HOW MUCH SHOULD A CONVERSATIONAL RECOMMENDER SYSTEM CONVERSE?

By

Akshit Kumar, Vahideh Manshadi and Akhilesh Tumu

May 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2535

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

How Much Should a Conversational Recommender System Converse?

Akshit Kumar
Yale University, akshit.kumar@yale.edu

Vahideh Manshadi
Yale University, vahideh.manshadi@yale.edu

Akhilesh Tumu
Yale University, akhilesh.tumu@yale.edu

Conversational recommender systems powered by generative AI can enhance personalization by facilitating information elicitation through follow-up questions. However, engaging in these conversations imposes a communication cost on users. As platforms with different objectives and monetization models deploy these systems, a central question is: *how does the platform's objective and sellers' strategic response shape the design of these systems in terms of their elicitation strategy?* We develop a parsimonious model of conversational elicitation in which interaction generates noisy preference information and imposes a communication cost borne by the user. A user-welfare-maximizing platform elicits more information when accurate niche matching yields large gains, even when niche users are rare. In contrast, under a conversion objective, for the same setting, the optimal strategy is to immediately recommend the same mainstream option to all users with no or minimal preference elicitation because the incremental conversion benefit from improved matching is bounded, while communication costs are borne by all users. When prices are endogenous and the platform earns a commission, increased elicitation is again optimal because improved screening raises equilibrium prices and platform revenue; however, these price responses can counteract consumer benefits and reduce user welfare. The model also highlights that the optimal elicitation intensity increases with preference heterogeneity, helping explain why conversational systems ask more in highly differentiated categories than in low-heterogeneity ones. We complement the theory with a dataset of long-form product queries that vary in length and informational content. Using our dataset and LLM-based user simulation, we quantify how additional information impacts user decisions and demonstrate that the magnitude of this impact depends on the degree of preference heterogeneity. Additionally, this dataset provides a testbed for measuring the (incremental) value of preference elicitation and may be of independent interest.

Key words: Conversational recommender systems, Preference elicitation, Human-AI Interaction

1. Introduction

Conversational interfaces are rapidly becoming a mainstream "front door" to discovery in digital markets. In e-commerce, Amazon has rolled out a generative-AI shopping assistant (Amazon

Rufus) that answers product questions and makes recommendations directly inside the chatbot environment[1]. In content streaming, Netflix has reported testing an AI-powered search feature that uses natural language to help users find relevant titles[2]. Together, these examples point to a broader shift: recommendation is moving from a one-shot ranking problem ("here are the top items") to an interactive decision problem in which the system can ask follow-up questions, reduce uncertainty about user preferences, and adapt recommendations in real time—capabilities that are fundamentally limited in traditional, non-interactive recommendation systems.

From the platform's perspective, however, asking questions is not free. Conversational elicitation trades off (i) improved matching and higher downstream satisfaction against (ii) conversational friction—time, attention, effort, and the risk of abandonment. In many conversational recommender systems, users must articulate preferences, respond to clarifying questions, and evaluate intermediate suggestions before reaching a final choice. User effort therefore becomes a first-order design constraint.

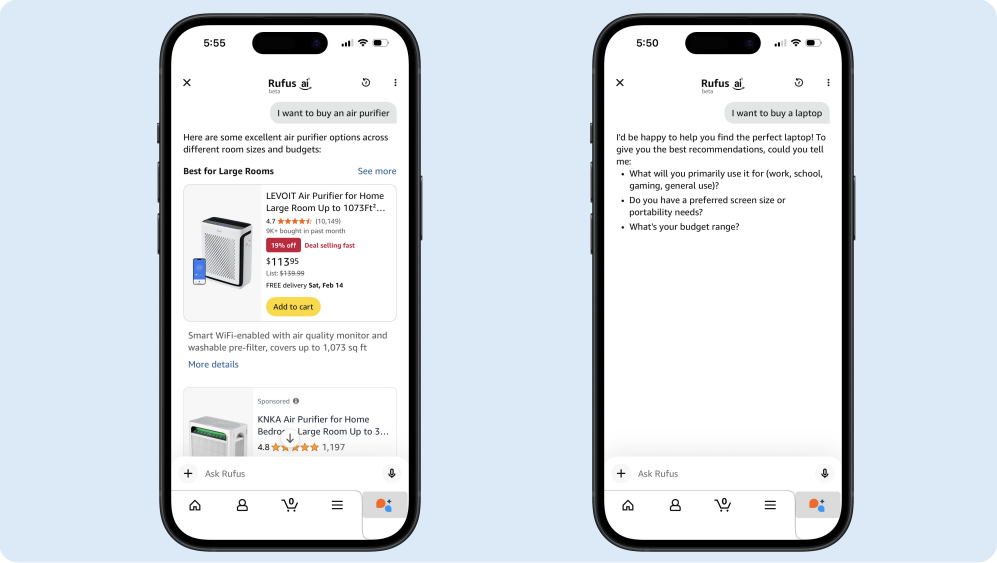


Figure 1 Two observed conversational strategies in Amazon Rufus. (Left) For air purifier, it immediately recommends products. (Right) For laptop, Rufus asks clarifying questions before recommending.

Figure 1 illustrates this design choice in practice. When we query Amazon Rufus for "laptop," it first asks targeted questions (use case, screen size, budget) before producing recommendations. When we query it for "air purifier," it skips elicitation and immediately suggests specific items. These behaviors are not cosmetic—they reflect different elicitation and recommendation strategies

[1] https://www.aboutamazon.com/news/retail/amazon-rufus

[2] https://about.netflix.com/en/news/unveiling-our-innovative-new-tv-experience

deployed by the same system across product categories. What drives these differences? A natural conjecture is that elicitation is most valuable in categories with higher horizontal differentiation (i.e., substantial heterogeneity in product types and in user preferences), where clarifying questions can sharply improve match quality.[3] In contrast, in categories with primarily vertical differentiation or relatively low heterogeneity, the marginal benefit of elicitation may be small, making quick recommendations (perhaps followed by a small number of questions) more attractive.

Yet heterogeneity alone may not fully determine optimal behavior, because the platform's objective matters as well. Preference elicitation is costly, and different platforms may internalize different costs and benefits. A content streaming platform's revenue may be driven by engagement and conversion, while for an e-commerce platform, the revenue may be driven by sales commission. This raises a broader question: do the same principles that guide the design and development of conversational recommendation systems in one domain like shopping (e.g., Amazon Rufus) carry over to another domain like content discovery (e.g., YouTube or Netflix) when the platform objectives or revenue models differ? Relatedly, can "more personalization" ever be counterproductive from the consumer's perspective—improving platform outcomes while reducing consumer welfare?

Despite rapid engineering progress and extensive systems research (e.g. Afzali et al. (2023), Bernard et al. (2025), Müller (2025)), there is limited understanding of when conversational elicitation should be used, how intensive it should be, how the elicitation and recommendation strategies depend on product and preference heterogeneity, and how these choices interact with platform objectives. This motivates our central research question: *How should a platform choose its conversational recommendation strategy as a function of market characteristics—preference heterogeneity, match value, and abandonment due to costly communication—and how do optimal strategies change with differing platform objectives?*

1.1. Our contribution

We summarize the key contributions of the paper below.

A parsimonious model of conversational elicitation with abandonment. We study a cold-start interaction in which a conversational recommender system (CRS) chooses (i) an elicitation intensity and (ii) a rule that maps conversational signals into a single recommended product. A user has latent type $\Theta \in \{\emptyset, 1, \ldots, n\}$: with probability $1-\alpha$ the user is well served by a mainstream option P, and with probability α the user matches one of n niches. Elicitation is summarized by a *fidelity* parameter f that determines the informativeness of a signal S about Θ through an $(n+1)$-ary symmetric channel. If the user reaches the recommendation stage (namely, does not

[3] Note that the user who captured these screenshots had no prior searches or purchases for these products on Amazon, suggesting that the platform's responses were not driven by the user's browsing or purchase history.

leave during the conversation process), she chooses whether to adopt the recommended product under a standard logit model.

The model incorporates two salient frictions. First, longer or more demanding interaction increases abandonment. We model conversation as gradually transmitting information, measured in *nats* (the natural-log unit of information), and assume a constant abandonment hazard *per nat*. Second, users incur effort costs while interacting and we model this by having a flow disutility parameter $\bar{\kappa}_u$ per nat.

Objective-dependent design and the returns to "hyper-personalization." Our central question is: *when should platforms invest in conversational elicitation as a hyper-personalization tool, and how does the answer depend on the platform's objective?* We study two platform objectives that capture common business models. Under (i) conversion/engagement monetization (fixed prices), the platform earns a constant payoff $R > 0$ when the user adopts the recommendation; expected payoff is discounted by survival. Under (ii) commission monetization, sellers set prices after observing CRS design, and the platform earns a take-rate τ on realized revenue. In both cases we track consumer welfare—expected surplus among surviving users minus effort costs borne during conversation—to assess alignment between platform-optimal design and consumer outcomes.

Our main theoretical results are organized around a rare-niche environment in which niche tastes are infrequent but the value of a correct niche match may be large. This regime provides a clean setting in which to separate two forces: (i) the potential surplus from correctly serving a small group of "rare-gems" users and (ii) the fact that conversational friction applies broadly through abandonment and effort.

The results yield clear implications for investment in conversational elicitation. First, for conversion or engagement-style objectives, recommend-first (minimal probing) is optimal in the rare-niche regime (Theorem 1). In contrast, the welfare-optimal policy would entail substantial elicitation (Proposition 1), implying a misalignment between engagement optimization and user welfare (Corollary 1). This provides a concrete interpretation for why platforms whose marginal revenue is tied mainly to getting users to consume *something* (e.g., session completion, watch-time, ad exposure) may rationally invest less in deep, high-resolution elicitation than a purely personalization-centric perspective would suggest.

Second, under commission monetization, elicitation can be platform optimal even when the fraction of niche seekers is small (Theorem 2). Here the benefit of conversation is not only improved matching; it is improved *screening*, which supports higher equilibrium prices and therefore higher commission revenue. This implies that marketplaces and e-commerce platforms can have stronger incentives to invest in hyper-personalization tools like conversational recommenders. At the same time, the induced consumer welfare effect can be positive or negative in the rare-niche regime

(Proposition 2). As a result, a platform may have incentives to invest in elicitation that may or may not be aligned with consumer welfare. We also provide comparative statics that map directly to within-platform cross-category conversational behavior observed in Figure 1: the impact of heterogeneity on the elicitation strategies (Proposition 3). Taken together, our theory has crisp design implications and provides a concrete lens on why engagement-driven platforms may rationally avoid deep elicitation, whereas commission-driven platforms may have incentives to invest in technologies that enable hyper-personalization such as recommendation chatbots.

An empirical testbed for incremental preference information. To complement the theory, we construct a synthetic evaluation dataset designed to measure the returns to additional preference information. Starting from long-form product queries in E-GEO (Bagga et al. 2025), we generate user personas with (i) a full "ground-truth" preference description and (ii) an ordered sequence of sentences that reveal preferences progressively, allowing us to simulate increasing elicitation depth (Section 4.1). We pair these personas with product listings from the Amazon Reviews dataset (Hou et al. 2024) to generate candidate recommendation sets, and use an LLM-based evaluator to produce purchase/no-purchase decisions and willingness-to-pay estimates, which we use as proxies for conversion and match-quality outcomes (Section 4.2). Our baseline empirical findings suggest that both conversion and revenue outcomes improve as more information is revealed, consistent with the simple premise that elicitation can sharpen matching. These gains, however, are substantially more significant for categories with more heterogeneous preferences (Figure 8). Building on this, we construct a case study: we simulate exogenous abandonment and compare the recommendation timing that maximizes conversion versus revenue outcomes (Section 4.4). This operationalization illustrates the central mechanism driving CRS elicitation choices: categories in which "niche" users have markedly higher simulated WTP exhibit a sharper divergence between conversion-oriented and revenue-oriented optimal elicitation (Figure 10). The dataset and evaluation protocol provide a practical testbed for quantifying the value of incremental preference information with differing heterogeneity and may be utilized to train/evaluate LLM-based retrieval and recommendation technologies.

1.2. Related Literature

Our paper connects to several strands of literature.

Conversational recommendation and Large language models. Conversational recommender systems (CRS) interleave *elicitation*—asking for attributes or feedback—with *recommendation* (Jannach et al. 2021, Peng et al. 2025). While early work viewed cold-start recommendation as a slot-filling or bandit problem (Christakopoulou et al. 2016, Li et al. 2018, Zhang et al. 2020), the advent of Large Language Models (LLMs) has fundamentally shifted the design space. Modern

LLMs can perform zero-shot recommendation (Sanner et al. 2023, He et al. 2023), purchase products on e-commerce websites (Allouah et al. 2025) and act as capable user simulators (Wang et al. 2023, Filippas et al. 2024, Sun et al. 2025, Manning and Horton 2025), though digital personas may exhibit distinct biases (Peng et al. 2026). We leverage these capabilities to construct our empirical testbed: rather than relying solely on static datasets, we use an LLM-as-a-judge (Zheng et al. 2023, Li et al. 2026) framework to simulate the *marginal* value of revealed preference information. This allows us to operationalize the trade-offs in our theoretical model using realistic product data, complementing our theoretical results.

Preference elicitation. There is a literature that studies optimal elicitation as a sequential decision problem, often focusing on how to select queries under limited interaction budgets. Examples include Boutilier (2002) (POMDP formulations of preference elicitation) and Blum et al. (2004) (query learning for preferences). In computational economics and mechanism design, preference elicitation appears in settings where eliciting full valuations is costly, e.g., Parkes (2005). Our model differs in both focus and structure: rather than optimizing query complexity for a single welfare objective, we characterize the *economic* forces that determine how much elicitation is justified when platforms optimize engagement or commission-based profit.

Strategic communication and costly information acquisition. Our model is closely related to classic work on costly information acquisition and discrete choice, including Weitzman (1979), McFadden (1974). Rational inattention provides a canonical information-theoretic cost framework: Sims (2003) introduces mutual-information costs in decision problems, Matějka and McKay (2015) microfound multinomial logit via costly information, and Caplin and Dean (2015) develop revealed-preference tools. Castro et al. (2024) study a model of human-AI interaction where users incur communication costs to reduce the entropy of their preference representation. Like Castro et al. (2024), we adopt an information-theoretic view where the cost of communicating a specific type depends on its rarity (or distinctiveness) relative to the prior. However, we depart from their framework by modeling this process sequentially—where users may abandon the platform before a transaction occurs—and by explicitly modeling the supply-side response (endogenous pricing) and the impact of differing platform objectives, which Castro et al. (2024) do not consider. Relatedly, Dong et al. (2025) analyze the interaction between a user and an AI agent in a high-dimensional product space. They model the trade-off between the user's costly preference communication and the agent's search/recommendation set size, finding that optimal strategies often require jointly optimizing message precision and recommendation diversity. Cao and Hu (2026) look at a similar problem, analyzing the optimal policies for reducing preference uncertainty between refining particular aspects of the user's preference and adjusting the recommendation assortment. Similarly, Park et al. (2025) study the "when to ask" problem under information asymmetry: the user knows

their specific preference, while the AI knows the population distribution. They highlight how the choice to query versus infer impacts fairness and welfare. Our work shares the structural view of communication as a costly resource found in these papers but differs in focus. While Dong et al. (2025) focus on the fidelity-search trade-off and Park et al. (2025) focus on fairness and distribution learning, we embed the elicitation decision into a market context. In our model, the primary friction is not just the cognitive load of a single turn, but the risk of *abandonment* (market exit) and the interaction with *monetization* (commission vs. conversion).

Platform objectives and welfare misalignment. Our analysis of how platform incentives shape product discovery connects to Besbes et al. (2024), who study the tension between engagement (maximizing observable clicks) and welfare (maximizing true utility). They identify a "rare-niche" regime where engagement-maximizing platforms under-serve users with idiosyncratic tastes because the measurable signal (click-through) does not capture the intense utility of a correct niche match. We find a conceptually similar misalignment in our "rare-niche" regime: under conversion/engagement incentives, platforms prefer a "recommend-first" strategy that ignores niches to minimize friction, even when deep elicitation is welfare-optimal. However, our work differs from Besbes et al. (2024) in three key respects. First, the mechanism driving misalignment is different: in their model, the friction is *observability* (the platform cannot see true utility); in our model, the friction is *communication costs* (the platform can see true utility, but only if it risks user abandonment to ask). Second, we extend the analysis to *commission-based monetization* with endogenous pricing. We show that unlike engagement-driven platforms, commission-driven platforms may have a strong incentive to serve rare niches—not out of benevolence, but because better screening supports higher equilibrium prices. Third, we complement our theoretical results with a novel LLM-based dataset that quantifies these trade-offs in real product categories, bridging the gap between abstract choice models and modern CRS implementation.

Monetization and pricing. Banchio et al. (2024) studies ad auctions in a conversational setting where the auctioneer chooses when to run the auction. The key trade-off is between match quality, which improves over time and with deeper elicitation, and market thickness, which decreases as low-quality ads are screened out. The authors show that optimal timing differs sharply between first-price and second-price auctions. Although our setting differs in several respects (e.g., there is no auction), our high-level message regarding how monetization models affect the optimal degree of elicitation is closely aligned in spirit. More broadly, the platform's decision to elicit preferences acts as a form of information design (Bergemann and Morris 2016), determining the market segmentation available to sellers.

Organization of the paper. Section 2 introduces the base model, interpreting multi-turn conversation as a costly information channel summarized by a fidelity choice. Section 3 presents the main results: optimal CRS policies under different objectives in the rare-niche regime, together with comparative statics in heterogeneity and product variety. Section 4 introduces our dataset and evaluation framework for quantifying the value of additional information, with an illustrative case study. Section 5 concludes. All proofs are deferred to the Appendix.

2. Model

We study a stylized cold-start interaction between a conversational recommender system (CRS) and a single arriving user. The CRS chooses (i) how intensively to elicit preferences via conversation and (ii) how to map conversational signals into a recommendation. The model is intentionally parsimonious: it is designed to isolate the core economic trade-off between *match quality* and *conversational friction* in a way that yields sharp comparative statics.

Users and products. There is one popular (mainstream) product type P and $n \geq 1$ niche product types indexed by $\{1,\dots,n\}$. The user has latent type $\Theta \in \{\emptyset, 1,\dots,n\}$. Type $\Theta = \emptyset$ indicates that no niche is a good match, while $\Theta = j$ indicates compatibility with niche j. The type prior is

$$\Pr(\Theta = \emptyset) = 1-\alpha, \qquad \Pr(\Theta = j) = \alpha/n, \quad j \in \{1,\dots,n\}, \tag{1}$$

where $\alpha \in (0,1)$ captures the prevalence of niche tastes.

Preferences and heterogeneity. The user's deterministic (base) utility from consuming product type j is

$$u_{\mathsf{B}}(\theta, j) = q(j) + \lambda\, \mu(\theta, j), \tag{2}$$

where $q(j)$ is intrinsic quality of product type j, $\mu(\theta, j)$ is match quality, and $\lambda \geq 0$ scales horizontal heterogeneity (the importance of match relative to intrinsic quality). We normalize niche intrinsic quality to zero and assume the popular product has higher intrinsic quality: $q(\mathsf{P}) = V_{\mathsf{P}} > 0$ and $q(j) = 0$ for $j \in \{1,\dots,n\}$. The product fit or match is defined by

$$\mu(\theta, \mathsf{P}) = 0 \quad \forall \theta, \qquad\qquad \mu(\theta, j) = \begin{cases} -1, & \theta \neq j, \\ V(\alpha), & \theta = j, \end{cases} \quad j \in \{1,\dots,n\}, \tag{3}$$

where mismatched niche recommendations impose a loss and a perfect niche match yields gain

$$V(\alpha) = V_{\mathsf{N}} \alpha^{-\beta}, \qquad V_{\mathsf{N}} > 0, \quad \beta \geq 0. \tag{4}$$

When $\beta = 0$, the niche match gain is bounded; when $\beta > 0$, the gain grows as the fraction of niche seekers α decreases. The outside option O has deterministic utility normalized to zero.

Conversation modeled as a noisy information channel. We model the conversation as a noisy information channel which provides a signal $S \in \{\emptyset, 1, \ldots, n\}$ about the user type θ. In particular, we model this through an $(n+1)$-ary channel with fidelity $f \in [1/(n+1), 1]$, i.e.,

$$\Pr(S = s \mid \Theta = \theta) = \begin{cases} f, & s = \theta, \\ (1-f)/n, & s \neq \theta. \end{cases} \tag{5}$$

When $f = 1/(n+1)$ the signal is uninformative (no effective elicitation); when $f = 1$ the signal fully reveals θ. This fidelity f captures the notion that more intensive conversation (high f) implies more accurate inference about the user type which in turn can possibly lead to high match fit.

Information elicitation, abandonment, and user effort. A central tension in conversational recommendation is that deeper elicitation improves match quality but consumes user attention and raises the risk of abandonment. We summarize the overall intensity of the interaction by the fidelity parameter f, which determines the informativeness of the conversational signal S about the user's latent type Θ. We measure the amount of preference information extracted by a conversation of fidelity f as $\mathcal{I}(f)$, in *nats* (the natural-log unit of information; 1 nat equals $1/\ln 2$ bits).

We consider two benchmarks for $\mathcal{I}(f)$. First, our main specification is the *true-prior mutual information* $\mathcal{I}^{\mathsf{MI}}(f) \equiv I_{\alpha,n}(f) = I(\Theta; S)$, computed under the population prior and the channel induced by fidelity f. This is the standard object in rational inattention and related models of costly information acquisition (e.g., Sims 2003, Matějka and McKay 2015, Castro et al. 2024) and is also closely related to information-cost formulations in persuasion and information design (e.g., Kamenica and Gentzkow 2011). Second, as a robustness benchmark we also consider a *capacity-based* measure, $\mathcal{I}^{\mathsf{cap}}(f) \equiv C_{n+1}(f) = \max_{P(\Theta)} I(\Theta; S)$, the Shannon capacity of the $(n+1)$-ary channel. The capacity benchmark captures a "design-time" notion of conversational burden: the interaction protocol must be able to reliably distinguish among $n+1$ latent states at fidelity f, regardless of how frequently each state arises in the population. This is useful when the platform's conversational interface is designed to be robust across environments with different priors (e.g., across categories, cohorts, or new markets in which niche prevalence is initially uncertain).

We interpret conversation as gradually transmitting information. Let $X \in [0, \mathcal{I}(f)]$ denote the cumulative information (in nats) transmitted up to a given point in the interaction. We assume the user faces a constant abandonment hazard *per nat* of information processed: for some $\eta > 0$,

$$\Pr(\text{user remains in the conversation through } x \text{ nats}) = \exp(-\eta x).$$

Thus the probability the user survives to the recommendation stage is $sur(f) = \exp(-\eta \mathcal{I}(f))$. We assume that the platform issues a final recommendation only at the end of the interaction; users who abandon early exit to the outside option without receiving a recommendation.[4]

[4] In practice, CRS may surface intermediate recommendations and users may exit early after seeing them. We abstract from such within-conversation stopping decisions to keep the analysis transparent; the reduced-form hazard captures the empirically salient fact that longer, more demanding interactions increase drop-off.

While the user remains in the conversation, she incurs a flow effort disutility of $\bar{\kappa}_u$ per nat. Therefore, the expected total effort loss—counting both users who ultimately receive a recommendation and those who abandon midstream—is the area under the survival curve:

$$\bar{\kappa}_u \int_0^{\mathcal{I}(f)} \exp(-\eta x)\, dx = \frac{\bar{\kappa}_u}{\eta}\Big(1-\exp(-\eta \mathcal{I}(f))\Big). \tag{6}$$

Note that as $\eta \to 0$, we have that $\bar{\kappa}_u(1-\exp(-\eta\mathcal{I}(f)))/\eta \to \bar{\kappa}_u \mathcal{I}(f)$ and this captures the standard information cost benchmark used in the literature Sims (2003), Castro et al. (2024), Dong et al. (2025).

Finally, the two information benchmarks lead to different implications in the rare-niche regime studied later. Under mutual information, $\mathcal{I}^{\mathsf{MI}}(f)$ can shrink with the niche prevalence α (indeed, full revelation satisfies $I_{\alpha,n}(1) = H(\Theta) = h(\alpha) + \alpha \ln n$ where $h(x)$ is binary entropy), so both effort and abandonment may vanish as niches become rare—capturing environments in which it is relatively easy to learn that a user is mainstream. Under the capacity benchmark, by contrast, $\mathcal{I}^{\mathsf{cap}}(f)$ depends on the required resolution $(n+1)$ and the target fidelity f, and does not decrease with α; this captures settings in which meaningful personalization requires a nontrivial interaction burden even if niche users are infrequent.

Choice at the recommendation stage. If the user reaches the recommendation stage and the platform recommends $j \in \{\mathsf{P}, 1, \ldots, n\}$ at price p_j, the user chooses between adopting j and taking the outside option. Utilities take the standard logit form: $\max\{u_{\mathsf{B}}(\theta, j) - p_j + \varepsilon_j, 0 + \varepsilon_{\mathsf{O}}\}$, where $\varepsilon_j, \varepsilon_{\mathsf{O}}$ are i.i.d. type-I extreme value shocks. Thus adoption conditional on reaching the choice stage is

$$\Pr(\text{adopt } j \mid \theta,\ \text{reach choice}) = \sigma(u_{\mathsf{B}}(\theta, j) - p_j) := \frac{\exp(u_{\mathsf{B}}(\theta, j) - p_j)}{1 + \exp(u_{\mathsf{B}}(\theta, j) - p_j)}. \tag{7}$$

Let $g(u_{\mathsf{B}}(\theta, j) - p_j) := \ln(1 + \exp(u_{\mathsf{B}}(\theta, j) - p_j))$ denote the inclusive value (expected utility) at the choice stage.

Timing. We analyze two environments: a fixed-price benchmark and an endogenous-pricing (commission) environment. The sequence of moves is:

(*i*) **(Commission environment only)** The platform commits to a CRS design (f, π) and a commission rate $\tau \in (0, 1)$. Sellers then simultaneously set prices $p = (p_{\mathsf{P}}, p_1, \ldots, p_n)$.

(*ii*) A user arrives. The platform conducts a conversation at intensity (fidelity) f. Conversation may end prematurely due to abandonment, as formalized above.

(*iii*) If the user reaches the recommendation stage, the platform observes a conversational signal S and recommends a single product $j = \pi(S)$.

(*iv*) The user chooses between adopting the recommendation and taking the outside option O.

Platform objectives and the welfare metric. Our analysis centers on how conversational design varies across *monetization models*. In particular, we focus on (i) platforms whose payoff is tied primarily to *engagement/conversion* (e.g., many streaming or ad-supported settings where the marginal value of a session is roughly independent of which title is consumed), and (ii) platforms whose payoff is tied to *transaction value* through a commission or take-rate (e.g., marketplaces and e-commerce). In both cases the platform chooses a CRS design (f, π): the conversation fidelity f and a recommendation rule π mapping signals into a single recommended product.

1. **Conversion / engagement payoff (fixed prices).** Suppose the platform earns a constant payoff $R > 0$ whenever the user adopts the recommended product, independent of which product is adopted. Since adoption requires both survival and a positive choice at the recommendation stage, the platform's expected payoff is

$$\Gamma(f, \pi) = R \cdot sur(f) \cdot \mathbb{E}\left[\sigma\left(u_{\mathsf{B}}(\Theta, \pi(S))\right)\right]. \tag{8}$$

2. **Commission payoff with endogenous pricing.** In the commission environment, each product $j \in \{\mathsf{P}, 1, \ldots, n\}$ is supplied by a seller who posts a price p_j. If the user adopts product j, the platform earns commission τp_j and the seller receives $(1-\tau)p_j$, where $\tau \in (0,1)$ is an exogenous take-rate. Given (f, π, p), demand for seller j is the probability a random user (i) survives elicitation, (ii) is recommended j, and (iii) adopts at price p_j:

$$D_j(f, \pi, p) := sur(f) \cdot \mathbb{E}[\mathbf{1}\{\pi(S) = j\} \cdot \sigma(u_{\mathsf{B}}(\Theta, j) - p_j)]. \tag{9}$$

   Seller j chooses p_j to maximize their expected revenue,

$$\Pi_j(p_j; f, \pi, p_{-j}) = (1-\tau)\, p_j \cdot D_j(f, \pi, p). \tag{10}$$

   Let $p^\star(f, \pi)$ denote a Nash equilibrium price vector. The platform's expected commission revenue is then

$$\Pi^{\mathsf{com}}(f, \pi) = \tau \sum_{j \in \{\mathsf{P}, 1, \ldots, n\}} p_j^\star(f, \pi) \cdot D_j\big(f, \pi, p^\star(f, \pi)\big). \tag{11}$$

Although the platform's design is chosen to maximize either (8) or (11), we also track user welfare to assess whether platform-optimal conversational policies align with consumer interests. With abandonment during elicitation, welfare equals expected choice-stage surplus among surviving users, minus expected effort losses incurred during elicitation by *all* users (including those who abandon midstream):

$$W(f, \pi; p) = sur(f) \cdot \mathbb{E}\left[g\big(u_{\mathsf{B}}(\Theta, \pi(S)) - p_{\pi(S)}\big)\right] - \frac{\bar{\kappa}_u}{\eta}\left(1 - sur(f)\right). \tag{12}$$

The second term is the cumulative effort cost along the conversation path, and it is incurred even in sessions that do not reach the recommendation stage.

Elicitation and Recommendation policies: recommend-first versus ask-then-recommend. A CRS design is a pair (f, π), where $f \in [1/(n+1), 1]$ is the *conversation fidelity* and $\pi : \{\emptyset, 1, \ldots, n\} \to \{\mathsf{P}, 1, \ldots, n\}$ maps the realized conversational signal into a single recommended product. We focus on two policy classes.

1. **Recommend-first (REC).** The platform conducts no effective elicitation, which is equivalent to setting $f = 1/(n+1)$ so that the signal is uninformative. Since $\mathcal{I}(f) = 0$ at this baseline, there is no abandonment due to conversation ($sur(f) = 1$). The platform then recommends the mainstream option P to every user.
2. **Ask-then-recommend (ATR(f)).** The platform runs a conversation calibrated to fidelity f, which induces survival probability $sur(f) = \exp(-\eta \mathcal{I}(f))$. If the user abandons before the recommendation stage, the session ends with the outside option. If the user reaches the recommendation stage, the platform observes the signal $S \in \{\emptyset, 1, \ldots, n\}$ and makes a single recommendation as follows:
   (a) If $S = \emptyset$, recommend P.
   (b) If $S = i \in \{1, \ldots, n\}$, compare recommending niche i to defaulting to P. Because the signal structure is symmetric, the posterior probability that $S = i$ is correct is the same for all niche labels and is increasing in f. The optimal action therefore takes a simple *posterior-threshold* form: recommend the indicated niche i if and only if the posterior correctness of the signal exceeds a cutoff, and otherwise recommend P. The cutoff depends on the platform's objective (conversion versus commission), because the payoff from correctly versus incorrectly routing a user differs across monetization models.

This formulation highlights the key design margin in our setting. Once the recommendation rule is reduced to "follow a niche signal only when it is sufficiently reliable," conversational design becomes essentially one-dimensional: the platform chooses f to trade off higher screening accuracy against greater conversational friction (effort and abandonment) that applies to all users.

Notation. Throughout the paper, we define

$$u_{\mathsf{P}} := V_{\mathsf{P}}, \qquad u_- := -\lambda, \qquad u_+(\alpha) := \lambda V_{\mathsf{N}} \alpha^{-\beta}, \qquad \sigma(x) := \frac{e^x}{1+e^x}, \qquad g(x) := \ln(1+e^x).$$

3. Main results under mutual-information–based abandonment

This section characterizes how a platform should design conversational elicitation when deeper interaction increases abandonment risk. We focus on two platform objectives—*conversion/engagement* and *commission revenue with endogenous pricing*—and we evaluate the induced *user welfare* to quantify the alignment (or misalignment) between platform incentives and consumer outcomes. This section focuses on the mutual information setting, with proofs of the

following results deferred to Appendix A. Our findings with the capacity-based benchmark thematically align with the structural insights discussed in this section, with full details on the distinctions available in Appendix B.

3.1. Rare niches with potentially large match value ($\alpha \to 0$)

The rare-niche regime isolates a clean force that is central to conversational design. Most users are well served by the mainstream option P, but a small fraction have niche tastes. At the same time, a correct niche match can be extremely valuable. We capture this with the scaling $u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{-\beta}$: β indexes how fast the value of a correct niche match grows as the fraction of niche users becomes small. This regime should be interpreted as a stylized *rare-gems* environment. Think of a small segment of users with highly unique tastes (obscure documentaries or specialized hobby equipment), for whom the value of being correctly matched is very large. The design question is whether a platform should invest in elicitation when the niche segment is small and deeper conversation increases abandonment.

3.1.1. Conversion/engagement: recommend-first, and welfare misalignment We first analyze engagement-style monetization. With fixed prices ($p_j \equiv 0$), the platform earns a constant payoff $R > 0$ whenever the user adopts the recommended product (Section 2).

THEOREM 1 **(Conversion-optimal recommend-first in the rare-niche regime)**. *Fix $n \geq 1$, $R > 0$, and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta)$. Assume $\mathcal{I}(f) = \mathcal{I}^{\mathsf{MI}}(f)$. There exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$, the conversion-optimal policy is* REC *(always recommend* P*).*

Discussion. Theorem 1 is a bounded-upside result. Conversion gains are measured in *probabilities*: even perfectly matching a niche user can raise adoption by at most a constant amount. Thus the maximal incremental conversion benefit from fully serving the niche segment is $O(\alpha)$. Any nontrivial elicitation reduces the mass of sessions that reach the recommendation stage by the survival factor $sur^{\mathsf{MI}}(f)$. Moreover, in the rare-niche regime the platform will recommend a niche only when the posterior clears a strictly positive cutoff (Lemma 2); because the prior probability of any given niche is α/n, acting on niche signals requires near-perfect fidelity. With $\mathcal{I}(f) = \mathcal{I}^{\mathsf{MI}}(f)$, the resulting information burden is of order $\alpha \ln(1/\alpha)$, so the survival loss is of the same order, i.e., $1 - sur^{\mathsf{MI}}(f) \approx \eta\alpha \ln(1/\alpha)$. For small α, this $\alpha \ln(1/\alpha)$ mass loss dominates any $O(\alpha)$ conversion gains, so recommend-first is optimal. In Figure 2a we plot the difference between the conversion obtained under the optimal elicitation policy and REC. For the representative model primitives detailed in Figure caption, we observe that for small enough α (see $\alpha \leq 0.3$ in Figure 2a), the difference is zero, i.e., REC is optimal which verifies Theorem 1.

Managerial implication. Our result should be interpreted comparatively, not literally. In our stylized setting, REC corresponds to *minimal* probing rather than "never asking questions." Theorem

1 says that when revenue is essentially tied to *getting the user to consume something* (watch-time, ad exposure, session completion), the platform has an incentive to front-load strong mainstream recommendations and economize on deep elicitation. This provides a concrete rationale for why engagement-driven systems may invest less in deep, high-resolution preference elicitation than one might expect from a purely personalization-centric engineering lens.

A welfare-optimal benchmark and misalignment. To quantify how conversion-driven design aligns with user interests, we compare REC (conversion-optimal policy for small α) to a benchmark policy optimizing user welfare as defined in (12).

PROPOSITION 1 **(User-welfare-maximizing benchmark)**. *Fix $n \geq 1$ and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta, \bar{\kappa}_u)$.*

(i) [Bounded Niche Value] *Let $\beta = 0$. There exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$, the user-welfare-maximizing CRS is* REC.

(ii) [Niche value scales with α] *Fix $\beta > 0$. There exists $\bar{\alpha}_\beta \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha}_\beta)$, user welfare strictly prefers elicitation to* REC *and the user-welfare-maximizing fidelity $f_W^\star(\alpha)$ satisfies $f_W^\star(\alpha) \to 1$ as $\alpha \to 0$.*

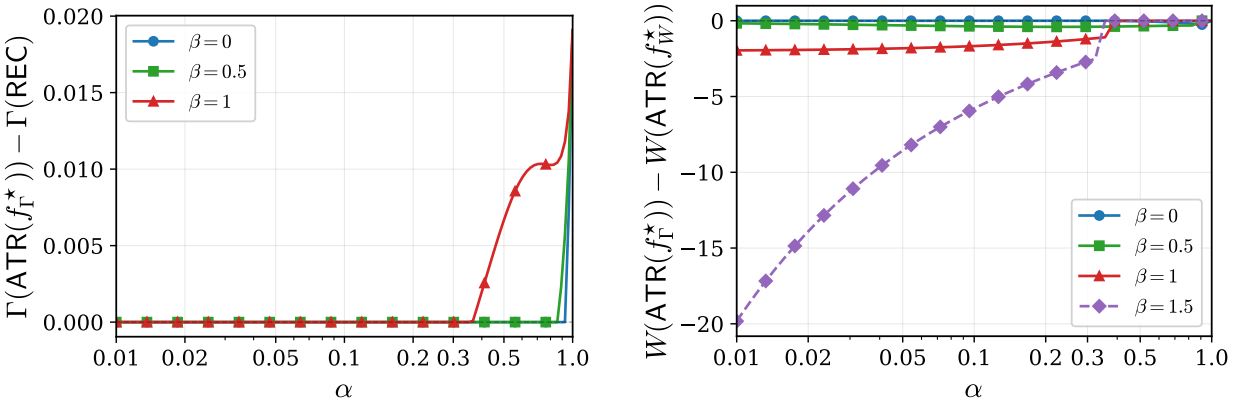
(iii) [Explicit Fidelity Scaling for $\beta = 1$] *Define $g_{\mathsf{P}} := g(V_{\mathsf{P}}), g_- := g(-\lambda), K_W := \bar{\kappa}_u + \eta\,(g_{\mathsf{P}} + \lambda V_{\mathsf{N}})$ and $D_W := g_{\mathsf{P}} + \lambda V_{\mathsf{N}} - g_-$. Then, for α sufficiently small, the user-welfare-maximizing fidelity satisfies $1 - f_W^\star(\alpha) = c_W^\star\,\alpha + o(\alpha)$, where $(c_W^\star)^{-1} = \exp(D_W/K_W) - 1$.*

COROLLARY 1 **(Welfare misalignment under engagement optimization)**. *Fix $n \geq 1$ and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \eta, \bar{\kappa}_u)$. Fix $\beta > 0$, then in the rare-niche regime where $\alpha \to 0$, the conversion-optimal design* REC *is strictly user-welfare-suboptimal. Let $W(\mathsf{ATR}(f_W^\star); \alpha)$ denote the optimal user welfare value. There exists a constant $c > 0$ and $\bar{\alpha}_\beta \in (0,1)$ such that for all $\alpha \leq \bar{\alpha}_\beta$, we have $W(\mathsf{REC}) - W(\mathsf{ATR}(f_W^\star); \alpha) \leq -c\alpha^{1-\beta}$.*

Discussion. Proposition 1 formalizes the core reason that objectives diverge. Welfare is measured in *utility levels* (inclusive values), so when niche match value explodes ($\beta > 0$), correctly matching a vanishingly small segment can generate large aggregate surplus. Conversion is measured in *probabilities*, so the upside from serving the niche segment remains bounded.

For the case of $\beta = 1$, Proposition 1 explicitly quantifies the optimal fidelity scaling of $f_W^\star = 1 - c_W^\star \alpha + o(\alpha)$ for user welfare. For $\alpha > 0$, the optimal fidelity is an interior solution and scales linearly with α. This explicit characterization also highlights how the optimal fidelity $f_W^\star$ scales with other primitives. For example, as the user effort disutility factor $\bar{\kappa}_u$ or the hazard rate η increases, the optimal fidelity level decreases, while as the heterogeneity parameter λ increases, the optimal fidelity level also increases.

Corollary 1 quantifies how the misalignment between user-welfare-optimal and conversion-optimal policies scales as $\alpha \to 0$. In Figure 2b, we plot the user welfare loss due to the conversion-optimal policy $\mathsf{ATR}(f^\star_\Gamma)$. Note that for α small enough, we have that $\mathsf{ATR}(f^\star_\Gamma) \equiv \mathsf{REC}$. For $\beta \in (0,1)$, this welfare loss decreases as α decreases (see the lines for $\beta = 0$ and $\beta = 0.5$). For $\beta = 1$, the loss converges to a constant (see $\beta = 1$ in Figure 2b), while for $\beta > 1$, this loss increases as α decreases (see $\beta = 1.5$ in Figure 2b) highlighting that in settings where matching niche users to items can deliver very large utility value, welfare aligns with deeper elicitation even at the cost of user abandonment. Corollary 1 makes the design implication explicit: in rare-gems environments, engagement-style monetization rationalizes "recommend-first" designs even when deeper elicitation would be consumer welfare improving.



(a) Value of Elicitation for Conversion Objective

(b) User Welfare Loss due to Conversion-Optimal Policy

Figure 2 (a) Difference between the optimal elicitation policy $\mathsf{ATR}(f^\star_\Gamma)$ **and** REC **for the conversion objective** (8) **as a function of** α **(log-scale) for different values of** $\beta \in \{0, 0.5, 1\}$**. (b) Difference between the user welfare obtained under the conversion-optimal elicitation and user-welfare-optimal elicitation as a function of** α **(log-scale) for different values of** $\beta \in \{0, 0.5, 1, 1.5\}$**. For both plots, we set** $V_{\mathsf{P}} = 1, V_{\mathsf{N}} = 2, \eta = 0.1, n = 5, \lambda = 1, R = 1$ **and** $\bar{\kappa}_u = 0.1$**.**

3.1.2. Commission monetization: elicitation becomes platform optimal, and user welfare can go either way We now endogenize prices. Each product $j \in \{\mathsf{P}, 1, \ldots, n\}$ is supplied by a seller who posts price p_j after observing (f, π), and the platform earns a take-rate $\tau \in (0,1)$ of realized seller revenue. Because the user sees only a single recommended product, each seller faces routed monopoly demand.

THEOREM 2 **(Commission-driven elicitation under MI abandonment).** *Fix* $n \geq 1$, $\tau \in (0,1)$, *and primitives* $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta)$. *Consider the commission objective* (11). *Assume that* $\mathcal{I}(f) = \mathcal{I}^{\mathsf{MI}}(f)$.

(i) [Bounded Niche Value] *Assume* $\beta = 0$*. There exists an* $\bar{\alpha} \in (0,1)$ *such that for all* $\alpha \in (0, \bar{\alpha})$*, the commission-optimal policy is* REC.

(ii) [Niche value scales with α] *Assume* $\beta > 0$*. For sufficiently small* α *the platform strictly prefers elicitation to* REC*, and any commission-optimal fidelity* $f^\star_\Pi(\alpha)$ *satisfies* $f^\star_\Pi(\alpha) \to 1$ *as* $\alpha \to 0$*.*

(iii) [Explicit Fidelity Scaling for $\beta = 1$] *The commission-optimal fidelity* $f^\star_\Pi$ *scales as* $1 - f^\star_\Pi(\alpha) = c^\star_\Pi \alpha + o(\alpha)$*, where* $(c^\star_\Pi)^{-1} = \exp(\eta^{-1}) - 1$*.*

Discussion. Commission monetization creates a distinct incentive: elicitation is valuable not only because it improves matching, but because it improves *screening*. Higher fidelity increases the posterior share of true niche users among those routed to a niche seller, which supports higher equilibrium prices. The platform captures a fraction τ of this additional revenue. The $\beta = 1$ case highlights the screening logic most sharply. In this regime, the platform can earn *non-vanishing* commission profit from a vanishing niche segment because niche willingness-to-pay scales like $1/\alpha$. Up to first order, the optimal fidelity $f^\star_\Pi$ depends only on η: the level of revenue scales out, and what remains is the survival elasticity embedded in abandonment sensitivity. When users are highly abandonment-sensitive (large η), the optimal fidelity $f^\star_\Pi$ decreases while when users are more patient (small η), the platform pushes fidelity closer to one. In Figure 3a we plot the difference between the optimal commission and the commission obtained under REC. For $\beta = 0$, we observe that for small enough α, the difference is zero, i.e., REC is commission optimal as Theorem 2 suggests. For $\beta > 0$, we observe that the difference is positive, affirming our result that for $\beta > 0$, the platform prefers non-trivial elicitation to REC.

Managerial implication. In environments where platform revenue is proportional to transaction value, investing in conversational screening can be privately optimal even in rare-gems settings. This rationalizes why commission-driven platforms may have stronger incentives to build high-resolution conversational interfaces and richer elicitation pipelines: information can be monetized through equilibrium price and margin.

Consumer welfare under commission. The same screening mechanism that raises platform revenue can reduce consumer welfare. With monopoly pricing, consumer surplus grows only logarithmically in willingness-to-pay, while prices (and seller revenue) grow roughly linearly (see Lemma 8). Thus the user welfare impact of commission-driven elicitation can be positive or negative in the rare-niche regime. Let $p_m(u)$ denote the monopoly price that maximizes $p \cdot \sigma(u - p)$ for a consumer with deterministic utility u, and define $p^\star_{\mathsf{P}} := p_m(u_{\mathsf{P}})$ and $p^\star_{\mathsf{N}} := p_m(u_+(\alpha))$.

PROPOSITION 2 **(Consumer welfare under the commission-optimal CRS)**. *Fix* $n \geq 1$*,* $\tau \in (0,1)$*, and primitives* $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta, \bar{\kappa}_u)$*. Let* $\Delta W^{\mathsf{com}}_\Pi(\alpha) := W^{\mathsf{com}}\Big(\mathsf{ATR}(f^\star_\Pi(\alpha))\Big) - W^{\mathsf{com}}(\mathsf{REC})$*. Then we have that,*

(i) [Bounded niche value] *If $\beta = 0$, then there exists $\bar{\alpha} \in (0,1)$ such that*

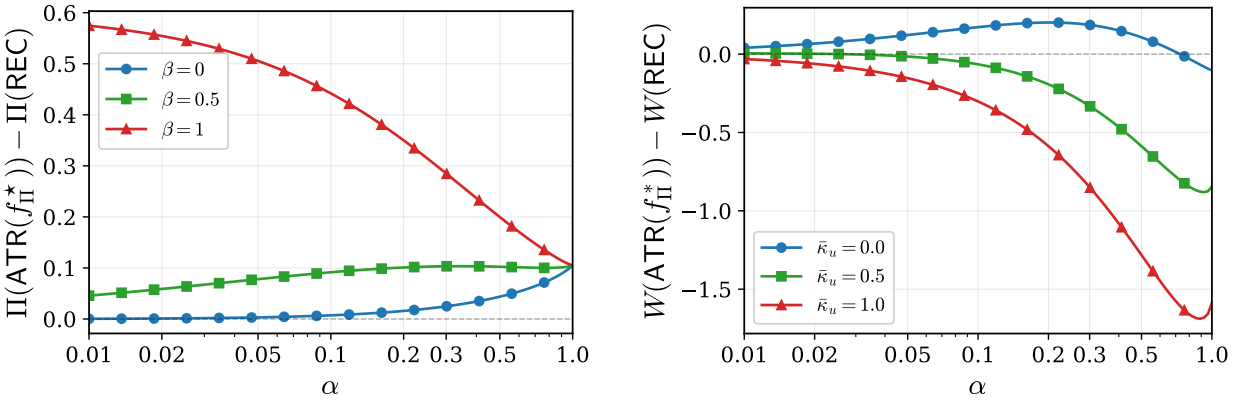
$$\Delta W_{\Pi}^{\mathsf{com}}(\alpha) = 0 \qquad \textit{for all } \alpha \in (0, \bar{\alpha}).$$

(ii) [Niche value scales with α] *Suppose $\beta > 0$ and $\lambda > 0$. Then*

$$\lim_{\alpha \to 0} \frac{\Delta W_{\Pi}^{\mathsf{com}}(\alpha)}{\alpha \ln(1/\alpha)} = \beta - \bar{\kappa}_u - \eta\, g\big(u_{\mathsf{P}} - p_{\mathsf{P}}^{\star}\big),$$

where $p_{\mathsf{P}}^{\star} = p_m(u_{\mathsf{P}})$. In particular, for sufficiently small α, consumer welfare under the commission-optimal CRS is higher than under REC *if $\beta > \bar{\kappa}_u + \eta\, g(u_{\mathsf{P}} - p_{\mathsf{P}}^{\star})$, and lower if the inequality is reversed.*

Discussion. Optimal screening raises niche users' surplus by about $\alpha \ln u_+(\alpha) \asymp \beta\, \alpha \ln(1/\alpha)$ (see Lemma 8). At the same time, elicitation reduces the mass of sessions that reach the recommendation stage, and it imposes an effort disutility proportional to $\bar{\kappa}_u$. The limit in Proposition 2 therefore has a simple interpretation: user welfare rises when the informational value of screening (captured by β) dominates both effort costs and abandonment losses. In Figure 3b, we illustrate when elicitation aligns with user welfare more than REC; we plot ΔW^{com} for fixed $\beta = 1, \eta = 0.1, V_{\mathsf{P}} = 1$ and vary $\bar{\kappa}_u \in \{0, 0.5, 1\}$. For small $\bar{\kappa}_u$, the limiting value is positive for small positive α and this is what we observe in Figure 3b while for large $\bar{\kappa}_u$, the limit becomes negative as observed for $\bar{\kappa}_u = 1$ in Figure 3b.



(a) Value of elicitation for commission objective

(b) User welfare (mis)alignment under commission objective

Figure 3 (a) Difference between the optimal elicitation policy and REC **for the commission-driven objective. (b) Difference between** ATR($f_\Pi^\star$) **and** REC **for user welfare with endogenous prices. We set $\bar{\kappa}_u = 0.1$ and $\beta = 1$ when not specified. For both plots, we have $V_{\mathsf{P}} = 1, V_{\mathsf{N}} = 2, \lambda = 1, \eta = 0.1$.**

3.2. Comparative statics: heterogeneity and variety

The rare-niche results fix (λ, n) and isolate $\alpha \to 0$ to obtain sharp contrasts across objectives. In this section, we fix α and understand the impact of two primitives that are particularly salient in practice: *heterogeneity* λ (how consequential matching is) and *variety* n (how many distinct niche labels must be distinguished).

When λ is small, niches are effectively irrelevant: both the upside from correct niche matching and the downside from mismatch are muted. When λ is large, matching becomes high-stakes: correct niche matches become very attractive while mismatches become unattractive, sharpening the value of screening. These forces interact differently with conversion and commission.

PROPOSITION 3 **(Impact of heterogeneity λ).** *Fix the primitives* $(\alpha, n, \beta, V_{\mathsf{P}}, V_{\mathsf{N}}, \eta, R, \tau)$. *Define the following quantities:*

$$sur^{\mathsf{MI}}(f) := \exp\left(-\eta I_{\alpha,n}(f)\right), \quad \sigma_{\mathsf{P}} := \sigma(u_{\mathsf{P}}), \quad q_0(f) = (1-\alpha)f + \alpha(1-f)/n,$$
$$\underline{f}_\infty := \frac{\sigma_{\mathsf{P}}\,(1-\alpha/n)}{\sigma_{\mathsf{P}}\,(1-\alpha/n) + \alpha(1-\sigma_{\mathsf{P}})}, \quad \Gamma_\infty(f) := R \cdot sur^{\mathsf{MI}}(f) A(f), \quad A(f) = \sigma_{\mathsf{P}} q_0(f) + \alpha f,$$

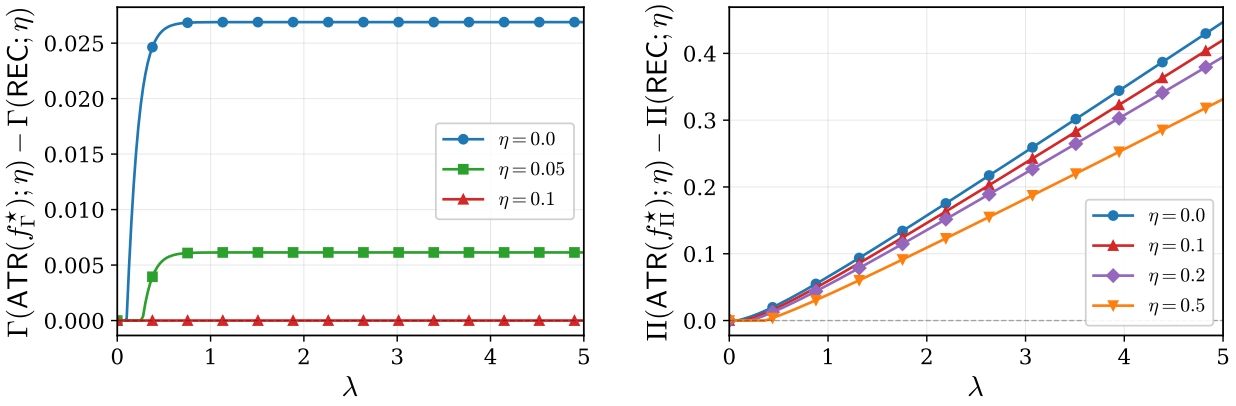
(i) ***Low heterogeneity.*** *There exists $\bar{\lambda} > 0$ such that for all $\lambda \in [0, \bar{\lambda}]$, the optimal policy under both the conversion objective* (8) *and the commission objective* (11) *is* REC.

(ii) ***High heterogeneity: conversion.*** *If $\max_{f \in [\underline{f}_\infty, 1]} \Gamma_\infty(f) \le R\sigma_{\mathsf{P}}$, then* REC *is conversion-optimal for all sufficiently large λ. Otherwise, elicitation is conversion-optimal for all such λ.*

(iii) ***High heterogeneity: commission.*** *There exists $\bar{\lambda} < \infty$ such that for all $\lambda \ge \bar{\lambda}$, $\Pi^{\mathsf{com}}(\mathsf{ATR}(1)) > \Pi^{\mathsf{com}}(\mathsf{REC})$.*

Discussion. High heterogeneity does *not* automatically imply deep elicitation under engagement monetization. Even if personalization can make the "right" item extremely compelling for some users, conversion gains remain bounded by probability saturation, while abandonment discounts all sessions. When baseline mainstream adoption is already high and/or when elicitation meaningfully reduces survival, even perfect screening cannot justify the abandonment discount—this is what the condition $\max_{f \in [\underline{f}_\infty, 1]} \Gamma_\infty(f) \le R\sigma_{\mathsf{P}}$ captures. In Figure 4, we see that for small enough hazard rate, i.e., $\eta = 0.05$, REC is optimal for small values of λ while for large enough λ, elicitation is optimal. However, for large enough hazard rate, i.e., $\eta = 0.1$, we see that for all λ, the optimal policy is REC, since the difference is zero. In contrast, commission monetization creates a stronger incentive to build screening capacity because prices (and hence platform revenue) respond to segmentation. In Figure 4b, we see that for different values of η, there exists some $\bar{\lambda}(\eta)$ such that for all $\lambda \le \bar{\lambda}(\eta)$, the optimal policy is REC while for all $\lambda > \bar{\lambda}(\eta)$, the optimal policy involves elicitation. Proposition

3 also connects directly to the motivating evidence in Figure 1: categories that are effectively low heterogeneity (e.g., air purifier) support immediate mainstream recommendations, while high-heterogeneity categories (e.g., laptop) can justify deeper elicitation.



(a) Value of elicitation for conversion objective

(b) Value of elicitation for commission objective

Figure 4 Difference between the optimal elicitation policy and REC **for (a) conversion-driven objective and (b) commission-based objective. For both plots, we have** $\alpha = 0.1, V_{\mathsf{P}} = 1, V_{\mathsf{N}} = 2, \beta = 1, R = 1, \tau = 0.3, \bar{\kappa}_u = 0.1$.

Variety raises the resolution burden of conversational inference: identifying one niche among n possibilities is inherently a $\log n$-bit task. When the catalog contains many distinct niches, conversation must either become higher-resolution (more questions, more precise questions), or it must fall back on coarse, robust recommendations. Proposition 4 formalizes this "resolution versus returns" force.

PROPOSITION 4 **(Impact of variety** n**).** *Fix* $(\alpha, \lambda, \beta, V_{\mathsf{P}}, V_{\mathsf{N}}, R, \tau, \eta)$. *There exists* $\bar{n} < \infty$ *such that for all* $n \geq \bar{n}$, REC *is optimal under both conversion and commission objectives.*

4. Dataset and Simulation Case Study

We complement our theoretical results by constructing a synthetic dataset of structured user preferences built on real long-form product queries found on Reddit (Section 4.1). Using this synthetic dataset and LLM reasoning models, we simulate user responses to product recommendation sets constructed using an increasing sequence of revealed user preferences (Section 4.2). We briefly discuss how this dataset and LLM responses can be used to quantify the value of information in categories with varying levels of preference heterogeneity (Section 4.3). We then operationalize this dataset of user needs and LLM evaluation to simulate a multi-turn interaction with abandonment, connecting results from our theoretical model with empirical behavior (Section 4.4).

4.1. User Preference Dataset Generation

Our synthetic dataset of user profiles is based on the E-GEO dataset consisting of real long-form product queries found on Reddit (Bagga et al. 2025). Figure 5 explains our synthetic dataset generation pipeline which starts by clustering the long-form product queries into different categories. We then select 40 categories with the largest number of queries. For each category, we prompt ChatGPT-5.2 (Thinking) with the corresponding queries to generate 100 different synthetic user profiles with different preferences and needs (see the prompt in Appendix D.2). Each generated user profile in a given category consists mainly of (i) a complete preference expressed in natural language and (ii) a "context trail" which is a sequence of (single) sentences revealing different aspects of their preference in the order of importance (see Figure 6). For further details, refer to Appendix D.

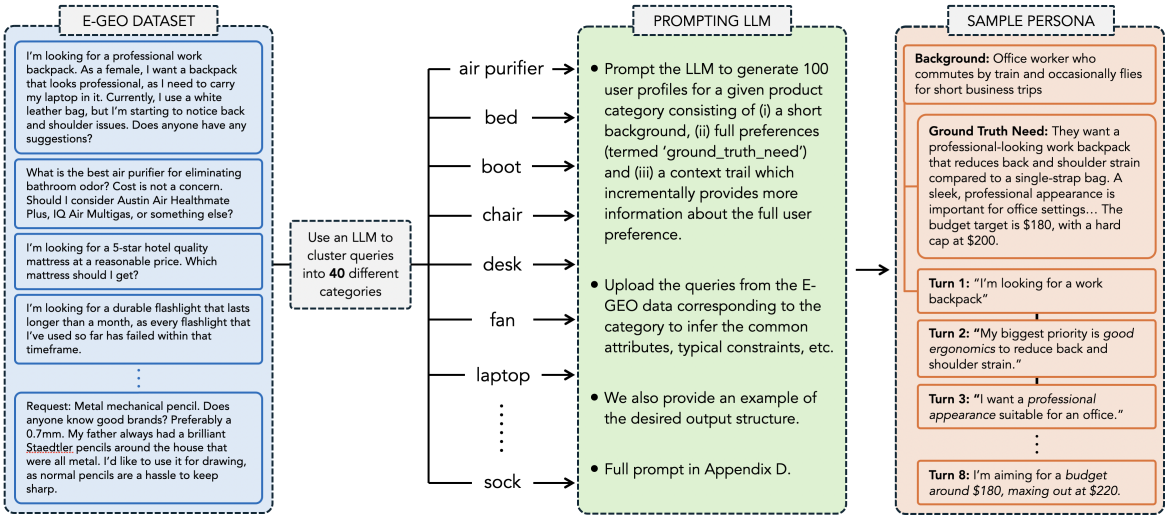


Figure 5 Dataset generation pipeline. Clustered E-GEO queries are provided to an LLM to generate preferences.

4.2. Recommendation Set Generation and LLM-based Evaluation

4.2.1. Recommendation Set Generation For each user profile in a given category, we construct an increasing sequence of partial user preferences by incrementally stitching the sentences in their "context trail". We append user preferences in the "de facto" importance ordering returned by the LLM during persona generation. This gives us a list of (partial) preferences which starts generic and becomes progressively more detailed and closer to the complete user preference (see Figure 6). We refer to these as partial user queries. This incremental revelation of user preference is intended to capture the incremental preference elicitation in a conversational recommendation system. We do acknowledge that in an actual CRS, the user may respond to questions asked by

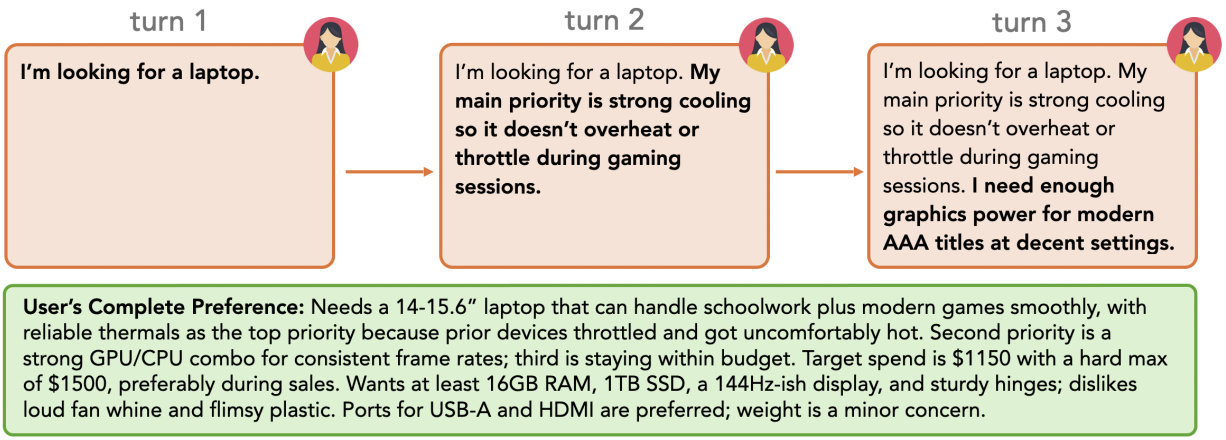


Figure 6 **An example of a user's ground truth needs, and the different queries representing each turn.**

the platform. However, our aim with this dataset is to understand and empirically test the value of incremental preference information. Incorporating the feedback loop is an interesting direction for future work.

For each category, we generate a product catalog from the Amazon Reviews dataset (Hou et al. 2024), which contains over 48 million products across 33 broad categories (e.g., Electronics, Appliances) through 2023. Using keywords specific to each category, we filter the dataset to create a smaller pool of potentially relevant products. For each of the partial user queries, we retrieve relevant products from this catalog of products using the `BAAI/bge-large-en-v1.5` sentence encoder for efficient nearest-neighbor search. We retain the top 5 products ranked by embedding cosine similarity, as well as an accumulated pool of the two highest-ranked unique products in each turn. As a result, later rounds contain larger recommendation sets, until a cap of 15 total products is reached. This increasing size of the recommendation set is intended to capture the feature that conversational recommender systems may recommend different products during the course of the conversation and hence the users' consideration set may expand over time.

4.2.2. LLM-based User Evaluation Prior work suggests that LLMs can realistically simulate human willingness-to-pay estimates (WTP) (Brand et al. 2023, Voltes-Dorta and Suau-Sanchez 2025). We use `gpt-5-mini` with `medium` reasoning effort to simulate purchase/no-purchase decisions and WTP. We provide the LLM model with the complete user preference and a recommendation set generated by a partial user query. We run the evaluation (see Prompt E.1 for the prompt used) across all 4000 user profiles with differing recommendation sets corresponding to the partial user queries.

4.3. Quantifying the Value of Information and Interplay with Heterogeneity

For each product category, we quantify preference heterogeneity using vector embeddings. We encode each user's full preference description into a vector embedding using the

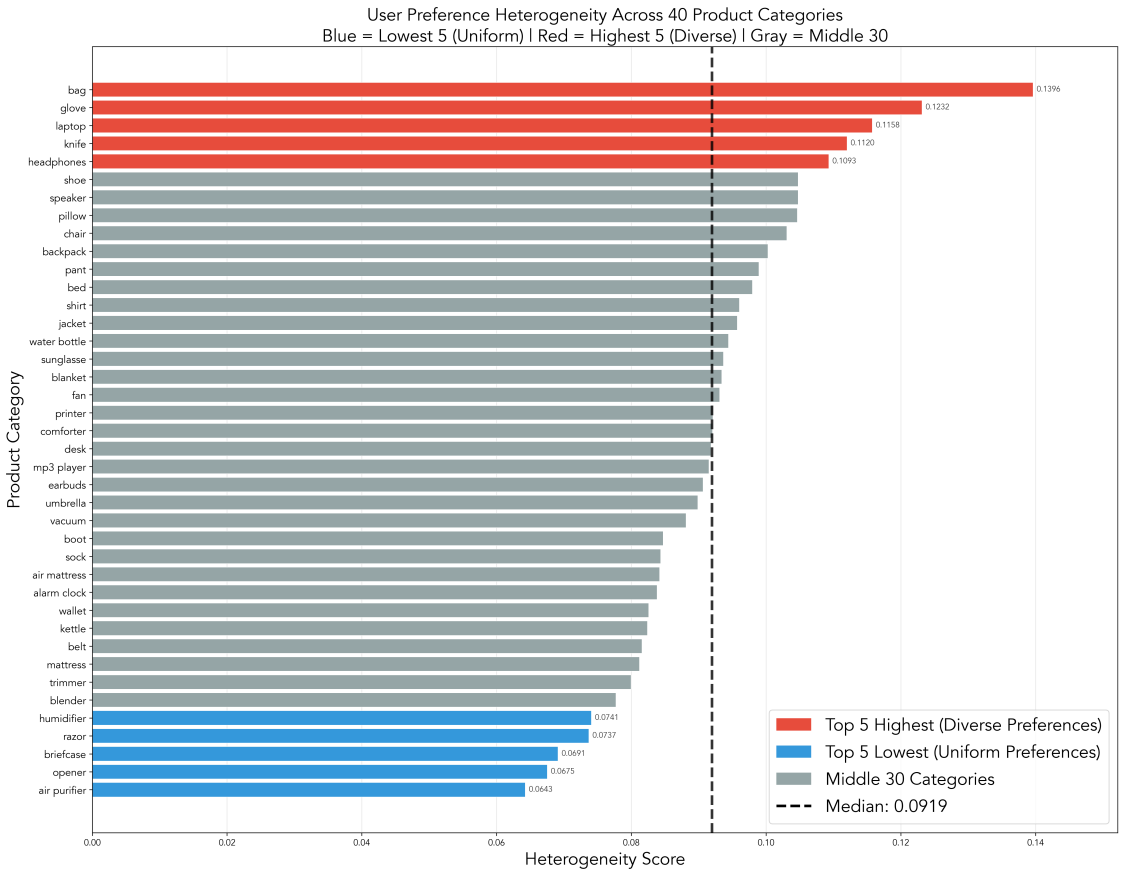
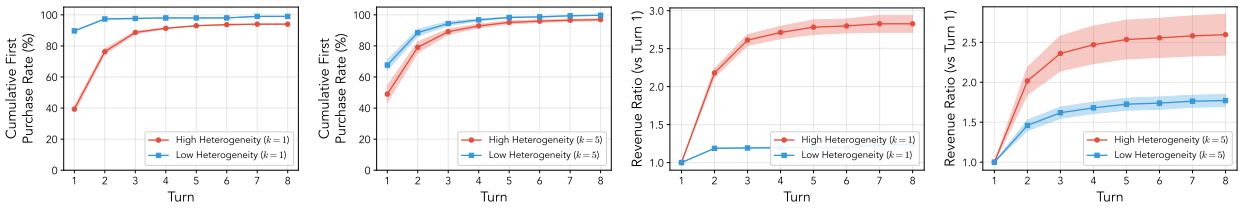


Figure 7 **The heterogeneity scores for all 40 categories, sorted from most heterogeneous preferences to least.**

`BAAI/bge-large-en-v1.5` sentence encoder, compute the category centroid as the mean embedding across users, and then measure each user's cosine distance to this centroid. We define category-level heterogeneity as the average of these distances (see Figure 7). Intuitively, smaller average distances indicate that users' stated needs cluster tightly around a common "typical" profile (more homogeneous preferences), whereas larger distances indicate a more dispersed preference distribution (greater heterogeneity).

Using the LLM-based simulation pipeline, we can quantify the marginal value of additional preference information under different outcome metrics and relate these returns to measured preference heterogeneity across categories. Figure 8 reports the incremental effects of revealing more preference statements on (i) purchase probability (a conversion proxy) and (ii) revenue/WTP-based outcomes (a commission-relevant proxy), averaged separately over the k most heterogeneous and the k most homogeneous categories. Two patterns stand out. First, both conversion and value-weighted outcomes improve as more information is revealed, consistent with the basic premise that

elicitation can sharpen matching.[5] Second, these gains are substantially larger in high-heterogeneity categories, echoing the within-platform variation illustrated in Figure 1: when user needs are more dispersed, incremental information has higher predictive value and yields larger improvements in downstream outcomes. In contrast, we see clear diminishing returns in low-heterogeneity categories—particularly for conversion—consistent with the theoretical comparative statics that low heterogeneity reduces the payoff to intensive elicitation.



(a) Purchase Rate ($k=1$) (b) Purchase Rate ($k=5$) (c) Revenue Ratio ($k=1$) (d) Revenue Ratio ($k=5$)

Figure 8 The average purchase rate and revenue ratio of the top/bottom categories and top/bottom 5 categories in terms of heterogeneity. Revenue is the sum of all WTPs, and is normalized to a ratio of the Turn 1 revenue for aggregation.

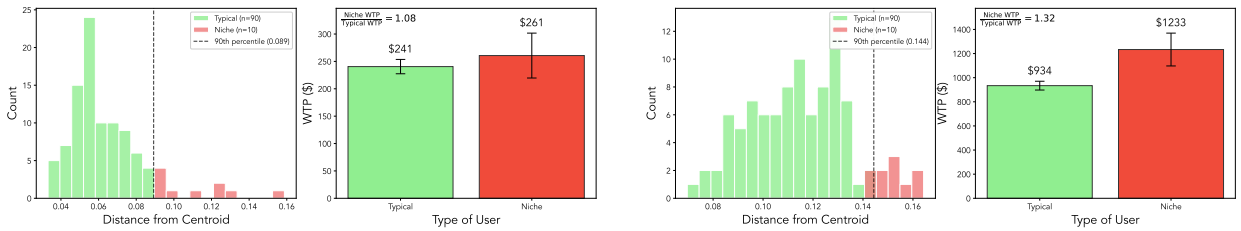
Beyond supporting the paper's qualitative mechanisms, this exercise illustrates the broader value of the dataset and evaluation protocol as a reusable testbed. Because preferences are revealed in controlled increments, the framework enables transparent measurement of "returns to elicitation" across categories, objectives, and conversational depths, and it can be used to run counterfactuals or benchmark conversational recommender designs under explicit information constraints. In Section 4.4, we operationalize the same ingredients in a simple abandonment simulation that maps the empirical incremental-value curves into an explicit when-to-recommend trade-off.

4.4. Simulation Environment Case Study

We use our user preference dataset and LLM-based outcomes to construct a simulation testbed and focus on two categories with starkly different niche needs and valuations: *laptops* and *air purifiers*.

Identifying niche users. Within each category, we embed each persona's natural-language requirements using `BAAI/bge-large-en-v1.5` and compute cosine distance to the category centroid. Users in the top decile of distances are classified as *niche* ($\alpha = 0.10$) and the remainder as *typical*. Niche laptop users exhibit substantially higher WTP at first purchase than typical users, while niche and typical air-purifier users have similar WTP (see Figure 9). In the language of our model, laptops resemble an environment in which the niche users obtain high value if correctly matched, whereas air purifiers resemble an environment with bounded niche value. Note that this is only an analogy rather than an identification of model primitives.

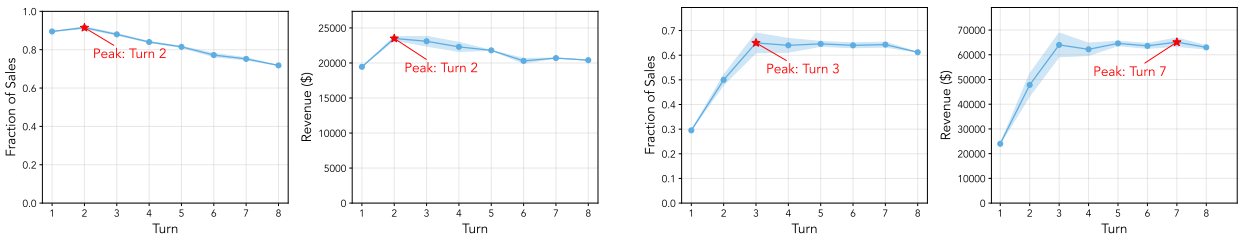
[5] Although each recommendation round includes both new and previously shown products, more than 85% of first purchases in each category are associated with newly added products, suggesting that the increase in purchase probability reflects improved recommendations.



(a) Air purifiers (b) Laptops

Figure 9 (Left) Distances from the category preference centroid; users above the 90th percentile are classified as niche ($\alpha = 0.10$). (Right) Mean WTP at the first purchase for typical vs. niche users, with 95% CI.

Abandonment simulation and the timing of recommendation. We simulate a platform decision that is analogous to the design question we study in this paper "how long to elicit before recommending." Fix a per-turn departure rate $\eta \in (0,1)$ and a recommendation turn $f \in \{1,\ldots,8\}$. All users start active at turn 1. For each turn $t < f$, each active user departs independently with probability η; departed users receive no recommendation. At turn f, the platform shows recommendations to all surviving users, and each user's purchase decision and WTP are drawn from our LLM evaluation at information level f (i.e., the f-sentence truncation described in Section 4.2). We record (i) the fraction of sales (purchases divided by the initial number of users) and (ii) total value-weighted revenue (sum of WTP among purchasers). The revenue measure is a proxy constructed from simulated WTP to reflect value-weighted objectives (in the spirit of commission monetization). We repeat the simulation 1,000 times for each (f, η).



(a) Air purifiers (b) Laptops

Figure 10 Abandonment simulation with per-turn departure rate $\eta = 4.5\%$ for (a) air purifier and (b) laptops.

Results and connection to the theory. The patterns align with the comparative predictions emphasized in the main results. For air purifiers, both sales volume and value-weighted revenue peak at a short interaction (turn 2) and then decline, consistent with limited marginal benefit from additional elicitation once attrition is accounted for. In terms of the theory, this category resembles an environment in which the payoff from identifying rare niche types is limited, and therefore both conversion-oriented and value-weighted objectives favor *minimal* elicitation. For laptops, the objective matters: sales peak earlier (turn 3) than value-weighted revenue (turn 7). This divergence

is the empirical analogue of the paper's central wedge: when a small segment has substantially higher valuation, deeper elicitation can be privately attractive under value-weighted monetization (the commission-like logic), even when it is not optimal for maximizing the number of conversions (the engagement logic). The underlying mechanism is the same as in the theory: longer interaction improves targeting and disproportionately increases outcomes for high-value niche users, but it also increases abandonment.

Scope. The simulation differs from the theory in several ways (elicitation is proxied by persona truncation rather than an $(n+1)$-ary channel; outcomes come from LLM-user simulations). We therefore interpret the case study as illustrative. Its role is to show, using empirically grounded variation in preference content and simulated abandonment, how the same basic trade-off can generate early-recommendation optima in low-value-tail categories and objective-dependent optima in high-value-tail categories.

5. Conclusion

Conversational recommender systems turn recommendation into an interactive design problem: the system can elicit preference information, but deeper interaction consumes attention and increases abandonment. This paper develops a parsimonious framework that connects conversational elicitation (modeled as noisy information acquisition) to downstream adoption and pricing, and uses it to study how optimal elicitation depends on the platform's monetization model.

Our results have direct design implications. Platforms whose marginal value comes mainly from getting a user to consume *something* (many subscription and ad-supported settings) should expect limited private returns to high-resolution elicitation in rare-gems environments; improvements in match quality translate into bounded changes in conversion, while abandonment affects the entire user base. By contrast, platforms whose monetization scales with transaction value have stronger incentives to invest in conversational screening: elicitation can pay for itself through equilibrium price responses. The analysis also clarifies why conversational behavior should vary across categories within a platform: heterogeneity increases the stakes of matching and strengthens incentives to learn, whereas high variety increases the information required to identify the right niche and pushes design toward coarse, robust recommendations and minimal probing.

Modeling limitations and future directions. The theory abstracts from several features that matter in deployed systems. First, we summarize multi-turn dialogue with a single fidelity choice and assume the recommendation arrives only at the end; modeling sequential question selection and endogenous stopping would connect more directly to progressive elicitation and early-exit behavior. Second, allowing structured attributes, non-uniform priors, and asymmetric confusability would yield sharper predictions about *which* questions are valuable and when elicitation should be targeted. Third, allowing content- and state-dependent hazards (e.g., perceived progress, latency)

would improve the mapping from model parameters to UX design choices. Fourth, richer supply-side structure (competition, sponsored placements, ranking, assortment) would clarify when elicitation intensifies competition versus when it increases seller market power.

Empirical testbed and limitations. Our empirical section is designed to measure the *marginal value of incremental preference information* in a controlled way by progressively revealing additional preference statements and re-evaluating outcomes. Because purchase decisions and WTP are generated by LLM-based evaluation, we do not interpret them as estimates of human demand primitives; rather, the dataset's value is methodological. It provides a reproducible testbed for benchmarking CRS designs across categories, for quantifying how recommendation quality and conversion proxies change with additional information, and for relating these patterns to measured preference dispersion. A few limitations are worth stating explicitly. First, LLM-based user simulation may reflect model-specific biases and prompt sensitivity; it is a proxy for human behavior rather than a substitute for field data. Second, our WTP measure is a model-based construct rather than observed payments, and it is best interpreted as an internally consistent ranking of value across conditions. Third, heterogeneity is measured via embedding geometry, which is informative but not equivalent to economic dispersion in tastes or valuations. These constraints mean that the empirical analysis is best viewed as descriptive and diagnostic, and as a complement to the clearly identified comparative statics in the theory.

References

Afzali J, Drzewiecki AM, Balog K, Zhang S (2023) Usersimcrs: A user simulation toolkit for evaluating conversational recommender systems. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 1160–1163.

Allouah A, Besbes O, Figueroa JD, Kanoria Y, Kumar A (2025) What is your ai agent buying? evaluation, biases, model dependence, & emerging implications for agentic e-commerce. *arXiv preprint arXiv:2508.02630* .

Bagga PS, Farias VF, Korkotashvili T, Peng T, Wu Y (2025) E-GEO: A testbed for generative engine optimization in e-commerce. URL `http://dx.doi.org/10.48550/arXiv.2511.20867`.

Banchio M, Mehta A, Perlroth A (2024) Ads in conversations. *arXiv preprint arXiv:2403.11022* .

Bergemann D, Morris S (2016) Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics* 11(2):487–522.

Bernard N, Joko H, Hasibi F, Balog K (2025) Crs arena: Crowdsourced benchmarking of conversational recommender systems. *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 1028–1031.

Besbes O, Kanoria Y, Kumar A (2024) The fault in our recommendations: On the perils of optimizing the measurable. *Proceedings of the 18th ACM Conference on Recommender Systems*, 200–208.

Blum A, Jackson J, Sandholm T, Zinkevich M (2004) Preference elicitation and query learning. *Journal of Machine Learning Research* 5(Jun):649–667.

Boutilier C (2002) A pomdp formulation of preference elicitation problems. *Eighteenth National Conference on Artificial Intelligence*, 239–246 (USA: American Association for Artificial Intelligence), ISBN 0262511290.

Brand J, Israeli A, Ngwe D (2023) Using llms for market research. *Harvard business school marketing unit working paper* 23-062.

Cao S, Hu M (2026) A solicit-then-suggest model of agentic purchasing. *Rotman School of Management Working Paper (forthcoming)* .

Caplin A, Dean M (2015) Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* 105(7):2183–2203.

Castro F, Gao J, Martin S (2024) Human-ai interactions and societal pitfalls. *Proceedings of the 25th ACM Conference on Economics and Computation*, 205, EC '24 (New York, NY, USA: Association for Computing Machinery), ISBN 9798400707049, URL `http://dx.doi.org/10.1145/3670865.3673482`.

Christakopoulou K, Radlinski F, Hofmann K (2016) Towards conversational recommender systems. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 815–824.

Dong J, Jhunjhunwala P, Kanoria Y (2025) Human-ai interaction in product recommendation. *NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*.

Filippas A, Horton JJ, Manning BS (2024) Large language models as simulated economic agents: What can we learn from homo silicus? *Proceedings of the 25th ACM Conference on Economics and Computation*, 614–615.

He Z, Xie Z, Jha R, Steck H, Liang D, Feng Y, Majumder BP, Kallus N, McAuley J (2023) Large language models as zero-shot conversational recommenders. *Proceedings of the 32nd ACM international conference on information and knowledge management*, 720–730.

Hou Y, Li J, He Z, Yan A, Chen X, McAuley J (2024) Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952* .

Jannach D, Manzoor A, Cai W, Chen L (2021) A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)* 54(5):1–36.

Kamenica E, Gentzkow M (2011) Bayesian persuasion. *American Economic Review* 101(6):2590–2615.

Li R, Ebrahimi Kahou S, Schulz H, Michalski V, Charlin L, Pal C (2018) Towards deep conversational recommendations. *Advances in neural information processing systems* 31.

Li W, Zhao M, Dong W, Cai J, Wei Y, Pocress M, Li Y, Yuan W, Wang X, Hou R, et al. (2026) Grading scale impact on llm-as-a-judge: Human-llm alignment is highest on 0-5 grading scale. *arXiv preprint arXiv:2601.03444* .

Manning BS, Horton JJ (2025) General social agents. *arXiv preprint arXiv:2508.17407* .

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105(1):272–298.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed., *Frontiers in Econometrics*, 105–142 (New York: Academic Press).

Müller M (2025) Advancing user-centric evaluation and design of conversational recommender systems. *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 1445–1450.

Park C, Donahue K, Raghavan M (2025) When to ask a question: Understanding communication strategies in generative ai tools. *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, 288–299.

Parkes DC (2005) Auction design with costly preference elicitation. *Annals of Mathematics and Artificial Intelligence* 44(3):269–302, URL `http://dx.doi.org/10.1007/s10472-005-4692-y`.

Peng Q, Liu H, Huang H, Yang Q, Shao M (2025) A survey on llm-powered agents for recommender systems. *arXiv preprint arXiv:2502.10050* .

Peng T, Gui G, Brucks M, Merlau DJ, Fan GJ, Sliman MB, Johnson EJ, Althenayyan A, Bellezza S, Donati D, Fong H, Friedman E, Guevara A, Hussein M, Jerath K, Kogut B, Kumar A, Lane K, Li H, Morwitz V, Netzer O, Perkowski P, Toubia O (2026) Digital twins as funhouse mirrors: Five key distortions. URL `https://arxiv.org/abs/2509.19088`.

Sanner S, Balog K, Radlinski F, Wedin B, Dixon L (2023) Large language models are competitive near cold-start recommenders for language-and item-based preferences. *Proceedings of the 17th ACM conference on recommender systems*, 890–896.

Sims CA (2003) Implications of rational inattention. *Journal of monetary Economics* 50(3):665–690.

Sun L, Fu S, Yao B, Lu Y, Li W, Gu H, Gesi J, Huang J, Luo C, Wang D (2025) Llm agent meets agentic ai: Can llm agents simulate customers to evaluate agentic-ai-based shopping assistants? *arXiv preprint arXiv:2509.21501* .

Voltes-Dorta A, Suau-Sanchez P (2025) Can large language models mimic airline passenger preferences? *Artificial Intelligence for Transportation* 3:100034.

Wang X, Tang X, Zhao WX, Wang J, Wen JR (2023) Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112* .

Weitzman ML (1979) Optimal search for the best alternative. *Econometrica* 47(3):641–654, ISSN 00129682, 14680262, URL `http://www.jstor.org/stable/1910412`.

Zhang X, Xie H, Li H, CS Lui J (2020) Conversational contextual bandit: Algorithm and application. *Proceedings of the web conference 2020*, 662–672.

Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing E, et al. (2023) Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36:46595–46623.

Appendix Table of Contents

Appendix A: Proofs for results in Section 3

This appendix provides proofs for the results in Section 3. Throughout, fix $n \geq 1$, $\alpha \in (0,1)$, and the primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta, \bar{\kappa}_u)$. Let $\Theta \in \{\emptyset, 1, \ldots, n\}$ be the user type with prior (1), and let $S \in \{\emptyset, 1, \ldots, n\}$ be generated by the symmetric channel (5) with fidelity $f \in [1/(n+1), 1]$.

Define the base utilities (with fixed prices $p_j \equiv 0$ unless stated otherwise)

$$u_{\mathsf{P}} := V_{\mathsf{P}}, \qquad u_- := -\lambda, \qquad u_+(\alpha) := \lambda V_{\mathsf{N}} \alpha^{-\beta}. \tag{13}$$

Let $\sigma(x) := \exp(x)/(1+\exp(x))$ and $g(x) := \ln(1+\exp(x))$.

Information, survival, and effort loss. Under the mutual-information benchmark,

$$\mathcal{I}^{\mathsf{MI}}(f) := I_{\alpha,n}(f) := I(\Theta; S) \quad \text{(nats)}.$$

Survival to the recommendation stage is

$$sur^{\mathsf{MI}}(f) = \exp\big(-\eta I_{\alpha,n}(f)\big), \tag{14}$$

and expected effort loss (counting abandoned sessions) is

$$L_u^{\mathsf{MI}}(f) = \bar{\kappa}_u \int_0^{I_{\alpha,n}(f)} e^{-\eta x}\, dx = \frac{\bar{\kappa}_u}{\eta}\big(1 - sur^{\mathsf{MI}}(f)\big). \tag{15}$$

A.1. Preliminaries: posterior and threshold rules

LEMMA 1 **(Posterior correctness)**. *Fix a fidelity f and a niche signal realization $S = s \in \{1, \ldots, n\}$. The posterior probability that the signal is correct is*

$$q(f, \alpha, n) := \Pr(\Theta = s \mid S = s) = \frac{\alpha f}{\alpha f + (1 - \alpha/n)(1-f)}. \tag{16}$$

Moreover, $q(f, \alpha, n)$ is strictly increasing in f.

Proof of Lemma 1: Fix $s \in \{1, \ldots, n\}$. By Bayes' rule,

$$\Pr(\Theta = s \mid S = s) = \frac{\Pr(\Theta = s)\Pr(S = s \mid \Theta = s)}{\sum_{t \in \{\emptyset, 1, \ldots, n\}} \Pr(\Theta = t)\Pr(S = s \mid \Theta = t)}.$$

Under (1), $\Pr(\Theta = s) = \alpha/n$. Under (5), $\Pr(S = s \mid \Theta = s) = f$ and $\Pr(S = s \mid \Theta = t) = (1-f)/n$ for every $t \neq s$. Hence,

$$\Pr(S = s) = \frac{\alpha}{n} f + \Big(1 - \frac{\alpha}{n}\Big)\frac{1-f}{n},$$

and substituting yields (16). Strict monotonicity in f follows by direct differentiation of (16):

$$\frac{\partial}{\partial f} q(f, \alpha, n) = \frac{\alpha(1 - \alpha/n)}{\big(\alpha f + (1-\alpha/n)(1-f)\big)^2} > 0. \quad \square$$

LEMMA 2 **(Threshold structure for conversion and user welfare (fixed prices))**. *Fix f and a niche signal $S = s \in \{1, \ldots, n\}$. Consider fixed prices $p_j \equiv 0$.*

(i) Under the welfare objective, conditional on $S = s$ the optimal recommendation is either P or niche s. Recommending niche s is optimal iff

$$q(f, \alpha, n) \;\geq\; q_W^\star(\alpha, \lambda) := \frac{g(u_{\mathsf{P}}) - g(u_-)}{g(u_+(\alpha)) - g(u_-)}. \tag{17}$$

(ii) Under the conversion objective, conditional on $S = s$ the optimal recommendation is either P or niche s. Recommending niche s is optimal iff

$$q(f, \alpha, n) \;\geq\; q^{\star}_{\Gamma}(\alpha, \lambda) := \frac{\sigma(u_{\mathsf{P}}) - \sigma(u_{-})}{\sigma(u_{+}(\alpha)) - \sigma(u_{-})}. \tag{18}$$

Proof of Lemma 2: We prove (i); (ii) is identical with g replaced by σ.

Fix $S = s$. If the platform recommends P, then the (choice-stage) expected surplus for this user is $g(u_{\mathsf{P}})$, independent of Θ because $u_{\mathsf{B}}(\theta, \mathsf{P}) = u_{\mathsf{P}}$ for all θ.

If instead it recommends niche s, then the user has base utility $u_{+}(\alpha)$ if $\Theta = s$ and u_{-} if $\Theta \neq s$. Hence conditional on $(S = s)$, expected surplus from recommending niche s equals

$$q(f, \alpha, n)\, g(u_{+}(\alpha)) + \big(1 - q(f, \alpha, n)\big)\, g(u_{-}).$$

Therefore, recommending niche s is optimal iff

$$q(f, \alpha, n)\, g(u_{+}) + (1 - q(f, \alpha, n))\, g(u_{-}) \;\geq\; g(u_{\mathsf{P}}),$$

which rearranges to (17) since g is strictly increasing and $g(u_{+}) > g(u_{-})$.

Finally, recommending any other niche $j \neq s$ yields a smaller probability of a correct match conditional on $S = s$ under the symmetric model, and weakly lowers expected surplus relative to recommending s (it replaces the u_{+} payoff on event $\{\Theta = s\}$ by u_{-}). Thus the only candidates are P and s. □

Corollary 2 **(Minimal fidelity for acting on niche signals).** *Fix (α, n) and $\mathsf{obj} \in \{W, \Gamma\}$. Let $q^{\star} := q^{\star}_{\mathsf{obj}}(\alpha, \lambda) \in (0, 1)$ be the corresponding posterior cutoff in Lemma 2. Define*

$$\underline{f}_{\mathsf{obj}}(\alpha, n, \lambda) := \frac{q^{\star}(1 - \alpha/n)}{q^{\star}(1 - \alpha/n) + \alpha(1 - q^{\star})}. \tag{19}$$

Then $q(f, \alpha, n) \geq q^{\star}$ iff $f \geq \underline{f}_{\mathsf{obj}}$. In particular, if $f < \underline{f}_{\mathsf{obj}}$ then the optimal recommendation rule ignores niche signals and recommends P even after $S \in \{1, \ldots, n\}$.

Proof of Corollary 2: Solve (16) for f under the inequality $q(f, \alpha, n) \geq q^{\star}$. Because $q(f, \alpha, n)$ is strictly increasing in f (Lemma 1), the set of f satisfying the inequality is an interval $[\underline{f}, 1]$. Algebra gives (19). The final claim follows from Lemma 2. □

A.2. Mutual-information expansions used in the rare-niche regime

Let $h(x) := -x \ln x - (1 - x) \ln(1 - x)$ denote the binary entropy function (in nats).

Lemma 3 **(Entropy of the type prior).** *Let Θ have prior (1). Then*

$$H(\Theta) = h(\alpha) + \alpha \ln n. \tag{20}$$

Moreover, as $\alpha \to 0$,

$$H(\Theta) = \alpha \ln(1/\alpha) + \alpha(1 + \ln n) + o(\alpha). \tag{21}$$

Proof of Lemma 3: Direct calculation yields (20). The expansion (21) follows from the standard Taylor expansion $h(\alpha) = \alpha \ln(1/\alpha) + \alpha + o(\alpha)$ as $\alpha \to 0$. □

LEMMA 4 **(Mutual-information closed form).** *Let $z := (1-f)/n$. The signal marginal satisfies*

$$q_0 := \Pr(S=\emptyset) = (1-\alpha)f + \alpha z, \qquad q_1 := \Pr(S=i) = \frac{\alpha}{n}f + \Big(1-\frac{\alpha}{n}\Big)z \quad (i=1,\dots,n),$$

and $q_0 + nq_1 = 1$. The mutual information is

$$I_{\alpha,n}(f) = \Big[-q_0 \ln q_0 - nq_1 \ln q_1\Big] - \Big[-f\ln f - (1-f)\ln\Big(\frac{1-f}{n}\Big)\Big]. \tag{22}$$

Proof of Lemma 4: By definition $I(\Theta;S) = H(S) - H(S \mid \Theta)$. Under the symmetric channel, $H(S \mid \Theta)$ equals the entropy of one row of the transition matrix, i.e. $-f\ln f - (1-f)\ln((1-f)/n)$. The marginal $H(S)$ uses the symmetry that $\Pr(S=i)$ is the same for all niches. Substituting yields (22). □

LEMMA 5 **(Rare-niche expansion for $f = 1-c\alpha$).** *Fix $n \geq 1$ and a constant $c > 0$. Let $f(\alpha) := 1-c\alpha$. Then as $\alpha \to 0$,*

$$I_{\alpha,n}(f(\alpha)) = \alpha\left[\ln\Big(\frac{1}{\alpha}\Big) + 1 + \ln\Big(\frac{n\,c^c}{(c+1)^{c+1}}\Big)\right] + o(\alpha). \tag{23}$$

Consequently,

$$sur^{\mathsf{MI}}(f(\alpha)) = 1 - \eta\alpha\ln(1/\alpha) - \eta\alpha\left[1 + \ln\Big(\frac{n\,c^c}{(c+1)^{c+1}}\Big)\right] + o(\alpha). \tag{24}$$

Proof of Lemma 5: The proof consists of five steps: 1) we derive a closed form expression for $I_{\alpha,n}(f)$, 2) we derive the relevant probabilities when $f(\alpha) = 1-c\alpha$, 3) expand each term in the closed form $I_{\alpha,n}(f)$, 4) combine terms and simplify, and 5) use the final expression to derive the survival function expression.

Fix $n \geq 1$ and $c > 0$. Throughout the proof, all logarithms are natural and all $o(\cdot), O(\cdot)$ statements are as $\alpha \to 0$.

Step 1: Closed-form expression for $I_{\alpha,n}(f)$. Recall $\Theta \in \{\emptyset, 1, \dots, n\}$ with prior $\Pr(\Theta = \emptyset) = 1-\alpha$ and $\Pr(\Theta = j) = \alpha/n$ for $j \in \{1,\dots,n\}$. The signal $S \in \{\emptyset, 1, \dots, n\}$ is generated by the symmetric channel

$$\Pr(S = s \mid \Theta = t) = \begin{cases} f, & s = t, \\ (1-f)/n, & s \neq t. \end{cases}$$

Since each row of the channel is a permutation of the same distribution, the conditional entropy $H(S \mid \Theta)$ does not depend on Θ and equals

$$\begin{aligned} H(S \mid \Theta) &= -\left(f\ln f + n\cdot\frac{1-f}{n}\ln\Big(\frac{1-f}{n}\Big)\right) \\ &= -f\ln f - (1-f)\ln\Big(\frac{1-f}{n}\Big). \end{aligned} \tag{25}$$

Let $P_0 := \Pr(S = \emptyset)$ and $P_1 := \Pr(S = i)$ for any fixed niche label $i \in \{1,\dots,n\}$. By the law of total probability,

$$\begin{aligned} P_0 &= \Pr(\Theta=\emptyset)\Pr(S=\emptyset \mid \Theta=\emptyset) + \sum_{j=1}^{n}\Pr(\Theta=j)\Pr(S=\emptyset \mid \Theta=j) \\ &= (1-\alpha)f + \alpha\cdot\frac{1-f}{n}. \end{aligned} \tag{26}$$

Similarly, for any niche label i,

$$\begin{aligned} P_1 &= \Pr(\Theta=i)\Pr(S=i \mid \Theta=i) + \Pr(\Theta\neq i)\Pr(S=i \mid \Theta\neq i) \\ &= \frac{\alpha}{n}f + \Big(1-\frac{\alpha}{n}\Big)\frac{1-f}{n}. \end{aligned} \tag{27}$$

The marginal entropy is therefore

$$H(S) = -P_0 \ln P_0 - nP_1 \ln P_1. \tag{28}$$

Thus mutual information is

$$\begin{aligned} I_{\alpha,n}(f) &= H(S) - H(S \mid \Theta) \\ &= -P_0 \ln P_0 - nP_1 \ln P_1 + f \ln f + (1-f) \ln\Big(\frac{1-f}{n}\Big). \end{aligned} \tag{29}$$

Step 2: Specialize to $f(\alpha) = 1 - c\alpha$ and expand P_0, P_1. Let $f(\alpha) = 1 - c\alpha$ and write $\delta(\alpha) := 1 - f(\alpha) = c\alpha$. Substituting into (26) gives the exact identity

$$\begin{aligned} P_0(\alpha) &= (1-\alpha)(1-c\alpha) + \alpha \cdot \frac{c\alpha}{n} \\ &= 1 - (1+c)\alpha + c\alpha^2\Big(1 + \frac{1}{n}\Big). \end{aligned} \tag{30}$$

Similarly, substituting into (27) yields the exact identity

$$\begin{aligned} P_1(\alpha) &= \frac{\alpha}{n}(1-c\alpha) + \Big(1 - \frac{\alpha}{n}\Big)\frac{c\alpha}{n} \\ &= \frac{(1+c)\alpha}{n} - c\alpha^2\Big(\frac{1}{n} + \frac{1}{n^2}\Big). \end{aligned} \tag{31}$$

In particular,

$$P_0(\alpha) = 1 - (1+c)\alpha + O(\alpha^2), \qquad P_1(\alpha) = \frac{(1+c)\alpha}{n} + O(\alpha^2). \tag{32}$$

Step 3: Expand each term in (29). We expand the four components

$$A(\alpha) := -P_0 \ln P_0, \qquad B(\alpha) := -nP_1 \ln P_1, \qquad C(\alpha) := f \ln f, \qquad D(\alpha) := (1-f)\ln\Big(\frac{1-f}{n}\Big),$$

so that $I_{\alpha,n}(f(\alpha)) = A(\alpha) + B(\alpha) + C(\alpha) + D(\alpha)$.

(a) Expansion of $A(\alpha) = -P_0 \ln P_0$. Let $x_\alpha := 1 - P_0(\alpha)$. From (30),

$$x_\alpha = (1+c)\alpha + O(\alpha^2), \qquad \text{and hence} \qquad P_0(\alpha) = 1 - x_\alpha.$$

Using the Taylor expansion $\ln(1-x) = -x - \frac{x^2}{2} + O(x^3)$ as $x \to 0$, we obtain

$$\begin{aligned} A(\alpha) &= -(1-x_\alpha)\ln(1-x_\alpha) = -(1-x_\alpha)\Big(-x_\alpha - \frac{x_\alpha^2}{2} + O(x_\alpha^3)\Big) \\ &= x_\alpha + O(x_\alpha^2) = (1+c)\alpha + O(\alpha^2). \end{aligned}$$

Thus

$$A(\alpha) = (1+c)\alpha + O(\alpha^2). \tag{33}$$

(b) Expansion of $B(\alpha) = -nP_1 \ln P_1$. Define

$$z := \frac{1+c}{n} > 0.$$

From (31), we can write

$$P_1(\alpha) = z\alpha\big(1 + \varepsilon_\alpha\big) \quad \text{with} \quad \varepsilon_\alpha = O(\alpha).$$

Then

$$\ln P_1(\alpha) = \ln(z\alpha) + \ln(1+\varepsilon_\alpha) = \ln\alpha + \ln z + O(\alpha),$$

since $\ln(1+\varepsilon_\alpha) = \varepsilon_\alpha + O(\varepsilon_\alpha^2) = O(\alpha)$. Therefore

$$\begin{aligned} P_1(\alpha)\ln P_1(\alpha) &= z\alpha(1+\varepsilon_\alpha)\Big(\ln\alpha + \ln z + O(\alpha)\Big) \\ &= z\alpha(\ln\alpha + \ln z) \;+\; z\alpha\,\varepsilon_\alpha(\ln\alpha + \ln z) \;+\; O(\alpha^2). \end{aligned}$$

Because $\varepsilon_\alpha = O(\alpha)$, the middle term is $O(\alpha^2\ln(1/\alpha)) = o(\alpha)$ (since $\alpha\ln(1/\alpha)\to 0$). Hence

$$P_1(\alpha)\ln P_1(\alpha) = z\alpha(\ln\alpha + \ln z) + o(\alpha).$$

Multiplying by $-n$ gives

$$\begin{aligned} B(\alpha) &= -nP_1(\alpha)\ln P_1(\alpha) \\ &= -n\cdot z\alpha(\ln\alpha + \ln z) + o(\alpha) = -(1+c)\alpha\ln\alpha - (1+c)\alpha\ln z + o(\alpha). \end{aligned} \tag{34}$$

Recalling $z = (1+c)/n$,

$$B(\alpha) = -(1+c)\alpha\ln\alpha - (1+c)\alpha\ln\Big(\frac{1+c}{n}\Big) + o(\alpha). \tag{35}$$

(c) Expansion of $C(\alpha) = f\ln f$. With $f(\alpha) = 1-c\alpha$, Taylor expansion yields

$$\ln f(\alpha) = \ln(1-c\alpha) = -c\alpha - \frac{c^2\alpha^2}{2} + O(\alpha^3),$$

so

$$\begin{aligned} C(\alpha) &= (1-c\alpha)\Big(-c\alpha - \frac{c^2\alpha^2}{2} + O(\alpha^3)\Big) \\ &= -c\alpha + O(\alpha^2). \end{aligned} \tag{36}$$

(d) Expansion of $D(\alpha) = (1-f)\ln((1-f)/n)$. Here $1-f(\alpha) = c\alpha$ exactly, so

$$D(\alpha) = c\alpha\ln\Big(\frac{c\alpha}{n}\Big) = c\alpha\big(\ln\alpha + \ln c - \ln n\big). \tag{37}$$

Step 4: Combine terms and simplify. Summing (33), (35), (36), and (37) gives

$$\begin{aligned} I_{\alpha,n}(f(\alpha)) = &\Big((1+c)\alpha + O(\alpha^2)\Big) + \Big(-(1+c)\alpha\ln\alpha - (1+c)\alpha\ln\Big(\frac{1+c}{n}\Big) + o(\alpha)\Big) \\ &+ \Big(-c\alpha + O(\alpha^2)\Big) + \Big(c\alpha\ln\alpha + c\alpha(\ln c - \ln n)\Big). \end{aligned}$$

Collecting the $\alpha\ln\alpha$ terms:

$$-(1+c)\alpha\ln\alpha + c\alpha\ln\alpha = -\alpha\ln\alpha = \alpha\ln(1/\alpha).$$

Collecting the linear-in-α (non-log) terms:

$$(1+c)\alpha - c\alpha = \alpha,$$

and the remaining constants from logarithms:

$$\begin{aligned}
&-(1+c)\alpha\ln\Big(\frac{1+c}{n}\Big)+c\alpha(\ln c-\ln n)\\
&\quad=-(1+c)\alpha(\ln(1+c)-\ln n)+c\alpha\ln c-c\alpha\ln n\\
&\quad=-(1+c)\alpha\ln(1+c)+\alpha\ln n+c\alpha\ln c.
\end{aligned}$$

Therefore

$$I_{\alpha,n}(f(\alpha))=\alpha\ln(1/\alpha)+\alpha\Big(1+\ln n+c\ln c-(1+c)\ln(1+c)\Big)+o(\alpha). \tag{38}$$

Finally, observe that

$$\ln n+c\ln c-(1+c)\ln(1+c)=\ln\Big(\frac{n\,c^c}{(1+c)^{1+c}}\Big),$$

so (38) is exactly (23).

Step 5: Expand survival. By definition, $sur^{\mathsf{MI}}(f)=\exp\big(-\eta I_{\alpha,n}(f)\big)$, so with $f(\alpha)=1-c\alpha$ we have

$$sur^{\mathsf{MI}}(f(\alpha))=\exp\big(-\eta I_{\alpha,n}(f(\alpha))\big).$$

From (23), $I_{\alpha,n}(f(\alpha))=\alpha\ln(1/\alpha)+O(\alpha)$, so $I_{\alpha,n}(f(\alpha))\to 0$ and moreover

$$I_{\alpha,n}(f(\alpha))^2=O\big(\alpha^2\ln^2(1/\alpha)\big)=o(\alpha)\qquad\text{as }\alpha\to 0,$$

since $\alpha\ln^2(1/\alpha)\to 0$. Using the Taylor expansion $\exp(-x)=1-x+O(x^2)$ as $x\to 0$ with $x=\eta I_{\alpha,n}(f(\alpha))$ yields

$$sur^{\mathsf{MI}}(f(\alpha))=1-\eta I_{\alpha,n}(f(\alpha))+o(\alpha).$$

Substituting (23) gives

$$sur^{\mathsf{MI}}(f(\alpha))=1-\eta\alpha\ln(1/\alpha)-\eta\alpha\left[1+\ln\Big(\frac{n\,c^c}{(c+1)^{c+1}}\Big)\right]+o(\alpha),$$

which is (24). □

LEMMA 6 **(A useful lower bound on $I_{\alpha,n}(f)$ when $1-f=O(\alpha)$).** *Fix $n\ge 1$ and $K<\infty$. There exists $\bar\alpha\in(0,1)$ such that for all $\alpha\in(0,\bar\alpha)$ and all f satisfying $1-f\le K\alpha$,*

$$I_{\alpha,n}(f)\;\ge\;\frac{1}{2}\,\alpha\ln(1/\alpha). \tag{39}$$

Proof of Lemma 6: Fix K. For α sufficiently small, $1-K\alpha\in[1/2,1]$. Under the symmetric channel family, a lower-fidelity channel can be obtained from a higher-fidelity channel by adding independent output noise (i.e., the family is ordered by Blackwell garbling). Hence by the data-processing inequality, $I_{\alpha,n}(f)$ is weakly increasing in f. Therefore, if $1-f\le K\alpha$ then $f\ge 1-K\alpha$ implies

$$I_{\alpha,n}(f)\;\ge\;I_{\alpha,n}(1-K\alpha).$$

Apply Lemma 5 with $c=K$ to obtain

$$I_{\alpha,n}(1-K\alpha)=\alpha\ln(1/\alpha)+O(\alpha)\qquad(\alpha\to 0).$$

Thus for α sufficiently small, $I_{\alpha,n}(1-K\alpha)\ge\frac{1}{2}\alpha\ln(1/\alpha)$, which yields (39). □

A.3. Proof of Theorem 1

Proof of Theorem 1: The proof consists of three steps: 1) we show that if the recommendation policy π never recommends a niche, then REC is optimal, 2) any optimal policy that does recommend a niche must have $1-f \le K\alpha$, and 3) such policies lose $O(\alpha \ln(1/\alpha))$ to abandonment, which dominates the improvement in conversion.

Let $\Gamma(f,\pi) = R\, sur^{\mathsf{MI}}(f)\, \mathbb{E}[\sigma(u_{\mathsf{B}}(\Theta, \pi(S)))]$ be the conversion objective with $p_j \equiv 0$.

Step 1: If π never recommends a niche, then REC *is optimal.* If $\pi(S) \equiv \mathsf{P}$ almost surely, then $u_{\mathsf{B}}(\Theta, \pi(S)) \equiv u_{\mathsf{P}}$ and

$$\Gamma(f,\pi) = R\, sur^{\mathsf{MI}}(f)\, \sigma(u_{\mathsf{P}}) \;\le\; R\,\sigma(u_{\mathsf{P}}) = \Gamma(\mathsf{REC}),$$

because $sur^{\mathsf{MI}}(f) \le 1$ with equality at $f = 1/(n+1)$ (no information, no abandonment). Thus among policies that never recommend a niche, REC is optimal.

Step 2: Any optimal policy that recommends a niche must have $1-f \le K\alpha$ for some constant K. Suppose the policy recommends a niche with positive probability. By symmetry and Lemma 2, this requires that after a niche signal $S = s$ the platform recommends niche s for some s, which in turn requires $q(f,\alpha,n) \ge q^\star_\Gamma(\alpha,\lambda)$.

We claim $q^\star_\Gamma(\alpha,\lambda)$ is bounded away from zero uniformly in small α. Indeed,

$$q^\star_\Gamma(\alpha,\lambda) = \frac{\sigma(u_{\mathsf{P}}) - \sigma(u_-)}{\sigma(u_+(\alpha)) - \sigma(u_-)} \;\ge\; \frac{\sigma(u_{\mathsf{P}}) - \sigma(u_-)}{1 - \sigma(u_-)} =: \underline{q}_\Gamma \in (0,1),$$

because $\sigma(u_+(\alpha)) \le 1$ and $\sigma(u_{\mathsf{P}}) > \sigma(u_-)$ (since $u_{\mathsf{P}} > u_-$). Hence $q(f,\alpha,n) \ge \underline{q}_\Gamma$ is necessary.

Rearranging (16) gives

$$1 - f = \frac{\alpha f (1-q)}{q(1-\alpha/n)} \;\le\; \frac{\alpha(1-q)}{q(1-\alpha/n)}.$$

Thus for $q \ge \underline{q}_\Gamma$ and α small (so $1-\alpha/n \ge 1/2$),

$$1 - f \;\le\; K\alpha, \qquad K := \frac{2(1-\underline{q}_\Gamma)}{\underline{q}_\Gamma}.$$

Step 3: Such policies incur an $O(\alpha \ln(1/\alpha))$ survival loss, which dominates the bounded conversion gain. By Lemma 6, for α small enough, $1-f \le K\alpha$ implies $I_{\alpha,n}(f) \ge \frac{1}{2}\alpha \ln(1/\alpha)$, hence

$$1 - sur^{\mathsf{MI}}(f) = 1 - \exp\big(-\eta I_{\alpha,n}(f)\big) \ge 1 - \exp\Big(-\frac{\eta}{2}\alpha \ln(1/\alpha)\Big).$$

Using $1 - e^{-x} \ge x/2$ for all sufficiently small $x > 0$, we get for α small:

$$1 - sur^{\mathsf{MI}}(f) \;\ge\; c_0\, \alpha \ln(1/\alpha) \quad \text{for some } c_0 > 0. \tag{40}$$

Next, we upper bound the adoption probability term. For any policy π and any (Θ, S),

$$\sigma(u_{\mathsf{B}}(\Theta, \pi(S))) \;\le\; 1.$$

Moreover, for $\Theta = \emptyset$ (mass $1-\alpha$), the adoption probability is maximized by recommending P, because $u_{\mathsf{B}}(\emptyset, \mathsf{P}) = u_{\mathsf{P}} > u_{\mathsf{B}}(\emptyset, j) = u_-$ for any niche j. Hence

$$\mathbb{E}[\sigma(u_{\mathsf{B}}(\Theta, \pi(S)))] \le (1-\alpha)\sigma(u_{\mathsf{P}}) + \alpha \cdot 1 = \sigma(u_{\mathsf{P}}) + \alpha(1 - \sigma(u_{\mathsf{P}})).$$

Therefore,

$$\begin{aligned}\Gamma(f,\pi)-\Gamma(\mathsf{REC}) &\le R\Big(sur^{\mathsf{MI}}(f)\big(\sigma(u_{\mathsf{P}})+\alpha(1-\sigma(u_{\mathsf{P}}))\big)-\sigma(u_{\mathsf{P}})\Big)\\ &= R\Big((sur^{\mathsf{MI}}(f)-1)\sigma(u_{\mathsf{P}})+sur^{\mathsf{MI}}(f)\alpha(1-\sigma(u_{\mathsf{P}}))\Big).\end{aligned}$$

Using $sur^{\mathsf{MI}}(f)\le 1$ and (40), we obtain for α small:

$$\Gamma(f,\pi)-\Gamma(\mathsf{REC}) \le R\Big(-c_0\sigma(u_{\mathsf{P}})\alpha\ln(1/\alpha)+\alpha(1-\sigma(u_{\mathsf{P}}))\Big)<0,$$

since $\alpha\ln(1/\alpha)\gg\alpha$ as $\alpha\to 0$. Thus any policy that recommends a niche is strictly worse than REC for α small. Combining with Step 1 completes the proof. □

A.4. Proof of Proposition 1

Note that the following lemma is only for use in part (iii) of Proposition 1, and uses a result from part (ii) which is independent.

LEMMA 7 **(Welfare-optimal fidelity is on the α-scale when $\beta=1$).** *Fix $n\ge 1$ and primitives $(V_{\mathsf{P}},V_{\mathsf{N}},\lambda,\eta,\bar\kappa_u)$ with $\lambda>0$. Let $\beta=1$, and $(f_W^\star(\alpha),\pi_W^\star(\alpha))$ be any welfare-maximizing CRS design, and define*

$$\delta_\alpha := 1-f_W^\star(\alpha),\qquad L:=\lambda V_{\mathsf{N}},\qquad g_{\mathsf{P}}:=g(u_{\mathsf{P}}),\qquad g_-:=g(u_-).$$

Also define

$$D_W := g_{\mathsf{P}}+L-g_-,\qquad K_W:=\bar\kappa_u+\eta(g_{\mathsf{P}}+L),$$

and

$$\Psi_W(c):=c(g_--g_{\mathsf{P}}-L)-K_W\left[1+\ln\Big(\frac{n\,c^c}{(1+c)^{1+c}}\Big)\right],\qquad c\ge 0,$$

where $0^0:=1$. Then there exist $K<\infty$ and $\bar\alpha\in(0,1)$ such that

$$0\le\delta_\alpha\le K\alpha\qquad \textit{for all } \alpha\in(0,\bar\alpha).$$

Moreover,

$$\frac{\delta_\alpha}{\alpha}\longrightarrow c_W^\star,$$

where $c_W^\star$ is the unique maximizer of Ψ_W on $[0,\infty)$, equivalently

$$\ln\Big(1+\frac{1}{c_W^\star}\Big)=\frac{D_W}{K_W},\qquad \textit{so}\qquad (c_W^\star)^{-1}=e^{D_W/K_W}-1.$$

Consequently,

$$1-f_W^\star(\alpha)=c_W^\star\,\alpha+o(\alpha).$$

Proof of Lemma 7 Part (ii) of Proposition 1 implies $\delta_\alpha\to 0$.

Step 1: for small gaps, the welfare-optimal rule follows niche signals. By Lemma 2, after a niche signal the welfare-optimal action is to recommend the indicated niche iff

$$q(f,\alpha,n) \geq q_W^\star(\alpha,\lambda) := \frac{g_{\mathsf{P}} - g_-}{g(u_+(\alpha)) - g_-}.$$

Since $\beta = 1$ and $u_+(\alpha) = L/\alpha$,

$$q_W^\star(\alpha,\lambda) = \frac{g_{\mathsf{P}} - g_-}{L}\,\alpha + o(\alpha).$$

Choose $\delta_0 \in (0,1/2)$ sufficiently small that

$$\frac{1-\delta_0}{2\delta_0} > \frac{g_{\mathsf{P}} - g_-}{L}.$$

If $\delta \in [0,\delta_0]$ and $\alpha \leq \delta_0$, then

$$q(1-\delta,\alpha,n) = \frac{\alpha(1-\delta)}{\alpha(1-\delta) + (1-\alpha/n)\delta} \geq \frac{\alpha(1-\delta_0)}{\alpha+\delta_0} \geq \frac{1-\delta_0}{2\delta_0}\,\alpha.$$

Hence every design with $\delta \in [0,\delta_0]$ follows niche signals for all sufficiently small α. Since $\delta_\alpha \to 0$, every welfare-optimal design does so as well.

Step 2: $\delta_\alpha = O(\alpha)$. For $\delta \in [0,\delta_0]$, let $\widetilde{W}_\alpha(\delta) := W_\alpha(\delta) + \bar{\kappa}_u/\eta$ denote shifted welfare under fidelity $f = 1-\delta$ and the follow-niche-signals rule. Then

$$\widetilde{W}_\alpha(\delta) = e^{-\eta J_\alpha(\delta)} A_\alpha(\delta), \qquad J_\alpha(\delta) := I_{\alpha,n}(1-\delta),$$

where

$$A_\alpha(\delta) := E_\alpha(\delta) + \frac{\bar{\kappa}_u}{\eta}$$

and

$$E_\alpha(\delta) = (1-\alpha)\big[(1-\delta)g_{\mathsf{P}} + \delta g_-\big] + \alpha\Big[(1-\delta)g(u_+(\alpha)) + \frac{\delta}{n}g_{\mathsf{P}} + \frac{(n-1)\delta}{n}g_-\Big].$$

Because $\alpha g(u_+(\alpha)) = L + o(\alpha)$, uniformly for $\delta \in [0,\delta_0]$,

$$A_\alpha(\delta) = A_0 - D_W\delta + O(\alpha), \qquad A_0 := g_{\mathsf{P}} + L + \frac{\bar{\kappa}_u}{\eta} > 0.$$

Therefore there exist constants $a_0, c_0 > 0$ and $\bar{\alpha}_1 \in (0,1)$ such that for all $\alpha < \bar{\alpha}_1$ and all $\delta \in [\alpha,\delta_0]$,

$$A_\alpha(\delta) \leq A_0\big(1 - a_0\delta + c_0\alpha\big), \qquad A_\alpha(\alpha) \geq A_0(1 - c_0\alpha). \tag{41}$$

After shrinking $\bar{\alpha}_1$ if necessary, we also have $A_\alpha(\delta) \geq A_0/2$ for all $\delta \in [\alpha,\delta_0]$ and all $\alpha < \bar{\alpha}_1$.

Next,

$$J_\alpha(\delta) = \alpha\ln(n/\delta) + O(\alpha) \qquad \text{uniformly for } \delta \in [\alpha,\delta_0], \tag{42}$$

hence

$$J_\alpha(\alpha) - J_\alpha(\delta) = \alpha\ln(\delta/\alpha) + O(\alpha) \qquad \text{uniformly for } \delta \in [\alpha,\delta_0]. \tag{43}$$

Since $\delta_\alpha \to 0$, the gap between welfare-optimal and full elicitation lies in $[0,\delta_0]$ for all sufficiently small α.

Let $\widetilde{W}_\alpha^{\text{cand}}$ denote shifted welfare for the feasible design $\delta = \alpha$. By (41),

$$\widetilde{W}_\alpha^{\text{cand}} \geq e^{-\eta J_\alpha(\alpha)} A_0(1 - c_0\alpha).$$

For any welfare-optimal gap $\delta \in [\alpha, \delta_0]$, (41) gives

$$\widetilde{W}_\alpha(\delta) \le e^{-\eta J_\alpha(\delta)} A_0(1 - a_0\delta + c_0\alpha).$$

Hence

$$\frac{\widetilde{W}_\alpha(\delta)}{\widetilde{W}_\alpha^{\text{cand}}} \le \frac{1 - a_0\delta + c_0\alpha}{1 - c_0\alpha} \exp\Big(\eta\big[J_\alpha(\alpha) - J_\alpha(\delta)\big]\Big).$$

Using (43), $\ln(1+y) \le y$, and $-\ln(1 - c_0\alpha) \le 2c_0\alpha$ for small α, we obtain

$$\ln \frac{\widetilde{W}_\alpha(\delta)}{\widetilde{W}_\alpha^{\text{cand}}} \le -a_0\delta + \eta\alpha \ln(\delta/\alpha) + C\alpha$$

for some constant $C < \infty$. Writing $x := \delta/\alpha \ge 1$, this becomes

$$\ln \frac{\widetilde{W}_\alpha(\delta)}{\widetilde{W}_\alpha^{\text{cand}}} \le \alpha\big[-a_0 x + \eta \ln x + C\big].$$

Because $-a_0 x + \eta \ln x \to -\infty$ as $x \to \infty$, there exists $K < \infty$ such that

$$-a_0 x + \eta \ln x + C < 0 \qquad \text{for all } x \ge K.$$

Therefore, for all sufficiently small α, no welfare-optimal gap $\delta \in [\alpha, \delta_0]$ can satisfy $\delta \ge K\alpha$. If $\delta_\alpha < \alpha$, then we simply have $\delta_\alpha \le K\alpha$ after enlarging K if necessary. Hence $\delta_\alpha = O(\alpha)$.

Step 3: identify the unique limit of δ_α/α. Set $c_\alpha := \delta_\alpha/\alpha$. By Step 2, $\{c_\alpha\}$ is bounded. Let $\alpha_k \downarrow 0$ and suppose $c_{\alpha_k} \to c \in [0, \infty)$. By Step 1, the optimal rule follows niche signals along this subsequence.

Using the exact expression for $E_\alpha(\delta)$ above with $\delta = c_{\alpha_k}\alpha_k$, we obtain

$$E_{\alpha_k}(\delta_{\alpha_k}) = g_{\mathsf{P}} + L + \alpha_k\big[-g_{\mathsf{P}}(1+c) + cg_- - cL\big] + o(\alpha_k). \tag{44}$$

We also need the near-perfect mutual-information expansion along bounded sequences. We use the following version of Lemma 5: for every $C < \infty$,

$$\sup_{0 \le c \le C} \left| \frac{1}{\alpha} I_{\alpha,n}(1 - c\alpha) - \left[\ln\Big(\frac{1}{\alpha}\Big) + 1 + \ln\Big(\frac{n\, c^c}{(1+c)^{1+c}}\Big)\right] \right| \longrightarrow 0 \qquad (\alpha \to 0), \tag{45}$$

where $0^0 := 1$. We can see that in the proof of Lemma 5, the expansions of P_0, P_1, $-P_0 \ln P_0$, $-nP_1 \ln P_1$, $f \ln f$, and $(1-f)\ln((1-f)/n)$ have remainders uniform on $c \in [0, C]$; for the logarithmic term, $P_1 = ((1+c)\alpha/n) + O(\alpha^2)$ uniformly on $[0, C]$, so

$$-nP_1 \ln P_1 = -(1+c)\alpha \ln \alpha - (1+c)\alpha \ln\Big(\frac{1+c}{n}\Big) + o(\alpha)$$

uniformly on $[0, C]$, and $c \ln c$ is continuous at 0 with value 0. Fix $C < \infty$ such that $c_{\alpha_k} \in [0, C]$ for all large k. Applying (45) with $c = c_{\alpha_k}$ gives

$$I_{\alpha_k,n}(1 - \delta_{\alpha_k}) = \alpha_k \left[\ln(1/\alpha_k) + 1 + \ln\Big(\frac{n\, c^c}{(1+c)^{1+c}}\Big)\right] + o(\alpha_k). \tag{46}$$

Here we also use that $c_{\alpha_k} \to c$ and the map $x \mapsto x \ln x - (1+x)\ln(1+x)$ is continuous on $[0, \infty)$.

Combining (44), (46), and

$$W = e^{-\eta I}\Big(E + \frac{\bar{\kappa}_u}{\eta}\Big) - \frac{\bar{\kappa}_u}{\eta},$$

we get

$$W_{\alpha_k}(\delta_{\alpha_k}) = C(\alpha_k) + \alpha_k \Psi_W(c) + o(\alpha_k), \tag{47}$$

where $C(\alpha)$ is independent of c.

For any fixed $d \geq 0$, the feasible design with gap $d\alpha$ also follows niche signals for all sufficiently small α, and the same calculation yields

$$W_\alpha(d\alpha) = C(\alpha) + \alpha \Psi_W(d) + o(\alpha),$$

with the same function $C(\alpha)$. Since δ_α is welfare-optimal,

$$W_{\alpha_k}(\delta_{\alpha_k}) \geq W_{\alpha_k}(d\alpha_k)$$

for every fixed $d \geq 0$ and all large k. Dividing by α_k and letting $k \to \infty$ shows that

$$\Psi_W(c) \geq \Psi_W(d) \qquad \text{for every } d \geq 0.$$

Thus every subsequential limit of $\{c_\alpha\}$ maximizes Ψ_W.

Finally,

$$\Psi'_W(c) = -D_W - K_W \ln\Big(\frac{c}{c+1}\Big), \qquad \Psi''_W(c) = -K_W\Big(\frac{1}{c} - \frac{1}{c+1}\Big) < 0 \quad (c > 0).$$

Also $\Psi'_W(c) \to +\infty$ as $c \downarrow 0$ and $\Psi'_W(c) \to -D_W < 0$ as $c \to \infty$, so Ψ_W has a unique maximizer $c^\star_W \in (0, \infty)$. Its first-order condition is

$$\ln\Big(1 + \frac{1}{c^\star_W}\Big) = \frac{D_W}{K_W},$$

equivalently

$$(c^\star_W)^{-1} = e^{D_W/K_W} - 1.$$

Since every subsequential limit of $\{c_\alpha\}$ equals $c^\star_W$, the full sequence converges:

$$\frac{\delta_\alpha}{\alpha} \to c^\star_W.$$

Therefore

$$1 - f^\star_W(\alpha) = \delta_\alpha = c^\star_W \, \alpha + o(\alpha),$$

as claimed. □

Proof of Proposition 1: We prove each part individually.

Write welfare (fixed prices) as

$$W(f, \pi) = sur^{\mathsf{MI}}(f) \cdot \mathbb{E}\big[g(u_{\mathsf{B}}(\Theta, \pi(S)))\big] - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur^{\mathsf{MI}}(f)\big). \tag{48}$$

Equivalently,

$$W(f, \pi) = sur^{\mathsf{MI}}(f)\Big(\mathbb{E}[g(u_{\mathsf{B}}(\Theta, \pi(S)))] + \frac{\bar{\kappa}_u}{\eta}\Big) - \frac{\bar{\kappa}_u}{\eta}. \tag{49}$$

Since the last term is constant, maximizing W is equivalent to maximizing the bracketed term multiplied by $sur^{\mathsf{MI}}(f)$.

Part (i): $\beta = 0$. Consider the candidate policy $\mathsf{ATR}(1)$ (perfect screening) and compare it to REC.

Under REC: $f = 1/(n+1)$ so $I_{\alpha,n}(f) = 0$ and $sur^{\mathsf{MI}}(f) = 1$; the platform recommends P always, hence

$$W(\mathsf{REC}) = g(u_{\mathsf{P}}).$$

Under ATR(1): $S = \Theta$ and $I_{\alpha,n}(1) = H(\Theta)$, so $sur^{\mathsf{MI}}(1) = \exp(-\eta H(\Theta))$. There are no false positives, and the expected (choice-stage) surplus equals $(1-\alpha)g(u_{\mathsf{P}}) + \alpha g(u_+)$. Thus

$$W(\mathsf{ATR}(1)) = sur^{\mathsf{MI}}(1)\big((1-\alpha)g(u_{\mathsf{P}}) + \alpha g(u_+)\big) - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur^{\mathsf{MI}}(1)\big).$$

Let $H(\Theta) = \alpha\ln(1/\alpha) + O(\alpha)$ by Lemma 3, so $1 - sur^{\mathsf{MI}}(1) = \eta H(\Theta) + o(H(\Theta))$. Using $\beta = 0$, u_+ is constant in α and hence $g(u_+)$ is bounded. A first-order expansion yields

$$W(\mathsf{ATR}(1)) - W(\mathsf{REC}) = \alpha\big(g(u_+) - g(u_{\mathsf{P}})\big) - \big(\bar{\kappa}_u + \eta g(u_{\mathsf{P}})\big)H(\Theta) + o\big(\alpha\ln(1/\alpha)\big).$$

The first term is $O(\alpha)$, while the second is $-\Theta(\alpha\ln(1/\alpha))$ with strictly negative coefficient. Hence $W(\mathsf{ATR}(1)) < W(\mathsf{REC})$ for all sufficiently small α. Since ATR(1) is the best possible elicitation policy in terms of avoiding false positives (it fully reveals type), no other elicitation policy can asymptotically overturn the $-\alpha\ln(1/\alpha)$ loss when niche gains are bounded. Formally, any policy that does not change recommendations relative to REC is weakly worse due to $sur^{\mathsf{MI}}(f) \le 1$ and the effort term, and any policy that recommends a niche with positive probability must (for small α) have $1 - f = O(\alpha)$ to avoid a constant mainstream mismatch loss, which by Lemma 6 implies an $O(\alpha\ln(1/\alpha))$ abandonment/effort penalty. Thus REC is welfare-optimal for α small.

Part (ii): $\beta > 0$**.** We again compare ATR(1) to REC. With $\beta > 0$, $u_+(\alpha) = \lambda V_{\mathsf{N}}\alpha^{-\beta} \to \infty$ and hence $g(u_+) = u_+ + o(1)$. Therefore

$$\alpha g(u_+(\alpha)) = \lambda V_{\mathsf{N}}\alpha^{1-\beta} + o(\alpha^{1-\beta}). \tag{50}$$

Meanwhile, $H(\Theta) = \alpha\ln(1/\alpha) + O(\alpha)$ and hence the total abandonment/effort penalty under $f = 1$ is $O(\alpha\ln(1/\alpha))$. Note that by (50), there is a cross term of order $O(\alpha^{2-\beta}\ln(1/\alpha))$ in the first-order expansion, which gives the slightly modified expression

$$W(\mathsf{ATR}(1)) - W(\mathsf{REC}) = \alpha\big(g(u_+) - g(u_{\mathsf{P}})\big) - \big(\bar{\kappa}_u + \eta g(u_{\mathsf{P}})\big)H(\Theta) + O\big(\alpha^{2-\beta}\ln(1/\alpha)\big). \tag{51}$$

The first term dominates the second because

$$\frac{\alpha^{1-\beta}}{\alpha\ln(1/\alpha)} = \frac{1}{\alpha^{\beta}\ln(1/\alpha)} \to \infty.$$

and dominates third term because

$$\frac{\alpha^{1-\beta}}{\alpha^{2-\beta}\ln(1/\alpha)} = \frac{1}{\alpha\ln(1/\alpha)} \to \infty$$

Hence for sufficiently small α, ATR(1) strictly dominates REC, so welfare strictly prefers elicitation.

To show $f^{\star}_W(\alpha) \to 1$, fix $\delta \in (0,1)$ and suppose $f \le 1 - \delta$. Then among mainstream users (mass $1-\alpha$), the probability of a niche signal is at least δ, and any policy that recommends niche on such signals delivers $g(u_-)$ instead of $g(u_{\mathsf{P}})$ on a $\Theta(1)$ fraction of sessions, causing a $\Theta(1)$ welfare loss. Alternatively, any policy that does *not* recommend niche on those signals is effectively REC but with $sur^{\mathsf{MI}}(f) < 1$ and a strictly positive

effort loss, hence is strictly worse than REC. Since ATR(1) dominates REC for small α and has $f=1$, no $f \le 1-\delta$ can be optimal for small α. Because δ is arbitrary, $f^\star_W(\alpha) \to 1$.

Part (iii): explicit scaling at $\beta=1$. Assume $\beta=1$ and restrict attention to the (asymptotically) optimal class where $f=1-c\alpha$ with $c>0$ constant and the signal rule follows niche signals. This restriction is without loss in the limit because Part (ii) implies $f^\star_W(\alpha)\to 1$, Lemma 7 justifies this scaling, and for f near 1 the welfare cutoff in Lemma 2 satisfies $q^\star_W(\alpha,\lambda)\to 0$ (since $g(u_+)\to\infty$), so acting on niche signals is optimal for all sufficiently small α.

Let $L := \lambda V_{\mathsf{N}}$. Following niche signals, the expected (choice-stage) surplus satisfies, uniformly over fixed c,

$$\mathbb{E}[g(u_{\mathsf{B}}(\Theta,\pi(S)))] = g(u_{\mathsf{P}}) + L + \alpha\Big[-g(u_{\mathsf{P}})(1+c) + cg(u_-) - cL\Big] + o(\alpha), \tag{52}$$

where we used $g(u_+(\alpha)) = u_+(\alpha)+o(1) = L/\alpha + o(1/\alpha)$ and $f=1-c\alpha$. Also, by Lemma 5,

$$I_{\alpha,n}(1-c\alpha) = \alpha\Big[\ln(1/\alpha) + B(c)\Big] + o(\alpha), \qquad B(c) := 1 + \ln\Big(\frac{n\,c^c}{(c+1)^{c+1}}\Big). \tag{53}$$

Hence $sur^{\mathsf{MI}}(f) = 1 - \eta\alpha\ln(1/\alpha) - \eta\alpha B(c) + o(\alpha)$.

Plugging (52) and (53) into (49), and keeping terms up to order α, yields

$$W(1-c\alpha,\pi) = \text{const} + \alpha\Big(\underbrace{c(g(u_-)-g(u_{\mathsf{P}})-L)}_{\text{misrouting vs. niche gain}} - \underbrace{\big(\bar\kappa_u + \eta(g(u_{\mathsf{P}})+L)\big)B(c)}_{\text{effort/abandonment}}\Big) - \eta(g(u_{\mathsf{P}})+L)\alpha\ln(1/\alpha) + o(\alpha).$$

The $\alpha\ln(1/\alpha)$ term is independent of c, so the optimal c maximizes

$$\Psi_W(c) := c(g(u_-)-g(u_{\mathsf{P}})-L) - K_W\,B(c), \qquad K_W := \bar\kappa_u + \eta(g(u_{\mathsf{P}})+L).$$

Differentiating gives

$$\Psi'_W(c) = (g(u_-)-g(u_{\mathsf{P}})-L) - K_W\frac{d}{dc}B(c) = (g(u_-)-g(u_{\mathsf{P}})-L) - K_W\ln\Big(\frac{c}{c+1}\Big),$$

since $B'(c) = \ln(c) - \ln(c+1) = \ln(c/(c+1))$. Setting $\Psi'_W(c)=0$ yields

$$\ln\Big(1+\frac{1}{c}\Big) = \frac{g(u_{\mathsf{P}})+L-g(u_-)}{K_W} = \frac{D_W}{K_W},$$

hence $c=c^\star_W$. Uniqueness follows because Ψ_W is strictly concave in c: $B''(c) = 1/c - 1/(c+1) > 0$ implies $-K_W B(c)$ is strictly concave and the linear term is linear. Therefore, for any welfare-optimal sequence $f^\star_W(\alpha)$ we must have $1-f^\star_W(\alpha) = c^\star_W\alpha + o(\alpha)$. □

A.5. Proof of Corollary 1

Proof of Corollary 1: The rate follows directly from (50) and (51), as the welfare of ATR(1) is a lower bound on the welfare-optimal elicitation. □

A.6. Proof of Theorem 2

Seller-side monopoly pricing. Because each user sees only the single recommended item versus the outside option, seller demands are independent across sellers. We therefore analyze each seller's monopoly problem separately.

LEMMA 8 **(Logit monopoly price and revenue for a single utility level)**. *Fix $u \in \mathbb{R}$. Consider a monopolist facing demand $D(p) = \sigma(u-p)$ and revenue $R(p) = pD(p)$ over prices $p \geq 0$. Then $R(\cdot)$ is strictly concave and has a unique maximizer*

$$p_m(u) = 1 + W\big(e^{u-1}\big), \tag{54}$$

where $W(\cdot)$ is the Lambert-W function. The corresponding monopoly revenue equals

$$r_m(u) := \max_{p \geq 0} p\sigma(u-p) = p_m(u)\sigma(u-p_m(u)) = W\big(e^{u-1}\big) = p_m(u) - 1. \tag{55}$$

Moreover, the logit inclusive value at the monopoly price satisfies

$$g(u - p_m(u)) = \ln p_m(u). \tag{56}$$

Finally, as $u \to \infty$, $p_m(u) \to \infty$ and $r_m(u)/u \to 1$.

Proof of Lemma 8: Differentiate $R(p) = p\sigma(u-p)$:

$$R'(p) = \sigma(u-p) - p\sigma'(u-p), \qquad \sigma'(x) = \sigma(x)\big(1-\sigma(x)\big).$$

Set $z = u - p$ and solve $R'(p) = 0$:

$$0 = \sigma(z) - p\sigma(z)(1-\sigma(z)) \quad \Longleftrightarrow \quad 1 = p(1-\sigma(z)).$$

Since $1 - \sigma(z) = 1/(1+e^z)$, the first-order condition is

$$p = 1 + e^{u-p}.$$

Let $y := p - 1$. Then $y = e^{u-1}e^{-y}$, hence $ye^y = e^{u-1}$ and $y = W(e^{u-1})$. This yields (54) and (55). Then $e^{u-p_m(u)} = y = p_m(u) - 1$ implies $g(u - p_m(u)) = \ln(1 + e^{u-p_m(u)}) = \ln(1+y) = \ln p_m(u)$, proving (56). Strict concavity follows because $R''(p) < 0$ for all p under logit demand (a direct calculation). The asymptotics use $W(x) \sim \ln x$ as $x \to \infty$. □

Proof of Theorem 2: The proof consists of 5 steps: 1) we write the commission revenue under REC, 2) we derive a lower bound on the commission revenue under perfect screening ATR(1), 3) compare this to REC for $\beta = 0, \beta > 0$, 4) show that $f^\star_\Pi(\alpha) \to 1$ when $\beta > 0$, and 5) show that $1 - f = \Theta(\alpha)$.

Fix a CRS design (f, π). Because users only see the recommended item versus the outside option, each seller j chooses p_j to maximize $(1-\tau)p_j D_j$, where D_j depends on p_j but not on p_{-j}. Thus sellers face independent monopoly problems and equilibrium prices are well-defined by pointwise monopoly pricing.

Step 1: Commission revenue under REC. Under REC, the platform does not elicit ($I_{\alpha,n} = 0$) so $sur^{\mathsf{MI}} = 1$ and it recommends P always. Seller P faces utility u_{P} for every consumer and therefore sets $p_{\mathsf{P}} = p_m(u_{\mathsf{P}})$. Platform commission revenue is

$$\Pi^{\mathsf{com}}(\mathsf{REC}) = \tau\, r_m(u_{\mathsf{P}}) = \tau r_{\mathsf{P}}. \tag{57}$$

Step 2: A lower bound on the platform's commission revenue under perfect screening ATR(1). Under ATR(1), $S = \Theta$ and $sur^{\mathsf{MI}}(1) = \exp(-\eta H(\Theta))$ with $H(\Theta)$ given by Lemma 3. Mainstream users (mass $1-\alpha$) are recommended P; niche users (mass α) are recommended their correct niche. Seller P again sets $p_m(u_{\mathsf{P}})$. Each niche seller faces only its own matched niche type with utility $u_+(\alpha)$ and therefore sets $p_m(u_+(\alpha))$. Hence platform commission revenue under ATR(1) equals

$$\Pi^{\mathsf{com}}(\mathsf{ATR}(1)) = \tau\, sur^{\mathsf{MI}}(1)\Big((1-\alpha)r_{\mathsf{P}} + \alpha r_m(u_+(\alpha))\Big). \tag{58}$$

Step 3: Compare $\mathsf{ATR}(1)$ *to* REC. Using (57)–(58),

$$\Pi^{\mathsf{com}}(\mathsf{ATR}(1)) - \Pi^{\mathsf{com}}(\mathsf{REC}) = \tau\Big(sur^{\mathsf{MI}}(1)\big((1-\alpha)r_{\mathsf{P}} + \alpha r_m(u_+(\alpha))\big) - r_{\mathsf{P}}\Big).$$

Write $sur^{\mathsf{MI}}(1) = 1 - \eta H(\Theta) + o(H(\Theta))$ with $H(\Theta) = \alpha\ln(1/\alpha) + O(\alpha)$. Then

$$\Pi^{\mathsf{com}}(\mathsf{ATR}(1)) - \Pi^{\mathsf{com}}(\mathsf{REC}) = \tau\Big(-\alpha r_{\mathsf{P}} + \alpha r_m(u_+(\alpha)) - \eta r_{\mathsf{P}} H(\Theta)\Big) + o\big(\alpha\ln(1/\alpha)\big), \tag{59}$$

since the term $\eta H(\Theta)\cdot\alpha r_m(u_+(\alpha))$ is of lower order (it is $O(\alpha^{2-\beta}\ln(1/\alpha)\cdot r_m(u_+))$, which will be dominated even if $\beta > 1$).

(i) Case $\beta = 0$**.** Then $u_+(\alpha)$ is constant in α and so is $r_m(u_+(\alpha))$. Thus the first two terms in (59) are $O(\alpha)$, while the third term is $-\Theta(\alpha\ln(1/\alpha))$ with strictly negative coefficient. Hence the difference is negative for sufficiently small α, so $\mathsf{ATR}(1)$ is worse than REC. Since any elicitation policy has $sur^{\mathsf{MI}}(f) \le 1$ and can at best increase the niche term by $O(\alpha)$ when u_+ is bounded, REC is optimal for α small.

(ii) Case $\beta > 0$**.** Then $u_+(\alpha) \to \infty$ and Lemma 8 implies $r_m(u_+(\alpha))/u_+(\alpha) \to 1$. Hence $\alpha r_m(u_+(\alpha)) = \lambda V_{\mathsf{N}}\alpha^{1-\beta} + o(\alpha^{1-\beta})$. This dominates $\alpha\ln(1/\alpha)$ for every $\beta > 0$:

$$\frac{\alpha^{1-\beta}}{\alpha\ln(1/\alpha)} = \frac{1}{\alpha^{\beta}\ln(1/\alpha)} \to \infty.$$

Therefore (59) is strictly positive for all sufficiently small α. Thus the platform strictly prefers elicitation to REC for small α.

Step 4: Show $f^\star_\Pi(\alpha) \to 1$ *when* $\beta > 0$. Fix $\delta \in (0,1)$. Consider any $f \le 1-\delta$. If the platform ignores niche signals, it is weakly dominated by REC because it has $sur^{\mathsf{MI}}(f) < 1$ but identical routing to P. If instead it recommends niches with positive probability at such f, then mainstream users (mass $1-\alpha$) are routed away from P on a set of probability at least δ, and since P yields strictly higher utility for mainstream users than any niche ($u_{\mathsf{P}} > u_-$), this strictly reduces purchase probabilities and hence commission revenue by a constant amount (independent of α) relative to a design with f closer to 1 that eliminates false positives. Because under MI the abandonment penalty of $f = 1$ shrinks to zero as $\alpha \to 0$, Step 3 implies that for α small, a design with $f = 1$ strictly dominates any $f \le 1-\delta$. Therefore, no $f \le 1-\delta$ can be optimal for small α. Since δ is arbitrary, $f^\star_\Pi(\alpha) \to 1$.

Step 5: Knife-edge $\beta = 1$ *and the* $1 - f = \Theta(\alpha)$ *expansion.* Assume **(iii)** $\beta = 1$ and write $u_+(\alpha) = L/\alpha$. By Step 4, $f^\star_\Pi(\alpha) \to 1$. We now refine the rate.

First, $1 - f^\star_\Pi(\alpha)$ cannot be $\gg \alpha$. If $1-f$ is of order strictly larger than α, then a non-negligible (relative to α) mass of mainstream users is routed to niches under any policy that uses niche signals, which destroys a strictly positive constant fraction of P-sales, reducing commission revenue by $\Theta(1-f)$. But the total incremental niche revenue under $\beta = 1$ is at most of constant order (because $\alpha u_+ = L$), so for sufficiently small α such a policy is dominated by a policy with $1 - f = O(\alpha)$. Hence $1 - f^\star_\Pi(\alpha) = O(\alpha)$.

Thus there exists a sequence c_α bounded above such that $f^\star_\Pi(\alpha) = 1 - c_\alpha\alpha + o(\alpha)$. We show $c_\alpha \to c^\star_\Pi$.

Consider the follow-niche-signals rule at $f = 1 - c\alpha$ and note that, when $u_+ = L/\alpha \to \infty$, the niche seller's monopoly revenue from correctly matched niche users is $r_m(u_+(\alpha)) = u_+(\alpha) + o(u_+(\alpha))$ (Lemma 8), so the aggregate niche revenue term is

$$\alpha f\, r_m(u_+(\alpha)) = fL + o(1).$$

Meanwhile, the loss of P-sales from false positives is of order $(1-f)r_{\mathsf{P}} = c\alpha r_{\mathsf{P}} + o(\alpha)$, and the survival term satisfies Lemma 5. Collecting terms yields the asymptotic expansion

$$\Pi^{\mathsf{com}}(1-c\alpha,\pi) = \tau\Big[(r_{\mathsf{P}}+L) - \eta(r_{\mathsf{P}}+L)\alpha\ln(1/\alpha) + \alpha\Big(-r_{\mathsf{P}} - c(r_{\mathsf{P}}+L) - \eta(r_{\mathsf{P}}+L)B(c)\Big) + o(\alpha)\Big],$$

where $B(c)$ is defined in (53). Since the $\alpha\ln(1/\alpha)$ term is independent of c, the asymptotically optimal c maximizes

$$\Psi_{\Pi}(c) := -c(r_{\mathsf{P}}+L) - \eta(r_{\mathsf{P}}+L)B(c).$$

Differentiating and using $B'(c) = \ln(c/(c+1))$ gives

$$\Psi'_{\Pi}(c) = -(r_{\mathsf{P}}+L) - \eta(r_{\mathsf{P}}+L)\ln\Big(\frac{c}{c+1}\Big).$$

Setting $\Psi'_{\Pi}(c) = 0$ yields $\eta\ln((c+1)/c) = 1$, hence $c = c^{\star}_{\Pi} = 1/(e^{1/\eta}-1)$. Strict concavity of $-B(c)$ implies uniqueness. Therefore $c_\alpha \to c^{\star}_{\Pi}$ and the result follows. □

A.7. Proof of Proposition 2

LEMMA 9 **(Rate of the commission-optimal fidelity)**. *Fix $\beta > 0$ and $\lambda > 0$, and assume the conditions of Theorem 2(ii). Let*

$$\delta_\alpha := 1 - f^{\star}_{\Pi}(\alpha).$$

Then there exist constants $K < \infty$ and $\bar{\alpha} \in (0,1)$ such that

$$0 \le \delta_\alpha \le K\alpha \qquad \textit{for all } \alpha \in (0,\bar{\alpha}).$$

Equivalently,

$$1 - f^{\star}_{\Pi}(\alpha) = O(\alpha).$$

Proof of Lemma 9: Fix $\beta > 0$ and $\lambda > 0$, and write $\delta := 1 - f \in [0,1)$. Let

$$r_{\mathsf{P}} := r_m(u_{\mathsf{P}}), \qquad r_- := r_m(u_-), \qquad r_+(\alpha) := r_m(u_+(\alpha)).$$

Because $\beta > 0$ and $\lambda > 0$, we have $u_+(\alpha) = \lambda V_{\mathsf{N}}\alpha^{-\beta} \to \infty$, so Lemma 8 implies that for all sufficiently small α,

$$r_+(\alpha) \ge r_{\mathsf{P}} > r_- > 0.$$

By Theorem 2(ii), for all sufficiently small α: (i) the platform strictly prefers elicitation to REC, and (ii) every commission-optimal fidelity satisfies $f^{\star}_{\Pi}(\alpha) \to 1$. Hence, for all small α, a commission-optimal design cannot recommend P after every signal: if it did, its commission payoff would be

$$\tau e^{-\eta I_{\alpha,n}(f)} r_{\mathsf{P}} < \tau r_{\mathsf{P}},$$

which is strictly below the payoff of REC whenever $f > 1/(n+1)$. By the posterior-threshold characterization of the optimal recommendation rule in the commission problem, the platform's action on niche signals is therefore all-or-nothing, and any commission-optimal design for small α must follow every niche signal:

$$\pi(\emptyset) = \mathsf{P}, \qquad \pi(i) = i, \quad i = 1,\dots,n.$$

Fix such an α, and consider the follow-niche-signals rule at fidelity $f = 1-\delta$. Define

$$a_\alpha(\delta) := \alpha(1-\delta), \qquad b_\alpha(\delta) := \Big(1-\frac{\alpha}{n}\Big)\delta, \qquad \widetilde{\alpha}_\delta := a_\alpha(\delta) + b_\alpha(\delta).$$

Here $a_\alpha(\delta)$ is the mass of correctly routed niche users, $b_\alpha(\delta)$ is the mass of false niche routes, and $\widetilde{\alpha}_\delta$ is the total mass of niche recommendations.

Under this rule, the popular seller receives mass $1-\widetilde{\alpha}_\delta$ and earns per-session revenue r_{P}. Each niche seller faces the same mixed pool, so total pre-survival seller revenue can be written as

$$\widehat{R}_\alpha(\delta) = (1-\widetilde{\alpha}_\delta)\, r_{\mathsf{P}} + \max_{p\ge 0}\Big\{ a_\alpha(\delta)\, p\, \sigma(u_+(\alpha)-p) + b_\alpha(\delta)\, p\, \sigma(u_- - p)\Big\}.$$

Hence the platform payoff is

$$\Pi_\alpha(\delta) = \tau e^{-\eta J_\alpha(\delta)} \widehat{R}_\alpha(\delta), \qquad J_\alpha(\delta) := I_{\alpha,n}(1-\delta).$$

Define also

$$R^0_\alpha := r_{\mathsf{P}} + \alpha\big(r_+(\alpha) - r_{\mathsf{P}}\big).$$

Step 1: Linear upper and lower bounds for pre-survival revenue.

Since $r_+(\alpha)$ is the maximum of $p \mapsto p\sigma(u_+(\alpha)-p)$ and r_- is the maximum of $p \mapsto p\sigma(u_- - p)$, we have

$$\begin{aligned}\widehat{R}_\alpha(\delta) &\le (1-\widetilde{\alpha}_\delta) r_{\mathsf{P}} + a_\alpha(\delta) r_+(\alpha) + b_\alpha(\delta) r_- \\ &= R^0_\alpha - \Big[\alpha\big(r_+(\alpha) - r_{\mathsf{P}}\big) + \Big(1-\frac{\alpha}{n}\Big)(r_{\mathsf{P}} - r_-)\Big]\delta. \end{aligned} \tag{60}$$

Choose $\bar{\alpha}_1 > 0$ such that $1-\alpha/n \ge 1/2$ for all $\alpha < \bar{\alpha}_1$. Then, for such α,

$$\alpha\big(r_+(\alpha) - r_{\mathsf{P}}\big) + \Big(1-\frac{\alpha}{n}\Big)(r_{\mathsf{P}} - r_-) \ge \alpha\big(r_+(\alpha) - r_{\mathsf{P}}\big) + \frac{r_{\mathsf{P}} - r_-}{2}.$$

Set

$$a_0 := \frac{r_{\mathsf{P}} - r_-}{2 r_{\mathsf{P}}} > 0.$$

Since $R^0_\alpha = r_{\mathsf{P}} + \alpha(r_+(\alpha) - r_{\mathsf{P}})$, we obtain

$$\alpha\big(r_+(\alpha) - r_{\mathsf{P}}\big) + \frac{r_{\mathsf{P}} - r_-}{2} \ge a_0 R^0_\alpha.$$

Thus (60) implies

$$\widehat{R}_\alpha(\delta) \le R^0_\alpha(1 - a_0\delta) \qquad \text{for all sufficiently small } \alpha. \tag{61}$$

For a lower bound, ignore the revenue generated by false niche routes:

$$\begin{aligned}\widehat{R}_\alpha(\delta) &\ge (1-\widetilde{\alpha}_\delta) r_{\mathsf{P}} + a_\alpha(\delta) r_+(\alpha) \\ &= R^0_\alpha - \Big[\alpha\big(r_+(\alpha) - r_{\mathsf{P}}\big) + \Big(1-\frac{\alpha}{n}\Big) r_{\mathsf{P}}\Big]\delta. \end{aligned} \tag{62}$$

Evaluating at $\delta = \alpha$ and noting that the bracket is at most R^0_α, we obtain

$$\widehat{R}_\alpha(\alpha) \ge R^0_\alpha(1-\alpha). \tag{63}$$

Step 2: Mutual-information comparison.

Let

$$H_n(x) := h(x) + x \ln n.$$

Under fidelity $f = 1 - \delta$, the total probability of a niche signal is

$$\widetilde{\alpha}_\delta = \alpha(1-\delta) + \Big(1 - \frac{\alpha}{n}\Big)\delta = \delta + \ell_\alpha(\delta),$$

where

$$\ell_\alpha(\delta) := \alpha\Big(1 - \Big(1 + \frac{1}{n}\Big)\delta\Big).$$

Hence

$$J_\alpha(\delta) = H_n(\widetilde{\alpha}_\delta) - H_n(\delta).$$

Choose

$$\delta_0 := \min\Big\{\frac{1}{4}, \frac{n}{2(n+1)}\Big\}.$$

Since $\delta_\alpha \to 0$ by Theorem 2(ii), there exists $\bar{\alpha}_2 > 0$ such that every commission-optimal δ_α lies in $[0, \delta_0]$ whenever $\alpha < \bar{\alpha}_2$. For $\delta \in [\alpha, \delta_0]$, we have

$$\frac{\alpha}{2} \le \ell_\alpha(\delta) \le \alpha,$$

because $1 - (1 + 1/n)\delta \in [1/2, 1]$ on this interval.

Now apply Taylor's theorem to $H_n(\delta + \ell_\alpha(\delta))$ around δ. There exists $\xi_{\alpha,\delta} \in (\delta, \delta + \ell_\alpha(\delta))$ such that

$$J_\alpha(\delta) = \ell_\alpha(\delta) H_n'(\delta) + \frac{1}{2}\ell_\alpha(\delta)^2 H_n''(\xi_{\alpha,\delta}),$$

where

$$H_n'(x) = \ln\Big(\frac{n(1-x)}{x}\Big), \qquad H_n''(x) = -\frac{1}{x(1-x)}.$$

Since $\delta \in [\alpha, \delta_0] \subset (0, 1/2)$ and $\xi_{\alpha,\delta} \ge \delta$, we have

$$|H_n''(\xi_{\alpha,\delta})| \le \frac{2}{\delta}.$$

Therefore

$$\frac{1}{2}\ell_\alpha(\delta)^2 H_n''(\xi_{\alpha,\delta}) = O\left(\frac{\alpha^2}{\delta}\right) = O(\alpha) \qquad \text{uniformly for } \delta \in [\alpha, \delta_0].$$

Also,

$$H_n'(\delta) = \ln(n/\delta) + \ln(1-\delta) = \ln(n/\delta) + O(\delta),$$

and

$$\ell_\alpha(\delta) = \alpha + O(\alpha\delta) \qquad \text{uniformly on } [\alpha, \delta_0].$$

Since $\delta \ln(n/\delta)$ is bounded on $(0, \delta_0]$, it follows that

$$\ell_\alpha(\delta) H_n'(\delta) = \alpha \ln(n/\delta) + O(\alpha) \qquad \text{uniformly for } \delta \in [\alpha, \delta_0].$$

Combining the last two displays yields

$$J_\alpha(\delta) = \alpha \ln(n/\delta) + O(\alpha) \qquad \text{uniformly for } \delta \in [\alpha, \delta_0]. \tag{64}$$

In particular,

$$J_\alpha(\alpha) - J_\alpha(\delta) = \alpha \ln(\delta/\alpha) + O(\alpha) \qquad \text{uniformly for } \delta \in [\alpha, \delta_0]. \tag{65}$$

Step 3: Compare an arbitrary large gap to the feasible design $\delta = \alpha$.

Let Π_α^{cand} denote the payoff from the feasible design with gap $\delta = \alpha$ and the follow-niche-signals rule. By (63),

$$\Pi_\alpha^{\text{cand}} \geq \tau e^{-\eta J_\alpha(\alpha)} R_\alpha^0 (1-\alpha).$$

Now let $\delta \in [\alpha, \delta_0]$ be the gap used by some commission-optimal design. As shown at the beginning of the proof, any commission-optimal design for small α must follow every niche signal, so (61) applies:

$$\Pi_\alpha(\delta) \leq \tau e^{-\eta J_\alpha(\delta)} R_\alpha^0 (1 - a_0\delta).$$

Hence

$$\frac{\Pi_\alpha(\delta)}{\Pi_\alpha^{\text{cand}}} \leq \frac{1-a_0\delta}{1-\alpha} \exp\Big(\eta\big[J_\alpha(\alpha) - J_\alpha(\delta)\big]\Big).$$

Using (65), there exists $C < \infty$ such that

$$\frac{\Pi_\alpha(\delta)}{\Pi_\alpha^{\text{cand}}} \leq \frac{1-a_0\delta}{1-\alpha} \exp\Big(\eta\alpha \ln(\delta/\alpha) + C\alpha\Big).$$

Taking logs and using $\ln(1-x) \leq -x$ and $-\ln(1-\alpha) \leq 2\alpha$ for small α,

$$\ln \frac{\Pi_\alpha(\delta)}{\Pi_\alpha^{\text{cand}}} \leq -a_0\delta + \eta\alpha \ln(\delta/\alpha) + (C+2)\alpha.$$

Writing $x := \delta/\alpha \geq 1$, this becomes

$$\ln \frac{\Pi_\alpha(\delta)}{\Pi_\alpha^{\text{cand}}} \leq \alpha\Big[-a_0 x + \eta \ln x + (C+2)\Big].$$

Because $-a_0 x + \eta \ln x \to -\infty$ as $x \to \infty$, there exists $K < \infty$ such that

$$-a_0 x + \eta \ln x + (C+2) < 0 \qquad \text{for all } x \geq K.$$

Therefore, whenever $\delta \geq K\alpha$,

$$\Pi_\alpha(\delta) < \Pi_\alpha^{\text{cand}},$$

so such a δ cannot be commission-optimal.

Finally, if a commission-optimal δ_α satisfies $\delta_\alpha < \alpha$, then trivially $\delta_\alpha \leq K\alpha$ after enlarging K if necessary. Hence, for all sufficiently small α,

$$\delta_\alpha \leq K\alpha.$$

This proves the lemma. □

LEMMA 10 **(Mixed logit monopoly with a nonvanishing high-type share)**. *Fix $u_- \in \mathbb{R}$ and $q_0 \in (0,1]$. For $u > 0$ and $q \in [q_0, 1]$, define $R_{u,q}(p) := p\Big[q\,\sigma(u-p) + (1-q)\,\sigma(u_- - p)\Big]$, $p \geq 0$. Let $p^\star(u,q)$ denote the unique maximizer of $R_{u,q}(\cdot)$. Then there exist constants $M < \infty$ and $u_0 < \infty$ such that for all $u \geq u_0$ and all $q \in [q_0, 1]$,*

$$|p^\star(u,q) - p_m(u)| \leq M.$$

Consequently, uniformly over $q \in [q_0, 1]$ as $u \to \infty$, $p^\star(u,q) = u - \ln u + O(1)$, $g(u - p^\star(u,q)) = \ln u + O(1)$, and $g(u_- - p^\star(u,q)) = o(1)$.

Proof of Lemma 10: Write $H_u(p) := p\,\sigma(u-p)$, $L(p) := p\,\sigma(u_- - p)$, so that $R_{u,q}(p) = qH_u(p) + (1-q)L(p)$. By Lemma 8, both H_u and L are strictly concave on $[0,\infty)$. Hence $R_{u,q}$ is strictly concave on $[0,\infty)$ for every $q \in [q_0, 1]$. Since $R_{u,q}(0) = 0$ and $R_{u,q}(p) \to 0$ as $p \to \infty$, the maximizer $p^\star(u,q)$ exists and is unique.

Let $p_m(u)$ be the pure high-type monopoly price from Lemma 8, and define

$$a_u := p_m(u) - 1 = e^{u - p_m(u)}.$$

Since $p_m(u) \to \infty$ by Lemma 8, we have $a_u \to \infty$ as $u \to \infty$.

Fix $y \in \mathbb{R}$ and set $p = p_m(u) + y$. Differentiating H_u gives

$$H_u'(p) = \sigma(u-p)\Big[1 - p\big(1 - \sigma(u-p)\big)\Big].$$

Using $a_u = e^{u-p_m(u)}$ and $p_m(u) = a_u + 1$, a direct substitution yields

$$H_u'(p_m(u)+y) = \frac{a_u e^{-y}\big[a_u(e^{-y}-1) - y\big]}{(1 + a_u e^{-y})^2}.$$

Hence, for every fixed compact set of y's,

$$H_u'(p_m(u)+y) \longrightarrow 1 - e^y \qquad \text{uniformly as } u \to \infty.$$

Choose any $M > 0$. Then there exists $u_1 < \infty$ and $c_M > 0$ such that for all $u \geq u_1$,

$$H_u'(p_m(u) - M) \geq 2c_M, \qquad H_u'(p_m(u) + M) \leq -2c_M. \tag{66}$$

Next, differentiate L:

$$L'(p) = \sigma(u_- - p) - p\sigma'(u_- - p).$$

Since $\sigma'(x) = \sigma(x)(1-\sigma(x)) \leq \sigma(x)$, we have

$$|L'(p)| \leq (1+p)\sigma(u_- - p) \leq (1+p)e^{u_- - p}.$$

Because $p_m(u) \to \infty$, it follows that

$$\sup_{p \in [p_m(u) - M,\, p_m(u) + M]} |L'(p)| \to 0 \qquad \text{as } u \to \infty.$$

Hence there exists $u_2 < \infty$ such that for all $u \geq u_2$,

$$\sup_{p \in [p_m(u) - M,\, p_m(u) + M]} |L'(p)| \leq q_0 c_M. \tag{67}$$

Let $u_0 := \max\{u_1, u_2\}$. For $u \geq u_0$ and $q \in [q_0, 1]$, combine (66) and (67):

$$R_{u,q}'(p_m(u) - M) = qH_u'(p_m(u) - M) + (1-q)L'(p_m(u) - M) \geq q_0(2c_M) - q_0 c_M > 0,$$

and similarly

$$R_{u,q}'(p_m(u) + M) = qH_u'(p_m(u) + M) + (1-q)L'(p_m(u) + M) \leq -q_0(2c_M) + q_0 c_M < 0.$$

Since $R_{u,q}$ is strictly concave, its unique maximizer must lie between these two points:

$$|p^\star(u,q) - p_m(u)| \leq M \qquad \text{for all } u \geq u_0,\ q \in [q_0, 1].$$

This proves the uniform $O(1)$ bracketing around $p_m(u)$. Now use Lemma 8. That lemma gives $r_m(u) = p_m(u) - 1$ and $r_m(u)/u \to 1$, so $p_m(u) = u + o(u)$. Also the first-order condition for $p_m(u)$ implies

$$u - p_m(u) = \ln(p_m(u) - 1) = \ln u + O(1).$$

Therefore

$$p_m(u) = u - \ln u + O(1),$$

and the same holds for $p^\star(u,q)$ because $|p^\star(u,q) - p_m(u)| \le M$ uniformly:

$$p^\star(u,q) = u - \ln u + O(1).$$

It follows that

$$u - p^\star(u,q) = \ln u + O(1),$$

uniformly over $q \in [q_0, 1]$. Since $u - p^\star(u,q) \to \infty$,

$$g(u - p^\star(u,q)) = \ln\bigl(1 + e^{u - p^\star(u,q)}\bigr) = u - p^\star(u,q) + o(1) = \ln u + O(1),$$

uniformly in q.

Finally, because $p^\star(u,q) \to \infty$ uniformly in q,

$$u_- - p^\star(u,q) \to -\infty \qquad \text{uniformly in } q,$$

so

$$g(u_- - p^\star(u,q)) = \ln\bigl(1 + e^{u_- - p^\star(u,q)}\bigr) \le e^{u_- - p^\star(u,q)} \to 0$$

uniformly over $q \in [q_0, 1]$. This proves the lemma. □

Proof of Proposition 2: Part (i) follows directly from Theorem 2(i): if $\beta = 0$, then REC is commission-optimal for all sufficiently small α, so $\Delta W_\Pi^{\mathsf{com}}(\alpha) = 0$ for all such α.

We now prove part (ii), fixing $\beta > 0$ and $\lambda > 0$.

Let

$$\delta_\alpha := 1 - f_\Pi^\star(\alpha).$$

By Lemma 9,

$$\delta_\alpha = O(\alpha).$$

By the posterior-threshold characterization of the commission-optimal recommendation rule, the platform's action on niche signals is common across niche labels. Since Theorem 2(ii) implies that, for all sufficiently small α, the commission-optimal design strictly outperforms REC, it cannot recommend P after every signal. Therefore, for all sufficiently small α,

$$\pi_\Pi^\star(\emptyset) = \mathsf{P}, \qquad \pi_\Pi^\star(i) = i, \quad i = 1, \dots, n.$$

Define

$$a_\alpha := \alpha(1 - \delta_\alpha), \qquad b_\alpha := \Bigl(1 - \frac{\alpha}{n}\Bigr)\delta_\alpha, \qquad \widetilde{\alpha}_\alpha := a_\alpha + b_\alpha.$$

Then a_α is the mass of correctly routed niche users, b_α is the mass of false niche routes, and $\widetilde{\alpha}_\alpha$ is the total mass of niche recommendations. Because $\delta_\alpha = O(\alpha)$,

$$a_\alpha = \alpha + O(\alpha^2), \qquad b_\alpha = O(\alpha), \qquad \widetilde{\alpha}_\alpha = O(\alpha). \tag{68}$$

Let $p^\star_{\mathsf{P}} := p_m(u_{\mathsf{P}})$. The popular seller always faces homogeneous utility u_{P} whenever it is recommended, so its equilibrium price is

$$p^\star_{\mathsf{P}} = p_m(u_{\mathsf{P}}),$$

and Lemma 8 gives

$$W^{\mathsf{com}}(\mathsf{REC}) = g(u_{\mathsf{P}} - p^\star_{\mathsf{P}}).$$

Under the commission-optimal design, symmetry implies that every niche seller faces the same mixed demand problem, so let $p^\star_{\mathsf{N},\alpha}$ denote the common niche price.

Step 1: The posterior correctness of a niche signal is bounded away from zero.

Conditional on a niche signal, the posterior correctness is

$$q_\alpha = \frac{a_\alpha}{a_\alpha + b_\alpha} = \frac{\alpha(1-\delta_\alpha)}{\alpha(1-\delta_\alpha) + (1-\alpha/n)\delta_\alpha}.$$

Since $\delta_\alpha \le K\alpha$ for some finite K and all sufficiently small α, there exists $\underline{q} > 0$ such that

$$q_\alpha \ge \underline{q} \qquad \text{for all sufficiently small } \alpha.$$

Therefore Lemma 10 applies to the niche seller's pricing problem with

$$u = u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{-\beta}, \qquad q = q_\alpha.$$

Because $\beta > 0$ and $\lambda > 0$, we have $u_+(\alpha) \to \infty$, so

$$g\big(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}\big) = \ln u_+(\alpha) + O(1) = \beta \ln(1/\alpha) + O(1), \tag{69}$$

and

$$g\big(u_- - p^\star_{\mathsf{N},\alpha}\big) = o(1). \tag{70}$$

Step 2: Choice-stage expected surplus under the commission-optimal design.

Let

$$g^{\mathrm{com}}_{\mathsf{P}} := g(u_{\mathsf{P}} - p^\star_{\mathsf{P}}).$$

Conditional on reaching the recommendation stage, expected surplus under the commission-optimal design is

$$E_\alpha := (1-\widetilde{\alpha}_\alpha) g^{\mathrm{com}}_{\mathsf{P}} + a_\alpha\, g\big(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}\big) + b_\alpha\, g\big(u_- - p^\star_{\mathsf{N},\alpha}\big).$$

Using (68), (69), and (70),

$$\begin{aligned} E_\alpha - g^{\mathrm{com}}_{\mathsf{P}} &= a_\alpha \Big(g(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}) - g^{\mathrm{com}}_{\mathsf{P}} \Big) + b_\alpha \Big(g(u_- - p^\star_{\mathsf{N},\alpha}) - g^{\mathrm{com}}_{\mathsf{P}} \Big) \\ &= \big(\alpha + O(\alpha^2)\big) \Big(\beta \ln(1/\alpha) + O(1) \Big) + O(\alpha) \cdot O(1) \\ &= \beta\, \alpha \ln(1/\alpha) + O(\alpha). \end{aligned}$$

Hence

$$E_\alpha = g_{\mathrm{P}}^{\mathrm{com}} + \beta\,\alpha\ln(1/\alpha) + O(\alpha). \tag{71}$$

Step 3: Mutual information and survival under the commission-optimal design.

Let

$$J_\alpha := I_{\alpha,n}(f^\star_\Pi(\alpha)).$$

Since $\delta_\alpha = O(\alpha)$, there exists $K < \infty$ such that

$$\delta_\alpha = c_\alpha \alpha \qquad \text{with } c_\alpha \in [0, K]$$

for all sufficiently small α. Also

$$\widetilde{\alpha}_\alpha = \alpha(1-\delta_\alpha) + \Big(1 - \frac{\alpha}{n}\Big)\delta_\alpha = \alpha\widetilde{c}_\alpha,$$

where

$$\widetilde{c}_\alpha := 1 + c_\alpha + O(\alpha) \in [0, K+2]$$

for all sufficiently small α.

Recall

$$J_\alpha = H_n(\widetilde{\alpha}_\alpha) - H_n(\delta_\alpha), \qquad H_n(x) := h(x) + x\ln n.$$

We claim that, for every fixed $C < \infty$,

$$H_n(c\alpha) = c\alpha\ln(1/\alpha) + O(\alpha) \qquad \text{uniformly over } c \in [0, C]. \tag{72}$$

To see this, for $x \downarrow 0$,

$$H_n(x) = x\ln(1/x) + x(1 + \ln n) + O(x^2).$$

Substituting $x = c\alpha$ gives

$$H_n(c\alpha) = c\alpha\ln(1/\alpha) + c\alpha\big(1 + \ln n - \ln c\big) + O(\alpha^2)$$

for $c > 0$, and the coefficient $c(1 + \ln n - \ln c)$ extends continuously to $c = 0$ with value 0. Hence the remainder is uniformly $O(\alpha)$ on $[0, C]$, proving (72).

Applying (72) with $c = c_\alpha$ and $c = \widetilde{c}_\alpha$ yields

$$J_\alpha = \widetilde{c}_\alpha\alpha\ln(1/\alpha) - c_\alpha\alpha\ln(1/\alpha) + O(\alpha) = \alpha\ln(1/\alpha) + O(\alpha),$$

because $\widetilde{c}_\alpha - c_\alpha = 1 + O(\alpha)$. Therefore

$$s_\alpha := e^{-\eta J_\alpha} = 1 - \eta\,\alpha\ln(1/\alpha) + O(\alpha), \tag{73}$$

and

$$1 - s_\alpha = \eta\,\alpha\ln(1/\alpha) + O(\alpha). \tag{74}$$

Step 4: Welfare expansion.

Welfare under the commission-optimal design is

$$W_{\Pi}^{\mathsf{com}}(\alpha) = s_\alpha E_\alpha - \frac{\bar\kappa_u}{\eta}(1-s_\alpha) = E_\alpha - (1-s_\alpha)\Big(E_\alpha + \frac{\bar\kappa_u}{\eta}\Big).$$

Using (71) and (74),

$$E_\alpha + \frac{\bar\kappa_u}{\eta} = g_{\mathsf{P}}^{\mathrm{com}} + \frac{\bar\kappa_u}{\eta} + o(1),$$

because $\alpha\ln(1/\alpha)\to 0$. Hence

$$\begin{aligned} W_{\Pi}^{\mathsf{com}}(\alpha) &= g_{\mathsf{P}}^{\mathrm{com}} + \beta\,\alpha\ln(1/\alpha) + O(\alpha) \\ &\quad - \Big(\eta\,\alpha\ln(1/\alpha) + O(\alpha)\Big)\Big(g_{\mathsf{P}}^{\mathrm{com}} + \frac{\bar\kappa_u}{\eta} + o(1)\Big). \end{aligned}$$

Since $O(\alpha) = o(\alpha\ln(1/\alpha))$, this simplifies to

$$W_{\Pi}^{\mathsf{com}}(\alpha) = g_{\mathsf{P}}^{\mathrm{com}} + \Big[\beta - \eta g_{\mathsf{P}}^{\mathrm{com}} - \bar\kappa_u\Big]\alpha\ln(1/\alpha) + o\big(\alpha\ln(1/\alpha)\big).$$

Finally, $W^{\mathsf{com}}(\mathsf{REC}) = g_{\mathsf{P}}^{\mathrm{com}}$, so

$$\Delta W_{\Pi}^{\mathsf{com}}(\alpha) = \Big[\beta - \bar\kappa_u - \eta\, g(u_{\mathsf{P}} - p_{\mathsf{P}}^\star)\Big]\alpha\ln(1/\alpha) + o\big(\alpha\ln(1/\alpha)\big).$$

Dividing by $\alpha\ln(1/\alpha)$ gives the stated limit, and the sign characterization follows immediately. □

A.8. Proof of Proposition 3

Proof of Proposition 3: We prove each part individually. Note that part (ii) has 5 steps.

Throughout, survival is $sur^{\mathsf{MI}}(f) = \exp(-\eta I_{\alpha,n}(f))$. We use that $\sigma(\cdot)$ is strictly increasing and continuous.

Part (i): low heterogeneity. As $\lambda\to 0$, we have $u_-(\lambda) = -\lambda \to 0$ and $u_+(\lambda) = \lambda V_{\mathsf{N}}\alpha^{-\beta}\to 0$. Since $u_{\mathsf{P}} = V_{\mathsf{P}} > 0$ is fixed, there exists $\bar\lambda > 0$ such that for all $\lambda\in[0,\bar\lambda]$,

$$u_{\mathsf{P}} \;>\; \max\{u_-(\lambda), u_+(\lambda)\}. \tag{75}$$

Conversion objective. Fix any fidelity f and recommendation rule π. Conditional on reaching the recommendation stage, the adoption probability when recommending P equals $\sigma(u_{\mathsf{P}})$ for every type, because match utility for P is 0 for all Θ. When recommending any niche item j, the deterministic utility is either $u_+(\lambda)$ (if $\Theta = j$) or $u_-(\lambda)$ (if $\Theta\neq j$). By (75) and monotonicity of σ,

$$\sigma\big(u_{\mathsf{B}}(\Theta,\pi(S))\big) \;\le\; \sigma(u_{\mathsf{P}}) \qquad \text{a.s.}$$

with strict inequality on any event where $\pi(S)\neq\mathsf{P}$. Hence for all (f,π),

$$\mathbb{E}\big[\sigma\big(u_{\mathsf{B}}(\Theta,\pi(S))\big)\big] \;\le\; \sigma(u_{\mathsf{P}}),$$

and equality requires $\pi(S)\equiv\mathsf{P}$ almost surely. Therefore, among all designs the best achievable conversion payoff is

$$\sup_{f,\pi}\Gamma(f,\pi) = \sup_f R\, sur^{\mathsf{MI}}(f)\,\sigma(u_{\mathsf{P}}).$$

Because $I_{\alpha,n}(f)\ge 0$ with equality if and only if $f = \frac{1}{n+1}$, we have $sur^{\mathsf{MI}}(f)\le 1$ with equality at $f=\frac{1}{n+1}$. Thus the conversion optimum is attained by REC.

Commission objective. Let $D(p;u) := \sigma(u-p)$ and define $r(u) := \max_{p\geq 0} p\, D(p;u)$. For any $u_1 < u_2$ and any $p \geq 0$, $D(p;u_1) \leq D(p;u_2)$, hence $pD(p;u_1) \leq pD(p;u_2)$ and therefore $r(u_1) \leq r(u_2)$: $r(\cdot)$ is increasing.

Fix any (f,π) and consider any seller j. Seller j faces a (possibly mixed) demand curve of the form $p \mapsto \mathbb{E}[D(p;u_{\mathsf{B}}(\Theta,j)) \mid \pi(S)=j]$ scaled by the recommendation probability and survival. By (75), we have $u_{\mathsf{B}}(\Theta,j) \leq u_{\mathsf{P}}$ almost surely, so for every p,

$$\mathbb{E}[D(p;u_{\mathsf{B}}(\Theta,j)) \mid \pi(S)=j] \;\leq\; D(p;u_{\mathsf{P}}).$$

Multiplying by p and taking the supremum over $p \geq 0$ yields that seller j's maximal revenue is weakly bounded by $r(u_{\mathsf{P}})$. Consequently, the platform's maximal expected commission revenue conditional on reaching the pricing stage is achieved by routing all demand to P, and as in the conversion case any elicitation strictly lowers survival unless $f = \frac{1}{n+1}$. Therefore REC is also optimal for the commission objective when λ is sufficiently small. This proves part (i).

Part (ii): high heterogeneity, conversion. Fix $\alpha \in (0,1)$ and let $\sigma_{\mathsf{P}} := \sigma(u_{\mathsf{P}})$.

Step 1: limit adoption probabilities. As $\lambda \to \infty$, $u_+(\lambda) \to +\infty$ and $u_-(\lambda) \to -\infty$, hence

$$\sigma(u_+(\lambda)) \to 1, \qquad \sigma(u_-(\lambda)) \to 0. \tag{76}$$

Step 2: acting on niche signals requires $f \geq \underline{f}_{\infty}$ in the limit. Fix fidelity f and niche signal $S=i$. Let $q(f,\alpha,n) := \Pr(\Theta = i \mid S = i)$. A direct Bayes calculation under the symmetric channel gives

$$q(f,\alpha,n) = \frac{\alpha f}{\alpha f + (1-\alpha/n)(1-f)}, \tag{77}$$

which is increasing in f. If the platform recommends niche i after $S=i$, then conditional on $S=i$ the adoption probability equals

$$q(f,\alpha,n)\,\sigma(u_+(\lambda)) + (1-q(f,\alpha,n))\,\sigma(u_-(\lambda)).$$

If it recommends P, adoption equals σ_{P}. Therefore recommending niche i after $S=i$ is optimal for conversion if and only if

$$q(f,\alpha,n) \;\geq\; q^{\star}_{\Gamma}(\lambda) := \frac{\sigma_{\mathsf{P}} - \sigma(u_-(\lambda))}{\sigma(u_+(\lambda)) - \sigma(u_-(\lambda))}. \tag{78}$$

By (76), $q^{\star}_{\Gamma}(\lambda) \to \sigma_{\mathsf{P}}$ as $\lambda \to \infty$.

Let

$$\underline{f}_{\infty} := \frac{\sigma_{\mathsf{P}}\,(1-\alpha/n)}{\sigma_{\mathsf{P}}\,(1-\alpha/n) + \alpha(1-\sigma_{\mathsf{P}})} \tag{79}$$

Define $\underline{f}_{\infty}$ by $q(\underline{f}_{\infty},\alpha,n) = \sigma_{\mathsf{P}}$; solving (77) yields (79).

Step 3: limiting conversion objective when niche signals are used. For any $f \geq \underline{f}_{\infty}$, the conversion-optimal rule (for large λ) recommends the niche indicated by S whenever $S \in \{1,\dots,n\}$, and recommends P when $S = \emptyset$. Under this rule, the probability of a correct niche recommendation is $\Pr(S = \Theta \in \{1,\dots,n\}) = \alpha f$. Let $q_0(f) = \Pr(S=\emptyset) = (1-\alpha)f + \alpha\frac{1-f}{n}$. Then the (conditional-on-survival) expected adoption probability equals

$$q_0(f)\,\sigma_{\mathsf{P}} + \alpha f\,\sigma(u_+(\lambda)) + \big(1 - q_0(f) - \alpha f\big)\sigma(u_-(\lambda)).$$

Using (76), this converges to $A(f) = \sigma_{\mathsf{P}} q_0(f) + \alpha f$. Multiplying by survival $sur^{\mathsf{MI}}(f)$ and R gives $\Gamma_{\infty}(f)$.

Step 4: REC*-versus-elicitation comparison at high* λ. Under REC, $I_{\alpha,n}(f)=0$, $sur^{\mathsf{MI}}(f)=1$, and the platform always recommends P, so $\Gamma^{\mathsf{REC}}=R\sigma_{\mathsf{P}}$. If $f<\underline{f}_{\infty}$, then for large λ the conversion-optimal rule ignores niche signals (by Step 2) and recommends P always, so the payoff is $R\cdot sur^{\mathsf{MI}}(f)\sigma_{\mathsf{P}}<R\sigma_{\mathsf{P}}$ whenever $I_{\alpha,n}(f)>0$. Hence any conversion-improving elicitation must use $f\in[\underline{f}_{\infty},1]$, and in the limit $\lambda\to\infty$ the best attainable elicitation payoff equals $\max_{f\in[\underline{f}_{\infty},1]}\Gamma_{\infty}(f)$. This proves the REC-versus-elicitation condition stated in part (ii).

Step 5: log-concavity and first-order condition. The map $f\mapsto A(f)$ is affine and strictly positive on $[\underline{f}_{\infty},1]$, hence $\log A(f)$ is concave there. Mutual information is convex in the channel $P(S\mid\Theta)$ for a fixed prior; since the symmetric channel law depends affinely on f, $f\mapsto I_{\alpha,n}(f)$ is convex. Therefore $\log sur^{\mathsf{MI}}(f)=-\eta I_{\alpha,n}(f)$ is concave. It follows that $\log\Gamma_{\infty}(f)=\log R+\log sur^{\mathsf{MI}}(f)+\log A(f)$ is concave on $[\underline{f}_{\infty},1]$, so Γ_{∞} is log-concave and admits at least one maximizer. If the maximizer is interior, differentiating $\log\Gamma_{\infty}$ gives the first-order condition $-\eta I'_{\alpha,n}(f)+A'(f)/A(f)=0$. This completes part (ii).

Part (iii): high heterogeneity, commission. Fix $\alpha\in(0,1)$ and take-rate $\tau\in(0,1)$. Let $r(u)=\max_{p\geq 0}p\,\sigma(u-p)$. We first show $r(u)\to\infty$ as $u\to\infty$: for any $u>0$, choosing $p=u/2$ yields revenue

$$p\,\sigma(u-p)=\frac{u}{2}\,\sigma(u/2)\ \xrightarrow[u\to\infty]{}\ \frac{u}{2},$$

so $r(u)\geq\frac{u}{2}\sigma(u/2)\to\infty$.

Consider the full-revelation policy $\mathsf{ATR}(1)$ with type-contingent routing: recommend P if $\Theta=\emptyset$ and recommend niche j if $\Theta=j$. Under this design, each seller faces routed demand from a single deterministic-utility segment (u_{P} for the popular seller and $u_{+}(\lambda)$ for each niche seller), scaled by survival $sur^{\mathsf{MI}}(1)$ and the segment mass. The seller's best-response price therefore solves $\max_{p\geq 0}p\,\sigma(u-p)$ and yields revenue $r(u)$. Summing expected platform commissions across sellers gives the lower bound (80).

$$\Pi^{\mathsf{com}}(\mathsf{ATR}(1))\ \geq\ \tau\,sur^{\mathsf{MI}}(1)\Big[(1-\alpha)\,r(u_{\mathsf{P}})+\alpha\,r(u_{+}(\lambda))\Big]. \tag{80}$$

Under REC, survival is 1 and the platform always recommends P, so the platform earns

$$\Pi^{\mathsf{com}}(\mathsf{REC})=\tau\,r(u_{\mathsf{P}}),$$

which is finite and independent of λ. Since $u_{+}(\lambda)=\lambda V_{\mathsf{N}}\alpha^{-\beta}\to\infty$ as $\lambda\to\infty$ and $r(u)\to\infty$ as $u\to\infty$, the lower bound (80) diverges with λ. Hence there exists $\bar{\lambda}<\infty$ such that for all $\lambda\geq\bar{\lambda}$, $\Pi^{\mathsf{com}}(\mathsf{ATR}(1))>\Pi^{\mathsf{com}}(\mathsf{REC})$. Therefore the commission-optimal design cannot coincide with REC for sufficiently large λ. This proves part (iii). □

A.9. Proof of Proposition 4

Proof of Proposition 4: The proof consists of 4 steps: 1) we show that if π never recommends a niche, then REC dominates, 2) any policy that recommends a niche requires fidelity bounded away from zero, 3) at any fixed positive fidelity, mutual information diverges to ∞ as $n\to\infty$, and 4) survival goes to zero, so REC dominates.

Consider any CRS design (f,π) with $f>1/(n+1)$.

Step 1: If π never recommends a niche, REC *dominates.* If $\pi(S) \equiv \mathsf{P}$ for all S, the recommendation-stage payoff is the same as under REC but survival satisfies $sur^{\mathsf{MI}}(f) \leq 1$ with strict inequality whenever $I_{\alpha,n}(f) > 0$, and effort loss is weakly positive. Hence REC weakly (and generically strictly) dominates any such design under all three objectives.

Step 2: Any policy that recommends a niche requires fidelity bounded away from zero, uniformly in n. Suppose π recommends a niche with positive probability. By Lemma 2, this requires $q(f,\alpha,n) \geq q^\star$ for the relevant objective-specific posterior cutoff $q^\star \in (0,1)$. Since $\sigma(u_+(\alpha)) \leq 1$ (for conversion) and $g(u_+(\alpha)) < \infty$ for fixed α (for welfare), the cutoff $q^\star$ is bounded away from zero uniformly in n.[6]

We have,

$$q(f,\alpha,n) = \frac{\alpha f}{\alpha f + (1-\alpha/n)(1-f)}.$$

Fix any $q^\star \in (0,1)$. The condition $q(f,\alpha,n) \geq q^\star$ rearranges to

$$f \;\geq\; \frac{q^\star(1-\alpha/n)}{q^\star(1-\alpha/n) + \alpha(1-q^\star)} \;=:\; \underline{f}(q^\star,\alpha,n).$$

Since $\alpha \in (0,1)$ is fixed and $1-\alpha/n \to 1$ as $n \to \infty$, we have $\underline{f}(q^\star,\alpha,n) \to q^\star/(q^\star + \alpha(1-q^\star)) > 0$. In particular, there exists $\underline{f}_0 > 0$ (independent of n) such that $f \geq \underline{f}_0$ is necessary for any niche recommendation, for all n large enough.

Step 3: At any fixed positive fidelity, mutual information diverges as $n \to \infty$. Fix any $f_0 \in (0,1)$. Using the closed form (22), let $a := (1-f_0)/n$ and compute

$$H(S \mid \Theta) = -f_0 \ln f_0 - (1-f_0)\ln\Big(\frac{1-f_0}{n}\Big) = -f_0 \ln f_0 - (1-f_0)\ln(1-f_0) + (1-f_0)\ln n.$$

Meanwhile, $H(S) \geq H(\Theta \mid S) \geq 0$ implies $H(S) \geq I_{\alpha,n}(f_0) \geq 0$, but more directly we can write $I_{\alpha,n}(f_0) = H(S) - H(S \mid \Theta)$. For the marginal signal probabilities,

$$q_1 = \Pr(S = i) = \frac{\alpha}{n} f_0 + \Big(1 - \frac{\alpha}{n}\Big)\frac{1-f_0}{n}, \qquad i \in \{1,\dots,n\}.$$

As $n \to \infty$, $nq_1 \to \alpha f_0 + (1-f_0) = 1 - (1-\alpha)f_0 \in (0,1)$, so $q_1 \sim (1-(1-\alpha)f_0)/n$. Similarly, $q_0 = \Pr(S = \emptyset) = (1-\alpha)f_0 + \alpha(1-f_0)/n \to (1-\alpha)f_0$. Hence

$$H(S) = -q_0 \ln q_0 - nq_1 \ln q_1 = -q_0 \ln q_0 - nq_1\big(\ln(nq_1) - \ln n\big) = -q_0 \ln q_0 - nq_1 \ln(nq_1) + nq_1 \ln n.$$

Since $q_0 \to (1-\alpha)f_0$ and $nq_1 \to 1-(1-\alpha)f_0$, the terms $-q_0 \ln q_0 - nq_1 \ln(nq_1)$ converge to finite limits. Subtracting $H(S \mid \Theta)$ from $H(S)$, the $\ln n$ contributions yield

$$I_{\alpha,n}(f_0) = \big(nq_1 - (1-f_0)\big)\ln n + O(1) = \alpha f_0 \ln n + O(1) \qquad \text{as } n \to \infty.$$

[6] For conversion, $q^\star_\Gamma = (\sigma(u_\mathsf{P}) - \sigma(u_-))/(\sigma(u_+) - \sigma(u_-)) \geq (\sigma(u_\mathsf{P}) - \sigma(u_-))/(1 - \sigma(u_-)) > 0$. For welfare and commission the argument is analogous.

Step 4: Survival vanishes, so REC *dominates.* Combining Steps 2 and 3, any design that recommends a niche must have $f \geq \underline{f}_0 > 0$, and at such fidelity

$$I_{\alpha,n}(f) \;\geq\; I_{\alpha,n}(\underline{f}_0) \;\geq\; \alpha \underline{f}_0 \ln n - C$$

for some constant $C < \infty$ and all n sufficiently large, where the first inequality uses the data-processing inequality (higher fidelity yields weakly higher mutual information). Therefore

$$sur^{\mathsf{MI}}(f) = \exp\big(-\eta\, I_{\alpha,n}(f)\big) \;\leq\; \exp\big(-\eta(\alpha \underline{f}_0 \ln n - C)\big) \;=\; e^{\eta C}\, n^{-\eta\alpha \underline{f}_0} \;\xrightarrow{n\to\infty}\; 0.$$

Under all three objectives, the payoff from any elicitation design is bounded above by $sur^{\mathsf{MI}}(f) \cdot M$ for some finite M that depends on the objective and the (fixed) primitives $(\alpha, \lambda, \beta, V_{\mathsf{P}}, V_{\mathsf{N}})$ but not on n:[7]

$$\text{Objective}(f, \pi) \;\leq\; sur^{\mathsf{MI}}(f) \cdot M \;\to\; 0 \qquad \text{as } n \to \infty.$$

Meanwhile, REC yields a strictly positive, n-independent payoff ($g(u_{\mathsf{P}})$ under welfare, $R\sigma(u_{\mathsf{P}})$ under conversion, $\tau r_m(u_{\mathsf{P}})$ under commission). Therefore, for all $n \geq \bar{n}$ sufficiently large, REC strictly dominates every design that recommends a niche. Combined with Step 1, REC is optimal. □

Appendix B: Main results under capacity-based abandonment

This section parallels Section 3 under the capacity benchmark, i.e. $\mathcal{I}(f) = \mathcal{I}^{\mathsf{cap}}(f)$. A CRS design consists of a *conversation fidelity* f and a *recommendation rule* π mapping conversational signals into a single recommended product. We again focus on two platform objectives—*conversion/engagement* and *commission revenue with endogenous pricing*—and we use the induced *user welfare* to quantify alignment (or misalignment) between platform incentives and consumer outcomes.

Capacity-based abandonment and effort. Recall that under the capacity benchmark, the amount of elicited information is the channel capacity

$$C_{n+1}(f) = \ln(n+1) + f \ln f + (1-f) \ln\Big(\frac{1-f}{n}\Big),$$

so survival to the recommendation stage and expected effort loss are

$$sur^{\mathsf{cap}}(f) = \exp\big(-\eta\, C_{n+1}(f)\big), \qquad \text{EffortLoss}^{\mathsf{cap}}(f) = \frac{\bar{\kappa}_u}{\eta}\big(1 - sur^{\mathsf{cap}}(f)\big).$$

Unlike mutual information, $C_{n+1}(f)$ depends on (n, f) but not on the population prior; in particular $C_{n+1}(1) = \ln(n+1)$ is constant in α. This implies that the interaction burden does not vanish when fraction of niche seekers approaches zero.

A reduction: posterior thresholds. As in the MI benchmark, $sur^{\mathsf{cap}}(f)$ depends only on f (not on the realized signal), so conditional on a chosen f the optimal recommendation rule retains the same posterior-threshold structure. After a niche signal $S = i$, the platform compares recommending niche i to recommending the mainstream option P, and the optimal action is characterized by an objective-specific posterior cutoff (Lemma 2). This reduction lets us state the main results as one-dimensional design choices over f, together with their welfare implications.

Recall

$$u_{\mathsf{P}} := V_{\mathsf{P}}, \qquad u_- := -\lambda, \qquad u_+(\alpha) := \lambda V_{\mathsf{N}} \alpha^{-\beta}, \qquad \sigma(x) := \frac{e^x}{1+e^x}, \qquad g(x) := \ln(1+e^x),$$

and recall that REC corresponds to $f = 1/(n+1)$ and always recommending P, while $\mathsf{ATR}(f)$ denotes an ask-then-recommend policy with fidelity f and the optimal posterior-threshold rule.

[7] Under welfare, $M = g(u_+(\alpha)) + \bar{\kappa}_u/\eta$; under conversion, $M = R$; under commission, $M = \tau\, r_m(u_+(\alpha))$.

B.1. Rare niches with potentially large match value ($\alpha \to 0$)

As in Section 3.1, the rare-niche regime isolates a "rare-gems" environment: most users are well served by the mainstream option P, but a vanishing fraction have niche tastes. At the same time, a correct niche match can be extremely valuable. We capture this with the scaling $u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{-\beta}$, where β indexes how fast niche value grows as niches become rarer.

B.1.1. Conversion/engagement: recommend-first, even more sharply under capacity With fixed prices ($p_j \equiv 0$), the conversion objective under capacity-based abandonment is

$$\Gamma(f, \pi) = R \cdot sur^{\mathsf{cap}}(f) \cdot \mathbb{E}[\sigma(u_{\mathsf{B}}(\Theta, \pi(S)))] .$$

THEOREM 3 **(Conversion-optimal recommend-first in the rare-niche regime (capacity))**. *Fix $n \geq 1$, $R > 0$, and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta)$. Assume $\mathcal{I}(f) = \mathcal{I}^{\mathsf{cap}}(f)$. There exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$, the conversion-optimal policy is* REC.

Discussion and mechanism. The logic is the same bounded-upside argument as under MI, but *stronger*. Even perfect niche matching can increase conversion by only a bounded amount, and this upside accrues on only an α-fraction of users. By contrast, under capacity, any nontrivial elicitation induces a *first-order* survival discount on the mainstream majority: if a policy ever acts on niche signals when α is small, it must drive posteriors high, which forces f close to 1 and hence $sur^{\mathsf{cap}}(f) \approx sur^{\mathsf{cap}}(1) = (n+1)^{-\eta} < 1$. Thus the platform pays a constant abandonment penalty on mainstream revenue to chase an $O(\alpha)$ conversion gain, so REC is optimal. In Figure 11a we plot the difference between the conversion obtained under the optimal elicitation policy and REC. For the representative model primitives detailed in Figure caption, we observe that for small enough α (see $\alpha \leq 0.3$ in Figure 11a), the difference is zero, i.e., REC is optimal which verifies Theorem 3.

Managerial interpretation. The theorem should be read comparatively: REC corresponds to *minimal* probing rather than "never asking." The result formalizes why engagement-driven platforms (watch-time, ad impressions, session completion) have incentives to front-load strong mainstream recommendations and economize on deep elicitation in rare-gems environments. Under capacity-style frictions, deep, high-resolution conversation imposes a *non-vanishing* abandonment cost, so conversion incentives become even more conservative than under MI.

User-welfare-maximizing benchmark and welfare misalignment under capacity We use user welfare to benchmark alignment. With fixed prices, welfare under capacity-based abandonment is

$$W^{\mathsf{cap}}(f, \pi) = sur^{\mathsf{cap}}(f) \cdot \mathbb{E}[g(u_{\mathsf{B}}(\Theta, \pi(S)))] - \frac{\bar{\kappa}_u}{\eta} \left(1 - sur^{\mathsf{cap}}(f)\right) .$$

PROPOSITION 5 **(Welfare-optimal benchmark)**. *Fix $n \geq 1$ and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta, \bar{\kappa}_u)$ in the fixed-price environment.*

1. [Subcritical Regime] *Let $\beta < 1$. There exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$, the user-welfare-maximizing policy is* REC.

2. [Supercritical Regime] *Let $\beta > 1$. There exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$, welfare strictly prefers elicitation to* REC. *Moreover, letting f_∞^{cap} denote the unique maximizer of $f \mapsto sur^{\mathsf{cap}}(f)\, f$ over $[1/(n+1), 1]$. The user-welfare-maximizing fidelity $f_W^\star(\alpha)$ satisfies $f_W^\star(\alpha) \to f_\infty^{\mathsf{cap}}$ where f_∞^{cap} is the unique solution to $\frac{1}{f} = \eta \ln(nf/1-f)$.*

3. [Critical Regime] *Fix $\beta = 1$. Let $L := \lambda V_{\mathsf{N}}$, define $g_{\mathsf{P}} := g(V_{\mathsf{P}})$ and $g_- := g(-\lambda)$, and set*

$$A := g_{\mathsf{P}} + L - g_- > 0.$$

Define the minimal fidelity for acting on niche signals *in the rare-niche limit,*

$$\underline{f}_W^{\mathsf{cap}} := \max\Big\{\frac{1}{n+1},\ \frac{g_{\mathsf{P}} - g_-}{g_{\mathsf{P}} - g_- + L}\Big\} \in \Big[\frac{1}{n+1}, 1\Big).$$

For any fixed $f \geq \underline{f}_W^{\mathsf{cap}}$, the user-welfare-optimal signal rule recommends the niche after any niche signal for all sufficiently small α. In this case, user welfare converges to the limiting envelope

$$\overline{W}^{\mathsf{cap}}(f) := sur^{\mathsf{cap}}(f) \cdot \big(g_- + fA\big) - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur^{\mathsf{cap}}(f)\big), \qquad f \in \Big[\underline{f}_W^{\mathsf{cap}}, 1\Big]. \tag{81}$$

Define the cutoff

$$\bar{\kappa}_u^{\mathsf{cap}} := \max\left\{0,\ \eta \cdot \sup_{f \in [\underline{f}_W^{\mathsf{cap}}, 1]} \frac{sur^{\mathsf{cap}}(f)\big(g_- + fA\big) - g_{\mathsf{P}}}{1 - sur^{\mathsf{cap}}(f)}\right\} \ \in\ [0, \infty). \tag{82}$$

Then there exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$:

(a) *If $\bar{\kappa}_u \geq \bar{\kappa}_u^{\mathsf{cap}}$, then* REC *is user-welfare-optimal.*

(b) *If $\bar{\kappa}_u < \bar{\kappa}_u^{\mathsf{cap}}$, then elicitation is welfare-improving and the user-welfare-maximizing fidelity satisfies*

$$f_W^\star(\alpha) \ \longrightarrow\ f_{\beta=1}^{\mathsf{cap}} \in \arg\max_{f \in [\underline{f}_W^{\mathsf{cap}}, 1]} \overline{W}^{\mathsf{cap}}(f).$$

If the maximizer is unique and interior, $f_{\beta=1}^{\mathsf{cap}}$ is characterized by the first-order condition

$$\frac{A}{g_- + \frac{\bar{\kappa}_u}{\eta} + Af} = \eta\, C'_{n+1}(f) = \eta \ln\Big(\frac{nf}{1-f}\Big).$$

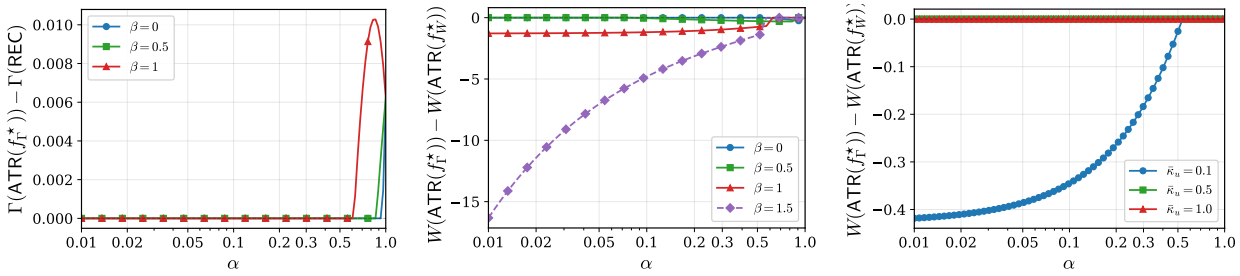
Discussion. Proposition 5 is the main conceptual contrast with the MI benchmark. Under MI, the information required to reveal type shrinks as $H(\Theta) \asymp \alpha \ln(1/\alpha)$, so abandonment and effort costs become asymptotically negligible; as a result, welfare can justify "hunting for rare gems" whenever $\beta > 0$. Under capacity, the interaction burden is a constant whenever elicitation is materially informative, so user welfare is more conservative: for $\beta < 1$ the aggregate niche gains $\alpha u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{1-\beta}$ vanish and cannot justify a constant survival/effort penalty. The knife-edge $\beta = 1$ is the unique regime in which the aggregate niche gains remain $O(1)$, yielding a genuine cost-benefit comparison and an explicit effort cutoff (82). When $\beta > 1$, aggregate niche gains explode, so welfare elicits even under capacity—but the *optimal fidelity does not necessarily converge to full revelation.* Instead it converges to the maximizer of $sur^{\mathsf{cap}}(f) \cdot f$, reflecting a clean trade-off between (i) increasing the probability of a correct match and (ii) preserving survival. Figures 11b and 11c numerically verify our results. For a fixed and small enough $\bar{\kappa}_u$, we observe that for $\beta < 1$, REC is user-welfare-optimal while for $\beta > 1$, elicitation is optimal for α sufficiently small. For fixed $\beta = 1$ and changing the disutility factor $\bar{\kappa}_u$, we observe that for small enough $\bar{\kappa}_u$, elicitation is user-welfare-optimal while as $\bar{\kappa}_u$ increases, REC becomes user-welfare-optimal.

COROLLARY 3 **(Welfare misalignment under engagement optimization with capacity).** *Fix $n \geq 1$ and primitives as above. In the rare-niche regime, conversion selects* REC *by Theorem 3. If either $\beta > 1$, or $\beta = 1$ and $\bar{\kappa}_u < \bar{\kappa}_u^{\mathsf{cap}}$, then* REC *is strictly welfare-suboptimal for sufficiently small α. Moreover, the welfare loss from conversion-driven design satisfies*

$$W^\star(\alpha) - W(\mathsf{REC}) = \Omega\big(sur^{\mathsf{cap}}(f_W^\star(\alpha)) \cdot \lambda V_{\mathsf{N}} \alpha^{1-\beta}\big),$$

so the wedge is of constant order at $\beta = 1$ and diverges at rate $\alpha^{1-\beta}$ when $\beta > 1$.

Managerial interpretation. Under capacity-like frictions, the "welfare gap" is not automatic: welfare itself will only invest in deep elicitation when rare matches are valuable *enough* to justify a non-vanishing abandonment/effort penalty. But when welfare does justify elicitation (e.g., high-stakes or extremely high-value rare matches), engagement monetization still selects recommend-first. This strengthens the message that an engineering goal of "maximizing personalization" need not align with revenue models based on bounded engagement probabilities.



(a) Value of Elicitation for Conversion Objective for $\bar{\kappa}_u = 0.1$

(b) User Welfare Loss due to Conversion-Optimal Policy for $\bar{\kappa}_u = 0.1$

(c) User Welfare Loss due to Conversion-Optimal Policy for $\beta = 1$

Figure 11 (a) Difference between the optimal elicitation policy $\mathrm{ATR}(f_\Gamma^\star)$ **and** REC **for the conversion objective** (8) **as a function of α for different values of $\beta \in \{0, 0.5, 1\}$ (b) Difference between the user welfare obtained under the conversion-optimal elicitation and user-welfare-optimal elicitation as a function of α for different values of $\beta \in \{0, 0.5, 1, 1.5\}$. (c) Difference between the user welfare obtained under the conversion-optimal elicitation and user-welfare-optimal elicitation as a function of α for different values of $\bar{\kappa}_u \in \{0.1, 0.5, 1.0\}$. For all the plots, we set $V_{\mathsf{P}} = 1, V_{\mathsf{N}} = 2, \eta = 0.1, n = 5, \lambda = 1, R = 1$**

THEOREM 4 **(Commission incentives under capacity-based abandonment).** *Fix $n \geq 1$, $\tau \in (0,1)$, and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta)$. Consider the commission objective* (11) *under* capacity-based *abandonment, i.e., $sur^{\mathsf{cap}}(f) = \exp\big(-\eta C_{n+1}(f)\big)$. Let $u_{\mathsf{P}} := V_{\mathsf{P}}$ and $u_+(\alpha) := \lambda V_{\mathsf{N}} \alpha^{-\beta}$. Let $r_m(u) := \max_{p \geq 0} p \cdot \sigma(u - p)$ denote the monopoly revenue under logit demand, and set $r_{\mathsf{P}} := r_m(u_{\mathsf{P}})$. Define f_∞^{cap} as the unique maximizer of $f \mapsto sur^{\mathsf{cap}}(f)\, f$ on $[1/(n+1), 1]$.*

(i) [Subcritical: $\beta < 1$]. *There exists $\bar{\alpha} \in (0,1)$ such that for all $\alpha \in (0, \bar{\alpha})$, the commission-optimal policy is* REC.

(ii) [Supercritical: $\beta > 1$]. *There exists* $\bar{\alpha} \in (0,1)$ *such that for all* $\alpha \in (0, \bar{\alpha})$*, the platform strictly prefers elicitation to* REC*. Moreover, the commission-optimal fidelity satisfies*

$$f^{\star}_{\Pi}(\alpha) \;\longrightarrow\; f^{\mathsf{cap}}_{\infty} \qquad \text{as } \alpha \to 0.$$

(iii) [Critical: $\beta = 1$]. *Let* $L := \lambda V_{\mathsf{N}}$ *and define*

$$M^{\mathsf{cap}} := \max_{f \in [1/(n+1),1]} sur^{\mathsf{cap}}(f)\, f = sur^{\mathsf{cap}}(f^{\mathsf{cap}}_{\infty})\, f^{\mathsf{cap}}_{\infty}.$$

Then

$$\Pi^{\mathsf{cap}}\big(\mathsf{ATR}(f^{\star}_{\Pi}(\alpha))\big) - \Pi^{\mathsf{cap}}(\mathsf{REC}) \;\longrightarrow\; \tau \max\Big\{0,\, (r_{\mathsf{P}} + L)M^{\mathsf{cap}} - r_{\mathsf{P}}\Big\} \qquad \text{as } \alpha \to 0.$$

In particular, elicitation asymptotically dominates REC *if and only if*

$$(r_{\mathsf{P}} + L)M^{\mathsf{cap}} \;>\; r_{\mathsf{P}} \quad\Longleftrightarrow\quad L \;>\; \Big(\frac{1}{M^{\mathsf{cap}}} - 1\Big)\, r_{\mathsf{P}}. \tag{83}$$

Moreover, if the strict inequality in (83) *holds, then elicitation is commission-optimal for all sufficiently small* α*, and*

$$f^{\star}_{\Pi}(\alpha) \;\longrightarrow\; f^{\mathsf{cap}}_{\infty}.$$

Discussion and mechanism. Commission monetization makes elicitation platform optimal because it improves *screening*. Higher fidelity increases the concentration of high-value niche users among those routed to niche sellers, which supports higher prices and higher seller revenue. The platform captures a fraction τ of this revenue. Under capacity, however, screening is not asymptotically cheap: elicitation imposes a non-vanishing survival penalty. This is why $\beta = 1$ yields the sharp cutoff (83): constant niche revenue L must be large enough to compensate for lost mainstream monopoly revenue due to abandonment. Figure 12a illustrates this result numerically. For $\beta < 1$, the gap between the optimal elicitation and REC decreases, indicating that for small enough α, REC is optimal for $\beta < 1$. For $\beta = 1$, we observe that the difference is always positive implying that elicitation is optimal in this instance.

Consumer welfare under commission (capacity). A key implication of capacity-based abandonment is that consumer welfare under commission becomes asymptotically *unambiguous* in the rare-niche limit. Under monopoly pricing, consumer surplus grows only logarithmically in willingness-to-pay, so the niche segment contributes at most $O(\alpha \ln(1/\alpha))$ surplus. But any nontrivial elicitation imposes an $O(1)$ survival discount on mainstream users (and an $O(1)$ effort loss), which dominates as $\alpha \to 0$.

PROPOSITION 6 **(Consumer welfare under the commission-optimal CRS (capacity))**. *Fix* $n \geq 1$, $\tau \in (0,1)$*, and primitives* $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta, \bar{\kappa}_u)$*. Let*

$$\Delta W^{\mathsf{com,cap}}_{\Pi}(\alpha) := W^{\mathsf{com,cap}}\Big(\mathsf{ATR}(f^{\star}_{\Pi}(\alpha))\Big) - W^{\mathsf{com,cap}}(\mathsf{REC}).$$

Then the following hold:

(i) [Subcritical niche value] *If* $\beta < 1$*, then there exists* $\bar{\alpha} \in (0,1)$ *such that*

$$\Delta W^{\mathsf{com,cap}}_{\Pi}(\alpha) = 0 \qquad \text{for all } \alpha \in (0, \bar{\alpha}).$$

(ii) [Supercritical niche value] *Suppose $\beta > 1$ and $\lambda > 0$. Let f_∞^{cap} be the unique maximizer of $f \mapsto sur^{\mathsf{cap}}(f)\, f$ on $[1/(n+1), 1]$, write $sur_\infty^{\mathsf{cap}} := sur^{\mathsf{cap}}(f_\infty^{\mathsf{cap}})$, and define $p_{\mathsf{P}}^\star := p_m(u_{\mathsf{P}})$. Then*

$$\lim_{\alpha \to 0} \Delta W_\Pi^{\mathsf{com,cap}}(\alpha) = -\big(1 - sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}}\big)\, g\big(u_{\mathsf{P}} - p_{\mathsf{P}}^\star\big) - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur_\infty^{\mathsf{cap}}\big) < 0.$$

In particular, for sufficiently small α, consumer welfare under the commission-optimal CRS is lower than under REC.

(iii) [Critical niche value] *Suppose $\beta = 1$. Let*

$$L := \lambda V_{\mathsf{N}}, \qquad r_{\mathsf{P}} := r_m(u_{\mathsf{P}}), \qquad M^{\mathsf{cap}} := \max_{f \in [1/(n+1),1]} sur^{\mathsf{cap}}(f)\, f = sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}},$$

where f_∞^{cap} is the unique maximizer of $f \mapsto sur^{\mathsf{cap}}(f)\, f$ on $[1/(n+1), 1]$, and let $p_{\mathsf{P}}^\star := p_m(u_{\mathsf{P}})$. If

$$(r_{\mathsf{P}} + L) M^{\mathsf{cap}} \le r_{\mathsf{P}},$$

then there exists $\bar{\alpha} \in (0,1)$ such that (breaking ties in favor of REC*)*

$$\Delta W_\Pi^{\mathsf{com,cap}}(\alpha) = 0 \qquad \text{for all } \alpha \in (0, \bar{\alpha}).$$

If instead $(r_{\mathsf{P}} + L) M^{\mathsf{cap}} > r_{\mathsf{P}}$, then

$$\lim_{\alpha \to 0} \Delta W_\Pi^{\mathsf{com,cap}}(\alpha) = -\big(1 - sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}}\big)\, g\big(u_{\mathsf{P}} - p_{\mathsf{P}}^\star\big) - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur_\infty^{\mathsf{cap}}\big) < 0.$$

Therefore, for sufficiently small α and $\beta = 1$, consumer welfare under the commission-optimal CRS is equal to REC *if $(r_{\mathsf{P}} + L) M^{\mathsf{cap}} \le r_{\mathsf{P}}$, and lower otherwise.*

Implication. Under capacity, whenever commission incentives induce deep elicitation in a rare-gems environment, consumer welfare eventually falls: the mainstream survival loss is first-order, while niche consumer surplus is second-order in α. This delivers a sharp version of the "personalization paradox" logic: screening can raise platform revenue through prices while reducing consumer welfare. Figure 12b ratifies our results numerically—observe that for small enough α, the difference is negative.

B.2. Comparative statics: heterogeneity and variety (capacity)

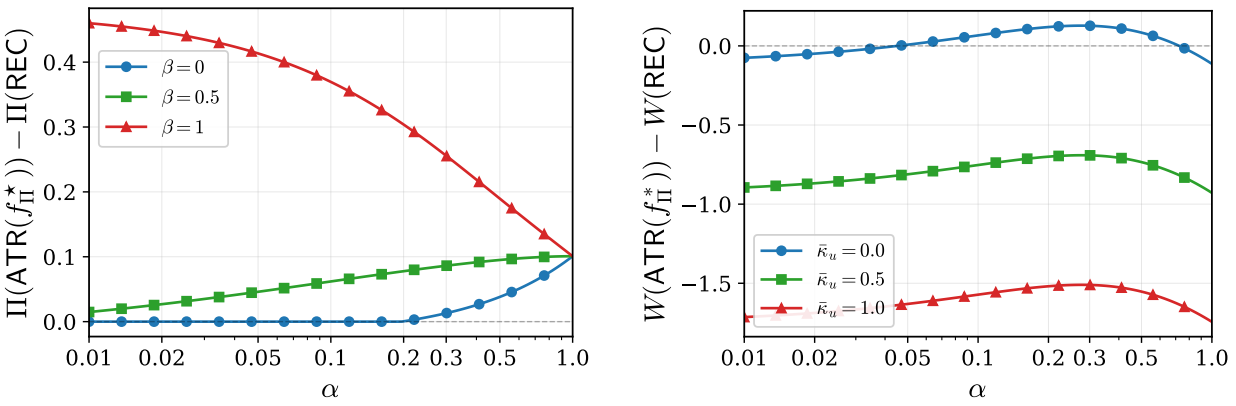
The qualitative comparative statics in λ and n mirror Section 3.2, but capacity shifts thresholds because the interaction friction is inherently tied to the resolution of the interface.

PROPOSITION 7 **(Impact of heterogeneity λ under capacity-based abandonment)**. *Fix primitives $(\alpha, n, \beta, V_{\mathsf{P}}, V_{\mathsf{N}}, \eta, R, \tau)$ with $\alpha \in (0,1)$ and $n \ge 1$. Let $\sigma_{\mathsf{P}} := \sigma(V_{\mathsf{P}})$ and let survival be $sur^{\mathsf{cap}}(f) := \exp(-\eta C_{n+1}(f))$.*

Define

$$q_0(f) := (1-\alpha) f + \alpha \frac{1-f}{n}, \qquad A_\infty(f) := \sigma_{\mathsf{P}}\, q_0(f) + \alpha f, \quad \underline{f}_\infty := \frac{\sigma_{\mathsf{P}}\,(1 - \alpha/n)}{\sigma_{\mathsf{P}}\,(1-\alpha/n) + \alpha(1-\sigma_{\mathsf{P}})}.$$

(i) ***Low heterogeneity.*** *There exists $\bar{\lambda} > 0$ such that for all $\lambda \in [0, \bar{\lambda}]$, the optimal CRS under both conversion and commission objectives is* REC.



(a) Value of elicitation for commission objective

(b) User welfare (mis)alignment under commission objective

Figure 12 (a) Difference between the optimal elicitation policy and REC for the commission-driven objective. We set $\bar{\kappa}_u = 0.1$. (b) Difference between ATR($f^\star_\Pi$) and REC for user welfare with endogenous prices. We set $\beta = 1$. For both plots, we have $V_\mathsf{P} = 1, V_\mathsf{N} = 2, \lambda = 1, \eta = 0.1$.

(ii) ***High heterogeneity: conversion.*** *Let*

$$\Gamma^{\mathsf{cap}}_\infty(f) := R\, sur^{\mathsf{cap}}(f)\, A_\infty(f), \qquad f \in \Big[\frac{1}{n+1}, 1\Big].$$

As $\lambda \to \infty$, any conversion-optimal design either coincides with REC *or uses $f \geq \underline{f}_\infty$ and follows niche signals. Moreover:*

(a) If $\max_{f \in [\underline{f}_\infty, 1]} \Gamma^{\mathsf{cap}}_\infty(f) \leq R\sigma_\mathsf{P}$, then REC *is conversion-optimal for all sufficiently large λ.*

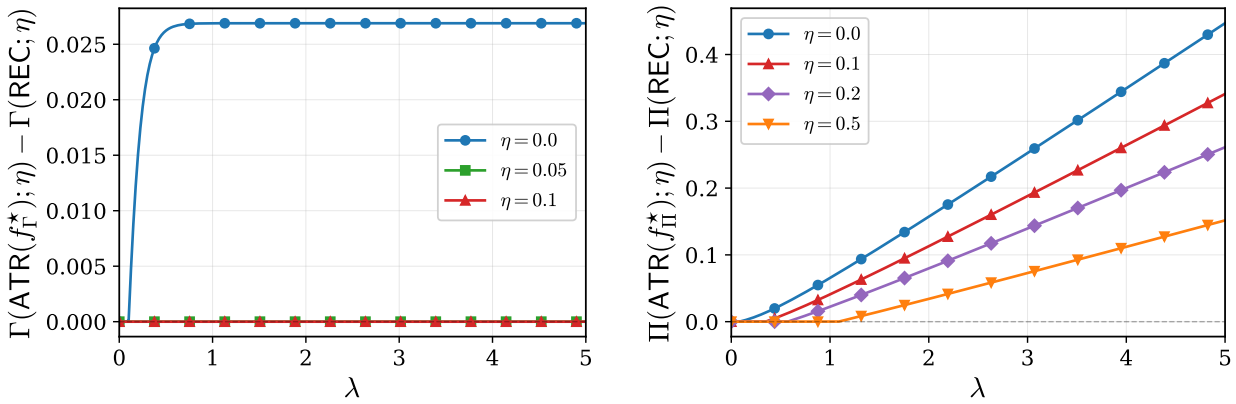
(b) If $\max_{f \in [\underline{f}_\infty, 1]} \Gamma^{\mathsf{cap}}_\infty(f) > R\sigma_\mathsf{P}$, then elicitation is conversion-optimal for all sufficiently large λ, and any sequence of optimal fidelities has limit points in $\arg\max_{f \in [\underline{f}_\infty, 1]} \Gamma^{\mathsf{cap}}_\infty(f)$.

(iii) ***High heterogeneity: commission (limiting fidelity).*** *There exists $\bar{\lambda} < \infty$ such that for all $\lambda \geq \bar{\lambda}$, the commission-optimal CRS elicits. Moreover, any sequence of commission-optimal fidelities has limit points in $\arg\max_{f \in [1/(n+1), 1]} sur^{\mathsf{cap}}(f)\, f$, and this maximizer is unique; denote it by f^{cap}_∞. It is characterized by the first-order condition*

$$\frac{1}{f^{\mathsf{cap}}_\infty} = \eta\, C'_{n+1}(f^{\mathsf{cap}}_\infty) = \eta \ln\Big(\frac{n f^{\mathsf{cap}}_\infty}{1 - f^{\mathsf{cap}}_\infty}\Big).$$

PROPOSITION 8 **(Variety under capacity: large n pushes toward recommend-first).** *Fix primitives $(\alpha, \lambda, \beta, V_\mathsf{P}, V_\mathsf{N}, R, \tau, \eta, \bar{\kappa}_u)$ with $\alpha \in (0,1)$ and $V_\mathsf{P} > 0$. Consider capacity-based elicitation burden $\mathcal{I}^{\mathsf{cap}}(f) = C_{n+1}(f)$ and survival $sur^{\mathsf{cap}}(f) = \exp(-\eta C_{n+1}(f))$. There exists $\bar{n} < \infty$ such that for all $n \geq \bar{n}$, the recommend-first policy* REC *(i.e., $f = 1/(n+1)$ and always recommend* P*) is optimal under (i) the conversion objective Γ (fixed prices), (ii) the commission objective Π^{com} (endogenous prices); and (iii) user welfare.*

Discussion. Variety increases the resolution burden of conversation: distinguishing among n niche labels is inherently a $\log(n)$-bit task. Under capacity-based abandonment, this logic is literal: $C_{n+1}(1) = \ln(n+1)$ and therefore survival under deep elicitation scales as $(n+1)^{-\eta}$. This makes high-variety categories especially hostile to deep elicitation unless match values are extreme.



(a) Value of elicitation for conversion objective

(b) Value of elicitation for commission objective

Figure 13 Difference between the optimal elicitation policy and REC for (a) conversion-driven objective and (b) commission-based objective. For both plots, we have $\alpha=0.1, V_{\mathsf{P}}=1, V_{\mathsf{N}}=2, \beta=1, R=1, \tau=0.3$.

Appendix C: Proofs of results in Appendix B

Throughout, fix $n\geq 1$ and primitives $(V_{\mathsf{P}}, V_{\mathsf{N}}, \lambda, \beta, \eta, \bar{\kappa}_u)$. Under capacity, survival is $sur^{\mathsf{cap}}(f) = \exp(-\eta C_{n+1}(f))$ where

$$C_{n+1}(f)=\ln(n+1)+f\ln f+(1-f)\ln\Big(\frac{1-f}{n}\Big).$$

Note $C_{n+1}(1/(n+1))=0$ and $C_{n+1}(1)=\ln(n+1)$. Define $u_{\mathsf{P}}=V_{\mathsf{P}}$, $u_-=-\lambda$, $u_+(\alpha)=\lambda V_{\mathsf{N}}\alpha^{-\beta}$, and $g(x)=\ln(1+e^x)$. For fixed prices, welfare is

$$W^{\mathsf{cap}}(f,\pi;\alpha)=sur^{\mathsf{cap}}(f)\cdot\mathbb{E}_\alpha\Big[g\big(u_{\mathsf{B}}(\Theta,\pi(S))\big)\Big]-\frac{\bar{\kappa}_u}{\eta}\big(1-sur^{\mathsf{cap}}(f)\big), \tag{84}$$

where $\mathbb{E}_\alpha[\cdot]$ is expectation under the type prior parameter α and the signal channel.

C.1. A useful asymptotic: when does welfare act on niche signals?

Let $q(f,\alpha,n)=\Pr(\Theta=s\mid S=s)$ be the posterior correctness after a niche signal (Lemma 1). Under the welfare objective, conditional on a niche signal $S=s$ the platform recommends niche s iff $q(f,\alpha,n)\geq q_W^\star(\alpha,\lambda)$ where

$$q_W^\star(\alpha,\lambda)=\frac{g(u_{\mathsf{P}})-g(u_-)}{g(u_+(\alpha))-g(u_-)}.$$

For small α, $u_+(\alpha)\to\infty$ whenever $\beta>0$, and hence $g(u_+(\alpha))=u_+(\alpha)+o(1)$. Thus

$$q_W^\star(\alpha,\lambda)=\frac{g(u_{\mathsf{P}})-g(u_-)}{\lambda V_{\mathsf{N}}}\cdot\alpha^\beta\ +\ o(\alpha^\beta). \tag{85}$$

On the other hand, for fixed $f<1$ and small α,

$$q(f,\alpha,n)=\frac{\alpha f}{\alpha f+(1-\alpha/n)(1-f)}=\frac{\alpha f}{1-f}+o(\alpha). \tag{86}$$

Combining (85)–(86) implies:

LEMMA 11 **(Near-perfect requirement when $\beta<1$; no such requirement when $\beta>1$).** *Fix $n\geq 1$ and any sequence $\alpha\to 0$.*

(i) If $\beta < 1$***,*** *then any sequence of fidelities* $f(\alpha)$ *such that* $q(f(\alpha), \alpha, n) \geq q^\star_W(\alpha, \lambda)$ *must satisfy*

$$1 - f(\alpha) = O(\alpha^{1-\beta}). \tag{87}$$

(ii) If $\beta > 1$***,*** *then for any fixed* $f > 1/(n+1)$*, the inequality* $q(f, \alpha, n) \geq q^\star_W(\alpha, \lambda)$ *holds for all sufficiently small* α*.*

Proof of Lemma 11: **Part (i)** Suppose $q(f, \alpha, n) \geq q^\star_W(\alpha, \lambda)$. Using (85), for sufficiently small α we have $q^\star_W(\alpha, \lambda) \geq c\,\alpha^\beta$ for some constant $c > 0$. Using the exact posterior formula (Lemma 1),

$$q(f, \alpha, n) = \frac{\alpha f}{\alpha f + (1 - \alpha/n)(1 - f)} \leq \frac{\alpha f}{(1 - \alpha/n)(1 - f)} \leq \frac{2\alpha}{1 - f}$$

for all small α (since $f \leq 1$ and $1 - \alpha/n \geq 1/2$). Thus $2\alpha/(1-f) \geq c\alpha^\beta$, i.e. $1 - f \leq (2/c)\alpha^{1-\beta}$, proving (87).

Part (ii) If $\beta > 1$, then $q^\star_W(\alpha, \lambda) = O(\alpha^\beta) = o(\alpha)$ by (85). Meanwhile for any fixed $f > 1/(n+1)$, (86) gives $q(f, \alpha, n) \sim (\alpha f)/(1 - f)$, which is of order α. Hence $q(f, \alpha, n) \geq q^\star_W(\alpha, \lambda)$ for all sufficiently small α. □

C.2. Proof of Theorem 3

Proof of Theorem 3: Under REC, survival is one and the platform recommends P, yielding revenue $\Gamma_{\mathsf{REC}} = R\sigma(u_{\mathsf{P}})$.

Fix any elicitation design (f, π) with $f > 1/(n+1)$. Then $sur^{\mathsf{cap}}(f) < 1$ is a constant independent of α. Adoption probabilities are bounded by one, and the maximal incremental adoption uplift from perfectly serving niche users is at most $\alpha(1 - \sigma(u_{\mathsf{P}}))$. Thus, uniformly over all designs,

$$\Gamma(f, \pi) \leq sur^{\mathsf{cap}}(f)\, R\Big(\sigma(u_{\mathsf{P}}) + \alpha(1 - \sigma(u_{\mathsf{P}}))\Big).$$

Hence

$$\Gamma_{\mathsf{REC}} - \Gamma(f, \pi) \geq R\sigma(u_{\mathsf{P}})(1 - sur^{\mathsf{cap}}(f)) - sur^{\mathsf{cap}}(f) R\alpha(1 - \sigma(u_{\mathsf{P}})).$$

The first term is a strictly positive constant (independent of α), while the second term is $O(\alpha)$. Therefore the difference is positive for sufficiently small α. Thus REC is conversion-optimal. □

C.3. Proof of Proposition 5

Proof of Proposition 5: This proof consists of 3 steps: 1) we establish a uniform upper bound on the expected utility, 2) we show that if no niche is ever recommended, REC is optimal, and 3) show when elicitation is welfare-optimal in each of the 3 regimes ($\beta < 1, \beta > 1, \beta = 1$).

Throughout, let $C(f) \equiv C_{n+1}(f)$ and $sur(f) \equiv sur^{\mathsf{cap}}(f) = \exp(-\eta C(f))$. Under fixed prices, welfare is

$$W^{\mathsf{cap}}(f, \pi; \alpha) = sur(f)\, \mathbb{E}_\alpha[g(u_{\mathsf{B}}(\Theta, \pi(S)))] - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur(f)\big),$$

where $\mathbb{E}_\alpha[\cdot]$ denotes expectation under the true prior with niche mass α and the channel with fidelity f.

Define $g_{\mathsf{P}} := g(u_{\mathsf{P}})$, $g_- := g(u_-)$, and $g_+(\alpha) := g(u_+(\alpha))$. Note that $g(\cdot)$ is increasing and $u_- \leq u_{\mathsf{P}}$, hence $g_- \leq g_{\mathsf{P}}$. Moreover, for $\beta > 0$, $g_+(\alpha) = u_+(\alpha) + o(1) = \lambda V_{\mathsf{N}} \alpha^{-\beta} + o(\alpha^{-\beta})$.

Step 1: A uniform upper bound. For any (f, π),

$$\mathbb{E}_\alpha[g(u_{\mathsf{B}}(\Theta, \pi(S)))] \leq (1 - \alpha)\, g_{\mathsf{P}} + \alpha\, g_+(\alpha) \leq g_{\mathsf{P}} + \alpha\, g_+(\alpha). \tag{88}$$

In particular, if $\beta < 1$ then $\alpha\, g_+(\alpha) = O(\alpha^{1-\beta}) \to 0$.

Step 2: if no niche is ever recommended, REC *is optimal.* Suppose $\pi(S) \equiv \mathsf{P}$ almost surely. Then $\mathbb{E}_\alpha[g(u_{\mathsf{B}}(\Theta, \pi(S)))] \equiv g_{\mathsf{P}}$ and

$$W^{\mathsf{cap}}(f, \pi; \alpha) = sur(f) g_{\mathsf{P}} - \frac{\bar{\kappa}_u}{\eta}(1 - sur(f)) \le g_{\mathsf{P}} = W^{\mathsf{cap}}(\mathsf{REC}),$$

with equality only if $sur(f) = 1$, i.e. $C(f) = 0$, i.e. $f = 1/(n+1)$ (the REC policy). Hence among designs that never recommend a niche, REC is optimal.

Step 3: any welfare-optimal design that recommends a niche must satisfy the welfare-optimal posterior threshold. Fix f. Since $sur(f)$ is constant across signals, conditional on a chosen f the welfare-maximizing recommendation rule is the posterior-threshold rule: after a niche signal $S = s$, recommending niche s is optimal iff $q(f, \alpha, n) \ge q^\star_W(\alpha, \lambda)$. Therefore, if a welfare-optimal design recommends a niche with positive probability, it must do so after some niche signal, and hence must satisfy

$$q(f, \alpha, n) \ge q^\star_W(\alpha, \lambda). \tag{89}$$

Part (i): $\beta < 1$ implies REC for all sufficiently small α. Consider any welfare-optimal design $(f(\alpha), \pi(\alpha))$.

If $\pi(\alpha)$ never recommends a niche for small α, Step 2 gives REC.

Otherwise, along a sequence $\alpha \to 0$ it recommends a niche with positive probability, hence by (89) we must have $q(f(\alpha), \alpha, n) \ge q^\star_W(\alpha, \lambda)$. For $\beta > 0$, $q^\star_W(\alpha, \lambda) = (g_{\mathsf{P}} - g_-)/(g_+(\alpha) - g_-) = \Theta(\alpha^\beta)$. For $\beta = 0$, $q^\star_W$ is a strictly positive constant. Using the posterior formula $q(f, \alpha, n) = \frac{\alpha f}{\alpha f + (1 - \alpha/n)(1-f)}$ and the inequality

$$\frac{\alpha f}{\alpha f + (1 - \alpha/n)(1-f)} \le \frac{\alpha f}{(1 - \alpha/n)(1-f)},$$

it follows that $q(f(\alpha), \alpha, n) \ge q^\star_W(\alpha, \lambda)$ implies

$$\frac{f(\alpha)}{1 - f(\alpha)} \;\ge\; c\,\alpha^{\beta - 1} \qquad \text{for some } c > 0 \text{ and all sufficiently small } \alpha.$$

When $\beta < 1$, the right-hand side diverges, hence $f(\alpha) \to 1$ and moreover $1 - f(\alpha) = O(\alpha^{1-\beta})$.

Therefore $C(f(\alpha)) \to C(1) = \ln(n+1)$ and $sur(f(\alpha)) \to (n+1)^{-\eta} =: sur_1 \in (0, 1)$. Combining this with the upper bound (88) and $\alpha g_+(\alpha) = O(\alpha^{1-\beta}) \to 0$ yields

$$W^{\mathsf{cap}}(f(\alpha), \pi(\alpha); \alpha) \le sur(f(\alpha))\big(g_{\mathsf{P}} + o(1)\big) - \frac{\bar{\kappa}_u}{\eta}\big(1 - sur(f(\alpha))\big) \to sur_1 g_{\mathsf{P}} - \frac{\bar{\kappa}_u}{\eta}(1 - sur_1) < g_{\mathsf{P}} = W^{\mathsf{cap}}(\mathsf{REC}),$$

a contradiction to optimality. Hence for all sufficiently small α, REC is welfare-optimal.

Part (ii): $\beta > 1$ implies elicitation is welfare-improving for small α, and $f^\star_W(\alpha) \to f^{\mathsf{cap}}_\infty$. Fix any $f \in (1/(n+1), 1]$. When $\beta > 1$, $q^\star_W(\alpha, \lambda) = \Theta(\alpha^\beta)$ while for fixed $f < 1$, $q(f, \alpha, n) = \Theta(\alpha)$; hence $q(f, \alpha, n) \ge q^\star_W(\alpha, \lambda)$ for all sufficiently small α. Thus the welfare-optimal signal rule recommends niche s after any niche signal $S = s$.

Under this rule, a niche user is correctly matched with probability f, so

$$\mathbb{E}_\alpha[g(u_{\mathsf{B}}(\Theta, \pi(S)))] \ge \alpha f\, g_+(\alpha) + O(1),$$

where $O(1)$ collects bounded contributions from mainstream users and mismatches. Since $g_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{-\beta} + o(\alpha^{-\beta})$, we have $\alpha f g_+(\alpha) = f \lambda V_{\mathsf{N}} \alpha^{1-\beta}(1 + o(1)) \to +\infty$. Multiplying by the constant survival factor $sur(f) >$

0 implies $W^{\mathsf{cap}}(f,\pi;\alpha)\to+\infty$ as $\alpha\to 0$, whereas $W^{\mathsf{cap}}(\mathsf{REC})=g_{\mathsf{P}}$ is constant. Therefore elicitation strictly dominates REC for all sufficiently small α.

To identify the limiting optimal fidelity, note that

$$\frac{W^{\mathsf{cap}}(f,\pi;\alpha)}{\lambda V_{\mathsf{N}}\alpha^{1-\beta}} = sur(f)\,f + o(1) \qquad (\beta>1),$$

uniformly over $f\in[1/(n+1),1]$. Hence any welfare-maximizing sequence $f_W^\star(\alpha)$ must converge to the unique maximizer of $f\mapsto sur(f)f$. Uniqueness follows because $\ln(sur(f)f)=\ln f-\eta C(f)$ is strictly concave on $(1/(n+1),1)$: $\ln f$ is strictly concave and $C''(f)=\frac{1}{f}+\frac{1}{1-f}>0$ implies C is strictly convex. The unique maximizer f_∞^{cap} satisfies the first-order condition

$$\frac{d}{df}\big(\ln f-\eta C(f)\big)=0 \iff \frac{1}{f}=\eta C'(f) \iff \frac{1}{f}=\eta\ln\Big(\frac{nf}{1-f}\Big),$$

using $C'(f)=\ln\big(\frac{nf}{1-f}\big)$. This completes part (ii).

Part (iii): $\beta=1$ yields a cutoff in $\bar\kappa_u$ and an $O(1)$ limiting problem. Let $\beta=1$ so $u_+(\alpha)=L/\alpha$ with $L=\lambda V_{\mathsf{N}}$. Then $\alpha f g_+(\alpha)\to fL$ for each fixed f. If f is such that the welfare-threshold condition holds for all sufficiently small α, the optimal signal rule follows niche signals and

$$\mathbb{E}_\alpha[g(u_{\mathsf{B}}(\Theta,\pi(S)))]=g_-+fA+o(1), \qquad A:=g_{\mathsf{P}}+L-g_->0.$$

Consequently,

$$W^{\mathsf{cap}}(f,\pi;\alpha)=sur(f)\big(g_-+fA\big)-\frac{\bar\kappa_u}{\eta}(1-sur(f))+o(1)=:\overline{W}^{\mathsf{cap}}(f)+o(1).$$

If instead the welfare threshold fails, the optimal rule recommends P after a niche signal, in which case $W^{\mathsf{cap}}(f,\pi;\alpha)\le sur(f)g_{\mathsf{P}}-\frac{\bar\kappa_u}{\eta}(1-sur(f))\le g_{\mathsf{P}}$, with strict inequality whenever $f>1/(n+1)$. Hence, in the rare-niche limit, REC can be beaten only by some f for which the niche-following rule is optimal.

Rearranging $\overline{W}^{\mathsf{cap}}(f)\ge g_{\mathsf{P}}$ gives the equivalent condition

$$\bar\kappa_u\le\eta\cdot\frac{sur(f)(g_-+fA)-g_{\mathsf{P}}}{1-sur(f)}.$$

Define the cutoff

$$\bar\kappa_u^{\mathsf{cap}}:=\max\Big\{0,\ \eta\sup_{f\in[\underline{f}_W^{\mathsf{cap}},1]}\frac{sur(f)(g_-+fA)-g_{\mathsf{P}}}{1-sur(f)}\Big\}\in[0,\infty].$$

If $\bar\kappa_u<\bar\kappa_u^{\mathsf{cap}}$, there exists f with $\overline{W}^{\mathsf{cap}}(f)>g_{\mathsf{P}}$, hence for all sufficiently small α, elicitation is welfare-improving. If $\bar\kappa_u\ge\bar\kappa_u^{\mathsf{cap}}$, then $\overline{W}^{\mathsf{cap}}(f)\le g_{\mathsf{P}}$ for all f, hence REC is welfare-optimal for all sufficiently small α. Finally, whenever elicitation is optimal, uniform convergence of $W^{\mathsf{cap}}(f,\pi;\alpha)$ to $\overline{W}^{\mathsf{cap}}(f)$ on the compact interval $[\underline{f}_W^{\mathsf{cap}},1]$ and Berge's maximum theorem imply $f_W^\star(\alpha)\to\arg\max_f\overline{W}^{\mathsf{cap}}(f)$. □

C.4. Proof of Corollary 3

Proof of Corollary 3: By Theorem 3, the conversion-optimal policy is REC for all sufficiently small α. If Proposition 5 implies that welfare strictly prefers elicitation for small α (either $\beta>1$, or $\beta=1$ with $\bar\kappa_u<\bar\kappa_u^{\mathsf{cap}}$), then REC is strictly welfare-suboptimal.

For the magnitude, let $f_W^\star(\alpha)$ denote the welfare-optimal fidelity. Under $\beta>1$, the proof of Proposition 5 shows $W^{\mathsf{cap}}(f_W^\star(\alpha),\pi;\alpha)$ has leading term $sur^{\mathsf{cap}}(f_W^\star(\alpha))\cdot f_W^\star(\alpha)\lambda V_{\mathsf{N}}\alpha^{1-\beta}$, while $W^{\mathsf{cap}}(\mathsf{REC})=g(u_{\mathsf{P}})$ is constant, yielding the stated $\Omega(\cdot)$ wedge. Under $\beta=1$, the wedge converges to $\max_f\overline{W}^{\mathsf{cap}}(f)-g(u_{\mathsf{P}})>0$ whenever $\bar\kappa_u<\bar\kappa_u^{\mathsf{cap}}$. □

C.5. Proof of Theorem 4

LEMMA 12 **(Mixed logit monopoly when the high-type share vanishes and qu stays finite).** *Fix $u_- \in \mathbb{R}$. For $u > 0$ and $q \in (0,1]$, define*

$$R_{u,q}(p) := p\Big[q\,\sigma(u-p) + (1-q)\,\sigma(u_- - p)\Big], \qquad p \geq 0,$$

and let

$$M(u,q) := \max_{p \geq 0} R_{u,q}(p).$$

Suppose $u \to \infty$, $q \to 0$, and $qu \to \chi \in [0,\infty)$. Then

$$M(u,q) \;\longrightarrow\; \max\{r_m(u_-), \chi\}.$$

If $\chi > r_m(u_-)$ and $p^\star(u,q)$ denotes any maximizer of $R_{u,q}(\cdot)$, then

$$p^\star(u,q) = u - o(u), \qquad g\big(u - p^\star(u,q)\big) = o(u), \qquad g\big(u_- - p^\star(u,q)\big) = o(1).$$

Proof of Lemma 12: Let $r_- := r_m(u_-)$.

Step 1: Lower bounds. Choose the low-type monopoly price $p_- := p_m(u_-)$. Then

$$R_{u,q}(p_-) = q\,p_-\,\sigma(u - p_-) + (1-q)\,p_-\,\sigma(u_- - p_-).$$

Because $q \to 0$ and $u \to \infty$, the first term vanishes and the second converges to r_-. Hence

$$\liminf M(u,q) \;\geq\; r_-.$$

Now choose the high-type monopoly price $p_u := p_m(u)$. Since $p_u \to \infty$ and $r_m(u) \sim u$ (Lemma 8),

$$R_{u,q}(p_u) = q\,r_m(u) + (1-q)\,p_u\,\sigma(u_- - p_u) \longrightarrow \chi.$$

Therefore

$$\liminf M(u,q) \;\geq\; \max\{r_-, \chi\}.$$

Step 2: upper bound. Fix $\varepsilon > 0$. Choose $P < \infty$ such that

$$\sup_{p \geq P} p\,\sigma(u_- - p) \leq \varepsilon.$$

If $p \leq P$, then

$$R_{u,q}(p) \leq qP + p\,\sigma(u_- - p) \leq qP + r_-.$$

Since $q \to 0$, for all sufficiently large u this is at most $r_- + \varepsilon$.

If $p \geq P$, then

$$R_{u,q}(p) \leq q\,r_m(u) + p\,\sigma(u_- - p) \leq q\,r_m(u) + \varepsilon.$$

Because $r_m(u)/u \to 1$ and $qu \to \chi$, for all sufficiently large u this is at most $\chi + 2\varepsilon$. Hence, for all large u,

$$M(u,q) \leq \max\{r_- + \varepsilon, \chi + 2\varepsilon\}.$$

Letting $\varepsilon \downarrow 0$ proves

$$M(u,q) \to \max\{r_-, \chi\}.$$

Step 3: the refinement when $\chi > r_-$. Assume now $\chi > r_-$, and fix $\varepsilon > 0$ sufficiently small that

$$r_- + 3\varepsilon < \chi - \varepsilon.$$

By Step 1, for all sufficiently large u,

$$M(u,q) \geq \chi - \varepsilon.$$

Take P as above with $\sup_{p \geq P} p\sigma(u_- - p) \leq \varepsilon$. If $p \leq P$, then for all large u,

$$R_{u,q}(p) \leq r_- + \varepsilon < \chi - \varepsilon,$$

so any maximizer $p^\star(u,q)$ must satisfy $p^\star(u,q) \geq P$. As P is arbitrary, this implies $p^\star(u,q) \to \infty$, and therefore

$$g\big(u_- - p^\star(u,q)\big) = o(1).$$

Next fix $\delta \in (0,1)$. For any $p \in [P, (1-\delta)u]$,

$$R_{u,q}(p) \leq q(1-\delta)u + \varepsilon.$$

Since $qu \to \chi$, for all large u the right-hand side is at most $(1-\delta)(\chi+\varepsilon)+\varepsilon$, which is strictly below $\chi - \varepsilon$ for ε small enough. Thus every maximizer satisfies

$$p^\star(u,q) > (1-\delta)u$$

for all sufficiently large u.

Finally, for any fixed $\delta > 0$ and any $p \geq (1+\delta)u$,

$$q\,p\,\sigma(u-p) \leq q\,p\,e^{u-p} \leq q(1+\delta)u\,e^{-\delta u} \to 0,$$

and because $p \to \infty$ the low-type term also tends to zero. Hence no maximizer can satisfy $p^\star(u,q) \geq (1+\delta)u$ for all large u. Therefore

$$p^\star(u,q) = u - o(u).$$

Since $u - p^\star(u,q) = o(u)$, we obtain

$$g\big(u - p^\star(u,q)\big) = o(u).$$

This proves the refinement. □

LEMMA 13 **(Total niche revenue when the high-type mass vanishes).** *Fix $u_- \in \mathbb{R}$. For $u > 0$ and $a, b \geq 0$, define*

$$\widehat{M}(u,a,b) := \max_{p \geq 0} p\Big[a\,\sigma(u-p) + b\,\sigma(u_- - p)\Big].$$

Suppose $u \to \infty$, $a \to 0$, $b \to \bar{b} \in [0,\infty)$, and

$$a\,u \to A \in [0,\infty).$$

Then

$$\widehat{M}(u,a,b) \;\longrightarrow\; \max\{\bar{b}\,r_m(u_-), A\}.$$

If moreover $A > \bar{b}\,r_m(u_-)$ and $p^\star(u,a,b)$ denotes any maximizer of the objective above, then

$$p^\star(u,a,b) = u - o(u), \qquad g\big(u - p^\star(u,a,b)\big) = o(u), \qquad g\big(u_- - p^\star(u,a,b)\big) = o(1).$$

Proof of Lemma 13: Let $r_- := r_m(u_-)$.

Case 1: $\bar{b} > 0$. For all sufficiently large u, we have $a + b > 0$. Define

$$q := \frac{a}{a+b}.$$

Then

$$\widehat{M}(u,a,b) = (a+b)\, M(u,q),$$

where $M(u,q)$ is the value function in Lemma 12. Because $a \to 0$ and $a + b \to \bar{b} > 0$, we have

$$q \to 0, \qquad q\,u = \frac{au}{a+b} \to \frac{A}{\bar{b}}.$$

Lemma 12 therefore yields

$$M(u,q) \to \max\left\{ r_-, \frac{A}{\bar{b}} \right\},$$

so

$$\widehat{M}(u,a,b) \to \bar{b} \max\left\{ r_-, \frac{A}{\bar{b}} \right\} = \max\{\bar{b}\, r_-, A\}.$$

If moreover $A > \bar{b}\, r_-$, then $A/\bar{b} > r_-$, and the refinement in Lemma 12 implies

$$p^\star(u,a,b) = u - o(u), \qquad g\big(u - p^\star(u,a,b)\big) = o(u), \qquad g\big(u_- - p^\star(u,a,b)\big) = o(1).$$

Case 2: $\bar{b} = 0$. Then $b \to 0$ and $a\,u \to A$. Choose the high-type monopoly price $p_m(u)$. Since $r_m(u) \sim u$,

$$\widehat{M}(u,a,b) \geq a\, r_m(u) + b\, p_m(u) \sigma(u_- - p_m(u)) \to A.$$

On the other hand, for every $p \geq 0$,

$$p\Big[a\,\sigma(u-p) + b\,\sigma(u_- - p)\Big] \leq a\, r_m(u) + b\, r_-,$$

so

$$\widehat{M}(u,a,b) \leq a\, r_m(u) + b\, r_- \to A.$$

Hence $\widehat{M}(u,a,b) \to A = \max\{0, A\}$.

If moreover $A > 0$, fix $\varepsilon \in (0, A/4)$. Because $\widehat{M}(u,a,b) \to A$, we have $\widehat{M}(u,a,b) \geq A - \varepsilon$ for all sufficiently large u. Choose $P < \infty$ so large that $\sup_{p \geq P} p\sigma(u_- - p) \leq \varepsilon$. If $p \leq P$, then

$$p\Big[a\,\sigma(u-p) + b\,\sigma(u_- - p)\Big] \leq aP + b\, r_- \to 0,$$

so for all sufficiently large u every maximizer satisfies $p^\star(u,a,b) \geq P$. As P is arbitrary, $p^\star(u,a,b) \to \infty$, hence

$$g\big(u_- - p^\star(u,a,b)\big) = o(1).$$

Next fix $\delta \in (0,1)$. For any $p \in [P, (1-\delta)u]$,

$$p\Big[a\,\sigma(u-p) + b\,\sigma(u_- - p)\Big] \leq a(1-\delta)u + \varepsilon \to (1-\delta)A + \varepsilon < A - \varepsilon$$

for all sufficiently large u, after choosing ε small enough. Thus every maximizer satisfies

$$p^\star(u,a,b) > (1-\delta)u$$

for all sufficiently large u.

Finally, for any fixed $\delta > 0$ and any $p \geq (1+\delta)u$,

$$a\, p\, \sigma(u-p) \leq a\, p\, e^{u-p} \leq a(1+\delta)u\, e^{-\delta u} \to 0,$$

and the low-type term also tends to zero because $b \to 0$ and $p \to \infty$. Hence no maximizer can satisfy $p^\star(u,a,b) \geq (1+\delta)u$ for all large u. Therefore

$$p^\star(u,a,b) = u - o(u), \qquad g\big(u - p^\star(u,a,b)\big) = o(u).$$

This completes the proof. □

Proof of Theorem 4: We prove each part individually after establishing properties of monopoly revenue and deriving an upper bound on total seller revenue. Throughout, let $C(f) := C_{n+1}(f)$ and $sur(f) := sur^{\mathsf{cap}}(f) = \exp(-\eta C(f))$.

Step 1: monopoly revenue facts. Let $r_m(u) := \max_{p \geq 0} p\,\sigma(u-p)$ and let $p_m(u)$ denote the unique maximizer. A standard calculation yields the first-order condition

$$1 = \frac{p}{1+e^{u-p}} \quad \Longleftrightarrow \quad e^{u-p} = p - 1,$$

so $p_m(u) = 1 + W(e^{u-1})$ and $r_m(u) = p_m(u)\sigma(u - p_m(u)) = p_m(u) - 1 = W(e^{u-1})$. In particular, $r_m(\cdot)$ is strictly increasing and $r_m(u) \sim u$ as $u \to \infty$. Also $r_m(u) \leq 1 + u$ for all u. We will use these properties repeatedly.

Under REC, the platform recommends P without elicitation, so $f = 1/(n+1)$, $C(f) = 0$, $sur(f) = 1$, and the P-seller posts $p_m(u_{\mathsf{P}})$ yielding seller revenue $r_{\mathsf{P}} := r_m(u_{\mathsf{P}})$. Hence

$$\Pi^{\mathsf{cap}}(\mathsf{REC}) = \tau\, r_{\mathsf{P}}. \tag{90}$$

Step 2: A useful upper bound on total seller revenue. Fix any CRS design (f,π) (and the induced equilibrium prices $p^\star(f,\pi)$). For any realized session that reaches the pricing stage, the seller revenue from the purchased item is always bounded above by $r_m(u_+(\alpha))$ if a *correctly matched* niche is recommended, and by r_{P} otherwise. Therefore, letting TP denote the event that a niche user is recommended her correctly matched niche,

$$\text{(seller revenue per arriving user)} \;\leq\; r_{\mathsf{P}} + \Pr(\mathsf{TP}) \cdot r_m(u_+(\alpha)). \tag{91}$$

Multiplying by survival $sur(f) \leq 1$ and by τ yields the bound

$$\Pi^{\mathsf{cap}}(f,\pi) \;\leq\; \tau \cdot sur(f)\Big(r_{\mathsf{P}} + \Pr(\mathsf{TP})\, r_m(u_+(\alpha))\Big). \tag{92}$$

Part (i): $\beta < 1$ implies REC for small α. We compare any CRS design to REC.

Case 1: the CRS never recommends a niche. Then TP never occurs, so by (92),

$$\Pi^{\mathsf{cap}}(f,\pi) \leq \tau \cdot sur(f)\, r_{\mathsf{P}} \leq \tau r_{\mathsf{P}} = \Pi^{\mathsf{cap}}(\mathsf{REC}),$$

with strict inequality whenever $f > 1/(n+1)$ because then $C(f) > 0$ and hence $sur(f) < 1$.

Case 2: the CRS recommends a niche with positive probability. At a fixed fidelity f, after any niche signal the only relevant actions are recommending P or recommending the indicated niche. By symmetry across

niche signals, the same action is taken after every niche signal. Thus if the CRS recommends a niche with positive probability, it follows every niche signal.

In any session, a correctly matched niche purchase can occur only for niche users, which have total mass α. Hence $\Pr(\mathsf{TP}) = \alpha f$ under the follow-niche-signals rule, and (92) implies

$$\Pi^{\mathsf{cap}}(f,\pi) \le \tau \cdot sur(f)\Big(r_{\mathsf{P}} + \alpha f\, r_m(u_+(\alpha))\Big).$$

Because $\beta < 1$ and $u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{-\beta} \to \infty$, we have $r_m(u_+(\alpha)) = O(u_+(\alpha)) = O(\alpha^{-\beta})$ and therefore

$$\alpha\, r_m(u_+(\alpha)) = O(\alpha^{1-\beta}) \longrightarrow 0.$$

Moreover, if $f > 1/(n+1)$ then $sur(f) < 1$ is bounded away from 1 on any set $f \ge 1-\delta$, and if $f \le 1-\delta$ then following niche signals routes a positive mass of low-value traffic to mismatched niches. Formally, fix any $\delta \in (0,1)$ and split $f \in [1/(n+1), 1]$ into $f \le 1-\delta$ and $f > 1-\delta$:

- If $f > 1-\delta$, then $sur(f) \le sur(1-\delta) < 1$, so for all sufficiently small α,

$$\tau \cdot sur(f)\Big(r_{\mathsf{P}} + \alpha r_m(u_+(\alpha))\Big) \le \tau\, sur(1-\delta)\Big(r_{\mathsf{P}} + o(1)\Big) < \tau r_{\mathsf{P}}.$$

- If $f \le 1-\delta$, then under the follow-niche-signals rule the popular seller receives mass

$$m_{\mathsf{P},\alpha} = (1-\alpha)f + \alpha\frac{1-f}{n},$$

the total mass of correctly matched niche sessions is αf, and the remaining niche-routed sessions all have utility u_-. Hence the total mass of mismatched niche sessions is

$$\ell_\alpha := 1 - m_{\mathsf{P},\alpha} - \alpha f = (1-f)\Big(1 - \frac{\alpha}{n}\Big).$$

Every mismatched niche session yields seller revenue at most $r_- := r_m(u_-)$, while every correctly matched niche session yields seller revenue at most $r_m(u_+(\alpha))$. Therefore seller revenue per arriving user is bounded above by

$$m_{\mathsf{P},\alpha} r_{\mathsf{P}} + \ell_\alpha r_- + \alpha f\, r_m(u_+(\alpha)) = r_{\mathsf{P}} - \ell_\alpha(r_{\mathsf{P}} - r_-) + \alpha f\, r_m(u_+(\alpha)).$$

Because $1 - f \ge \delta$, we have $\ell_\alpha \ge \delta(1 - \alpha/n)$, so for all sufficiently small α,

$$\ell_\alpha(r_{\mathsf{P}} - r_-) \ge \frac{\delta}{2}(r_{\mathsf{P}} - r_-).$$

Since $\alpha f\, r_m(u_+(\alpha)) \le \alpha r_m(u_+(\alpha)) = o(1)$, it follows that

$$\Pi^{\mathsf{cap}}(f,\pi) \le \tau\Big(r_{\mathsf{P}} - \frac{\delta}{2}(r_{\mathsf{P}} - r_-) + o(1)\Big) < \tau r_{\mathsf{P}}$$

for all sufficiently small α.

Thus in both subcases $\Pi^{\mathsf{cap}}(f,\pi) < \Pi^{\mathsf{cap}}(\mathsf{REC})$ for all sufficiently small α. Combining Case 1 and Case 2 yields the claim.

Part (ii): $\beta > 1$ implies elicitation is optimal and $f^\star_\Pi(\alpha) \to f^{\mathsf{cap}}_\infty$. Fix any $f \in [1/(n+1), 1]$ and consider the ask-then-recommend policy that (i) recommends P after $S = \emptyset$ and (ii) recommends niche i after $S = i$. Under this policy, the event TP occurs whenever $\Theta = i$ and $S = i$, so $\Pr(\mathsf{TP}) = \alpha f$. If each niche seller were to post the monopoly price $p_m(u_+(\alpha))$, each TP-session would generate revenue $r_m(u_+(\alpha))$. Since sellers best

respond, their equilibrium revenue is *at least* this amount from the TP-sessions. Therefore the platform can guarantee commission profit

$$\Pi^{\mathsf{cap}}(f,\pi) \;\geq\; \tau \cdot sur(f) \cdot \alpha f \cdot r_m(u_+(\alpha)). \tag{93}$$

Because $\beta > 1$, we have $\alpha r_m(u_+(\alpha)) \sim \alpha u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{1-\beta} \to \infty$, so (93) diverges for any fixed $f > 0$ with $sur(f) > 0$. In particular, for sufficiently small α, this exceeds $\Pi^{\mathsf{cap}}(\mathsf{REC}) = \tau r_{\mathsf{P}}$. Hence elicitation strictly dominates REC for all sufficiently small α.

To characterize the limiting optimal fidelity, combine the lower bound (93) with the upper bound (92). For any design (f,π), the probability of a correctly matched niche recommendation is at most αf under any policy that recommends niche i only after observing $S = i$. Thus $\Pr(\mathsf{TP}) \leq \alpha f$, and (92) yields

$$\Pi^{\mathsf{cap}}(f,\pi) \;\leq\; \tau \cdot sur(f) \Big(r_{\mathsf{P}} + \alpha f \, r_m(u_+(\alpha)) \Big).$$

Divide both sides by $\alpha r_m(u_+(\alpha))$ and use $\alpha r_m(u_+(\alpha)) \to \infty$ to obtain

$$\limsup_{\alpha \to 0} \; \frac{\Pi^{\mathsf{cap}}(f,\pi)}{\alpha r_m(u_+(\alpha))} \;\leq\; \tau \cdot sur(f)\, f.$$

On the other hand, (93) implies for the follow-signal policy,

$$\liminf_{\alpha \to 0} \; \frac{\Pi^{\mathsf{cap}}(f,\pi)}{\alpha r_m(u_+(\alpha))} \;\geq\; \tau \cdot sur(f)\, f.$$

Hence the asymptotic maximization problem reduces to maximizing $sur(f)f$ over $f \in [1/(n+1), 1]$. Because $\ln(sur(f)f) = \ln f - \eta C(f)$ is strictly concave in f, the maximizer is unique; denote it by f_∞^{cap}. Moreover,

$$\frac{d}{df}\Big(\ln f - \eta C(f)\Big)\Big|_{f=1/(n+1)} = n+1 > 0, \qquad \frac{d}{df}\Big(\ln f - \eta C(f)\Big) \longrightarrow -\infty \quad \text{as } f \uparrow 1,$$

so the unique maximizer lies in the interior:

$$f_\infty^{\mathsf{cap}} \in \Big(\frac{1}{n+1}, 1\Big).$$

Standard argmax-continuity then implies $f_\Pi^\star(\alpha) \to f_\infty^{\mathsf{cap}}$.

Part (iii): $\beta = 1$ and the optimal-fidelity comparison. Let $\beta = 1$, so $u_+(\alpha) = L/\alpha$ with $L = \lambda V_{\mathsf{N}}$, and let

$$r_- := r_m(u_-).$$

For any fidelity f, the optimal recommendation rule either recommends P after every niche signal or recommends the indicated niche after every niche signal. Indeed, after a niche signal the only candidates are P and the indicated niche, and by symmetry the same action is taken after every niche signal.

If the platform recommends P after every niche signal, then its commission payoff is

$$\tau \, sur(f) \, r_{\mathsf{P}} \leq \tau r_{\mathsf{P}},$$

with strict inequality whenever $f > 1/(n+1)$. Hence such a policy can never asymptotically beat REC.

It therefore remains to analyze the follow-niche-signals rule. Let $\alpha_k \downarrow 0$ be any sequence and let $f_k \in [1/(n+1), 1]$ satisfy

$$f_k \to \bar{f} \in [1/(n+1), 1].$$

Under the follow-niche-signals rule at (α_k, f_k), define

$$m_{\mathsf{P},k} := (1-\alpha_k)f_k + \alpha_k \frac{1-f_k}{n}, \qquad a_k := \alpha_k f_k, \qquad b_k := \Big(1 - \frac{\alpha_k}{n}\Big)(1-f_k).$$

The popular seller receives mass $m_{\mathsf{P},k}$ and therefore contributes

$$m_{\mathsf{P},k}\, r_{\mathsf{P}} \longrightarrow \bar{f}\, r_{\mathsf{P}}.$$

Each niche seller faces revenue

$$R_{\mathsf{N},k}(p) = \frac{1}{n}\, p\Big[a_k\, \sigma(u_+(\alpha_k) - p) + b_k\, \sigma(u_- - p)\Big],$$

so the *total* niche revenue across all niche sellers is exactly

$$\widehat{M}\big(u_+(\alpha_k), a_k, b_k\big),$$

where $\widehat{M}(\cdot,\cdot,\cdot)$ is the value function in Lemma 13. Now

$$a_k \to 0, \qquad b_k \to 1-\bar{f}, \qquad a_k\, u_+(\alpha_k) = \alpha_k f_k \cdot \frac{L}{\alpha_k} \to \bar{f}\, L.$$

Therefore Lemma 13 implies

$$\widehat{M}\big(u_+(\alpha_k), a_k, b_k\big) \longrightarrow \max\{(1-\bar{f})r_-, \bar{f}L\}.$$

Hence the platform payoff under the follow-niche-signals rule satisfies

$$\Pi_k^{\mathsf{follow}} \longrightarrow \Phi(\bar{f}) := \tau\, sur(\bar{f})\Big[\bar{f}\, r_{\mathsf{P}} + \max\{(1-\bar{f})r_-, \bar{f}L\}\Big].$$

This establishes the limiting objective for *every* convergent sequence $f_k \to \bar{f}$, including the case $\bar{f} = 1$ and sequences that approach 1 at an α-dependent rate.

Now consider the two branches in $\Phi(\bar{f})$. If

$$\bar{f}\, L \le (1-\bar{f})r_-,$$

then

$$\Phi(\bar{f}) \le \tau\, sur(\bar{f})\big(\bar{f}\, r_{\mathsf{P}} + (1-\bar{f})r_-\big) < \tau r_{\mathsf{P}},$$

because $r_- < r_{\mathsf{P}}$ and either $\bar{f} < 1$, in which case $\bar{f}\, r_{\mathsf{P}} + (1-\bar{f})r_- < r_{\mathsf{P}}$, or $\bar{f} = 1$, in which case the branch inequality forces $L = 0$ and $sur(1) < 1$. Therefore any asymptotically profitable elicitation must lie in the high-type branch

$$\bar{f}\, L > (1-\bar{f})r_-,$$

where

$$\Phi(\bar{f}) = \tau(r_{\mathsf{P}} + L)\, sur(\bar{f})\, \bar{f}.$$

Define

$$M^{\mathsf{cap}} := \max_{f \in [1/(n+1),1]} sur(f)\, f = sur(f_\infty^{\mathsf{cap}})\, f_\infty^{\mathsf{cap}}.$$

We now split cases.

Let $\alpha_k \downarrow 0$ be any sequence and let (f_k, π_k) be commission-optimal designs at α_k. By compactness, after passing to a subsequence we may assume $f_k \to \bar{f} \in [1/(n+1), 1]$. If π_k ignores niche signals for infinitely many k, then

$$\Pi^{\mathsf{cap}}(f_k, \pi_k) \le \tau\, sur(f_k)\, r_{\mathsf{P}},$$

so

$$\limsup_{k\to\infty} \Pi^{\mathsf{cap}}(f_k, \pi_k) \le \tau\, sur(\bar{f})\, r_{\mathsf{P}} \le \tau r_{\mathsf{P}}.$$

If instead π_k follows every niche signal for infinitely many k, then the same calculation as above applies with $f = f_k$ and yields

$$\lim_{k\to\infty} \Pi^{\mathsf{cap}}(f_k, \pi_k) = \Phi(\bar{f}).$$

Therefore every subsequential limit of optimal commission payoffs is bounded above by

$$\max\left\{\tau r_{\mathsf{P}}, \sup_{f\in[1/(n+1),1]} \Phi(f)\right\}.$$

Case 1: $(r_{\mathsf{P}} + L)M^{\mathsf{cap}} < r_{\mathsf{P}}$. For every f in the high-type branch,

$$\Phi(f) \le \tau(r_{\mathsf{P}} + L)M^{\mathsf{cap}} < \tau r_{\mathsf{P}}.$$

We already saw that the low-type branch also satisfies $\Phi(f) < \tau r_{\mathsf{P}}$ for every $f > 1/(n+1)$, while the ignore-signals branch yields at most τr_{P} with equality only at $f = 1/(n+1)$, i.e., under REC. Hence no eliciting design can be optimal for all sufficiently small α, so REC is commission-optimal for all sufficiently small α. In particular,

$$\Pi^{\mathsf{cap}}\big(\mathsf{ATR}(f^\star_\Pi(\alpha))\big) - \Pi^{\mathsf{cap}}(\mathsf{REC}) \longrightarrow 0.$$

Case 2: $(r_{\mathsf{P}} + L)M^{\mathsf{cap}} > r_{\mathsf{P}}$. Because $sur(f) \le 1$, we have

$$M^{\mathsf{cap}} \le f^{\mathsf{cap}}_\infty.$$

The strict inequality $(r_{\mathsf{P}} + L)M^{\mathsf{cap}} > r_{\mathsf{P}}$ is equivalent to

$$\frac{r_{\mathsf{P}}}{r_{\mathsf{P}} + L} < M^{\mathsf{cap}}.$$

Since $r_- < r_{\mathsf{P}}$, this gives

$$\frac{r_-}{r_- + L} < \frac{r_{\mathsf{P}}}{r_{\mathsf{P}} + L} < M^{\mathsf{cap}} \le f^{\mathsf{cap}}_\infty,$$

and therefore

$$f^{\mathsf{cap}}_\infty L > (1 - f^{\mathsf{cap}}_\infty) r_-.$$

So f^{cap}_∞ lies in the high-type branch, and

$$\Phi(f^{\mathsf{cap}}_\infty) = \tau(r_{\mathsf{P}} + L)M^{\mathsf{cap}} > \tau r_{\mathsf{P}}.$$

The feasible follow-niche-signals rule at fidelity f^{cap}_∞ therefore has asymptotic payoff strictly above τr_{P}, so elicitation strictly dominates REC for all sufficiently small α. Moreover, any sequence of commission-optimal designs must have all of its limit points in the high-type branch, because the low-type branch and the

ignore-signals branch are asymptotically below τr_{P}. On the high-type branch the fixed-f limiting objective is exactly

$$\tau(r_{\mathsf{P}} + L)\, sur(f)\, f,$$

whose unique maximizer is $f_{\infty}^{\mathsf{cap}}$. Therefore every subsequential limit of $\{f_{\Pi}^{\star}(\alpha)\}$ equals $f_{\infty}^{\mathsf{cap}}$, and hence

$$f_{\Pi}^{\star}(\alpha) \longrightarrow f_{\infty}^{\mathsf{cap}}.$$

Combining the upper bound obtained from the subsequence argument with the feasible lower bound from the follow-niche-signals rule at $f = f_{\infty}^{\mathsf{cap}}$ yields

$$\Pi^{\mathsf{cap}}\big(\mathsf{ATR}(f_{\Pi}^{\star}(\alpha))\big) - \Pi^{\mathsf{cap}}(\mathsf{REC}) \longrightarrow \tau\big((r_{\mathsf{P}} + L)M^{\mathsf{cap}} - r_{\mathsf{P}}\big).$$

Case 3: $(r_{\mathsf{P}} + L)M^{\mathsf{cap}} = r_{\mathsf{P}}$. The same upper bound shows that no eliciting design can have asymptotic payoff strictly above τr_{P}, while REC attains τr_{P} exactly. Hence

$$\Pi^{\mathsf{cap}}\big(\mathsf{ATR}(f_{\Pi}^{\star}(\alpha))\big) - \Pi^{\mathsf{cap}}(\mathsf{REC}) \longrightarrow 0.$$

This proves part (iii). □

C.6. Proof of Proposition 6

LEMMA 14 **(Mixed logit monopoly when the high-type share vanishes but $qu \to \infty$).** *Fix $u_- \in \mathbb{R}$. For $u > 0$ and $q \in (0,1]$, define*

$$R_{u,q}(p) := p\Big[q\,\sigma(u-p) + (1-q)\,\sigma(u_- - p)\Big], \qquad p \geq 0,$$

and let $p^{\star}(u,q)$ denote the unique maximizer of $R_{u,q}(\cdot)$.

Suppose $u \to \infty$, $q \to 0$, and $qu \to \infty$. Then

$$p^{\star}(u,q) = p_m(u) + O(1).$$

Consequently,

$$g\big(u - p^{\star}(u,q)\big) = \ln u + O(1), \qquad g\big(u_- - p^{\star}(u,q)\big) = o(1).$$

Proof of Lemma 14: Write

$$H_u(p) := p\,\sigma(u-p), \qquad L(p) := p\,\sigma(u_- - p),$$

so that

$$R_{u,q}(p) = qH_u(p) + (1-q)L(p).$$

By Lemma 8, both H_u and L are strictly concave on $[0,\infty)$. Hence $R_{u,q}$ is strictly concave and has a unique maximizer.

Let $p_m(u)$ be the unique maximizer of H_u. Fix any $M > 0$. Using the first-order condition

$$e^{u - p_m(u)} = p_m(u) - 1,$$

a direct calculation shows that for every fixed $z \in \mathbb{R}$,

$$H_u'\big(p_m(u) + z\big) = \frac{(p_m(u)-1)e^{-z}}{1 + (p_m(u)-1)e^{-z}} \left[1 - \frac{p_m(u) + z}{1 + (p_m(u)-1)e^{-z}}\right].$$

Since $p_m(u) \to \infty$ as $u \to \infty$, the right-hand side converges uniformly for $z \in [-M, M]$ to

$$1 - e^z.$$

Therefore there exist constants $c_M > 0$ and $u_0 < \infty$ such that for all $u \geq u_0$,

$$H'_u\big(p_m(u) - M\big) \geq c_M, \qquad H'_u\big(p_m(u) + M\big) \leq -c_M. \tag{94}$$

Next, for $p \in [p_m(u) - M,\ p_m(u) + M]$, Lemma 8 and (54) imply

$$p_m(u) = u - \ln u + O(1),$$

hence

$$p = u - \ln u + O(1).$$

Therefore

$$|L'(p)| \leq C(1+p)e^{u_- - p} \leq C' u^2 e^{-u} = o(q),$$

for some constants $C, C' < \infty$, because $qu \to \infty$ implies $q \geq 1/u$ for all sufficiently large u, and thus

$$\frac{u^2 e^{-u}}{q} \leq u^3 e^{-u} \to 0.$$

Combining this with (94) gives, for all sufficiently large u,

$$R'_{u,q}\big(p_m(u) - M\big) = q\, H'_u\big(p_m(u) - M\big) + (1-q)L'\big(p_m(u) - M\big) > 0,$$

and

$$R'_{u,q}\big(p_m(u) + M\big) = q\, H'_u\big(p_m(u) + M\big) + (1-q)L'\big(p_m(u) + M\big) < 0.$$

Since $R_{u,q}$ is strictly concave, its unique maximizer must lie in

$$[p_m(u) - M,\ p_m(u) + M].$$

Because M was arbitrary but fixed, this proves

$$p^\star(u, q) = p_m(u) + O(1).$$

Finally, Lemma 8 gives $p_m(u) = u - \ln u + O(1)$, so

$$u - p^\star(u,q) = \ln u + O(1) \implies g\big(u - p^\star(u,q)\big) = \ln u + O(1).$$

We also have

$$u_- - p^\star(u,q) \to -\infty \implies g\big(u_- - p^\star(u,q)\big) = o(1).$$

□

Proof of Proposition 6: **Part (i):** This follows directly from Theorem 4(i): if $\beta < 1$, then REC is commission-optimal for all sufficiently small α, so $\Delta W_{\Pi}^{\mathsf{com},\mathsf{cap}}(\alpha) = 0$ for all such α.

Part (ii). We now prove part (ii), fixing $\beta > 1$ and $\lambda > 0$.

Let

$$f_\alpha := f_\Pi^\star(\alpha), \qquad s_\alpha := sur^{\mathsf{cap}}(f_\alpha), \qquad a_\alpha := \alpha f_\alpha, \qquad b_\alpha := \Big(1-\frac{\alpha}{n}\Big)(1-f_\alpha), \qquad m_{\mathsf{P},\alpha} := (1-\alpha)f_\alpha + \alpha\frac{1-f_\alpha}{n}.$$

By Theorem 4(ii),

$$f_\alpha \to f_\infty^{\mathsf{cap}}, \qquad s_\alpha \to sur_\infty^{\mathsf{cap}} := sur^{\mathsf{cap}}(f_\infty^{\mathsf{cap}}).$$

By the first-order characterization in the proof of Theorem 4(ii),

$$f_\infty^{\mathsf{cap}} \in \Big(\frac{1}{n+1}, 1\Big).$$

Because $f_\alpha \to f_\infty^{\mathsf{cap}}$,

$$m_{\mathsf{P},\alpha} = f_\alpha + O(\alpha), \qquad a_\alpha = O(\alpha), \qquad b_\alpha = 1 - f_\alpha + O(\alpha). \tag{95}$$

Since Theorem 4(ii) also implies that for all sufficiently small α the commission-optimal design strictly outperforms REC, it cannot recommend P after every signal. After any niche signal the optimal action is either P or the indicated niche (see the logic of Lemma 2). Because the posterior correctness is the same for every niche signal, the same action is taken after every niche signal. Hence, for all sufficiently small α,

$$\pi_\Pi^\star(\emptyset) = \mathsf{P}, \qquad \pi_\Pi^\star(i) = i, \quad i = 1, \dots, n.$$

Let

$$p_{\mathsf{P}}^\star := p_m(u_{\mathsf{P}}), \qquad g_{\mathsf{P}}^{\mathrm{com}} := g(u_{\mathsf{P}} - p_{\mathsf{P}}^\star).$$

Under REC, the popular seller sets price $p_{\mathsf{P}}^\star$, so

$$W^{\mathsf{com},\mathsf{cap}}(\mathsf{REC}) = g_{\mathsf{P}}^{\mathrm{com}}.$$

Step 1: Pricing asymptotics. Because every user routed to P has deterministic utility u_{P}, the popular seller's revenue under the commission-optimal design is exactly

$$R_{\mathsf{P},\alpha}(p) = m_{\mathsf{P},\alpha}\, p\, \sigma(u_{\mathsf{P}} - p).$$

Therefore its equilibrium price is $p_{\mathsf{P},\alpha}^\star = p_{\mathsf{P}}^\star$ for every α. Hence

$$g(u_{\mathsf{P}} - p_{\mathsf{P},\alpha}^\star) = g_{\mathsf{P}}^{\mathrm{com}}. \tag{96}$$

Now define

$$q_\alpha := \frac{a_\alpha}{a_\alpha + b_\alpha}.$$

Since $a_\alpha \to 0$ and $b_\alpha \to 1 - f_\infty^{\mathsf{cap}} > 0$, we have $q_\alpha \to 0$. Additionally, $a_\alpha + b_\alpha \le 1$ and $f_\alpha \ge 1/(n+1)$ gives

$$q_\alpha \ge a_\alpha = \alpha f_\alpha \ge \frac{\alpha}{n+1}.$$

Because $\beta > 1$ and $u_+(\alpha) = \lambda V_{\mathsf{N}} \alpha^{-\beta}$,

$$q_\alpha \, u_+(\alpha) \geq \frac{\lambda V_{\mathsf{N}}}{n+1} \alpha^{1-\beta} \longrightarrow \infty.$$

For a niche seller, revenue is

$$R_{\mathsf{N},\alpha}(p) = \frac{a_\alpha + b_\alpha}{n} \, p \Big[q_\alpha \, \sigma(u_+(\alpha) - p) + (1 - q_\alpha) \, \sigma(u_- - p) \Big].$$

Multiplying by the positive constant $(a_\alpha + b_\alpha)/n$ does not affect the maximizer, so Lemma 14 applies with

$$u = u_+(\alpha), \qquad q = q_\alpha.$$

Hence

$$p^\star_{\mathsf{N},\alpha} = p_m(u_+(\alpha)) + O(1), \tag{97}$$

and

$$g(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}) = \ln u_+(\alpha) + O(1) = \beta \ln(1/\alpha) + O(1), \tag{98}$$

while

$$g(u_- - p^\star_{\mathsf{N},\alpha}) = o(1). \tag{99}$$

Step 2: Choice-stage expected surplus under the commission-optimal design. Conditional on reaching the recommendation stage, expected surplus under the commission-optimal design is

$$E_\alpha := m_{\mathsf{P},\alpha} \, g(u_{\mathsf{P}} - p^\star_{\mathsf{P},\alpha}) + a_\alpha \, g(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}) + b_\alpha \, g(u_- - p^\star_{\mathsf{N},\alpha}).$$

Using (95), (96), (98), and (99),

$$E_\alpha = \big(f_\alpha + O(\alpha)\big) g^{\mathrm{com}}_{\mathsf{P}} + O(\alpha \ln(1/\alpha)) + \big(1 - f_\alpha + O(\alpha)\big) o(1).$$

Because $\alpha \ln(1/\alpha) \to 0$, we obtain

$$E_\alpha = f_\alpha g^{\mathrm{com}}_{\mathsf{P}} + o(1). \tag{100}$$

Step 3: Welfare limit. Welfare under the commission-optimal design is

$$W^{\mathsf{com},\mathsf{cap}}_{\Pi}(\alpha) = s_\alpha E_\alpha - \frac{\bar{\kappa}_u}{\eta}(1 - s_\alpha).$$

Combining this with (100) yields

$$W^{\mathsf{com},\mathsf{cap}}_{\Pi}(\alpha) = s_\alpha f_\alpha g^{\mathrm{com}}_{\mathsf{P}} - \frac{\bar{\kappa}_u}{\eta}(1 - s_\alpha) + o(1).$$

Since $f_\alpha \to f^{\mathsf{cap}}_\infty$ and $s_\alpha \to sur^{\mathsf{cap}}_\infty$,

$$\lim_{\alpha \to 0} W^{\mathsf{com},\mathsf{cap}}_{\Pi}(\alpha) = sur^{\mathsf{cap}}_\infty f^{\mathsf{cap}}_\infty g^{\mathrm{com}}_{\mathsf{P}} - \frac{\bar{\kappa}_u}{\eta}(1 - sur^{\mathsf{cap}}_\infty).$$

Subtracting $W^{\mathsf{com},\mathsf{cap}}(\mathsf{REC}) = g^{\mathrm{com}}_{\mathsf{P}}$ gives

$$\lim_{\alpha \to 0} \Delta W^{\mathsf{com},\mathsf{cap}}_{\Pi}(\alpha) = -\big(1 - sur^{\mathsf{cap}}_\infty f^{\mathsf{cap}}_\infty\big) g^{\mathrm{com}}_{\mathsf{P}} - \frac{\bar{\kappa}_u}{\eta}(1 - sur^{\mathsf{cap}}_\infty).$$

This limit is strictly negative. Indeed, if $f_\infty^{\mathsf{cap}} = 1/(n+1)$, then

$$sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}} = \frac{1}{n+1} < 1.$$

If instead $f_\infty^{\mathsf{cap}} > 1/(n+1)$, then $C_{n+1}(f_\infty^{\mathsf{cap}}) > 0$, so

$$sur_\infty^{\mathsf{cap}} = e^{-\eta C_{n+1}(f_\infty^{\mathsf{cap}})} < 1,$$

and therefore again $sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}} < 1$. Since $g_{\mathsf{P}}^{\mathrm{com}} > 0$, the first term is strictly negative, while the second term is weakly negative for every $\bar{\kappa}_u \geq 0$. Therefore, for all sufficiently small α,

$$\Delta W_\Pi^{\mathsf{com},\mathsf{cap}}(\alpha) < 0.$$

This proves part (ii).

Part (iii). Let

$$L := \lambda V_{\mathsf{N}}, \qquad r_{\mathsf{P}} := r_m(u_{\mathsf{P}}), \qquad M^{\mathsf{cap}} := sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}},$$

where f_∞^{cap} is the unique maximizer of $f \mapsto sur^{\mathsf{cap}}(f) f$.

If $(r_{\mathsf{P}} + L) M^{\mathsf{cap}} \leq r_{\mathsf{P}}$, then Theorem 4(iii) implies that REC is commission-optimal for all sufficiently small α. Hence

$$\Delta W_\Pi^{\mathsf{com},\mathsf{cap}}(\alpha) = 0$$

for all such α.

Assume now that $(r_{\mathsf{P}} + L) M^{\mathsf{cap}} > r_{\mathsf{P}}$. Let

$$f_\alpha := f_\Pi^\star(\alpha), \qquad s_\alpha := sur^{\mathsf{cap}}(f_\alpha), \qquad a_\alpha := \alpha f_\alpha, \qquad b_\alpha := \Big(1 - \frac{\alpha}{n}\Big)(1 - f_\alpha), \qquad m_{\mathsf{P},\alpha} := (1-\alpha) f_\alpha + \alpha \frac{1 - f_\alpha}{n}.$$

By Theorem 4(iii),

$$f_\alpha \to f_\infty^{\mathsf{cap}}, \qquad s_\alpha \to sur_\infty^{\mathsf{cap}},$$

and for all sufficiently small α the commission-optimal design strictly outperforms REC, so it follows every niche signal. Therefore

$$m_{\mathsf{P},\alpha} = f_\alpha + O(\alpha), \qquad a_\alpha = O(\alpha), \qquad b_\alpha = 1 - f_\alpha + O(\alpha). \tag{101}$$

The popular seller again sets $p_{\mathsf{P},\alpha}^\star = p_{\mathsf{P}}^\star$, so

$$g(u_{\mathsf{P}} - p_{\mathsf{P},\alpha}^\star) = g_{\mathsf{P}}^{\mathrm{com}}.$$

Now define

$$q_\alpha := \frac{a_\alpha}{a_\alpha + b_\alpha}.$$

Because $a_\alpha \to 0$ and $b_\alpha \to 1 - f_\infty^{\mathsf{cap}} > 0$, we have

$$q_\alpha \to 0.$$

Then

$$q_\alpha \, u_+(\alpha) = \frac{a_\alpha}{a_\alpha + b_\alpha} \cdot \frac{L}{\alpha} \longrightarrow \frac{f_\infty^{\mathsf{cap}} L}{1 - f_\infty^{\mathsf{cap}}}.$$

As in the proof of Theorem 4(iii),

$$\frac{f_\infty^{\mathsf{cap}} L}{1 - f_\infty^{\mathsf{cap}}} > r_-, \qquad r_- := r_m(u_-).$$

For a niche seller, revenue is

$$R_{\mathsf{N},\alpha}(p) = \frac{a_\alpha + b_\alpha}{n}\, p\Big[q_\alpha\, \sigma(u_+(\alpha) - p) + (1 - q_\alpha)\, \sigma(u_- - p)\Big].$$

Let $p^\star_{\mathsf{N},\alpha}$ denote the corresponding equilibrium niche price. Lemma 12 now applies and yields

$$p^\star_{\mathsf{N},\alpha} = u_+(\alpha) - o(u_+(\alpha)), \qquad g(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}) = o(u_+(\alpha)), \qquad g(u_- - p^\star_{\mathsf{N},\alpha}) = o(1).$$

Conditional on reaching the recommendation stage, expected surplus under the commission-optimal design is

$$E_\alpha := m_{\mathsf{P},\alpha}\, g(u_{\mathsf{P}} - p^\star_{\mathsf{P},\alpha}) + a_\alpha\, g(u_+(\alpha) - p^\star_{\mathsf{N},\alpha}) + b_\alpha\, g(u_- - p^\star_{\mathsf{N},\alpha}).$$

Using (101), we obtain

$$E_\alpha = \big(f_\alpha + O(\alpha)\big) g_{\mathsf{P}}^{\mathrm{com}} + a_\alpha\, o(u_+(\alpha)) + \big(1 - f_\alpha + O(\alpha)\big) o(1).$$

Because $a_\alpha = O(\alpha)$ and $u_+(\alpha) = L/\alpha$,

$$a_\alpha\, o(u_+(\alpha)) = o(\alpha u_+(\alpha)) = o(1).$$

Therefore

$$E_\alpha = f_\alpha g_{\mathsf{P}}^{\mathrm{com}} + o(1).$$

It follows that

$$W_\Pi^{\mathsf{com},\mathsf{cap}}(\alpha) = s_\alpha E_\alpha - \frac{\bar\kappa_u}{\eta}(1 - s_\alpha) = s_\alpha f_\alpha g_{\mathsf{P}}^{\mathrm{com}} - \frac{\bar\kappa_u}{\eta}(1 - s_\alpha) + o(1).$$

In the limit, this becomes

$$\lim_{\alpha \to 0} W_\Pi^{\mathsf{com},\mathsf{cap}}(\alpha) = sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}} g_{\mathsf{P}}^{\mathrm{com}} - \frac{\bar\kappa_u}{\eta}(1 - sur_\infty^{\mathsf{cap}}),$$

and subtracting $W^{\mathsf{com},\mathsf{cap}}(\mathsf{REC}) = g_{\mathsf{P}}^{\mathrm{com}}$ yields

$$\lim_{\alpha \to 0} \Delta W_\Pi^{\mathsf{com},\mathsf{cap}}(\alpha) = -\big(1 - sur_\infty^{\mathsf{cap}} f_\infty^{\mathsf{cap}}\big) g_{\mathsf{P}}^{\mathrm{com}} - \frac{\bar\kappa_u}{\eta}(1 - sur_\infty^{\mathsf{cap}}).$$

This limit is strictly negative by the same argument as in part (ii). Hence, for all sufficiently small α,

$$\Delta W_\Pi^{\mathsf{com},\mathsf{cap}}(\alpha) < 0.$$

In the specific case of $(r_{\mathsf{P}} + L)M^{\mathsf{cap}} = r_{\mathsf{P}}$, Theorem 4(iii) shows the commission payoff gap tends to zero, so we assume the policy is REC. □

C.7. Proof of Proposition 7

Proof of Proposition 7: We prove each part individually. Throughout the proof, write $C(f) := C_{n+1}(f)$ and $sur(f) := sur^{\mathsf{cap}}(f) = \exp(-\eta C(f))$. For the $(n+1)$-ary symmetric channel (5), the Shannon capacity is

$$C(f) = \ln(n+1) \;+\; f\ln f \;+\; (1-f)\ln\Big(\frac{1-f}{n}\Big), \qquad f \in \Big[\frac{1}{n+1}, 1\Big]. \tag{102}$$

In particular, $C(1/(n+1)) = 0$, $C(f) > 0$ for $f > 1/(n+1)$, and therefore $sur(f) = 1$ if and only if $f = 1/(n+1)$ while $sur(f) < 1$ for any $f > 1/(n+1)$.

We also use the base utilities

$$u_{\mathsf{P}} = V_{\mathsf{P}}, \qquad u_-(\lambda) = -\lambda, \qquad u_+(\lambda) = \lambda V_{\mathsf{N}} \alpha^{-\beta},$$

and $\sigma(x) = \frac{e^x}{1+e^x}$.

Part (i): Low heterogeneity ($\lambda \to 0$).

Conversion. Fix $\alpha, n, \beta, V_{\mathsf{P}}, V_{\mathsf{N}}, \eta, R$. For any niche j and any type θ, the niche utility satisfies

$$u_{\mathsf{B}}(\theta, j) \in \big[u_-(\lambda), u_+(\lambda)\big] = \big[-\lambda, \lambda V_{\mathsf{N}} \alpha^{-\beta}\big].$$

As $\lambda \to 0$, both endpoints converge to 0, hence by continuity of $\sigma(\cdot)$,

$$\sup_{\theta, j \in \{1,\dots,n\}} \sigma\big(u_{\mathsf{B}}(\theta, j)\big) \;\longrightarrow\; \sigma(0) = \frac{1}{2}.$$

Since $V_{\mathsf{P}} > 0$, we have $\sigma(u_{\mathsf{P}}) > \frac{1}{2}$. Choose $\bar\lambda > 0$ such that for all $\lambda \in [0, \bar\lambda]$,

$$\sup_{\theta, j \in \{1,\dots,n\}} \sigma\big(u_{\mathsf{B}}(\theta, j)\big) \;\le\; \sigma(u_{\mathsf{P}}). \tag{103}$$

Consider any CRS design (f, π). Conditional on reaching the recommendation stage, adoption probability is $\sigma(u_{\mathsf{B}}(\Theta, \pi(S)))$. Hence by (103),

$$\mathbb{E}\big[\sigma\big(u_{\mathsf{B}}(\Theta, \pi(S))\big)\big] \;\le\; \sigma(u_{\mathsf{P}}),$$

with equality only if $\pi(S) \equiv \mathsf{P}$ a.s. Therefore the conversion payoff satisfies

$$\Gamma(f, \pi) = R\, sur(f)\, \mathbb{E}\big[\sigma\big(u_{\mathsf{B}}(\Theta, \pi(S))\big)\big] \le R\, sur(f)\, \sigma(u_{\mathsf{P}}) \le R\, \sigma(u_{\mathsf{P}}),$$

where the second inequality uses $sur(f) \le 1$ with equality only at $f = 1/(n+1)$. Thus $\Gamma(f, \pi) \le R\sigma(u_{\mathsf{P}})$, with equality only at REC (which uses $f = 1/(n+1)$ and recommends P). Hence REC is conversion-optimal for all $\lambda \in [0, \bar\lambda]$.

Commission. Let $r_m(u) := \max_{p \ge 0} p\,\sigma(u - p)$ denote the monopoly revenue under logit demand. For any $u' > u$ and any $p \ge 0$, $p\sigma(u' - p) > p\sigma(u - p)$, hence $r_m(\cdot)$ is strictly increasing. In particular, $r_m(u_{\mathsf{P}}) > r_m(0)$.

Choose $\bar\lambda > 0$ (possibly smaller than above) such that

$$r_m\big(u_+(\lambda)\big) \;<\; r_m(u_{\mathsf{P}}) \qquad \forall\, \lambda \in [0, \bar\lambda]. \tag{104}$$

Now fix any design (f, π) and consider any seller j. For any recommended niche $j \in \{1, \dots, n\}$ and any type θ, we have $u_{\mathsf{B}}(\theta, j) \le u_+(\lambda)$; therefore for any price $p \ge 0$,

$$p\,\sigma\big(u_{\mathsf{B}}(\theta, j) - p\big) \;\le\; p\,\sigma\big(u_+(\lambda) - p\big),$$

and taking expectations over Θ and maximizing over p yields that the seller-side revenue per recommendation of niche j is at most $r_m(u_+(\lambda))$. In contrast, since $u_{\mathsf{B}}(\theta, \mathsf{P}) = u_{\mathsf{P}}$ for all θ, the seller-side revenue per recommendation of P equals $r_m(u_{\mathsf{P}})$.

Thus, for $\lambda \in [0, \bar{\lambda}]$ the platform maximizes commission revenue by recommending P in every session that reaches the recommendation stage (because (104) makes P strictly dominant in revenue per recommendation). Given that choice, the platform's expected commission is $\tau \cdot sur(f)\, r_m(u_{\mathsf{P}})$, which is maximized by setting $sur(f) = 1$, i.e. $f = 1/(n+1)$. Hence REC is commission-optimal for all $\lambda \in [0, \bar{\lambda}]$. This proves part (i).

Part (ii): High heterogeneity ($\lambda \to \infty$), conversion.

Fix $(\alpha, n, \beta, V_{\mathsf{P}}, V_{\mathsf{N}}, \eta, R)$. Write $\sigma_{\mathsf{P}} := \sigma(u_{\mathsf{P}}) \in (0,1)$. Define $q_0(f) := \Pr(S = \emptyset) = (1-\alpha)f + \alpha(1-f)/n$ and $A_\infty(f) := \sigma_{\mathsf{P}} q_0(f) + \alpha f$ as in the proposition.

Step 1: Threshold convergence and the limiting fidelity cutoff. For each λ, the posterior-threshold property for the conversion objective implies that after a niche signal $S = i$, the platform compares recommending niche i versus recommending P and recommends the niche iff $q(f, \alpha, n) \geq q^\star_\Gamma(\lambda)$, where

$$q^\star_\Gamma(\lambda) = \frac{\sigma(u_{\mathsf{P}}) - \sigma(u_-(\lambda))}{\sigma(u_+(\lambda)) - \sigma(u_-(\lambda))}.$$

Since $u_-(\lambda) \to -\infty$ and $u_+(\lambda) \to +\infty$ as $\lambda \to \infty$, we have $\sigma(u_-(\lambda)) \to 0$ and $\sigma(u_+(\lambda)) \to 1$, hence

$$q^\star_\Gamma(\lambda) \;\longrightarrow\; \sigma_{\mathsf{P}} \qquad (\lambda \to \infty). \tag{105}$$

Let $\underline{f}_\Gamma(\lambda)$ denote the minimal fidelity such that $q(f, \alpha, n) \geq q^\star_\Gamma(\lambda)$. By the closed form for $\underline{f}$ (solving $q(f, \alpha, n) = q^\star_\Gamma(\lambda)$),

$$\underline{f}_\Gamma(\lambda) = \frac{q^\star_\Gamma(\lambda)\,(1 - \alpha/n)}{q^\star_\Gamma(\lambda)\,(1-\alpha/n) + \alpha\,(1 - q^\star_\Gamma(\lambda))}.$$

Combining this with (105) yields

$$\underline{f}_\Gamma(\lambda) \;\longrightarrow\; \underline{f}_\infty = \frac{\sigma_{\mathsf{P}}\,(1-\alpha/n)}{\sigma_{\mathsf{P}}\,(1-\alpha/n) + \alpha(1-\sigma_{\mathsf{P}})}. \tag{106}$$

Step 2: Payoff form for $f < \underline{f}_\Gamma(\lambda)$. If $f < \underline{f}_\Gamma(\lambda)$, then the optimal rule ignores niche signals and recommends P even after $S \in \{1, \ldots, n\}$. Thus $\pi(S) \equiv \mathsf{P}$ and

$$\Gamma(f, \pi) = R\, sur(f)\, \sigma_{\mathsf{P}} \;\leq\; R\sigma_{\mathsf{P}},$$

with equality if and only if $f = 1/(n+1)$ (i.e. REC).

Step 3: Payoff form for $f \geq \underline{f}_\Gamma(\lambda)$ and uniform convergence. If $f \geq \underline{f}_\Gamma(\lambda)$, then after any niche signal $S = i$ the optimal rule recommends niche i. Under this follow-the-niche-signal rule, conditional on reaching the recommendation stage,

$$\begin{aligned}\mathbb{E}\big[\sigma\big(u_{\mathsf{B}}(\Theta, \pi(S))\big)\big] &= \sigma_{\mathsf{P}} \Pr(S = \emptyset) \;+\; \sigma(u_+(\lambda)) \Pr(\Theta = S \in \{1, \ldots, n\}) + \qquad (107)\\ &\qquad \sigma(u_-(\lambda)) \Pr(S \in \{1, \ldots, n\}, \Theta \neq S)\\ &= \sigma_{\mathsf{P}} q_0(f) \;+\; \alpha f\, \sigma(u_+(\lambda)) \;+\; (1-f)\Big(1 - \frac{\alpha}{n}\Big)\sigma(u_-(\lambda)). \qquad (108)\end{aligned}$$

Define

$$A_\lambda(f) := \sigma_{\mathsf{P}} q_0(f) \;+\; \alpha f\, \sigma(u_+(\lambda)) \;+\; (1-f)\Big(1-\frac{\alpha}{n}\Big)\sigma(u_-(\lambda)).$$

Then $\Gamma(f,\pi) = R\,sur(f)\,A_\lambda(f)$. Moreover, as $\lambda \to \infty$ we have $\sigma(u_+(\lambda)) \to 1$ and $\sigma(u_-(\lambda)) \to 0$, so for all $f \in [1/(n+1), 1]$,

$$\big|A_\lambda(f) - A_\infty(f)\big| \le \alpha\big|\sigma(u_+(\lambda)) - 1\big| \;+\; \Big(1-\frac{\alpha}{n}\Big)\big|\sigma(u_-(\lambda))\big| \;\xrightarrow[\lambda\to\infty]{}\; 0.$$

Hence $A_\lambda(\cdot) \to A_\infty(\cdot)$ uniformly on $[1/(n+1), 1]$, and therefore

$$\sup_{f\in[1/(n+1),1]} \Big|R\,sur(f)\,A_\lambda(f) - \Gamma^{\mathsf{cap}}_\infty(f)\Big| \;\xrightarrow[\lambda\to\infty]{}\; 0, \qquad \Gamma^{\mathsf{cap}}_\infty(f) := R\,sur(f)\,A_\infty(f). \tag{109}$$

Step 4: Optimality comparison. Let $B := R\sigma_{\mathsf{P}}$ denote the REC payoff. Combining Steps 2–3, for each λ the conversion optimum equals

$$\max\Big\{B,\; \max_{f\in[\underline{f}_\Gamma(\lambda),1]} R\,sur(f)\,A_\lambda(f)\Big\}.$$

Using (106) and (109), the inner maximization converges to $\max_{f\in[\underline{f}_\infty,1]} \Gamma^{\mathsf{cap}}_\infty(f)$. Therefore:

- If $\max_{f\in[\underline{f}_\infty,1]} \Gamma^{\mathsf{cap}}_\infty(f) \le B$, then for all sufficiently large λ we have $\max_{f\in[\underline{f}_\Gamma(\lambda),1]} R\,sur(f)\,A_\lambda(f) \le B$, hence REC is conversion-optimal.
- If $\max_{f\in[\underline{f}_\infty,1]} \Gamma^{\mathsf{cap}}_\infty(f) > B$, then pick any $f^\dagger \in \arg\max_{f\in[\underline{f}_\infty,1]} \Gamma^{\mathsf{cap}}_\infty(f)$. By (106), $f^\dagger \in [\underline{f}_\Gamma(\lambda), 1]$ for all sufficiently large λ, and by (109), $R\,sur(f^\dagger)\,A_\lambda(f^\dagger) > B$ for all sufficiently large λ. Hence elicitation is conversion-optimal for all sufficiently large λ.

Finally, let $\lambda_k \to \infty$ and let f_k be any sequence of conversion-optimal fidelities with $f_k \ge \underline{f}_\Gamma(\lambda_k)$ (i.e. in the elicitation region). Compactness of $[1/(n+1), 1]$ yields a convergent subsequence $f_{k_\ell} \to \bar{f}$. By (106), $\bar{f} \ge \underline{f}_\infty$. Moreover, optimality of f_{k_ℓ} implies for every $f \in [\underline{f}_\Gamma(\lambda_{k_\ell}), 1]$,

$$R\,sur(f_{k_\ell})\,A_{\lambda_{k_\ell}}(f_{k_\ell}) \;\ge\; R\,sur(f)\,A_{\lambda_{k_\ell}}(f).$$

Taking $\ell \to \infty$ and using uniform convergence (109) yields $\Gamma^{\mathsf{cap}}_\infty(\bar{f}) \ge \Gamma^{\mathsf{cap}}_\infty(f)$ for all $f \in [\underline{f}_\infty, 1]$. Thus $\bar{f} \in \arg\max_{f\in[\underline{f}_\infty,1]} \Gamma^{\mathsf{cap}}_\infty(f)$, proving the argmax claim in part (ii).

Part (iii): High heterogeneity ($\lambda \to \infty$), commission.

Fix $\alpha, n, \beta, V_{\mathsf{P}}, V_{\mathsf{N}}, \eta, \tau$. Let $r_m(u) := \max_{p\ge 0} p\,\sigma(u-p)$ and define $r_{\mathsf{P}} := r_m(u_{\mathsf{P}})$. Also define $r_{\mathsf{N}}(\lambda) := r_m(u_+(\lambda))$. Since $u_+(\lambda) \to \infty$ as $\lambda \to \infty$, we have $r_{\mathsf{N}}(\lambda) \to \infty$; for instance, choosing $p = u_+(\lambda)/2$ gives $r_{\mathsf{N}}(\lambda) \ge \frac{u_+(\lambda)}{2}\sigma\big(u_+(\lambda)/2\big) \to \infty$.

Step 1: Elicitation eventually dominates REC. Consider the feasible elicitation policy that uses $f = 1$ and recommends P when $S = \emptyset$ and niche j when $S = j$. Under $f = 1$, all niche users are perfectly screened and routed to their matching niche seller, so with survival $sur(1) = (n+1)^{-\eta}$, the platform's commission revenue is at least

$$\tau\, sur(1) \cdot \alpha\, r_{\mathsf{N}}(\lambda),$$

since an α-fraction of users are routed to a matching niche seller and yield seller revenue $r_{\mathsf{N}}(\lambda)$. Because $r_{\mathsf{N}}(\lambda) \to \infty$, this lower bound diverges as $\lambda \to \infty$. In contrast, REC yields commission τr_{P} (a finite constant). Hence there exists $\bar{\lambda} < \infty$ such that for all $\lambda \ge \bar{\lambda}$, some elicitation policy strictly dominates REC, and therefore the commission-optimal CRS does not coincide with REC.

Step 2: Leading-order objective and limiting fidelity. Fix any $f \in [1/(n+1), 1]$ and consider the (feasible) follow-the-niche-signal rule that recommends P when $S = \emptyset$ and recommends niche S otherwise. For each niche seller j, let $m_j(f) = \Pr(\Theta = j, S = j) = \frac{\alpha}{n} f$ denote the mass of correctly routed users to j (recall the user must also survive, which contributes a factor $sur(f)$ common to all sellers). At any posted price $p \geq 0$, the seller-j revenue contribution from mismatched users is bounded by

$$p \Pr(\Theta \neq j,\, S = j)\, \sigma(u_-(\lambda) - p) \;\leq\; p\, \sigma(u_-(\lambda)) \;\leq\; p\, e^{u_-(\lambda)} = p\, e^{-\lambda}.$$

At the same time, the matched-user revenue term equals $p\, m_j(f)\, \sigma(u_+(\lambda) - p)$. Since the logit revenue $p\sigma(u_+(\lambda) - p)$ is maximized at some finite price of order $u_+(\lambda)$, the seller's optimal price under the mixture differs from $p_m(u_+(\lambda))$ by a vanishing amount and the maximal revenue satisfies

$$\max_{p \geq 0} p\Big(m_j(f)\sigma(u_+(\lambda) - p) + \Pr(\Theta \neq j,\, S = j)\sigma(u_-(\lambda) - p)\Big) = m_j(f)\, r_{\mathsf{N}}(\lambda) \;+\; o\big(r_{\mathsf{N}}(\lambda)\big),$$

where the $o(\cdot)$ term is uniform in f because the mismatch term is uniformly bounded by $O(u_+(\lambda)e^{-\lambda}) = o(r_{\mathsf{N}}(\lambda))$. Summing across $j = 1, \ldots, n$ yields total niche-seller revenue

$$\sum_{j=1}^{n} m_j(f)\, r_{\mathsf{N}}(\lambda) \;+\; o\big(r_{\mathsf{N}}(\lambda)\big) = \alpha f\, r_{\mathsf{N}}(\lambda) \;+\; o\big(r_{\mathsf{N}}(\lambda)\big),$$

uniformly in f. Adding the (bounded) P-seller revenue contribution and multiplying by the survival factor $sur(f)$ and take rate τ gives that the platform's commission payoff under this rule satisfies

$$\Pi_\lambda^{\mathsf{com}}(f) = \tau \cdot sur(f)\, \Big(\alpha f\, r_{\mathsf{N}}(\lambda)\Big) \;+\; o\big(r_{\mathsf{N}}(\lambda)\big), \qquad \text{uniformly in } f \in \Big[\frac{1}{n+1}, 1\Big]. \tag{110}$$

Since $r_{\mathsf{N}}(\lambda) \to \infty$, the asymptotic maximization of $\Pi_\lambda^{\mathsf{com}}(f)$ over f reduces to maximizing $sur(f)f$. Formally, divide (110) by $\tau\alpha r_{\mathsf{N}}(\lambda)$ to obtain uniform convergence

$$\sup_{f \in [1/(n+1), 1]} \left| \frac{\Pi_\lambda^{\mathsf{com}}(f)}{\tau\alpha r_{\mathsf{N}}(\lambda)} - sur(f) f \right| \;\xrightarrow[\lambda \to \infty]{}\; 0.$$

Hence any sequence of commission-optimal fidelities has limit points in $\arg\max_{f \in [1/(n+1), 1]} sur(f)f$. To show uniqueness, consider $\phi(f) := \ln f - \eta C(f)$ on $(1/(n+1), 1)$. Using (102), $C(\cdot)$ is C^2 with

$$C'(f) = \ln\Big(\frac{nf}{1-f}\Big), \qquad C''(f) = \frac{1}{f(1-f)} > 0, \tag{111}$$

so $-\eta C(\cdot)$ is strictly concave and $\ln f$ is strictly concave; therefore ϕ is strictly concave. Moreover, $\phi'(1/(n+1)) = 1/f - \eta C'(f)\big|_{f=1/(n+1)} = n + 1 > 0$ and $\phi'(f) \to -\infty$ as $f \uparrow 1$ (since $C'(f) \to +\infty$), so ϕ has a unique maximizer $f_\infty^{\mathsf{cap}} \in (1/(n+1), 1)$ characterized by $\phi'(f_\infty^{\mathsf{cap}}) = 0$, i.e.

$$\frac{1}{f_\infty^{\mathsf{cap}}} = \eta C'(f_\infty^{\mathsf{cap}}) = \eta \ln\Big(\frac{n f_\infty^{\mathsf{cap}}}{1 - f_\infty^{\mathsf{cap}}}\Big).$$

This proves part (iii) and completes the proof. □

C.8. Proof of Proposition 8

Proof of Proposition 8: The proof consists of 4 steps: 1) we show a lower bound for the capacity and the decay rate of survival, 2) we show that if niches are never recommended, REC is optimal, 3) for large n, recommending a niche requires a fidelity bounded away from 0, and 4) for large n, any niche-recommending design is dominated by REC.

Write $C_n(f) := C_{n+1}(f)$ and $sur_n(f) := sur^{\mathsf{cap}}(f) = \exp(-\eta C_n(f))$ to emphasize the dependence on n. Recall the $(n{+}1)$-ary symmetric channel (5) and define

$$u_{\mathsf{P}} := V_{\mathsf{P}}, \qquad u_- := -\lambda, \qquad u_+(\alpha) := \lambda V_{\mathsf{N}} \alpha^{-\beta}, \qquad \sigma(x) := \frac{e^x}{1+e^x}, \qquad g(x) := \ln(1+e^x).$$

Throughout the proof, α is fixed, hence $u_+(\alpha)$ is a finite constant independent of n.

Step 1: capacity lower bound and survival decay in n. For the $(n{+}1)$-ary symmetric channel, the Shannon capacity is

$$C_n(f) = \ln(n{+}1) + f\ln f + (1-f)\ln\Big(\frac{1-f}{n}\Big), \qquad f \in \Big[\frac{1}{n+1}, 1\Big]. \tag{112}$$

Rewriting,

$$C_n(f) = \ln(1+1/n) \;+\; f\ln n \;+\; \big[f\ln f + (1-f)\ln(1-f)\big].$$

Since $\ln(1+1/n) \geq 0$ and $f\ln f + (1-f)\ln(1-f) \geq -\ln 2$ for all $f \in [0,1]$, for every $n \geq 2$ and every $f \in [1/(n+1), 1]$,

$$C_n(f) \;\geq\; f\ln n - \ln 2. \tag{113}$$

Consequently,

$$sur_n(f) = \exp(-\eta C_n(f)) \;\leq\; 2^{\eta}\, n^{-\eta f} \qquad \forall n \geq 2. \tag{114}$$

Step 2: if niches are never recommended, REC *is optimal.* Consider any design (f,π) that *never* recommends a niche (i.e., $\pi(S) \equiv \mathsf{P}$). Then:

- *Conversion:* $\Gamma(f,\pi) = R \cdot sur_n(f)\,\sigma(u_{\mathsf{P}}) \leq R\sigma(u_{\mathsf{P}}) = \Gamma(\mathsf{REC})$, with strict inequality if $f > 1/(n+1)$ (since then $C_n(f) > 0$ and $sur_n(f) < 1$).
- *Commission:* the P seller's monopoly revenue is $r_{\mathsf{P}} := r_m(u_{\mathsf{P}}) > 0$. If P is always recommended, platform commission is $\Pi^{\mathsf{com}}(f,\pi) = \tau \cdot sur_n(f)\, r_{\mathsf{P}} \leq \tau r_{\mathsf{P}} = \Pi^{\mathsf{com}}(\mathsf{REC})$, again strict if $f > 1/(n+1)$.
- *Welfare:* $W(f,\pi) = sur_n(f)\, g(u_{\mathsf{P}}) - \frac{\bar{\kappa}_u}{\eta}(1 - sur_n(f)) \leq g(u_{\mathsf{P}}) = W(\mathsf{REC})$, strict if $f > 1/(n+1)$.

Thus, among designs that never recommend niches, REC is uniquely optimal for each objective. Therefore, to beat REC under any objective, a design must recommend a niche with positive probability.

Step 3: recommending a niche requires a fidelity bounded away from 0 *for large* n. Let $q(f,\alpha,n)$ denote the posterior correctness probability after a niche signal,

$$q(f,\alpha,n) = \Pr(\Theta = s \mid S = s) = \frac{\alpha f}{\alpha f + (1-\alpha/n)(1-f)}.$$

For fixed (α, f), $q(f,\alpha,n)$ is decreasing in n and converges to

$$q_\infty(f) := \lim_{n\to\infty} q(f,\alpha,n) = \frac{\alpha f}{\alpha f + (1-f)}. \tag{115}$$

(The monotonicity follows because $(1-\alpha/n)(1-f)$ increases in n.)

Fix an objective $\mathsf{obj} \in \{\Gamma, W, \Pi^{\mathsf{com}}\}$. If $u_+(\alpha) \le u_{\mathsf{P}}$, then recommending a niche is weakly dominated (even under perfect information) for that objective, so by Step 2 REC is optimal for all n and we are done. Henceforth assume $u_+(\alpha) > u_{\mathsf{P}}$ so that acting on niche signals is potentially beneficial.

Under conversion and welfare, the standard posterior-threshold property implies: after a niche signal, recommending the niche is optimal if and only if $q(f,\alpha,n) \ge q^\star_{\mathsf{obj}}$, where $q^\star_\Gamma \in (0,1)$ and $q^\star_W \in (0,1)$ are the cutoffs

$$q^\star_\Gamma = \frac{\sigma(u_{\mathsf{P}}) - \sigma(u_-)}{\sigma(u_+) - \sigma(u_-)}, \qquad q^\star_W = \frac{g(u_{\mathsf{P}}) - g(u_-)}{g(u_+) - g(u_-)}.$$

For commission, define the niche-seller monopoly revenue when the routed population has utility u_+ with probability q and u_- with probability $1-q$:

$$r_{\mathsf{N}}(q) := \max_{x \ge 0}\; x\big(q\,\sigma(u_+ - x) + (1-q)\,\sigma(u_- - x)\big).$$

For each x, the term in parentheses is affine and strictly increasing in q because $\sigma(u_+ - x) > \sigma(u_- - x)$. Taking a supremum over x preserves strict monotonicity, so $r_{\mathsf{N}}(q)$ is strictly increasing in q. Moreover $r_{\mathsf{N}}(0) = r_m(u_-)$ and $r_{\mathsf{N}}(1) = r_m(u_+)$. Since $u_- > u_{\mathsf{P}}$ is false (indeed $u_- < u_{\mathsf{P}}$), monotonicity of $r_m(\cdot)$ implies $r_m(u_-) < r_m(u_{\mathsf{P}}) =: r_{\mathsf{P}}$. Also $u_+ > u_{\mathsf{P}}$ implies $r_m(u_+) > r_{\mathsf{P}}$. Therefore there exists a unique $q^\star_\Pi \in (0,1)$ such that $r_{\mathsf{N}}(q^\star_\Pi) = r_{\mathsf{P}}$, and a commission-maximizing platform recommends a niche only if $q(f,\alpha,n) \ge q^\star_\Pi$. (Commission rate τ cancels in this comparison.)

Thus, for each objective obj there is a threshold $q^\star_{\mathsf{obj}} \in (0,1)$ such that recommending any niche requires

$$q(f,\alpha,n) \;\ge\; q^\star_{\mathsf{obj}}. \tag{116}$$

Let $q^\star := q^\star_{\mathsf{obj}}$ and define the limiting minimal fidelity $f^\infty(q^\star)$ by solving $q_\infty(f) = q^\star$:

$$f^\infty(q^\star) := \frac{q^\star}{\alpha + q^\star(1-\alpha)} \in (0,1). \tag{117}$$

Since $q_\infty(\cdot)$ is continuous and strictly increasing and $q_\infty(f^\infty(q^\star)) = q^\star$, we have $q_\infty(f) < q^\star$ for all $f < f^\infty(q^\star)$. Fix $f_0 := \frac{1}{2} f^\infty(q^\star) > 0$. Then $q_\infty(f_0) < q^\star$. By (115), there exists $n_0 < \infty$ such that for all $n \ge n_0$,

$$q(f_0,\alpha,n) \;<\; q^\star. \tag{118}$$

Because $q(f,\alpha,n)$ is increasing in f, (118) implies that for all $n \ge n_0$, any $f < f_0$ violates (116), hence cannot rationalize niche recommendation under the optimal rule. Therefore, for all $n \ge n_0$, any design that recommends a niche with positive probability must satisfy

$$f \;\ge\; f_0 \;>\; 0, \tag{119}$$

where f_0 depends only on primitives and the objective, but not on n.

Step 4: for large n, any niche-recommending design is dominated by REC. Fix $n \geq \max\{2, n_0\}$ and consider any design that recommends a niche with positive probability. By (119), it must have $f \geq f_0$. Then (114) gives $sur_n(f) \leq 2^\eta n^{-\eta f_0}$.

Now bound each objective from above by a constant times $sur_n(f)$. Since $\sigma(\cdot) \leq 1$, conversion payoff satisfies

$$\Gamma(f, \pi) \leq R\, sur_n(f) \;\leq\; R\, 2^\eta n^{-\eta f_0}.$$

Commission payoff is τ times seller revenue; with a single recommended item per session and logit demand, seller revenue per surviving session is uniformly bounded by $r_{\max} := \max\{r_m(u_{\mathsf{P}}), r_m(u_+), r_m(u_-)\} < \infty$, hence

$$\Pi^{\mathsf{com}}(f, \pi) \leq \tau\, r_{\max}\, sur_n(f) \;\leq\; \tau r_{\max}\, 2^\eta n^{-\eta f_0}.$$

Welfare satisfies $g(\cdot) \leq g_{\max} := \max\{g(u_{\mathsf{P}}), g(u_+), g(u_-)\} < \infty$, and the effort term is nonpositive, so

$$W(f, \pi) \leq sur_n(f)\, g_{\max} \;\leq\; g_{\max}\, 2^\eta n^{-\eta f_0}.$$

Since $n^{-\eta f_0} \to 0$ as $n \to \infty$, we can choose $\bar{n}_\Gamma$ so that $R\, 2^\eta \bar{n}_\Gamma^{-\eta f_0} < R\sigma(u_{\mathsf{P}}) = \Gamma(\mathsf{REC})$, choose $\bar{n}_\Pi$ so that $\tau r_{\max} 2^\eta \bar{n}_\Pi^{-\eta f_0} < \tau r_{\mathsf{P}} = \Pi^{\mathsf{com}}(\mathsf{REC})$, and choose $\bar{n}_W$ so that $g_{\max} 2^\eta \bar{n}_W^{-\eta f_0} < g(u_{\mathsf{P}}) = W(\mathsf{REC})$. Let $\bar{n} := \max\{\bar{n}_\Gamma, \bar{n}_\Pi, \bar{n}_W, n_0, 2\}$.

Then for all $n \geq \bar{n}$, every niche-recommending design yields strictly lower conversion payoff, commission payoff, and welfare than REC. By Step 2, any design that never recommends a niche is also weakly dominated by REC. Hence REC is optimal for all three objectives when $n \geq \bar{n}$. □

Appendix D: Dataset Creation

This appendix contains details on the creation of the dataset used in Section 4, including prompts used. The base dataset used is the E-GEO dataset.

D.1. Classification and Category Selection

The original queries from E-GEO were classified using ChatGPT-5.2 (Thinking). The prompt is given below:

Prompt D.1: Initial Classification Prompt

```
I want you to go through these queries one by one, identify what product category are they looking for in that query and then cluster the
queries according to that category. Example categories are shoes, staplers, scissors, headphones, usb cable, etc. I want you to give a
downloadable json file which as the product category and then all the queries that are for that product category.
```

After the classification, this JSON was merged with the original E-GEO dataset to add the "query_category" field. Then, categories were ordered according to frequency. Within the top 60 categories, some similar categories were merged together:

Combined Term	Initial Categories
backpack	backpack, laptop backpack, laptop bag
bag	bag, bags, messenger bag
boot	boot, work boot
bed	bed, bed frame
chair	chair, office chair, desk chair, computer chair
glove	glove, work glove, winter glove
knife	knife, kitchen knife, pocket knife
speaker	speaker, bluetooth speaker
trimmer	trimmer, beard trimmer

Then, the top 40 categories were taken, ignoring categories that did not directly correspond to a product category or were improperly categorized, i.e. categories named "outdoor", "preferably", "travel", and "winter". From this, 40 separate JSON files were created, each with the queries for a single category. A random subsample of at least 20 queries was hand-checked to verify that the classification was accurate. This set of files is available in the Github.

D.2. Persona Generation

For each category, we generated 100 synthetic personas from ChatGPT-5.2 Thinking. To avoid output token limits from the model, personas were generated 20 at a time and concatenated into a final list, with all 100 personas being generated from a single conversation instance. On occasion, the LLM would generate fewer than 20 personas, in which case we would instruct it to "proceed until {nearest multiple of 20}".

The prompt is given below in prompt D.2. We would provide the category name for the product category and category slug fields, and upload the JSON file consisting of the relevant queries for the dataset.

Prompt D.2: Persona Generation Prompt

```
‘‘‘{category}_queries.json‘‘‘
You are generating a research-quality synthetic dataset for evaluating Conversational Recommender Systems (CRS).

You will be given one JSON file uploaded in this chat. Each entry contains a real-world user request for the product category. Your job is
to learn, from these JSONs, what attributes, pain points, jargon, constraints, and budgets real users mentionthen generate realistic
synthetic personas and multi-round preference-revelation trails. If the dataset does not have sufficient features mentioned, feel free to
make up attributes which might be relevant for the product.

====================
PARAMETERS (EDIT THIS)
====================

PRODUCT_CATEGORY: "<INSERT_CATEGORY>"
CATEGORY_SLUG: "<insert_slug>"

TOTAL_PERSONAS: 100
PERSONAS_PER_BATCH: 20

NUM_ROUNDS: 8
HETEROGENEITY: "<high>"
LANGUAGE: "English"

HIGHLIGHT_MARKER: "[[...]]"

====================
BATCHING + CONTROL FLOW
====================

- Generate personas in batches of exactly PERSONAS_PER_BATCH personas.
- Personas must be generated sequentially by ID.
  - Batch 1: <CATEGORY_SLUG>_001 to <CATEGORY_SLUG>_020
  - Batch 2: <CATEGORY_SLUG>_021 to <CATEGORY_SLUG>_040
  - Continue until TOTAL_PERSONAS is reached.
- After each batch:
  - Output ONLY the valid JSON for that batch.
  - STOP generation immediately.
  - Wait for the user to reply with "proceed"
- Do NOT generate the next batch unless proceed is received.
- After the final batch, stop with no additional output.

====================
INPUT DATA (JSON FILES)
====================

1) Read ALL uploaded JSON file(s).
2) Treat each rows query text as REFERENCE_DATA.
3) Do NOT copy any request verbatim.
4) Do NOT closely paraphrase any single request. Generalize and recombine.

Internally (do NOT output this analysis), infer:
- Common attributes and feature vocabulary for PRODUCT_CATEGORY
- Common failure modes / complaints / why Im replacing this
- Typical contexts of use
- Typical constraints (compatibility, sizing, comfort, durability, materials, etc.)
- Typical budget ranges and how people talk about budgets, how important the budget is for them
- Realistic ways preferences are revealed over multiple chat turns
```

```
====================
OUTPUT (STRICT, PER BATCH)
====================

Return ONLY valid JSON with EXACTLY this structure:

{
  "personas": [
    {
      "id": "<CATEGORY_SLUG>_001",
      "gender": "female | male | non-binary | unspecified",
      "background": "1 sentence, 1225 words. Lifestyle/context only. No names, no
      exact addresses.",
      "ground_truth_need": "1 paragraph, 70120 words. Fully specified latent requirement including usage context, dealbreakers, constraints,
budget logic (target + hard max), and a few secondary preferences/dislikes, with clear indication which aspects are more important.",
      "context_trail": [
        "Round 1: one sentence.",
        "Round 2: one sentence.",
        "Round 3: one sentence.",
        "Round 4: one sentence.",
        "Round 5: one sentence.",
        "Round 6: one sentence.",
        "Round 7: one sentence.",
        "Round 8: one sentence."
      ]
    }
  ]
}

Hard requirements:
- Exactly PERSONAS_PER_BATCH personas per batch.
- IDs must be sequential and continuous across batches.
- Each persona must have exactly NUM_ROUNDS entries in context_trail.
- Each context_trail entry must be exactly ONE sentence.
- The context_trail is NOT cumulative; it is one user message per round. I will concatenate rounds myself.
- Nothing in context_trail may contradict ground_truth_need.

====================
CRITICAL: IMPORTANCE-RANKED REVELATION + HIGHLIGHTING
====================

For EACH persona, do this internally:

1) From the ground_truth_need, derive a ranked list of the top (NUM_ROUNDS - 1) preferences/constraints in STRICT decreasing order of
importance:
   - P1 = most important (dealbreaker / primary driver)
   - P2 = second most important
   - ...
   - P(NUM_ROUNDS-1) = least important among the revealed set

2) Map them to rounds:
   - Round 2 reveals P1
   - Round 3 reveals P2
   - ...
   - Round NUM_ROUNDS reveals P(NUM_ROUNDS-1)

Highlighting rule:
- In EVERY round from Round 2 to Round NUM_ROUNDS:
  - Include EXACTLY ONE highlighted span using [[...]]
  - That highlighted span must be the NEW preference introduced in that round
  - Do NOT highlight anything else
- Round 1 must have NO highlight

Language guidance:

- Each round must make the relative importance of the newly introduced preference clear (priority order is unambiguous), using natural
language derived from REFERENCE_DATA.
- Across personas, use specific language to describe the preferences, including:
  - statements of priority ("My main priority is...")
  - negative constraints (I cant deal with)
  - tradeoff framing (Im willing to sacrifice X if)
  - experiential framing (In the past, Ive disliked)
  - future-use framing (This matters because Ill be)
- When you have created the NUM_ROUNDS, evaluate it and adjust word to avoid repetitiveness in phrasing and syntax---the sentences together
should be natural and accurate to the persona.

Budget rule:
- Budget must appear in ground_truth_need.
- Budget must appear as one of the preferences in a highlighted span.
- When budget is a higher priority item or a critical deciding factor for the user, it should appear in an earlier round, like other
preferences. More budget-conscious users, for example, would put their budget in very early rounds.
- You should infer the budget priority from the user's background and ground truth needs, e.g. students may have budget as a primary
concern, and will express it early in round 2 or 3.

====================
CONTEXT TRAIL CONTENT RULES
====================

Round 1 (generic, under-specified):
```

```
- Always: a short request for PRODUCT_CATEGORY.
- Example pattern: Im looking for <product>., Need <product>,
  I want to buy <product>, I want <product>.

Rounds 2 to NUM_ROUNDS:
- Each round adds EXACTLY ONE new preference (the highlighted one).
- The rest of the sentence may provide a tiny bit of glue words, but do not sneak in additional constraints.
- Across personas, vary which attribute becomes P1 (budget, durability, comfort, compatibility, size, aesthetics, brand avoidance, etc.)as
supported by REFERENCE_DATA and HETEROGENEITY.

====================
GROUND TRUTH NEED RULES
====================

ground_truth_need is the latent answer key. It must:
- Be internally consistent.
- Include usage context, constraints, dealbreakers, and budget logic.
- Contain more detail than the context trail (it can include extra secondary preferences not revealed).
- Reflect realistic attribute vocabulary and complaint patterns learned from REFERENCE_DATA---different customers should have unique
phrasing and syntax.
- Should clearly establish priorities, i.e. which preferences are more vs. less important, including budget---for some users, the most
important thing may be having a low price.

====================
HETEROGENEITY CONTROL
====================

LOW (or <=0.3):
- Most personas share a common core of 24 typical constraints from REFERENCE_DATA.
- Differences are small: minor budget shifts, small feature tweaks, mild preference ordering changes.

MEDIUM (or 0.30.7):
- Mix common-core personas with several distinct subgroups.
- Noticeable variation, but still anchored around typical category needs.

HIGH (or >=0.7):
- Create clearly distinct segments with different primary drivers and contexts of use.
- Include some niche constraints that still feel realistic and are supported by REFERENCE_DATA patterns.
- Budgets vary widely as supported by REFERENCE_DATA.

====================
FINAL CHECK (DO THIS SILENTLY)
====================

Before outputting each batch:
- Exactly PERSONAS_PER_BATCH personas.
- Correct sequential IDs.
- Exactly NUM_ROUNDS one-sentence turns per persona.
- Exactly ONE [[...]] highlight in each turn from Round 2 onward.
- Round 2 highlight is most important, then decreasing importance each round.
- No contradictions or near-verbatim reuse from REFERENCE_DATA.

Now produce ONLY the first batch and STOP. Wait for the user to say "proceed" before continuing.
```

Appendix E: Details for the Dataset and Simulation Testbed

This section contains details on the calculations and evaluations used in Section 4.

E.1. Heterogeneity

To calculate the heterogeneity of a set of personas, we first obtain an embedding of each persona's `ground_truth_need` x_i using the bge-large-en-v1.5 encoder sentence encoder. Then, we calculate the centroid by finding the mean vector $\bar{x} = (1/n)\sum_i x_i$, and measure the distance to a vector as $d(x_i, x_j) = 1 - \frac{x_i^\top x_j}{\|x_i\|\|x_j\|}$, or $1 -$ cosine similarity. The heterogeneity is calculated as the average distance of personas to the centroid $\bar{x}$. Figure 7 shows the heterogeneity for all 40 categories.

E.2. Products and Recommendations

We retrieve the Amazon products by identifying potential configurations of the products in the Amazon Review database. For example, "belt" is found with the Amazon category `Clothing_Shoes_and_Jewelry` with the keywords `belt`, `belts`, `leather belt`. The full configuration is found in the Github repository. We use the product title, features, and description to create the embeddings—when showing the recommendation set for evaluation, the average rating and the number of ratings is also included.

Recommendations are generated according to the procedure in Section 4.2. All persona-recommendation pairs are prepared in advance of any evaluations.

E.3. Evaluations

The LLM-as-a-judge evaluations were performed in parallel calls to `gpt-5-mini` using the OpenAI Batch API. The prompt used is shown in Prompt E.1. To account for some of the randomness in model outputs, we average outcomes over 3 runs for each persona-recommendation pair. Note that revenue graphs show the total willingness to pay of all personas, and those who do not purchase are assumed to generate zero revenue.

E.3.1. Prompt for LLM-based user evaluation of recommendation set

Prompt E.1: LLM-as-a-judge Prompt

```
system_prompt: f"""You are a customer with the following specific needs and preferences:
{ground_truth_need}
Evaluate the products based on these needs. Consider whether a product could reasonably meet your requirements, even if not every detail is
explicitly confirmed in the description."""
user_prompt: f"""Based on your needs, here are {len(products)} recommended products:
{products_text}
Please evaluate these products and respond in the following JSON format:
{{
    "decision": "PURCHASE" or "NO_PURCHASE",
    "product_number": <1-{len(products)} if purchasing, null if not>,
    "product_title": "<title of chosen product if purchasing, null if not>",
    "willingness_to_pay": <dollar amount you'd pay based on how well it fits your needs, null if not purchasing>,
    "reasoning": "<brief explanation of your decision, 2-3 sentences>"
}}
Choose NO_PURCHASE only if none of these products would be reasonable for your needs at all."""
```

E.3.2. Evaluation Pipeline

Figure 14 shows the evaluation pipeline in detail.

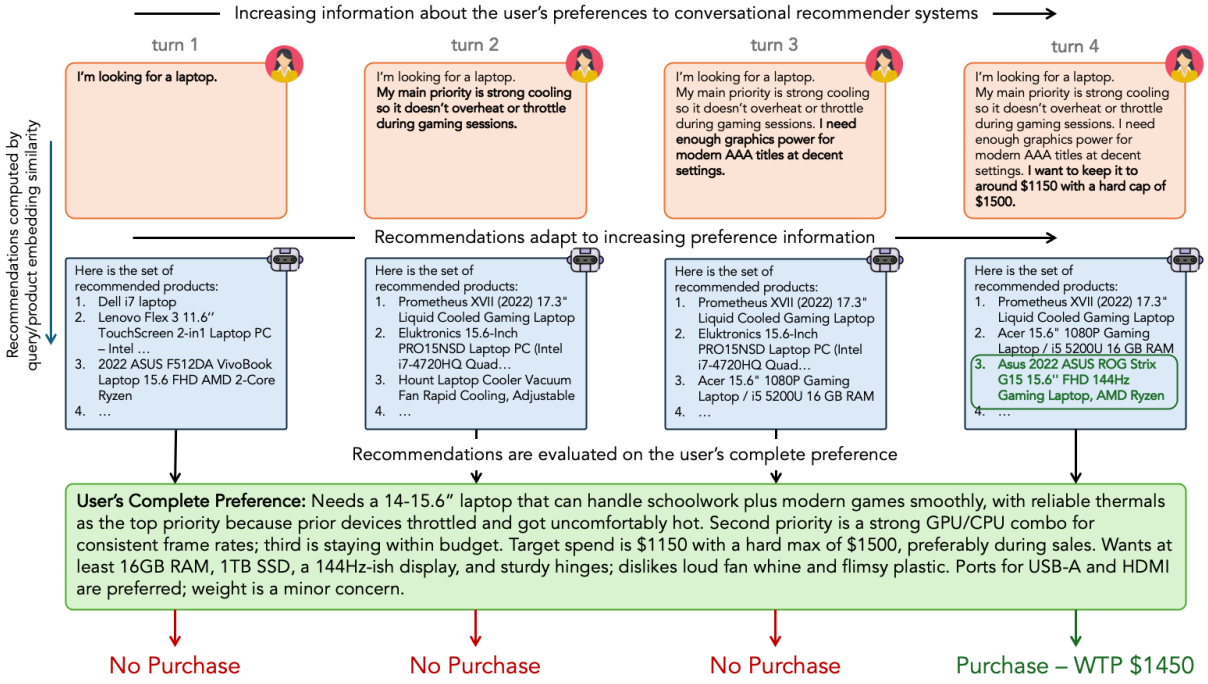


Figure 14 The procedure for evaluation. The queries become more detailed, leading to different recommendation sets. All sets are then evaluated against the user's complete preference to determine outcomes.