Demand for Conversational AI

Elliott Ash, Francesco Capozza, Sergio Galletta

October 6, 2026

Abstract

Relational uses of conversational AI register as barely demanded in stated-preference data, even as the market for AI companions grows quickly. We estimate demand across five purposes (administrative assistance, studying, wellness, friendship, and romance) in a preregistered within-respondent conjoint experiment ($N = 1{,}989$) that independently randomizes price, privacy, personalization, modality, and safety policy. Purpose dominates every design attribute: respondents require $7.80 per month to accept a romantic companion in place of an administrative assistant, six to seven times the largest non-price feature effect. Making the product free does not close this gap: the zero-price premium is roughly halved for a romantic companion. Second-order beliefs are compressed across purposes relative to own interest, and the ratio of the belief–interest gap to the purpose penalty is statistically indistinguishable across the two relational purposes, as a single social-image cost parameter implies. That ratio doubles as the welfare statistic and does not depend on the price slope: if the entire gap is suppression, social image accounts for 39.5% of the romantic penalty and 46.6% of the friendship penalty. A hierarchical Bayes demand model recovers the taste distribution and shows that social-image sensitivity shifts relational tastes more than instrumental ones while leaving price and attribute sensitivities unchanged; belief shrinkage cannot account for the pattern, but the structural estimates imply that only 10–15% of the belief–interest gap is suppression, so most of the measured relational penalty is intrinsic taste.

Keywords: conversational AI; conjoint analysis; willingness to pay; social image; second-order beliefs; hierarchical Bayes.
JEL: C81; C93; D82; M31

Ash: Center for Law and Economics, ETH Zurich and CEPR, Zurich, Switzerland
Capozza: Department of Public Economics, University of Barcelona, IEB, CESifo, Barcelona, Spain.
Galletta: Department of Economics and Law La Sapienza, Rome, Italy.

The study received IRB/ethics approval from *[ETH, protocol # 356].* The design and analysis plan were preregistered on AsPredicted (`https://aspredicted.org/hj4wz8.pdf`). The authors declare no competing interests.

1 Introduction

Ask people what they would pay for an AI that keeps them company and the answer is: less than nothing. In our data, respondents require $7.80 a month in compensation to accept a romantic companion in place of an administrative assistant. Meanwhile the companion market is one of the fastest-growing categories in consumer software. Somebody is wrong, and it is worth working out who.

The growth is not in dispute. Between 2022 and mid-2025 the number of dedicated AI companion applications rose by 700%, reaching 337 active revenue-generating products by July 2025, with over 220 million lifetime downloads and an 88% year-on-year increase in downloads in the first half of 2025 (Perez, 2025). Purpose-built companions such as Replika and Character.AI attract tens of millions of users, and over half of Character.AI's 20 million monthly actives are under 24 (SQ Magazine, 2025). Table A.1 documents the landscape by primary purpose, pricing, and user base.

Measured by the standard tools of demand analysis, relational AI looks close to unwanted. Stated interest in a romantic companion is the lowest of any purpose we study, and the $7.80 penalty is six to seven times the value of any non-price design feature we vary. A manager reading these numbers would conclude that the relational segment is not worth entering. A regulator would conclude that it generates little value. Both would be reading a market that is visibly growing, and both would be reading it wrong.

This paper argues that a substantial part of the measured relational penalty is not a preference parameter but a *social-image cost*. Reporting interest in a romantic AI companion, even on an anonymous survey, risks revealing an undesirable type: that one is lonely, socially awkward, or unable to sustain human relationships. Stated demand therefore falls below latent demand, and it does so specifically for relational purposes, where the signaling cost is high, and not for instrumental ones, where admitting use signals efficient technology adoption. That creates a measurement problem, and it is the paper's central object. The demand estimate a firm or a regulator would obtain from any stated-preference instrument, and from most revealed-preference instruments, is a composite of intrinsic taste and social-image cost. The two look identical in the data. They imply opposite things about what to do next: one says the segment is small, the other says the segment is hidden.

We field a preregistered, within-respondent, rating-based conjoint experiment (Hainmueller et al., 2014, 2015) in a broadly representative U.S. sample ($N = 1{,}989$). Each respondent evaluates five profiles, one per purpose (Admin, Study, Wellness, Friend, Romantic), with five independently randomized design attributes: monthly price ($10, $5, free), interaction modality, personalization depth, privacy protections, and safety policy. Alongside own interest we elicit second-order beliefs (SOBs): respondents' estimates of how many out of 100 similar people would want to try the same profile. We estimate a hierarchical Bayes demand model that recovers

the population distribution of willingness to pay and admits an individual social-image sensitivity parameter as a second-level covariate.

The design choice that does the work is the joint elicitation of own interest and second-order beliefs *on the same randomized profile.* Own interest is suppressed by the respondent's own social-image cost; beliefs about others are not, or are less so. The wedge between them, differenced against an instrumental baseline to remove scale artifacts, is our measure of suppression. Because both the purpose penalty and the belief–interest gap are generated by the same underlying stigma parameter, the model implies a cross-equation restriction that is testable rather than imposed.

First, purpose dominates demand, and the instrumental–relational split is externally corroborated. Relative to Admin, a Friend companion reduces interest by 0.572 points on a 0–10 scale and a Romantic companion by 1.262 points; Study raises interest by 0.153 points and Wellness is statistically indistinguishable from Admin. The same instrumental–relational separation appears in two behavioral sources: U.S. Google Trends search intensity in 2025 and the anonymized ChatGPT usage distribution of Chatterji et al. (2025). We use these sources to validate the ordinal structure only; their scales are not comparable to ours, and they are themselves shaped by product positioning and, plausibly, by the same stigma we study.

Second, price is the strongest non-purpose lever but cannot close the purpose gap. The \$5-to-free gain is roughly four times the \$10-to-\$5 gain, consistent with zero-price effects (Shampanier et al., 2007). Critically, the free-pricing premium is *attenuated* for relational purposes: the free-versus-\$10 premium is 1.96 interest points for Admin against 0.94 for Romantic, an interaction of -1.02 ($t = -6.3$), while the corresponding \$5-versus-\$10 interactions are all indistinguishable from zero, the attenuation is specific to the zero-price margin, although the \$5 interactions are imprecise enough that we cannot exclude proportional attenuation there. Under any additive linear model the price effect is constant across purposes by construction; the data reject this. Removing the monetary barrier leaves the non-monetary barrier intact.

Third, perceived social demand differs systematically from expressed private demand, in the pattern a single-factor signaling model predicts. Own interest spans 1.42 points across purposes; beliefs span only 0.80 points on the comparable scale. The ratio of the Admin-relative belief–interest gap to the Admin-relative purpose penalty is 0.466 (bootstrap s.e. 0.067) for Friend and 0.395 (0.033) for Romantic, and we cannot reject equality (difference 0.071, bootstrap $p = 0.217$). This is what one parameter governing both the suppression of stated interest and the compression of beliefs implies; Section 2.4 is explicit about the alternatives it does not exclude.

Fourth, the suppression is real but does not account for most of the penalty. Following Allcott and Taubinsky (2015), we value the suppression at the ratio of the belief–interest gap to the price slope, giving \$3.08/month for a romantic companion at the benchmark scale conversion, 39.5% of the measured \$7.80 Admin–Romantic penalty if the whole gap is suppression, and 30% once

we remove the part that simple belief shrinkage would generate. A hierarchical Bayes model with an individual signaling-sensitivity covariate recovers a correction from a different moment. In the specification in which no rating enters both sides, it delivers +\$0.30/month for Romantic and +\$0.25/month for Friend, implying that only about 0.10 (Romantic) to 0.15 (Friend) of the belief–interest gap is suppression. We are explicit about what this is and is not: it recovers the *individual* suppression component under a fixed norm, not the additional surplus that would be released if aggregate adoption shifted the norm itself, which is not identified from individual stated-preference data.

Section 2 develops the framework. Rather than treating stigma as an exogenous taste shifter, we follow Bernheim (1994), in whose model social costs arise endogenously: the cost of deviating from a norm is proportional to how clearly the deviation reveals an undesirable predisposition. The first mechanism is *purpose-as-signal*. Admin or Study use signals efficient technology adoption; romantic use signals unmet social needs. This generates the purpose gradient and two testable restrictions we take to the data: an attenuated zero-price premium for relational purposes, and an approximately constant ratio of belief–interest gap to purpose penalty across them. The second is a *coordination failure* at the equilibrium rather than the individual level. Each agent adopts optimally given the prevailing norm, but none internalizes that adoption itself lowers the signaling cost for others. The low-adoption equilibrium is individually stable and socially suboptimal, the structural inverse of the social-media collective trap of Bursztyn et al. (2025a). There, active users impose costs on non-users and the market over-adopts; here, each adopter lowers the social cost for the next and the market under-adopts. Both are invisible to standard surplus measures.

Three literatures meet here. On *measurement*, recent work estimates demand for large language models from revealed-preference data (Kireyev and Vitorino, 2025; Caoui, 2025). That approach avoids hypothetical bias but cannot vary design attributes independently of competing products, cannot price purposes that are underrepresented in platform logs *precisely because of* the social-image concerns we document, and cannot recover the gap between current behavior and behavior under a shifted norm. The two approaches are complements; the convergence of our instrumental–relational split with Google Trends and with the ChatGPT usage distribution supports reading the survey gradient as market structure rather than survey artifact. On *social image*, we connect the conformity and stigma literature (Bernheim, 1994; Bursztyn et al., 2025b; Reif, 2025; Carvajal et al., 2025) and work on algorithm aversion (Dietvorst et al., 2015; Castelo et al., 2019; Dargnies et al., forthcoming) to a demand-estimation setting where the social cost is attached to the *category* rather than to the act of using an algorithm as such. On *welfare*, we combine Bernheim and Rangel (2009), who specifies the ancillary condition (social observability) whose removal recovers welfare-relevant preferences, with Allcott and Taubinsky (2015)'s money-metric conversion. The distortion here is a social-image cost imposed on accurate self-reporting, not a decision error relative to the agent's own preferences, which

shifts the policy implication toward reducing observability rather than paternalistic correction.

Methodologically, we combine a stated-preference conjoint, second-order belief elicitation on the same profiles, and a hierarchical Bayes demand model (Rossi et al., 2005; Moshary et al., 2025) that recovers the full cross-respondent distribution of willingness to pay and embeds the signaling-sensitivity parameter as a structural covariate. To our knowledge this is the first paper to estimate the mean and distribution of demand for conversational AI jointly with a bound on its social-image distortion.

The results depart from the standard attribute-optimization playbook. Because purpose effects dwarf feature effects, positioning is the highest-use design choice. Because the zero-price premium is attenuated exactly where the demand gap is largest, price cannot buy its way past a non-monetary barrier, and free tiers for relational products should be engineered as conversion systems rather than discounts. And because stated demand understates latent demand for relational purposes, the returns to reducing observability (discreet interaction modes, private billing descriptors, strong privacy defaults, de-stigmatizing bundling under wellness or study framing) may exceed the returns to further feature investment. Section 8 develops these points and attaches money-metric magnitudes.

Section 2 presents the model, defines every parameter and estimand, and states the identifying assumptions. Section 3 describes the design and sample. Section 4 sets out identification and estimation. Section 5 reports the reduced-form estimates, external validation, and the two mechanism tests. Section 6 estimates the hierarchical Bayes demand model. Section 7 develops the welfare analysis and counterfactuals. Section 8 draws out the managerial implications. Section 9 concludes.

2 Conceptual Framework and Model

This section sets out the choice environment, defines every primitive and parameter used in the paper, derives the four hypotheses we test, and states what is and is not identified. Table 1 collects the notation for reference.

2.1 The choice environment

Agents and alternatives. A population of consumers is indexed by $i = 1, \ldots, N$. A conversational AI product (an *alternative*) is fully described by a pair $(p, \mathbf{d})$, where

- $p \in \mathscr{P} = \{\mathsf{Admin}, \mathsf{Study}, \mathsf{Wellness}, \mathsf{Friend}, \mathsf{Romantic}\}$ is the product's *purpose*, the stated use case around which it is positioned. Purposes partition into *instrumental* uses, directed at external task completion (Admin, Study), and *relational* uses, directed at social and emotional experience (Friend, Romantic), with Wellness intermediate: it addresses personal

well-being but through informational and coaching mechanisms closer to instrumental help than to social bonding (Ruggiero, 2000; Ta et al., 2020; Brandtzaeg et al., 2022; Laestadius et al., 2024).

- $\mathbf{d} = (m, \iota, \kappa, \rho, \varsigma)$ is the vector of *design attributes*: monthly price $m \in \{0, 5, 10\}$ in dollars; interaction modality ι; personalization depth κ; privacy protections ρ; and safety policy ς, each with three levels (Table 2).

We write $\mathbf{x}_{it} \in \mathbb{R}^K$ for the design vector of profile t faced by respondent i: an intercept, four purpose indicators (baseline Admin), the numeric price m_{it}, and two indicators for each of the four non-price attributes (baselines Text + Voice, Advanced personalization, High privacy, Adults-only safety), so $K = 14$. Price enters *numerically*, not as a set of level indicators; there is accordingly no omitted price level.

Observability. Let $o \in [0, 1]$ denote the *observability* of adoption: the probability that a socially relevant audience learns that the consumer uses a product of purpose p. Observability is a property of the adoption environment, not of the product. We do not randomize it, and $\bar{o}$ denotes the prevailing regime implicit in the survey; privacy protections ρ shift perceived observability but are not the same object. Section 2.7 is explicit about the consequences.

The choice. Each respondent evaluates $T = 5$ profiles, one per purpose, and reports (i) own interest $r_{it} \in [0, 10]$ and (ii) a second-order belief $s_{it} \in [0, 100]$, the number out of 100 similar people who would want to try that profile. Two features of this environment matter for what follows. First, the instrument is a *rating*, not a choice from a menu; there is no outside option. Second, ratings are ordinal by construction and cardinal only under assumption. Both constrain the welfare analysis, and we carry the constraint explicitly rather than assuming it away (Section 7).

2.2 Utility, reporting, and the social-image cost

Latent utility. Respondent i's latent, welfare-relevant utility from profile t is

$$U_{it} \;=\; \alpha(p_t) \;+\; \mathbf{d}_{it}'\beta_i \;+\; \varepsilon_{it}, \tag{1}$$

where $\alpha(p_t)$ is the *intrinsic purpose valuation*, the value of purpose p_t absent any social consideration; β_i is respondent i's vector of design-attribute part-worths, whose price component we write $\beta_i^m < 0$; and ε_{it} is an idiosyncratic taste shock with $E[\varepsilon_{it} \mid \mathbf{x}_{it}] = 0$, which holds by randomization.

Reporting. What the conjoint records is not U_{it} but reported interest,

$$r_{it} \;=\; U_{it} \;-\; \lambda_i\,\sigma(p_t, o) \;+\; \eta_{it}, \tag{2}$$

where

- $\sigma(p,o) \geq 0$ is the *signaling cost*: the social-image penalty attached to being known to want a product of purpose p under observability o. It is increasing in o, with $\sigma(p,0) = 0$; increasing in how sharply purpose p discriminates undesirable types; and, as Section 2.5 adds, decreasing in aggregate adoption.
- $\lambda_i \geq 0$ is respondent i's *social-image sensitivity*, the weight placed on the inferred social judgment relative to consumption utility. Write $\bar{\lambda} \equiv E[\lambda_i]$.
- η_{it} is reporting noise, $E[\eta_{it} \mid \mathbf{x}_{it}] = 0$.

The pure preference benchmark is the special case $\sigma \equiv 0$, under which $r_{it} = U_{it} + \eta_{it}$ and reported interest is a sufficient statistic for welfare-relevant utility. The signaling model breaks that equality, and does so *differentially*: the wedge is large where $\sigma(p,\bar{o})$ is large and near zero where it is not.

What the conjoint coefficient measures. Let $\hat{\beta}(p)$ denote the population coefficient on purpose p relative to Admin in a regression of r_{it} on $\mathbf{x}_{it}$. Taking expectations of (2) under randomization,

$$\hat{\beta}(p) \;=\; \underbrace{\big[\alpha(p) - \alpha(\mathsf{Admin})\big]}_{\Delta\alpha_p} \;-\; \bar{\lambda}\;\underbrace{\big[\sigma(p,\bar{o}) - \sigma(\mathsf{Admin},\bar{o})\big]}_{\Delta\sigma_p}. \tag{3}$$

The purpose coefficient is a *composite* of an intrinsic taste difference and a social-image cost difference. Neither component is separately identified from (3) alone. This is the identification problem the rest of the paper addresses, and it is worth stating in its managerial form: a firm that reads $\hat{\beta}(\mathsf{Romantic})$ as a taste parameter will invest in changing the product, when the binding constraint may be the environment in which the product is consumed.

Because $\sigma(\mathsf{Romantic},\bar{o}) \gg \sigma(\mathsf{Admin},\bar{o})$, we obtain our first hypothesis.

Hypothesis 1. The purpose gradient in stated interest satisfies $\mathsf{Romantic} < \mathsf{Friend} < \mathsf{Wellness} \approx \mathsf{Admin} < \mathsf{Study}$, and is steeper than the gradient in intrinsic valuations $\alpha(\cdot)$.

The ordering is directly testable. The second clause is not: because $\alpha(\cdot)$ is not separately identified from (3), it is disciplined only through the restrictions that follow.

2.3 Restriction 1: price and social-image costs do not substitute

Equation (2) is additive, so on the latent index a price reduction shifts reported interest by $\beta^m \Delta m$ regardless of purpose, exactly as in the pure preference model. The two models separate only once the participation margin is taken seriously. Setting $m = 0$ removes the monetary barrier but leaves $\lambda_i \sigma(p_t, \bar{o})$ intact, so a respondent deterred from a romantic companion partly by cost and partly by social image becomes more willing when it is free, but only by the monetary component; if $\lambda_i \sigma$ alone keeps her below the margin, the price cut moves nothing. For an administrative assistant, where $\sigma \approx 0$, the price reduction translates fully into higher stated interest. Formally, let reported interest be $\tilde{r}_{it} = \max\{0, r_{it}\}$, so that the rating scale has a floor. Then $\partial \tilde{r} / \partial m$ shrinks in magnitude as σ rises, i.e. $\partial^2 \tilde{r} / \partial m \, \partial \sigma > 0$.

This prediction is not unique to signaling: any force that lowers the latent index for relational purposes, including intrinsic distaste ($\Delta \alpha_p < 0$), pushes more mass below the floor and attenuates the observed price response in the same way. The attenuation therefore rejects the additive *linear* model, but it discriminates signaling from intrinsic taste only if it survives on the latent index, i.e. in a censored model that removes the floor. Section 5.3 reports both.

Hypothesis 2. The zero-price demand premium is attenuated for relational purposes: the interest gain from \$10/month to free is largest for Admin and Study and smallest for Romantic.

An analogous restriction holds for privacy. If social-image costs operate through observability, a design feature that lowers perceived observability reduces σ *specifically where* σ *is non-zero.* The pure preference model predicts no purpose $\times$ privacy interaction; the signaling model predicts High Privacy $\times$ Romantic > 0.

Hypothesis 3. Privacy protections raise interest differentially for relational purposes: High Privacy $\times$ Romantic > 0, and no instrumental purpose exhibits a positive privacy interaction.

2.4 Restriction 2: the cross-equation restriction on beliefs

Second-order beliefs supply the additional moment that (3) lacks. The logic is that reporting a belief about *others'* demand exposes the respondent to a weaker social-image cost than reporting *own* demand. We do not assume beliefs are uncontaminated, only that they are less contaminated. Formally, let $\mu \geq 0$ be the sensitivity governing belief reporting and assume

$$0 \leq \mu < \bar{\lambda}. \tag{A1}$$

Let $\varphi > 0$ be the *scale conversion factor* mapping the 0–100 belief measure into 0–10 interest units, so that φs_{it} is denominated in interest points. The benchmark is $\varphi = 1/10$; we report

sensitivity over $\varphi \in [0.08, 0.12]$. Beliefs then satisfy

$$\varphi\, s_{it} \;=\; U_{it} \;-\; \mu\, \sigma(p_t, \bar{o}) \;+\; \xi_{it}, \qquad E[\xi_{it} \mid \mathbf{x}_{it}] = 0. \tag{4}$$

Subtracting (2) from (4) and averaging over respondents defines the *belief–interest gap* at purpose p,

$$G(p) \;\equiv\; \varphi\, \bar{s}(p) - \bar{r}(p) \;=\; (\bar{\lambda} - \mu)\, \sigma(p, \bar{o}), \tag{5}$$

and its *Admin-relative* counterpart, which is the object we use throughout,

$$G^*(p) \;\equiv\; G(p) - G(\mathsf{Admin}) \;=\; (\bar{\lambda} - \mu)\, \Delta\sigma_p. \tag{6}$$

Differencing against Admin removes any component of the belief–interest divergence that is common to all purposes (baseline optimism about others' technology enthusiasm, or a fixed artifact of the 0–100 versus 0–10 scale) and is the reason we work with G^* rather than G.

Dividing (6) by the purpose penalty from (3) gives the restriction we test:

$$R(p) \;\equiv\; \frac{G^*(p)}{-\hat{\beta}(p)} \;=\; \frac{(\bar{\lambda} - \mu)\, \Delta\sigma_p}{-\Delta\alpha_p + \bar{\lambda}\, \Delta\sigma_p}. \tag{7}$$

Hypothesis 4. $R(p)$ is constant across relational purposes: $R(\mathsf{Friend}) = R(\mathsf{Romantic})$.

Three remarks on (7) are important and are, to our knowledge, not made carefully elsewhere in the conjoint literature that uses belief elicitation.

(i) The common value of the ratio is not $(\bar{\lambda} - \mu)/\bar{\lambda}$ in general. It equals that only in the knife-edge case $\Delta\alpha_p = 0$, i.e. only if romance and admin have identical intrinsic valuations, which is neither plausible nor something we wish to assume. Since $\Delta\alpha_p \leq 0$ (instrumental purposes are intrinsically preferred), the observed ratio is a *lower bound*:

$$R(p) \;\leq\; \frac{\bar{\lambda} - \mu}{\bar{\lambda}} \qquad \text{whenever } \Delta\alpha_p \leq 0. \tag{8}$$

(ii) What Hypothesis 4 does and does not test. Constancy of $R(p)$ is implied by a single-factor signaling structure. It is also consistent with $\Delta\alpha_p \propto \Delta\sigma_p$, and with beliefs that are simply shrunk toward a common anchor (Section 5.5). The test therefore rejects models in which intrinsic and social gradients move non-proportionally across purposes; it does not, on its own, rule out every alternative without social image. It is nonetheless a genuine over-identifying restriction: it is not imposed in estimation, it involves two different outcome variables measured on different scales, and it could fail in the data.

(iii) The correction is bounded below by construction. Because (6) identifies $(\bar{\lambda} - \mu)\Delta\sigma_p$ while welfare depends on the full suppression $\bar{\lambda}\Delta\sigma_p$, any correction built from G^* recovers only

the fraction $(\bar{\lambda}-\mu)/\bar{\lambda} < 1$ of the individual suppression cost. The full individual correction satisfies

$$\text{correction}^{\text{full}} = \frac{\bar{\lambda}}{\bar{\lambda}-\mu} \cdot \text{correction}^{G^*} \;>\; \text{correction}^{G^*}. \tag{9}$$

2.5 Mechanism 2: endogenous norms and the under-adoption trap

The first mechanism concerns individual behavior given a fixed norm. The second concerns the norm. Stigma is endogenous to adoption: when romantic AI use is rare, it is readily read as revealing unmet social needs; when it is common, the same behavior carries no such inference. Let $A \in [0,1]$ denote aggregate adoption of purpose p and extend the signaling cost to $\sigma(p,o,A)$ with $\partial\sigma/\partial A < 0$. Each adopter confers a positive externality on subsequent potential adopters, and no individual internalizes it.

Fixed-point structure. Adoption is individually rational for i if and only if $U_i - \lambda_i\sigma(p,\bar{o},A) \geq 0$, i.e. if and only if

$$\theta_i \;\equiv\; \frac{U_i}{\lambda_i} \;\geq\; \sigma(p,\bar{o},A), \tag{10}$$

where θ_i is respondent i's *social tolerance*: the social-image cost she is willing to bear per unit of consumption utility. Note that the relevant heterogeneity is in the ratio U_i/λ_i, not in λ_i alone; respondents with $U_i \leq 0$ have $\theta_i \leq 0$ and never adopt, so the mechanism operates on the subpopulation with positive latent utility.[1]

Let $G_\theta(\cdot)$ be the CDF of θ_i in the population. Equilibrium adoption solves the fixed point

$$A^* \;=\; 1 - G_\theta\big(\sigma(p,\bar{o},A^*)\big). \tag{11}$$

Because σ is decreasing in A, the right-hand side is increasing in A, and (11) may admit multiple solutions. A low-adoption and a high-adoption equilibrium can both be stable: in each, prevailing adoption validates the stigma level that sustains it, and neither is escapable by unilateral choice.

Comparative statics. Three follow directly and map onto instruments.

(i) Escape threshold. The basin boundary depends on the shape of G_θ. If most consumers have high social tolerance (G_θ concentrated at large θ), the low equilibrium is fragile and a small coordinated nudge triggers a cascade. If G_θ has a heavy left tail, the low equilibrium is robust.

(ii) Privacy-preserving design substitutes partially for norm change. Reducing effective observability from $\bar{o}$ to $\underline{o}$ lowers σ at every A, shifting the adoption map upward and potentially

[1] In the data the Romantic purpose has the lowest mean interest (≈ 1.86 on 0–10), the largest penalty (-1.262), and 54.9% of ratings at the floor, which suggests a non-trivial share with $U_i \leq 0$ even under full anonymity. The welfare statements below are averages over the subpopulation for whom the social-image cost, not intrinsic disinterest, is the binding constraint.

eliminating the low-adoption equilibrium without any aggregate coordination.

(iii) Price and privacy work through different channels. Price enters U_i but not σ. Setting $m = 0$ raises θ_i for everyone with $U_i > 0$, which also shifts the adoption map upward, so a large enough price cut could in principle carry the economy across the basin boundary through the induced fall in $\sigma(\cdot, A)$. The difference from (ii) is one of efficiency, not possibility: a price cut must raise U_i by enough to outweigh the full stigma cost for marginal adopters, whereas lowering observability attacks σ directly, and the zero-price floor bounds how far price can go. Our data identify only the partial-equilibrium response at the prevailing norm (Hypothesis 2); they cannot tell us whether free pricing would tip the equilibrium.

2.6 Four demand concepts

The framework separates four objects that standard conjoint practice conflates. Keeping them distinct is essential to reading Section 7.

1. **Stated demand** r_{it}: interest as elicited, conditioning on the prevailing norm A^* and observability $\bar{o}$. Distorted downward by $\lambda_i \sigma(p_t, \bar{o}, A^*)$.
2. **Latent individual demand** $U_{it} + \eta_{it}$: interest under fully private adoption ($o \to 0$, so $\sigma = 0$), holding the norm fixed. Directly identified only by a randomized observability design; our G^*-based correction recovers it partially, and from below by (9).
3. **Latent social demand**: interest under a normalized equilibrium A^{**} at which $\sigma(p, \bar{o}, A^{**}) \approx 0$. Its gap from latent individual demand is the *equilibrium trap premium* $\Delta W > 0$, which requires a collective-adoption treatment analogous to the collective-deactivation design of Bursztyn et al. (2025a) and is *not* identified here.
4. **Intrinsic demand** $\alpha(p)$: the purpose valuation absent all social considerations. Identified only under functional-form restrictions we are unwilling to impose; we do not estimate it.

Two consequences govern the welfare analysis. Our correction recovers latent *individual*, not latent *social*, demand, so for a given suppression share π every welfare number we report is a lower bound in two distinct senses, (9) and $\Delta W > 0$. Uncertainty about π itself runs the other way (Section 7). And the two concepts call for different instruments: latent individual demand is recoverable through privacy-preserving design, latent social demand only through norm-shifting or platform coordination.

2.7 Identification: what we assume and what it buys

We state the assumptions in one place because each does specific work and each is a point of attack.

(A0) Randomization. $\mathbf{x}_{it} \perp (\varepsilon_{it}, \eta_{it}, \xi_{it}, \lambda_i)$. Guaranteed by design; delivers (3) and all attribute part-worths.

(A1) Belief attenuation. $0 \leq \mu < \bar{\lambda}$ in (A1). Delivers the sign of the correction and, via (9), its lower-bound property. *Not testable here.* If instead $\mu > \bar{\lambda}$, respondents suppress beliefs more than own interest, the correction has the wrong sign. We regard this as implausible but note it.

(A2) Common baseline divergence. $G(\mathsf{Admin})$ contains only components common to all purposes. Delivers the interpretation of G^* as purpose-specific suppression. Partially testable: the Admin gap should be uncorrelated with purpose-specific covariates, and the $R(p)$ ratios should be stable across subgroups. We report the second in Section 5.7, and Section 5.5 tests the most important alternative source of a purpose-specific gap, belief shrinkage.

(A3) Cardinality and quasilinearity. Ratings are affine in utility with respondent-specific location absorbed by fixed effects, and utility is linear in money over $[0, 10]$ dollars. Delivers the money-metric conversion $-\hat{\beta}_j/\hat{\beta}^m$. *This is the weakest link*, and the zero-price discontinuity we document is direct evidence against linearity in money; we therefore report every dollar figure under three price slopes (Section 7) and a scale-free robustness check using within-respondent standardized ratings.

(A4) Scale conversion. φ maps the belief scale into the interest scale linearly through the origin. Delivers cardinal comparability of s and r. We never rely on a single φ. Because φ scales the belief term but not the interest term in G^*, neither $R(p)$ nor the difference $R(\mathsf{Friend}) - R(\mathsf{Romantic})$ is invariant to it; we therefore report the ratio test, and all welfare numbers, over $\varphi \in [0.08, 0.12]$.

What is not identified. Three things, and we do not claim otherwise. First, $\alpha(p)$ and $\sigma(p, \bar{o})$ are not separately identified without (A1); the conjoint alone recovers only the composite (3). Second, observability o is never randomized, so $\partial\sigma/\partial o$ is identified only through the privacy attribute, which is an imperfect proxy; Hypothesis 3 is accordingly a directional test, not a point estimate of the observability elasticity. Third, because the instrument contains no outside option, the *level* of consumer surplus is not identified, only differences in willingness to pay across profiles. Section 7 respects this.

3 Conjoint Survey Design

The within-respondent structure and randomized attributes identify the purpose gradient (Hypothesis 1) and zero-price attenuation (Hypothesis 2); the SOB elicitation tests the cross-equation restriction (Hypothesis 4); and the obfuscated follow-up checks whether the purpose gradient is

Table 1: Notation

Symbol	Definition
Choice environment	
i, t	respondent; profile ($t = 1, \ldots, 5$, one per purpose)
$p \in \mathscr{P}$	purpose: Admin, Study, Wellness, Friend, Romantic
$\mathbf{d}_{it} = (m, \iota, \kappa, \rho, \varsigma)$	design attributes: price, modality, personalization, privacy, safety
$m_{it} \in \{0, 5, 10\}$	monthly price in dollars (enters numerically)
$\mathbf{x}_{it} \in \mathbb{R}^{14}$	full design vector (intercept, 4 purpose dummies, price, 8 attribute dummies)
$o \in [0,1]$, $\bar{o}$	observability of adoption; prevailing regime
$A \in [0,1]$, A^*, A^{**}	aggregate adoption; equilibrium; normalized counterfactual
Outcomes	
$r_{it} \in [0, 10]$	stated own interest
$s_{it} \in [0, 100]$	second-order belief (out of 100 similar people)
Preference parameters	
U_{it}	latent, welfare-relevant utility, eq. (1)
$\alpha(p)$	intrinsic purpose valuation (not separately identified)
β_i, β_i^m	individual attribute part-worths; price part-worth
$\Delta\alpha_p$, $\Delta\sigma_p$	Admin-relative intrinsic and signaling-cost differences
Social-image parameters	
$\sigma(p, o, A) \geq 0$	signaling cost of purpose p
$\lambda_i \geq 0$, $\bar{\lambda}$	individual social-image sensitivity; population mean
$\mu \geq 0$	sensitivity governing belief reporting; (A1): $\mu < \bar{\lambda}$
$\theta_i = U_i/\lambda_i$	social tolerance; $G_\theta(\cdot)$ its CDF
$\hat{\lambda}_i$	empirical proxy for λ_i, eq. (16)
Derived and auxiliary	
$\hat{\beta}(p)$	estimated Admin-relative purpose coefficient, eq. (3)
$G(p)$, $G^*(p)$	belief–interest gap; Admin-relative gap, eqs. (5)–(6)
$R(p) = G^*(p)/[-\hat{\beta}(p)]$	cross-equation ratio, eq. (7)
φ	scale conversion, 0–100 belief $\to$ 0–10 interest; benchmark $1/10$
$\pi \in [0,1]$	share of G^* attributable to genuine suppression
$\Delta W > 0$	unidentified equilibrium trap premium
Structural model (Section 6)	
$\mathbf{z}_i \in \mathbb{R}^{n_z}$	respondent-level covariates in the population regression
Δ $(n_z \times K)$	population regression coefficients
$\mathbf{V}_\beta$ $(K \times K)$	covariance of individual taste deviations
σ_i^2	individual-level error variance

specific to the conjoint instrument.

3.1 Sample

Data come from an online conjoint experiment of a broadly representative U.S. sample recruited via Prolific (Bursztyn et al., 2023; Enke et al., 2023). Following the preregistered protocol (AsPredicted), we excluded respondents failing attention or LLM-usage checks and those completing the survey unrealistically quickly. Of 1,993 completed responses, three fail the attention check and one is a Qualtrics preview record, leaving a final sample of $N = 1{,}989$ respondents and 9,945 profile-level observations. Every estimate in the paper uses this sample and is produced by a single replication script.

The sample broadly reflects the U.S. population on age, region, ethnicity, and gender (Appendix Table A.3). As typical of Prolific, respondents are slightly younger, more educated, and higher-income than the general population, but closely match the U.S. on relationship status and AI use frequency (YouGov, 2025).[2] The sample over-represents respondents familiar with AI companions.[3]

Statistical power. Each respondent evaluates five profiles, yielding 9,945 profile-level observations, roughly 3,300 per level of a balanced three-level attribute. The respondent-clustered standard errors on attribute contrasts are 0.050–0.057 (Table A.6), so a two-sided test at $\alpha = 0.05$ has 80% power against effects of about $2.8 \times 0.053 \approx 0.15$ interest points. The smallest attribute effects we report (standard privacy, -0.124; basic personalization, -0.160) sit close to this threshold, which is why we also report multiple-testing adjustments (Table A.9). In the smallest subgroup used in Section 5.7 ($N = 543$ companion users), standard errors roughly double and the minimum detectable effect is about 0.35–0.40 points; purpose $\times$ attribute interactions, which split each purpose further, are powered only for effects of about 0.45 points.

3.2 Conjoint profiles and randomized attributes

Following Hainmueller et al. (2015), we created five hypothetical AI app profiles corresponding to five purposes: *Admin*, *Friend*, *Romantic*, *Study*, and *Wellness*, with order randomized across respondents. The purposes span *instrumental* uses (directed at external task completion) and *relational* uses (directed at social and emotional experience) (Ruggiero, 2000; Brandtzaeg et al., 2022). Wellness is classified as intermediate: it addresses personal well-being but through informational and coaching mechanisms closer to instrumental help than social bonding (Ta et al., 2020; Laestadius et al., 2024).

[2] PEW Research, 2025.
[3] NORC, 2025.

Table 2: Conjoint attributes and randomized levels

Attribute	Levels (randomized)
Monthly price	$10/month; $5/month; Free
Interaction modality	Text only; Text + Voice; Text + Voice + optional photo sharing
Personalization depth	Basic; Moderate; Advanced
Privacy protections	Limited; Standard; High
Safety policy	Strict (PG-13); Balanced; Adults-only

Notes: Levels independently randomized within each profile. Baseline (omitted) levels differ by specification; see table notes.

Five attributes were independently randomized within each profile: monthly price, interaction modality, personalization depth, privacy protections, and safety policy (Table 2).[4] Any residual ambiguity in how respondents interpret the attribute levels would most likely attenuate attribute coefficients toward zero, so the attribute magnitudes are, if anything, conservative.

The vignette template was:

*Meet an AI companion designed for **{Purpose}**. You can interact via **{Interaction}**. It offers **{Personalization}** personalization and follows **{Safety}** safety/content rules. Its privacy policy is: **{Privacy}**. Price: **{Price}**.*

The design yielded 1,215 unique profiles from all combinations of $3^5 \times 5$ attributes and purposes, with each level appearing in roughly one-third of profiles (Table A.4).

3.3 Outcomes

Respondents reported two outcomes per profile: (i) own interest in trying the app (0–10), our main outcome for testing Hypotheses 1 and 2; and (ii) second-order beliefs about demand, elicited as the number out of 100 similar people who would want to try it (0–100), our primary instrument for Hypothesis 4.

To discourage inattention, participants were told *ex ante* that they would receive a personalized AI app recommendation based on their ratings, delivered via Prolific after completion. This partial incentivization, in which responses determine personalized information received, reduces disengaged responding without inducing the strategic distortions that monetary incentive schemes can produce (Kessler et al., 2019; Cullen et al., 2023; Kessler et al., 2024; Low, 2014; Macchi, 2023; Colonnelli et al., 2024; Exley et al., 2024). We do not claim full incentive compatibility: respondents already certain of their preferences derive little marginal value from the recommendation.

[4]Independent randomization of personalization depth and privacy protections is by design: it identifies each attribute's separate marginal value, which is the policy-relevant quantity for platforms choosing design dimensions independently.

The interest rating measures *stated purchase intent* (Morwitz, 2008), which can diverge from realized adoption due to hypothetical bias and social desirability. We therefore treat estimates as capturing the *preference structure* over purposes and attributes rather than predicting market shares. Two behavioral benchmarks support this framing: Google Trends search intensity in 2025 shows the same instrumental–relational separation (Section 5.2), and Chatterji et al. (2025) document a usage distribution dominated by instrumental tasks from anonymized ChatGPT sessions. These benchmarks validate the ordinal structure of demand. They cannot adjudicate how much of the relational penalty is social-image suppression, because behavioral data are themselves shaped by the same stigma and by product positioning; that question is taken up in Sections 5.5 and 7.

3.4 Obfuscated Follow-Up

The signaling framework predicts that self-presentation distortions are largest for relational purposes. The obfuscated follow-up is designed to test whether the conjoint purpose gradient is an artifact of the specific conjoint framing, or whether it replicates under a qualitatively different elicitation instrument. Critically, this is a test of *instrument-invariance under constant suppression*: if social-image costs operate similarly across both instruments, convergence rules out conjoint-specific framing effects but cannot rule out suppression in both instruments simultaneously. We return to this distinction in Section 5.7, where we use cross-respondent heterogeneity to sharpen the test.

Three to four weeks after the main survey (which ran 24–29 November 2025; the follow-up data were collected on 21 December 2025), 1,082 respondents completed an obfuscated follow-up with an altered interface, question order, and screen count to make the two surveys appear unrelated; 1,071 of them (53.8% of the analysis sample) match to a main-survey record on the self-reported panel ID. Respondents ranked the five purposes from 1 (least favorite) to 5 (most favorite), a qualitatively different task (ranking vs. rating) that limits the scope for shared method variance to explain any convergence. All 1,082 submitted rankings are valid permutations of $\{1,\ldots,5\}$. Because the two instruments differ in format and context but plausibly share exposure to similar social-image costs, convergence of purpose rankings speaks to whether the conjoint gradient is an artifact of conjoint-specific framing; it cannot distinguish latent preferences from suppression that operates in both instruments. That question is addressed instead through the SOB-based welfare correction in Section 7 and the heterogeneity test in Section 5.9.

4 Estimation

Section 2.7 set out the identifying assumptions. This section describes the estimators that map them into the four hypotheses.

Step 1 (Hypothesis 1): the purpose gradient. We estimate rating-based conjoint models (Hainmueller et al., 2014). The pooled specification is

$$r_{it} = \gamma + \mathbf{x}_{it}'\beta + \varepsilon_{it}, \tag{12}$$

with standard errors clustered at the respondent level. The respondent fixed-effects specification is

$$r_{it} = a_i + \mathbf{x}_{it}'\beta^* + u_{it}, \tag{13}$$

which absorbs respondent-specific baseline enthusiasm and supports reading each coefficient as a within-person part-worth utility. Because attributes are randomized within respondent, (12) and (13) should be, and are, nearly identical; we report (13) throughout. Under (A0), $\hat{\beta}^*$ recovers the composite in (3).

We use single-profile ratings rather than paired or menu-based choices because we require cardinal valuations for attributes with no observable market price: a binary selection probability cannot be converted into dollars without additional functional-form assumptions (Rao, 2014). The cost of this choice is that the instrument has no outside option, so participation levels and the *level* of consumer surplus are not identified. We treat this as a design limitation rather than an inference to be finessed (Sections 2.7 and 7), and it is the first thing we would change in a follow-up design.

The fixed-effects estimator addresses interpersonal incomparability in ratings by purging stable individual-level *location*. It does not purge individual-level *scale*. Since our social-image proxy $\hat{\lambda}_i$ is a difference of two instruments measured on different scales, scale heterogeneity is a live confound for every individual-level result. We therefore re-run all individual-level analyses on within-respondent standardized ratings (z-scored across each respondent's five profiles) as a scale-free check (Table 3). The money-metric estimates are essentially unchanged: $-\$7.71$ for Romantic and $-\$3.20$ for Friend on z-scored ratings, against $-\$7.80$ and $-\$3.48$ on the raw scale. Individual scale heterogeneity is therefore not driving the results.

Step 2 (Hypothesis 2): zero-price attenuation. We estimate purpose-specific price effects using the `cjbart` estimator of Duch and Robinson (2023), which recovers heterogeneous treatment effects via Bayesian Additive Regression Trees without imposing linearity or additivity, the appropriate tool here precisely because the null we are testing *is* additivity. Under the pure preference model the \$5 $\rightarrow$ free boost is constant across purposes; `cjbart` tests whether it is instead systematically smaller where σ is larger.

Step 3 (Hypothesis 3): the privacy interaction. We augment (13) with the full set of privacy $\times$ purpose interactions. Because the model makes a signed directional prediction for a single cell, we report one-sided tests for that cell and two-sided tests elsewhere, and we treat the result

Table 3: Scale-free robustness: within-respondent z-scored ratings

	Coefficient (z)	WTP, z-scale	WTP, raw scale
Study	0.105 (0.031)	+1.22	+0.99
Wellness	−0.023 (0.031)	−0.27	−0.50
Friend	−0.274 (0.030)	−3.20	−3.48
Romantic	−0.660 (0.030)	−7.71	−7.80
Price ($/month)	−0.086 (0.003)	—	—
Limited retention	−0.114 (0.027)	−1.33	−1.24
Basic personalization	−0.089 (0.027)	−1.03	−1.04
Balanced safety	−0.097 (0.027)	−1.13	−1.14

Notes: Dependent variable is each respondent's interest rating z-scored across their own five profiles, which purges individual-specific scale as well as location. Respondent fixed effects; respondent-clustered standard errors. Restricted to the 1,609 respondents with non-degenerate rating vectors. The raw-scale column is from the full-sample numeric-price model.

as directional evidence rather than a point estimate of $\partial\sigma/\partial o$.

Step 4 (money-metric conversion). We re-estimate (13) with numeric price $m_{it} \in \{0, 5, 10\}$,

$$r_{it} = a_i + \beta^m m_{it} + \sum_k \beta_k x_{kit} + e_{it}, \tag{14}$$

and define willingness to pay for attribute level j as the marginal rate of substitution between that attribute and money,

$$\widehat{\mathrm{WTP}}_j = -\frac{\hat{\beta}_j}{\hat{\beta}^m}, \tag{15}$$

with delta-method confidence intervals computed from the respondent-clustered covariance (Rao, 2014; Green and Srinivasan, 1990; Louviere et al., 2000). The price slope is large and precisely estimated ($\hat{\beta}^m = -0.161$), ensuring numerical stability of the ratio.

Equation (15) rests on (A3), and (A3) is under visible strain: the interest response to price is markedly nonlinear, with the \$5 → free gain roughly four times the \$10 → \$5 gain. The average slope $\hat{\beta}^m$ therefore averages a locally steep response near zero with a locally flat one at higher prices, and $\widehat{\mathrm{WTP}}_j$ inherits whichever is the relevant margin. We consequently report every dollar figure under three mappings (the average slope, the local \$10 → \$5 slope, and the local \$5 → free slope) and state which margin each is appropriate for. The local \$5 → free slope is the right one when the counterfactual is a freely available product, which is the empirically relevant case for consumer AI.

Step 5 (Hypothesis 4): the cross-equation restriction. We compute the Admin-relative belief–interest gap $G^*(p)$ from (6) and the ratio $R(p) = G^*(p)/[-\hat{\beta}(p)]$ from (7) for the two relational purposes, and test $R(\mathsf{Friend}) = R(\mathsf{Romantic})$ by a nonparametric bootstrap over respondents

(2,000 replications) that propagates sampling error in both the numerator and the denominator. The test is invariant to the price slope and to any common rescaling of both instruments, but not to φ, which scales only the belief term in $G^*(p)$. We therefore report it across $\varphi \in [0.08, 0.12]$.

Step 6 (individual social-image sensitivity). We construct a respondent-level proxy for λ_i as the Admin-relative belief–interest gap averaged over relational purposes,

$$\hat{\lambda}_i = \frac{1}{2} \sum_{p \in \{\mathsf{Rom},\mathsf{Friend}\}} \Big[\big(\varphi s_{ip} - r_{ip}\big) - \big(\varphi s_{i,\mathsf{Admin}} - r_{i,\mathsf{Admin}}\big) \Big]. \tag{16}$$

Two points about (16) deserve emphasis, because they are easy to get wrong. First, $\hat{\lambda}_i$ estimates the *product* $(\lambda_i - \mu)\overline{\Delta\sigma}$, not λ_i: it is a scaled sensitivity, in interest points, and its units are not comparable to λ_i in (2). Second, $\hat{\lambda}_i$ contains $-r_{ip}$ for exactly the relational purposes whose coefficients we later regress on it. Any specification that uses $\hat{\lambda}_i$ to explain relational-purpose tastes is therefore *mechanically* predisposed to find a negative loading. We address this directly in Section 6.5 by building $\hat{\lambda}_i$ from one relational purpose, differenced against Study rather than Admin, and testing its loading on the other, so that no shared r_{ip} term appears on both sides. The full-sample version is reported only as a consistency check.

5 Results

Two features bear on interpreting conjoint estimates as preference measures. First, our primary claims concern *relative* valuations, the preference structure shaping adoption, rather than absolute interest levels or market share predictions. Relative valuations are robust to uniform hypothetical bias, which affects all purposes equally. Second, two behavioral benchmarks corroborate the ordinal structure before the mechanism tests: Google Trends search intensity in 2025 separates instrumental from relational purposes in the same way (Section 5.2), and Chatterji et al. (2025) document a usage mix dominated by instrumental tasks.

5.1 Purpose gradient

Figure 1 shows the highest interest for instrumental applications: *Admin* (≈ 3.13), *Study* (≈ 3.28), *Wellness* (≈ 3.07); and substantially lower interest for relational applications: *Friend* (≈ 2.56) and *Romantic* (≈ 1.86).

Figure 2 and Appendix Table A.6 report conjoint estimates. Pooled OLS and FE estimates are nearly identical, as expected under within-respondent randomization. Coefficients in this subsection come from the specification with price-level indicators; the money-metric results from Section 5.8 onward use numeric price, whose other coefficients differ only in the third

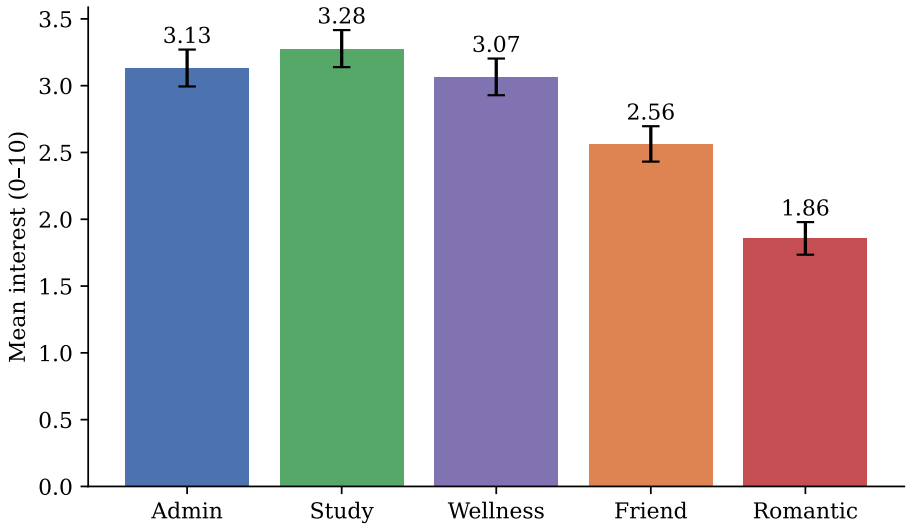


Figure 1: Average interest by companion purpose. Bars show means; whiskers show 95% confidence intervals.

decimal (e.g. -1.260 against -1.262 for Romantic, -0.562 against -0.572 for Friend).

Effect of App Purpose. Relative to *Admin*, *Friend* reduces interest by 0.572 points (s.e. 0.059; 95% CI $[-0.688, -0.456]$) and *Romantic* by 1.262 points (s.e. 0.060; 95% CI $[-1.380, -1.144]$). *Study* increases interest by 0.153 points (s.e. 0.058; 95% CI $[0.039, 0.267]$) and *Wellness* is indistinguishable from zero (coef. -0.093, s.e. 0.057; 95% CI $[-0.205, 0.019]$). The Romantic penalty corresponds to roughly two-fifths of the pooled baseline interest (intercept $= 2.879$), confirming Hypothesis 1.

Effect of Price. The \$10→\$5 gain is 0.321 points (s.e. 0.050); the \$5→Free gain is 1.291 points, roughly four times larger (Appendix Fig. B.2), consistent with zero-price effects (Shampanier et al., 2007).

Other attributes. Relative to high privacy, standard privacy reduces interest by 0.124 points (s.e. 0.050; 95% CI $[-0.222, -0.026]$) and limited retention by 0.206 (s.e. 0.052; 95% CI $[-0.308, -0.104]$). Relative to advanced personalization, basic personalization reduces interest by 0.160 (s.e. 0.052; 95% CI $[-0.262, -0.058]$). Relative to the adults-only safety baseline, balanced safety reduces interest by 0.179 (s.e. 0.052; 95% CI $[-0.281, -0.077]$). Interaction modality effects are near zero (text only: -0.046, s.e. 0.053; 95% CI $[-0.150, 0.058]$).

Magnitude in context. Scaled by the standard deviation of ratings (3.09), the Romantic penalty is $d \approx 0.41$ and the Friend penalty $d \approx 0.19$. The free-versus-\$10 contrast ($+1.612$ points) is $d \approx 0.52$, the only design effect of comparable size. Non-price attribute effects are $d \approx 0.04$–

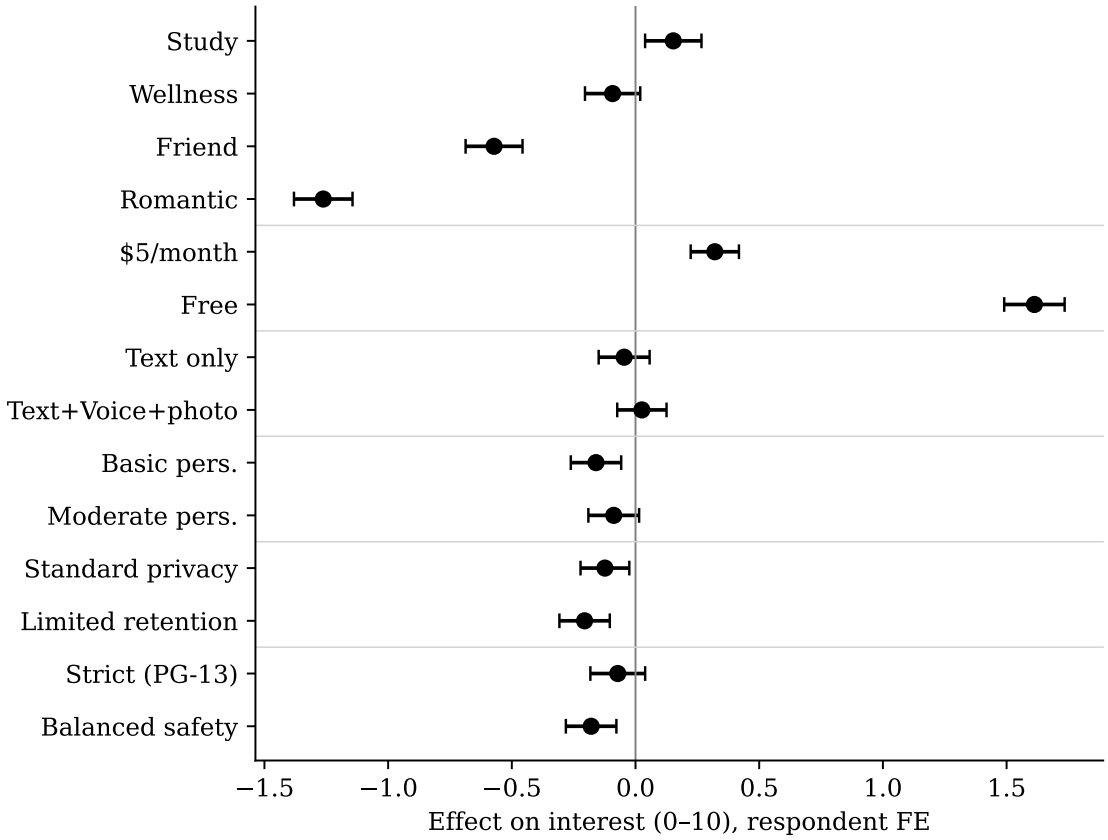


Figure 2: Conjoint part-worth utilities (respondent fixed effects). Points show coefficients; whiskers show 95% confidence intervals.

0.07: small, but those for privacy, personalization, and safety survive false-discovery control (Appendix Table A.9).

Censoring at the floor. A large share of ratings sits at the bottom of the scale: 54.9% for Romantic, 41.0% for Friend, and 30–33% for the three instrumental purposes (the top of the scale is used by under 4% throughout). The linear estimates above are therefore attenuated relative to the latent index, and uniformly so in the direction that understates the purpose gradient. A two-limit Tobit on $[0, 10]$ recovers latent penalties of -2.345 for Romantic and -0.921 for Friend, 1.6–1.9$\times$ the linear estimates, with a price slope of -0.233 (Section 6.6). Every sign and the full ordinal structure are preserved. Mass at the floor is consistent with, though not specific to, a signaling model with heterogeneous λ_i, and it connects to the respondents who give identical ratings to all five profiles discussed in Section 5.9.

5.2 External validity: Google Trends (2025)

Google Trends reports relative search interest (0–100), normalized by geography and time (Google; Google News Initiative). We downloaded weekly U.S. search interest for 2025 using five purpose-matched queries: "AI Assistant" (Admin), "AI study" (Study), "AI Therapy" (Wellness), "AI friend" (Friend), and "AI romantic partner" (Romantic). Values reported as "$<$1" are coded as 0.5.

Figure B.5 shows a stable ordering throughout 2025: instrumental categories dominate, relational categories remain far lower, and the romantic-partner query is near zero most weeks. Average search interest (Figure B.6) is highest for Admin (≈ 23.9) and Study (≈ 22.8), followed by Wellness (≈ 11.7), Friend (≈ 4.2), and Romantic (≈ 0.1).

The instrumental purposes lead and the relational purposes trail in both sources; the one ordinal difference is that Admin edges out Study in Trends, whereas Study edges out Admin in the conjoint. We do not compare magnitudes. Trends indices are normalized to the peak query and depend heavily on which phrase stands in for each purpose ("AI romantic partner" is a much rarer formulation than, say, "AI girlfriend"), so the size of the Trends gap says little about the size of the conjoint gap.

Two further limitations bound interpretive scope. Search volume captures attention rather than purchase intent, and social-image concerns may depress search for relational products in the same direction as survey responses. Chatterji et al. (2025) provide a complementary benchmark that does not rely on self-report: ChatGPT usage from hundreds of millions of sessions is dominated by instrumental tasks, with relational and companionship uses a small share. Because ChatGPT is positioned as a general-purpose assistant rather than a companion, its usage mix reflects positioning as well as demand. The three sources therefore agree on the ordinal instrumental–relational split, which is the only claim we rest on them.

5.3 Zero-price premium and identity costs

Hypothesis 2 states that the zero-price demand premium should be attenuated for high-signaling-cost purposes.

Pattern. We estimate the interaction directly, adding purpose × Free and purpose × \$5 terms to (13) with \$10/month as the omitted price level (Table 4). The purpose × Free interactions are large and precisely estimated: -1.026 for Romantic (s.e. 0.162; $t = -6.32$) and -0.482 for Friend (0.152; $t = -3.17$), against -0.036 for Study and -0.224 for Wellness, neither significant. The implied free-versus-\$10 premium falls monotonically along the purpose gradient: $+1.964$ points for Admin, $+1.929$ for Study, $+1.741$ for Wellness, $+1.483$ for Friend and $+0.938$ for Romantic.

Every purpose × \$5 interaction is indistinguishable from zero (largest $|t| = 0.95$). We do not read this as evidence that the attenuation is confined to the zero-price margin. If the \$5 premium were attenuated in the same proportion as the free premium (the romantic free premium is 48% of the Admin premium), the Romantic × \$5 interaction would be about -0.21, well inside the confidence interval of the estimate (-0.104, s.e. 0.146). The design cannot distinguish a margin-specific from a proportional attenuation. The `cjbart` estimates in Figure 3 reproduce the free-price pattern without imposing linearity or additivity.

Floor censoring. As Section 2 notes, attenuation on the observed scale is also what one expects mechanically when more than half of romantic ratings sit at zero: a price cut cannot raise a rating that is still below the floor, whether the latent index is low because of social image or because of intrinsic distaste. The discriminating test is the same interaction on the latent index, which Table 5 reports. For Romantic the attenuation survives: in a two-limit Tobit the Romantic × Free interaction is -1.203 (s.e. 0.390; $t = -3.1$), and the latent free premium for a romantic companion is 54% of the Admin premium (s.e. 12.6 points), essentially the same proportion as on the observed scale (48%). @@CENSINT@@ For Friend it does not: the latent interaction is -0.232 (s.e. 0.371), so the observed-scale attenuation for friendship companions cannot be distinguished from a floor effect. We therefore read Hypothesis 2 as supported for the purpose with the largest signaling cost and unresolved for the intermediate one.

What this rules out. Under the additive pure preference model,

$$r_{it} = \gamma + \beta_{\text{purpose}}(p_t) + \beta_{\text{price}}(\text{price}_{it}) + \mathbf{x}'_{it}\beta + \varepsilon_{it},$$

the price effect $\beta_{\text{price}} \cdot \Delta\text{price}$ is constant across purposes by construction. The data reject this: the free-price boost is 1.03 points larger for Admin than for Romantic ($t = 6.32$). The linear additive model cannot accommodate this without a purpose × price interaction. For Romantic

Table 4: Purpose × price interactions (respondent fixed effects)

	× Free	× $5/month
Free / $5 main effect (Admin)	+1.964 (0.115)	+0.407 (0.102)
× Study	−0.036 (0.159)	−0.138 (0.153)
× Wellness	−0.224 (0.153)	−0.060 (0.142)
× Friend	−0.482 (0.152)	−0.130 (0.136)
× Romantic	−1.026 (0.162)	−0.104 (0.146)
Implied premium relative to $10/month		
Admin	+1.964	+0.407
Study	+1.929	+0.269
Wellness	+1.741	+0.347
Friend	+1.483	+0.277
Romantic	+0.938	+0.302

Notes: Respondent fixed effects; all non-price attributes included; omitted price level $10/month; omitted purpose Admin. Respondent-clustered standard errors in parentheses. $N = 9{,}945$ profiles from 1,989 respondents. Every purpose × $5 interaction is indistinguishable from zero (largest $|t| = 0.95$).

Table 5: Purpose × price interactions on the observed scale and on the latent index

	Linear FE	Two-limit Tobit
Free (Admin)	1.964 (0.115)	2.623 (0.256)
$5/month (Admin)	0.407 (0.102)	0.467 (0.259)
Study × Free	−0.036 (0.159)	−0.186 (0.347)
Wellness × Free	−0.224 (0.153)	−0.038 (0.361)
Friend × Free	−0.482 (0.152)	−0.232 (0.371)
Romantic × Free	−1.026 (0.162)	−1.203 (0.390)
Study × $5	−0.138 (0.153)	−0.127 (0.362)
Wellness × $5	−0.060 (0.142)	0.202 (0.368)
Friend × $5	−0.130 (0.136)	0.049 (0.371)
Romantic × $5	−0.104 (0.146)	−0.061 (0.385)
Romantic free premium / Admin free premium	0.477	0.541 (0.126)

Notes: Column 1 repeats Table 4 (respondent FE). Column 2 is a pooled two-limit Tobit on $[0, 10]$ with the same regressors, respondent-clustered sandwich standard errors. The Tobit removes the floor, so an interaction that survives there is not a mechanical consequence of censoring.

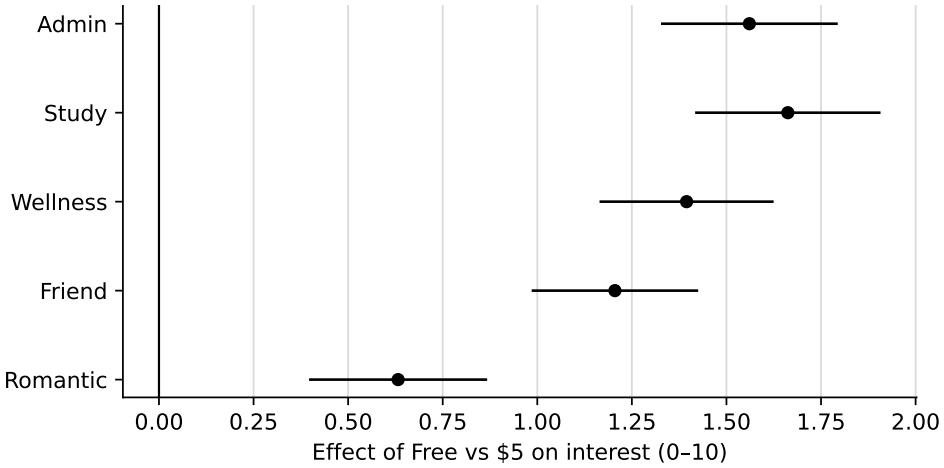


Figure 3: Heterogeneous treatment effect of the $5-to-Free price change by purpose, from the `cjbart` estimator, which imposes neither linearity nor additivity. Points show purpose-specific effects; whiskers show 95% confidence intervals. The pattern reproduces the parametric interactions in Table 4.

the interaction also survives on the latent index, so it is not a mechanical consequence of the rating floor.

A second inconsistency comes from a separate estimating equation. Under pure preferences, respondents' beliefs about others should track their own interest across purposes. Instead, the average respondent rates her own interest in a romantic companion at 1.86 out of 10 yet believes 34 of 100 similar people would try it, and the belief–interest gap is largest precisely where the purpose penalty is largest (Section 5.5). The two patterns, zero-price attenuation and SOB compression, use different outcome variables and sources of variation, and both are consistent with a non-monetary component of the purpose penalty.

Money-metric magnitude. Under the local $5→Free slope, the implied discount for accepting Romantic over Admin is −$4.88/month; under $10→$5, it is −$19.65/month (Table A.7). The penalty is large under every mapping. Product strategies for relational AI therefore require interventions addressing identity and stigma, positioning, social normalization, or low-observability design, not pricing architecture alone.

5.4 Privacy, observability, and the signaling mechanism

The signaling model predicts High Privacy × Romantic > 0 because privacy reduces effective observability o_{it}, lowering the signaling cost specifically where it is non-zero. No such interaction is predicted under the pure preference model, where the purpose penalty is a fixed parameter.

Augmenting the baseline FE specification with full privacy × purpose interactions (Table A.5), the Limited Retention × Romantic coefficient is −0.208 (s.e. 0.160; one-sided

$p = 0.10$), equivalently High Privacy $\times$ Romantic $= +0.208$ against a limited-retention baseline. It is the largest such interaction among the four non-baseline purposes and carries the predicted sign; the Friend interaction is also negative (-0.049), while the two instrumental purposes carry the opposite sign ($+0.075$ for Study, $+0.127$ for Wellness). None is statistically distinguishable from zero. We therefore treat Hypothesis 3 as directionally supported and statistically unresolved. The design is not well powered for this test: each purpose $\times$ privacy cell contains roughly 660 observations, the interaction standard errors are about 0.16, and the minimum detectable interaction is therefore about 0.45 points, more than twice the largest privacy main effect. A randomized observability treatment would be the appropriate instrument.

5.5 Second-order beliefs and the cross-equation restriction

Table A.8 reports mean own interest (0–10) and mean beliefs about others' interest (0–100) for each purpose.

Although the two measures operate on different scales, own interest is a preference-intensity rating (0–10) while SOBs are a frequency estimate (0–100), direct subtraction would impose an untested linear equivalence assumption. We therefore work with Admin-differenced gaps, which remove any fixed offset between the two scales, and report every quantity that depends on the scale conversion over $\varphi \in [0.08, 0.12]$ around the $/10$ benchmark.

Own interest spans Study (3.28) to Romantic (1.86), a spread of 1.42 points. Beliefs span Study (42.44) to Romantic (34.46), a spread of only 7.98 points on the 0–100 scale. The gradient is considerably flatter in beliefs than in own interest. For Romantic companions, own interest is the lowest of all five purposes yet the corresponding belief, that roughly 34 of 100 similar people would try the app, remains relatively elevated, producing the starkest private-vs-social-demand divergence. This pattern is consistent with social-image concerns and norm uncertainty (Bursztyn et al., 2025b; Reif, 2025).

Table A.12 reports the proportionality ratios and a bootstrap test. The ratio for Friend is $G^*(\text{Friend})/[-\hat{\beta}(\text{Friend})] = 0.262/0.561 = 0.466$ (bootstrap s.e. 0.067; 95% CI $[0.334, 0.595]$); for Romantic it is $0.498/1.260 = 0.395$ (0.033; $[0.328, 0.457]$). The difference is 0.071 (bootstrap s.e. 0.058; $Z = 1.23$; $p = 0.217$; 95% CI $[-0.043, 0.179]$), from 2,000 respondent-level replications that re-estimate the fixed-effects denominator in each draw. We cannot reject equal ratios, which is what a single stigma parameter governing both the purpose penalties and the compression of beliefs implies. The conclusion does not depend on the scale conversion: the difference is 0.057 ($p = 0.226$) at $\varphi = 0.08$ and 0.085 ($p = 0.217$) at $\varphi = 0.12$.

Social image or belief shrinkage? Constant $R(p)$ is also what one obtains, with no social image at all, if beliefs about others are shrunk toward a purpose-independent anchor: if $\varphi s_{ip} = w\, r_{ip} + (1-w)\, c_i$ with $0 < w < 1$, then $G^*(p) = (1-w)(-\Delta r_p)$ and $R(p) \approx 1 - w$ for every

purpose. Compression toward the middle of the scale is a familiar feature of frequency estimates, so we test this alternative directly (Table 6).

The sharpest test uses the randomized attributes. Under shrinkage, beliefs absorb the same fraction w of *any* shift in own interest, whatever its source. Under the signaling model, a shift in intrinsic utility (which is what price, privacy, or personalization produce) passes through to beliefs about others, while the purpose-specific social-image component does not. The data side clearly with the second account. Moving from \$10 to free raises φs by 0.90 times as much as it raises own interest (95% CI $[0.86, 0.96]$), and the non-purpose attributes jointly pass through at the same rate (0.90). The relational purpose penalties pass through at only 0.59 ($[0.51, 0.66]$). The difference, 0.32 (s.e. 0.045), is what compression specific to relational purposes looks like; uniform shrinkage cannot generate it.

The instrumental purposes add a second, more more detailed, piece of evidence. Our baseline model assumes $\sigma(\text{Study}) \approx \sigma(\text{Wellness}) \approx \sigma(\text{Admin})$ and hence $G^* \approx 0$ for both. It does not hold exactly: $G^*(\text{Study}) = -0.125$ (s.e. 0.047) and $G^*(\text{Wellness}) = -0.092$ (0.045), jointly different from zero ($p = 0.021$). Shrinkage cannot account for this either: it predicts a small negative gap for Study (-0.060), which respondents like slightly more than Admin, but a small *positive* gap for Wellness ($+0.027$), which they like slightly less; the observed Wellness gap has the opposite sign ($p < 0.001$ against the shrinkage prediction). Both negative gaps are what a *signed* social-image term implies: stated interest in using AI to study or to look after one's health is mildly inflated relative to beliefs about others, i.e. these uses carry a small social-image *premium* relative to administrative assistance. The model accommodates this by letting $\Delta\sigma_p < 0$ for the two instrumental purposes; it leaves the relational tests and the welfare statistic unchanged, because both are defined relative to Admin.

Finally, we can remove whatever shrinkage there is. Subtracting the part of the relational gap that pass-through at the price rate would generate gives shrinkage-adjusted ratios of $R^{\text{adj}}(\text{Friend}) = 0.36$ and $R^{\text{adj}}(\text{Romantic}) = 0.30$, still statistically equal ($p = 0.27$). These are the conservative versions of the welfare statistic used in Section 7.

It is worth noting explicitly what these ratios are, because they reappear as the paper's welfare statistic. By (27), $R(p)$ is exactly the share of the measured purpose penalty that social-image suppression accounts for, at $\pi = 1$. The mechanism test and the money-metric correction are therefore the same number. Because $R(p)$ is a ratio of two Admin-relative differences, it does not involve the price slope, though it does depend on φ. Section 7 develops the implication.

The absolute belief–interest gap for Admin (1.092 points on the 0–10 scale) is itself non-trivial, despite the model predicting $\sigma(\text{Admin}, \bar{o}) \approx 0$. Three interpretations are consistent with this finding. First, a universal baseline optimism: respondents systematically believe others are more enthusiastic about new technology than themselves, regardless of purpose. Second, a mild social-image cost even for instrumental AI use, consistent with evidence on algorithm

Table 6: Distinguishing social-image suppression from belief shrinkage

	Estimate	95% CI
Panel A: Pass-through of own-interest shifts into beliefs ($\varphi\,\partial s/\partial r$)		
Randomized price (free vs. \$10)	0.903	[0.858, 0.961]
All randomized non-purpose attributes	0.897	[0.848, 0.951]
Purpose: Friend vs. Admin	0.543	[0.418, 0.680]
Purpose: Romantic vs. Admin	0.606	[0.541, 0.667]
Purpose: relational pooled	0.586	[0.513, 0.664]
Difference, price minus relational	0.317	[0.230, 0.408]
Panel B: Admin-relative gaps for instrumental purposes		
$G^*(\mathsf{Study})$, observed	-0.125	$[-0.217, -0.033]$
prediction under shrinkage	-0.060	
$G^*(\mathsf{Wellness})$, observed	-0.092	$[-0.180, -0.004]$
prediction under shrinkage	0.027	
Joint test $G^*(\mathsf{Study}) = G^*(\mathsf{Wellness}) = 0$, p	0.021	
Panel C: Shrinkage-adjusted ratios $R^{\mathrm{adj}}(p) = [\varphi\Delta s_p - w\Delta r_p]/[-\hat{\beta}(p)]$, w = price pass-through		
$R^{\mathrm{adj}}(\mathsf{Friend})$	0.360	[0.223, 0.498]
$R^{\mathrm{adj}}(\mathsf{Romantic})$	0.297	[0.223, 0.383]
Equality, p	0.274	

Notes: Panel A: ratio of the FE effect of each source of variation on φs to its effect on r; under pure shrinkage toward a common anchor every ratio equals the same w, under the signaling model randomized attributes pass through while purpose effects are compressed. Panel B: shrinkage predicts $G^*(p) = \kappa(-\Delta r_p)$ with κ calibrated on the relational purposes; the signaling model with $\sigma(\mathsf{Study}) = \sigma(\mathsf{Wellness}) = \sigma(\mathsf{Admin})$ predicts zero. Panel C removes the part of the belief–interest gap that shrinkage at the price pass-through rate would generate. Respondent-level bootstrap, 1,000 replications (2,000 for Panel B).

aversion (Dietvorst et al., 2015) and underreporting of AI use generally (Ling et al., 2025). Third, a scale artifact whereby the 0–100 SOB measure and the 0–10 interest measure are not linearly equivalent even after rescaling, producing a mechanical gap at all purposes. The welfare correction in Section 7 subtracts the Admin gap precisely to remain agnostic among these explanations: whatever drives the Admin divergence, it is treated as a baseline component common to all purposes and removed before estimating the suppression cost specific to relational AI.

5.6 Who is more interested? Purpose-specific determinants

Table 7 regresses each purpose-specific interest rating on respondent characteristics controlling for randomized attributes and display order. A one-unit increase in AI use frequency (0–4) raises interest by 0.68–0.82 points, implying a $+2.7$ to $+3.3$ point shift from the lowest to highest AI-use category.

Current AI companion users are more interested across all purposes (coefficient $+1.09$ to $+1.46$, more than one point above non-users). Romantic interest is low among non-users but markedly higher among current users (Appendix Fig. B.3). Age is negatively associated with demand across all purposes (steepest for Romantic: coef. -0.013/year), and men report higher romantic interest than women (coef. $+0.313$, s.e. 0.119). Relationship status shows little association with romantic interest (coef. $+0.064$, s.e. 0.128). The income gradient, lower-income respondents more interested, most steeply for relational companions (-0.029 and -0.031 per \$10k for Friend and Romantic, -0.021 to -0.022 for Study and Wellness, -0.009 for Admin), is discussed in the distributional note of Section 7.

5.7 Robustness checks

Companion-user overrepresentation.

The sample contains 27.3% current users vs. $\approx 16\%$ AI-companion use in the general population (YouGov, 2025).[5] Table A.11 re-estimates the core FE specification separately for users ($N = 543$) and non-users ($N = 1{,}446$). The Romantic penalty is larger among users: -1.569 (s.e. 0.138) against -1.161 (0.064) among non-users. In a pooled FE model that interacts every attribute with user status, the difference is -0.408 (s.e. 0.152; $p = 0.007$). The Friend penalty, by contrast, is essentially identical across the two groups (-0.609 vs. -0.569; difference -0.040, s.e. 0.146). The larger romantic penalty among users is not what lower social-image sensitivity among people who have already adopted would predict on its own; it reflects users' much higher

[5] Our item asks whether the respondent "currently use[s] an AI app to have conversations", which includes general-purpose assistants and is therefore broader than companion-app use; the population comparison should be read with that in mind. We keep the label "companion users" for brevity.

Table 7: Determinants of interest by purpose (respondent-level correlates)

	Admin	Study	Wellness	Friend	Romantic
Age (years)	-0.005	-0.011^{***}	-0.006	-0.009^{**}	-0.013^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)
Male	-0.151	0.086	-0.025	0.041	0.313^{***}
	(0.127)	(0.129)	(0.129)	(0.124)	(0.119)
Income (per \$10k)	-0.009	-0.021^{**}	-0.022^{**}	-0.029^{***}	-0.031^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
AI use frequency (0–4)	0.824^{***}	0.807^{***}	0.778^{***}	0.702^{***}	0.683^{***}
	(0.070)	(0.071)	(0.073)	(0.072)	(0.068)
Uses AI conversation app	1.369^{***}	1.282^{***}	1.253^{***}	1.464^{***}	1.093^{***}
	(0.165)	(0.166)	(0.167)	(0.167)	(0.159)
Stable partner	0.118	0.267^{*}	0.338^{**}	0.256^{*}	0.064
	(0.136)	(0.139)	(0.139)	(0.133)	(0.128)
N	1,960	1,960	1,960	1,960	1,960
R^2	0.227	0.221	0.203	0.209	0.169

Notes: OLS regressions with one observation per respondent. Controls include all randomized profile attributes and display-order dummies. Standard errors are HC1 robust.

baseline interest in every AI purpose and their much larger response to free pricing ($+2.09$ against $+1.44$), against which the romantic profile is compared. The model's object is not the penalty but its ratio to the belief–interest gap, and that ratio is the same in both groups. The cross-equation ratios of (7) are stable across subgroups and equality is never rejected within any of them at the 5% level (Table 8): $R(\mathsf{Friend})/R(\mathsf{Romantic})$ is $0.407/0.391$ among companion users ($p = 0.90$) and $0.482/0.393$ among non-users ($p = 0.20$), with similar patterns by gender, age and partnership status (the smallest p-value, 0.09, is for respondents at or below median age). The single-factor signaling structure is not an artifact of sample composition.

Table 8: Cross-equation ratio $R(p)$ by subgroup

Subgroup	N	$R(\mathsf{Friend})$	$R(\mathsf{Romantic})$	p (equality)
All	1,989	0.466 (0.065)	0.395 (0.033)	0.206
Companion users	543	0.407 (0.142)	0.391 (0.053)	0.900
Non-users	1,446	0.482 (0.081)	0.393 (0.042)	0.196
Male	963	0.471 (0.111)	0.448 (0.050)	0.805
Female	1,026	0.459 (0.088)	0.355 (0.046)	0.178
Age $\leq$ median	1,015	0.507 (0.116)	0.338 (0.054)	0.094
Age $>$ median	974	0.443 (0.085)	0.448 (0.045)	0.942
Partnered	1,230	0.418 (0.092)	0.390 (0.041)	0.726
Not partnered	759	0.547 (0.106)	0.407 (0.059)	0.140

Notes: $R(p) = G^{*}(p)/[-\hat{\beta}(p)]$ at $\varphi = 1/10$. Bootstrap standard errors in parentheses (1,000 respondent-level replications within each subgroup, fixed-effects denominator re-estimated in each draw); p-values are for $R(\mathsf{Friend}) = R(\mathsf{Romantic})$ within subgroup. The "All" row uses 1,000 replications; the text reports the 2,000-replication version.

Multiple testing.

Table A.9 reports Benjamini-Hochberg and Bonferroni adjusted p-values (Benjamini and Hochberg, 1995). Free pricing, Romantic, and Friend effects remain highly significant under both adjustments. Limited retention, balanced safety, basic personalization, standard privacy, and the Study premium survive FDR control; limited retention and balanced safety also survive Bonferroni. Interaction modality remains indistinguishable from zero.

Carryover effects.

Including vignette display position indicators (Table A.10) leaves conjoint coefficients essentially unchanged. Order effects are within ± 0.06 points and statistically insignificant.

Respondents' reasoning.

Open-ended explanations classified using a language model (Table 9) reinforce the quantitative findings. The dominant themes are instrumental value (55.1% learning and information; 32.0% administrative productivity) and wellness (30.6%). Relational purposes are discussed primarily through reservations: 59.6% mention romance to explain why it is inappropriate, 13.8% prefer real human connection, and 8.2% describe stigma or discomfort ("weird," "creepy").

Table 9: Reasons mentioned in open-ended explanations

Reason	%
Romance / dating / intimacy (mostly as a reason *against*)	59.6
Learning / study / research / information	55.1
Admin & organization / productivity	32.0
Wellness (general / habits / self-improvement)	30.6
Loneliness / companionship	14.6
Authenticity / prefer real people	13.8
Social stigma / discomfort ("creepy", awkward)	8.2
Already have partner / friends / no need	7.7
Mental health & emotional support	7.6
General reluctance / dislike of AI	4.8
Accuracy / reliability concerns	3.6
Physical health / fitness / diet / sleep	3.3
Fun / curiosity / entertainment	2.4
Privacy / data security concerns	1.5
Moral / ethical objection	1.3
Dependency / overreliance concerns	0.4

Notes: Shares of respondents whose open-ended explanation mentions each theme, classified with a language model; categories are not mutually exclusive. Representative verbatim responses for each category are reported in Appendix Table A.16.

Table 10: Willingness to pay for AI companion features (respondent FE)

Attribute level	Coefficient	WTP ($/month)	95% CI
Study	0.159	0.99	[0.26, 1.71]
Wellness	−0.081	−0.50	[−1.20, 0.20]
Friend	−0.561	−3.48	[−4.24, −2.72]
Romantic	−1.260	−7.80	[−8.72, −6.89]
Text only	−0.045	−0.28	[−0.92, 0.36]
Text + Voice + photo	0.029	0.18	[−0.44, 0.80]
Basic personalization	−0.169	−1.04	[−1.68, −0.41]
Moderate personalization	−0.100	−0.62	[−1.26, 0.02]
Standard privacy	−0.120	−0.75	[−1.36, −0.13]
Limited retention	−0.201	−1.24	[−1.88, −0.61]
Strict (PG-13) safety	−0.073	−0.45	[−1.15, 0.24]
Balanced safety	−0.185	−1.14	[−1.78, −0.50]
Price ($/month)	−0.161		
	(0.006)		

Notes: WTP from the respondent-FE model with numeric price ($0/$5/$10 per month). Baselines: Admin (purpose), Text+Voice (interaction), Advanced (personalization), High privacy, Adults-only (safety). Delta-method confidence intervals, respondent-clustered covariance.

5.8 Willingness to pay for features

We re-estimate the FE model with numeric price ($0, $5, $10), yielding a slope of $\hat{\beta}^m = -0.161$ interest points per $1 increase. WTP estimates (equation 15, delta-method standard errors) are reported in Table 10 and Figure B.1a.

The largest valuations pertain to purpose: respondents require ≈ $3.48/month to accept Friend over Admin and ≈ $7.80/month to accept Romantic (95% CI $[-8.72, -6.89]$); they pay ≈ $0.99/month more for Study (95% CI $[0.26, 1.71]$). Limited retention is valued at −$1.24/month ($[-1.88, -0.61]$) and basic personalization at −$1.04/month ($[-1.68, -0.41]$). Interaction modality valuations are near zero.

The demand curves (Figure B.1b) show the largest price drop between Free and $5, stable purpose ordering across all price levels, and a persistent absolute gap between instrumental and relational companions even at free, confirming price cannot close the purpose gap.

Because the price-utility function is nonlinear (the free gain is ≈ 4× the $10→$5 gain), WTP estimates are sensitive to the local price interval. Table A.7 reports three mappings: the Romantic WTP ranges from −$19.65 ($10→$5 slope) to −$4.88 ($5→Free slope), with the headline −$7.80 at the average slope. All mappings preserve the same ordinal purpose ranking.

5.9 Within-subject stability and the suppression mechanism

Instrument-invariance. The follow-up reproduces the purpose gradient in a sharper form. Mean ranks are Study 3.73, Wellness 3.67, Admin 3.58, Friend 2.67 and Romantic 1.35 on the 1–5 favorability scale, and the concentration at the bottom is extreme: 81.6% of respondents place Romantic last and only 1.8% place it first, against 34.2% who place Study first. The rank correlation between the five conjoint means and the five follow-up means is 0.90: the instrumental–relational separation and the position of Friend and Romantic are identical across instruments, and the single discrepancy is a swap of Admin and Wellness, two purposes the conjoint cannot distinguish either ($\hat{\beta}(\text{Wellness}) = -0.093$, s.e. 0.057).

At the individual level, we compute for each matched respondent the Spearman correlation ρ_i between their five conjoint interest ratings and their five follow-up ranks. Among the $N = 844$ respondents with non-degenerate conjoint interest vectors, the remaining 227 gave identical ratings to all five purposes and cannot be ranked, the mean ρ is 0.281 (median 0.354; 69.3% have $\rho > 0$). The purpose gradient therefore replicates across a qualitatively different elicitation instrument, weeks later, which makes conjoint-specific framing an unlikely source. It does not establish that the gradient reflects latent rather than suppressed preferences, since social-image costs may operate similarly in both instruments.

Heterogeneity test of the suppression mechanism. We use cross-respondent variation in estimated social-image sensitivity to distinguish constant suppression from differential observability across instruments. Under the signaling model, suppression is driven by individual sensitivity λ_i. If the two instruments differ in perceived observability, high-λ_i respondents should express less-suppressed preferences in the lower-observability instrument, producing *weaker* alignment (negative δ_1). If suppression is constant across instruments, alignment should be uniform across λ_i (zero δ_1) or, if high-suppression respondents have more extreme and stable preference orderings, potentially positive.

We use the Admin-differenced social-image proxy $\hat{\lambda}_i$ defined in equation (16), and estimate

$$\rho_i \;=\; \delta_0 + \delta_1 \hat{\lambda}_i + \mathbf{z}_i'\gamma + \nu_i, \tag{17}$$

where $\mathbf{z}_i$ includes AI use frequency and a companion-app use indicator. Standard errors are HC1 robust.

Table 11 reports the results. The estimated coefficient is $\hat{\delta}_1 = 0.053$ (s.e. 0.010; $t = 5.38$; $p < 0.001$) using the Admin-differenced proxy of (16), and 0.053 (s.e. 0.009; $t = 6.11$) using the undifferenced gap. The quartile pattern is monotone in both: mean ρ rises from 0.092 in the lowest-$\hat{\lambda}_i$ quartile to 0.238, 0.354 and 0.441 in successive quartiles.

This is *inconsistent* with the differential-observability hypothesis, which predicts $\delta_1 < 0$. It

is consistent with constant suppression, in which respondents who suppress most strongly do so stably across instruments, but it is not strong evidence for it, for the reason given next.

Two qualifications. First, the positive δ_1 is not merely preference extremity, though the two are not orthogonal: the correlation between $\hat{\lambda}_i$ and the within-respondent variance of conjoint ratings is 0.200, and controlling for that variance leaves $\hat{\delta}_1 = 0.049$ ($p < 0.001$). The effect operates through sensitivity rather than dispersion, but dispersion is not orthogonal to it. Second, the 227 constant-rating respondents have a *higher* undifferenced gap than the analysis sample (1.975 vs. 1.278), consistent with extreme suppression as a distinct response pattern; under the Admin-differenced proxy the comparison reverses (-0.128 vs. 0.529), but mechanically so, since a respondent who gives identical ratings to all five profiles has the own-interest terms cancel out of (16), leaving only the (negative) belief difference. We therefore do not read the constant-rating group as evidence either way.

A mechanical component. The proxy $\hat{\lambda}_i$ in (16) contains $-r_{i,\mathsf{Rom}}$ and $-r_{i,\mathsf{Friend}}$ and $+r_{i,\mathsf{Admin}}$. A respondent with high $\hat{\lambda}_i$ is therefore, by construction, one whose conjoint ratings place the relational purposes well below Admin, which is exactly the modal follow-up ranking (81.6% place Romantic last). A positive δ_1 is thus partly mechanical, for the same reason Section 6.5 guards against in the HB model. Column (4) of Table 11 removes it: ρ_i is computed over the three instrumental purposes only, and the proxy is the undifferenced relational gap $\hat{\lambda}_i^0$, so no rating enters both sides. The loading falls to 0.028 (s.e. 0.013; $p = 0.03$; $N = 789$) and remains monotone across quartiles of the proxy (-0.03, -0.01, 0.05, 0.14). The positive association therefore survives without shared ratings, although most of the original coefficient was mechanical. We read it as evidence against differential observability, and as modest support for a stable, respondent-level suppression trait.

6 A Structural Demand Model

The fixed-effects estimates recover population-average part-worths. Three questions they cannot answer motivate a structural model. What is the *distribution* of willingness to pay across consumers, which determines whether a relational product has a viable niche even when the average consumer rejects it? How does social-image sensitivity shift the taste distribution, as opposed to shifting the mean? And what happens to demand under design and pricing regimes we did not field? We answer these with a hierarchical Bayes (HB) demand model (Rossi et al., 2005; Allenby and Rossi, 1998; Moshary et al., 2025).

Table 11: Follow-up alignment and social-image sensitivity

Dependent variable	(1) ρ_i, all five	(2) ρ_i, all five	(3) ρ_i, all five	(4) ρ_i, instrumental
$\hat{\lambda}_i$, Admin-differenced	0.053 (0.010)		0.049 (0.010)	
$\hat{\lambda}_i^0$, undifferenced		0.053 (0.009)		0.028 (0.013)
AI use frequency (0–4)	−0.015 (0.018)	0.004 (0.018)	−0.018 (0.018)	0.013 (0.026)
Uses AI conversation app	−0.102 (0.041)	−0.079 (0.042)	−0.109 (0.041)	−0.019 (0.060)
Variance of own ratings			0.006 (0.003)	
Shares ratings with proxy	Yes	Yes	Yes	No
N	844	844	844	789
R^2	0.056	0.053	0.060	0.006

Notes: OLS with HC1 standard errors. In columns (1)–(3) the dependent variable is Spearman's ρ_i between a respondent's five conjoint ratings and five follow-up ranks, for the 844 matched respondents with non-degenerate ratings. In column (4) it is ρ_i computed over the three instrumental purposes only, for the 789 respondents whose three instrumental ratings are not all equal. $\hat{\lambda}_i$ is defined in (16); $\hat{\lambda}_i^0 = \frac{1}{2}\sum_{p\in\{\text{Rom,Friend}\}}(\varphi s_{ip} - r_{ip})$ is its undifferenced counterpart, which contains no instrumental rating. Column (4) therefore shares no rating between the proxy and the outcome.

6.1 Specification

Individual level. For respondent i and profile t,

$$r_{it} = \mathbf{x}_{it}'\beta_i + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_i^2), \tag{18}$$

with $\mathbf{x}_{it} \in \mathbb{R}^K$, $K = 14$, exactly as defined in Section 2.1: an intercept, four purpose indicators (baseline Admin), the *numeric* price m_{it}, and two indicators each for modality, personalization, privacy, and safety (baselines Text + Voice, Advanced, High, Adults-only). β_i is the respondent-specific taste vector of Section 2.2, with price component β_i^m.

Population level. Heterogeneity is modeled as

$$\beta_i = \mathbf{z}_i'\Delta + \eta_i, \qquad \eta_i \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_\beta), \tag{19}$$

where $\mathbf{z}_i \in \mathbb{R}^{n_z}$ collects respondent covariates, Δ is the $n_z \times K$ matrix of population regression coefficients, and $\mathbf{V}_\beta$ is the $K \times K$ covariance of taste deviations. The baseline sets $\mathbf{z}_i = (1, \tilde{a}_i, m_i, f_i/4, c_i, q_i, \tilde{y}_i)'$, where $\tilde{a}_i$ and $\tilde{y}_i$ are standardized age and income, m_i is a male indicator, $f_i \in \{0,\ldots,4\}$ is AI-use frequency, c_i indicates current AI-companion use, and q_i indicates a stable partner. Section 6.5 augments $\mathbf{z}_i$ with a social-image sensitivity measure.

Willingness to pay. Because price enters numerically, WTP for attribute level j relative to its baseline is the MRS

$$\text{WTP}_{i,j} = -\frac{\beta_i^j}{\beta_i^m}. \tag{20}$$

We evaluate (20) *within each retained MCMC draw* and then summarize the resulting posterior, rather than plugging in posterior means. This matters: $\text{WTP}_{i,j}$ is a ratio of approximately normal quantities whose denominator has non-negligible mass near zero, so its posterior is Cauchy-like and its mean need not exist. We therefore report posterior *medians* and equal-tailed credible intervals throughout, and never a posterior mean WTP.

6.2 Priors and estimation

The model takes the standard conjugate priors,

$$\Delta \sim \mathcal{MN}\left(\bar{\Delta}, \mathbf{V}_\beta \otimes \mathbf{A}_\Delta^{-1}\right), \tag{21}$$

$$\mathbf{V}_\beta \sim \mathcal{IW}(\nu, \mathbf{V}_0), \tag{22}$$

$$\sigma_i^2 \sim \chi^{-2}(\nu_e, s_e^2), \tag{23}$$

with $\bar{\Delta} = \mathbf{0}$, $\mathbf{A}_\Delta = 0.01 \cdot \mathbf{I}_{n_z}$ (diffuse), $\nu = K + 4 = 18$, $\nu_e = 3$, and $s_e^2 = 1$. Because $E[\mathbf{V}_\beta] = \mathbf{V}_0/(\nu - K - 1)$, setting $\mathbf{V}_0 = (\nu - K - 1)\mathbf{I}_K = 3\,\mathbf{I}_K$ yields a prior centered on unit taste variance, which is the intended calibration on the 0–10 rating scale.[6]

Estimation uses the Gibbs sampler in `rhierLinearModel` from the `bayesm` package (Rossi, 2019), cycling through (i) $\{\beta_i\} \mid (\Delta, \mathbf{V}_\beta, \sigma_i^2)$; (ii) $\Delta \mid (\{\beta_i\}, \mathbf{V}_\beta)$; (iii) $\mathbf{V}_\beta \mid (\{\beta_i\}, \Delta)$; and (iv) $\{\sigma_i^2\} \mid (\beta_i, \mathbf{x}_i, \mathbf{r}_i)$. We run three independent chains of 9,000 iterations from dispersed starting values, discard the first 3,000 of each as burn-in, and retain every sixth draw, giving 3,000 posterior draws in total. Convergence is assessed by trace inspection and by split-$\hat{R}$ and effective sample size computed across chains for every element of $\bar{\mathbf{z}}'\Delta$ and $\text{diag}(\mathbf{V}_\beta)$. For the linear model the maximum $\hat{R}$ is 1.03 across population means and 1.05 across variances (minimum ESS 84). The censored model of Section 6.6 mixes more slowly because of its data-augmentation step, so we run it for 30,000 iterations per chain (10,000 burn-in, thinned by 20): the maximum $\hat{R}$ is 1.03 across population means and at most 1.004 for the Romantic, Friend and price variances; the largest value, 1.055, is for the text-only variance. These values are adequate for the location and dispersion statistics we report, though not all fall below the strictest 1.01 threshold.

What is and is not identified at the individual level. With $n_i = 5$ observations and $K = 14$ parameters, each respondent's likelihood is rank-deficient: β_i is not identified from respondent i's data alone and its posterior is dominated by the population distribution. What *is* well identified, from $N = 1{,}989$ respondents, is the pair $(\Delta, \mathbf{V}_\beta)$. We therefore read Figure 5 as the estimated *population taste distribution*, not as 1,989 separately measured consumers, and we avoid individual-level claims that the design cannot support. Because $\mathbf{V}_\beta$ is identified across

[6]The calibration is not innocuous with $n_i = 5$ observations per respondent. Setting instead $\mathbf{V}_0 = \nu\mathbf{I}_K$, which implies $E[\mathbf{V}_\beta] = 6\mathbf{I}_K$, shrinks individual tastes less and moves the population means marginally away from zero (Romantic -0.998 against -0.920; price -0.152 against -0.147). All substantive conclusions are unchanged.

respondents, statements about dispersion and about the mass in the upper tail are legitimate; statements about *which* respondent occupies the tail are not.

Model fit. We report a holdout exercise in which the model is estimated on four of each respondent's five profiles and used to predict the fifth. Holding out one randomly chosen profile per respondent, the HB model achieves a root mean squared error of 1.990 on the held-out ratings (mean absolute error 1.425; Pearson $r = 0.759$; Spearman 0.765), against 2.094 / 1.563 / 0.739 / 0.751 for the fixed-effects model estimated on the same four profiles and 2.338 / 1.597 / 0.664 / 0.682 for a respondent-mean benchmark. The held-out ratings have a standard deviation of 3.050, so the model explains roughly 57% of out-of-sample variance and improves on both benchmarks.

6.3 Population-level estimates

Table A.13 compares HB population means, evaluated at $\bar{\mathbf{z}}'\widehat{\Delta}$, with the fixed-effects estimates. The qualitative picture is identical: the purpose gradient is preserved ($\mathsf{Romantic} < \mathsf{Friend} < \mathsf{Wellness} < \mathsf{Admin} < \mathsf{Study}$), the price coefficient is negative, and privacy and personalization carry the expected signs.

The HB population means are mildly attenuated relative to the fixed-effects estimates: $\mathsf{Romantic}$ is -0.920 (95% CrI $[-1.020, -0.820]$) against -1.260, a ratio of 0.73, and the price coefficient is -0.147 $[-0.159, -0.135]$ against -0.161, a ratio of 0.91. We attribute most of the residual attenuation to floor censoring rather than to prior shrinkage: the linear-normal likelihood in (18) is misspecified when more than half of romantic ratings sit at zero. Consistent with this, the less informative prior in the footnote above moves the estimates only marginally, whereas relaxing the likelihood (Section 6.6) reverses the ordering, producing a latent penalty larger than the linear FE estimate.

Note that a common multiplicative attenuation would leave WTP ratios (20) unaffected, which is one reason the HB and FE money-metric estimates remain comparable even where the raw coefficients are not. We nonetheless treat the FE estimates as the benchmark for population-*mean* WTP, following Moshary et al. (2025), and reserve the HB model for what it is uniquely suited to: the distribution, the covariance structure of tastes, and counterfactuals.

6.4 The distribution of willingness to pay

Figure 5 plots the posterior distribution of $\mathrm{WTP}_{i,\mathsf{Romantic}}$, $\mathrm{WTP}_{i,\mathsf{Friend}}$, and $\mathrm{WTP}_{i,\mathsf{Study}}$ relative to the Admin baseline. Three features are salient.

First, the distributions are heavy-tailed and overlapping. About a fifth of individual price-coefficient draws are zero or positive, for which WTP is not defined; among the rest, the posterior

median WTP for a romantic companion relative to admin is −$5.06/month (Table A.14). A statistic that does not depend on the price coefficient is the share of the taste distribution that prefers a romantic companion to an administrative assistant at equal price: 9.0% under the linear model and 14.1% under the censored likelihood (Table 13). The commercially relevant fact is not the mean but this mass: a product that the average consumer values at large negative amounts can still support a viable segment.

Second, the upper tail is concentrated among current AI-companion users, whose mean predicted utility for a free romantic companion (4.73 interest points) is double that of non-users (2.37 points). Whether this reflects selection on taste or learning through use is not identified here and is a first-order question for follow-up work with panel data.

Third, dispersion itself varies by purpose. The standard deviation of the Romantic taste coefficient across respondents is 0.646 (from $\hat{\mathbf{V}}_\beta$), against 0.400 for Friend, 0.390 for Study and 0.324 for Wellness; under the censored likelihood the Romantic dispersion is 2.55 against 1.26 for Friend and 0.94 for Study. The romantic dispersion is what the signaling model predicts: where $\sigma(p,\bar{o})$ is large, heterogeneity in λ_i translates into heterogeneity in reported taste, whereas where $\sigma \approx 0$ it does not.

6.5 Social-image sensitivity in the taste distribution

We let social-image sensitivity enter (19) directly, so that $\Delta[\lambda, \cdot]$ measures how a one-standard-deviation increase in sensitivity shifts each taste parameter, holding demographics fixed.

Avoiding a mechanical result. As flagged in Section 4, the natural proxy $\hat{\lambda}_i$ from (16) contains $-r_{ip}$ for both relational purposes, and $\beta_i^{\mathsf{Romantic}}$ is estimated principally from $r_{i,\mathsf{Romantic}}$. It also contains $+r_{i,\mathsf{Admin}}$, and every purpose coefficient is measured relative to Admin. Regressing purpose tastes on such a proxy delivers negative loadings *by construction*, whether or not any signaling mechanism exists. We therefore build proxies from one relational purpose and use them to explain the other, and we difference them against Study rather than Admin: $\hat{\lambda}_i^{\mathsf{F}} = (\varphi s_{i,\mathsf{Friend}} - r_{i,\mathsf{Friend}}) - (\varphi s_{i,\mathsf{Study}} - r_{i,\mathsf{Study}})$ explains $\beta_i^{\mathsf{Romantic}}$, and $\hat{\lambda}_i^{\mathsf{R}}$, built the same way from the Romantic profile, explains $\beta_i^{\mathsf{Friend}}$. In these specifications no rating enters both the proxy and either the target relational coefficient or the Wellness coefficient, which serves as the clean instrumental comparison. Both proxies are standardized; in raw units they have means 0.387 and 0.623 and standard deviations 2.19 and 2.37 on the 0–10 interest scale.

Results. Table A.15 reports the preferred specifications alongside the Admin-differenced versions used in earlier drafts. The mechanical component is large. With Admin-differenced proxies, every purpose coefficient loads at -0.4 to -0.6 or more, including Study and Wellness, and the intercept loads at $+0.7$ to $+0.9$: this is the shared Admin rating. With Study-differenced

proxies the picture is cleaner. Sensitivity measured on the Friend profile shifts β^{Romantic} by -0.32 $[-0.43, -0.22]$ per standard deviation, against -0.12 $[-0.21, -0.01]$ for β^{Wellness}; sensitivity measured on the Romantic profile shifts β^{Friend} by -0.17 $[-0.28, -0.07]$, against $+0.02$ $[-0.08, 0.12]$ for Wellness. Every design attribute loads at essentially zero in all four specifications, including price (-0.010 to -0.014).

The prediction the specification tests is sharp and asymmetric: social-image sensitivity should shift *relational purpose* coefficients and leave *attribute* sensitivities (price, privacy, personalization, safety) unchanged. Nothing in the estimation imposes this. A model in which the belief–interest gap simply measures general response style, scale use, or optimism about others would shift all coefficients or none. The excess loading on the two purposes with high $\sigma(p, \bar{o})$, and its absence on price, is the structural counterpart of the cross-equation restriction in (7). It is, however, much smaller than the Admin-differenced specifications suggested.

A structural suppression correction. Shifting a respondent from the sample mean of the raw proxy to zero, a change of $\text{mean}(\hat{\lambda})/\text{sd}(\hat{\lambda}) \approx 0.18$ to 0.26 standard deviations, and recomputing the population-mean WTP within each draw raises romantic WTP by $+$\$0.30/month (95% CrI $[0.16, 0.46]$) and friendship WTP by $+$\$0.25/month $[0.06, 0.44]$. Using only the excess over the Wellness loading gives $+$\$0.17 $[0.01, 0.32]$ and $+$\$0.28 $[0.09, 0.47]$. The Admin-differenced proxies of earlier drafts give $+$\$0.65 and $+$\$0.85, and a proxy pooled across both relational purposes $+$\$1.58 and $+$\$1.35: the difference is the mechanical component, not suppression.

The exercise removes only the component of suppression that the belief–interest gap detects, which by (9) is a fraction $(\bar{\lambda} - \mu)/\bar{\lambda} < 1$ of individual suppression, and none of the equilibrium premium ΔW. In those two senses it understates individual suppression.

6.6 Relaxing the normal likelihood

A large share of ratings sits at the floor of the scale: 54.9% for Romantic, 41.0% for Friend, and 30–33% for the three instrumental purposes, against under 4% at the ceiling. Equation (18) treats these as interior observations, which attenuates every coefficient toward zero and does so most where the mass at the floor is largest. We therefore re-estimate the model with a two-limit censored likelihood on $[0, 10]$, augmenting the sampler with a draw of the latent rating r_{it}^{*} from a truncated normal at each iteration and leaving the rest of the hierarchy unchanged.

Table 13 reports the result. Relative to the linear HB model the purpose gradient roughly triples, Romantic moves to -2.895 $[-3.166, -2.644]$ and Friend to -1.024 $[-1.193, -0.847]$, and the price coefficient steepens to -0.273 $[-0.294, -0.253]$. Because both numerator and denominator scale up, the money-metric estimates move less than the coefficients: romantic WTP is $-$\$10.60/month $[-11.9, -9.4]$ against $-$\$6.24 under the linear likelihood. The full ordinal structure and every attribute sign are preserved, and the censored estimates line up with a

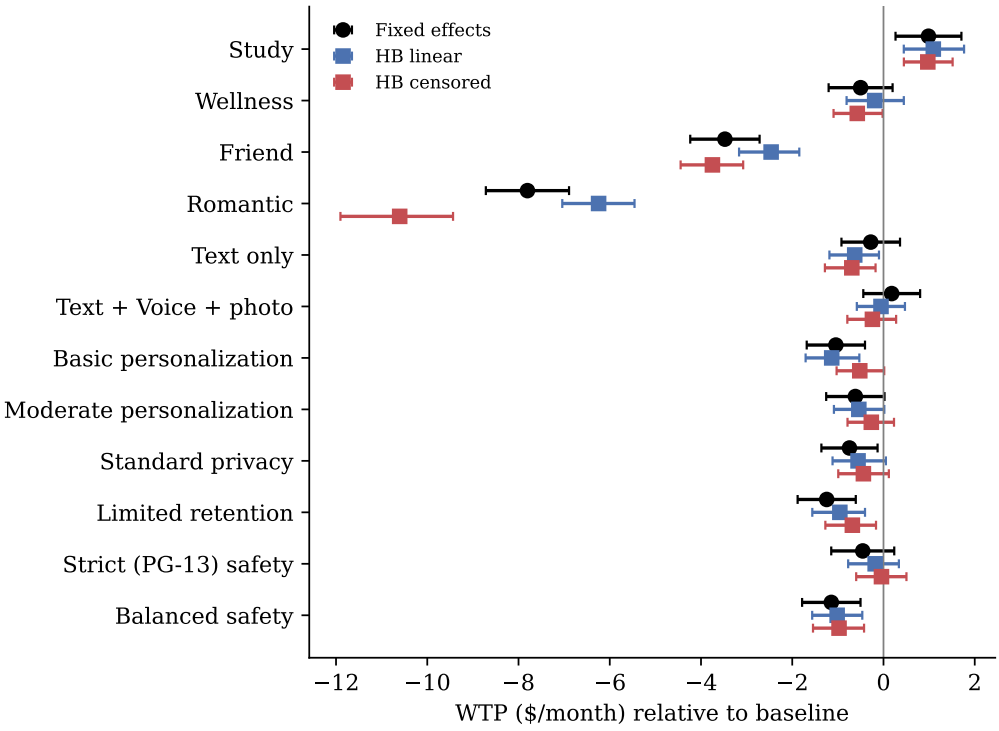


Figure 4: Willingness to pay by attribute, from the fixed-effects model and from the HB population mean under both the linear and the censored likelihood. WTP is the monthly dollar amount a respondent would need to be compensated to accept the listed attribute level relative to its baseline. Baselines are Admin (purpose), High privacy, Advanced personalization, Adults-only safety and Text + Voice; price enters numerically and therefore has no omitted level. Whiskers are 95% credible intervals for the HB estimates. The purpose gradient is preserved across all three estimators, and the censored model scales it up uniformly.

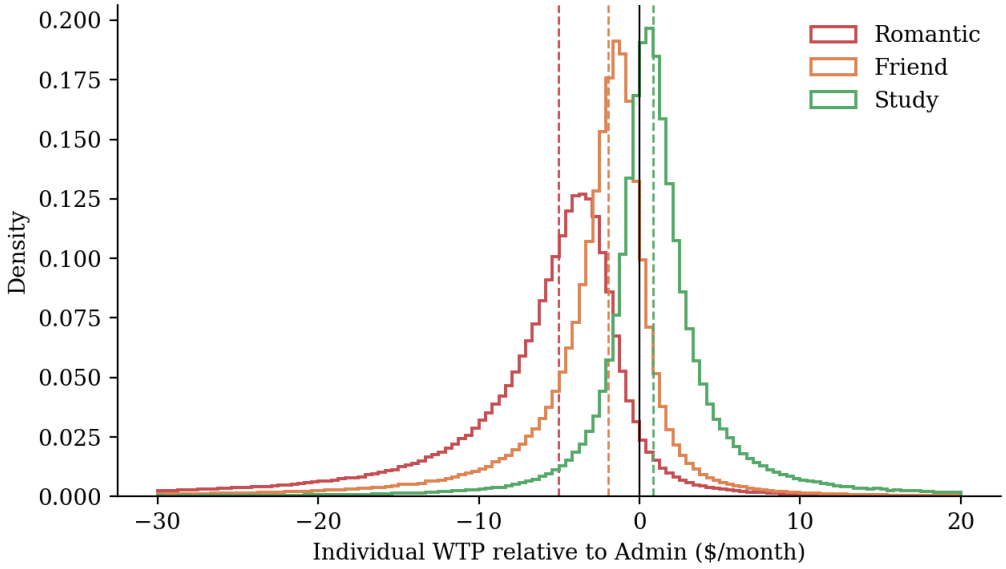


Figure 5: Posterior distribution of individual willingness to pay for Romantic, Friend and Study companions relative to an Admin baseline, from the linear HB model, evaluated within each retained draw and pooled across draws and respondents, among draws with a negative individual price coefficient. Dashed lines mark medians. Horizontal range truncated for display.

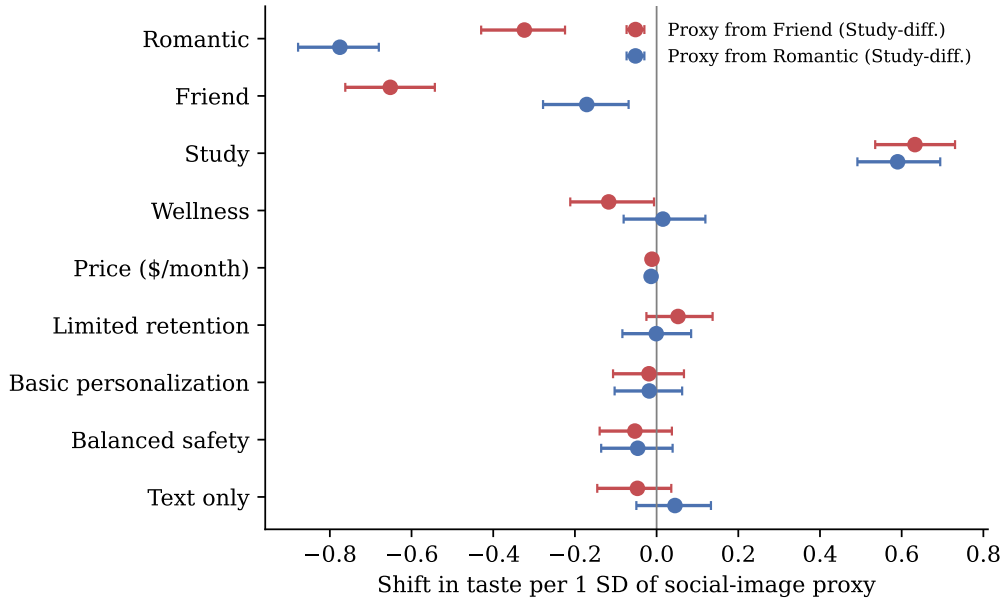


Figure 6: Effect of social-image sensitivity on taste parameters, from the HB signaling extension with Study-differenced proxies (columns 3–4 of Table A.15). Each point is a row of $\Delta[\lambda, \cdot]$: the change in population-mean taste per one-standard-deviation increase in the proxy, with 95% credible intervals. The Study loadings are mechanical (the proxy contains the Study rating); the informative comparison is between the target relational purpose and Wellness. Every design attribute, including price, is indistinguishable from zero.

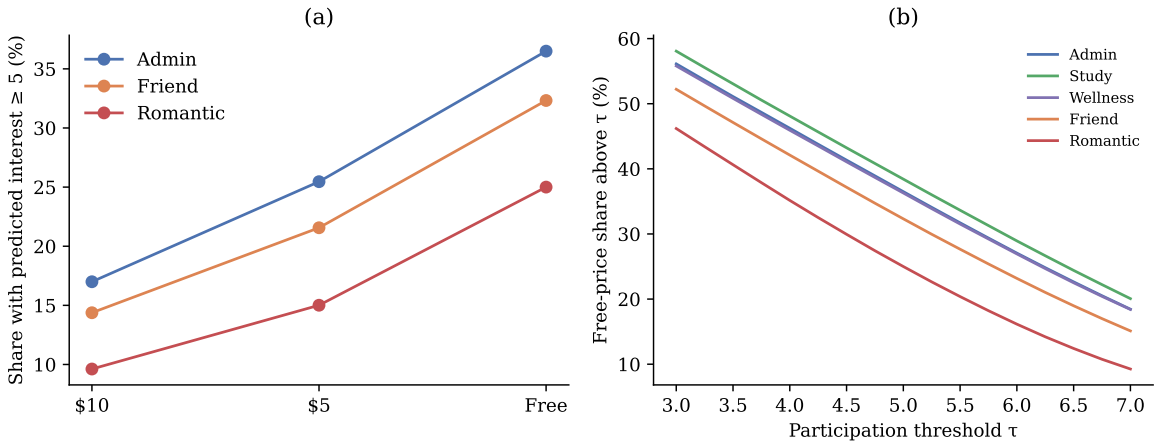


Figure 7: Counterfactual demand from the HB taste distribution. Panel (a) gives the share of the distribution above the benchmark participation threshold $\tau = 5$ at each price. Panel (b) traces the free-price share for each purpose across the full range of thresholds, showing that the ordering and the relational gap are invariant to where the participation margin is set.

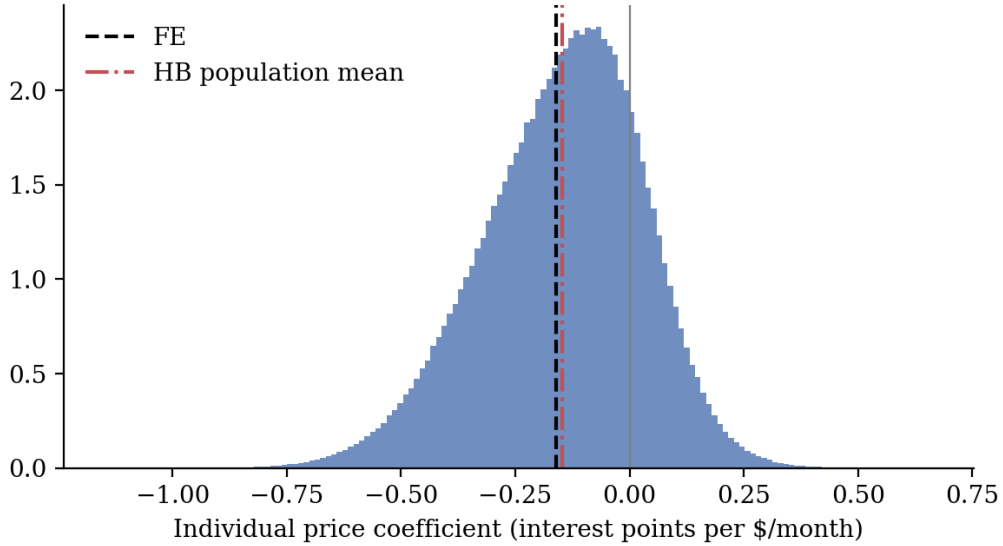


Figure 8: Distribution of individual price coefficients β_i^m from the HB model, in interest points per dollar per month. The dashed line marks the fixed-effects estimate and the dot-dashed line the HB population mean. Roughly a fifth of the distribution lies within 0.05 of zero: for these respondents the money-metric ratio in (20) is poorly determined, which is why WTP is summarized by posterior medians rather than means.

Table 12: Counterfactual demand under alternative price and privacy regimes

Scenario	Mean predicted interest	Share above threshold (%) $\tau = 5$	Share above threshold (%) $\tau = 4$	Users / non-users mean interest
Admin AI × \$10	2.46	17.0	25.3	3.91 / 1.92
Admin AI × \$5	3.20	25.5	36.1	4.88 / 2.57
Admin AI × Free	3.93	36.5	46.2	5.85 / 3.21
Romantic AI × \$10	1.54	9.6	15.8	2.78 / 1.07
Romantic AI × \$5	2.28	15.0	23.5	3.76 / 1.72
Romantic AI × Free	3.01	25.0	35.2	4.73 / 2.37
Romantic AI × Free, limited retention	2.87	23.1	33.3	4.52 / 2.25
Friend AI × \$5	2.83	21.6	31.2	4.56 / 2.19
Friend AI × Free	3.57	32.3	42.1	5.54 / 2.83
Study AI × Free	4.10	38.4	48.1	6.06 / 3.36
Wellness AI × Free	3.91	36.3	45.9	5.81 / 3.19

Notes: Predicted interest $\mathbf{x}'\beta_i$ from the linear HB taste distribution on the 0–10 scale, over 600 retained draws and all respondents. Share above threshold is the fraction with predicted interest at least τ. Because the instrument contains no outside option, τ is calibrated rather than estimated and levels should not be read as adoption forecasts; differences across scenarios are the object of interest. Non-varying attributes at reference levels (High privacy, Advanced personalization, Adults-only safety, Text + Voice).

two-limit Tobit fitted without heterogeneity (-2.345, -0.920 and -0.233 respectively), which is what one expects if censoring rather than the random-coefficient structure is what separates the two specifications.

Two consequences carry forward. The linear money-metric estimates are conservative, so the dollar figures in Section 7 understate the latent penalty. And the dispersion of relational tastes is larger than the linear model implies: the standard deviation of the Romantic taste coefficient rises from 0.646 to 2.548, and the share of the distribution preferring a romantic companion to Admin rises from 9.0% to 14.1%. The viable relational segment is wider under the correctly specified likelihood, not narrower.

Adding purpose × price interactions to the censored model (one chain of 15,000 iterations) confirms the latent-index result of Section 5.3 with individual heterogeneity: the Romantic × Free interaction is -1.22 $[-1.62, -0.73]$, a latent romantic free premium of 55% of the Admin premium, and the Friend × Free interaction is -0.59 $[-0.91, -0.33]$, a premium of 79%.

6.7 Counterfactuals

We use the posterior draws to compute predicted demand under pricing and privacy regimes we did not field. Because the instrument contains no outside option, the participation margin is not identified from the data, and we are explicit about the consequence: we define adoption as predicted interest $\geq \tau$ and treat τ as a *calibrated*, not estimated, parameter. Levels therefore

Table 13: Linear and censored hierarchical Bayes estimates

	Population mean		WTP ($/month)	
	Linear	Censored	Linear	Censored
Study	0.161	0.265	+1.09	+0.97
Wellness	−0.028	−0.157	−0.19	−0.57
Friend	−0.363	−1.024	−2.46	−3.75
Romantic	−0.920	−2.895	−6.24	−10.60
Limited retention	−0.141	−0.187	−0.96	−0.68
Basic personalization	−0.167	−0.142	−1.13	−0.52
Balanced safety	−0.150	−0.266	−1.02	−0.97
Price ($/month)	−0.147	−0.273	—	—
SD of Romantic taste	0.646	2.548		
Share preferring Romantic to Admin	9.0%	14.1%		
Max. $\hat{R}$ (population means)	1.03	1.03		

Notes: Population means evaluated at $\bar{\mathbf{z}}$ (posterior medians); WTP is the ratio of population means evaluated within each draw. The censored model uses a two-limit likelihood on $[0, 10]$ with data augmentation of the latent rating: three chains of 30,000 iterations, 10,000 burn-in, thinned by 20. "Share preferring Romantic" is the share of individual draws with $\beta_i^{\text{Romantic}} > 0$, which does not depend on the sign of the individual price coefficient.

depend on τ; differences across scenarios are the object of interest.

Table 12 reports results at the benchmark $\tau = 5$ and at $\tau = 4$, and Figure 7b traces the free-price share for every purpose across $\tau \in [3, 7]$. The mapping from τ to the free-Admin share is monotone and smooth, $\tau = 4.64$, 5.16, 5.68, 6.23 and 6.81 reproduce shares of 40, 35, 30, 25 and 20 percent respectively, so a reader who prefers a different external benchmark for general-purpose assistant adoption (YouGov, 2025) can recalibrate the levels directly without re-estimating the model.

Four results stand out.

First, free pricing does not close the purpose gap. Moving a romantic companion from $10 to free raises the share above threshold from 9.6% to 25.0%, which still leaves it 11.5 points below the 36.5% share for a free administrative assistant. The relational gap at free pricing ranges between 9 and 11 points across $\tau \in [3, 7]$ and never closes. This is Hypothesis 2 in share terms.

Second, the price gradient is steeper for instrumental purposes at the thresholds that matter. The Admin share rises 19.5 points between $10 and free against 15.4 points for Romantic at $\tau = 5$. The ranking holds for $\tau \geq 3.75$; below that it reverses slightly, because the Admin distribution is already largely above the margin.

Third, privacy moves utility but only modestly moves the participation margin. High privacy raises predicted romantic utility by 0.14 points relative to limited retention, and the corresponding share moves from 23.1% to 25.0% at zero price. The effect is between 1.3 and 1.9 points across the threshold range: real in continuous utility, but too small on its own to generate large threshold

crossings.

Fourth, heterogeneity by current companion use is large. Mean predicted utility for a free romantic companion is 4.73 interest points among current users against 2.37 among non-users. This is the segment structure that the aggregate purpose penalty conceals.

7 Welfare: How Much of the Relational Penalty Is Social Image?

The estimates so far document a steep purpose gradient, a zero-price premium that is attenuated exactly where the gradient is steepest, a compression of second-order beliefs relative to own interest whose cross-purpose ratio is statistically constant, and a taste distribution in which social-image sensitivity loads on relational purposes and nowhere else. This section asks what those findings imply for the welfare-relevant demand for relational AI.

The question has a specific managerial and regulatory form. A firm reading $\widehat{\text{WTP}}(\text{Romantic}) = -\$7.80$/month concludes that consumers dislike the category and that the fix is a better product. A regulator reading the same number concludes the category generates no value worth protecting. Both readings assume that the estimated purpose coefficient is $\Delta\alpha_p$. Equation (3) says it is $\Delta\alpha_p - \bar{\lambda}\Delta\sigma_p$. This section bounds the second term.

7.1 What is identified, and what we therefore report

We begin with a constraint that shapes everything below. Our instrument is a rating with no outside option and no scale anchor. It identifies *differences* in willingness to pay across profiles; it does not identify the *level* of consumer surplus from any single profile, because the utility of not adopting is never elicited. Any statement of the form "consumer surplus for a free romantic companion is $\$X$" requires an anchor the design does not provide.[7]

We therefore state every welfare quantity as a difference against the instrumental baseline, and we report it in two equivalent forms: in dollars per month, and as a *share of the measured relational penalty*. The share form turns out to be the more robust of the two, for a reason developed below.

[7]The point is easy to see arithmetically. The estimated interest gain from a \$10 price cut is 1.612 points, and dividing by the price slope of 0.1612 points per dollar returns \$10.00, an identity, not an estimate. Any surplus level constructed this way is definitionally equal to the price change. Adding an outside option to the instrument, as in Moshary et al. (2025), would identify participation shares and compensating variation directly; we regard this as the single most valuable change to the design and discuss it in Section 9.

The identified standard measure is the Admin-relative money-metric penalty,

$$\Delta\mathrm{CS}(p) \;=\; \frac{\hat{\beta}(p)}{|\hat{\beta}^{m}|}, \tag{24}$$

which equals $-\$7.80$/month for Romantic and $-\$3.48$/month for Friend at the average price slope, and $-\$4.88$ and $-\$2.18$ at the local \$5 $\rightarrow$ free slope. This is a biased welfare measure in the sense of Bernheim and Rangel (2009): it inherits the distortion generated by an identifiable ancillary condition, here social observability.

7.2 The corrected measure

Following Allcott and Taubinsky (2015), the money-metric value of removing a distortion is the ratio of the distortion's effect on the outcome to the price slope. The distortion's effect is the Admin-relative belief–interest gap $G^{*}(p)$ from (6). Writing $\Delta s_p \equiv \bar{s}(p) - \bar{s}(\mathsf{Admin})$ and $\Delta r_p \equiv \bar{r}(p) - \bar{r}(\mathsf{Admin})$,

$$G^{*}(p) \;=\; \varphi\,\Delta s_p \;-\; \Delta r_p. \tag{25}$$

Both differences are negative (relational purposes attract less own interest *and* lower beliefs about others) so $G^{*}(p)$ is *decreasing* in φ. With $\Delta s_{\mathsf{Rom}} = -7.775$, $\Delta r_{\mathsf{Rom}} = -1.276$, $\Delta s_{\mathsf{Friend}} = -3.065$ and $\Delta r_{\mathsf{Friend}} = -0.568$, the benchmark $\varphi = 1/10$ gives $G^{*}(\mathsf{Romantic}) = 0.498$ and $G^{*}(\mathsf{Friend}) = 0.262$ interest points.[8]

Differencing against Admin removes whatever drives the baseline divergence (optimism about others, mild algorithm aversion, or a scale artifact) and is what assumption (A2) buys.

The corrected penalty is then

$$\Delta\mathrm{CS}^{\mathrm{corr}}(p;\pi,\varphi) \;=\; \frac{\hat{\beta}(p) + \pi\, G^{*}(p;\varphi)}{|\hat{\beta}^{m}|}, \tag{26}$$

where $\pi \in [0,1]$ is the share of $G^{*}(p)$ attributable to genuine suppression rather than residual pluralistic ignorance or scale artifacts.

7.3 The suppression share is invariant to the price slope

The dollar form of (26) depends on $\hat{\beta}^{m}$, which assumption (A3) puts under strain: the price response is sharply nonlinear, and the "right" slope differs by margin. But the *share* of the relational penalty removed by the correction does not depend on the slope at all, because the

[8] Note that the own-interest difference enters (25) unscaled: φ converts the 0–100 belief measure into interest points and applies to Δs_p only. Applying it to both terms would leave the benchmark $\varphi = 1/10$ unchanged but reverse the direction of the sensitivity.

slope divides both numerator and denominator:

$$\underbrace{\frac{\pi G^*(p)}{-\hat{\beta}(p)}}_{\text{share of penalty explained}} = \pi R(p), \tag{27}$$

where $R(p)$ is exactly the cross-equation ratio of (7) that Section 5.5 tests. The mechanism test and the welfare statistic are the same object. This is the paper's central quantitative result. It does not use the price slope at all, so (A3) is irrelevant to it, and it is a ratio of two Admin-relative differences, so any fixed offset between the belief and interest scales cancels. It does depend on φ, and we report it over the full range.

At the benchmark $\varphi = 1/10$ and $\pi = 1$,

$$R(\text{Romantic}) = \frac{0.498}{1.260} = 39.5\% \text{ (bootstrap s.e. 3.3)}, \qquad R(\text{Friend}) = \frac{0.262}{0.561} = 46.6\% \text{ (6.7)},$$

and we cannot reject equality (difference 0.071, bootstrap s.e. 0.058, $Z = 1.23$, $p = 0.217$; 2,000 respondent-level replications that propagate sampling error in both the numerator and the fixed-effects denominator). Across $\varphi \in [0.08, 0.12]$ the ranges are 27.2%–51.9% for Romantic and 35.7%–57.6% for Friend (Figure 9).

Headline. If the entire Admin-relative belief–interest gap is genuine suppression ($\pi = 1$), social-image costs account for roughly *two-fifths* of the measured demand penalty for a romantic AI companion and a similar share for a friendship companion: in dollars at the average price slope, \$3.08 of the \$7.80/month romantic penalty; at the local \$5 → free slope, \$1.93 of \$4.88/month. This is not a bound in either direction. Two forces push the true share up: beliefs are themselves partly suppressed ($\mu > 0$), and the equilibrium premium ΔW is excluded. Two push it down: part of the gap is belief shrinkage rather than suppression (removing it lowers the romantic share to 30% and the friendship share to 36%; Table 6), and the structural estimates below put π at roughly 0.10 for Romantic and 0.15 for Friend, i.e. about \$0.30 and \$0.25 per month. Within the component the belief–interest gap identifies, the large majority of the relational penalty is intrinsic taste.

7.4 What pins down π, and what lies outside it

The correction is only as good as π, which is not directly observed. Three pieces of evidence bear on it and on how the corrected figure relates to the full distortion.

(i) The structural estimate. The HB signaling extension (Section 6.5) recovers the suppression from a different moment, cross-respondent covariation between social-image sensitivity and relational tastes, rather than from the level of the belief–interest gap. In the specification in which no rating enters both sides (Study-differenced proxies), shifting a respondent from the

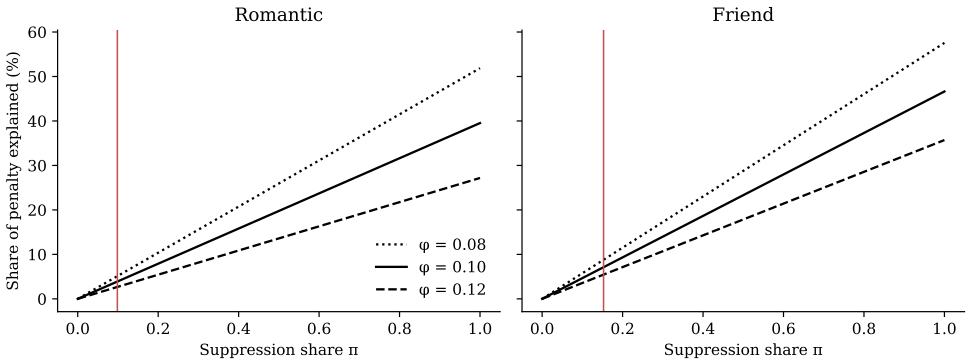


Figure 9: Share of the Admin-relative purpose penalty attributable to social-image suppression, as a function of the suppression share π and the scale conversion φ. The quantity plotted is $\pi R(p)$ from (27), which does not involve the price slope and is therefore invariant to the nonlinearity in the price response. The vertical line marks the suppression share implied by the structural estimates of Section 6.5 (Study-differenced proxies), separately for each purpose.

sample mean of the social-image proxy to zero raises romantic WTP by \$0.30/month (95% CrI $[0.16, 0.46]$) and friendship WTP by \$0.25/month $[0.06, 0.44]$. Against penalties of \$7.80 and \$3.48 these are shares of 3.8% and 7.2%, implying $\hat{\pi} \approx 0.10$ and $\hat{\pi} \approx 0.15$ respectively. Admin-differenced proxies, which share a rating with every purpose coefficient, give 0.21 and 0.53; the difference is mechanical. The belief-gap and covariance-based approaches therefore agree on sign but not on level. We regard $\pi \approx 0.1$–0.15 as the best structural estimate, $\pi R^{\text{adj}}(p)$ at $\pi = 1$ as a shrinkage-adjusted upper reference, and $R(p)$ as the unadjusted one.

(ii) Bounds from (A1). By (9), even $\pi = 1$ recovers only the fraction $(\bar{\lambda} - \mu)/\bar{\lambda}$ of individual suppression, because beliefs are themselves partially suppressed. The true individual suppression share is therefore at least $\pi R(p)$, and strictly greater whenever $\mu > 0$.

(iii) The equilibrium premium. The individual correction conditions on the prevailing adoption level A^*. Moving to a normalized equilibrium A^{**} lowers σ for everyone and releases additional surplus $\Delta W > 0$ (Section 2.5), which is not identified from individual stated-preference data and would require a collective-adoption treatment analogous to Bursztyn et al. (2025a). The full welfare-relevant penalty is therefore

$$\Delta \text{CS}^{\text{full}}(p) \;=\; \Delta \text{CS}^{\text{corr}}(p;\pi,\varphi) \;+\; \underbrace{\left[\frac{\bar{\lambda}}{\bar{\lambda}-\mu} - 1\right] \cdot \frac{\pi G^*(p)}{|\hat{\beta}^m|}}_{\text{belief-suppression residual}} \;+\; \underbrace{\Delta W}_{\text{equilibrium premium}}, \tag{28}$$

with both correction terms strictly positive. Conditional on π, every number we report is thus a lower bound on the true correction; the uncertainty about π itself runs in the opposite direction.

7.5 Sensitivity and censoring

Table 14 reports the corrected romantic penalty over the (π, φ) grid in all three forms. Under no combination in the plausible range does the correction eliminate the relational penalty: even at $\pi = 1$ and $\varphi = 0.08$, social image accounts for just over half of it.

Table 14: Corrected Admin-relative Romantic penalty across (π, φ)

	Suppression share π				
φ	0.00	0.25	0.50	0.75	1.00
Panel A. Share of the 1.260-point penalty explained (slope-invariant)					
0.08	0.0%	13.0%	25.9%	38.9%	51.9%
0.10	0.0%	9.9%	19.8%	29.6%	39.5%
0.12	0.0%	6.8%	13.6%	20.4%	27.2%
Panel B. Corrected penalty, \$/month, average slope $\lvert\hat{\beta}^m\rvert = 0.161$					
0.08	−7.80	−6.79	−5.78	−4.77	−3.76
0.10	−7.80	−7.03	−6.26	−5.49	−4.72
0.12	−7.80	−7.27	−6.74	−6.21	−5.68
Panel C. Corrected penalty, \$/month, local \$5→Free slope $= 0.258$					
0.08	−4.88	−4.24	−3.61	−2.98	−2.35
0.10	−4.88	−4.39	−3.91	−3.43	−2.95
0.12	−4.88	−4.54	−4.21	−3.88	−3.55

Notes: From (26) with $\hat{\beta}(\text{Romantic}) = -1.260$ and $G^*(\varphi) = \varphi\Delta s - \Delta r$. Panel A reports $\pi R(\text{Romantic})$, which does not involve the price slope; Panels B and C convert to dollars at the average and local slopes.

Censoring and the scale of the correction. Because more than half of romantic ratings sit at the floor of the rating scale, the linear estimates that enter (26) are attenuated relative to the latent index. The censored model of Section 6.6 recovers a latent romantic penalty of -2.895 against -1.260 and a price slope of -0.273 against -0.161, implying a latent romantic WTP of $-\$10.60$/month. The ordinal structure and every attribute sign are unchanged, so the qualitative claims are unaffected, but the money-metric magnitudes in Panels B and C should be read as conservative.

The share in Panel A requires more care, because $G^*(p)$ is measured on the observed rating scale while the censored penalty is measured on the latent index. Scaling the numerator by the same factor that separates the two purpose coefficients would leave the share unchanged; leaving it unscaled would reduce it to roughly 17%. Which is correct depends on whether the belief–interest gap is itself censored, and beliefs, elicited on a 0–100 scale with only 2.8% of responses at zero, are far less censored than own interest. The honest reading is therefore that 39.5% is an upper estimate of the observed suppression share and the latent-scale share is somewhat lower, which reinforces rather than weakens the conclusion that the majority of the relational penalty is intrinsic.

7.6 Relation to the social-media mirror case

The structure inverts Bursztyn et al. (2025a) along every dimension. Their correction *subtracts* welfare from a positive standard surplus by internalizing costs that active users impose on non-users; ours *adds* welfare to a negative standard measure by recovering costs that the prevailing norm imposes on potential adopters. The difference in sign drives a difference in instruments: the social-media coordination failure calls for instruments that reduce adoption (usage caps, abstention subsidies, taxation of attention), whereas the relational-AI coordination failure calls for instruments that facilitate it (privacy-preserving design, discreet usage modes, norm-shifting communication).

What the two cases share is more important than what separates them. Both adjustments are invisible to standard surplus measures. Both require off-equilibrium behavior that individual stated-preference data cannot identify. And both describe individually stable equilibria that no individual choice can correct, which is why in each case the policy lever is aggregate, not because consumers err relative to their own preferences, but because they optimize against a collectively suboptimal social equilibrium.

7.7 Distributional note

The heterogeneity results in Section 5.6 show that lower-income respondents are disproportionately interested in relational companions (≈ -0.028 interest points per \$10k for Friend and Romantic, against ≈ -0.005 for Admin). If the suppression share $\pi R(p)$ is uniform across income, the distortion falls disproportionately on the segment most interested in relational products, and that segment has higher marginal utility of income and is least able to pay for privacy-preserving premium tiers. Under any welfare weighting that favors lower-income consumers, the distortion therefore matters more than its average dollar value suggests. We do not have the power to estimate π separately by income, and flag this as the distributional question a larger design should address.

8 Managerial Implications

Positioning is the highest-use design choice. The largest demand shifts come from purpose, not features. Relative to an Admin assistant, a Romantic companion lowers interest by 1.262 points and a Friend companion by 0.572 points (Table A.6), equivalent to roughly \$7.80 and \$3.48 per month in required compensation. The largest non-price attribute effect, limited data retention, is worth about \$1.24: a sixth of the romantic penalty and a third of the friendship penalty. The signaling model in Section 2 offers one reason: purpose operates as a category cue that activates identity and social-image concerns before any attribute-level evaluation, so

productivity, study, and wellness propositions sit closer to mass-market demand, while explicitly relational products require sharper targeting, distinct governance, and positioning that manages the signaling cost rather than ignoring it.

Pricing architecture matters disproportionately, but free cannot close the purpose gap. Moving from \$10 to \$5 per month raises interest by 0.321 points; moving to free raises it by 1.612 points, consistent with zero-price effects (Shampanier et al., 2007). Free is therefore not a large discount but a different regime. The limit is equally sharp. Section 5.3 shows the free premium is *attenuated* precisely for relational purposes, because the residual barrier is non-monetary and a price of zero leaves it intact. Two implications follow. The counterfactual simulations in Table 12 make the point concretely: moving a Romantic companion from \$10 to free raises the simulated share above the participation threshold from 9.6% to 25.0%, still 11.5 points below a free Admin assistant. Free entry points are therefore an acquisition instrument and should be engineered as a conversion system rather than a one-time discount, with trial duration, upgrade timing, and onboarding intensity targeted rather than uniform (Lambrecht and Misra, 2017; Yoganarasimhan et al., 2023). And monetization of relational users should run through premium privacy, personalization, or bundled instrumental features rather than willingness to pay for relational use alone.

Privacy and personalization are priced; modality is not. Relative to high privacy, limited retention and standard privacy are worth roughly $-\$1.24$ and $-\$0.75$ per month; basic rather than advanced personalization costs about \$1.04 per month. Interaction modality has small and imprecisely estimated effects. These magnitudes are economically meaningful at consumer app price points, and they push against a naive reading of the privacy paradox (Acquisti and Grossklags, 2005): where retained data is emotionally sensitive, privacy salience is high enough to move stated choice. Privacy-by-design (data minimization, retention controls, transparent settings) is thus a competitive attribute with measurable value rather than a compliance constraint, and for top-of-funnel consideration it dominates adding sensory channels.

Safety policy is part of the value proposition. Relative to an adults-only baseline, the balanced safety regime lowers interest by 0.179 points ($\approx \$1.15$/month), while the strict/PG-13 regime is smaller and statistically indistinguishable from zero. Governance choices therefore affect perceived usefulness and fit, not only risk exposure: constraints that users read as misaligned with their intended use reduce demand independently of any specific blocked behavior. Safety policy is best treated as a design dimension with clearly communicated boundaries, user-facing controls where feasible, and defaults tested by segment.

The belief wedge implies low-observability design. Respondents believe others are more interested than they are, and the wedge, net of the instrumental baseline, is largest for relational purposes (about 0.50 interest points for Romantic). For a firm, the operative implication is that stated demand *understates* the addressable market, though by less than the raw wedge suggests: the structural model in Section 6.5 puts the individual suppression cost at about \$0.30/month for a Romantic companion, and in the taste distribution of Section 6.4 about 9% (linear) to 14% (censored) prefer a romantic companion to an administrative assistant at equal price, concentrated among current companion users. Design and distribution choices that lower perceived social-evaluation costs (discreet interaction modes, private billing descriptors, strong privacy defaults, de-stigmatizing bundles under wellness or study framing) therefore act on the binding constraint in a way that further feature investment does not.

Bottom line. Demand is shaped most by positioning (instrumental versus relational), pricing architecture (especially the discontinuity at free), and trust-related design (privacy and governance). The portfolio strategy that follows is to scale instrumental products with mass-market pricing and credible privacy signaling, and to treat relational companions as a segmented business whose growth depends on reducing observability and converting free users, not on price alone. These implications are drawn from stated preferences over hypothetical profiles and describe the structure of demand rather than market shares; they are best read as a ranking of levers, not as a forecast of uptake.

9 Conclusion

Conversational AI is beginning to mediate routine coordination, learning, and emotional support. This paper measures demand across those uses and shows that the demand estimate a firm or a regulator would obtain from any conventional instrument is a composite of two things that call for different responses.

The clearest pattern is a steep purpose gradient. Instrumental purposes attract substantially higher interest than relational companionship, and respondents require \$7.80 per month in compensation to accept a romantic companion in place of an administrative assistant, six to seven times the value of any non-price design feature we vary. Price is the largest non-purpose lever and is sharply nonlinear, with the \$5-to-free gain roughly four times the \$10-to-\$5 gain, but it cannot close the purpose gap: the zero-price premium is attenuated precisely for relational purposes, because the residual barrier is non-monetary and a price of zero leaves it intact. Beyond price, consumers value privacy and personalization; interaction modality does not move demand.

The central result is that part of the relational penalty is social image rather than taste, and that the part can be bounded. Second-order beliefs are compressed across purposes relative to own

interest, and the ratio of the Admin-relative belief–interest gap to the Admin-relative purpose penalty is statistically indistinguishable across the two relational purposes, the over-identifying restriction that a single stigma parameter implies. That same ratio *is* the welfare statistic: if the whole gap is suppression, about two-fifths of the measured romantic penalty, and a similar share of the friendship penalty, is a cost imposed by the prevailing social norm rather than a preference. The statistic does not involve the price slope, which insulates it from the nonlinearity that makes dollar-denominated conjoint welfare estimates fragile. It omits two components that would raise it (partial suppression of beliefs themselves, and the equilibrium premium released by a shift in the norm) while the belief-shrinkage adjustment and the structural estimates, which imply that only 10–15% of the gap is suppression, lower it. On balance, most of the relational penalty is taste.

A hierarchical Bayes demand model recovers the population taste distribution and locates the suppression precisely where the theory puts it. Social-image sensitivity shifts relational purpose coefficients more than instrumental ones and leaves price, privacy, personalization, and safety sensitivities unchanged, an asymmetry that a general response-style or scale-use explanation does not obviously generate. The distribution also reveals what the mean conceals: between 9% and 14% of the taste distribution prefers a romantic companion to an administrative assistant at equal price, concentrated among current companion users, so a category the average consumer rejects can still support a segment.

For firms, positioning dominates features, free pricing is an acquisition instrument rather than a solution to a non-monetary barrier, and the highest-return investments for relational products are those that reduce perceived observability rather than those that add capability. For regulators and platforms, the coordination failure runs in the opposite direction from the one familiar from social media: relational AI sits in a low-adoption equilibrium that is individually stable and socially suboptimal, and the instruments that address it facilitate rather than restrict adoption. For researchers, the general lesson is methodological. Wherever a product category carries a social-image cost, and consumer research is full of them, demand estimates from stated *and* revealed preference are jointly contaminated, because the same stigma that suppresses survey responses suppresses search, download, and purchase. Eliciting beliefs about others alongside own demand costs one extra question per profile and bounds the contamination.

Conversational AI will diffuse first through instrumental use cases while relational companionship faces stronger headwinds. Those headwinds are mostly taste, but partly endogenous: they reflect a social norm that suppresses stated below latent demand, and a market that responds to stated demand. Measuring demand together with perceived norms, and correcting for the wedge between them, gives firms and regulators a way to tell the two apart.

References

Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 3(1):26–33, 2005. doi: 10.1109/MSP.2005.22.

AI Companion Guides. AI companion pricing guide 2026: Every platform compared. `https://aicompanionguides.com/blog/ai-companion-pricing-guide-2026/`, 2026. Accessed June 2026.

Hunt Allcott and Dmitry Taubinsky. Evaluating behaviorally-motivated policy: Experimental evidence from the lightbulb market. *American Economic Review*, 105(5):2501–2538, 2015. doi: 10.1257/aer.20131564.

Greg M. Allenby and Peter E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1–2):57–78, 1998. doi: 10.1016/S0304-4076(98)00055-4.

Backlinko. ChatGPT statistics: How many people use ChatGPT? `https://backlinko.com/chatgpt-stats`, 2026. Accessed June 2026.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

B. Douglas Bernheim. A theory of conformity. *Journal of Political Economy*, 102(5):841–877, 1994.

B. Douglas Bernheim and Antonio Rangel. Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124(1): 51–104, 2009. doi: 10.1162/qjec.2009.124.1.51.

Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. My ai friend: How users of a social chatbot understand their human–ai friendship. *Human Communication Research*, 48(3): 404–429, 2022.

Leonardo Bursztyn, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. Justifying dissent. *The Quarterly Journal of Economics*, 138(3):1403–1451, 2023.

Leonardo Bursztyn, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth. When product markets become collective traps: The case of social media. *American Economic Review*, 115(12):4105–4136, 2025a.

Leonardo Bursztyn, Alex Imas, Rafael Jiménez-Durán, Aaron Leonard, and Christopher Roth. Social dynamics of ai adoption. Technical report, National Bureau of Economic Research, 2025b.

El Hadi Caoui. Estimating consumer preferences for LLMs: Evidence from LMArena. Working paper, SSRN, 2025. Posted December 26, 2025. Available at SSRN: https://ssrn.com/abstract=5970054.

Daniel Carvajal, Catalina Franco, and Siri Isaksson. Will artificial intelligence get in the way of achieving gender equality? Revised version in NHH Discussion Paper Series, 2025.

Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019. doi: 10.1177/0022243719851788.

Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. How people use chatgpt. Technical report, OpenAI, September 2025. URL https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf. Working paper.

Emanuele Colonnelli, Bo Li, and Ernest Liu. Investing with the government: A field experiment in china. *Journal of Political Economy*, 132(1):248–294, 2024.

Zoë Cullen, Will Dobbie, and Mitchell Hoffman. Increasing the demand for workers with a criminal record. *The Quarterly Journal of Economics*, 138(1):103–150, 2023.

Marie-Pierre Dargnies, Rustamdjan Hakimov, and Dorothea Kübler. Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*, forthcoming. Forthcoming.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2015. doi: 10.1037/xge0000033.

Raymond M. Duch and Thomas S. Robinson. How to detect heterogeneity in conjoint experiments. *Journal of Politics*, 86(2):412–427, 2023. doi: 10.1086/727597.

Electroiq. AI companions statistics by usage, market size, apps and facts. https://electroiq.com/stats/ai-companions-statistics/, 2025. Accessed June 2026.

Benjamin Enke, Thomas Graeber, and Ryan Oprea. Confidence, self-selection, and bias in the aggregate. *American Economic Review*, 113(7):1933–1966, 2023.

Christine L Exley, Raymond Fisman, Judd B Kessler, Louis-Pierre Lepage, Xiaomeng Li, Corinne Low, Xiaoyue Shan, Mattie Toma, and Basit Zafar. Information-optional policies and the gender concealment gap. Technical report, National Bureau of Economic Research, 2024.

Google. Faq about google trends data. https://support.google.com/trends/answer/4365533. Accessed 2025-12-12.

Google News Initiative. Google trends: Understanding the data. https://newsinitiative.withgoogle.com/resources/trainings/google-trends-understanding-the-data/. Accessed 2025-12-12.

Paul E. Green and V. Srinivasan. Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54(4):3–19, 1990. doi: 10.1177/002224299005400402.

Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1):1–30, 2014. doi: 10.1093/pan/mpt024.

Jens Hainmueller, Dominik Hangartner, and Teppei Yamamoto. Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8):2395–2400, 2015.

Judd B. Kessler, Corinne Low, and Colin D. Sullivan. Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review*, 109(11):3713–3744, 2019.

Judd B Kessler, Corinne Low, and Xiaoyue Shan. Lowering the playing field: Discrimination through sequential spillover effects. *Review of Economics and Statistics*, pages 1–28, 2024.

Pavel Kireyev and Maria Ana Vitorino. User preferences for large language model refusals: Implications for moderation and market structure. Working paper, London School of Economics and Political Science, 2025. LSE Research Online. Unpublished manuscript. Available at https://eprints.lse.ac.uk/130606/.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *New Media & Society*, 26 (10):5923–5941, 2024.

Anja Lambrecht and Kanishka Misra. Fee or free: When should firms charge for online content? *Management Science*, 63(4):1150–1165, 2017. doi: 10.1287/mnsc.2015.2383.

Yier Ling, Alex Kale, and Alex Imas. Underreporting of ai use: The role of social desirability bias. Technical Report 2025-144, Becker Friedman Institute for Economics, University of Chicago, 2025. 46 pages, Posted 4 December 2025.

Jordan J. Louviere, David A. Hensher, and Joffre D. Swait. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, Cambridge, 2000. doi: 10.1017/CBO9780511753831.

Corinne Low. Pricing the biological clock: Reproductive capital on the us marriage market. *Wharton School of Business 17*, 2014.

Elisa Macchi. Worth your weight: experimental evidence on the benefits of obesity in low-income countries. *American Economic Review*, 113(9):2287–2322, 2023.

Vicki G. Morwitz. Purchase intentions and their role in consumer decision making. In Curtis P. Haugtvedt, Paul M. Herr, and Frank R. Kardes, editors, *Handbook of Consumer Psychology*, pages 576–604. Lawrence Erlbaum Associates, New York, 2008.

Sarah Moshary, Bradley T. Shapiro, and Sara Drango. Preferences for firearms. *American Economic Review: Insights*, 7(3):340–356, 2025. doi: 10.1257/aeri.20230435.

Sarah Perez. AI companion apps on track to pull in $120M in 2025. `https://techcrunch.com/2025/08/12/ai-companion-apps-on-track-to-pull-in-120m-in-2025`, 2025. Data from Appfigures. Accessed June 2026.

Vithala R. Rao. *Applied Conjoint Analysis*. Springer, Berlin and Heidelberg, 2014. doi: 10.1007/978-3-540-87753-0.

Julian Reif. The social penalty for unconventional behavior. Working paper, 2025.

Peter E. Rossi. *bayesm: Bayesian Inference for Marketing/Micro-Econometrics*, 2019. URL `https://CRAN.R-project.org/package=bayesm`. R package version 3.1-4.

Peter E. Rossi, Greg M. Allenby, and Robert McCulloch. *Bayesian Statistics and Marketing*. John Wiley & Sons, Chichester, 2005. doi: 10.1002/0470863692.

Nikola Roza. Replika AI: Statistics, facts and trends guide for 2026. `https://nikolaroza.com/replika-ai-statistics-facts-trends/`, 2026. Accessed June 2026.

Thomas E. Ruggiero. Uses and gratifications theory in the 21st century. *Mass Communication and Society*, 3(1):3–37, 2000. doi: 10.1207/S15327825MCS0301_02.

Kristina Shampanier, Nina Mazar, and Dan Ariely. Zero as a special price: The true value of free products. *Marketing Science*, 26(6):742–757, 2007. doi: 10.1287/mksc.1060.0254.

SQ Magazine. Character AI statistics 2026: Shocking user growth. `https://sqmagazine.co.uk/character-ai-statistics/`, 2025. Accessed June 2026.

Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, 22(3): e16235, 2020.

WeavAI. Candy AI review 2026: Features, pricing & guide. https://weavai.app/blog/en/2026/04/20/candy-ai-review-2026-features-pricing-guide/, 2026. Accessed June 2026.

WTMF Insights. Best AI friend apps for emotional support in 2026: An honest comparison. https://wtmf.ai/blog/best-ai-friend-apps-for-emotional-support-in-2026-an-honest-comparison, 2026. Accessed June 2026.

Hema Yoganarasimhan, Ebrahim Barzegary, and Abhishek Pani. Design and evaluation of optimal free trials. *Management Science*, 69(6):3220–3240, 2023. doi: 10.1287/mnsc.2022.4394.

YouGov. Yougov survey: Artificial intelligence. https://d3nkl3psvxxpe9.cloudfront.net/documents/Artificial_Intelligence_poll_results_gmJf8SM.pdf, August 2025. Fieldwork: Aug 8–11, 2025. U.S. adult citizens, N=1,122. Poll results PDF.

For online publication only:

Demand for Conversational AI

A Additional Tables

Table A.1: Overview of conversational AI platforms by category, pricing, and user base (2025–2026).

Platform	Developer	Category	Pricing	User base	Founded
Wellness / Wellbeing					
Wysa	Touchkin	Wellness	Free; $29.99/mo (premium)	6M+ users	2016
Woebot	Woebot Health	Wellness	Free (consumer); enterprise licensing	∼1.5M (B2B)	2017
Youper	Youper Inc.	Wellness	Free; $9.99/mo (premium)	∼3M users	2016
Headspace	Headspace	Wellness	Free trial; $12.99/mo or $69.99/yr	70M+ app users	2024[a]
Study / Education					
Khanmigo	Khan Academy	Study	Free (teachers); $4/mo (families)	135M (platform)	2023
Duolingo Max	Duolingo	Study	Free; Super $6.99/mo; Max $30/mo	40M DAU (all tiers)	2023
Admin / Productivity					
MS Copilot	Microsoft	Admin, Study	Free; $30/mo (M365 Copilot)	Enterprise-scale	2023

Continued on next page

Table A.1 – continued

Platform	Developer	Category	Pricing	User base	Founded
ChatGPT	OpenAI	Admin, Study, Wellness	Free; $20/mo (Plus); $25/mo (Pro)	900M+ weekly	2022
Friendship					
Pi	Inflection AI	Friendship, Wellness	Free	∼4M users	2023
Snapchat My AI	Snap Inc.	Friendship, Admin	Free; Snapchat+ $3.99/mo	150M users	2023
Kindroid	Kindroid Inc.	Friendship, Romantic	Free; ∼$11.99/mo	1.2M downloads	2023
Romantic					
Replika	Luka, Inc.	Romantic, Friendship, Wellness	Free; $19.99/mo; $299.99 (lifetime)	25M registered	2017
Character.AI	Character Tech.	Romantic, Friendship, Wellness	Free; $9.99/mo (c.ai+)	20M monthly	2022
Candy AI	EverAI	Romantic, Friendship	Free; $5.99/mo (annual); $12.99/mo	11.6M monthly	2023
Xiaoice	Microsoft China	Romantic, Friendship, Wellness	Free (China market)	660M+ registered	2014

[a] AI coaching layer launched 2024; app founded 2010.

Note. User figures are best available public estimates and may reflect different reporting periods (MAU, DAU, or registered accounts). "Free" indicates a meaningful no-cost tier alongside paid tiers. Sources: Backlinko (2026), Electroiq (2025), Roza (2026), WTMF Insights (2026), WeavAI (2026), AI Companion Guides (2026).

Table A.2: Summary of survey waves

	Wave 1 Main Conjoint	Wave 2 Obfuscated Follow-Up
N	1,989	1,082
Timing	24–29 Nov. 2025	3–4 weeks after Wave 1
Task	Rate each of five AI app profiles (0–10)	Rank five AI purposes (1–5)
Outcomes elicited	Own interest; second-order beliefs (0–100)	Purpose ranking
Attributes randomized	Price, modality, personalization, privacy, safety	None
Purpose	Estimate demand and WTP; test signaling predictions	Robustness check against hypothetical bias and demand effects
Incentivization	Personalized app recommendation delivered post-survey	None

Notes: Both waves recruited from the same Prolific panel. Wave 2 respondents represent 54% of the Wave 1 analysis sample. The survey interface, question order, and number of screens were altered between waves to make the two surveys appear unrelated.

Table A.3: Summary statistics

a, Continuous variables

Variable	*N*	Mean	SD	Min	Median	Max
Age (years)	1,989	46.88	15.51	20	47	87
Household income ($1,000s)	1,960	83.31	70.12	5	65	625
AI use frequency (0–4)	1,989	1.34	1.03	0	1	4

b, Categorical variables

Variable	Category	Share (%)
Gender	Female	50.4
Race/ethnicity	White	73.0
Education	4-year college degree	36.2
	High school graduate	24.7
	Postgraduate degree (MA, MBA, MD, JD, PhD, etc.)	16.6
	2-year college degree	14.2
	Post-high school vocational certification	7.5
	Did not graduate from high school	0.8
Employment	Working full time now	52.4
	Working part time now	16.6
	Retired	12.7
	Unemployed	8.2
	Taking care of home or family	4.7
	Student	2.5
	Other	1.6
	Receiving benefits or insurance	1.1
	Temporarily laid off	0.2
Census region	South	39.7
	Midwest	21.0
	West	20.1
	Northeast	19.2
Party	Independent	40.3
	Democrat	31.2
	Republican	27.7
	Other	0.5
	Not sure	0.4
AI use frequency	Sometimes	56.6
	Never	15.7
	Most of the time	11.7
	About half the time	11.0
	Always	5.0
Uses AI conversation app	No	72.7
	Yes	27.3
Stable partner	Yes	61.8
	No	36.6
	Prefer not to say	1.5

Table A.4: Attribute balance across all profiles

Attribute	Level	Share of profiles (%)
Price	Free	33.3
	$10/month	34.1
	$5/month	32.6
Interaction	Text + Voice	33.1
	Text + Voice + photo	33.5
	Text only	33.4
Personalization	Advanced	33.0
	Basic	33.5
	Moderate	33.4
Privacy	High privacy	33.7
	Limited retention	33.4
	Standard privacy	32.9
Safety	Adults only	32.9
	Balanced	33.0
	Strict (PG-13)	34.1

Table A.5: Privacy × purpose interactions (respondent FE)

Interaction	Coef.	SE	p (two-sided)	p (one-sided)
Limited retention × purpose (baseline: High privacy)				
Limited retention × Study	0.075	0.166	0.653	—
Limited retention × Wellness	0.127	0.157	0.417	—
Limited retention × Friend	−0.049	0.158	0.756	0.378
Limited retention × Romantic	−0.208	0.160	0.193	0.097
Standard privacy × purpose (baseline: High privacy)				
Standard privacy × Study	0.004	0.175	0.981	—
Standard privacy × Wellness	−0.058	0.159	0.717	—
Standard privacy × Friend	−0.086	0.161	0.592	0.296
Standard privacy × Romantic	−0.074	0.165	0.653	0.326

Notes: Dependent variable is interest (0–10). The specification adds all privacy × purpose interactions to the respondent-FE conjoint model with price-level indicators. Baselines: Admin, High privacy, $10/month, Text + Voice, Advanced personalization, Adults-only safety. One-sided p-values test the signaling prediction (negative interaction, i.e. lower privacy hurts relational purposes more) and are reported only for relational purposes. Respondent-clustered standard errors. $N = 1{,}989$ respondents; 9,945 observations.

Table A.6: Conjoint regressions: pooled OLS and respondent fixed effects

	(1) Pooled OLS	(2) Respondent FE	(3) FE + order
Study	0.155^{***}	0.153^{***}	0.152^{***}
	(0.058)	(0.058)	(0.058)
Wellness	-0.086	-0.093	-0.093
	(0.057)	(0.057)	(0.057)
Friend	-0.569^{***}	-0.572^{***}	-0.571^{***}
	(0.059)	(0.059)	(0.059)
Romantic	-1.257^{***}	-1.262^{***}	-1.261^{***}
	(0.060)	(0.060)	(0.060)
$5/month	0.290^{***}	0.321^{***}	0.320^{***}
	(0.072)	(0.050)	(0.050)
Free	1.514^{***}	1.612^{***}	1.612^{***}
	(0.080)	(0.062)	(0.062)
Text only	-0.171^{**}	-0.046	-0.047
	(0.074)	(0.053)	(0.053)
Text + Voice + photo	-0.073	0.026	0.026
	(0.073)	(0.051)	(0.051)
Basic personalization	-0.204^{***}	-0.160^{***}	-0.162^{***}
	(0.074)	(0.052)	(0.052)
Moderate personalization	-0.100	-0.088^{*}	-0.091^{*}
	(0.073)	(0.052)	(0.052)
Standard privacy	-0.038	-0.124^{**}	-0.123^{**}
	(0.074)	(0.050)	(0.050)
Limited retention	-0.114	-0.206^{***}	-0.206^{***}
	(0.074)	(0.052)	(0.052)
Strict (PG-13) safety	-0.103	-0.072	-0.070
	(0.076)	(0.056)	(0.056)
Balanced safety	-0.236^{***}	-0.179^{***}	-0.177^{***}
	(0.074)	(0.052)	(0.052)
Constant	2.879		
	(0.123)		
Position dummies	No	No	Yes
Observations	9,945	9,945	9,945
Respondents	1,989	1,989	1,989

Notes: Dependent variable is the interest rating (0-10). Baseline levels are Admin (purpose), $10/month (price), Text+Voice (interaction), Advanced (personalization), High privacy, Adults-only (safety), and position 1 (order). Column (1) is pooled OLS with a constant. Columns (2)-(3) include respondent fixed effects (within estimator) and therefore have no intercept. Standard errors are clustered by respondent.

Table A.7: Sensitivity of WTP to the price-to-utility mapping

	\$10→\$5 slope (0.0641)	\$5→Free slope (0.2584)	Average slope (0.1614)
Study	−2.48	−0.62	0.99
Wellness	1.26	0.31	−0.50
Friend	8.75	2.17	−3.48
Romantic	19.65	4.88	−7.80
Text only	0.70	0.17	−0.28
Text + Voice + photo	−0.45	−0.11	0.18
Basic personalization	2.63	0.65	−1.04
Moderate personalization	1.55	0.39	−0.62
Standard privacy	1.88	0.47	−0.75
Limited retention	3.13	0.78	−1.24
Strict (PG-13) safety	1.15	0.28	−0.45
Balanced safety	2.88	0.71	−1.14

Notes: Price effects are non-linear: the jump from paid to free is larger than the difference between \$10 and \$5. This table reports alternative money-metric valuations using (i) the local slope between \$10 and \$5, (ii) the local slope between \$5 and free, and (iii) the average slope between \$10 and free.

Table A.8: Beliefs about others' demand by purpose

Purpose	Own interest (0–10)	Belief (0–100)	Belief/10	Gap $G(p)$	$G^*(p)$
Admin	3.13	42.24	4.22	1.092	—
Study	3.28	42.44	4.24	0.967	−0.125 (0.047)
Wellness	3.07	40.66	4.07	1.000	−0.092 (0.045)
Friend	2.56	39.17	3.92	1.353	0.262 (0.047)
Romantic	1.86	34.46	3.45	1.590	0.498 (0.050)

Table A.9: Multiple-testing adjustments for conjoint FE coefficients

Coefficient	Estimate	p	p (BH)	p (Bonferroni)
Study	0.153	0.009	0.015	0.121
Wellness	−0.093	0.104	0.133	1.000
Friend	−0.572	0.000	0.000	0.000
Romantic	−1.262	0.000	0.000	0.000
$5/month	0.321	0.000	0.000	0.000
Free	1.612	0.000	0.000	0.000
Text only	−0.046	0.382	0.411	1.000
Text + Voice + photo	0.026	0.613	0.613	1.000
Basic personalization	−0.160	0.002	0.004	0.030
Moderate personalization	−0.088	0.093	0.130	1.000
Standard privacy	−0.124	0.014	0.022	0.195
Limited retention	−0.206	0.000	0.000	0.001
Strict (PG-13) safety	−0.072	0.204	0.238	1.000
Balanced safety	−0.179	0.001	0.001	0.008

Notes: Table reports unadjusted p-values and adjusted p-values using Benjamini-Hochberg (BH) false discovery rate control and Bonferroni correction for the 14 conjoint coefficients in the respondent fixed-effects model.

Table A.10: Order effects (display position) in the FE model

Display position	Coefficient	SE
Position 2	0.057	0.055
Position 3	0.009	0.059
Position 4	−0.037	0.058
Position 5	−0.056	0.059

Notes: Coefficients are relative to the first displayed profile (position 1). Standard errors are respondent-clustered.

Table A.11: Conjoint estimates by AI conversation-app use (respondent FE)

	Users ($N = 543$)		Non-users ($N = 1{,}446$)		Diff.
Variable	Coef.	SE	Coef.	SE	(U−NU)
Purpose (baseline: Admin)					
Friend	−0.609	0.132	−0.569	0.064	−0.040 (0.146)
Romantic	−1.569	0.138	−1.161	0.064	−0.408*** (0.152)
Study	0.091	0.127	0.169	0.064	−0.079 (0.142)
Wellness	−0.228	0.125	−0.053	0.062	−0.174 (0.140)
Price (baseline: $10/month)					
$5/month	0.369	0.115	0.299	0.053	0.070 (0.127)
Free	2.091	0.140	1.437	0.067	0.654*** (0.155)
Privacy (baseline: High)					
Standard privacy	−0.234	0.117	−0.088	0.054	−0.145 (0.128)
Limited retention	−0.405	0.123	−0.139	0.055	−0.267** (0.134)
Cross-equation ratios					
$R(\mathsf{Friend})$	0.407		0.482		
$R(\mathsf{Romantic})$	0.391		0.393		

Notes: Each column pair is a respondent-FE conjoint regression (price-level indicators) for the indicated subgroup. Users are respondents who report currently using an AI app to have conversations. The difference column is the coefficient on the user interaction in a pooled FE model that interacts every attribute with user status, with respondent-clustered standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.12: Cross-equation restriction: proportionality ratios and bootstrap test

	Friend	Romantic	Difference
Panel A: Components ($\varphi = 1/10$)			
Purpose penalty (FE, rel. Admin)	0.561	1.260	
SOB gap $G^*(p)$ (rescaled 0–10)	0.262	0.498	
Panel B: Ratio $R(p) = G^*(p)/[-\hat{\beta}(p)]$			
Ratio	0.466	0.395	0.071
Bootstrap SE	0.067	0.033	0.058
95% CI	[0.334, 0.595]	[0.328, 0.457]	[−0.043, 0.179]
p-value, equality ($\varphi = 0.10$)			0.217
Difference / p at $\varphi = 0.08$			0.057 / 0.226
Difference / p at $\varphi = 0.12$			0.085 / 0.217

Notes: Purpose penalties are numeric-price FE conjoint coefficients (baseline Admin), sign-reversed. $G^*(p) = [\varphi\bar{s}(p) - \bar{r}(p)] - [\varphi\bar{s}(\mathsf{Admin}) - \bar{r}(\mathsf{Admin})]$. Respondent-level bootstrap ($B = 2{,}000$) re-estimating the FE denominator and the belief and interest means jointly in each draw. $N = 1{,}989$ respondents.

Table A.13: Fixed-effects vs. HB population-mean estimates

	Fixed effects		HB population mean	
Variable	Coef.	SE	Median	95% CrI
Study	0.159	0.059	0.161	[0.066, 0.261]
Wellness	−0.081	0.058	−0.028	[−0.120, 0.067]
Friend	−0.561	0.059	−0.363	[−0.463,-0.273]
Romantic	−1.260	0.061	−0.920	[−1.020,-0.820]
Price ($/month)	−0.161	0.006	−0.147	[−0.159,-0.135]
Text only	−0.045	0.053	−0.093	[−0.174,-0.015]
Text + Voice + photo	0.029	0.051	−0.008	[−0.086, 0.071]
Basic personalization	−0.169	0.052	−0.167	[−0.249,-0.079]
Moderate personalization	−0.100	0.053	−0.080	[−0.157, 0.003]
Standard privacy	−0.120	0.051	−0.082	[−0.165, 0.008]
Limited retention	−0.201	0.052	−0.141	[−0.231,-0.061]
Strict (PG-13) safety	−0.073	0.057	−0.027	[−0.113, 0.050]
Balanced safety	−0.185	0.052	−0.150	[−0.229,-0.068]
Respondents	1,989		1,989	

Notes: FE standard errors clustered by respondent (numeric price). HB population means evaluated at the average covariate vector $\bar{\mathbf{z}}$; posterior medians and 95% equal-tailed credible intervals from three chains (3,000 retained draws). Baselines: Admin, Text + Voice, Advanced personalization, High privacy, Adults-only safety.

Table A.14: Distribution of individual WTP relative to Admin (linear HB, US$/month)

	p5	p10	p25	p50	p75	p90	p95	Share $\beta_i > 0$
Romantic	−35.6	−19.5	−9.2	−5.1	−2.8	−1.0	0.7	9.0%
Friend	−17.8	−9.6	−4.2	−1.9	−0.5	1.1	3.0	20.3%
Study	−5.7	−2.8	−0.6	0.9	2.7	6.7	12.6	64.9%

Notes: Quantiles of $\text{WTP}_{i,j} = -\beta_i^j / \beta_i^m$ evaluated within each retained draw and pooled over draws and respondents, restricted to draws with $\beta_i^m < 0$ (80% of all draws; the remainder have a wrong-signed or zero price coefficient and no finite WTP). The last column is the share of all draws with a positive purpose coefficient, i.e. preferring the purpose to Admin at equal price.

Table A.15: Signaling extension: effect of social-image sensitivity on taste parameters (HB second-level coefficients)

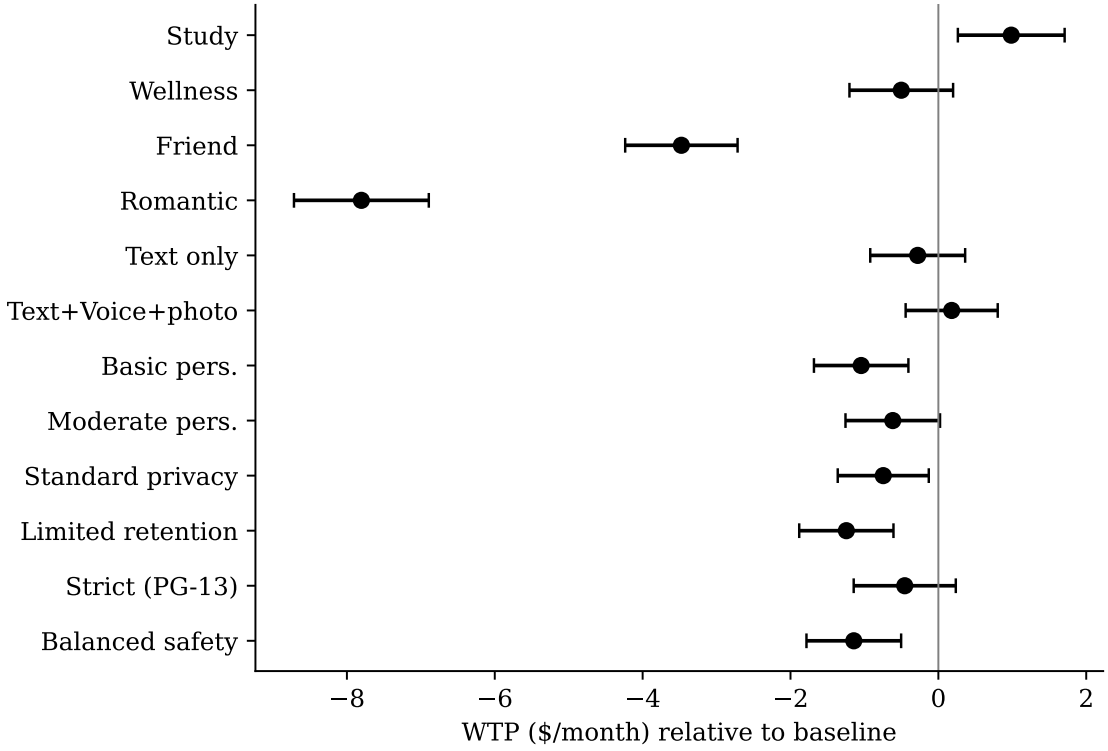
	(1)	(2)	(3)	(4)
Proxy built from	Friend	Romantic	Friend	Romantic
Differenced against	Admin	Admin	Study	Study
Romantic	-0.827	$-1.262^{\dagger}$	-0.324	$-0.775^{\dagger}$
Friend	$-1.181^{\dagger}$	-0.600	$-0.652^{\dagger}$	-0.171
Study	-0.472	-0.406	$0.633^{\dagger}$	$0.590^{\dagger}$
Wellness	-0.605	-0.394	-0.117	0.015
Price ($/month)	-0.010	-0.014	-0.011	-0.013
Limited retention	-0.027	-0.076	0.052	-0.001
Basic personalization	0.037	0.008	-0.019	-0.018
Balanced safety	0.030	-0.006	-0.053	-0.046
Text only	-0.026	-0.002	-0.047	0.045
Intercept	$0.682^{\dagger}$	$0.923^{\dagger}$	$0.147^{\dagger}$	$0.354^{\dagger}$
Target relational purpose	Romantic	Friend	Romantic	Friend
Correction, level ($/month)	0.65 [0.53, 0.76]	0.85 [0.68, 1.03]	0.30 [0.16, 0.46]	0.25 [0.06, 0.44]
Correction, vs. Wellness ($/month)	0.13 [0.02, 0.24]	0.25 [0.10, 0.40]	0.17 [0.01, 0.32]	0.28 [0.09, 0.47]

Notes: Each column adds one standardized belief–interest proxy as a second-level covariate in (19); entries are posterior medians of the corresponding row of Δ. †The proxy contains a rating that also identifies this coefficient, so the loading is mechanical. In columns (3)–(4) neither the target relational coefficient nor the Wellness coefficient shares a rating with the proxy; these are our preferred specifications. The correction shifts the proxy from its sample mean to zero and reports the change in the population-mean WTP for the target purpose, within draws; the "vs. Wellness" version uses the difference between the target and Wellness loadings. 95% credible intervals in brackets.

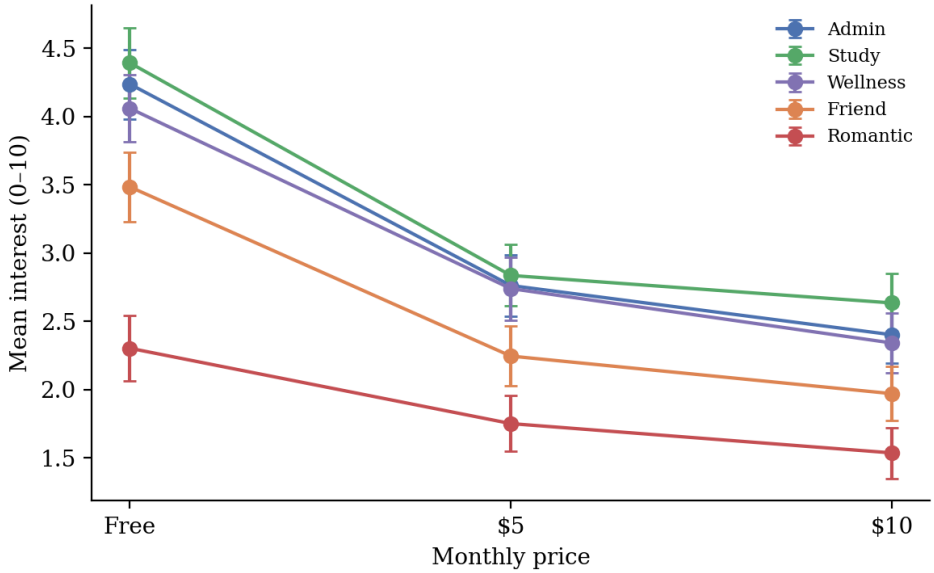
Table A.16: Reasons mentioned in open-ended explanations: representative verbatim responses

Reason	%	Example
Romance / dating / intimacy	59.6	I would prefer to use AI for functional purposes and use real people for intimate or romantic relationships. Wellness advice is ok; having a romantic relationship with AI is not.
Learning / study / research / information	55.1	I ranked these based on what I use these bots for. First I learn/study. Second I use them for general wellbeing. Third, administrative tasks are still important but not as much for me. Friend and romance were last because I already have friends and I have a partner. This was my thought process during the ranking.
Admin & organization / productivity	32.0	I have a great partner and don't need romantic AI. I also have many friends and feel a friendship AI would not help me much. I'd be open to using AI for productivity, admin tasks, and occasionally use it to learn about new things and also for work.
Wellness (general / habits / self-improvement)	30.6	I ranked wellness first because a wellness bot could support healthy habits and self-improvement in a way that feels useful. I then put study and admin next because they are practical, while friendship and romance feel less appropriate for AI interactions, which is why they both came in in fourth and fifth place.
Loneliness / companionship	14.6	I would use the AI for learning and wellness but I do not think I would use it for companionship. I might eventually talk to the AI like a friend.
Authenticity / prefer real people	13.8	I ranked the AI bots based on how appropriate I felt it was for an AI to fill that role. I felt the utilitarian purposes were best suited to AI, as a bot can help you with your tasks quickly, but friendship and romance should primarily come from real people rather than technology.
Social stigma / discomfort ("creepy", awkward)	8.2	I think it's weird to get personal with a bot but I can see it for studying. I wouldn't feel comfortable using it as a romantic partner. But if it can help me practically with school then why not.
Already have partner/friends / no need	7.7	I already have a partner and friends, so I don't need AI for romance or friendship. I'd rather focus on practical uses like studying or admin, and in terms of taking care of my health and that is why I have a doctor.
Mental health & emotional support	7.6	I ran wellness first because I have anxiety and a wellness/mental health bot could help with coping strategies and emotional support, but I am not comfortable or interested in ai fulfilling a romantic purpose.
General reluctance / dislike of AI	4.8	I chose wellness first because I could see some use there, but overall I don't want to use AI much. I'm not interested in a relationship and the others of no interest.
Accuracy / reliability concerns	3.6	I don't have enough trust in A.I when it comes to the social aspects of things. A.I isnt reliable when it come to friend or romantic advice.
Physical health / fitness / diet / sleep	3.3	I like to be able to use it for physical health, like exercise or diet or sleep advice, and for learning and other interests. I would rather have human friends, or a human romance.
Fun / curiosity / entertainment	2.4	I don't want or need a friend or romantic partner. I would use admin for schedules and tasks, and I like study because I enjoy getting helpful summaries of information that satisfies my curiosity.
Privacy / data security concerns	1.5	I would only want to use it for impersonal tasks because I worry about privacy and data security. I don't want to share personal feelings or relationships. Analyzing data or writing something for work is no issue.
Moral/ethical objection (unethical, wrong)	1.3	ai chatbots are unnatural and unethical. It would make me feel isolated and disconnected.
Dependency / overreliance concerns	0.4	I ranked the purposes so that romance came last because romantic intimacy risks emotional dependency without real reciprocity.

B Additional Figures



(a) WTP for feature changes relative to baselines.



(b) WTP curves across AI purposes.

Figure B.1: Willingness-to-pay patterns for AI companion features and purposes. Points show implied WTP or average interest; whiskers show 95% confidence intervals.

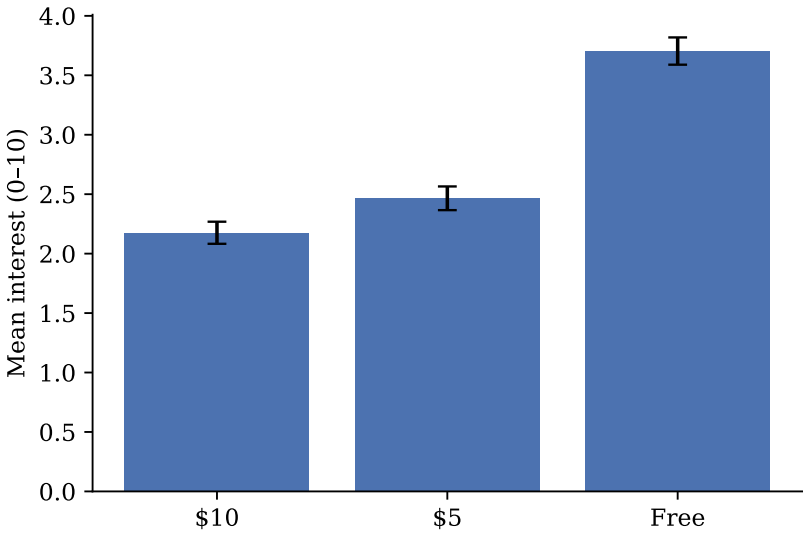


Figure B.2: Average interest by price level (pooled across purposes and other attributes). Bars show means; whiskers show 95% confidence intervals.

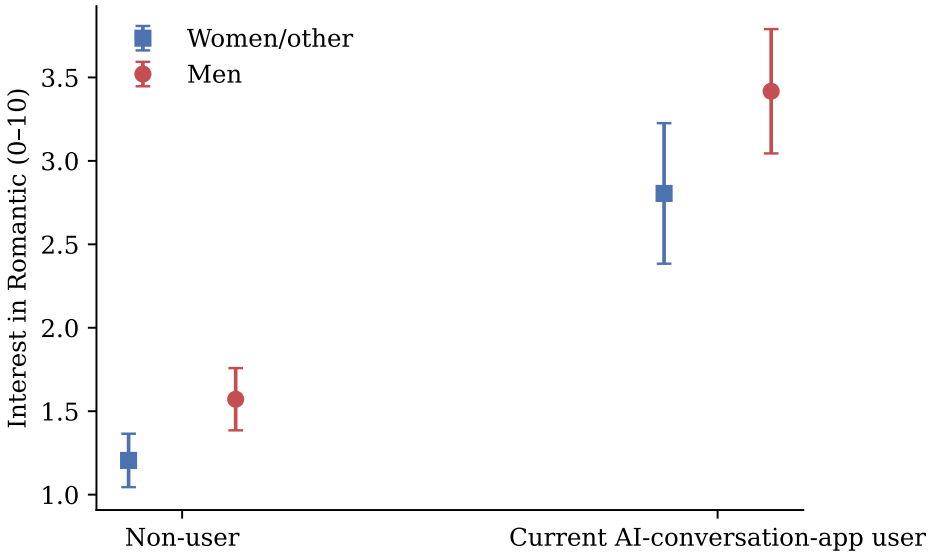


Figure B.3: Romantic-companion interest by gender and current AI companion use. Points show subgroup means; whiskers show 95% confidence intervals.

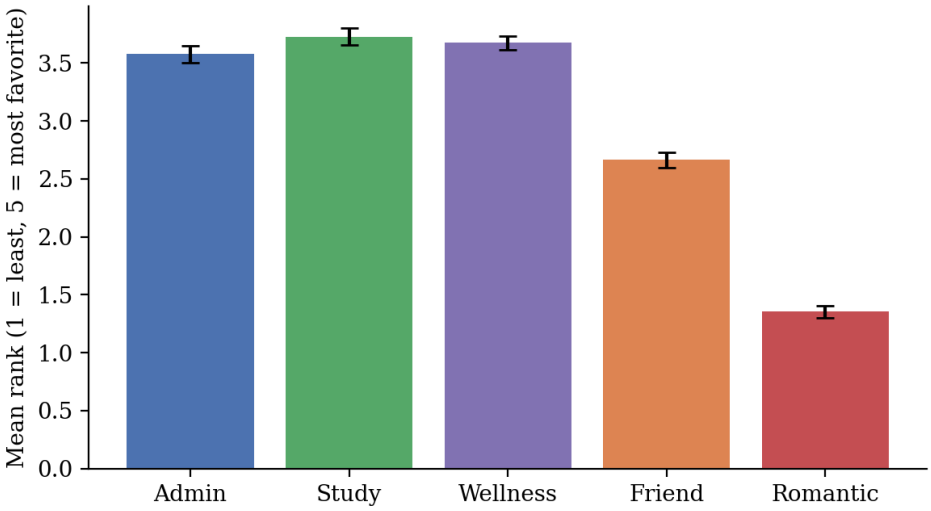


Figure B.4: Rankings of AI purposes on a scale from 1 (least favorite) to 5 (most favorite). Histograms show means; whiskers show 95% confidence intervals.

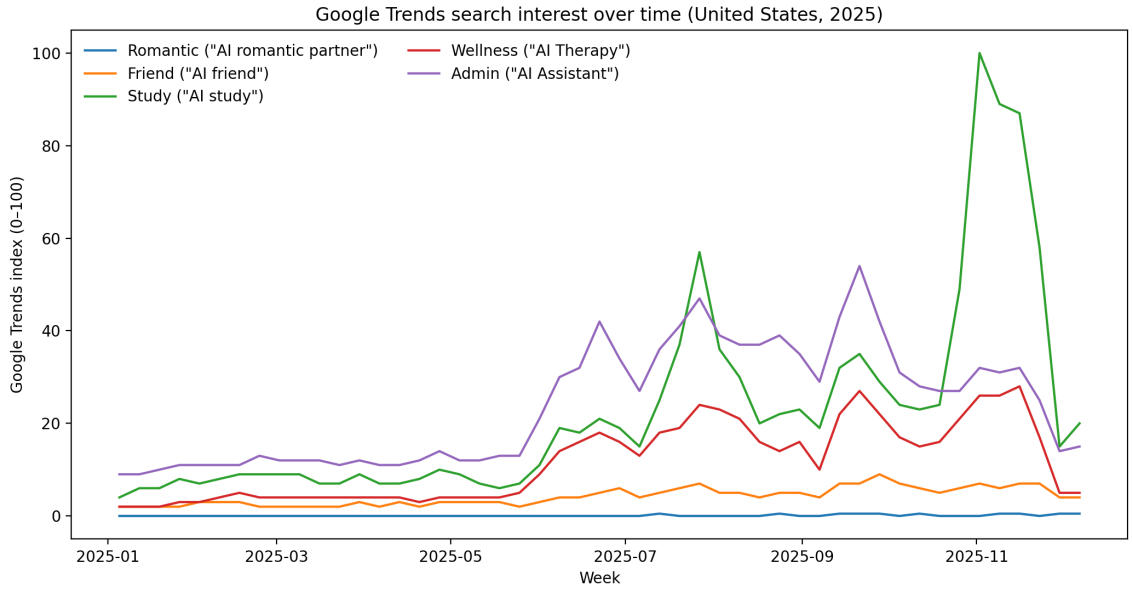


Figure B.5: Google Trends search interest over time (United States, 2025). Values are normalized (0-100) and represent relative (not absolute) search volume.

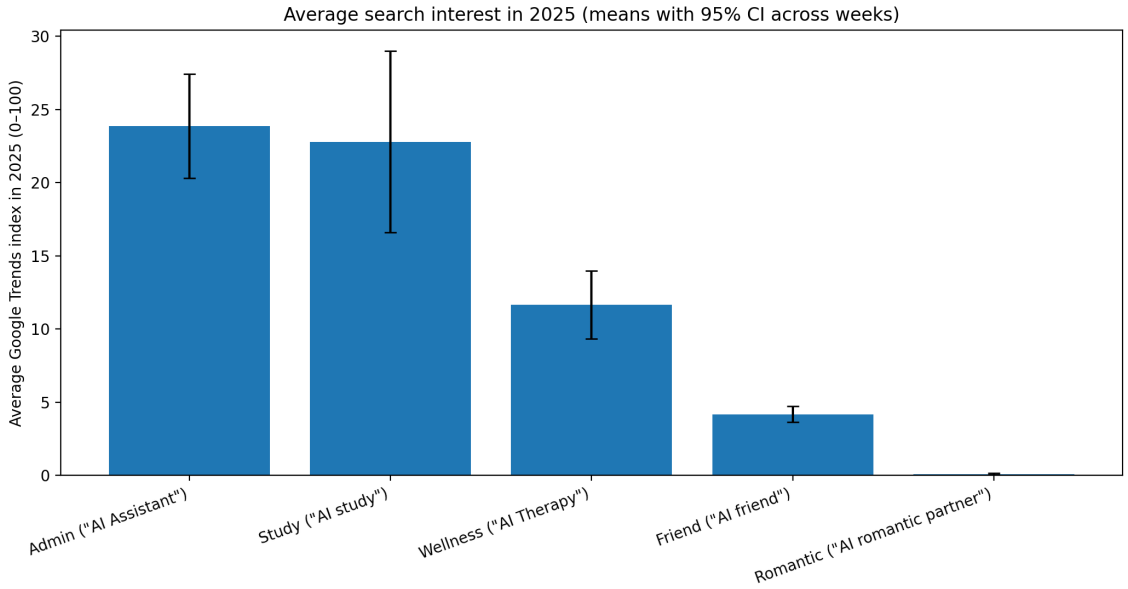


Figure B.6: Average Google Trends search interest in 2025 by query (means with 95% confidence intervals across weeks).

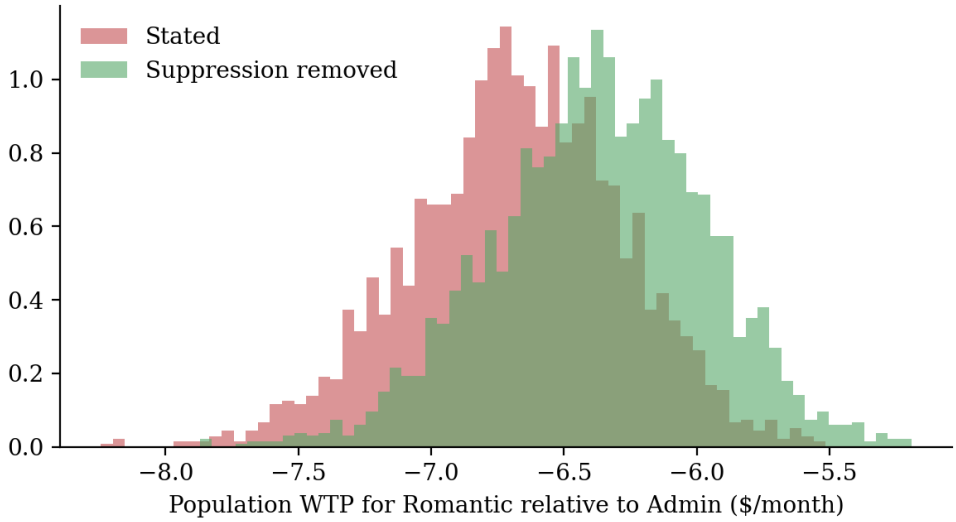


Figure B.7: Posterior distribution of the population-mean WTP for a Romantic AI companion relative to Admin, as stated (red) and after removing the measured social-image suppression (green), from the preferred signaling specification (Study-differenced proxy built from the Friend profile). The shift, +$0.30/month at the median, is the individual suppression cost that the belief–interest gap detects.

C Instructions

The full instructions for the main survey are available here, and those for the follow-up survey here.