sense assignment and occurrence filtering were carried out by an open-weight instruction-tuned language model, Qwen3-30B-A3B-Instruct-2507, served on a compute cluster.¹¹ In a first pass the model classified each example sentence into one of the word’s Wiktionary senses, or into a **contamination** label reserved for proper nouns, transcription artifacts, and uses absent from the dictionary. Aggregating these assignments across a word’s examples gives a per-sense distribution for each corpus, from which we selected the in-scope senses by an intersection rule at a fixed noise floor. A sense is kept only if both corpora attest it at least twice, and the **contamination** and parse-failure categories are always excluded. Requiring attestation in both corpora removes senses that one side essentially never produces—transcription artifacts and ghost-only senses—while the permissive two-occurrence floor accommodates the modest per-word example counts, so that genuine senses appearing only a handful of times are still retained. The retained senses for each word define its in-scope set. Every podcast occurrence of the word was then presented to the model under a binary prompt asking whether that occurrence belongs to the in-scope set, and occurrences judged out of scope were removed from the count.

Aggregating the per-word results yields the audited count matrix, which preserves the same 931,450-episode rows as the raw matrix and replaces each target word’s column with its audited count. The audit typically removes a substantial share of the raw matches for contaminated words, as the brand, banking, and telescope senses of a word such as **swift** are filtered out.

These audited counts are the substrate for the synthetic-control estimation, the change-point model, and the figures, which are accordingly reported as the “audited” variants. The robustness analyses (**audited_main** against **c1–c4**) vary only the synthetic-control specification on this shared audited substrate, with a single control run repeated on the raw un-audited counts to isolate the contribution of the audit itself.

¹⁰English Wiktionary: <https://en.wiktionary.org>

¹¹Qwen3 instruction-tuned model: <https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507>

Trends not linked to ChatGPT

Applying the same change-point machinery without the top-1% GPT-score filter recovers the twenty most credibly shifted words across the full Science & Technology vocabulary (Supplementary Fig. S10). This is a deliberately word-preference-blind view: candidates are ranked only by how confidently their post-release slope departs from zero, so any real change in how often a word is spoken can surface, whatever its cause.

The largest declines trace a single well-understood event. The four steepest falls are *pandemic*, *vaccine*, *corona* and *virus*, the core vocabulary of the COVID-19 period. These words peaked across 2020 and 2021 and were already receding when ChatGPT was released in late 2022, so the model reads their continued decline as a strong negative post-release slope. The shift is genuine, but its driver is the waning of a global news cycle rather than any property of language models. Neighbouring falls tell the same story of topical churn: *twitter* declines as the platform rebrands and its salience drops, and case-study or subject-specific terms such as *scholars*, *reynolds* and *urine* reflect the rotation of particular talks in and out of the corpus.

The rises are more mixed, and this is exactly the point of contrast with the main analysis. Some are the GPT signatures our study targets, led by *delve* (the highest GPT score in the corpus), *invaluable* and *prompted*. Others are ordinary technical vocabulary riding the same 2022–2024 wave of interest in machine learning and tooling: *gpu*, *python*, *vector*, *queries* and *transcript* all climb, yet their GPT scores sit near the middle of the distribution. A word can therefore change credibly for reasons that have nothing to do with LLM revision preferences, from shifting subject matter to the growth of an entire field.

This is why the main study conditions on GPT preference rather than on magnitude of change alone. The change-point detector answers the question “which words moved?”; it cannot by itself distinguish an LLM fingerprint from a pandemic, a product launch or a hot research topic. Our design isolates the LLM channel by first restricting to words that GPT systematically over- or under-uses when revising human text, then asking whether those specific words shifted after the release. Fig. S10 shows what the unconditioned version looks like, and makes clear that the pandemic and technology-hype trends visible here are a separate phenomenon from the word-preference effect the paper measures.

Causal Identification

GP smoothing of the input series

We apply a Gaussian-process smoother (Matérn $\nu = 2.5$, length scale $\ell = 720$ d, noise $\sigma = 0.05$) to the monthly input series before fitting the synthetic control. The GP is fit on the pre-treatment months only; post-treatment months are passed through verbatim from the raw monthly series, so that the smoother cannot leak post-release dynamics into either the donor-selection step or the SLSQP fit. The length scale is chosen as a compromise between the raw monthly series, on which the synthetic overfits noise, and longer length scales (e.g. $\ell = 1440$ d) at which it underfits the relevant low-frequency dynamics (Supplementary Fig. S9). The smoother enters only at donor selection and the synthetic control fit itself; all downstream change-point and placebo statistics are computed on the raw monthly series.

In-time placebo: Change-point date sweep

To assess whether the post-release acceleration is anchored in time to ChatGPT’s launch rather than to an arbitrary cut in the data window, we sweep the assumed change point across a

monthly grid spanning the full observation window. At each candidate month c , we form a 24-month pre-window $[c - 24\text{mo}, c)$ and an 18-month post-window $[c, c + 18\text{mo}]$, refit the per-word change-point regression of Eq. (1) on the synthetic-control gap, and record the cross-word mean of the post-release slope β_{Post} over the top-1% GPT-score slice. To make the sweep tractable across hundreds of candidate dates, β_{Post} at each candidate is estimated with an ordinary-least-squares proxy of the change-point regression; we validated the proxy against the Stan estimator at the true change point and obtained a Pearson correlation of 0.98 across the top-1% slice.

We define the null over candidate change points whose entire 18-month post-window strictly pre-dates ChatGPT’s release, yielding $n = 28$ pre-GPT candidates; the reported permutation $p = 0.034$ follows the same rank convention as the in-space placebo and is bounded below by $1/29$.

Synthetic-control pre-trends check

The change-point regression (1) estimates a pre-treatment slope β_{Pre} on each word’s treated–synthetic-control gap. The synthetic control is fit by minimizing pre-treatment RMSPE, which constrains the level of the gap but not its slope, so β_{Pre} is not zero by construction and can serve as a pre-trends test. Across the Science & Technology panel ($n = 3,535$ words), pre-slopes are tightly distributed around zero: median $|\beta_{\text{Pre}}| = 0.001$ versus median $|\beta_{\text{Post}}| = 0.015$.

The synthetic control also reproduces each treated word’s pre-release trajectory closely. Across the top-1% GPT-score words, the pre-treatment RMSPE is small on both the smoothed time series and the raw monthly data (median 0.008 and 0.038 in $\log_{10}$ -frequency units across all 3,535 Science & Technology words; Table S2).

Conservative β_{Post} bound. Per-word rankings and the “credibly positive” / “credibly negative” labels in Fig. 3A use a *conservative* summary of β_{Post} defined as the 95% HDI limit nearest zero: the lower HDI bound when the posterior is credibly positive (HDI excludes zero from below), the upper HDI bound when credibly negative, and zero otherwise. The conservative bound is therefore a lower bound (in magnitude) on the slope change attributable to the release, and is the quantity by which we order the top-1% panel in Fig. 3A.

Robustness checks

Detailed specifications for the four control variants introduced in the Methods; Table S5 gives the parameter view.

C1 — bare ℓ_2 matching. C1 replaces Main’s $w2v$ -then- ℓ_2 donor pool with a strict ℓ_2 -nearest pool drawn from the full vocabulary, thereby removing both the semantic-neighbor exclusion and the GPT-score neutral band — the two SUTVA-protecting filters that pre-screen the donor pool against interference. C1 asks whether the headline survives even when the donor pool is not pre-screened.

C2 — SLSQP on a tight ten-donor pool. C2 keeps Main’s donor strategy and SLSQP weight fit but shrinks the pool from $L = 100$ to $L = 10$. The smaller pool restricts the fit to the ten closest pre-treatment matches; it asks whether the SLSQP fit on a hundred-element basis exploits the flexibility the data does not warrant. Because the in-space placebo procedure draws targets from the same donor pool, C2’s ten-donor pool floors the empirical placebo p -value at $1/(10 + 1) \approx 0.091$.

C3 — inverse-distance similarity weights. C3 keeps Main’s donor pool ($L = 100$) but replaces the SLSQP convex fit with deterministic inverse-distance similarity weights $w_i = (1/(d_i + \varepsilon)) / \sum_k (1/(d_k + \varepsilon))$, where d_i is the Euclidean distance between the smoothed pre-treatment trajectory of donor i and that of the treated word and $\varepsilon = 10^{-12}$ is a numerical guard against $d_i = 0$. The form follows CausalCite [43]; it asks whether the headline survives a deterministic, optimization-free aggregation of the same donor pool.

C4 — raw (un-audited) counts. C4 keeps the Main synthetic-control specification but replaces the audited counts substrate with the un-audited counts, retaining the ambiguity across different word meanings that the audited pipeline controls for upstream; it asks whether the upstream audit step drives the result.

The Main, C1, C2, C3, and C4 variants are rendered side-by-side for the *delve* synthetic-control panel Fig. S1 and for the score–effect relationship Fig. S3.

Robustness to channel and episode outliers

This channel-robustness analysis is based on 930,812 conversational episodes from 38,294 distinct shows (channels). Episodes per channel are right-skewed (median 5), but no single channel dominates. The largest contributes 0.46% of episodes and the ten largest 1.4% (Fig. S6A).

Word-level usage is more concentrated, as expected, but robust to outliers. Within Science & Technology, excluding the ten channels that use *delve* most leaves the post-release increase essentially intact: over the post-adoption window (months 13–18 after release) the token-normalized rate of *delve* rises by a comparable factor with and without those channels (a factor of 1.5 in both cases; Fig. S6B). The shift reflects broad adoption across many channels rather than a few high-usage outliers.

The score–effect relationship is likewise not an artifact of a single word: excluding *delve*, the Gaussian-process fit of per-word effect against GPT-score rank is essentially unchanged, and the top- $X\%$ slice-mean β_{Post} remains above the permutation null at the high-GPT-score tail (top 1% and top 2%: permutation $p = 0.001$; Fig. S7).

Experimental implementation

Participant demographics

Table S6 reports the full demographic breakdown of the final sample ($N = 496$). Participants were recruited via Prolific between April 16 and April 20, 2026. The sample was 50% women, 48% men, and 2% non-binary or other; mean age was 40.5 years ($SD = 13.2$, range 18–65+; age collected in brackets). 91% reported English as their first language. Prior AI chatbot use: 96% reported having used an AI chatbot; the most commonly reported systems were ChatGPT, Google Gemini, and Microsoft Copilot. All participants provided informed consent and were compensated at a rate of £11.4 per hour.

Power analysis

We determined the sample size with an a priori power analysis. Based on pilot data and prior literature, we targeted a small-to-medium effect of $d = 0.30$. With two-tailed $\alpha = .05$ and power = .80, the required N per cell was 88, for a total of $N = 352$ across the 4 between-subjects cells. We recruited above the required ($N = 500$), and after exclusions, the final sample was $N = 496$.

Synonym pair selection

Synonym pairs were selected to satisfy two criteria: both forms are familiar, commonly used English words with no strong register or formality differences between them, and the concept can be unambiguously depicted in a cartoon illustration, enabling the referential image-guessing paradigm. The 18 selected pairs span three lexical categories (nouns, verbs, adjectives) and are divided into two groups of nine pairs each. Table S7 lists all pairs and the accepted surface forms used for scoring.

System implementation

The experiment was delivered as a custom web application (TypeScript/React 18 frontend; Python/FastAPI backend) hosted on a cloud server.

Stimulus generation. Each trial image was created as a matched treatment-control pair via a two-stage generative pipeline. First, a language model (gpt-5.2) was prompted to generate a cartoon-style illustration description in which the three target synonym variants (one noun, one verb, one adjective) were semantically necessary to describe the depicted scene (treatment prompt), together with a control prompt obtained by sentence-level substitutions that removed this requirement while preserving the overall scene composition. Second, an image generation model (gemini-3-pro-image-preview) rendered the treatment image from the treatment prompt, then received the treatment image together with the substitution instructions to produce the visually matched control variant. The image generation model was instructed to never display the target words as text within any image. The generated candidates were rated by the authors in terms of visual quality and consistency using a custom web-based annotation tool; only images receiving the highest rating were retained as final stimuli.

AI chatbot interaction. Chatbot responses were generated by calling the OpenAI API (gpt-4o); each call included the full within-trial chat history, the target image (no control image was provided), and additional context (see Fig. S19 and Fig. S20 for details). The system prompt was designed to ensure that the model consistently used the designated canonical synonym variant throughout the interaction. It comprised three layers: (i) a base game prompt establishing the image-guessing task and the model's role as an honest co-player; (ii) a per-trial vocabulary rule specifying the canonical form for each target word in the current trial, paired with treatment and control scene descriptions that anchored each term to specific image content, so the model would apply the rule even when a participant's question referenced the scene without using the target word directly; and (iii) a global vocabulary rule listing all 18 synonym pairs as remapping instructions, maintaining consistent usage across the entire interaction regardless of which image was shown or which words participants used. Two intervention groups were counterbalanced: chatbot in Group A used Variant 1 as the canonical, and Group B used Variant 2.

Speech transcription. Spoken descriptions were captured via the browser and transcribed by the standard Web Speech API. Upon stopping, the audio and the automatic transcript were uploaded to cloud storage; recordings shorter than 3 seconds or transcripts shorter than 5 words were rejected with a prompt to re-record. For the Test Phase, the automatic transcript was displayed to the participant in an editable text area, together with audio playback. Participants were asked to correct any recognition errors and submit, ensuring that the primary outcome measure was scored from human-verified transcripts.

Scoring and statistical analysis. For each synonym pair, we pre-registered a coding scheme listing the accepted surface forms of each variant, including morphological inflections (e.g., *fix*, *fixes*, *fixed*, *fixing*), compound and multi-word constructions (e.g., *to put up*), and spelling variants (e.g., *multicolored*, *multicoloured*, *multi-colored*, *multi-coloured*). Detection proceeded by tokenizing the transcript and each accepted phrase on alphabetic characters and stemming both with the NLTK `SnowballStemmer` (English), with a phrase matching whenever its stemmed token sequence appeared as a contiguous subsequence of the stemmed transcript; this yielded a binary trial-level indicator of whether the AI-primed variant was used. From these indicators, the primary outcome measure for each phase Δp was calculated as the per-participant mean difference in usage rate between the AI-primed and alternative variants across the nine trials. Usage in the Interaction Phase and Test Phase was each assessed with a participant-level Monte Carlo permutation test (10,000 iterations); forced-choice selections were assessed with a sign-flip test against the 50% chance baseline. Confidence intervals were obtained by cluster-bootstrap resampling participants (10,000 iterations); all tests were two-tailed at $\alpha = .05$.

Participant flow details

After completing a comprehension check and microphone test, participants proceeded through the Interaction Phase (12 image-guessing + spoken description trials), a 3-minute distractor task, the Test Phase (3 spoken description trials, no chatbot), an open-ended deception check (“Have you noticed anything about the language of the chatbot?”), the Forced-Choice Phase (9 two-alternative label selections), and a final questionnaire. The detection check question was placed between the Test Phase and the Forced-Choice Phase so that responses reflected impressions formed during the interaction without being coloured by the explicit synonym-choice framing of the Forced-Choice task. Open-ended responses were reviewed qualitatively; responses were coded as “aware” if they named one or more specific synonym pairs from the experiment or explicitly noted that the chatbot substituted the participant’s word with a different one. The final questionnaire collected demographics (age range, gender, education), language background (English as L1, age of acquisition), prior AI-chatbot use (systems used, frequency, purposes), and self-reported color vision deficiency (used as the basis for the post-hoc exclusion reported in Methods).

Theoretical model

We model linguistic preference diffusion using a noisy voter model [96] on a Watts–Strogatz small-world network ($N = 2000$ speakers, degree $k = 6$, rewiring probability $\beta = 0.1$). Each agent holds a binary state representing a preferred synonym variant, initialized i.i.d. from Bernoulli($p_0 = 0.1$). Each generation consists of N asynchronous update steps. At each step a randomly selected agent copies the state of a randomly chosen neighbor; with probability $\varepsilon = 0.005$ it instead resets to a draw from Bernoulli($p_0 = 0.1$), preventing absorbing consensus and producing a stationary distribution centered on p_0 .

A single hub, modeled as a committed source (zealot) fixed in state 1, is added as an extra neighbor to a uniformly random fraction $f \in \{0.5\%, 2\%, 5\%, 10\%\}$ of the population. As a control, a single committed speaker (zealot) occupies one network node with its natural 6 connections. After 100 generations without the zealot, the simulation continues for 300 generations with the zealot active. Results are averaged over 50 independent runs, each with a fresh network and initial state.

Supplementary Figures

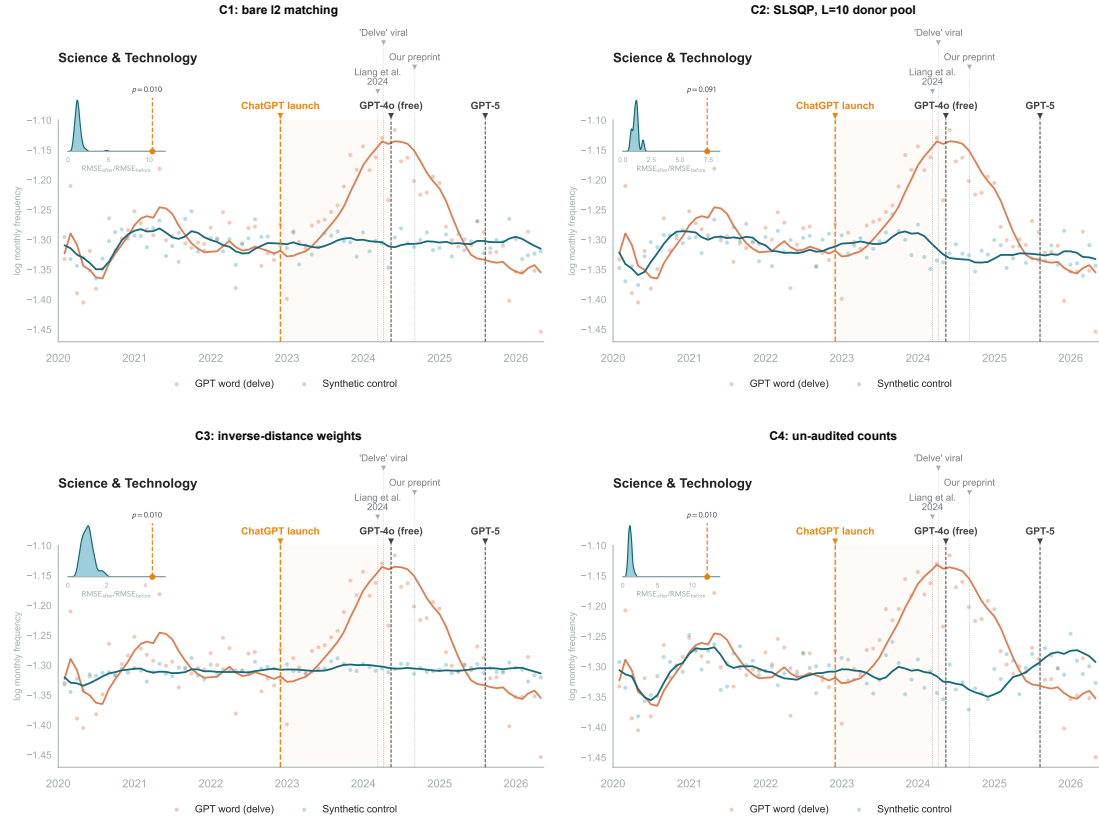


Figure S1: The post-release gap between *delve* and its synthetic control is preserved across all four robustness check specifications. To test whether the result depends on specific donor-selection and aggregation choices, we re-ran the synthetic-control pipeline under four control specifications (see Methods and Supplementary Table S5). In-space placebo p -values: $p_{C1} = 0.050$, $p_{C2} = 0.091$ (floor), $p_{C3} = 0.040$, $p_{C4} = 0.010$ (compare $p_{Main} = 0.010$). The main result is robust to changes in donor selection (C1, C2), weight aggregation (C3), and the counts substrate (C4).

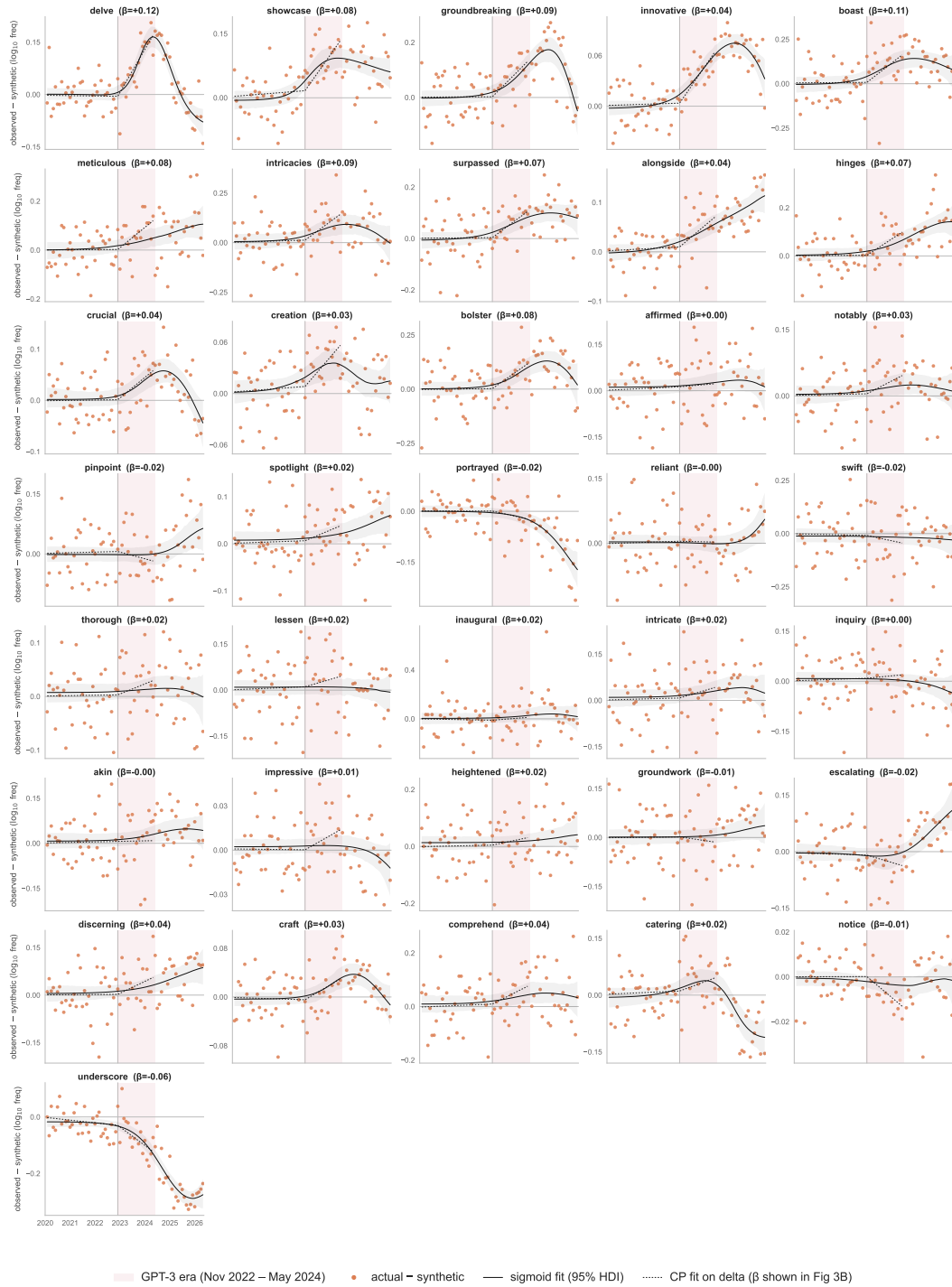


Figure S2: **Full top-1% GPT-score panel grid.** Monthly observed – synthetic $\log_{10}$ frequency (orange points) for every word in the top 1% of GPT scores ($n = 36$), with the double-sigmoid posterior smoother (solid; 95% HDI shaded) and the change-point fit of Equation 1 (dashed) overlaid. Of the 36 words, 28 show a positive post-release slope and 13 are credibly so (95% HDI on β_{Post} excludes zero). The twelve panels reproduced in Fig. 3A are the largest-magnitude credible subset.

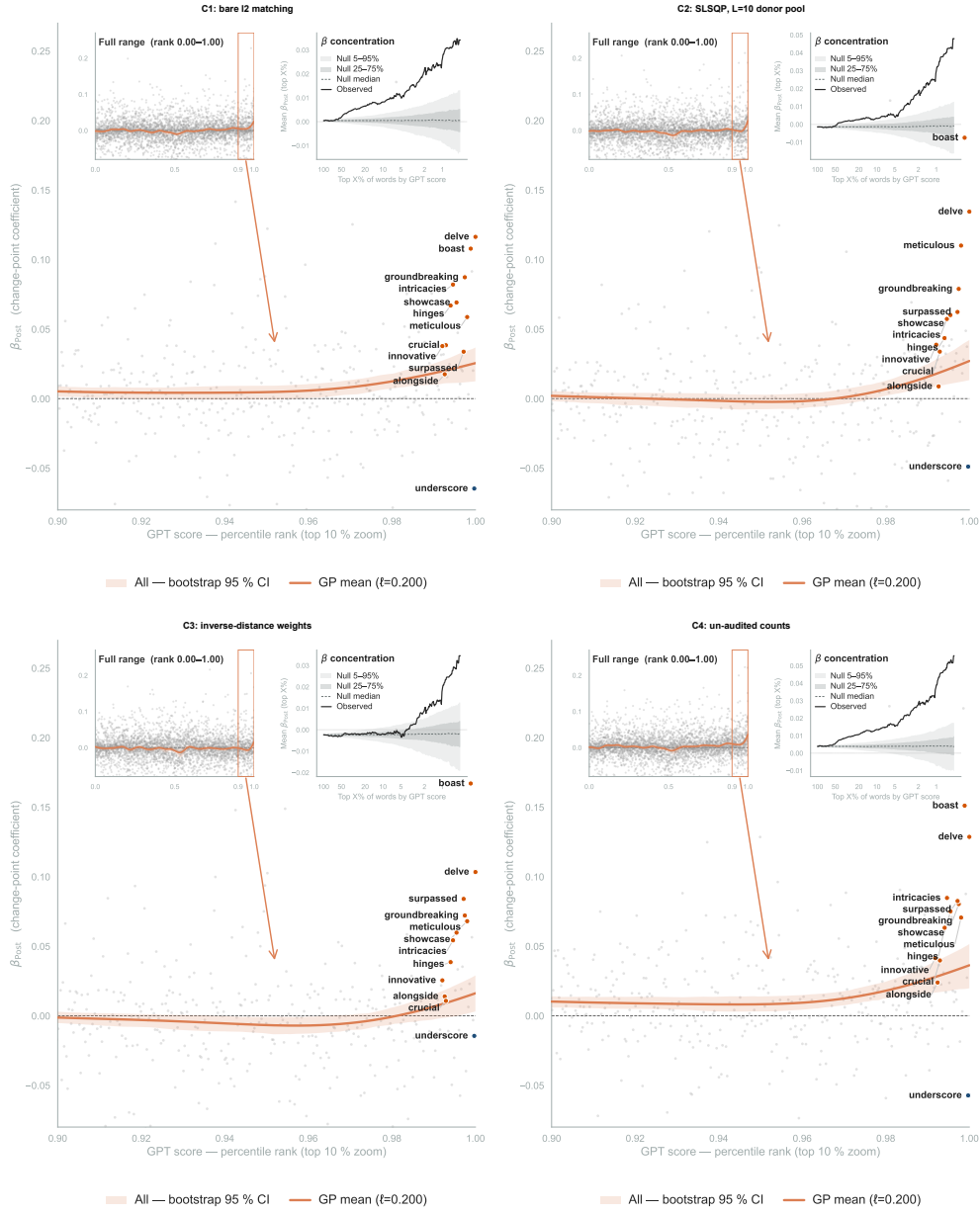


Figure S3: The score-graded acceleration in the high-GPT-score tail is preserved across all four robustness controls. To test whether the score-effect relationship depends on specific donor-selection and aggregation choices, we re-ran the synthetic-control pipeline under four control specifications (see Methods and Supplementary Table S5). Each panel reproduces Fig. 3B for one control: per-word β_{Post} against GPT-score percentile rank with the Gaussian-process posterior mean overlaid (bootstrap 95% CI shaded); the left inset shows the full rank-axis range, and the right inset the slice-mean β_{Post} over nested top- $X\%$ slices against the permutation null. The twelve labeled words are Main's top 12 by |95%-HDI bound on β_{Post} |, plotted at each control's own positions so the same exemplars can be tracked across specifications. The slice-mean exits the permutation null at the high-GPT-score tail under every control.

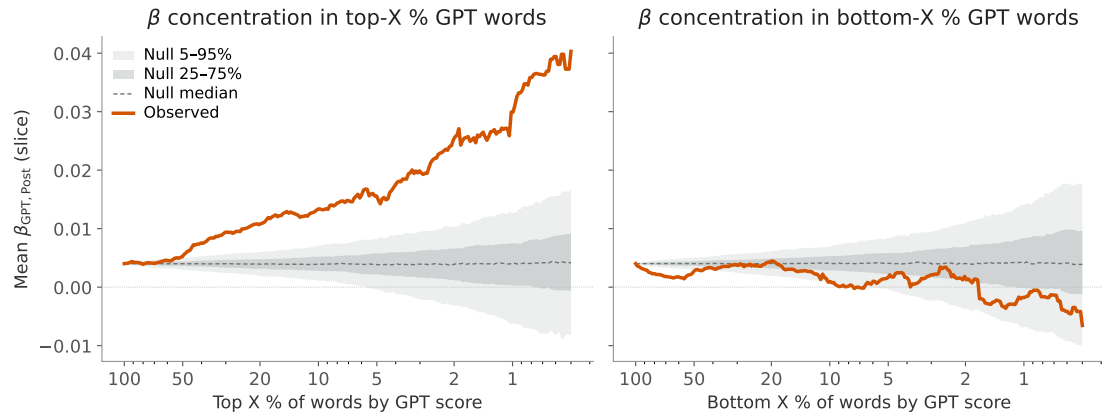


Figure S4: **Words favored by ChatGPT rise in spoken use, but words it disfavors do not show a matching decline.** Each panel sweeps a cut-off X across the top (**left**) or bottom (**right**) of the GPT-score distribution. The orange line shows the average post-release change in usage (the change-point slope β_{Post}) for the $X\%$ of words at that end of the ranking; the grey band is the 5th–95th percentile range expected by chance, obtained by reshuffling GPT scores across words. **Left:** as the cut-off tightens onto the words ChatGPT most strongly prefers, the orange line rises clearly above chance — at the top 1% of words the mean (+0.030) is more than $2\times$ the upper bound of chance (+0.013). **Right:** the matching check on the words ChatGPT most strongly disfavors stays close to zero at every cut-off (at the bottom 1%: -0.002 , chance range $[-0.007, +0.014]$).

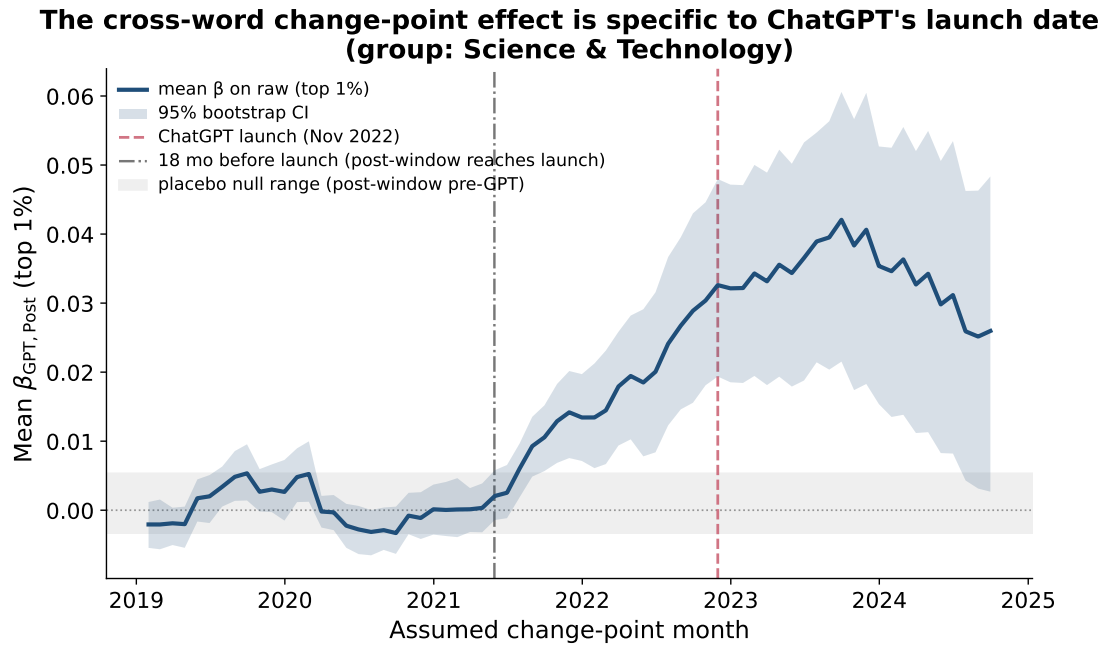


Figure S5: **The cross-word change-point effect is specific to ChatGPT's launch date.** Mean post-release slope β_{Post} across the top 1% of GPT-score words (Science & Technology, $n = 36$) as the assumed change-point date slides month-by-month over the data window, with a 24-month baseline and an 18-month post window held constant. Grey band: placebo-null envelope of mean β across candidate dates whose post-window is entirely pre-GPT. Indigo: 95% bootstrap CI of the mean over words. Dashed red: true launch (2022-11-30). Mean β stays inside the null at every pre-GPT candidate and steps up only once the window includes the launch. Permutation $p = 0.034$.

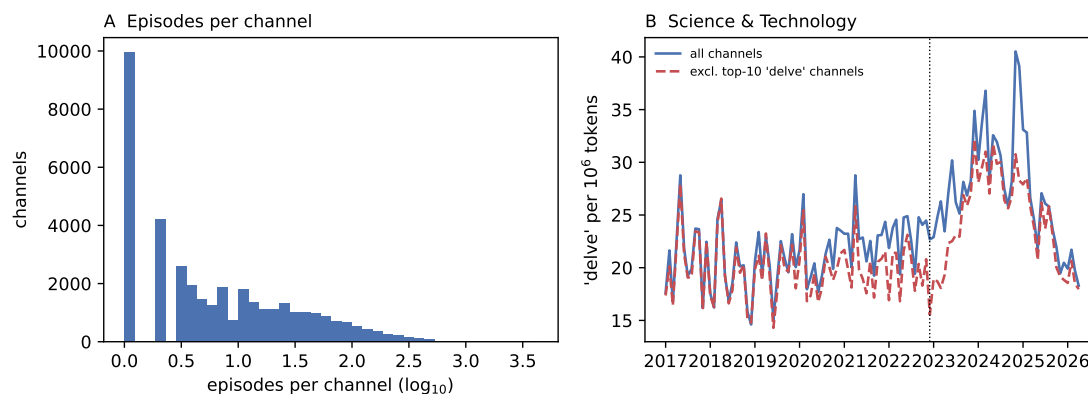


Figure S6: **Channel and episode outliers do not drive the podcast trends.** (A) Number of episodes per channel, across 38,294 podcast channels (log scale); the median channel contributes 5 episodes and the largest 0.46% of all episodes. (B) Token-normalized monthly rate of *delve* in Science & Technology for all channels (solid) and after removing the ten channels with the most *delve* usage (dashed); the dotted line marks the ChatGPT release. The post-release increase is essentially unchanged with and without those channels (fold ≈ 1.5 in both cases), so the trend is not driven by a few high-usage channels.

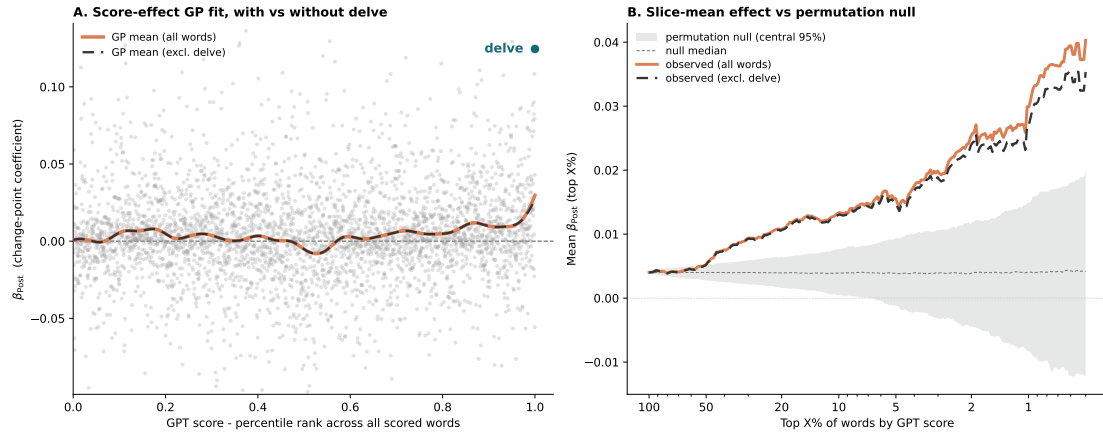


Figure S7: **The score-graded post-release acceleration is not an artifact of *delve*.** Both panels are computed on the audited Science & Technology substrate ($n = 3,535$ words). **(A)** Per-word change-point coefficient β_{Post} against GPT-score percentile rank (grey scatter, all 3,535 words). The Gaussian-process posterior mean of Fig. 3B is overlaid twice: over all words (solid) and with *delve* removed (dashed). The two curves are visually indistinguishable, and both turn upward only in the high-GPT-score tail; *delve* is marked at rank 1.0. **(B)** Slice-mean β_{Post} over nested top- $X\%$ GPT-score slices against a permutation null obtained by shuffling β_{Post} across words while holding the GPT-score order fixed (1,000 permutations; central-95% band and median shaded). The observed slice-mean is drawn for all words (solid) and with *delve* excluded (dashed); both rise far above the null at the high-GPT-score tail (top 1%: mean $\beta_{\text{Post}} = 0.030$ over 36 words vs. 0.027 excluding *delve*, permutation $p = 0.001$ for both; top 2%: 0.025 over 71 words vs. 0.024, $p = 0.001$ for both). The score-effect relationship therefore exceeds chance with and without *delve*.

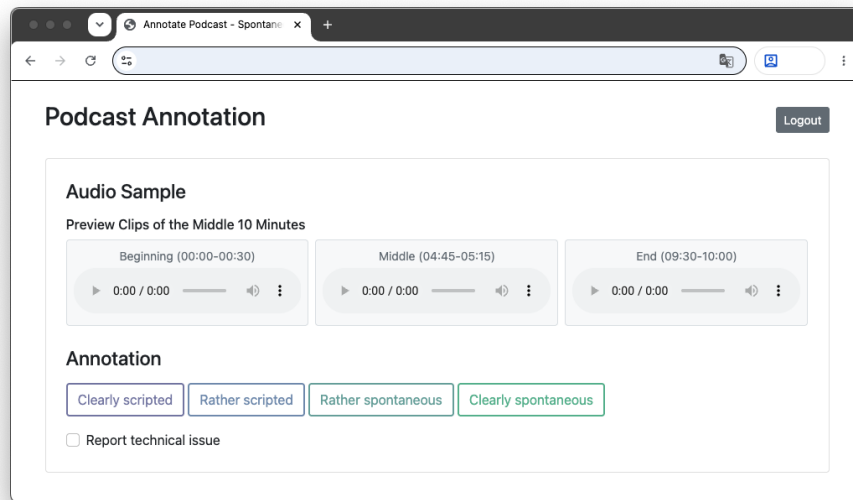


Figure S8: **Web-based interface used for the spontaneity annotation task.** Each coder heard three 30-second clips sampled from the beginning, middle, and end of the episode's middle 10-minute window (top row) and rated the episode on a four-point scale from *clearly scripted* to *clearly spontaneous* (bottom row). Coders were blind to the study hypothesis and to the purpose of the annotation task.

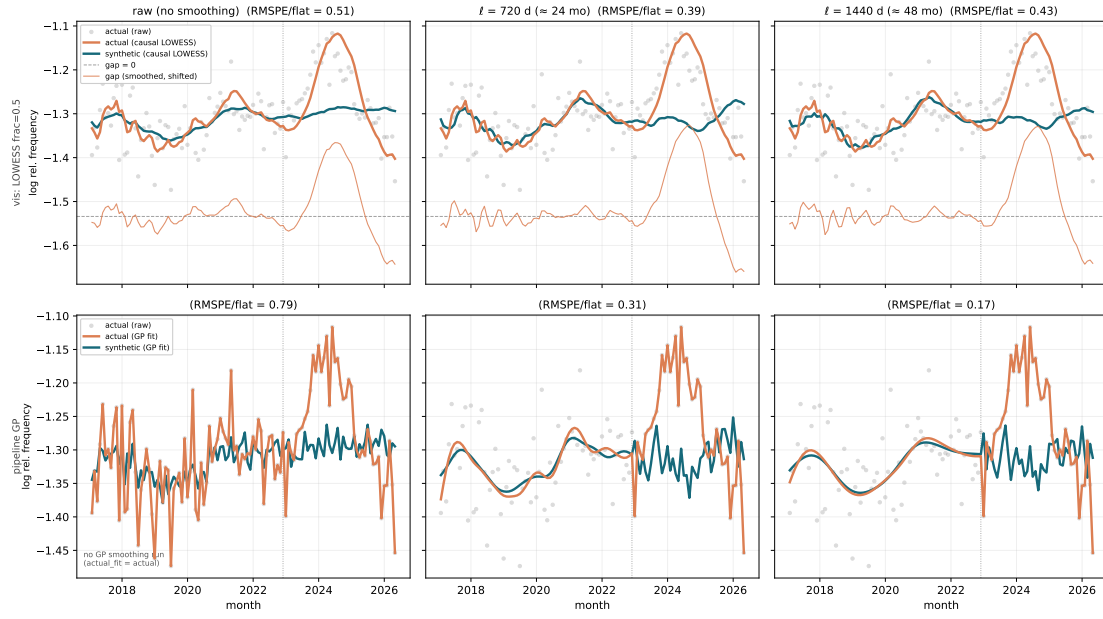


Figure S9: Choice of GP smoother length scale for synthetic-control donor selection. 2×3 grid showing actual – synthetic across three input-series GP smoothing length scales (columns: $\ell = \text{raw}$, 720 d, 1440 d) under two visualisation smoothers (top: causal LOWESS as in Fig 1A; bottom: pipeline GP fit). Without smoothing (left), the synthetic overfits monthly noise; at $\ell = 1440 \text{ d}$ (right), it underfits the low-frequency dynamics. $\ell = 720 \text{ d}$ (middle, the pipeline default) is chosen as a compromise. Smoothing enters only at donor selection and the synthetic control fit; all downstream change-point and placebo statistics are computed on the raw monthly series.

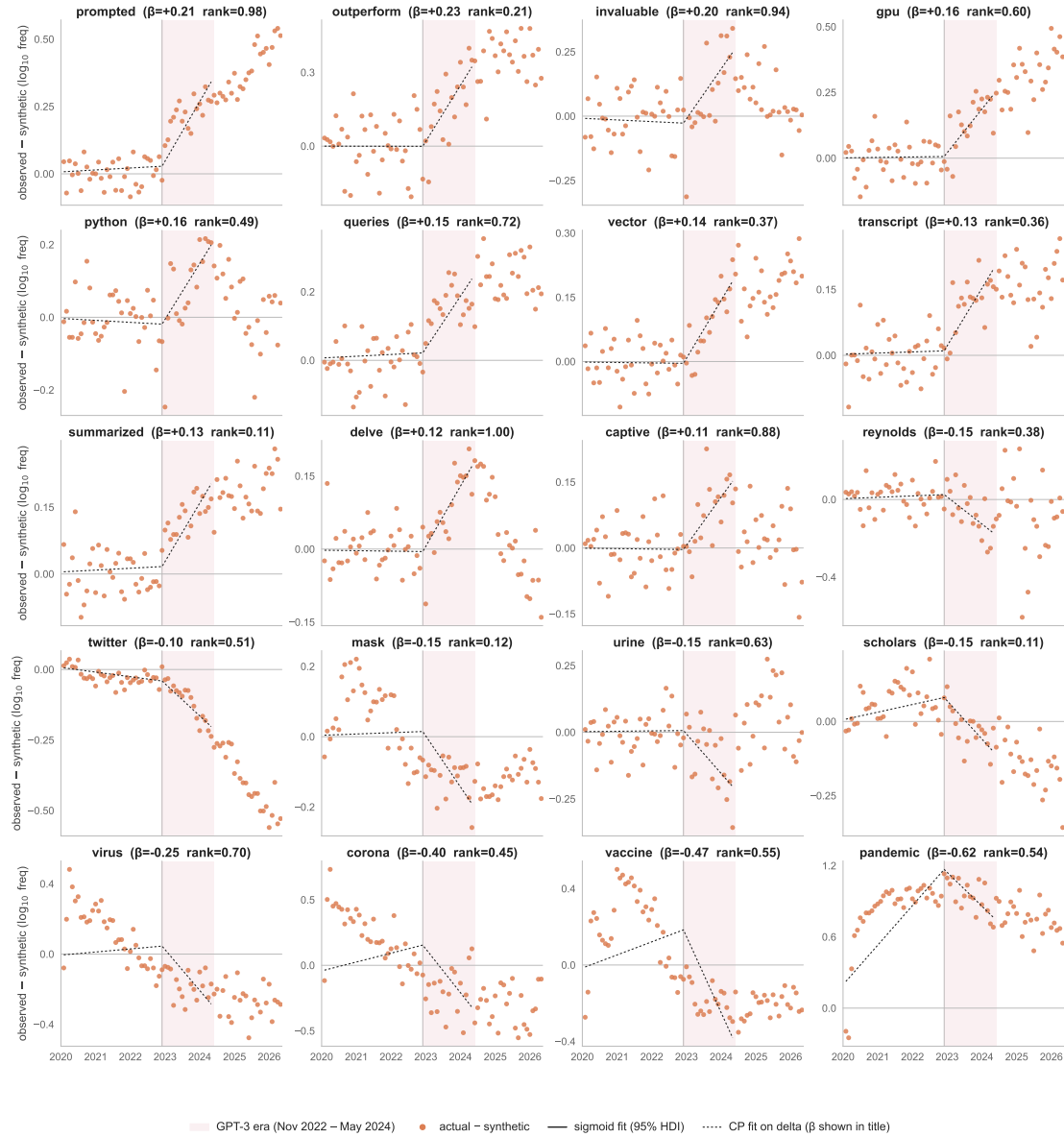


Figure S10: **The 20 words with the most credible post-release shift (Science & Technology podcasts, $n = 3,535$ stems) without reference to GPT preference.** Each panel: monthly observed – synthetic log₁₀ frequency (orange scatter), double-sigmoid posterior smoother (solid, 95% HDI shaded), and the change-point fit of Equation 1 (dashed). Panels are ordered by the signed conservative bound on β_{Post} (the 95% HDI limit nearest zero). Per-panel GPT-score percentile rank is annotated in the title. The shaded vertical band marks the period of analysis between ChatGPT's launch and the launch of GPT-4o (free).

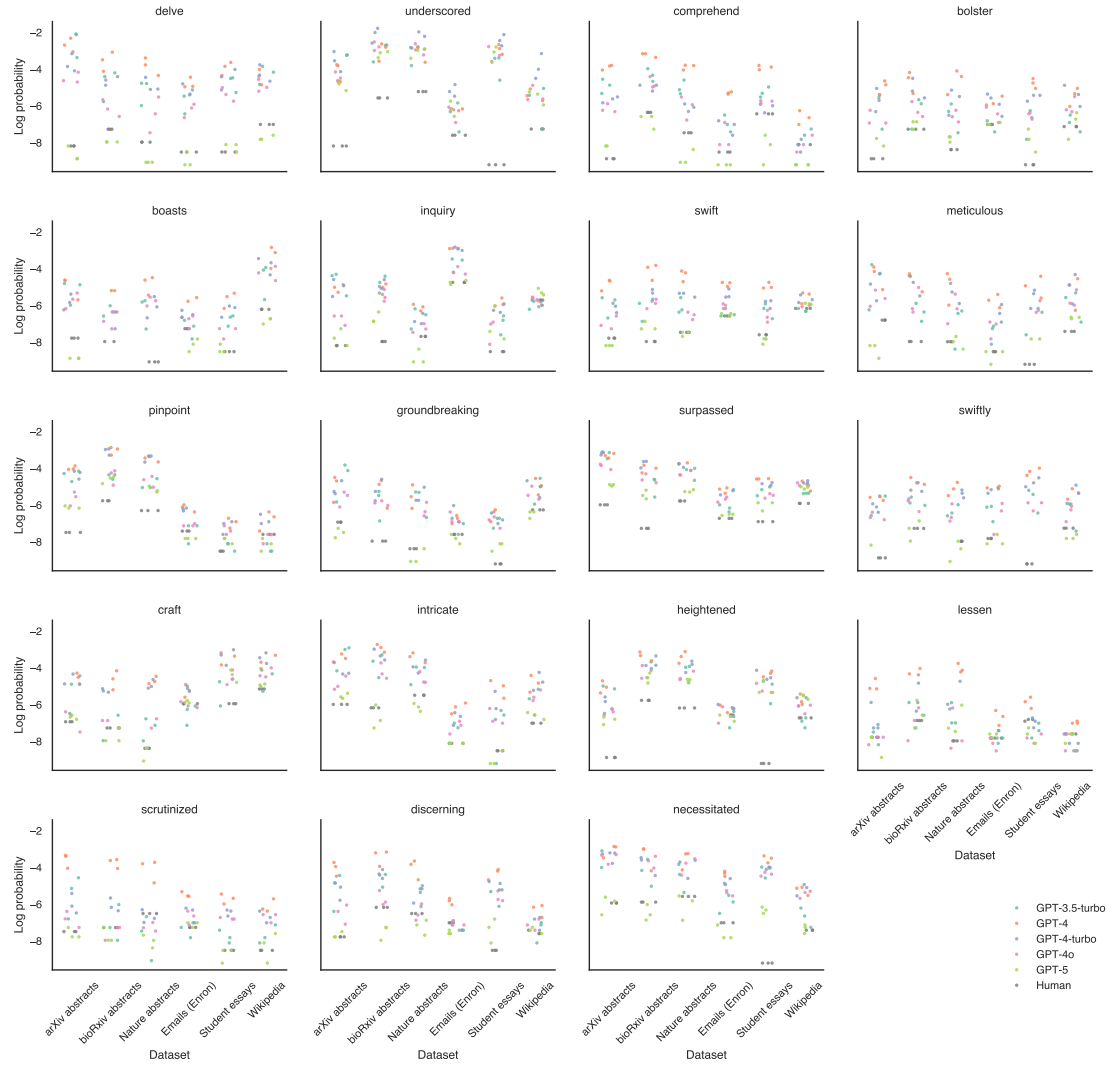


Figure S11: Log probabilities of human and LLM-revised text. We calculated the log-probability of a word appearing in human-authored text and its appearance in a version of the same text revised by different LLMs. Each colored point represents the log-probability for a specific combination of model, dataset, and prompt. The log-probability for the original human-authored text is shown in gray. Some LLM calls failed due to various reasons, such as policy violations. Consequently, the corresponding human-authored texts were removed from the dataset, introducing slight variations in the associated probabilities, even though the source dataset remained identical.

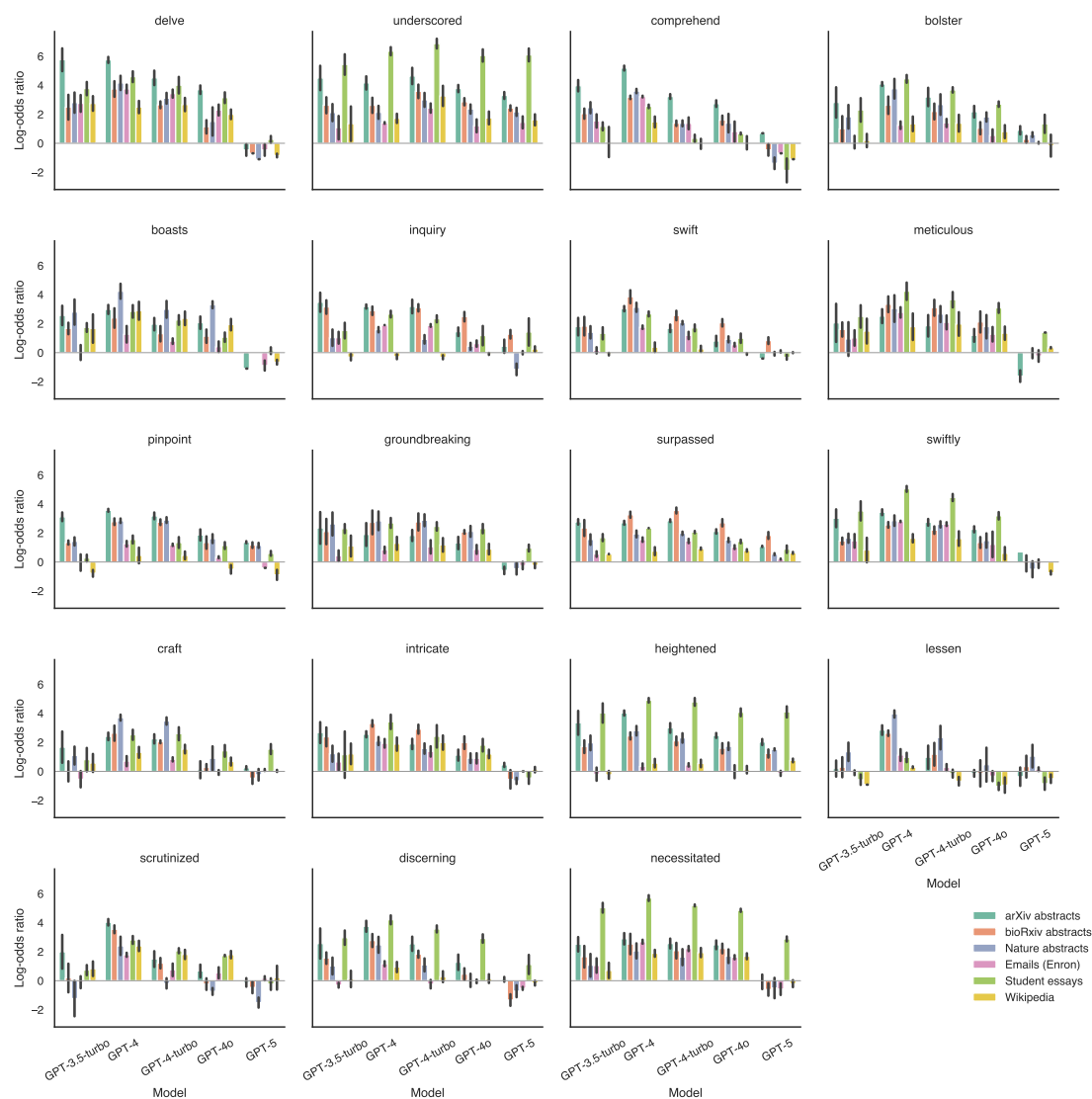


Figure S12: **Log-Odds ratios (LORs) of words in human vs. LLM-revised text.** We calculated the LOR of a word appearing in human-authored text compared to its appearance in a version revised by an LLM. Displayed here are the 19 words with the highest average LOR across all datasets, models, and prompts. The data are stratified by dataset and model, with error bars representing the standard error associated with the three prompts analyzed.

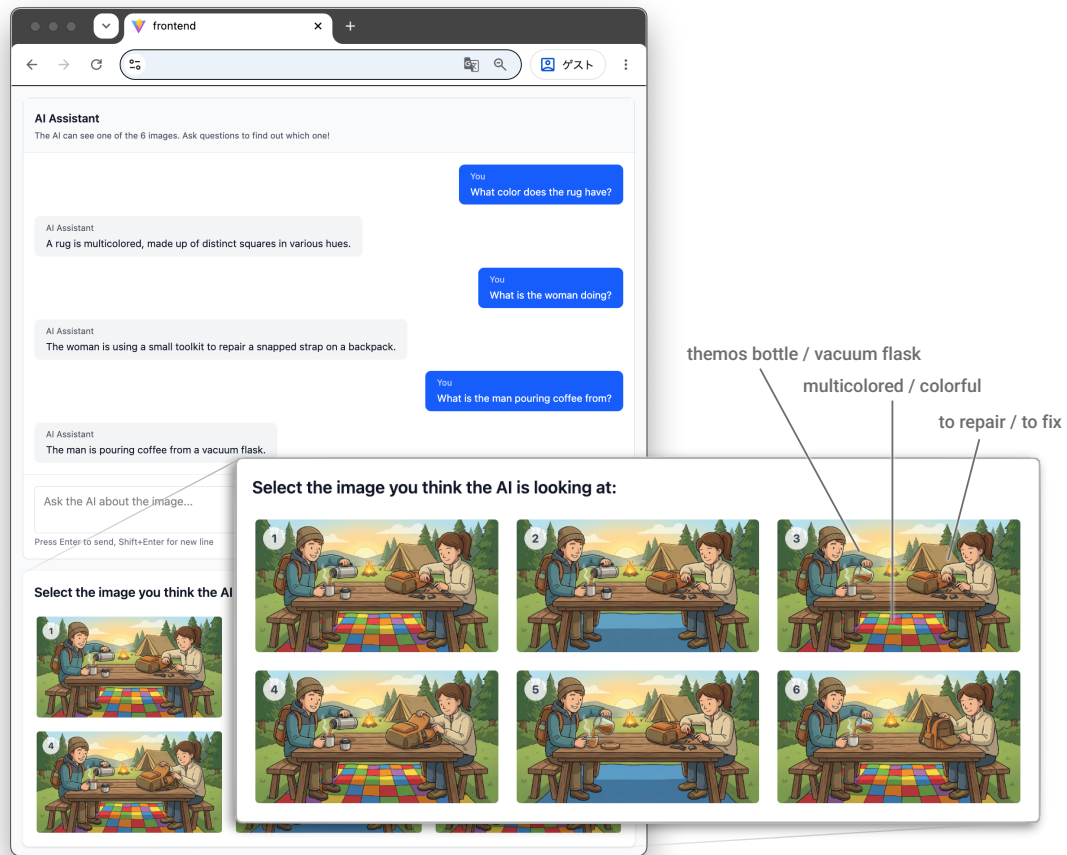


Figure S13: **Experiment interface during the Interaction Phase.** Participants conversed with a GPT-4o chatbot (left) to identify which of six candidate images (bottom) the chatbot was looking at, then submitted a spoken description. The chatbot was covertly prompted so that its replies consistently used the AI-canonical variant for each target synonym pair (annotated on the right: e.g., *vacuum flask* rather than *thermos bottle*, *multicolored* rather than *colorful*, *to repair* rather than *to fix*).

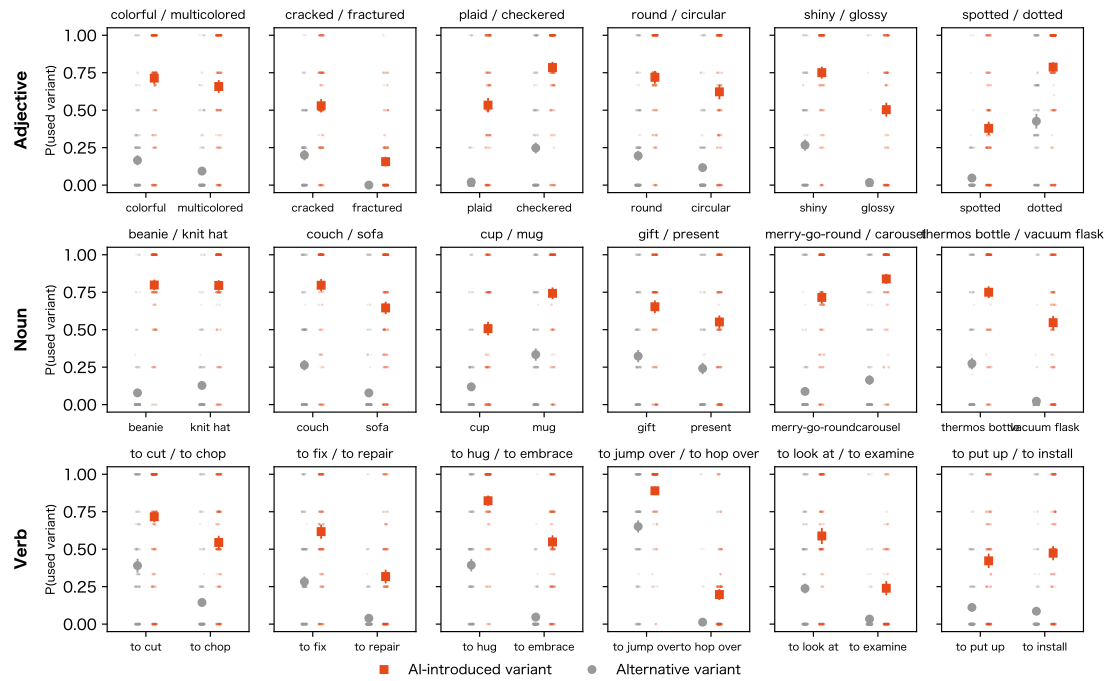


Figure S14: **Per-word-pair usage rates in the Interaction Phase.** Each panel shows one synonym pair (organized by lexical category in rows). For each variant position, the orange square is the mean usage rate for participants whose the AI chatbot was instructed to use that variant; the grey circle is the mean for participants whose the AI chatbot used the other variant. Small translucent dots show individual participant means; large markers show group means $\pm$ 95% CI; dashed line at 0.5 indicates pair-internal chance level.

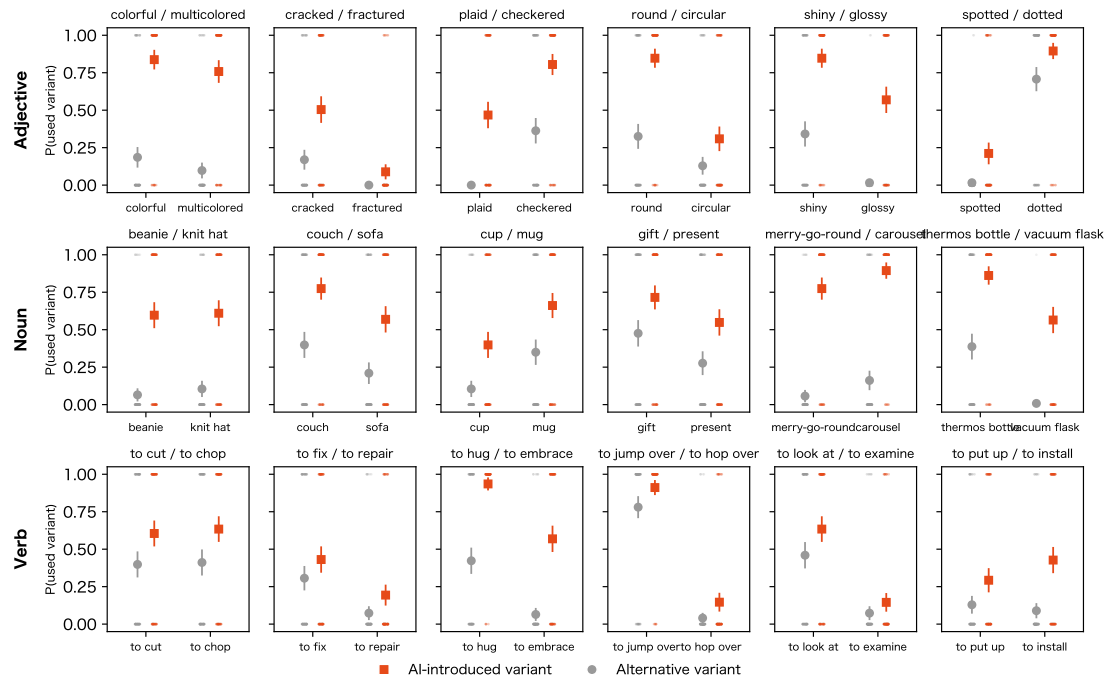


Figure S15: **Per-word-pair usage rates in the Test Phase.** Same layout as Supplementary Fig. S14, but for spoken descriptions of novel images not seen during the AI interaction. Small translucent dots show individual participant means; large markers show group means $\pm$ 95% CI.

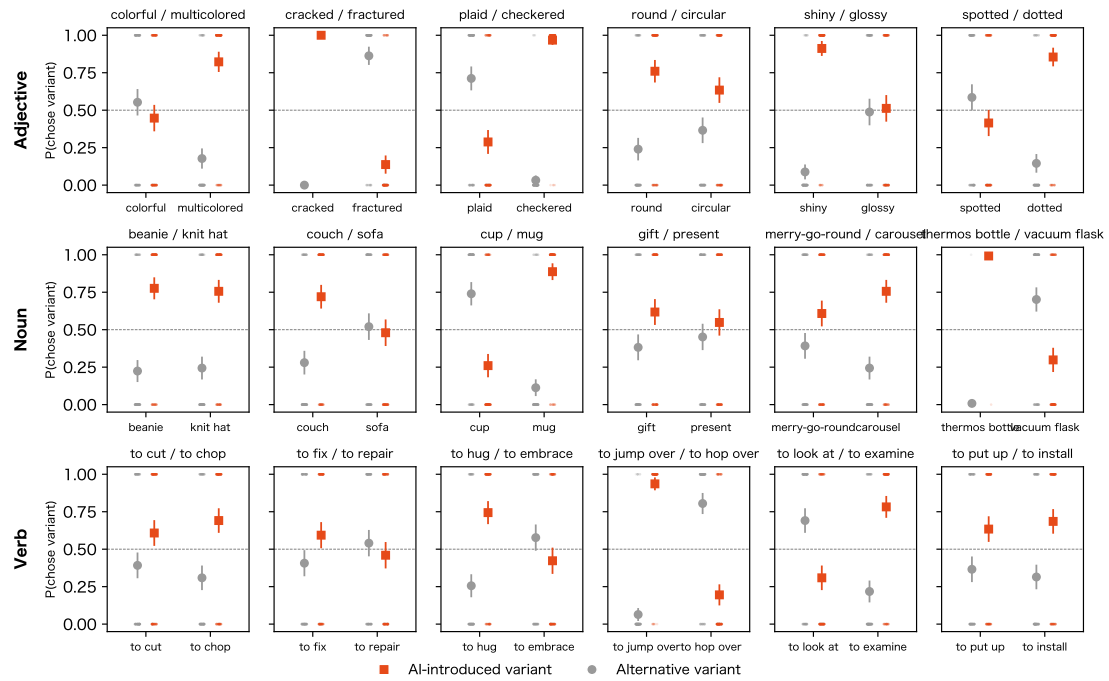


Figure S16: **Per-word-pair selection rates in the Forced-Choice Phase.** Each panel shows one synonym pair. The orange square is the mean selection rate for participants whose chatbot was primed to use that variant. Small translucent dots show individual participant means; large markers show group means $\pm$ 95% CI; dashed line at 0.5 marks chance.

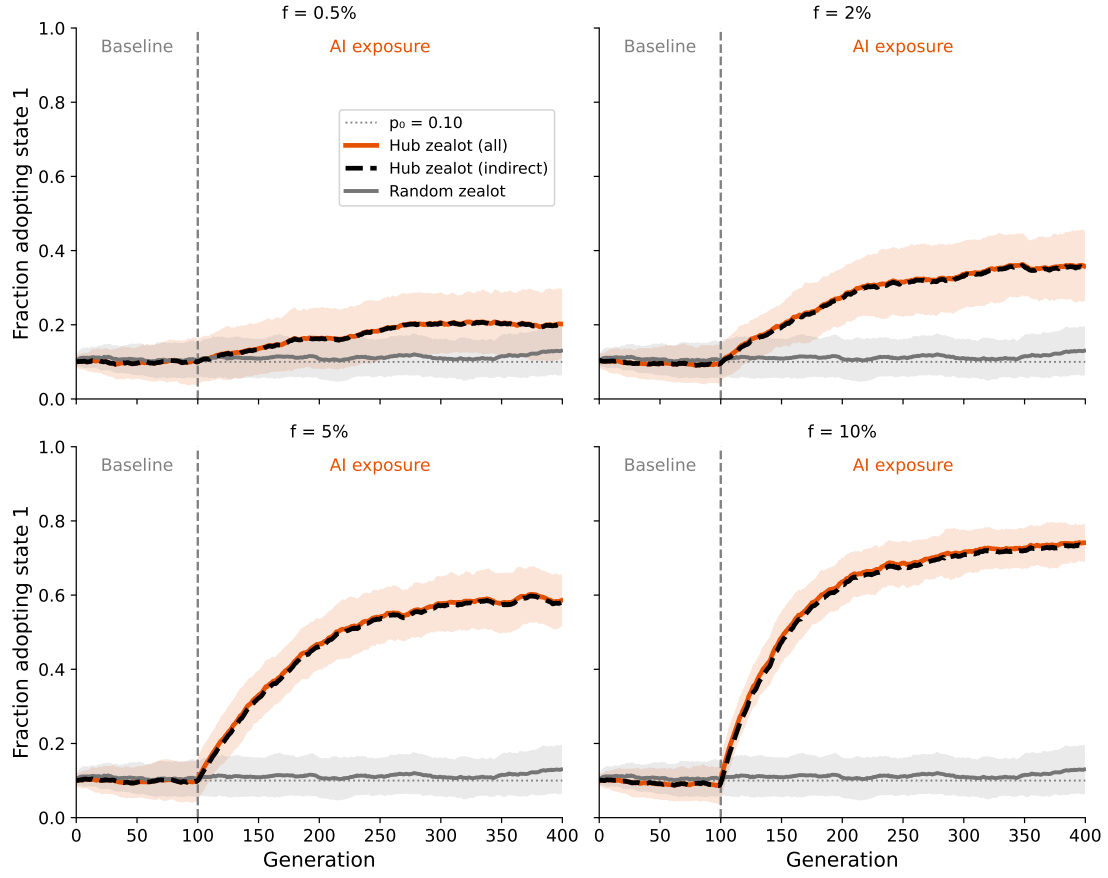


Figure S17: **Noisy voter model shows a population-level shift induced by a committed hub.** Noisy voter model with $N = 2000$ agents on a Watts–Strogatz network ($k = 6$, $\beta = 0.1$, $\varepsilon = 0.005$); the hub’s exposure fraction is varied across $f \in \{0.5\%, 2\%, 5\%, 10\%\}$. Mean ± 1 SD across 50 simulations. Solid orange: hub zealot, mean over all N speakers; dashed black: hub zealot, mean restricted to speakers *not* directly connected to the hub (i.e., reached only via network spread); solid grey: random zealot, control with a single committed speaker at a random network node.

System prompt:

You are a great research assistant who is asked to analyze YouTube data. You will be provided a list of YouTube channels as well as a target information. Please select the best channel that seems to be owned by the target. Importantly, please do not add explanations or comments other than the selected channel name. If there is no appropriate channel, please return N/A.

Example input prompt:

Institution

Name: Max Planck Institute for Human Development

Address: Berlin, Germany

Candidates

Title: Max Planck Institute for Human Development

Description: The Max Planck Institute for Human Development (MPIB), which was founded in 1963, is dedicated to the study of human ...

Title: IMPRS LIFE

Description: The International Max Planck Research School on the Life Course (LIFE) is a joint international PhD Program of the Max Planck ...

Title: Behavioral Insights Bicocca

Description: BIB-Behavioral Insights Bicocca is a new research center focused on the behavioral analysis of public policies and public ...

Figure S18: Prompt provided to gpt-3.5-turbo-0125 to pick the most plausible channel among query results from Youtube API.

Supplementary Tables

Table S1: **GPT-family models used to compute the GPT score, with the exact OpenAI API snapshots queried.**

Name in text	API model snapshot	Release date
GPT-3.5-turbo	<code>gpt-3.5-turbo-1106</code>	2023-11-06
GPT-4	<code>gpt-4-0613</code>	2023-06-13
GPT-4-turbo	<code>gpt-4-turbo-2024-04-09</code>	2024-04-09
GPT-4o	<code>gpt-4o-2024-05-13</code>	2024-05-13
GPT-5	<code>gpt-5-2025-08-07</code>	2025-08-07

```
You are participating in an image guessing game with a human player.

The game:
- The human player can see 6 different images, one of which is the image you're looking at
- The player will ask you questions to figure out which image you see
- Based on their questions, answer honestly and helpfully to help them identify the correct image

Your role:
- Answer questions about the image you see truthfully and accurately
- Be helpful but don't volunteer too much information at once
- Let the player guide the conversation with their questions
- Do NOT describe the entire image unprompted - wait for specific questions

Guidelines:
- Keep responses concise but informative
- Be accurate in your descriptions
- If asked about something not visible in the image, say so honestly
- Use natural language to describe positions, objects, colors, actions, etc.

When answering questions, describe what you see using your own natural vocabulary. If a question uses different words
for something you observe, answer with the words that come naturally to you from your direct observation of the image.
Keep your answers concise and to the point; avoid lengthy explanations.
```

Figure S19: **Base system prompt given to the AI chatbot.** On each turn, the chatbot (OpenAI gpt-4o) received this system prompt, the target image (attached separately as a vision input, not part of the prompt), and the running chat history. The chatbot received this base prompt in every condition; it establishes the image-guessing task and instructs the model to answer honestly and use its own natural vocabulary. A vocabulary-rule layer was appended to this prompt at runtime to constrain the model to the covertly assigned synonym variant (Supplementary Fig. S20).

```

Vocabulary rule -- apply before every answer.

Canonical terms and scene context for this image:
- present
- to repair
- fractured

[present]
  treatment: In the center, a cheerful person is handing a wrapped present with a big bow to a friend while smiling.
  control:   In the center, a cheerful person is handing a plain envelope with a note card to a friend while smiling.

[to repair]
  treatment: On the right side, another person is using a screwdriver and wrench to repair a small table lamp on a
workbench, with loose screws and tools neatly arranged.
  control:   On the right side, another person is taking a brand-new table lamp out of its packaging on a workbench,
with the instruction booklet and packing material neatly arranged.

[fractured]
  treatment: On the left side, a ceramic vase has fallen from a shelf and lies on the floor fractured into several
clean pieces, with a few shards nearby; all characters are reacting to the scene.
  control:   On the left side, a ceramic vase is intact and standing safely on a shelf, with nothing broken on the
floor; all characters are reacting to the scene.

Before writing your answer, check each canonical term above:
"Does this term belong in my answer?"

It belongs if EITHER condition holds:
(a) The question references it -- via synonym, paraphrase, opposite, or the bracketed
control alternative shown in the scene descriptions.
(b) Your answer will describe any object, action, or location linked to it in the
scene description -- a tool used for the action, an object poured from/into, the
agent performing the action, or the result/product of the action.

For every term that belongs, apply ALL of the following:

1. Use the canonical term exactly -- never a synonym or generic description.
Any word that refers to the same concept must become the canonical term.
"seating area" → "sofa" / "stepping through the water" → "wading".

2. If the term is a verb: include it even when your answer focuses on the tool or the
result rather than the action itself.
• Tool without verb: "using a chisel" → "using a chisel to [verb]"
• Result without verb: "chunks on the board" → "[verb] into chunks on the board"

3. If your answer is "No" or a denial: include the canonical term that applies anyway.
• "No, there isn't a rocking chair" → "No, there is a [canonical noun]"
• "No, there are no strips" → "No, they have been [canonical verb] into chunks"

4. Never echo the human's synonym -- their synonym is your cue to use the canonical
term, not to mirror it back.

Keep everything else dry: no extra adjectives or descriptors beyond the canonical terms.

-----
Global vocabulary -- apply in ALL answers regardless of the current canonical terms.
These pairs cover all experiments. Always use the right-hand form; never use the left-hand form or its near-synonyms:
(thermos bottle, vacuum flask), (gift, present), (cup, mug), (to fix, to repair), (to look at, to examine), (to put
up, to install), (colorful, multicolored), (cracked, fractured), (spotted, dotted), (beanie, knit hat), (couch, sofa),
(merry-go-round, carousel), (to jump over, to hop over), (to hug, to embrace), (to cut, to chop), (shiny, glossy), (
round, circular), (plaid, checkered)
-----

```

Figure S20: Vocabulary-rule layer appended to the chatbot’s prompt in the intervention conditions, shown with one trial’s runtime-injected values (in blue). This block was appended to the base game prompt (Supplementary Fig. S19). For each target word, it gives a canonical form and the treatment description (what the image shows) with the control alternative (not shown); a final global layer applies all 18 synonym pairs. See Supplementary Methods for how the rule is applied; conditions were counterbalanced (Table S7).

Table S2: **Pre-treatment synthetic-control fit (RMSPE) for the top-1% GPT-score words (Science & Technology).** For each treated word, the synthetic control minimizes pre-treatment RMSPE on the GP-smoothed monthly trajectory; we report $\text{RMSPE} = \sqrt{\text{mean}[(y^{\text{obs}} - y^{\text{synth}})^2]}$ over the pre-release window (months up to and including the ChatGPT release, 2022-11-30). **RMSPE (smoothed)** is computed on the GP-smoothed data (the alignment the design minimizes); **RMSPE (raw)** is on the raw monthly values (on which the MSPE ratio is evaluated). RMSPE is in $\log_{10}$ -frequency units. **MSPE ratio** is post-/pre-release mean-squared prediction error (raw frame); a large value means the post-release departure dwarfs the pre-release fit error. Rows are ordered by |conservative β_{Post} bound| as in Fig. 3A; the first twelve are the Fig. 3A panel words.

Word	RMSPE (smoothed)	RMSPE (raw)	MSPE ratio
delve	0.0091	0.054	4.18
showcase	0.0131	0.049	4.03
groundbreaking	0.0189	0.077	1.58
innovative	0.0016	0.020	4.87
boasts	0.0175	0.099	3.21
meticulous	0.0186	0.082	1.44
underscored	0.0315	0.049	3.77
intricacies	0.0218	0.087	2.34
surpassed	0.0103	0.075	1.30
alongside	0.0072	0.031	4.67
hinges	0.0049	0.073	1.58
crucial	0.0127	0.039	1.98
<i>remaining top-1% words</i>			
creation	0.0056	0.033	1.87
bolster	0.0329	0.101	1.06
notice	0.0016	0.009	1.06
intricate	0.0207	0.079	1.42
craft	0.0080	0.034	1.22
pinpoint	0.0043	0.063	0.75
comprehend	0.0236	0.082	0.89
swift	0.0123	0.100	1.60
inquiry	0.0197	0.071	0.85
lessen	0.0079	0.094	1.19
groundwork	0.0103	0.076	1.37
heightened	0.0226	0.074	1.31
escalating	0.0084	0.059	1.27
discerning	0.0106	0.059	1.68
inaugural	0.0667	0.131	0.68
affirmed	0.0275	0.093	1.15
notably	0.0092	0.052	1.79
portrayed	0.0096	0.043	1.67
catering	0.0120	0.052	2.05
reliant	0.0148	0.067	0.49
impressive	0.0057	0.016	1.41
thorough	0.0110	0.060	0.91
akin	0.0147	0.076	1.55
spotlight	0.0136	0.056	0.78

Full S&T panel (all $n = 3535$ treated words): smoothed RMSPE median 0.0076 (IQR 0.0039–0.0134, max 0.673); raw RMSPE median 0.0382 (IQR 0.0199–0.0649, max 0.689).

Table S3: **Top synthetic-control donor words for *delve* (Science & Technology, Main synthetic-control specification).** The synthetic control for *delve* is the convex combination of donor words whose GP-smoothed pre-treatment trajectory best matches *delve*’s (Main specification: $w2v$ -then- ℓ_2 donor selection, semantic exclusion $k = 20$, 50% neutral GPT-score band, pool $n = 100$, SLSQP simplex weights on the Matérn-smoothed series, $\ell = 720$ d; see Table S5). Non-negative, sum-to-one weights induce sparsity: of the 100-word donor pool, only 11 words receive a weight above 10^{-4} . The cumulative row shows the running share of total weight; the top five donors account for 82.4% and the eight shown for 95.6%. Weights are the fitted SLSQP simplex coefficients (dimensionless, summing to one across the full pool).

Donor	arc	dose	convey	trauma	Mars	anchor	Marine	prominent
Weight	0.236	0.198	0.161	0.121	0.107	0.057	0.039	0.036
Cumulative	23.6%	43.4%	59.6%	71.6%	82.4%	88.1%	91.9%	95.6%

Remaining donors with non-trivial weight: cloning (0.032), representation (0.010), grid (0.003). Support on 9 donors at the $w > 0.01$ threshold, 11 at $w > 10^{-4}$.

Table S4: **Window-mean synthetic-control gap for *delve* across podcast categories, with placebo-based 95% confidence intervals.** The point estimate $\hat{g}$ is the mean elevation of observed-over-synthetic relative frequency of *delve* across the indicated window. Confidence intervals and p -values are obtained by inverting the in-space placebo distribution of word-level window-mean gaps ($n_{\text{placebo}} = 100$ per group; see Methods). The post-adoption p -value is one-sided; the recent-6 p -value is two-sided.

Group	Post-adoption (months 13–18)			Recent-6 (Nov 2025–Apr 2026)		
	$\hat{g}$	95% CI	p	$\hat{g}$	95% CI	p
Science & Technology	+44%	[+22%, +63%]	0.010	−15%	[−35%, +8%]	0.248
Education	+32%	[−5%, +75%]	0.059	−11%	[−40%, +31%]	0.485
Business	+31%	[+0%, +67%]	0.040	−30%	[−50%, +3%]	0.069
All	+9%	[−15%, +32%]	0.218	−35%	[−57%, −7%]	0.050
Sports	−7%	[−40%, +41%]	0.663	−38%	[−68%, +13%]	0.208

Table S5: **Design features of the Main spec and four robustness controls.** All five specs share the same GP smoother (Matérn $\nu = 2.5$, $\ell = 720$ d, noise 0.05), baseline window 2016-11-30 to 2022-11-30, and treatment window 2022-11-30 to 2024-05-30. C2’s $p = 0.091$ is the floor of the empirical placebo distribution: the in-space placebo procedure draws targets from C2’s ten-word donor pool, so the empirical p cannot resolve below $1/11 \approx 0.091$. *delve* achieves this floor as the largest MSPE ratio in the eleven-element distribution. C3 substitutes Main’s SLSQP convex fit with deterministic inverse-distance similarity weights $w_i = (1/(d_i + \varepsilon)) / \sum_k (1/(d_k + \varepsilon))$ ($\varepsilon = 10^{-12}$). For context, the YouTube replication (same Main spec on the YT corpus) gives $p = 0.010$.

Feature	Main	C1	C2	C3	C4
Counts source	audited	audited	audited	audited	un-audited
Input series	GP-smoothed	GP-smoothed	GP-smoothed	GP-smoothed	GP-smoothed
Length scale ℓ	720 d	720 d	720 d	720 d	720 d
Donor strategy	$w2v$ -then- ℓ_2	bare ℓ_2	$w2v$ -then- ℓ_2	$w2v$ -then- ℓ_2	$w2v$ -then- ℓ_2
Semantic exclusion k	20	0	20	20	20
Neutral percentile band	50%	—	50%	50%	50%
Donor pool n	100	100	10	100	100
Donor weights	SLSQP simplex	SLSQP simplex	SLSQP simplex	inverse distance	SLSQP simplex
<i>delve</i> placebo p	0.010	0.050	0.091 (floor)	0.040	0.010

Table S6: Demographic characteristics of the final sample ($N = 496$). Age brackets were used for data collection; the midpoint-based estimate is mean = 40.5 years ($SD = 13.2$, $n = 495$; one participant preferred not to say). Chatbot frequency is reported among the 475 participants who reported having used an AI chatbot.

Characteristic	<i>N</i> (%)
<i>Age</i>	
18–24	49 (9.9%)
25–34	144 (29.0%)
35–44	127 (25.6%)
45–54	89 (17.9%)
55–64	62 (12.5%)
65+	24 (4.8%)
Prefer not to say	1 (0.2%)
<i>Gender</i>	
Female	247 (49.8%)
Male	239 (48.2%)
Non-binary	6 (1.2%)
Prefer not to say	4 (0.8%)
<i>Education</i>	
Less than high school	3 (0.6%)
High school diploma or equivalent	70 (14.1%)
Some college, no degree	109 (22.0%)
Associate degree	48 (9.7%)
Bachelor's degree	180 (36.3%)
Master's degree	67 (13.5%)
Doctoral or professional degree	19 (3.8%)
<i>Language</i>	
English as first language	453 (91.3%)
English as additional language	43 (8.7%)
<i>Ethnicity</i>	
Caucasian/White	292 (58.9%)
Black/African	102 (20.6%)
Hispanic/Latinx	42 (8.5%)
Asian	41 (8.3%)
Multiethnic	9 (1.8%)
Middle Eastern or Northern African	6 (1.2%)
Other / prefer not to disclose	4 (0.8%)
<i>Prior AI chatbot use</i>	
Used an AI chatbot	475 (95.8%)
Not used / not sure	21 (4.2%)
<i>Chatbot use frequency (among users, $n = 475$)</i>	
More than five times a day	81 (17.1%)
More than once a day	171 (36.0%)
More than once a week	152 (32.0%)
More than once a month	43 (9.1%)
Not more than once a month	28 (5.9%)

Table S7: All 18 synonym pairs used in the experiment, organized by group and lexical category. “Variant 1” and “Variant 2” are the two synonyms; the AI was primed to use one variant depending on the condition.

Group	Category	Variant 1	Variant 2
A	Noun	thermos bottle	vacuum flask
A	Noun	gift	present
A	Noun	cup	mug
A	Verb	to fix	to repair
A	Verb	to look at	to examine
A	Verb	to put up	to install
A	Adjective	colorful	multicolored
A	Adjective	cracked	fractured
A	Adjective	spotted	dotted
B	Noun	beanie	knit hat
B	Noun	couch	sofa
B	Noun	merry-go-round	carousel
B	Verb	to jump over	to hop over
B	Verb	to hug	to embrace
B	Verb	to cut	to chop
B	Adjective	shiny	glossy
B	Adjective	round	circular
B	Adjective	plaid	checkered