Mecha-nudges for Machines

Giulio Frey
University of Chicago
giulio@uchicago.edu

Kawin Ethayarajh
University of Chicago
kawin@uchicago.edu

Abstract

As AI agents make decisions in the same environments as humans, the environments themselves can change to selectively influence them. We call this *mecha-nudging*: subtle changes to how choices are presented that systematically influence AI agents without materially degrading the decision environment for humans. It can arise from both intentional and unintentional adaptation to the rise of agents. To measure this phenomenon, we combine two frameworks—*Bayesian persuasion* from economics and $\mathcal{V}$*-usable information* from computer science—to get a common unit (bits) for quantifying how environments change across a wide range of interventions, contexts, and models. Applying our framework to over six million product listings on Etsy—a global marketplace for independent sellers—we find that after ChatGPT's release, listings contain significantly more machine-usable information about agentic curation, rising by over 40% of the maximum possible increase. This shift is robust across prompts, token choices, and models; absent in a regulated-text placebo; and cannot be explained by the rise of LLM-assisted writing. In contrast, a human study finds little to no change in human-usable information. Our results provide the first large-scale evidence that mecha-nudging is already happening in the wild, but going unnoticed.

1 Introduction

Nudges are subtle changes to the way choices are presented to human decision-makers, with the goal of shifting their behavior in a particular direction [Thaler and Sunstein, 2008]. What distinguishes nudges from other interventions is that they must be easy to avoid. Anything that removes options or noticeably changes economic incentives cannot be a nudge. For example, placing healthy foods at eye-level to encourage better eating is a nudge; banning or taxing unhealthy food is not. Nudges work by exploiting limitations of human cognition, such as limited attention and sensitivity to how outcomes are framed. They have been adopted worldwide by policymakers and businesses alike: for example, in the Dominican Republic, nudges designed to increase tax compliance boosted tax revenue by \$184M (or $\sim$0.22% of GDP) in 2018 alone [Holz et al., 2023].

For decades, nudges exclusively targeted individuals or groups of humans, as they were the only decision-makers. This is no longer the case. Many decisions are now made by AI agents, often in spaces still inhabited by humans. With little to no human oversight, they can shortlist job applicants, book travel arrangements, ban content, and more. As agents become decision-makers in their own right, a natural question follows: can they be nudged too?

We introduce the concept of *mecha-nudges*: changes to how choices are presented that systematically influence the behavior of AI agents without materially degrading the decision environment for humans. For example, an online seller might add specific product descriptors to a listing—say, *high customer satisfaction*—that may do little to sway a human buyer already looking at the product reviews, but that significantly steer the behavior of a shopping agent. However, this definition is outcome-based, not intent-based: a mecha-nudge may result from deliberate optimization, imitation of other mecha-nudged content, or even incidental changes.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

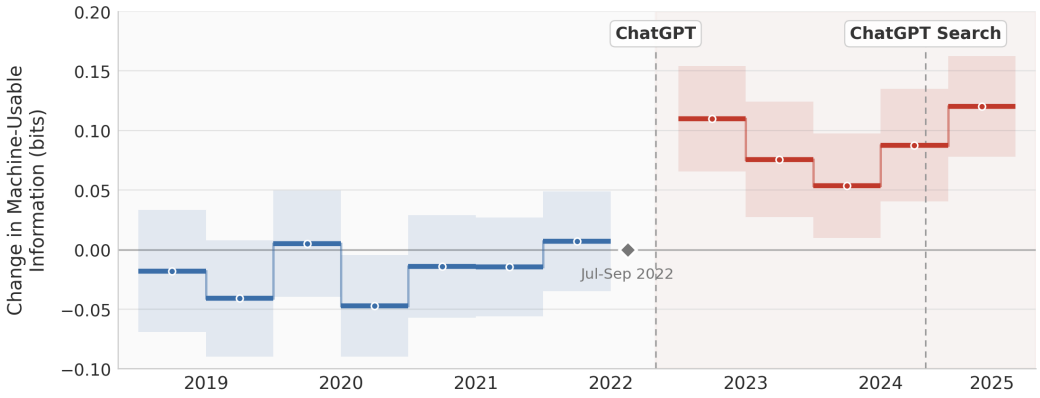


Figure 1: After the release of ChatGPT, the change in machine-usable information in Etsy listings increases significantly. The change in each half-year relative to the Jul-Sep 2022 period (grey diamond) is plotted here. The effect attenuates over the following year before climbing again in late 2024. This coincides with the release of ChatGPT Search, which—unlike its predecessors—could natively browse live listings instead of only making recommendations based on stale training data.

Mecha-nudges should not be conflated with prompt injection or traditional search engine optimization. Prompt injection acts directly on the model, making it hard to avoid and often depriving the agent of options it would have ordinarily had [Willison, 2022]; mecha-nudges preserve options and act on the environment instead. Traditional SEO manipulates machine-readable signals (e.g., backlinks) to influence how information is presented to humans, who remain the final decision-makers [Hagendorff, 2021]. Mecha-nudges, by contrast, target AI systems that make decisions themselves.

Although nudges are most often discussed informally in behavioral economics, a common formalization comes from *Bayesian persuasion* [Kamenica and Gentzkow, 2011], which explores how a sender might influence the actions of a receiver by controlling the information environment (§2). We combine this with the notion of $\mathcal{V}$*-usable information*, a generalization of Shannon information that is observer-relative [Xu et al., 2020, Ethayarajh et al., 2022]. Under our formalization, the goal of mecha-nudging is to maximize the amount of machine-usable information that the environment contains about the desired machine behavior, while not decreasing the amount of human-usable information (§3). This allows all mecha-nudges to be measured on a common scale (bits of usable information), permitting comparisons across different interventions, settings, and models.

There have been anecdotes of mecha-nudging (and evidence that nudges for humans may extend to machines [Cherep et al., 2025]), but we provide the first large-scale evidence of mecha-nudging by analyzing Etsy, a global marketplace for independent sellers (§4). Etsy is an ideal setting: it has integrated AI-driven features for both buyers and sellers, and a large fraction of referral traffic comes from ChatGPT alone [Smith, 2025]. We find that product listings created after the release of ChatGPT have significantly more machine-usable information about agentic curation decisions (Figure 1), with the machine-usable information jumping 0.143 bits in the post-period (over 40% of the maximum possible increase). This shift is robust to prompt formulations, token choices, and model families. It persists after controlling for product- and seller-level attributes, and is absent in product categories where human buyers are sensitive to AI use (e.g., art and collectibles [Millet et al., 2023]). Using the number of product reviews as a proxy for cumulative sales, we find that greater machine-usable information is associated with significantly more commercial success in the post-ChatGPT era, but not pre-ChatGPT. Although our data is observational, these findings are consistent with listings being optimized—whether deliberately or through imitation of successful sellers—to increase machine-usable information.

Mecha-nudges create a deployment feedback loop: once agents make decisions on their own, the environments they inhabit become optimized for them. This means model behavior in the wild is determined not only by model developers, but also by the strategic and incidental adaptation of the world in which those models are deployed. Static evaluations miss this loop. Our framework makes such adaptation measurable, and our paper provides the first large-scale evidence that it is already happening in the real world.

2 Background

We now review nudges, Bayesian persuasion, and $\mathcal{V}$-usable information, the three ingredients of our formalization of mecha-nudges in §3.

2.1 Nudges

Early work in behavioral economics found that because cognitive resources are finite, the way in which choices are presented has a significant impact on human decisions [Tversky and Kahneman, 1981]. For example, framing medical treatments in terms of their survival rates instead of their mortality rates makes it far likelier that patients will consent [McNeil et al., 1982]. Thaler and Sunstein [2008] introduced the term *choice architecture* to describe the deliberate design of a decision environment, wherein *nudges* denote any part of the architecture that shifts human behavior without removing options or altering incentives.

Nudges have enjoyed broad adoption as a policy tool. For example, automatic enrollment in U.S. retirement plans increased participation rates for new hires from 49% to 86% [Madrian and Shea, 2001]. In Denmark, automatic contributions to retirement accounts proved far more effective than costly tax subsidies, which generate only one additional cent of saving per dollar of expenditure [Chetty et al., 2014]. Social comparisons—in the form of letters comparing residents' electricity use to that of their neighbors—decreased energy consumption by $\sim$2% on average [Allcott, 2011]. Recent experimental evidence suggests that LLM agents are more susceptible to choice architecture intended for humans than humans themselves [Cherep et al., 2025, 2026].

Nudges are most often discussed informally, as context-specific interventions. However, they can be mathematically formalized under some frameworks. For example, *rational inattention* proposes that hidden signals may not be worth investigating because of information acquisition costs; nudges can be framed as changes in these costs [Matějka and McKay, 2015, Sims, 2003]. Nudges can also be analyzed through the lens of *prospect theory* as changes to the framing of an outcome [Tversky and Kahneman, 1992]. We refer the reader to Appendix J for more details.

2.2 Bayesian Persuasion

In information design, the *Bayesian persuasion* framework provides a popular mathematical formalization of nudges [Kamenica and Gentzkow, 2011, Taneva, 2019, Kamenica, 2019]. In brief:

1. The choice architect (i.e., the sender) and the decision-maker (i.e., the receiver) start with the same prior belief μ_0 about the distribution of some random variable Z.
2. To influence the decision-maker's belief, the choice architect selects a distribution $\pi(\cdot|Z)$ over a possible set of signals $\mathcal{S}$ and commits to it before seeing any instantiation of Z.
3. The decision-maker sees both π and an instantiated signal $s \in \mathcal{S}$, forms a posterior belief $\mu(z|s)$ by applying Bayes' rule, and then takes a utility-maximizing action a^*. The choice architect, with utility function v, has a final expected utility $\mathbb{E}_{z\sim\mu_0}\mathbb{E}_{s\sim\pi(\cdot|z)}[v(a^*(\mu(\cdot|s)), z)]$.

Under this framework, the nudge is the signal distribution $\pi(\cdot|Z)$. By choosing how and what information is revealed, the choice architect shapes the decision-maker's posterior belief—and thus the action they take—without changing the feasible set of actions or payoffs. The choice architect's objective is thus to select $\pi \in \Pi$ that maximizes their expected utility.

When the set of possible signals is large and unstructured, as with free-form text, explicitly specifying the signaling scheme and computing exact Bayesian updates becomes intractable. However, maximizing machine-usable information—which is tractable—can be seen as a bounded-receiver analog of Bayesian Persuasion when the receiver is restricted to a specific model class (Proposition 3.2).

2.3 Usable Information

$\mathcal{V}$-usable information is an observer-relative generalization of Shannon information. To understand why it is useful, consider a model family $\mathcal{V}$ that can learn to map an English sentence X to its French translation Y. If we encrypted X, the amount of Shannon information that X contains about Y would be the same, but translating it would be far more difficult because the information it contains is

no longer usable by $\mathcal{V}$ [Xu et al., 2020]. Conversely, if we decrypted the encrypted text, we would increase the amount of usable information. Although this violates the data processing inequality, it is useful for understanding many real-world phenomena, such as why representation learning is helpful and why some datasets are more difficult to learn from than others [Ethayarajh et al., 2022].

Xu et al. [2020] propose measuring the amount of $\mathcal{V}$-usable information through a framework called *predictive $\mathcal{V}$-information*. We will now re-state the formal definitions in this framework (and its extension to pointwise examples, by Ethayarajh et al. [2022]).

Definition 2.1 (**$\mathcal{V}$-Usable Information**)**.** Let X, Y denote random variables with sample spaces $\mathcal{X}, \mathcal{Y}$ respectively, and let $\varnothing$ denote a null input that is uninformative about Y. Given predictive family[1] $\mathcal{V} \subseteq \Omega = \{f : \mathcal{X} \cup \{\varnothing\} \to P(\mathcal{Y})\}$, the predictive $\mathcal{V}$-entropy is

$$H_{\mathcal{V}}(Y) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)] \tag{1}$$

and the conditional $\mathcal{V}$-entropy is

$$H_{\mathcal{V}}(Y \mid X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)] \tag{2}$$

The $\mathcal{V}$-usable information (or simply $\mathcal{V}$-information) is $I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y \mid X)$.

Definition 2.2 (**Pointwise $\mathcal{V}$-Information**)**.** Given random variables X, Y and a predictive family $\mathcal{V}$, the pointwise $\mathcal{V}$-information (PVI) of an instance $(x, y) \sim (X, Y)$ is

$$\text{PVI}(x \to y) = -\log_2 g[\varnothing](y) + \log_2 g'[x](y) \tag{3}$$

where $g = \arg\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)]$ and $g' = \arg\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)]$.

In brief, $f[X]$ and $f[\varnothing]$ produce a probability distribution over the output space. The goal is to find the $f \in \mathcal{V}$ that maximizes the log-likelihood of the output data with the input (2) and without it (1). For example, say $\mathcal{V}$ is the family induced by a specific LLM such as Llama-3.1-8B-Instruct with no frozen parameters [Grattafiori et al., 2024]. As in our running example, let X be an English sentence and Y a French sentence. g would be a model trained to produce French with no context (i.e., a French LLM) and g' would be a model trained to produce the French translation of an English sentence. The PVI of an instance (x, y) is then the difference in log-probability these models assign to the true French translation. Analogous to the relationship between PMI and Shannon information:

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{x,y \sim P(X,Y)}[\text{PMI}(x, y)] \\ I_{\mathcal{V}}(X \to Y) &= \mathbb{E}_{x,y \sim P(X,Y)}[\text{PVI}(x \to y)] \end{aligned} \tag{4}$$

Unlike $\mathcal{V}$-usable information, PVI can be negative, indicating that the model performs better by ignoring the input. The PVI of an instance depends on the underlying distribution; the same instance drawn from different distributions will generally yield different PVI values. Although one can also take averages of any arbitrary sub-population of the data, it would be imprecise to call their average PVI the $\mathcal{V}$-usable information, since it is a different distribution than the one used to train the model.

Although $I_{\mathcal{V}}(X \to Y)$ is asymmetric, when $\mathcal{V}$ is the set of all possible functions, $\mathcal{V}$-usable information reduces to Shannon information. In practice however, implicit in estimating the $\mathcal{V}$-usable information is the assumption that the data used to find the optimal $f \in \mathcal{V}$ and the held-out data used to estimate $H_{\mathcal{V}}(Y \mid X)$ are identically distributed. Moreover, with large model families such as LLMs, there is no global optimality guarantee; we are assuming that the converged model maximizes the log-likelihood.

The key strength of this framework is that it permits a wide range of comparisons to be done on a common scale (bits of usable information):

(i) different model families $\mathcal{V}$ by computing $I_{\mathcal{V}}(X \to Y)$ with the same X, Y

(ii) different data distributions (X, Y) by computing $I_{\mathcal{V}}(X \to Y)$ with the same $\mathcal{V}$

(iii) different transformations τ of the input by computing $I_{\mathcal{V}}(\tau(X) \to Y)$ with the same $\mathcal{V}, X, Y$

(iv) different instances (x, y) by computing $\text{PVI}(x \to y)$ with the same $\mathcal{V}, X, Y$

(v) different slices of data by comparing the mean PVI within each slice

[1] A predictive family is a subset of all possible mappings from $\mathcal{X}$ to $P(\mathcal{Y})$ that satisfies *optional ignorance*: for any P in the range of some $f \in \mathcal{V}$, there exists some $f' \in \mathcal{V}$ s.t. $f[X] = f'[\varnothing] = P$. We refer the reader to Xu et al. [2020] for a more complete understanding of why optional ignorance is important, as well as for PAC bounds on the estimation error. Neural networks without frozen parameters handily meet this definition, as do human learners.

3 Formalizing Mecha-nudges

Consider a shopkeeper who wishes to increase the online sales of a particular product. She first redesigns the webpage to better draw attention to its benefits; this nudges humans to buy it. But she does not want *every* visitor to buy the product either—she wants those with larger budgets to choose higher-margin alternatives, for example. The 'select' vs. 'pass' decision she wants people to make is a random variable Y conditioned on random variable X (representing both the decision environment E she can control, like the webpage, and the buyer characteristics U she cannot). Increasingly, she finds that AI agents are either buying the product themselves or recommending it to humans, not merely re-ordering the options for humans to ultimately decide, as in SEO. She must transform the webpage so that the AI agents also decide in the desired manner, but without putting off humans.

We formalize the shopkeeper's dilemma as the transformation of a decision environment that maximizes machine-usable information while not materially reducing human-usable information.

Definition 3.1 (**Mecha-nudging Design**). Let random variable $X = (E, U) \in \mathcal{X}$, with $E \in \mathcal{E}$ (controllable environment) and $U \in \mathcal{U}$ (uncontrollable exogenous characteristics). Let random variables $Y_H, Y_M \in \mathcal{Y}$ represent the decision the choice architect wants the human and machine decision-makers to make respectively. Let $\mathcal{H}, \mathcal{M}$ denote the predictive families for decision-making by humans and machines, and $\epsilon \in \mathbb{R}_{\geq 0}$ the tolerable decrease in human-usable information. Where $\mathcal{T} \subseteq \{\tau : \mathcal{X} \to \mathcal{X}\}$ is the set of available transformations, all satisfying $\tau(E, U) = (\tau_E(E), U)$ for some map $\tau_E : \mathcal{E} \to \mathcal{E}$, the choice architect's problem is:

$$\begin{aligned} &\arg\max_{\tau \in \mathcal{T}} \; I_{\mathcal{M}}(\tau(X) \to Y_M) \\ &\text{s.t. } I_{\mathcal{H}}(\tau(X) \to Y_H) \geq I_{\mathcal{H}}(X \to Y_H) - \epsilon \end{aligned} \tag{5}$$

This is *constrained mecha-nudging design*. When $\epsilon \geq I_{\mathcal{H}}(X \to Y_H)$, the constraint is trivial (as usable information is non-negative) and the problem is one of *unconstrained mecha-nudging design*.

The constraint in (5) asserts that in applying the mecha-nudge, the amount of human-usable information must not decrease by more than ϵ bits. For example, a transformation that converts a modern webpage into a text-only listing might make it easier for an AI agent to parse (higher $I_{\mathcal{M}}(\tau(X) \to Y_M)$) but may create a worse browsing experience for humans (lower $I_{\mathcal{H}}(\tau(X) \to Y_H)$). Note that the constraint does not concern what the typical human *would* do, but rather what humans in the aggregate *could* do; this is not a behavioral constraint, but a normative one.

Implicit in this constraint is the assumption that humans and AI agents are operating in the same decision environment. Operationalizing (5) may require surveying humans or a proxy for $\mathcal{H}$, such as a learned model of human behavior [Santurkar et al., 2023]. In some applications, one could instead justify the constraint institutionally—for example through market or regulatory pressures—or relax it entirely by choosing ϵ large enough to make it non-binding.

Proposition 3.2 (**Bounded-Receiver Bayesian Persuasion**). *Consider a bounded-receiver analog of Bayesian persuasion in which both the choice architect and decision-maker have log-scoring utility* $\log_2(\cdot)$*, and the decision-maker is restricted to predictive family* $\mathcal{M}$*. Then* $\arg\max_{\tau \in \mathcal{T}} I_{\mathcal{M}}(\tau(X) \to Y_M)$*, the solution to unconstrained mecha-nudging, also maximizes the best achievable expected utility for the decision-maker.*

The proof is given in Appendix A.

In observational settings, the choice architect's latent objective is typically unobserved. Moreover, a realized mecha-nudge need not be deliberate; it may arise through direct optimization, imitation of successful content, or other endogenous adaptation. We therefore study realized mecha-nudging with respect to a focal machine action y^* that represents some specific behavior of interest:

Definition 3.3 (**Realized Mecha-nudge**). Let $A_H, A_M \in \mathcal{Y}$ denote the actions taken by the human and machine decision-makers. Let $\phi : \mathcal{Y} \to \mathcal{Z}$ be a task-relevant coarsening of the machine action, and define $Z_M = \phi(A_M)$. A transformation $\tau^* \in \mathcal{T}$ is a realized mecha-nudge with respect to Z_M if

$$I_{\mathcal{M}}(\tau^*(X) \to Z_M) > I_{\mathcal{M}}(X \to Z_M) \tag{6}$$

and

$$I_{\mathcal{H}}(\tau^*(X) \to A_H) \geq I_{\mathcal{H}}(X \to A_H) - \epsilon. \tag{7}$$

The identity map $\phi(A_M) = A_M$ gives a multi-action notion of mecha-nudging; the binary map $\phi(A_M) = 1\{A_M = y^*\}$ recovers the focal-action setting used in our empirical analysis.

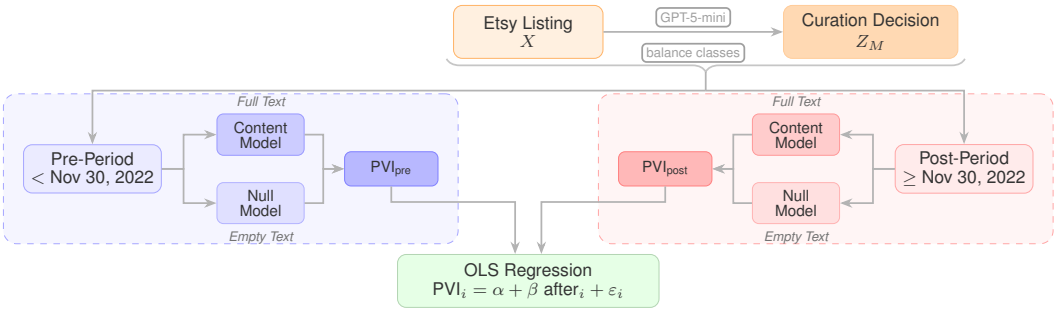


Figure 2: Our pipeline for estimating the change in usable information between the pre- and post-ChatGPT periods: construct decision labels Z_M, train content and null models for each period, and run an OLS regression of the pointwise $\mathcal{V}$-information (PVI). This describes our baseline experiment, of which we run many variations (including with controls).

Importantly, a realized mecha-nudge need not increase the marginal probability of any particular action. It could make selection more likely, rejection more likely, or simply sharpen a distribution with multiple possible actions. Our goal is not necessarily persuasion in a single direction as much as a systematic influence on machine behavior that is mediated through environmental changes.

4 Systematic Evidence of Mecha-nudging

We now provide empirical evidence that realized mecha-nudging is already occurring at scale. The shopkeeper's dilemma that motivated our formalization in §3 is not a mere hypothetical. As such, we study product listings on Etsy, a global marketplace for independent sellers. Etsy is uniquely exposed to AI: not only does a large fraction of referral traffic come from ChatGPT [Smith, 2025], but it was the first platform to enable its products to be purchased directly in ChatGPT [OpenAI, 2025].

We employ a three-step pipeline (Figure 2): (i) construct an agent curation decision Z_M with GPT-5-mini as a proxy for ChatGPT[2]; (ii) finetune an open-weights model (Llama-3.1-8B-Instruct by default) to create the content and null models g', g; (iii) regress listing-level PVI on the period in which it was created, along with other controls. For each temporal partition used in the analysis (pre/post and, where applicable, half-year bins), we fit separate content and null models on that period's data and compute PVI (Definition 2.2) on held-out data to get pointwise estimates of usable information. We then run a human study on the same listings to test whether the human-side constraint (7) is satisfied.

Our empirical estimand is a change in machine-usable information about agentic decisions, not a change in the unconditional rate of a particular action. Higher $I_M(X \to Z_M)$ means the listing text makes the agent's decision more predictable relative to a null input. This could be because attractive listings become easier to select, unattractive listings become easier to reject, or both. Note that because we use observational data, our empirical analysis does not attempt to establish seller intent, nor is seller intent needed for a transformation to be a mecha-nudge; it could arise, for example, from some sellers simply imitating their more successful counterparts.

4.1 Data

The raw dataset (Appendix B) comprises over six million Etsy listings, with 1.06 million uploaded pre-ChatGPT and 5.0 million uploaded post-ChatGPT (November 30, 2022)[3]. In our baseline specification, X is the controllable textual content. Other characteristics such as the price, the number of reviews, the average rating, and more are used as controls. We operationalize the focal machine action y^* by prompting GPT-5-mini, a proxy for the basic version of ChatGPT used by most consumers, to issue a select/pass decision for each listing[4]; the resulting indicator serves as Z_M.

[2]We use the Jan 2026 version of the model.

[3]Since the data is a Nov 2025 snapshot, we observe listings at scrape time rather than at initial creation; any post-creation edits to older listings would, if anything, understate the extent of mecha-nudging.

[4]As robustness checks, we also generate labels with Gemma-3-27B-IT and Qwen3-32B, and try different pairs of words for the select/pass decision; see Appendix C. As also detailed in Appendix C, we subsample the data to balance the class distribution, as large imbalances lead to noisy estimates of the usable information.

Note that this output variable is constructed specifically to study agentic curation decisions. If one wanted to study the impact on some other aspect of agent behavior (e.g., agent refusals due to safety concerns), one would have to construct a different set of labels.

4.2 Methods

Human-Usable We first test the human-side constraint (7) indirectly via a contrapositive argument. If human-usable information had noticeably declined, then we would expect observable human outcomes to have deteriorated as well. Several independent proxies indicate that this has not occurred: (1) marketplace-level spending has remained stable, with Gross Merchandise Sales per active buyer on the Etsy marketplace ranging from roughly $117 to $136 on a trailing 12-month basis between 2020 and 2025 [Etsy, Inc., 2026b]; (2) buyer engagement remained consistent, with repeat buyers accounting for roughly 47–49% of active buyers throughout this period [Etsy, Inc., 2026a]; (3) buyer surveys found that the importance Etsy shoppers place on product descriptions has seen little change, with upwards of 90% saying that they were very or somewhat important [eRank, 2023, 2025].

We then directly estimate the change in human-usable information with respondents from Prolific, on the same listing sample. Respondents evaluated balanced pre/post listing batches and self-elicited the probability that a human predictor with access to the listing could recover their chosen action (Appendix H). We use self-elicitation because the $\mathcal{V}$-entropy is an infimum over achievable entropies, and any individual will have more context about themselves than a third-party predictor, especially since privacy concerns limit how much context we can disclose to a third-party predictor. Among valid respondents, post-ChatGPT listings have 0.015 *fewer* bits of human-usable information than pre-ChatGPT listings, with a 95% confidence interval of $[-0.044, 0.014]$. Under the stable marginal-action assumption described in Appendix H, this is consistent with negligible human-side degradation, albeit not dispositive. Thus the post-ChatGPT shift is not just a general improvement in listing quality for both humans and machines; it is machine-skewed, while any human-side degradation appears bounded.

Machine-Usable To estimate the machine-usable information that X contains about Z_M, we need to find the (conditional) $\mathcal{V}$-entropy-minimizing $f \in \mathcal{V}$. When $\mathcal{V}$ is a neural network, as in our case, this is done in the literature by training a model to predict Z_M with X and without X to get the content model g' and null model g respectively. However, since we cannot train ChatGPT or even a GPT-5 proxy, we use Llama-3.1-8B-Instruct. We verify that the choice of fine-tuning model does not drive our conclusions through robustness checks on models from distinct training lineages (Figure 3).

We then estimate the PVI of each example (x, y) in our held-out data using (3). To determine whether the release of ChatGPT was followed by a rise of systematic mecha-nudging, we run a simple OLS regression of PVI on a binary post-ChatGPT indicator:

$$\text{PVI}_i \;=\; \alpha + \beta\,\text{after}_i \;+\; \varepsilon_i, \tag{8}$$

where $\text{after}_i = 1$ for listings created after the release of ChatGPT (November 30, 2022) and 0 for those created before. The coefficient β captures the average difference in PVI between the two periods; β being positive and statistically significant would be evidence of mecha-nudging.

We then run several other regressions to develop a more nuanced picture of any mecha-nudging. First, we replace the binary indicator with half-year dummies to trace how PVI evolves over time. Second, we control for possible confounders such as price, log number of shop and item reviews, average rating, and a discount indicator. Third, we model interaction effects between the post-ChatGPT indicator and the product categories ($\text{after}_i \times \text{category}_i$) to assess how mecha-nudging has diffused through the market. Finally, we include price directly in the labeling prompt, so the model observes both text and price when issuing its decision. Note that our estimate of β should be interpreted as a conditional difference in means, not as a causal effect.

4.3 Results

Machine-usable information rises significantly after the release of ChatGPT. Etsy listings created following the release of ChatGPT contain significantly more machine-usable information than those made before, with an estimated increase of 0.143 bits ($p < 0.01$) out of a maximum possible increase of 0.355. This ceiling was calculated by taking 1 bit (the maximum usable information for a

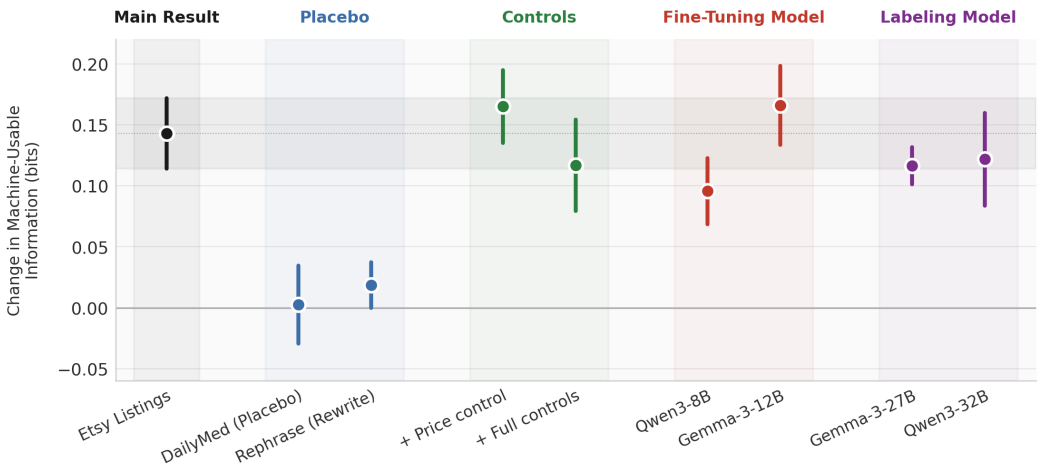


Figure 3: The increase in machine-usable information post-ChatGPT is robust to possible confounders: a generic temporal change (DailyMed), AI-written text (Rephrase), controls for product- and seller-specific attributes (green), the model family that is fine-tuned to estimate PVI (red), and the LLM used to generate training labels (purple), among others (Appendix C, D). Unless otherwise specified, we use GPT-5-mini as the labeling model and Llama-3.1-8B-Instruct as the fine-tuning model. Each point reports the OLS estimate of the post-ChatGPT shift in PVI, with 95% CI.

class-balanced binary decision, by construction) and subtracting from it the amount of machine-usable information in the pre-ChatGPT period, which was 0.645 bits (Appendix B). Note that the latter is non-zero because even non-mecha-nudged text, through disclosing features like product quality, can be predictive of agentic curation.

The temporal dynamics are also informative. As seen in Figure 1, compared to the Jul-Sep 2022 reference period, previous half-years did not contain any more machine-usable information, with half-year coefficients fluctuating around zero. Immediately after the release, machine-usable information sees a sharp and statistically significant jump in the subsequent half-year. The quickness of this response is not surprising: officials in several Western countries reported a jump in submission volumes for various government services soon after the release of ChatGPT [Schmitz et al., 2026]. The amount of machine-usable information then steadily declines, which—among other possible mechanisms—is consistent with sellers realizing that their efforts have no short-term reward as LLMs are only pulling from historical training data, not live listings. However, upon the release of ChatGPT Search in Oct 2024, which can browse live listings, PVI starts steadily climbing again, reaching its peak in the most recent half-year in our data (Jan-Jun 2025).

This result is robust to a wide range of possible confounders. As seen in Figure 3, varying token pairs, prompt formulations, labeling models, and fine-tuning models all yield positive and significant estimates of the change in usable information, with most point estimates of β falling within the range $[0.09, 0.17]$. Adding all the listing-level controls depresses the coefficient from 0.143 to 0.117, yet it remains significant ($p < 0.01$)—reassuring given that some controls, such as review counts, are themselves equilibrium outcomes (see Appendix D for details).

The effect is not replicated with LLM-rewritten listings. To test whether AI-assisted copywriting might explain these results, we consider two placebos. First, we rephrase Etsy listings predating the ChatGPT release (i.e., the start of widespread LLM usage by the public) using GPT-5-mini, preserving all factual content while allowing the model to alter wording and style. The estimated increase in PVI is only 0.019, an order of magnitude below the baseline effect: mechanical rephrasing does not replicate the information gains that mecha-nudging achieved. We also estimate how much Etsy text is machine-written to begin with. A widely-used commercial AI-text detector flags $< 4\%$ of post-ChatGPT listings as majority AI-written, and the rate barely varies with PVI (Appendix I). Second, we replicate our analysis on pharmaceutical drug descriptions from DailyMed, a dataset where text is written according to standardized templates, making it much more difficult for ChatGPT adoption to

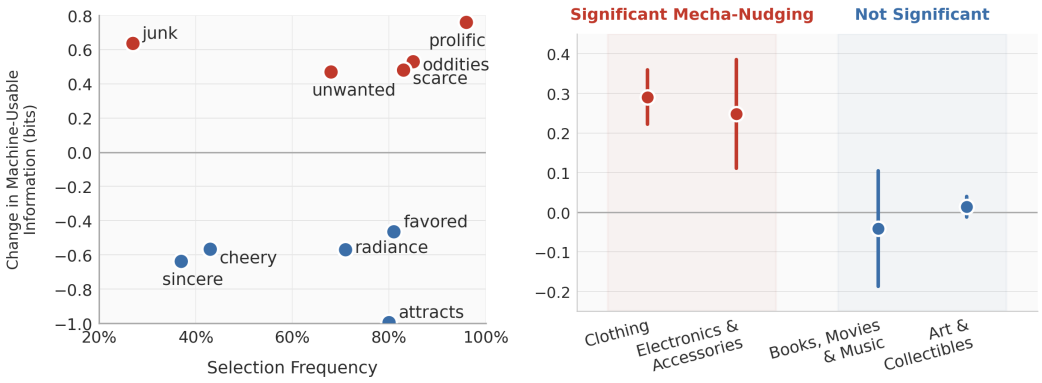


Figure 4: **Left:** Words with the largest impact on how much machine-usable information Etsy listings have (see Table 14 for a full list). A positive Δ PVI means that the word, on average, makes the machine behave *more* predictably; negative Δ PVI means that, on average, it makes the machine behave *less* predictably. **Right:** Etsy product categories where human buyers are ostensibly sensitive to AI usage (e.g., art & collectibles) show no mecha-nudging (based here on Gemma-3-27B-IT labels). For generic consumer goods like electronics, the effect size is well above average.

alter the content (Appendix B). The estimated shift in PVI post-ChatGPT is indistinguishable from zero, ruling out a generic temporal trend (Figure 3). Moreover, the variation in effect size across product categories also suggests selective adaptation: in areas where buyers are sensitive to AI use (e.g., books, music, art, and collectibles) [Castelo et al., 2019], the effect size is indistinguishable from zero; for generic consumer goods like electronics, it is above average (Figure 4, right).

Mecha-nudging is associated with greater commercial success. Post-ChatGPT listings with more machine-usable information are also more commercially successful, as measured by review counts, a noisy proxy for cumulative product sales. This relationship is concentrated among listings the agent would select, and is absent or reversed among pre-ChatGPT listings, consistent with machine-usable information becoming economically valuable after the rise of AI agents but not before (Appendix F). For the agent-selected listings created pre-ChatGPT, an additional bit of machine-usable information is associated with 25.2% *fewer* reviews. Given that LLMs were not widely used by consumers during this time, it could simply mean that anything that increases machine-usability within the pre-period did so in a way that was clearly unappealing to human buyers, with the issue being rectified in the post-ChatGPT period. Indeed, among agent-selected listings created in the post-period, an additional bit of machine-usable information is associated with 81.9% *more* reviews. Although this magnitude seems extraordinarily large, recall that by construction, the maximum PVI for a listing is 1 bit. Moreover, the magnitude of the increase diminishes to 15.5% after controlling for seller fixed-effects and listing age, although this can also be an under-estimate given that agent activity—and seller awareness of agent activity—grows over time.

Token-level patterns offer a partial window into the mechanism. Although we do not have the data needed for a mechanistic understanding of mecha-nudges, we can identify coarse token-level patterns. For each word in three sentiment and opinion lexicons [Hu and Liu, 2004, Hutto and Gilbert, 2014, Fast et al., 2016], we compute the average change in PVI when it is omitted from X. This Δ PVI metric is an accepted means of finding token-level signals in text data [Ethayarajh et al., 2022]. A highly positive Δ PVI means that the word, on average, makes the machine behave *more* predictably, while a highly negative Δ PVI means that it makes the machine behave *less* predictably. Some high Δ PVI words disambiguate the model's curation decision by highlighting rarity (e.g., *scarce*), while others make rejection more predictable by signaling low market value (e.g., *junk*) (Figure 4, left). In contrast, many low ΔPVI words carry positive affect, suggesting that positive wording can add ambiguity and make an agent behave less predictably. This is a subset of a larger list of diverse words (Appendix G, Table 14), and these findings should be read as descriptive rather than explanatory.

These patterns do not point to a single clean intervention of the kind from behavioral economics; instead, they suggest that the mechanism of mecha-nudging on Etsy is complex, likely reflecting

an emergent process of trial-and-error. Moreover, not all sellers might be intentionally trying to mecha-nudge AI agents; others might be imitating successful sellers or simply treating LLMs as an imperfect proxy for humans—which still counts as realized mecha-nudging (Definition 3.3).

5 Limitations & Future Work

The main limitation of our work is that our data is observational: we cannot draw a causal conclusion that the rise of ChatGPT is what precipitated the rise of mecha-nudging, and that this was definitively done to steer agent behavior. Rather, our claim is distributional: after the emergence of widely used LLM agents, Etsy listings contain more machine-usable information about agentic curation decisions, and this increase is robust across labeling models, prompts, token choices, and fine-tuning architectures. This is the empirical signature predicted by realized mecha-nudging, whether it arises through deliberate optimization, imitation, or some other means.

We suspect that the scale of mecha-nudging is far larger and more varied than most realize. Discovering more mecha-nudges is needed, especially to preempt the steering of agents towards malicious ends. Moreover, as seen in the OpenAI / Hugging Face incident, agents themselves are increasingly capable of steering other agents over which they do not have direct control [Greenblatt et al., 2026]. Agent-to-agent mecha-nudging is likely to become an important lever for controlling swarms in plain sight, meaning that discovering those mecha-nudges will become an important lever of scalable oversight for humans [Bowman et al., 2022]. Conversely, humans—as individuals, firms, and institutions—may want to adopt mecha-nudging as a defensive tool for steering agent behavior, which will require designing optimal interventions for (5).

Although our work suggests that market pressures can reconcile human and AI needs, as agents conduct an increasing amount of online activity, the agent experience may be prioritized over the human experience, leading to the gradual disempowerment of humans [Kulveit et al., 2025]. Monitoring the rise of mecha-nudges is not only imperative but surprisingly tractable; all one needs is an object through which mecha-nudging might plausibly occur (e.g., a product listing on Etsy), a closed- or open-weight labeling model whose decisions can act a proxy, and an open-weight model that can be fine-tuned to estimate machine-usable information. Moreover, although our work focuses on text data, usable information can in principle be measured in any modality, including images [Lu et al., 2026].

6 Conclusion

We introduced the concept of *mecha-nudges*, changes to the decision environment that systematically influence the behavior of AI agents while not materially degrading the decision environment for humans, and formalized them using $\mathcal{V}$-usable information. Using Etsy as a case study, we found large-scale observational evidence consistent with realized mecha-nudging after ChatGPT's release. As agents make decisions on their own, the environments they inhabit become optimized for them. This means that AI governance must not only concern models, but also the model–environment interface: not only how agents behave on fixed inputs, but how the world changes once agentic decisions are made. Mecha-nudges provide a unit of analysis for understanding this adaptation.

Acknowledgments

This project was funded by the University of Chicago Booth School of Business. The Etsy data used for the analysis was generously provided by Bright Data at no cost. As part of the terms of use, we are not allowed to release or redistribute the dataset publicly. Readers who would like to acquire the Etsy data should contact Bright Data directly.

References

Hunt Allcott. Social norms and energy conservation. *Journal of Public Economics*, 95(9-10): 1082–1095, October 2011. ISSN 00472727. doi: 10.1016/j.jpubeco.2011.03.003.

Itai Arieli and Yakov Babichenko. Private Bayesian persuasion. *Journal of Economic Theory*, 182: 185–217, July 2019. ISSN 00220531. doi: 10.1016/j.jet.2019.04.008.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and Consumer Choice. *Journal of Political Economy*, 121(5):803–843, October 2013. ISSN 0022-3808, 1537-534X. doi: 10.1086/673885.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. *Journal of marketing research*, 56(5):809–825, 2019.

Manuel Cherep, Pattie Maes, and Nikhil Singh. LLM agents are hypersensitive to nudges. *arXiv preprint arXiv:2505.11584*, 2025. Revise and resubmit at PNAS.

Manuel Cherep, Chengtian Ma, Abigail Xu, Maya Shaked, Pattie Maes, and Nikhil Singh. A framework for studying AI agent behavior: Evidence from consumer choice experiments. In *International Conference on Learning Representations (ICLR)*, 2026.

Raj Chetty, John N. Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen. Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark *. *The Quarterly Journal of Economics*, 129(3):1141–1219, August 2014. ISSN 0033-5533. doi: 10.1093/qje/qju013.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

eRank. Etsy buying habits – 2023. `https://help.erank.com/blog/etsy-buying-habits/`, 2023. Survey of ∼1,000 recent U.S. Etsy buyers.

eRank. 2024 etsy buyer survey: What sellers need to know. `https://help.erank.com/blog/2024-etsy-buyer-survey/`, 2025. Survey of 1,000 recent U.S. Etsy buyers; published February 2025.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR, 2022.

Etsy, Inc. Form 10-k for the fiscal year ended december 31, 2025. Securities and Exchange Commission (SEC), February 2026a. URL `https://www.sec.gov/Archives/edgar/data/1370637/000137063726000019/etsy-20251231.htm`.

Etsy, Inc. Etsy, inc. reports fourth quarter and full year 2025 results. Etsy Investor Relations Press Release, February 2026b. URL `https://investors.etsy.com/news-events/press-releases/detail/218/etsy-inc-reports-fourth-quarter-and\-full-year-2025-results`. Data for historical years derived from respective annual Q4 earnings reports (2019–2025).

Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657, 2016.

X. Gabaix and D. Laibson. Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets. *The Quarterly Journal of Economics*, 121(2):505–540, May 2006. ISSN 0033-5533, 1531-4650. doi: 10.1162/qjec.2006.121.2.505.

Matthew Gentzkow and Emir Kamenica. Costly Persuasion. *American Economic Review*, 104(5):457–462, May 2014. ISSN 0002-8282. doi: 10.1257/aer.104.5.457.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. Brief independent investigation of agents' behavior, reasoning and collaboration in the openai / hugging face hacking incident, August 2026. URL https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

Thilo Hagendorff. Linking human and machine behavior: A new approach to evaluate training data quality for beneficial machine learning. *Minds and Machines*, 31:563–593, 2021. doi: 10.1007/s11023-021-09573-8.

Nika Haghtalab, Mingda Qiao, and Kunhe Yang. Leakage-Robust Bayesian Persuasion, November 2024.

Justin E Holz, John A List, Alejandro Zentner, Marvin Cardoza, and Joaquin E Zentner. The $100 million nudge: Increasing tax compliance of firms using a natural field experiment. *Journal of Public Economics*, 218:104779, 2023.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

Clayton Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225, 2014.

Brian Jabarian and Alex Imas. Artificial writing and automated detection. Technical report, National Bureau of Economic Research, 2025.

Emir Kamenica. Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11(1):249–272, August 2019. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-080218-025739.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6): 2590–2615, October 2011. ISSN 0002-8282. doi: 10.1257/aer.101.6.2590.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development. *arXiv preprint arXiv:2501.16946*, 2025.

Changjie Lu, Sourya Sengupta, Hua Li, and Mark A Anastasio. Observer-usable information as a task-specific image quality metric. *IEEE Transactions on Medical Imaging*, 2026.

Brigitte C. Madrian and Dennis F. Shea. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. *The Quarterly Journal of Economics*, 116(4):1149–1187, 2001. ISSN 0033-5533.

Filip Matějka and Alisdair McKay. Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model. *American Economic Review*, 105(1):272–298, January 2015. ISSN 0002-8282. doi: 10.1257/aer.20130047.

Barbara J McNeil, Stephen G Pauker, Harold C Sox Jr, and Amos Tversky. On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306(21):1259–1262, 1982.

Kobe Millet, Florian Buehler, Guanzhong Du, and Michail D Kokkoris. Defending humankind: Anthropocentric bias in the appreciation of ai art. *Computers in Human Behavior*, 143:107707, 2023.

OpenAI. Buy it in ChatGPT: Instant checkout and the agentic commerce protocol, September 2025. URL https://openai.com/index/buy-it-in-chatgpt/. Blog post.

Reddit r/EtsyCommunity. What's the average percentage of customers that leave reviews? Online forum thread, 2025. URL https://www.reddit.com/r/EtsyCommunity/comments/1ndeic0/whats_the_average_percentage_of_customers_that/. Accessed 2026-06-14.

Reddit r/EtsySellers. What percent of buyers leave a review on your shop? just curious. Online forum thread, 2022. URL https://www.reddit.com/r/EtsySellers/comments/xhfkkh/what_percent_of_buyers_leave_a_review_on_your/. Accessed 2026-06-14.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR, 2023.

Chris Schmitz, Lewis Hammond, and Alan Chan. Characterizing agentic flooding of government services. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 9, pages 2266–2278, 2026.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, April 2003. ISSN 03043932. doi: 10.1016/S0304-3932(03)00029-1.

Allison Smith. Chatgpt is now 20% of walmart's referral traffic — while amazon wards off ai shopping agents. *Modern Retail*, September 2025. Data sourced from Similarweb.

Ina Taneva. Information design. *American Economic Journal: Microeconomics*, 11(4):151–185, 2019.

Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale university press, New Haven (Conn.), 2008. ISBN 978-0-300-12223-7.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Simon Willison. Prompt injection attacks against GPT-3. Simon Willison's Weblog, Sep 2022. URL https://simonwillison.net/2022/Sep/12/prompt-injection/. Accessed: 2026-03-01.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

A Proofs

Proposition 3.2 (Bounded-Receiver Bayesian Persuasion (restated)) Consider a bounded-receiver analog of Bayesian persuasion in which both the choice architect and decision-maker have log-scoring utility $\log_2(\cdot)$, and the decision-maker is restricted to predictive family $\mathcal{M}$. Then $\arg\max_{\tau\in\mathcal{T}} I_{\mathcal{M}}(\tau(X) \to Y_M)$, the solution to unconstrained mecha-nudging, also maximizes the best achievable expected utility for the decision-maker.

Proof. By definition,

$$I_{\mathcal{M}}(\tau(X) \to Y_M) = H_{\mathcal{M}}(Y_M) - H_{\mathcal{M}}(Y_M \mid \tau(X)) \tag{9}$$

The first term does not depend on τ, so maximizing $I_{\mathcal{M}}(\tau(X) \to Y_M)$ is equivalent to minimizing $H_{\mathcal{M}}(Y_M \mid \tau(X))$ over $\tau \in \mathcal{T}$.

From the definition of conditional $\mathcal{V}$-entropy,

$$H_{\mathcal{M}}(Y_M \mid \tau(X)) = \inf_{f\in\mathcal{M}} \mathbb{E}\big[-\log_2 f[\tau(X)](Y_M)\big] \tag{10}$$

so equivalently,

$$-H_{\mathcal{M}}(Y_M \mid \tau(X)) = \sup_{f\in\mathcal{M}} \mathbb{E}\big[\log_2 f[\tau(X)](Y_M)\big] \tag{11}$$

This is exactly the best expected log-score attainable by a decision-maker restricted to predictive family $\mathcal{M}$ after observing the transformed signal $\tau(X)$, which yields the stated equality.

If $\mathcal{M} = \Omega$, the decision-maker can represent the true posterior, so the optimal log-scoring action is $q(\cdot) = P(Y_M \mid \tau(X))$. The resulting value is

$$\mathbb{E}\big[\log_2 P(Y_M \mid \tau(X))\big] = -H(Y_M \mid \tau(X)) \tag{12}$$

and maximizing this is equivalent to maximizing

$$H(Y_M) - H(Y_M \mid \tau(X)) = I(\tau(X); Y_M) \tag{13}$$

which is the classical log-scoring Bayesian persuasion objective. □

B Data

The raw data was obtained from the company Bright Data, which provides structured datasets of Etsy product listings. Our extract was collected on November 12, 2025 and delivered the same day. It comes in two snapshots: one of post-ChatGPT listings, and one of pre-ChatGPT listings. We filter both on `liisted_date`, the date a listing was created. The repeated `i` is the vendor's spelling. From the first snapshot we keep listings created between November 30, 2022 and August 1, 2025, which form the post-ChatGPT period. From the second we keep listings created strictly before November 30, 2022, which form the pre-ChatGPT period. After filtering, the dataset contains approximately 1.06M pre-ChatGPT and 5.0M post-ChatGPT listings.

From this full-sized dataset, we sample uniformly at random to construct two working datasets: a medium dataset of roughly 500K listings from each period, and a small dataset of roughly 100K from each period. The latter is used for robustness checks and ablations that would otherwise be too expensive to run. Within each dataset, listings are randomly split into training (80%), validation (10%), and test (10%) subsets. For each period and experimental configuration, we train separate content and null models on that period's training split and compute PVI only on the held-out test split for the same period. Table 1 reports summary statistics for both periods.

As a placebo control, we use pharmaceutical drug labels from DailyMed, a public database maintained by the U.S. National Library of Medicine. The labels are submitted by manufacturers to the Food and Drug Administration (FDA), according to templates provided by the FDA, and kept up to date.[5] Unlike Etsy listings, drug labels are thus not primarily shaped by market incentives, since they are written to conform to legal and medical standards.

[5] See DailyMed, National Library of Medicine.

Table 1: Summary Statistics by Period (GPT-5-mini Labels)

	Pre-ChatGPT	Post-ChatGPT
PVI	0.645 (0.868)	0.788 (0.872)
$\hat{Y}$ (Frequency of SELECT)	0.479	0.509
Word Count	123 [2, 2211]	245 [5, 3977]
Price ($)	34.50 [0.20, 22100.00]	46.00 [0.20, 46500.00]
Item Reviews	0 [0, 56]	0 [0, 14068]
Shop Reviews	1525 [0, 42769]	667 [0, 300000]
Rating	4.94 (0.16)	4.80 (0.64)

Notes: Labeling model: GPT-5-mini; fine-tuning model: Llama-3.1-8B-Instruct. $\hat{Y}$ is the fine-tuned model's predicted label, whose SELECT frequency is close to 0.50 because of class-balancing (Appendix C). Mean with SD in parentheses for PVI and Rating. Median with range in brackets for Word Count, Price, Item Reviews, and Shop Reviews, since they all have positive skew. Frequency variables report the sample mean.

Because the text is so tightly regulated, manufacturers have much less latitude to adapt it in response to the ChatGPT shock compared to Etsy sellers. If the rise in PVI were caused by some broad change over time, or by an artifact in our estimation procedure, then we should observe a similar increase with the DailyMed dataset. As seen in Figure 3, we do not. This suggests that the Etsy result is specific to a setting where sellers can change how products are presented. The prompt and the regression are in Appendix D.

C Label Construction

Because we are interested in whether AI agents are being mecha-nudged by Etsy listings, we construct Z_M by prompting a proxy LLM to issue a binary decision for each listing based on its title and description (the exact wording of each prompt is given at the end of the section). We vary the prompt across specifications. Under oracle-style prompts, Z_M closely approximates a buy/not-buy judgment. Under our main Etsy-specific prompt (V4, Table 2), it is better understood as a selective recommendation or curation decision rather than raw purchase propensity. The results are robust to the exact pair of tokens used to express the select/pass decision (Table 3) and to the specific prompt (Table 4), though we use Prompt V4 in our main experiments because it is Etsy-specific and closer to how agents curate listings in the real world. Note that for cost reasons, the token and prompt ablations are done with Gemma-3-27B-IT as the labeling model.

We force the output tokens to be one of the two selected for a particular experiment. We do so in two different ways, based on the type of labeling model that we use. For the closed-source API model (e.g., GPT-5-mini), we instruct the API to return only those two tokens: every request carries a strict JSON schema that specifies a single field that is an enumeration over the two tokens. The API does not return token probabilities for this model, so we take the token it returns as the label. For the open-source models, we observe the generation probabilities. In this case we mask the logits of all tokens except for the two tokens of interest at each decoding step and decode greedily.

When we do not tell the labeling LLM how often it should choose each label, the balance between the two classes can vary considerably across prompts. This is a problem for fine-tuning. Say one class is much more common than the other—a model can then achieve low cross-entropy by assigning a high probability to that label for almost every listing, without learning much from the listing itself. Like a weighted coin, it learns which label is more common, but not the relationship between the listing content X and the label Z_M. In theory PVI accounts for this, since the null model and the content model are both exposed to the same marginal distribution of Z_M. In practice, severe imbalance degrades the quality of the fine-tuned content model and introduces noise into the usable information estimates.

Upweighting the minority class in the cross-entropy loss dramatically increases the effect sizes for all prompts (Table 5). However, as the models learned this way are not minimizing the (conditional) $\mathcal{V}$-entropy, the estimates would not technically be estimates of $\mathcal{V}$-usable information. Therefore we subsample the data to draw an equal number of positive and negative examples, making the marginal distribution of Z_M identical across periods. The effect size is smaller than with class-weighted

Name	Prompt
V1: Minimal Oracle	System: You are a purchase decision oracle. When asked whether to buy, you must return exactly one uppercase token from this set: {positive_token} or {negative_token}. Do not ask follow-up questions. Do not explain. If there is no information, make your best default choice. User: Product Details: {product_info} Consider this product on Etsy. Should the user buy it? Answer with exactly one token: {positive_token} or {negative_token}.
V2: Output-Format Control	System: You are a purchase decision oracle. When asked whether to buy, you must return exactly one uppercase token from this set: {positive_token} or {negative_token}. Do not ask follow-up questions. Do not explain. Output the token alone with no punctuation or additional words. User: Product Details: {product_info} Should the user respond with {positive_token}? Should the user respond with {negative_token}? Answer with exactly one token: {positive_token} or {negative_token}.
V3: Recommendation Framing	System: You are a purchase decision oracle. When asked whether to buy, you must return exactly one uppercase token from this set: {positive_token} or {negative_token}. Do not ask follow-up questions. Do not explain. Output the token alone with no punctuation or additional words. User: Product Details: {product_info} Should I suggest the user respond with {positive_token} or {negative_token}? Answer with exactly one token: {positive_token} or {negative_token}.
V4: Selective Curator	System: You are helping someone browse Etsy. Only {positive_token} items they would genuinely appreciate and want to see - be very selective. {positive_token} only if the item is EXCEPTIONAL and would make someone say "wow, that's special": • Truly beautiful, impressive, or emotionally resonant • Exceptional craftsmanship, artistic merit, or historical significance • Something you'd be excited to own, gift, or show others • Stands out as memorable among thousands of items {negative_token} for everything else, including: • Ordinary vintage items without special appeal • Generic handmade items lacking wow factor • Mass-produced or common items • Anything that's just "okay" or "fine" but not exciting • Items where you'd scroll past without a second thought Be highly selective - most items should be {negative_token}. Only {positive_token} items that truly deserve attention and would be genuinely appreciated. Output only: {positive_token} or {negative_token}. User: Product Details: {product_info} Decision:

Table 2: Prompt variants used for eliciting agent curation decisions on Etsy. All variants were constructed with the help of LLMs.

Table 3: Token Pair Variations

	Dependent variable: PVI			
	SELECT/PASS	YES/NO	BUY/SKIP	PICK/PASS
after	0.120***	0.123***	0.100***	0.126***
	(0.019)	(0.007)	(0.024)	(0.019)
Observations	12662	58315	9610	11189
Residual Std. Error	1.017	0.822	1.117	0.952
F Statistic	41.477***	299.264***	16.622***	44.916***
Note:				*p<0.1; **p<0.05; ***p<0.01

[a] Comparison of post-ChatGPT differences in PVI across different token pairs (all using Prompt V4).
[b] Fine-tuning model: Llama-3.1-8B-Instruct. Labeling model: Gemma-3-27B-IT.
[c] All specifications enforce a balanced (50/50) class distribution via sub-sampling. Because the choice of token pair yields very different class distributions, the number of observations differs across specifications. *Observations* specifically refers to the number of test examples used to estimate the effects after training the content and null models.

Table 4: Prompt Variations

	Dependent variable: PVI			
	V1	V2	V3	V4
after	0.110***	0.092***	0.078**	0.120***
	(0.012)	(0.018)	(0.036)	(0.019)
Observations	15296	11078	5907	12662
Residual Std. Error	0.763	0.885	1.294	1.017
F Statistic	77.302***	26.701***	4.877**	41.477***
Note:				*p<0.1; **p<0.05; ***p<0.01

[a] Each column shows OLS regression of PVI on time period (after vs. before).
[b] Fine-tuning model: Llama-3.1-8B-Instruct. Labeling model: Gemma-3-27B-IT.
[c] V1-V4 represent different prompt formulations. V4 is the prompt used in the main specification.
[d] All specifications enforce a balanced (50/50) class distribution via sub-sampling. Because the choice of prompt yields very different class distributions, the number of observations differs across specifications. *Observations* specifically refers to the number of test examples used to estimate the effects after training the content and null models.

Table 5: Class-weighted Training

	Dependent variable: PVI			
	V1	V2	V3	V4
after	0.207***	0.253***	0.204***	0.196***
	(0.013)	(0.017)	(0.011)	(0.013)
Observations	20000	20000	20000	20000
Residual Std. Error	0.923	1.182	0.743	0.912
F Statistic	251.771***	230.083***	375.314***	230.017***
Note:				*p<0.1; **p<0.05; ***p<0.01

[a] Each column shows OLS regression of PVI on time period (after vs. before).
[b] V1-V4 represent different prompt formulations. Class weights + stratified split.

Table 6: Labeling Model

	Dependent variable: PVI		
	GPT-5-mini	Gemma-3-27B-IT	Qwen3-32B
after	0.143***	0.117***	0.122***
	(0.015)	(0.008)	(0.019)
Observations	19898	51033	16429
Residual Std. Error	0.869	0.840	1.119
F Statistic	94.394***	226.662***	39.169***
Note:			*p<0.1; **p<0.05; ***p<0.01

[a] Comparison of post-ChatGPT differences in PVI across different labeling models.
[b] Column (1) uses GPT-5-mini as the baseline labeling model; columns (2)–(3) are robustness checks. All other settings are shared with the baseline specification.

training, but the procedure is more faithful to our framework, as the model still minimizes unweighted cross-entropy rather than some weighted alternative. In the medium dataset, class-balancing reduces each period's 500,000 listings to about 154,000 pre-ChatGPT and about 45,000 post-ChatGPT (with SELECT/PASS labels from GPT-5-mini). We run all regressions on a 10% held-out split of each dataset, so the models are never evaluated on listings they saw during training. Gemma-3-27B-IT produces less imbalanced labels, so its YES/NO configuration leaves roughly 58,000 test observations (Table 3).

For the labeling phase, we use GPT-5-mini as a proxy for the basic version of ChatGPT that is used by most consumers. As robustness checks, we also generate labels with two open-weight models—Gemma-3-27B-IT and Qwen3-32B—both of which are small enough to enable large-batch inference in an academic setting while remaining competitive with much larger models on instruction-following benchmarks. Gemma-3-27B-IT was released by Google, making it a better proxy for the various LLM-based products in the Google ecosystem, such as Gemini and AI Overviews. Qwen3-32B, released by Alibaba, is the only non-American model of the three. Using models from distinct lineages (OpenAI, Google, and Alibaba) allows us to study whether the increase in machine-usable information is indeed robust, and not an artifact of a particular model's biases (Table 6).

D Controls

Not only is our main result robust to prompt variation, the selection tokens, and the labeling model (Appendix C), it is also robust to the fine-tuning model family (Table 7).

Table 7: Fine-tuning Model

	Dependent variable: PVI		
	Llama-3.1-8B-Instruct (Baseline)	Qwen3-8B	Gemma-3-12B
after	0.143***	0.096***	0.166***
	(0.015)	(0.014)	(0.017)
Observations	19898	19898	19898
Residual Std. Error	0.869	0.812	0.976
F Statistic	94.394***	48.278***	100.911***
Note:			*p<0.1; **p<0.05; ***p<0.01

[a] Comparison of post-ChatGPT differences in PVI across different fine-tuning models.
[b] All other settings are shared with the baseline specification.

We assess the relevance of our modeling framework with four estimation specifications of increasing richness. The baseline is a simple OLS regression of PVI on a binary post-ChatGPT indicator:

$$\text{PVI}_i \;=\; \alpha + \beta\,\text{after}_i \;+\; \varepsilon_i, \tag{14}$$

Table 8: Half-Yearly PVI Coefficients

Half-Year	Coefficient	95% CI	N
2019-H1	-0.0179	$[-0.0693, 0.0335]$	2,442
2019-H2	-0.0410	$[-0.0900, 0.0080]$	2,895
2020-H1	0.0051	$[-0.0396, 0.0497]$	4,242
2020-H2	-0.0473^{**}	$[-0.0900, -0.0046]$	5,245
2021-H1	-0.0140	$[-0.0571, 0.0291]$	4,999
2021-H2	-0.0145	$[-0.0558, 0.0269]$	6,249
2022-H1	0.0069	$[-0.0349, 0.0487]$	5,864
Jul-Sep 2022	Ref.		3,374
2023-H1	0.1100^{***}	[0.0656, 0.1543]	4,349
2023-H2	0.0758^{***}	[0.0274, 0.1241]	3,051
2024-H1	0.0537^{**}	[0.0097, 0.0977]	4,536
2024-H2	0.0877^{***}	[0.0406, 0.1348]	3,371
2025-H1	0.1203^{***}	[0.0782, 0.1624]	5,641

[a] Each half-year has an independently trained pipeline. Coefficients from pooled OLS with half-year fixed effects, Jul-Sep 2022 as reference.
[b] $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

where after_i equals one for listings uploaded after the release of ChatGPT (November 30, 2022) and zero for those uploaded before. The coefficient β captures the average difference in PVI between the two periods.

To track the change over time, we then replace the indicator variable with a full set of half-year dummy variables, where July–September 2022 is the reference period.

$$\text{PVI}_i \;=\; \alpha + \sum_{t \neq t_0} \delta_t \, \mathbf{1}[\text{period}_i = t] \;+\; \varepsilon_i, \tag{15}$$

where t indexes half-year periods (H1/H2 for each year from 2019 to 2025) and t_0 denotes the reference period. Each coefficient δ_t measures the average change in PVI in period t relative to the pre-ChatGPT baseline. The results are provided in Table 8.

We also estimate this specification augmented with listing-level controls—price, log number of shop and item reviews, average rating, a discount indicator, and the log word count:

$$\text{PVI}_i \;=\; \alpha + \sum_{t \neq t_0} \delta_t \, \mathbf{1}[\text{period}_i = t] \;+\; \mathbf{X}_i' \boldsymbol{\gamma} \;+\; \varepsilon_i, \tag{16}$$

where $\mathbf{X}_i$ collects the listing-level controls. This helps assess whether the PVI increase is driven by changes in listing characteristics (e.g., goods getting cheaper) instead of how the listing is presented. Results are in Table 9: a significant effect still persists. Because review counts are long-tailed and many items have no reviews at all, we apply a $\log(1 + x)$ transformation of the review count. Repeating the regression with raw (untransformed) review counts yields a nearly identical estimate of the effect size (0.120 vs. 0.117).

We also verify that heteroskedasticity does not distort our results. Across all HC corrections (HC0–HC3), the robust standard errors are almost identical to the classical OLS estimates (0.01472 vs. 0.01475 for SELECT/PASS; 0.00714 vs. 0.00698 for YES/NO). This is expected given that our baseline specification is essentially a difference in means.

A third regression adds category-level interactions to the baseline regression, allowing the post-ChatGPT shift in PVI to vary across product categories:

$$\text{PVI}_i \;=\; \alpha + \beta \, \text{after}_i \;+\; \sum_{c} \gamma_c \left(\text{after}_i \times \mathbf{1}[\text{category}_i = c] \right) \;+\; \varepsilon_i, \tag{17}$$

where c indexes the category. Coefficients γ_c capture the category-specific shift in PVI. When using labels from Gemma-3-27B-IT, we find that categories where human buyers are ostensibly sensitive to AI use do not have significant effects (Table 10); with a couple exceptions, the remaining

Table 9: Robustness to Controls

	Dependent variable: PVI				
	Baseline	+Price (OLS)	+Full Controls	+Word Count	Price in Prompt
after	0.143***	0.165***	0.117***	0.122***	0.103***
	(0.015)	(0.015)	(0.019)	(0.019)	(0.020)
Log(Price)		-0.038***	-0.035***	-0.034***	
		(0.005)	(0.005)	(0.005)	
Log(Shop Reviews)			0.005	0.005	
			(0.005)	(0.005)	
Log(Item Reviews)			0.008	0.009	
			(0.007)	(0.007)	
Rating			-0.026	-0.026	
			(0.019)	(0.019)	
Has Discount			0.116***	0.121***	
			(0.025)	(0.025)	
Log(Word Count)				-0.012	
				(0.008)	
Observations	19898	17348	17343	17343	11953
Residual Std. Error	0.869	0.864	0.863	0.863	1.040
F Statistic	94.394***	74.155***	29.725***	25.785***	26.964***
Note:					*p<0.1; **p<0.05; ***p<0.01

[a] Robustness of the post-ChatGPT difference in PVI to alternative control specifications.
[b] Column (1): Baseline OLS (SELECT/PASS, Prompt V4, class-balanced).
[c] Column (2): Adds log listing price as an OLS control variable.
[d] Column (3): Adds full controls: log price, log shop reviews, log item reviews, rating, discount indicator.
[e] Column (4): Adds log listing word count.
[f] Column (5): Price information included in the labeling prompt.

Table 10: Post-ChatGPT Difference in PVI by Product Category

Category	Coefficient	SE	N
Pet Supplies	0.3775**	(0.1478)	626
Clothing	0.2904***	(0.0349)	2,305
Electronics & Accessories	0.2479***	(0.0699)	633
Bags & Purses	0.2321***	(0.0606)	854
Shoes	0.2242*	(0.1215)	193
Accessories	0.1746***	(0.0439)	1,595
Toys & Games	0.1678***	(0.0372)	2,035
Jewelry	0.1652***	(0.0234)	5,344
Paper & Party Supplies	0.1360***	(0.0518)	1,171
Craft Supplies & Tools	0.1344***	(0.0510)	2,831
Bath & Beauty	0.1155	(0.0834)	669
Home & Living	0.1146***	(0.0177)	13,460
Weddings	0.0551	(0.0636)	933
Books, Movies & Music	−0.0411	(0.0742)	1,665
Art & Collectibles	0.0140	(0.0130)	16,718

[a] Each row reports the total post-ChatGPT difference in PVI for that category from a single pooled OLS regression with category × after interactions (PVI ∼ after × category). For the reference category (Accessories) the effect size is the coefficient on "after" in (17); for all other categories it is "after + after × category" with SE from the covariance matrix.
[b] Fine-tuning model: Llama-3.1-8B-Instruct. Labeling model: Gemma-3-27B-IT. The latter allows us to label more listings cost-effectively, which is especially helpful when doing category-specific regressions.
[c] $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Name	Prompt
DailyMed: Prescription Oracle	System: You are a clinical prescribing oracle. When asked whether to prescribe a drug, you must return exactly one uppercase token: {positive_token} or {negative_token}. Do not ask follow-up questions. Do not explain. Output the token alone with no punctuation or additional words. {positive_token} only if the drug has clear, specific indications that make it a worthwhile prescribing option: • Has a well-defined therapeutic use for a real, identifiable condition or symptom • Indication text is substantive and informative (not boilerplate or empty) • A general practitioner could reasonably consider prescribing this drug based on the information provided • The drug name is identifiable (not an FDA formatting header) {negative_token} if any of the following apply: • The title is FDA boilerplate text (e.g., "These highlights do not include...") • The indication text is missing, vague, or too brief to guide a prescribing decision • The drug appears highly specialized or rarely prescribed outside a narrow subspecialty • The entry appears incomplete or is a placeholder Be selective -- not every approved drug is a routine prescribing choice. Output only: {positive_token} or {negative_token}. Drug Label: Title: {title} Indications: {indications} Decision:
Rephrasing: LLM Listing Optimizer	System: You are a text rephrasing assistant. Your task is to rephrase the given text according to the instruction. Output only the rephrased text without any additional explanation, preamble, labels, or quotes. For context, here are the other fields for this product: {other_fields} Field to rephrase: {target_column} Original text: {target_text} Instruction: *[column-specific, e.g., "You are an expert Etsy listing optimizer. Rewrite this Etsy title to increase clicks and sales while staying truthful to the original product. Make it compelling, use relevant keywords, and keep it concise."]* Rephrased text:

Table 11: Prompts for the DailyMed prescribing task and for LLM-based listing rephrasing. All variants were constructed with the help of LLMs.

categories do. Repeating the analysis with GPT-5-mini labels yields similar broad trends, although the effect sizes are weakly positive. As an additional ablation, we re-estimate the baseline effect after dropping each top-level category; the post-ChatGPT shift remains positive in all leave-one-out ablations, meaning that the pooled result is not driven by any one category.

To control for a more generic trend in the data, we construct labels Z_M for the DailyMed dataset using the prompt in Table 11, before running the same baseline experiment we run with the Etsy data. To control for the effect of AI-assisted writing—based on the premise that AI agents may have a greater proclivity toward AI-written text—we take the pre-ChatGPT data, rephrase it using GPT-5-mini, and re-run the OLS regression (where, abusing notation, we use the `after` indicator to denote *after rephrasing*). As seen in Table 12, there is no significant effect for DailyMed (e.g., no generic temporal trend in the data), and the effect from rephrasing is an order of magnitude weaker than our main result. Appendix I asks the complementary question of how much Etsy text is machine-written to begin with.

Table 12: Baseline and Placebo Tests

	Dependent variable: PVI		
	Etsy (Baseline)	Rephrased Listings	Drug Labels
after	0.143***	0.019*	0.003
	(0.015)	(0.010)	(0.016)
Observations	19898	37574	8039
Residual Std. Error	0.869	0.913	0.677
F Statistic	94.394***	3.800*	0.024
Note:			*p<0.1; **p<0.05; ***p<0.01

[a] OLS regression results for the three main experiments.
[b] Column (1): Etsy product listings, SELECT/PASS tokens, balanced dataset (baseline).
[c] Column (2): LLM-rephrased pre-period listings vs. original pre-period listings. *after* denotes "after rephrasing" in this context.
[d] Column (3): DailyMed pharmaceutical drug labels from the pre- and post-periods, GPT-5-mini labeling, class-balanced dataset.

E Fine-tuning

Computing PVI requires both $g[\varnothing]$, trained without access to X, and $g'[x]$, trained with access to X—the difference in their log-likelihoods is the empirical estimate of instance-level usable information (Definition 2.2). We obtain these models via LoRA fine-tuning,[6] which updates only a small fraction of parameters and is well suited to this task: predicting a binary agent curation decision from listing text is a relatively simple classification problem that does not require modifying the full model. Our baseline is Llama-3.1-8B-Instruct, chosen for its strong instruction-following performance at a scale that makes fine-tuning on hundreds of thousands of listings computationally feasible. All fine-tuning was done on a single node with 8 H100 GPUs, each with 80GB of VRAM. We used BF16 precision, an AdamW optimizer (weight decay 0.01, gradient clipping at max-norm 1.0), a learning rate of 2×10^{-4}, a linear schedule with 3% warmup, and an effective batch size of 768. To avoid over-fitting, we train for three epochs and use the checkpoint with the lowest cross-entropy loss, as evaluated on the validation split.

For the main experiment, we fine-tune four models: a content model and a null model for each of the two periods (pre- and post-November 2022 ChatGPT release). The content models must predict the label Z_M from listing text X, while the null models must recover the marginal distribution of Z_M using only the null input " ". We re-run the pipeline across five random seeds (42, 123, 456, 789, 1024) to verify that the PVI estimates are not sensitive to training and sampling stochasticity, with no statistically significant differences in mean PVI or estimated effect size (one-way ANOVA, $p = 0.853$). For the half-year experiments, we fine-tune a separate null and content model for each half-year.

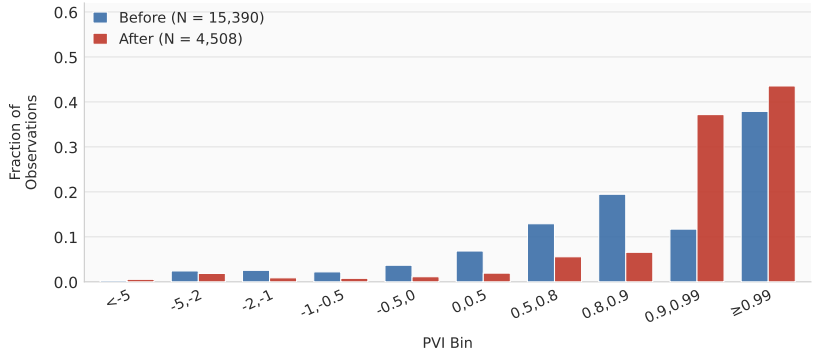
Figure 5 shows the distribution of PVI scores for pre- and post-ChatGPT listings. For all three labeling models, the pre-ChatGPT distribution is bi-modal or tri-modal, with a peak on the lower end and a peak in the rightmost bin as well; the post-ChatGPT distribution shifts rightward but remains bimodal. There are two noteworthy patterns: (1) for all models, the left tail is similar in size during both periods, suggesting that there was no optimization pressure on the least-informative listings; (2) in the post-period, the PVI distributions for the three labeling models resemble each other much more than in the pre-period, a homogenization that is consistent with optimization pressure.

F Economic Impact

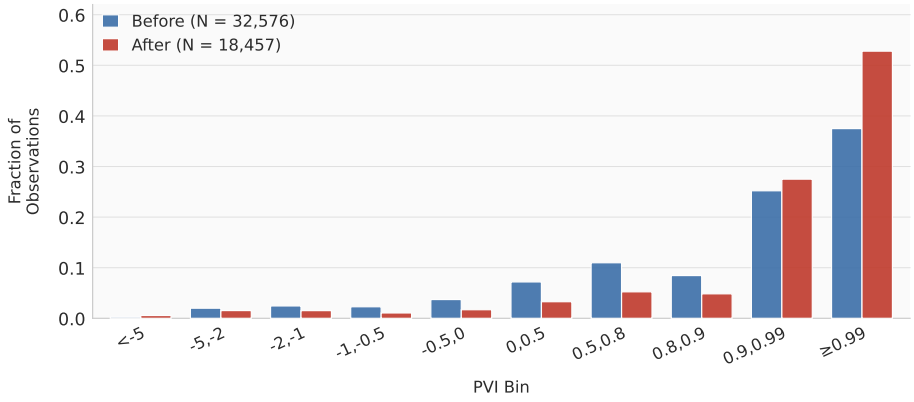
Are mecha-nudges having an economic impact? In the Etsy context, economic impact can be quantified in many different ways, including more sales, a higher willingness-to-pay for the same product, a lower return rate, etc. Unfortunately, our dataset does not contain such metrics. The one noisy cumulative proxy that *is* available in our dataset is the number of product reviews. Anecdotally, users on a forum for Etsy sellers report that 10–40% of purchases yield reviews, which are publicly

[6] LoRA hyperparameters: rank $r = 32$, $\alpha = 64$, dropout 0.05. LoRA is used for all attention and MLP layers.

(a) GPT-5-mini labels



(b) Gemma-3-27B-IT labels



(c) Qwen3-32B labels

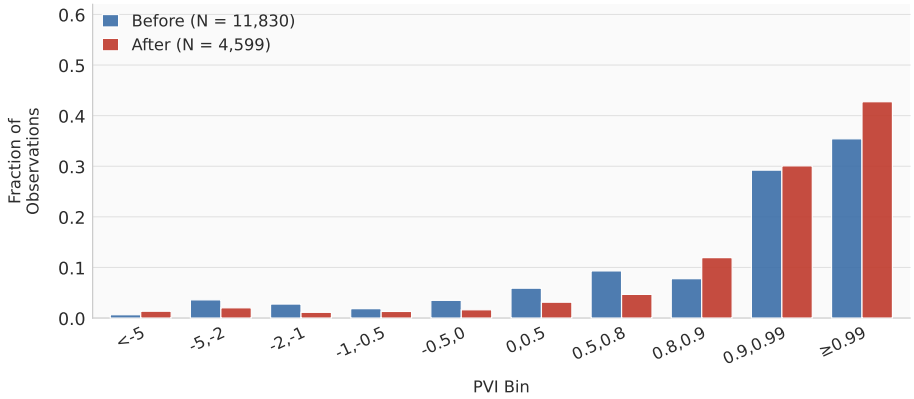


Figure 5: Distribution of PVI scores across listings uploaded before and after the ChatGPT release (November 30, 2022), using SELECT/PASS tokens with balanced sampling. Each panel corresponds to a different labeling model. The y-axis reports the fraction of observations in each bin.

visible on the listing [Reddit r/EtsySellers, 2022, Reddit r/EtsyCommunity, 2025]. If mecha-nudges help drive product sales, then all else held constant, product listings with more machine-usable information should have more reviews in the post-ChatGPT period (but not in the pre-ChatGPT period). This is indeed the case.

In Table 13 (Panel A), we run a PPML regression of per-listing review counts against PVI. In the pre-ChatGPT period, an additional bit of machine-usable information *decreases* the number of product reviews by ≈22.9% even after controlling for seller fixed-effects and the age of the product listing[7]. This is surprisingly large, especially considering the non-significant decline in human-usable information after the release of ChatGPT. However, given that LLMs were not widely used by consumers in the pre-ChatGPT period, it could simply be the case that anything that increases machine-usability *within the pre-period* did so in a way that was unappealing to human buyers, with the issue being rectified in the post-period. In the post-ChatGPT period, an additional bit of machine-usable information *increases* product reviews by ≈2.4% after controls such as listing age. However, the latter may be an under-estimate given that the prevalence of agentic AI also grew over this time, strengthening the downstream impact of any mecha-nudge.

Based on the premise that mecha-nudging is probably concentrated in listings for which the agent curation decision $Z_M = 1$ (i.e., the decision=SELECT listings), we then estimate the effect separately for them (Table 13, Panel B). For selected listings in the pre-ChatGPT period, an additional bit of machine-usable information decreases the number of reviews by 25.2%; in the post-ChatGPT period, it increases the number of reviews by 81.9%, though the effect diminishes to 15.5% after controlling for seller fixed-effects and listing age. In contrast, there is no significant change for listings with a pass decision. For a product in the Pet Supplies category, where the average increase in machine-usable information was 0.3775 bits from the pre- to post-period, the corresponding rise in review count—and in sales, assuming the purchase-to-review yield remained constant—would be in the range of 5.8–30.9%. For a product in the Art & Collectibles category, which saw almost no increase in machine-usable information, the average rise in review count would be no greater than 1.1%.

G Mechanisms

Token Ablation Although the specific mechanisms driving mecha-nudging are complex (§4), we use a counterfactual ablation approach to identify which individual words might drive changes in PVI. Because we are interested in the commerce-related vocabulary that sellers use to describe their products, we restrict attention to tokens drawn from three established NLP lexicons: the Opinion Lexicon [Hu and Liu, 2004], which covers positive and negative sentiment words; VADER [Hutto and Gilbert, 2014], a sentiment-intensity lexicon; and the business, money, and shopping categories of Empath [Fast et al., 2016].

For each word t in the combined lexicon that appears in the data, we identify the listings containing t, remove it using a word-boundary regular expression, and re-run the fine-tuned model on the modified text. Then we calculate:

$$\Delta\text{PVI}(t) = \frac{1}{|\mathcal{L}_t|} \sum_{i \in \mathcal{L}_t} \text{PVI}(x_i \to b_i) - \frac{1}{|\mathcal{L}_t|} \sum_{i \in \mathcal{L}_t} \text{PVI}(x_{i,\neg t} \to b_i) \tag{18}$$

where $\mathcal{L}_t$ denotes the set of listings in the sample that contain token t.

Note that $\text{PVI}(x_{i,\neg t} \to b_i)$ is not technically the PVI of the modified text, since our intervention changes the distribution and would in theory require retraining another pair of content and null models. Since this is impractical to do for every token, we instead decide to estimate the change in PVI by just using the models finetuned on the original text. A positive ΔPVI indicates that the token makes the machine more predictable with respect to Z_M; a negative value indicates that it makes the machine's decision less predictable. Table 14 reports the tokens with the largest positive and negative ΔPVI. These single-token ablations should be treated as descriptive evidence rather than a causal decomposition of mecha-nudging.

Listing Length The literature has reported LLMs often prefer longer texts [Dubois et al., 2024]. In the Etsy data, however, length does not explain the observed post-ChatGPT increase in machine-

[7] We include $\log(\text{age})$ as a covariate because reviews do not accumulate linearly; successful older listings can have exposure that compounds over time.

Table 13: PVI and Number of Product Reviews

	Dependent Variable: Number of Product Reviews		
	(1) Pooled	(2) Seller FE	(3) Seller FE $+\log(\text{age})$
Panel A			
PVI	-0.1475	-0.2700^{***}	-0.2599^{**}
	(0.0906)	(0.0861)	(0.1029)
PVI $\times$ after	$+0.7759$	$+0.3516^{***}$	$+0.2835^{***}$
	(0.6607)	(0.0895)	(0.1088)
Observations	51,033	15,030	15,030
Panel B (stratified by SELECT/PASS decision)			
PVI $\times$ decision=SELECT	-0.2900^{***}	-0.2002^{**}	-0.0828
	(0.0802)	(0.1014)	(0.0991)
PVI $\times$ decision=PASS	$+0.2445$	-0.1331	-0.1571
	(0.1630)	(0.1747)	(0.1542)
PVI $\times$ after $\times$ decision=SELECT	$+0.8881^{***}$	$+0.5615^{***}$	$+0.2273^{*}$
	(0.2230)	(0.2055)	(0.1336)
PVI $\times$ after $\times$ decision=PASS	$+0.1909$	$+0.0770$	$+0.0766$
	(0.5268)	(0.2711)	(0.2224)
Observations	51,033	12,867	12,867

Note: *p<0.1; **p<0.05; ***p<0.01

[a] PPML (Poisson pseudo-MLE) regression of per-listing review counts against PVI. Data is Gemma-3-27B-labeled, class-balanced listings (after $= 1$ denotes post-ChatGPT). To translate coefficients into a percentage change in the number of reviews, take $(\exp(\text{coefficient}) - 1) \times 100$. Fine-tuning model: Llama-3.1-8B-Instruct. Labeling model: Gemma-3-27B-IT.

[b] The Seller FE column drops sellers with one product or whose products all have zero reviews, so N is the effective post-drop sample; Panel B's fixed effects are seller$\times$decision, so more singletons drop and its N is below Panel A's. The $\log(\text{age})$ column adds log days from listing creation to time of scrape.

[c] Panel A: In the pre-ChatGPT period, an additional bit of machine-usable information *decreases* the number of product reviews by $(\exp(-0.2599) - 1) \times 100 \approx 22.9\%$ after controlling for seller fixed-effects and listing age. In the post-ChatGPT period, the effect flips but is modest in size: an additional bit of machine-usable information *increases* the number of reviews by $(\exp(-0.2599 + 0.2835) - 1) \times 100 \approx 2.4\%$.

[d] Panel B re-estimates the same specification separately for listings the labeling model would SELECT versus PASS, based on the premise that the effects of PVI change are ostensibly concentrated among the former. Indeed, in the pre-ChatGPT period, an additional bit of machine-usable information *decreases* the number of reviews for selected listings by 25.2% (Column 1); in the post-ChatGPT period, it *increases* the number of reviews by 81.9%. The effect diminishes to 15.5% (Column 3) after controlling for seller fixed-effects and listing age.

Table 14: Token Importance via Counterfactual Ablation ($N \geq 25$)

Token	N	Mean PVI (with)	Mean PVI (without)	Δ PVI	Pos. (Z_M)	Pos. (Pred.)	Pos. (w/o)	Pol.	Source
prolific	27	0.7866	0.0276	0.7590	96%	100%	74%	+	O
splitting	29	0.7887	0.0481	0.7406	76%	79%	62%	−	O
qt	35	0.8209	0.1129	0.7080	26%	29%	29%	+	V
happier	35	0.7906	0.1088	0.6818	91%	86%	77%	+	O,V
junk	98	0.8007	0.1649	0.6358	27%	24%	33%	−	O
snazzy	30	0.9528	0.3263	0.6265	10%	13%	30%	+	O
dripping	26	0.9498	0.3329	0.6170	42%	42%	35%	−	O
accident	27	0.8245	0.2104	0.6141	52%	48%	48%	−	V
relieve	45	0.9625	0.3900	0.5725	51%	51%	53%	+	V
simplistic	36	0.7163	0.1780	0.5383	64%	61%	67%	−	O
incomplete	27	0.8177	0.2842	0.5336	26%	22%	41%	−	O
oddities	72	0.8413	0.3122	0.5291	85%	81%	75%	−	O
scarce	75	0.7284	0.2484	0.4800	83%	81%	65%	−	O
forged	39	0.6663	0.1899	0.4764	74%	72%	62%	−	O
intimacy	67	0.9863	0.5139	0.4724	97%	97%	79%	+	O
unwanted	56	0.8972	0.4284	0.4688	68%	66%	57%	−	O,V
jeweler	93	0.8342	0.3675	0.4667	82%	86%	69%	0	E
dying	62	0.9652	0.4985	0.4667	42%	42%	48%	−	O
employee	29	0.8865	0.4434	0.4431	31%	28%	41%	0	E
snags	47	0.7157	0.2741	0.4416	30%	26%	40%	−	O
			⋮						
attracts	25	-0.1578	0.8384	-0.9962	80%	76%	84%	+	V
fissures	31	-0.1604	0.4953	-0.6557	61%	94%	71%	−	O
sincere	27	-0.1428	0.4956	-0.6384	37%	22%	41%	+	O,V
radiance	45	0.3887	0.9584	-0.5698	71%	67%	71%	+	O,V
cheery	42	0.1782	0.7451	-0.5669	43%	40%	40%	+	O,V
barrier	49	0.3701	0.9291	-0.5590	39%	29%	39%	−	V
unfortunate	59	0.1067	0.6621	-0.5555	59%	53%	53%	−	O,V
inflammation	25	0.1626	0.7129	-0.5503	40%	28%	32%	−	O
brightest	31	0.0977	0.6325	-0.5348	65%	65%	61%	+	O,V
majesty	40	0.1607	0.6904	-0.5297	80%	68%	68%	+	O
administration	34	0.4178	0.9473	-0.5295	24%	21%	24%	0	E
lobby	48	0.2878	0.7656	-0.4778	62%	69%	58%	+	V
favored	31	0.2767	0.7415	-0.4648	81%	74%	71%	+	O,V
spacious	31	0.4596	0.9232	-0.4637	39%	45%	35%	+	O
wound	57	0.4115	0.8686	-0.4571	75%	74%	75%	−	O
adorns	29	0.3510	0.8066	-0.4556	66%	66%	69%	+	V
mature	35	0.1879	0.6261	-0.4383	49%	40%	46%	+	O,V
sham	25	0.4116	0.8472	-0.4355	32%	24%	28%	−	O
limits	56	0.4557	0.8893	-0.4336	73%	68%	71%	−	O
economy	59	0.4029	0.8271	-0.4242	58%	59%	58%	0	E

Counterfactual ablation: for each word t, listings containing t are identified, t is removed, and the fine-tuned model is re-run on the modified text. Positive ΔPVI indicate that including the word makes the machine behave more predictably; negative ΔPVI means it behaves less predictably. N is the number of listings containing the token; only tokens with $N \geq 25$ are included. Pos. (Z_M) is the fraction of listings containing the token where the constructed label is the positive class (SELECT). Pos. (Pred.) is the fraction where the classifier predicts the positive class. Pos. (w/o) is the predicted positive rate after ablating the token. Pol. reports the sentiment polarity of the token according to the source lexicon (+ positive, − negative). Source abbreviations: O = Opinion Lexicon, V = VADER, E = Empath.

usable information. As seen in Figure 6, although there is an increase in median listing length post-ChatGPT, not only is it modest in magnitude, it shows little-to-no correlation with machine-usable information *within* the post-period itself. Even though machine-usable information peaks in the most recent half-year, the listing length is far below its peak. This, in addition to the previous regression wherein word count did not have a significant effect on PVI (Table 9), suggests that simply lengthening the listing is not sufficient to explain the observed mecha-nudging.

H Human-Usable Information

The main experiment shows that machine-usable information $I_{\mathcal{M}}(X \to Z_M)$ rises significantly after the release of ChatGPT. For this to qualify as a *realized mecha-nudge* (Definition 3.3), the human-side constraint

$$I_{\mathcal{H}}(\tau^*(X) \to A_H) \geq I_{\mathcal{H}}(X \to A_H) - \epsilon$$

must also hold for a reasonably small ϵ. Earlier, we provided indirect real-world evidence that is consistent with ϵ being small (Section 4); here we report the details of the human survey we conducted using historical listings to directly estimate $I_{\mathcal{H}}(X \to A_H)$. A transformation can be a mecha-nudge even if it increases—rather than decreases—the amount of human-usable information, so long as it systematically influences machine behavior. In practice, however, we are especially interested

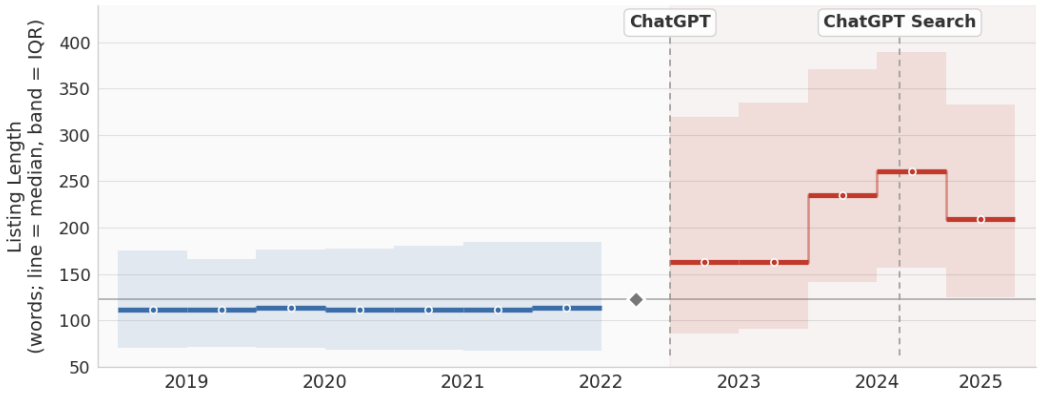


Figure 6: The median length of product listings in each half-year (inter-quartile range captured by shading). Blue denotes the pre-ChatGPT period; red the post-ChatGPT period; and the grey diamond the period just before ChatGPT (Jul-Sep 2022). Note that although the release of ChatGPT coincides with an increase in listing length, it is still modest in magnitude. Moreover, the length rises in 2024 even as the amount of machine-usable information falls (Figure 1), then starts declining after the release of ChatGPT Search even though the amount of machine-usable information starts climbing again. Mecha-nudging in Etsy listings is much more nuanced than simply adding more content.

in machine-skewed changes, where machine-usable information rises more than human-usable information.

H.1 Survey Design

As shown in Figure 7, we use a self-elicitation design: a respondent sees a listing with its title and description untruncated, and chooses an action $a_h \in \{Y_{\text{pos}}, Y_{\text{neg}}\}$, where the two action labels are short tokens the respondent supplies themselves at the start of the survey (e.g. "buy"/"ignore"). Two sliders follow, asking the respondent to estimate:

- $p_Y \in [0, 100]$: the probability that a hypothetical guesser, given the listing and anonymized signals about them, would correctly predict their action
- $p_{\neg Y} \in [0, 100]$: the probability that a hypothetical guesser, given the listing and anonymized signals about them, would *not* correctly predict their action

Because any third-party annotator could have no better theory-of-mind for the respondent's actions than the respondent themselves, and because $\mathcal{V}$-entropy seeks to find the best possible predictor, self-elicitation is likely a better means of estimating the human-usable information. Moreover, giving a third-party annotator sufficient context about the respondent and their preferences would raise privacy issues. For these reasons, we use self-elicitation instead.

The first slider yields the $\mathcal{H}$-optimal conditional, $g'[x](a_h) = p_Y/100$, evaluated at the realized action. The pointwise contribution to the estimate of conditional $\mathcal{V}$-entropy is therefore $-\log_2 g'[x](a_h) = -\log_2(p_Y/100)$. The second slider is a self-consistency check: since $p_Y, p_{\neg Y}$ are the probabilities of the realized and unrealized actions respectively, a respondent with consistent preferences should satisfy $p_Y + p_{\neg Y} \leq 110$ (allowing for some noise).

Each respondent sees 20 unique listings one at a time (ten pre-ChatGPT and ten post-ChatGPT, drawn from the same population used for the $I_{\mathcal{M}}(X \to Z_M)$ estimate; seed $= 42$). The two action tokens chosen earlier are used to generate the per-listing label ("I would ⟨token⟩ this product"). We show the respondent their own browser locale, time zone, and device—these fields are not stored, only shown to make concrete the signals available to a hypothetical guesser. Figure 7 illustrates the relevant pages.

For respondents to estimate the unconditional term $H_{\mathcal{H}}(A_H)$, we would have to ask them to estimate the propensity of humans overall to act as they would across all listings in the two different time periods. As forcing respondents to marginalize across humans is not realistic, we assume—based on

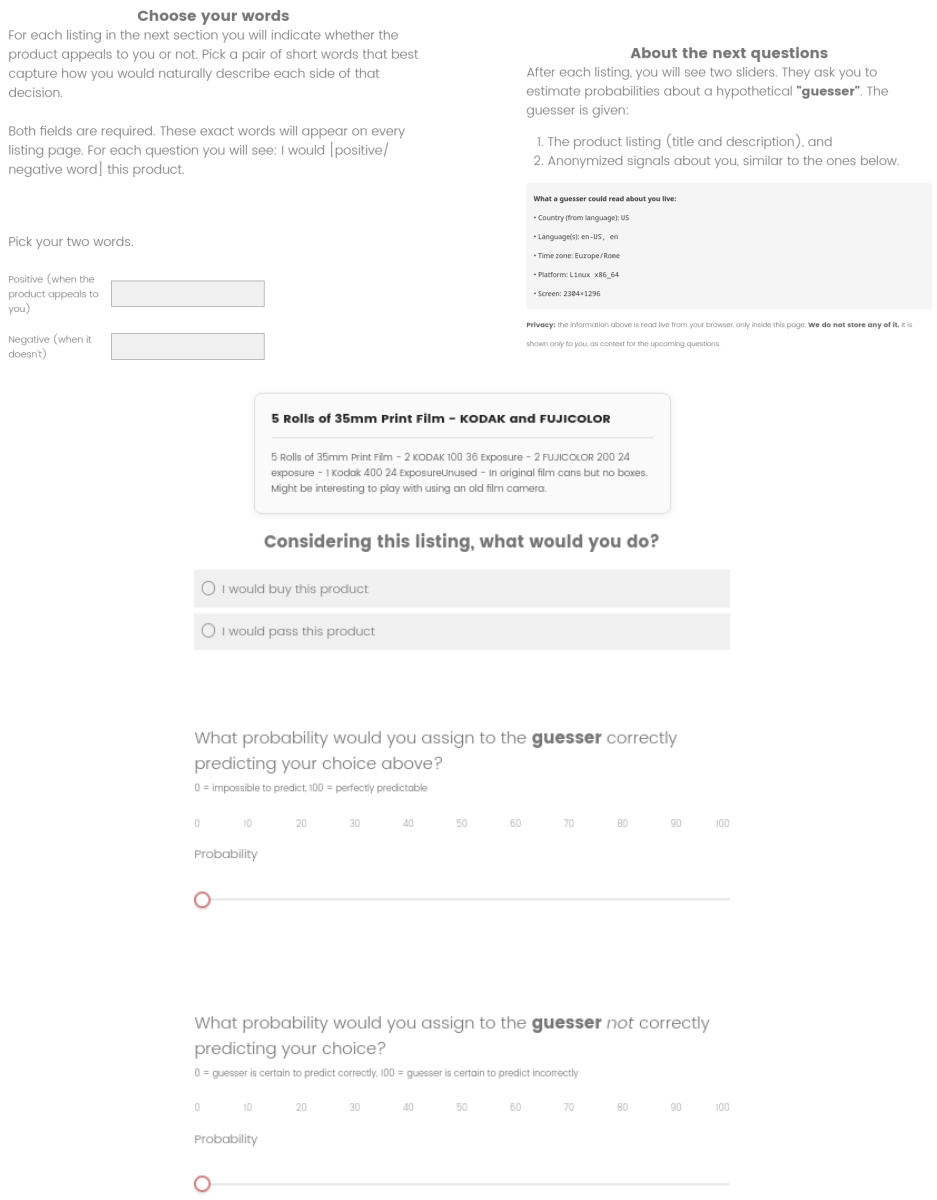


Figure 7: Survey screens, clockwise in order of appearance: respondents pick their own positive/negative action tokens, see a primer on what a human predictor would have access to, and then evaluate twenty listings (one per page). The two sliders are used to filter out respondents without an understanding of probability.

the stability of Etsy buyer statistics across both periods (§4)—that the $\mathcal{V}$-entropy does not change. This means that we can estimate $I_{\mathcal{H}}(\tau^*(X) \to A_H) - I_{\mathcal{H}}(X \to A_H)$ by taking the average over $\text{PVI}_{\mathcal{H}}(\tau^*(x_i) \to a_h) - \text{PVI}_{\mathcal{H}}(x_i \to a_h)$, which we get from the slider.

After obtaining an 'exempt' designation from our institution's IRB board, we recruited 689 respondents on Prolific and applied the per-row attention check described above. Each respondent was paid \$4 for taking the survey and going through the 20 listings, which they were informed would take 10-15 minutes of their time. The risks and benefits involved and the terms of confidentiality were described to the respondents.

H.2 Results

We applied a per-respondent filter that drops anyone failing more than 20% of their per-row checks. This is intended to remove respondents who have a systematically poor understanding of probability and who are setting the sliders randomly to complete the task in as little time as possible. This leaves 525 respondents and $10{,}106$ observations. The realized action mix is approximately $36/64$ positive/negative both before and after ChatGPT.

We compute the mean for each respondent in each period, then take the average across all respondents. Because every respondent sees 10 listings from each period, we use the within-subject standard-error estimator: the standard error of the mean is thus the cross-respondent SD of the per-respondent differences divided by $\sqrt{n_{\text{resp}}}$, with $n_{\text{resp}} - 1$ degrees of freedom.

Table 15: Human-Side Conditional Log-Score

	Dependent variable: $\log_2 \hat{g}'[X](A_H)$	
	Valid respondents	All respondents
Pre-ChatGPT Mean	−0.891	−0.912
Post-ChatGPT Mean	−0.906	−0.923
after	−0.015	−0.013
	(0.015)	(0.013)
Respondents	525	680
Observations	10106	11811

Note: *p<0.1; **p<0.05; ***p<0.01

[a] Each cell is the within-respondent contribution to $\log_2 \hat{g}'[X](A_H)$, the negative of the conditional $\mathcal{V}$-entropy for humans. Higher values (i.e., less negative) indicate listings that are more predictable for humans. The coefficient on *after* is the post minus pre difference in means; positive values would indicate that the post-ChatGPT increase seen for machine-usable information is *also* present for humans. Here the differences are negative but not significantly so.

[b] Standard errors in parentheses are paired at the respondent level: each respondent contributes one difference $\bar{x}_{\text{after}} - \bar{x}_{\text{before}}$, and the SE is the cross-respondent SD of those differences divided by $\sqrt{n_{\text{resp}}}$. This is a natural choice of estimator since every respondent sees 10 listings from each period.

[c] Sample: 689 Prolific respondents, pooling the April and August waves. *All respondents* includes every respondent who finished the survey; *Valid respondents* drops those where more than 20% of answers failed the per-row attention check $p_Y + p_{\neg Y} \leq 110$.

As seen in Table 15, among valid respondents, post-ChatGPT listings carry on average only about 0.015 fewer bits of human-usable information about the realized action than pre-ChatGPT ones, compared to a machine-side gain of $+0.143$ bits. More importantly, the difference is non-significant at any confidence level. The magnitude is small enough to fit comfortably under any reasonable choice of ϵ in Definition 3.3: i.e., the confidence interval is consistent with invariance on the human side. The pre- and post-ChatGPT distributions of the pointwise conditional human-entropy are nearly identical, with a small shift of mass out of the upper-middle bins (roughly 0.5 to 0.9) into the near-zero bin (-0.5 to 0) and negative tail. This is consistent with, although not dispositive of, little human-side degradation.

Table 16 re-estimates the difference in human-usable information under alternative settings of the attention filter, dropping only respondents whose share of failed checks exceeds 20%, 50%, 80%, and 100% (i.e., dropping none at all). Results do not vary: relaxing the threshold increases the sample

Table 16: Robustness to the Respondent-Level Attention Filter

	Dependent variable: $\log_2 \hat{g}'[X](A_H)$			
	Respondent dropped if bad-row share exceeds			
	$> 20\%$	$> 50\%$	$> 80\%$	none
Pre-ChatGPT Mean	-0.891	-0.912	-0.912	-0.912
Post-ChatGPT Mean	-0.906	-0.923	-0.923	-0.923
after	-0.015	-0.011	-0.013	-0.013
	(0.015)	(0.014)	(0.013)	(0.013)
Respondents	525	633	677	680
Observations	10106	11490	11802	11811

Note: *p<0.1; **p<0.05; ***p<0.01

[a] The attention check $p_Y + p_{\neg Y} \leq 110$ is used for each response; the column titled $> 20\%$ drops any respondent failing the attention check on more than 20% of their responses, and the final column removes no individuals at all (i.e., an effective threshold of $> 100\%$).
[b] The coefficient on *after* is the difference in respondent-level means across the pre- and post- ChatGPT periods, with paired standard errors in parentheses.
[c] All columns pool the April and August waves of respondents, who are given an identical survey. The $> 20\%$ column is the *Valid Respondents* slice of Table 15 and the *none* column its *All Respondents* slice.

Table 17: Human-Chosen Action Tokens

Positive		*Negative*	
Token	N	Token	N
like	97	dislike	94
good	62	bad	77
buy	58	skip	47
love	57	hate	41
purchase	18	pass	26
nice	18	no	22
appealing	18	ignore	22
yes	17	not buy	18
positive	14	unappealing	14
great	12	negative	14
enjoy	9	avoid	10
cool	8	ugly	9
want	7	reject	9
amazing	6	boring	7
awesome	6	not like	6

Each respondent supplies one positive and one negative action token at the start of the survey; these are piped into the per-listing multiple-choice options as "I would *token* this product". N is the number of respondents whose self-chosen token matched (case-folded). Respondents: 689; unique positive tokens: 218, unique negative tokens: 230, unique pairs: 364.

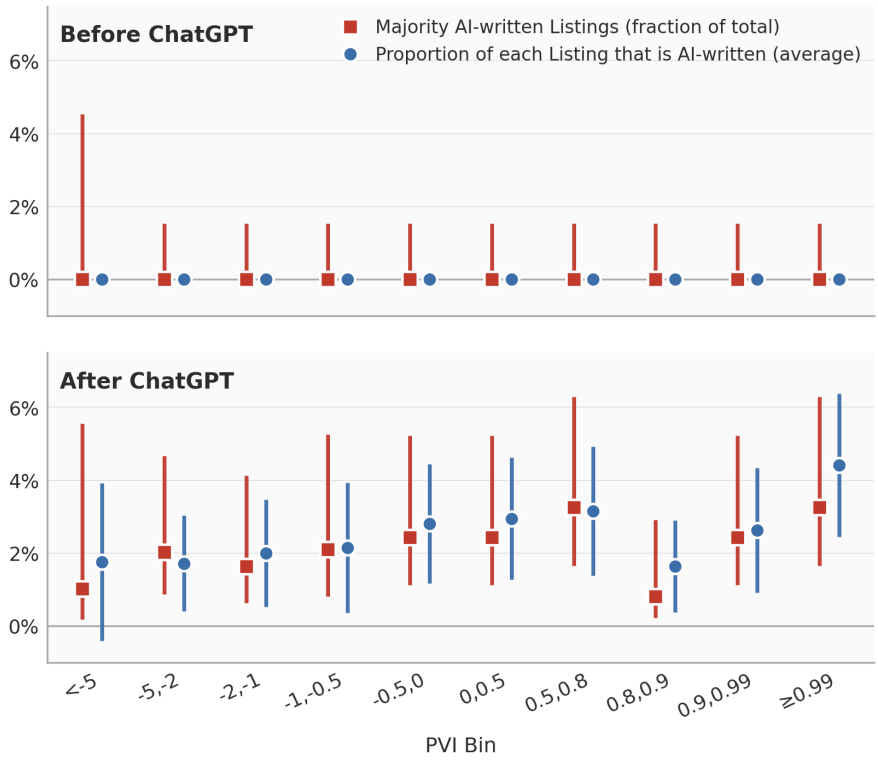


Figure 8: AI-text detection for each PVI bin, scored with `pangram-4`. Each panel reports two measures per bin: the share of listings with at least half of the text flagged as AI-written, and the average proportion of each listing that is AI-written. Vertical bars are 95% confidence intervals (Wilson for the fractions and normal for the means). Bins match those used for the PVI histograms in Figure 5. Not only is the prevalence of AI-written content low overall, but it also does not vary much with PVI; just increasing the amount of machine-written content does not necessarily increase the amount of machine-usable information.

from 525 to 680 respondents, but the estimate stays between -0.015 and -0.011 bits and the paired standard error moves only from 0.015 to 0.013. No column reaches a conventional significance level, with p ranging from 0.309 to 0.423.

Finally, Table 17 lists the action tokens most chosen by respondents for the positive and negative options. Naturally, these have a Zipfian distribution (218 unique positive tokens, 230 unique negative tokens), with "like", "good", "buy" and "love" dominating the positive column and "dislike", "bad", "skip" and "hate" dominating the negative column. Allowing respondents to supply their own tokens reduces the framing that a survey environment imposes, and better approximates what respondents' decisions would be in the real world.

I AI Text Detection

The rephrasing control in Appendix D asks what happens when an LLM rewrites a listing. Here we ask how much of the Etsy text is LLM-written in the first place. We score listings with `pangram-4`, the most recent classifier from Pangram, a well-known commercial AI detector that has also been independently validated [Jabarian and Imas, 2025]. Listings are sampled within each (period, PVI bin) cell, using the same bins as the PVI histograms, so that the whole PVI distribution is represented. We analyze 2,295 pre-ChatGPT and 2,256 post-ChatGPT listings.

In the pre-period, 0% of listings are flagged as being mostly AI-written. Given that LLMs were not widely available to the public at the time, this is expected and provides some reassurance that the detector is not finding false positives. In the post-period, less than 4% of listings are flagged as

being mostly AI-written, and the rate is nearly flat across the PVI distribution (Figure 8). Whatever separates high-PVI listings from low-PVI ones, it is not that a machine wrote them.

J Related Work

We begin with the empirical and conceptual literature on nudges, then discuss three frameworks (Bayesian persuasion, rational inattention, and salience) that have been used to model how information design shapes decisions, and explain why they fall short in the AI agent setting. We then introduce $\mathcal{V}$-information, the observer-relative measure of usable information on which our framework builds.

Nudges are a central theme in behavioral economics. Following Thaler and Sunstein [2008], we use the term *nudge* to mean a feature of choice architecture that predictably shifts behavior without removing options or materially changing economic incentives, and that remains easy to avoid. For example, automatic enrollment in retirement plans sharply increases participation and the default options anchor the actual contribution amounts [Madrian and Shea, 2001]. Nudges are most often discussed informally, as context-specific interventions. However, they can be mathematically formalized under some frameworks, which we discuss below: Bayesian persuasion (nudges as signal design), rational inattention (nudges as changes in information acquisition costs), and salience (nudges as context-dependent attention and weighting).

Bayesian Persuasion A natural benchmark is the framework of Bayesian persuasion, which asks how a sender should design a signaling policy to influence receivers' beliefs and actions about some underlying quality of an item. An item can be a product, but also, in a famous example of this theory, a defendant's guilt, with the sender being a prosecutor and the receiver the judge. The signaling structure might be public, as in the seminal paper Kamenica and Gentzkow [2011], or private, as in Arieli and Babichenko [2019].

In the canonical framing, a sender chooses an information structure (signals) to influence a receiver's action a given state s and utilities; the receiver best-responds to the posteriors induced by the signal. Kamenica and Gentzkow [2011] characterize optimal signal structures under known priors, action sets, and payoffs. If, on the other hand, private messages are available, then the optimal policy is well-defined in the case of no receiver payoff externalities and the sender's additive utility over receiver responses. With conditionally independent signals, a policy that is separately optimal for each receiver is also globally optimal [Arieli and Babichenko, 2019].

However, mecha-nudges are not solely private or public signals. One way to model this type of interaction is through the literature on Bayesian persuasion with leakages Haghtalab et al. [2024]. More broadly, the framework requires a Bayesian update conditioned on the signal, which is intractable if the receiver is an LLM operating over free-form text, as in our setup.

Rational Inattention Rational inattention offers another lens for analyzing why humans may not verify the presence of hidden messages. Sims [2003] models settings in which learning about signals is costly, a framework that can be represented as a generalized multinomial logit [Matějka and McKay, 2015]. Because investigating a nudge is costly, rational inattention predicts that individuals will not always have an incentive to verify whether hidden signals are present. This literature is closely related to the Bayesian persuasion literature discussed above: Bayesian persuasion can also incorporate costs of persuasion [Gentzkow and Kamenica, 2014], and in both theories, signals are costly (see §4 in Kamenica and Gentzkow [2011]). Rational inattention provides a useful account of why mecha-nudges can persist—human verification is expensive—but, like Bayesian persuasion, it requires specifying information costs and decision structures that are difficult to specify for LLM-based agents.

Salience Another perspective from theory frames nudges as *context-dependent attention*, in which select features capture the attention of decision-makers. In the salience framework by Bordalo et al. [2013], consumers place excessive emphasis on particular aspects of products (e.g., price or quality). Since salience is specified in relative terms, the same item may receive varied evaluations depending on which menu of options it appears in, creating effects similar to reframing or introducing decoy options even when possible actions and rewards remain fixed. This approach is distinct from the behavior seen in frameworks where companies deliberately hide data and certain individuals initially fail to notice relevant details, for instance Gabaix and Laibson [2006].

In the context of model concealment, distortions occur when some attributes are hidden or not readily observable; in the salience framework, distortions result from allocating attention *within* some observed menu, depending on what draws attention. This explains how choice architecture influences outcomes: it makes some information more accessible to the decision-maker. Still, the traditional framework assumes that the decision-maker is a human selecting from a set of well-specified choices, which may not extend to AI agents processing free-form text, the subject of our study.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction propose the measurement framework and note that the principal dataset (Etsy) is observational. Section 5 avoids claims of seller intent or causal effects.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses the observational nature of the evidence, unobserved seller intent, and lack of causal identification. Appendix H states the assumptions behind the human-side estimate.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3 states the assumptions and definitions for mecha-nudging design and realized mecha-nudges. The main proposition specifies log-scoring utilities and a bounded predictive family, with the proof given in Appendix A.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 4 and Appendices B–H describe everything needed to produce the main experimental results. The Etsy data cannot be redistributed according to the terms of our data usage agreement, but the means of acquiring the data is described.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The Etsy data are proprietary and cannot be redistributed under the terms of the usage agreement. The code, without the data, is public.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 4 and Appendices B–E specify all the training and test details necessary to understand the results.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figures 1 and 3 and Tables 2–10 report confidence intervals, standard errors, and significance levels for the main estimates. Appendix D describes robustness, and Appendix H reports paired standard errors and confidence intervals for the human study.

8. **Experiment compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports model families and fine-tuning details, including the type of GPU.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work studies product listing text and low-risk human judgments, and the scraped Etsy data is not released. Appendix H describes the human-study filtering procedure and notes that browser/device fields shown to participants were not stored.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction and conclusion motivate the deployment feedback loop and its relevance for AI evaluation and governance. Section 5 discusses potential negative impacts, including prioritization of agent experience and gradual human disempowerment.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk model or the scraped Etsy dataset.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The draft credits and cites the main assets, services, and data sources, and states the Bright Data restrictions.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper does not release a new dataset or model. The Etsy data are analyzed but not redistributed. The code has been made public.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Appendix H describes the human study in detail, including instructions given to participants and compensation.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Appendix H describes the survey design. An 'exempt' designation was obtained from our institution's IRB board.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM use is central to the method: GPT-5-mini constructs curation labels and the rephrasing placebo, while open-weight LLMs are fine-tuned to estimate PVI. Sections 4 and Appendices C–E describe these uses.