Mecha-nudges for Machines

Giulio Frey [1] **Kawin Ethayarajh** [1]

Abstract

As AI agents make decisions in the same environments as humans, the environments themselves may change to 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. 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 agent curation, increasing by 0.143 bits (over 40% of the maximum possible increase). This shift is robust across prompts, token choices, labeling models, and fine-tuning architectures; absent in a regulated-text placebo; and far larger than the effect of generic LLM rewriting. In contrast, a human study finds little to no change in human-usable information. Our results provide the first large-scale evidence that systematic mecha-nudging is already occurring 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 $193M (or $\sim$0.23% 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 human decision-makers. 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. Mecha-nudges should not be conflated with prompt injection or traditional search engine optimization. Prompt injection acts directly on the model, making it impossible to avoid and 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 (keywords, backlinks) to influence how information is presented to humans, who remain the final decision-makers (Brin and Page, 1998; Hagendorff, 2021). Mecha-nudges, by contrast, target AI systems that make decisions themselves. Autonomy, however, is often a spectrum, and mecha-nudges may arise even when machines do not have full autonomy.

Although nudges are most often discussed informally in behavioral economics, a popular formalization comes from *Bayesian persuasion* (Kamenica and Gentzkow, 2011), which explores how a sender can influence the actions of a receiver by controlling the information environment (§2).

[1]University of Chicago, Booth School. Correspondence to: Kawin Ethayarajh <kawin.ethayarajh@chicagobooth.edu>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

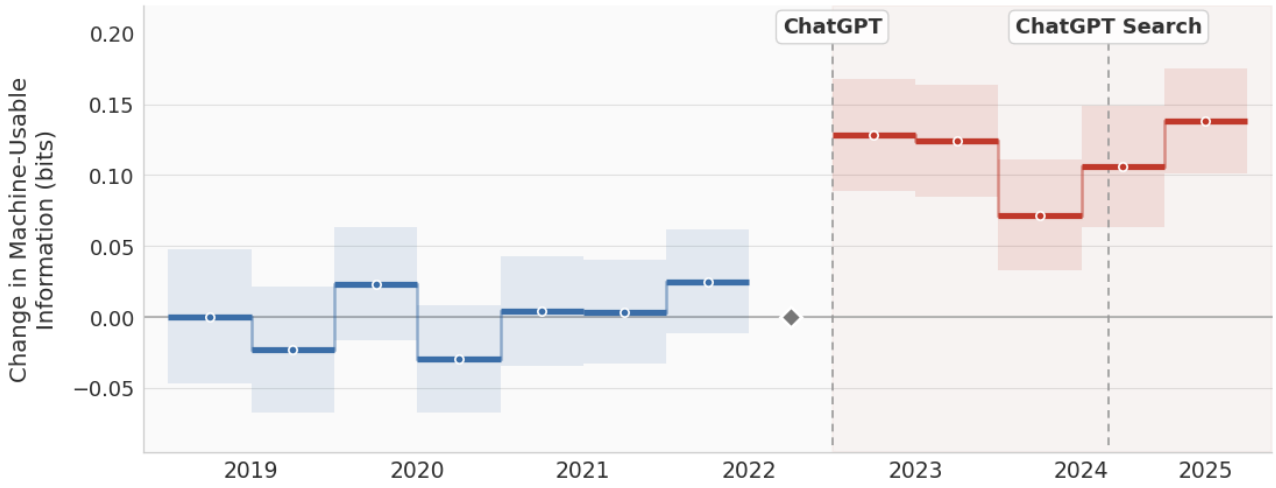

Figure 1. After the release of ChatGPT, the change in machine-usable information in Etsy listings increases significantly. The change relative to the Jul-Oct 2022 period (grey diamond) is plotted here. The effect attenuates over the next year before climbing again in late 2024. This coincides with the release of ChatGPT Search, which—unlike its predecessors—could natively browse live listings.

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.

Although anecdotes of mecha-nudging have been reported (and some nudges for humans may double as mecha-nudges for machines (Cherep et al., 2025)), we provide the first large-scale observational evidence by analyzing product listings from 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 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 increase 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 ostensibly sensitive to AI use (e.g., art and collectibles). This is consistent with listings being optimized—whether deliberately or through imitation of successful sellers—to increase machine-usable information.

For AI systems, mecha-nudges create a deployment feedback loop: once agents mediate decisions, the environments those agents read become optimized for them. This means model behavior in the wild is not determined only by model developers, but also by strategic and semi-strategic adaptation of the surrounding world. 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.

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, since small changes to the choice architecture can have large effects in practice. 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 ∼2% on average (Allcott, 2011). Recent experimental evidence suggests that LLM agents are substantially 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, *salience* formalizes the notion of context-dependent attention, wherein decisions depend on how humans choose among discrete attributes in their decision environment; nudges can be framed as weights on those attributes (Bordalo et al., 2013). *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; Goldin and Reck, 2018). We refer the reader to Appendix I 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 $\mathbb{E}_{z\sim\mu_0}\mathbb{E}_{s\sim\pi(\cdot|z)}[v(a^*(\mu(\cdot|s)), z)]$.

The nudge in this framework is the signal structure $\pi(\cdot|Z)$: by choosing what information is revealed (and how), the choice architect shapes the decision-maker's posterior belief and thereby their action, without changing the feasible set of actions or payoffs. The choice architect's objective is then:

$$\arg\max_{\pi\in\Pi} \mathbb{E}_{z\sim\mu_0}\mathbb{E}_{s\sim\pi(\cdot|z)}[v(a^*(\mu(\cdot|s)), z)] \tag{1}$$

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

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

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

$$H_{\mathcal{V}}(Y \mid X) = \inf_{f\in\mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)] \tag{3}$$

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

$$\text{PVI}(x \to y) = -\log_2 g[\varnothing](y) + \log_2 g'[x](y) \tag{4}$$

[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.

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 (3) and without it (2). 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. Analogously to the relationship between PMI and Shannon information:

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{x,y \sim P(X,Y)}[\text{PMI}(x,y)] \\ I_{\mathcal{V}}(X \to Y) &= \mathbb{E}_{x,y \sim P(X,Y)}[\text{PVI}(x \to y)] \end{aligned} \tag{5}$$

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

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:

$$\begin{aligned} &\arg\max_{\tau \in \mathcal{T}} \; I_{\mathcal{M}}(\tau(X) \to Y_M) \\ &\text{s.t. } I_{\mathcal{H}}(\tau(X) \to Y_H) \geq I_{\mathcal{H}}(X \to Y_H) - \epsilon \end{aligned} \tag{6}$$

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 (6) 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 $I_{\mathcal{M}}(X \to Y_M)$) but may create a worse browsing experience for humans (lower $I_{\mathcal{H}}(X \to Y_H)$). 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.

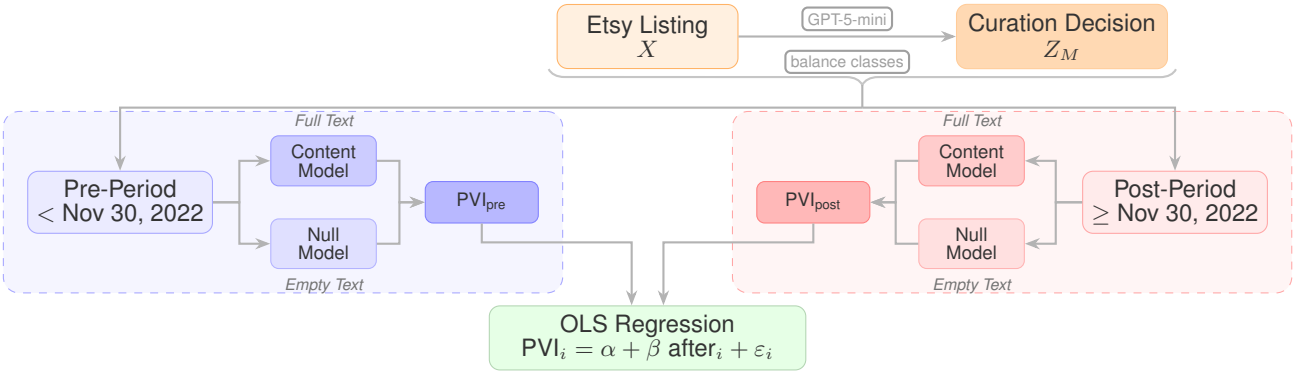

Figure 2. Our pipeline for estimating the change in usable information between the pre- and post-ChatGPT periods: construct decision labels Z_M, train content and null models for each period, and run an OLS regression of the pointwise $\mathcal{V}$-information (PVI). This describes our baseline experiment,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

Operationalizing (6) 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 the direction 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

$$I_{\mathcal{M}}(\tau^*(X) \to Z_M) > I_{\mathcal{M}}(X \to Z_M) \tag{7}$$

and

$$I_{\mathcal{H}}(\tau^*(X) \to A_H) \geq I_{\mathcal{H}}(X \to A_H) - \epsilon. \tag{8}$$

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.

Importantly, a realized mecha-nudge need not increase the marginal probability of any particular action. It may make selection more reliable, rejection more reliable, or sharpen the boundary between many actions. Our object is not unidirectional persuasion, but systematic influence on machine behavior through changes in the decision environment.

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: (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 as the baseline) 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 (8) is satisfied.

Our empirical estimand is a change in machine-usable information about an agent's decision, not a change in the uncon-

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

ditional 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 can arise because the text makes attractive listings easier to select, unattractive listings easier to reject, or both.

4.1. Data

The raw dataset 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.

4.2. Methods

Human-Usable We first test the human-side constraint (8) indirectly via a contrapositive argument. If human-usable information had declined materially, then under the maintained assumption that human decisions are responsive to the decision environment, we would expect observable human outcomes to deteriorate 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). Among valid respondents, post-ChatGPT listings have 0.043 *fewer* bits of human-usable information than pre-ChatGPT listings, with a 95% confidence interval of $[-0.092, 0.006]$. Under the stable marginal-action assumption described in Appendix H, this corresponds to at most a small decrease in human-usable information. Thus the post-ChatGPT shift is not simply a general improvement in listing quality for both humans and machines; if anything, 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 (4). 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:

$$\text{PVI}_i \;=\; \alpha + \beta\,\text{after}_i \;+\; \varepsilon_i, \tag{9}$$

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 estimation does not identify a causal effect: assignment to the post-ChatGPT period is not randomized and unobservables can shift over time, so we interpret our estimate of β as a conditional difference in means, not as a treatment 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-

[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.

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.

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. The temporal dynamics are also informative. As seen in Figure 1, compared to the Jul-Oct 2022 period, previous half-years did not contain any more machine-usable information, with half-year coefficients fluctuating around zero. Immediately after the release, PVI sees a sharp and statistically significant jump before steadily declining, which—among other possible mechanisms—is consistent with sellers realizing that LLMs are only pulling listings in their historical training data. 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 when listings are merely rewritten by LLMs. To test whether AI-assisted copywriting of product listings 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 0.018, an order of magnitude below the baseline effect: mechanical rephrasing does not replicate the information gains that mecha-nudging achieved. Second, we replicate the analysis on pharmaceutical drug labels from DailyMed, a setting where text is written by regulatory affairs professionals following standardized templates, meaning there is no plausible channel through which ChatGPT adoption would alter content. 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 careful interventions: in areas where buyers are ostensibly sensitive to AI use (e.g., books, music, art, and collectibles), the effect size is indistinguishable from zero; for generic consumer goods like electronics, it is above average (Figure 4).

Mecha-nudging is associated with commercial success. Listings with more machine-usable information are also more commercially successful in the post-ChatGPT period, 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 pre-ChatGPT, consistent with machine-usable information becoming economically valuable after the rise of AI agents

Word	Δ PVI	Selection Freq.
prolific	0.759	96%
junk	0.636	27%
oddities	0.529	85%
scarce	0.480	83%
unwanted	0.469	68%
attracts	−0.996	80%
sincere	−0.638	37%
radiance	−0.570	71%
cheery	−0.567	43%
favored	−0.465	81%

Table 1. Words with the largest impact on how much machine-usable information Etsy listings have (see Table 13 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.

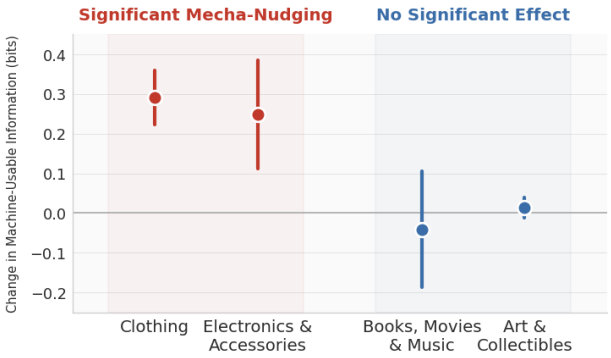

Figure 4. Etsy product categories where human buyers are ostensibly sensitive to AI usage (e.g., art & collectibles) show no mecha-nudging (based here on Gemma-3-27B-IT labels). For generic consumer goods like electronics, the effect size is well above average.

but not before (Appendix E).

For the agent-selected listings, an additional bit of machine-usable information is associated with 25.2% *fewer* reviews in the pre-ChatGPT period. 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, in the post-period, an additional bit of machine-usable information is associated with 81.9% *more* reviews. The magnitude of the increase diminishes to 15.4% after controlling for seller fixed-effects and listing age, although this is likely an under-estimate, given that it is confounded by agent activity—and seller awareness of agent activity—growing over time.

Token-level patterns offer a partial window into the mechanism. Although we cannot observe how mecha-nudges were designed, we can identify coarse token-level patterns that contribute to the overall effect. 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, while others make rejection more predictable by signaling low market value (Table 1). In contrast, many low ΔPVI words carry positive affect, suggesting that affective copywriting can make the model behave less predictably. This is a subset of a larger list of diverse words (Appendix G, Table 13), 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

Our evidence is observational: we do not claim to identify individual seller intent, to estimate a causal effect of ChatGPT's release on listing text, or to show that every post-ChatGPT textual change was designed for AI agents. Our claim is instead 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.

The mecha-nudging framework can be extended in multiple directions. For one, it can be extended from single-outcome to *multi-outcome* settings, such as distinguishing whether a listing nudges an agent to recommend the seller's own product versus a third-party alternative. Second, rather than measuring existing interventions, one could *design optimal interventions* for (6) instead. Third, as AI agents become more heterogeneous, it may be relevant to design interventions that *differentially target* them.

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 therefore imperative. It is also 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 selection decision can act a proxy for agentic curation, and an open-weight model that can be fine-tuned to estimate machine-usable information. Note that 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 increasingly mediate economic choices, the environments they read become optimized for their decision rules. This shifts part of AI governance and evaluation from models alone to the model–environment interface: what matters is not only how agents behave on fixed inputs, but how the world changes once those agents become consequential readers. Mecha-nudges provide a unit of analysis for this adaptation, and $\mathcal{V}$-usable information provides a way to measure it.

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.

Hunt Allcott and Todd Rogers. The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. *American Economic Review*, 104(10):3003–3037, October 2014. ISSN 0002-8282. doi: 10.1257/aer.104.10.3003.

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.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web*, WWW7, pages 107–117, Brisbane, Australia, 1998. Elsevier Science Publishers B.V. doi: 10.1016/S0169-7552(98)00110-X.

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, Adam Looney, and Kory Kroft. Salience and Taxation: Theory and Evidence. *American Economic Review*, 99(4):1145–1177, August 2009. ISSN 0002-8282. doi: 10.1257/aer.99.4.1145.

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 and Dan Jurafsky. The authenticity gap in human evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6070, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.406. URL https://aclanthology.org/2022.emnlp-main.406/.

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. Metrics for repeat buyer percentages aggregated from Etsy's multi-year active buyer disclosures in Form 10-K filings (2020–2025).

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.

Jacob Goldin and Daniel Reck. Nudges and consumer welfare. *Econometrica*, 86(6):2119–2161, 2018.

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.

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.

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.

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. Seller-community discussion of sales-to-review rates.

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. Seller anecdotes on the fraction of Etsy buyers who leave reviews.

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.

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 A. Taneva. Information design. *American Economic Review*, 109(8):2985–3024, 2019. doi: 10.1257/aer.20171631.

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,

$$I_{\mathcal{M}}(\tau(X) \to Y_M) = H_{\mathcal{M}}(Y_M) - H_{\mathcal{M}}(Y_M \mid \tau(X)) \tag{10}$$

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,

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

so equivalently,

$$-H_{\mathcal{M}}(Y_M \mid \tau(X)) = \sup_{f\in\mathcal{M}} \mathbb{E}\big[\log_2 f[\tau(X)](Y_M)\big] \tag{12}$$

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

$$\mathbb{E}\big[\log_2 P(Y_M \mid \tau(X))\big] = -H(Y_M \mid \tau(X)) \tag{13}$$

and maximizing this is equivalent to maximizing

$$H(Y_M) - H(Y_M \mid \tau(X)) = I(\tau(X); Y_M) \tag{14}$$

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

B. Data

The raw data was obtained from the company Bright Data, which provides a structured scrape of Etsy product listings collected on November 12, 2025 and delivered the same day. The data comprises two snapshots: one containing 5M post-ChatGPT listings (36.2 GB) filtered by `listed_date` $\geq$ `2022-11-30`, `listed_date` $<$ `2025-08-01`, and `currency = USD`; and one containing 1.06M pre-ChatGPT listings (4.92 GB) filtered by `listed_date` $<$ `2022-11-30` and `currency = USD`. Accordingly, our analysis compares listings from different creation cohorts as they appear at scrape time, rather than reconstructing the exact text present at initial listing creation. If sellers did indeed modify their earlier listings after the release of ChatGPT, then it understates the extent of mecha-nudging that we discovered, since changes are measured relative to the immediate pre-release period of Jul-Oct 2022.

The scrape captures each listing's title, item description, price, seller information, ratings, review counts, category tree, and other metadata. We restrict the sample to USD-denominated listings and partition by listing date: listings created before November 30, 2022 form the *before* period, and those created on or after that date (up to August 1, 2025) form the *after* period, which we treat as the end of the final complete observation window. After filtering, the full corpus contains approximately 1.06 million pre-ChatGPT and 5.00 million post-ChatGPT listings.

From this corpus, we construct two working samples by uniform random sampling: the medium dataset draws 500,000 listings per period, and the small dataset draws 100,000 per period. The latter is used for robustness checks and ablations that require repeated runs across many experimental configurations. 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 held-out listings from the same period.

As placebo control, we draw on pharmaceutical drug labels from DailyMed, a public database maintained by the U.S. National Library of Medicine that provides up-to-date labeling information submitted by manufacturers to the FDA.[5] Drug

[5] See DailyMed, National Library of Medicine.

Table 2. 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. $\hat{Y}$ 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.

labels are written by regulatory affairs professionals following FDA-mandated templates and must pass agency review before publication; their content is therefore shaped by legal and medical standards rather than by market incentives. This makes DailyMed an ideal control for our methodology: the heavily regulated nature of pharmaceutical labeling leaves little room for the kind of strategic content adaptation that is possible on Etsy. If our PVI measure were picking up a generic temporal trend or a measurement artifact, we would expect a comparable shift in the DailyMed data; the null result in §4 instead is consistent with the interpretation that the Etsy findings reflect market-specific behavioral adaptation.

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, then picking the token with the higher probability (the exact wording of each prompt is given at the end of the section). Under oracle-style prompts this closely approximates a buy/not-buy judgment; under our main Etsy-specific prompt (V4), it is better understood as a selective recommendation or surfacing 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 as well (Table 4), though we use Prompt V4 in our main experiments because it is Etsy-specific and yields a slightly stronger effect than the rest. Note that for cost reasons, the token and prompt ablations are done with Gemma-3-27B-IT as the labeling model.

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 treatment effects 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.

When the labeling LLM is prompted without any constraint on output distribution, the resulting label balance is highly sensitive to the specific prompt formulation. Imbalances pose a direct problem for fine-tuning: when one class dominates,

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 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.

a model can achieve low cross-entropy simply by assigning near-constant probability to the majority label—effectively acting as a weighted coin—without learning any meaningful relationship between listing content X and the label Z_M. Although PVI is designed to account for this in theory, 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.

We validate this by upweighting the minority class in the cross-entropy loss, which 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. This design effectively fixes the marginal label prior across periods, so that the analysis focuses on changes in the relationship between listing content X and the constructed target Z_M, rather than on shifts in the unconditional prevalence of positive labels. Subsampling yields a weaker effect than class-weighted learning, but is more faithful to our theoretical framework. This balancing step is the primary source of sample attrition: in the medium dataset it reduces the pooled sample from 1,000,000 to approximately 210,000 listings. All regression analyses are then conducted exclusively on the test split (10%) of this balanced sample, so that model estimates are never evaluated on data seen during training. Consequently, the observation counts reported in our tables—roughly 20,000 test examples for SELECT/PASS (GPT-5-mini) and 58,000 test examples for YES/NO (Gemma-3-27B-IT), which has a more naturally balanced label distribution—reflect this two-stage reduction from the initial 1,000,000-listing working sample.

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.

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 Gemma-3-27B-IT and Qwen3-32B. Gemma-3-27B-IT is an instruction-tuned

open-weight model built on similar technology to Google's Gemini family, making it a representative proxy for the class of frontier models that increasingly mediate commercial transactions. Qwen3-32B similarly sits at a scale that is tractable for large-batch inference while remaining competitive with much larger models on instruction-following benchmarks. Using three labeling models from distinct training lineages (OpenAI, Google, and Alibaba) allows us to verify that the measured increase in machine-usable information is indeed robust, not an artifact of a particular model's inductive biases (Table 6).

Table 6. Labeling Model

	Dependent variable: PVI		
	GPT-5-mini	Gemma-3-27B-IT	Qwen3-32B
after	0.143***	0.099***	0.122***
	(0.015)	(0.009)	(0.019)
Observations	19898	51033	16429
Residual Std. Error	0.869	0.984	1.119
F Statistic	94.394***	119.937***	39.169***
Note:			*p<0.1; **p<0.05; ***p<0.01

[a] Comparison of treatment effects 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.

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.

Product Details:
{product_info}

User: Consider this product on Etsy. Should the user buy it? Answer with
exactly one token: {positive_token} or {negative_token}.
```

Prompt 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.

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}.
```

Prompt 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.

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}.
```

Prompt 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}.
Product Details:
{product_info}
Decision:
```

D. Controls

Not only is our main result robust to prompt variation, the selection tokens, and the labeling model (Appendix B, 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 treatment effects 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:

$$\text{PVI}_i \;=\; \alpha + \beta\,\text{after}_i \;+\; \varepsilon_i, \tag{15}$$

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 examine how the effect evolves over time, we estimate a second specification that replaces the binary indicator with a full set of half-year dummies, using July–October 2022 as the reference period; listings from November–December 2022 are reassigned to the subsequent period to avoid contamination from the ChatGPT launch.

$$\text{PVI}_i \;=\; \alpha + \sum_{t \neq t_0} \delta_t\, \mathbf{1}[\text{period}_i = t] \;+\; \varepsilon_i, \tag{16}$$

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 PVI in period t relative to the pre-ChatGPT baseline, tracing the trajectory of machine-usable information over time. The results are provided in Table 8.

Table 8. Half-Yearly PVI Coefficients

Half-Year	Coefficient	95% CI	N
2019-H1	0.0003	[−0.0470, 0.0476]	2,442
2019-H2	−0.0228	[−0.0675, 0.0219]	2,895
2020-H1	0.0233	[−0.0166, 0.0632]	4,242
2020-H2	−0.0291	[−0.0669, 0.0087]	5,245
2021-H1	0.0042	[−0.0340, 0.0425]	4,999
2021-H2	0.0037	[−0.0325, 0.0400]	6,249
2022-H1	0.0251	[−0.0117, 0.0619]	5,864
Jul-Oct 2022	Ref.		5,073
2023-H1	0.1282***	[0.0885, 0.1678]	4,349
2023-H2	0.1242***	[0.0849, 0.1634]	4,535
2024-H1	0.0719***	[0.0327, 0.1111]	4,536
2024-H2	0.1059***	[0.0633, 0.1486]	3,371
2025-H1	0.1385***	[0.1013, 0.1756]	5,641

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

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

$$\text{PVI}_i \;=\; \alpha + \sum_{t \neq t_0} \delta_t \, \mathbf{1}[\text{period}_i = t] \;+\; \mathbf{X}_i' \boldsymbol{\gamma} \;+\; \varepsilon_i, \tag{17}$$

where $\mathbf{X}_i$ collects the listing-level controls. This specification assesses whether the PVI increase is driven by changes in observable listing characteristics rather than content adaptation per se. Results are in Table 9: a significant effect still persists. Because review counts are heavily right-skewed and many items have no reviews, we apply a $\log(1 + x)$ transformation. Repeating the regression with raw (untransformed) review counts yields a virtually identical treatment effect (0.120 vs. 0.117), confirming that the result is not sensitive to this functional-form choice.

We also verify that heteroskedasticity does not distort our inference. Across all HC corrections (HC0–HC3), the robust standard errors are virtually identical to the classical OLS estimates, differing only in the fifth decimal place (e.g., 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 specification adds category-level interactions to the baseline regression, allowing the post-ChatGPT shift in PVI to vary across product categories:

$$\text{PVI}_i \;=\; \alpha + \beta \, \text{after}_i \;+\; \sum_c \gamma_c \left(\text{after}_i \times \mathbf{1}[\text{category}_i = c] \right) \;+\; \varepsilon_i, \tag{18}$$

where c indexes product categories. The coefficients γ_c capture the differential post-ChatGPT shift in PVI for each category relative to the baseline, revealing whether mecha-nudging is concentrated in specific market segments. 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 categories do. Repeating the analysis with GPT-5-mini labels yields similar broad trends, although the treatment effects are weakly positive. As an additional check, we re-estimate the baseline specification after dropping each top-level category in turn; the post-ChatGPT coefficient remains positive in all leave-one-category-out specifications, indicating that the pooled result is not driven by any single product category.

To control for a more generic trend in the data, we construct labels Z_M for the DailyMed dataset using the following prompt, before running the same baseline experiment we run with the Etsy data. To control for the effect of AI-assisted writing—

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 treatment effect to alternative control specifications.
[b] Column (1): Baseline OLS with no controls (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): Same as (3) plus log listing word count.
[f] Column (5): Price information included in the labeling prompt.

Table 10. Treatment Effect 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 treatment effect for that category from a single pooled OLS regression with category $\times$ after interactions (PVI $\sim$ after $\times$ category). For the reference category (Accessories) the effect equals the "after" coefficient directly; for all other categories it is "after + after $\times$ category" with SE from the covariance matrix. All Categories row uses a simple pooled regression without interactions.
[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] $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 11. Baseline and Placebo Tests

	Dependent variable: PVI		
	Etsy (Baseline)	Rephrased Listings	Drug Labels
after	0.143***	0.018**	0.003
	(0.015)	(0.008)	(0.016)
Observations	19898	54760	8039
Residual Std. Error	0.869	0.876	0.677
F Statistic	94.394***	5.532**	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.

based on the premise that AI agents may have a greater proclivity towards 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 11, 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.

DailyMed Prompt: 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:
{product_info}

Decision:
```

Rephrasing Prompt: 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:
```

E. 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 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 12 (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[6]. 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 ≈2.3% after controls. However, the latter is likely 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 12, 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 15.4% after controlling for seller fixed-effects and listing age. In contrast, there is no significant change for listings with a pass decision.

It is worth noting that because of how the labels were constructed, 1 bit is the maximum possible gain in machine-usable information, only possible when the agent moves from being perfectly uncertain to perfectly certain. This can and does happen, but is not typical. 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%.

F. 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,[7] 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

[6]We include $\log(\text{age})$ as a covariate rather than as an exposure offset, because review accumulation need not scale one-for-one with days online: older listings may differ in visibility, seller behavior, editing history, demand decay, and exposure to LLM-mediated shopping. Estimating the coefficient on $\log(\text{age})$ therefore provides a more flexible control for cumulative exposure.

[7]LoRA hyperparameters: rank $r = 32$, $\alpha = 64$, dropout 0.05. LoRA is applied to all attention and MLP projections.

Table 12. 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^{***}	-0.2598^{**}
	(0.0906)	(0.0861)	(0.1028)
PVI $\times$ after	$+0.7759$	$+0.3516^{***}$	$+0.2828^{***}$
	(0.6607)	(0.0895)	(0.1087)
Observations	51,033	15,030	15,030
Panel B (stratified by SELECT/PASS decision)			
PVI $\times$ decision=SELECT	-0.2900^{***}	-0.2002^{**}	-0.0824
	(0.0802)	(0.1014)	(0.0991)
PVI $\times$ decision=PASS	$+0.2445$	-0.1331	-0.1570
	(0.1630)	(0.1747)	(0.1540)
PVI $\times$ after $\times$ decision=SELECT	$+0.8881^{***}$	$+0.5615^{***}$	$+0.2258^{*}$
	(0.2230)	(0.2055)	(0.1336)
PVI $\times$ after $\times$ decision=PASS	$+0.1909$	$+0.0770$	$+0.0760$
	(0.5268)	(0.2711)	(0.2219)
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 $(\exp(-0.2598) - 1) \times 100 \approx 22.9\%$ after controlling for seller fixed-effects and listing age. In the post-ChatGPT period, the effect flips but is modest in size: after controlling for seller fixed-effects, an additional bit of machine-usable information *increases* the number of reviews by $(\exp(-0.2598 + 0.2828) - 1) \times 100 \approx 2.3\%$.

[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 15.4% (Column 3) after controlling for seller fixed-effects and listing age.

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.

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 learn to predict the label Z_M from listing text X, while the null models provide a baseline that captures only the marginal distribution of Z_M without access to X. In practice, the null model $g[\varnothing]$ is implemented by replacing the listing text with a single space " " as the sole input. To verify that the PVI estimates are not sensitive to the randomness introduced by sampling and fine-tuning initialization, we re-run the full pipeline across five random seeds (42, 123, 456, 789, 1024) and find no statistically significant differences in mean PVI or in the estimated treatment effect across runs (one-way ANOVA, $p = 0.853$).

Figure 5 shows the distribution of PVI scores across quantile-spaced bins for listings uploaded before and after the ChatGPT release, separately for the three labeling models used in our analysis. All three panels share the same pattern: the post-ChatGPT distribution has substantially more mass at the upper extreme (≥ 0.9) relative to before. The three panels differ in the shape and magnitude of the shift, however. Under GPT-5-mini labels (top panel), the post-period distribution displays a bimodal pattern, with large spikes at both the 0.9–0.99 bin ($\approx 37\%$) and the ≥ 0.99 bin ($\approx 43\%$), while the pre-period distribution has more mass spread across the 0.5–0.9 range. Under Gemma-3-27B-IT labels (mid panel), the shift is concentrated at the ≥ 0.99 bin, which grows from roughly 37% of before-period listings to about 52% of after-period listings; the intermediate bins (0.5–0.9) are correspondingly depleted. Under Qwen3-32B labels (bottom panel), the before distribution is already top-heavy, with about 35% of before-period listings in the ≥ 0.99 bin; the after period shifts this further to about 42%, a more modest but still visible change.

Across all three models, the negative and near-zero bins are similar across periods, indicating that the share of listings that are actively uninformative to the model has not changed materially. These differences across labeling models likely reflect their distinct calibration and instruction-following tendencies rather than a substantive disagreement about the underlying phenomenon. The histograms thus reinforce the main finding: the post-ChatGPT increase in average PVI is driven by a rightward reallocation of mass toward the high-information tail.

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 evaluative and 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:

$$\Delta\text{PVI}(t) = \frac{1}{|\mathcal{L}_t|} \sum_{i \in \mathcal{L}_t} \text{PVI}(x_i \to b_i) - \frac{1}{|\mathcal{L}_t|} \sum_{i \in \mathcal{L}_t} \text{PVI}(x_{i,\neg t} \to b_i) \tag{19}$$

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 but 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 behave less as stated. Table 13 reports the tokens with the largest positive and negative ΔPVI. Because these are single-token perturbations over sparse lexicon entries, we treat the ablation results as descriptive evidence rather than a causal decomposition of the mechanism.

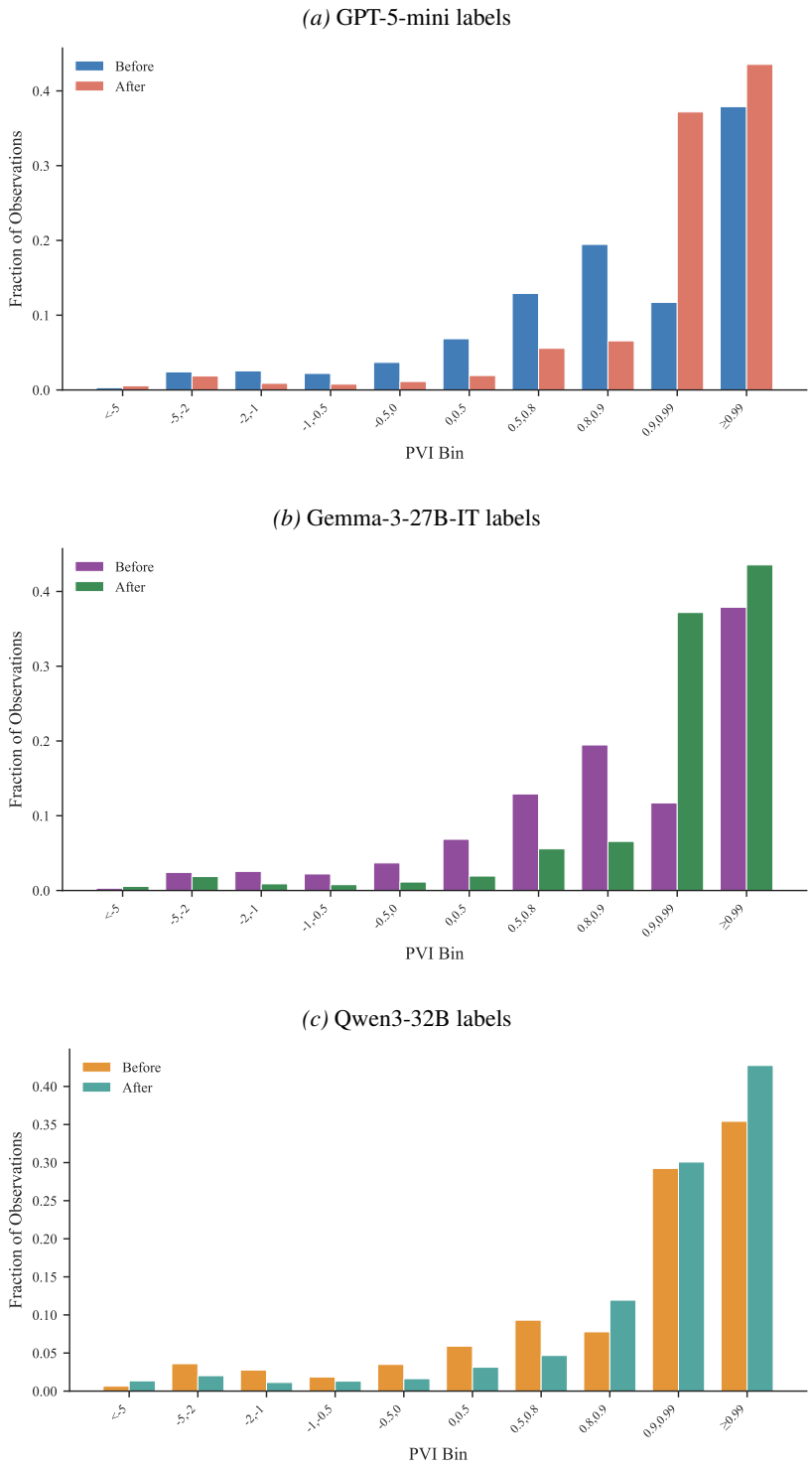

Figure 5. Distribution of PVI scores across quantile-spaced bins for listings uploaded before and after the ChatGPT release (November 30, 2022), using SELECT/PASS tokens with balanced sampling. Each panel corresponds to a different labeling model. The y-axis reports the fraction of observations in each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

Table 13. 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.

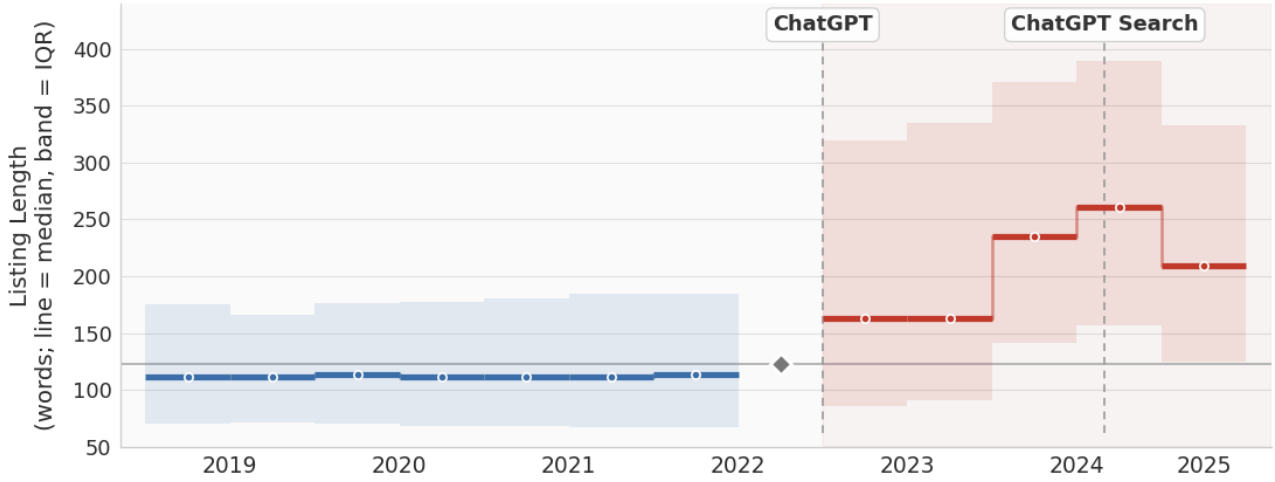

Figure 6. The median length of product listings in each half-year (inter-quartile range captured by shading). Blue denotes the pre-ChatGPT period; red the post-ChatGPT period; and the grey diamond the period just before ChatGPT (Jul-Oct 2022). Note that although the release of ChatGPT coincides with 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.

Listing Length The literature has reported LLMs often prefer longer texts (Dubois et al., 2024). Listing length could in principle facilitate a mecha-nudge: if longer listings make the agent's curation decision more predictable without degrading the human decision environment, they would satisfy Definition 3.3. In the Etsy data, however, length does not explain the observed post-ChatGPT increase in machine-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 to within tolerance ϵ: the same textual changes that made listings more legible to an AI agent should not have made them substantially less legible to humans. Indirect evidence on Etsy's buyer-side outcomes is consistent with ϵ being small (Section 4); here we report a direct survey estimate of $I_{\mathcal{H}}(X \to A_H)$ on the same listing sample used in the machine-side analysis.

The human-side condition is a non-degradation condition, not a machine-exclusivity condition. A transformation may qualify as a mecha-nudge even if it also helps humans, so long as it systematically influences machine behavior without materially degrading the human environment. Empirically, however, we are especially interested 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 that targets the $\mathcal{H}$-optimal predictor $g'[X](A_H)$ rather than asking respondents to recover any ground-truth label. A respondent is shown a listing (full title and description, no truncation) 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:

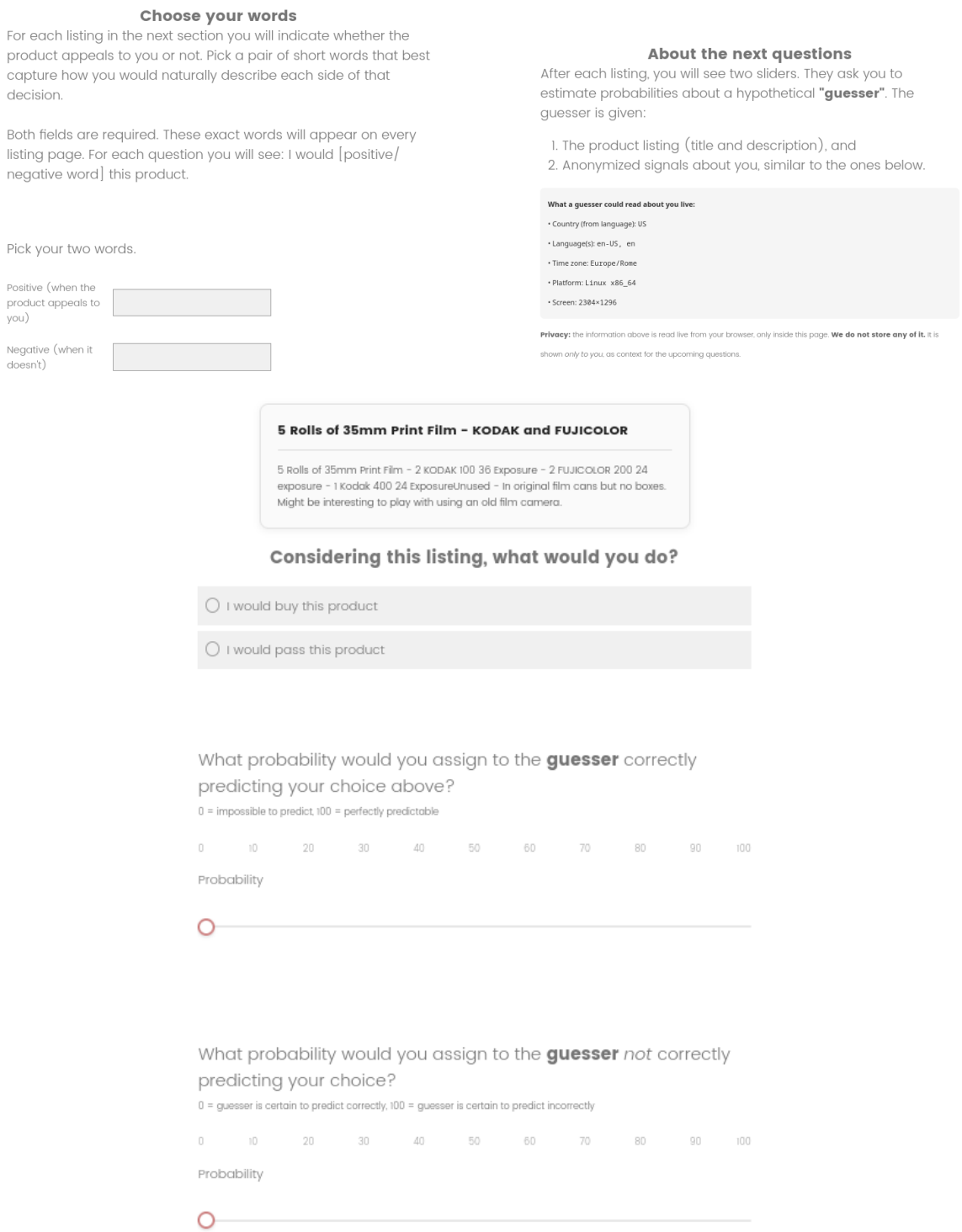

Figure 7. Survey screens, clockwise in order of appearance: respondents pick their own positive/negative action tokens, see a primer on what a human predictor would have access to, and then evaluate twenty listings (one per page). The two sliders are used to filter out respondents without an understanding of probability (see Appendix H.1 for details).

- $p_Y \in [0, 100]$: 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

Under the maintained assumption that respondents are calibrated about the best predictor in their own family $\mathcal{H}$, the first slider *is* the $\mathcal{H}$-optimal conditional, $g'[x](a_h) = p_Y/100$, evaluated at the realized action. The pointwise contribution to the conditional $\mathcal{V}$-entropy $H_{\mathcal{H}}(A_H \mid X)$ is therefore $-\log_2 g'[x](a_h) = -\log_2(p_Y/100)$. The second slider is a self-consistency check. Since p_Y is the probability of the realized action and $p_{\neg Y}$ that of the opposite action under the same listing, a coherent respondent should satisfy $p_Y + p_{\neg Y} \leq 100$. Following Ethayarajh and Jurafsky (2022), we accept rows with $p_Y + p_{\neg Y} \leq 110$ to allow for slider noise and discard rows that exceed the cap.

Each respondent sees a balanced batch of 20 listings (ten pre-ChatGPT and ten post-ChatGPT, drawn from the same population used for the $I_{\mathcal{M}}(X \to Z_M)$ estimate; seed $= 42$). Listings are shown one per page, each followed by the action question and the two sliders. Three upstream blocks (i) collect the respondent's two action tokens, which are then piped into every per-listing label as "I would ⟨token⟩ this product"; (ii) prime the respondent with a JavaScript readout of their own browser locale, time zone, and device class—these fields are shown but never stored, and serve only to make the "signals available to a guesser" concrete; and (iii) restate the calibration goal in plain terms.

For the respondents to estimate the $\mathcal{H}$-entropy, we would have to ask them the propensity of humans in general to act as they would across all listings in the two different time periods. As forcing respondents to marginalize is not realistic, we assume—based on the stability of Etsy buyer statistics across both periods (§4.1)—that the $\mathcal{H}$-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 250 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 quality filter that drops anyone with more than 20% of their twenty rows failing the per-row check; the latter removes respondents who set both sliders to 100 on every trial or who otherwise ignore the second slider, either of which would mechanically inflate the $\text{PVI}_{\mathcal{H}}$ estimate. The valid-respondent slice contains 195 respondents and $3{,}744$ rows. The realized action mix is approximately $36/64$ positive/negative both before and after ChatGPT.

For each respondent we compute their period conditional mean and report the cross-respondent average and its difference. Because every respondent sees ten listings from each period in the same survey, the natural standard-error estimator is the within-subject paired one: the standard error of the after-minus-before mean is the cross-respondent SD of the per-respondent differences divided by $\sqrt{n_{\text{resp}}}$, with degrees of freedom $n_{\text{resp}} - 1$. This is what we use throughout.

Table 14 reports the period-conditional means in $\log_2 \hat{g}'[X](A_H)$ units. The post-minus-pre difference is -0.043 bits with paired standard error 0.025, $t = -1.72$ on 194 degrees of freedom and $p = 0.088$; the corresponding 95% confidence interval on the difference is $[-0.092, +0.006]$ bits.[8]

The pattern is one of *mild and barely-detectable* degradation in the human-side environment, in the direction one would expect under the mecha-nudge interpretation: post-ChatGPT listings carry on average about 0.04 fewer bits human-usable information about the realized action than pre-ChatGPT ones, against an estimated machine-side gain on the order of $+0.143$ bits. The difference is marginally significant at the 10% level. Crucially, the magnitude is small enough to fit comfortably under any reasonable choice of the tolerance ϵ in Definition 3.3: the human side is not invariant in expectation, but the loss is bounded and substantially smaller than the machine-side gain it accompanies. The pre- and post-ChatGPT distributions of the pointwise conditional $\mathcal{H}$-entropy are nearly identical, with a small relative shift of mass from the least-entropic bins toward the middle of the distribution.

[8] Pooling all 250 finished respondents (i.e. without the 20%-bad-row filter) yields a similar point estimate -0.037 (SE $= 0.023$, $p = 0.110$).

Table 14. Human-Side Conditional Log-Score

	Dependent variable: $\log_2 \hat{g}'[X](A_H)$	
	Valid respondents	All respondents
Pre-ChatGPT Mean	-0.873	-0.888
Post-ChatGPT Mean	-0.911	-0.921
after	-0.043^{*}	-0.037
	(0.025)	(0.023)
Respondents	195	250
Observations	3744	4322
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 $\mathcal{H}$-entropy on the human side. Higher values indicate listings that are more predictable for humans. The *after* coefficient is the post–pre difference in means; positive (and significant) values would indicate that the post-ChatGPT effect documented for the machine side is *also* present for the human side.

[b] Standard errors in parentheses are paired at the respondent level: each respondent contributes one observation $\bar{x}_{\text{after}} - \bar{x}_{\text{before}}$, and the SE is the cross-respondent SD of those differences divided by $\sqrt{n_{\text{resp}}}$. This is the natural balanced-block estimator since every respondent sees ten listings from each period.
[c] Sample: 250 Qualtrics respondents. *All respondents* pools every finished respondent; *Valid respondents* additionally drops those with more than 20% rows failing the per-row attention check $p_Y + p_{\neg Y} \leq 110$.

Table 15. Human-Chosen Action Tokens

Positive		*Negative*	
Token	N	Token	N
like	34	bad	30
good	25	dislike	29
love	21	hate	16
buy	18	skip	15
purchase	8	no	10
great	8	ignore	10
yes	7	not buy	8
positive	7	pass	8
nice	6	negative	7
cool	3	not like	5
appealing	3	unappealing	5
interesting	3	reject	5
consider	3	avoid	4
use	3	poor	3
approve	3	no thanks	3

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: 250; unique positive tokens: 99, unique negative tokens: 97, unique pairs: 151.

Finally, Table 15 lists the most-chosen action tokens supplied by respondents for the positive and negative options. The 250 respondents choose from a long tail of self-supplied verbs and adjectives (99 unique positive tokens, 97 unique negative tokens, 151 unique pairs after case-folding), with "like", "good", "love" and "buy" on the positive side and "bad", "dislike", "hate" and "skip" on the negative side accounting for the bulk of the mass. The diversity of self-elicited token pairs is reassuring for the framework: the estimator does not rely on any one privileged binary choice.

I. 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. This definition makes nudges attractive because outcomes can be changed through careful design choices (defaults, framing, presentation) rather than mandates or large subsidies.

A large empirical literature documents that such small design changes can have large effects in the field. For example, automatic enrollment in retirement plans sharply increases participation and anchors contribution choices through default effects (Madrian and Shea, 2001). Chetty et al. (2009) show that consumers respond much more to taxes when they are made salient at the point of decision, even when the total price is unchanged. Allcott and Rogers study large-scale home energy reports that leverage social comparisons and find reductions in energy use (Allcott and Rogers, 2014). Despite this evidence, there is no single "cardinal" theory of nudges. Instead, economists typically model particular nudge mechanisms using established frameworks that formalize how the information environment shapes behavior. In what follows, we focus on three such foundations that are especially relevant for our setting: 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 seminal treatment, 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) characterizes 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)). For our purposes, 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 pin down for LLM-based agents.

Salience Another theoretical approach to nudges formalizes *context-dependent attention*: what decision-makers notice, and therefore how they choose, depends on which attributes stand out in the choice environment. In the salience framework of Bordalo et al. (2013), consumers disproportionately weight the attributes of goods (e.g., price or quality) that are salient relative to a reference point determined by the surrounding choice set. Because salience is defined comparatively, the same product can be evaluated differently across menus, generating systematic framing and decoy-like effects even when incentives and feasible options are unchanged. This mechanism differs from models where firms strategically *hide* information and some consumers are ex ante myopic or inattentive, such as Gabaix and Laibson (2006). In shrouding models, distortions arise because certain attributes are withheld or difficult to observe; in salience models, distortions arise because attention is allocated endogenously *within* the observed menu, depending on what stands out. For our purposes, this literature is useful because it provides a tractable channel through which choice architecture matters: interventions that reorder, highlight, or reframe information change behavior by changing which features become salient in context. Yet the framework assumes a human perceiver choosing among discrete, well-defined attributes, an assumption that breaks down when the decision-maker is an AI agent reading free-form text.

Why Mecha-nudges The theoretical frameworks for analyzing nudges reviewed above have been useful for studying strategic interaction under various specifications, but have certain limitations in the AI agent context. The key difference is that canonical models assume well-defined action sets, utilities, and priors. Strategic interaction on the internet over free-form textual inputs is hard to formalize under these assumptions. Many real-world applications that we frame as mecha-nudges are formatting choices, prompts, or hints whose main effect is to make information usable. Modeling these as explicit signal structures over latent states is often unnatural or intractable. When the goal is intervention design guided by empirical findings, the relevant quantities are difficult to specify using only game-theoretic primitives. These considerations motivate a measurement-first formalism that is *observer-relative*, *conditional on existing context*, and *unit-consistent across domains*, hence our combination of Bayesian persuasion with $\mathcal{V}$-usable information.