MEASURING META-CULTURAL COMPETENCY IN ACTION: A BEHAVIORAL FRAMEWORK FOR LLM-BASED PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalization requires a system to know when it has enough information about a user and when it should ask for more. These abilities correspond to two capacities of meta-cultural competence: variational awareness, or sensitivity to the limits of one's knowledge about a user, and explication strategy, or the ability to elicit missing information efficiently. Existing evaluations of meta-cultural competence mostly measure static model-implied distributions. They do not test whether models can use uncertainty to guide interaction. We propose a dynamic explore-exploit framework for evaluating meta-cultural competence behaviorally in conversational recommender systems. The framework measures whether models explore when initial user context is limited, exploit when context is sufficient, and ask informative questions when they explore. We instantiate it as a multi-turn movie recommendation task using simulated users from MovieLens-1M, and evaluate 11 models from 6 families. While static evaluation reports a plateau in meta-cultural competence beyond a certain model size, our dynamic framework detects statistically significant differences across models above that threshold. This separates them into distinct performance tiers, and we find evidence that model size increases meta-cultural competence in action. These findings suggest that measuring meta-cultural competency through distributions and behavior captures different aspects of model capacity.

1 INTRODUCTION

Meta-cultural competence in AI systems describes their ability to adapt across users and contexts. It includes *variational awareness*, the ability to recognize how users may vary across domains and dimensions and how uncertain the system is about the user it faces, and *explication strategy* , the ability to elicit missing information through interaction (Sharifian, 2013; Leung et al., 2013; Noshadi & Dabbagh, 2015; Saha et al., 2025). Variational awareness guides proper explication strategy, by informing whether and when to ask questions, as well as what questions to ask. These capacities are crucial for large language model (LLM)-based personalization systems, such as conversational recommendation systems (CRS), which must do two things to personalize well: (i) estimate their knowledge boundary, or determine whether they know enough about a user to act, and (ii) acquire missing information efficiently when their current knowledge is insufficient.

Existing evaluations of meta-cultural competence in LLMs have mostly studied it statically, by measuring whether model-implied distributions over possible responses align with real-world cultural distributions (Saha et al., 2025; 2026). However, such evaluations capture only what models represent, and not how they behave in user-facing settings where assumptions must be revised through interaction. For example, a model may represent uncertainty over several possible responses and still fail to ask for clarification before acting. Conversely, it may also state uncertainty but behave as if it were certain. Prior work has similarly shown that LLMs can be miscalibrated and that their decisions do not always reflect their stated uncertainty (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Wang et al., 2026). Conversational recommendation provides a natural setting to study this gap because the system must repeatedly choose between asking for more information and making a recommendation. However, most CRS evaluations focus on final prediction or recommendation quality, rather than whether the system explores appropriately (Wang et al., 2023; Li et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025). Final accuracy alone does not fully measure meta-cultural competence: a model may reach the correct prediction through a shortcut, or make an incorrect prediction despite taking reasonable steps under limited information. Thus, neither representational alignment nor final recommendation quality establishes whether a model manages uncertainty competently during interaction.

In this paper, we propose a dynamic *explore-exploit* framework for evaluating meta-cultural competence behaviorally in CRS. The framework introduces three metrics that jointly assess the two capacities we study: variational awareness and explication strategy. These capacities are measured by whether a model *explores* when prior information is limited and *exploits* when sufficient information is available, and whether the model asks questions that efficiently improve downstream prediction. Together, these metrics evaluate how a model manages uncertainty throughout an interaction, rather than only whether its final prediction is correct.

We evaluate our framework with 11 models from 6 model families using a multi-turn movie recommendation task constructed from the MovieLens-1M dataset (Harper & Konstan, 2015). We vary the amount and type of user information available before interaction and evaluate the models' decisions over repeated interactions with simulated users. While static evaluation reports a plateau in meta-cultural competence beyond a certain model size (Saha et al., 2026), our dynamic framework separates the models into distinct performance tiers above that threshold. Moreover, within every open-weight family, the larger variant significantly outperforms the smaller one. The one exception is the API-accessed Gemini-2.5 family, where the lower-tier Flash variant significantly outperforms the higher-tier Pro. These results suggest that static knowledge distributions and interactive behavior capture different aspects of meta-cultural competence, and that one cannot be reliably inferred from the other. Although our experiments focus on movie recommendation, the same explore-exploit decision appears in many LLM-mediated settings where systems personalize under uncertainty, including education, search, writing support, health information, and cross-cultural communication (Kasneci et al., 2023; Radlinski & Craswell, 2017; Lee et al., 2022; Singhal et al., 2023; Saha et al., 2025). The code and datasets used in this paper can be found in this repository.

2 BACKGROUND

We briefly introduce the three core concepts used in the paper: the explore-exploit trade-off, meta-cultural competence, and conversational recommender systems.

Explore-Exploit trade-off is central to sequential decision-making, where systems must choose between acquiring information and acting on what they currently know (Kaelbling et al., 1996). It is commonly formalized through multi-armed bandit problems (Robbins, 1952; Auer et al., 2002; Lattimore & Szepesvári, 2020; Slivkins, 2024). An agent can explore by selecting actions that improve its estimate of the reward distribution, or exploit by selecting the action that currently appears best. The same trade-off also appears in Bayesian optimization (Shahriari et al., 2016), where an acquisition function must balance learning about an unknown objective with optimizing it under costly evaluations. In user-facing systems, the cost of exploration also includes the burden placed on the user in terms of their time, attention, and cognitive effort (Rashid et al., 2002; Boutilier et al., 2003; Sweller, 1988; Jannach et al., 2021).

Meta-cultural Competency refers to the ability to communicate and adapt across cultural contexts (Sharifian, 2013). For AI systems, it is described through three capacities (Saha et al., 2025). First, **variational awareness** is the ability to recognize that behaviors, preferences, and interpretations vary across users and contexts, and to estimate one's familiarity with the current user. Second, **explication strategy** is the ability to ask clarifying questions when the system lacks sufficient information, preferably with as few user inputs as possible. Third, **negotiation strategy** is the ability to resolve misunderstandings through joint interaction. In this work, we focus on variational awareness and explication strategy, which directly govern whether and how a system seeks information.

These two capacities map naturally onto the explore-exploit trade-off. Variational awareness concerns whether the system recognizes the limits of its current understanding. This awareness motivates its explication strategy, informing its decision about whether to explore or exploit, and what it asks when it decides to explore. Since explication is sample-efficient, it also captures whether the system can reduce uncertainty without placing unnecessary burden on the user. From this per-

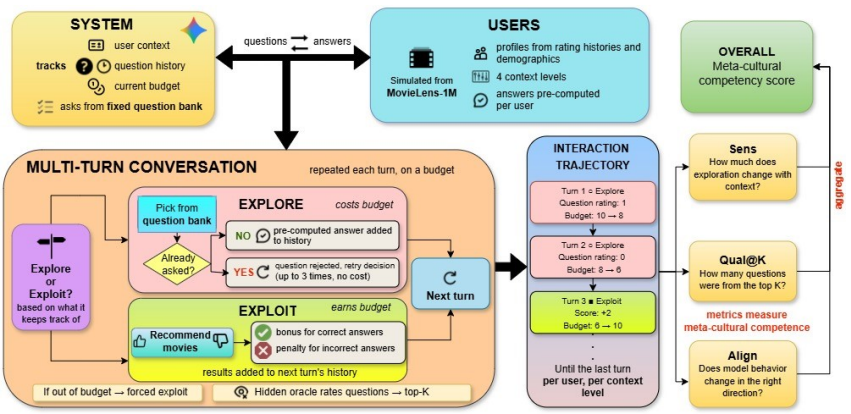

Figure 1: Evaluation game framework

spective, a meta-culturally competent personalization system should estimate its uncertainty, ask informative questions, and stop asking when its current evidence is sufficient.

Conversational recommender systems aim to elicit user preferences through dialogue and recommend suitable items based on the information elicited (Christakopoulou et al., 2016; Sun & Zhang, 2018; Lei et al., 2020a; Jannach et al., 2021). Many CRSs operate in a multi-turn setting, where the system repeatedly asks about preferences, recommends items, and updates its strategy from user feedback (Lei et al., 2020b; Zhang et al., 2022; Sepliarskaia et al., 2018; Graus & Willemsen, 2015; Li et al., 2021). This makes CRSs a natural testbed for explore-exploit behavior. At each turn, the system can explore by asking about the user's preferences, or exploit by recommending items from its current understanding. Conversational-bandit approaches make this connection explicit by modeling recommendation as sequential decision-making under uncertainty (Li et al., 2010; Zhang et al., 2020; Li et al., 2021).

This setting also exposes the connection to meta-cultural competence. A CRS must decide when to ask and what to ask (Lei et al., 2020a; Deng et al., 2021). These decisions are part of the system's explication strategy and are guided by its variational awareness: does the system know when it lacks enough information, and how it can efficiently reduce this uncertainty? Traditional CRSs often depend on the coverage of user interaction data available during training, which makes cold-start settings difficult. LLM-based CRSs offer a different possibility because they can interpret free-form preferences and produce recommendations without task-specific training (Friedman et al., 2023), while showing competitive recommendation performance (He et al., 2023). However, it remains unclear whether they manage the interaction itself well: exploring when they know little, committing when they know enough, and asking informative questions when they choose to explore. Recommendation performance alone cannot answer this question, as it does not reveal whether a model relied on prior assumptions, adapted to the user through interaction, or asked useful questions. Hence, we evaluate LLM-based CRSs through the two capacities of meta-cultural competence that correspond to the explore-exploit trade-off: variational awareness and explication strategy. We operationalize these capacities as behavioral metrics, discussed next.

3 Framework

Our proposed framework evaluates a system's meta-cultural competency as a behavioral property under uncertainty. The setup is a finite-horizon sequential evaluation game between a system θ and a set of m users $\mathcal{U} = \{u_1, \ldots, u_m\}$. The game resembles a contextual bandit setting, where at each turn the system must choose between *exploring:* gathering more information and *exploiting:* acting on its current knowledge. Let $\mathcal{L} = \{\ell_1, \ldots, \ell_c\}$ denote an ordered set of c context levels, with $\ell_1 < \cdots < \ell_c$, where a higher context level means that the system receives more information about the user before the interaction begins. We evaluate each user $u \in \mathcal{U}$ under different context levels $\ell \in \mathcal{L}$. For each pair (u, ℓ), the system starts with a budget B_0 and interacts with the user for R

rounds. At turn t, the system observes a state $s_t = (\ell, h_t, B_{t-1})$, where h_t is the interaction history so far and B_{t-1} is the remaining budget.

The system then makes a two-stage decision. First, it chooses a mode $z_t \in \{\text{explore}, \text{exploit}\}$. If $z_t = \text{explore}$, the system is presented with a set of k candidate questions $\mathcal{Q} = \{q_1, \ldots, q_k\}$ and selects one question $q \in \mathcal{Q}$ to gather more information about the user. Let $H_t \subseteq \mathcal{Q}$ denote the set of questions already asked before turn t. If $q \notin H_t$, the user's answer is added to the interaction history at the end of the round. Since asking questions is costly, exploration incurs a cost λ_-, which we deduct from the budget: $B_t = B_{t-1} - \lambda_-$. If $q \in H_t$, we reject the selection and ask the system to choose again, either by selecting a different question or by revising its mode choice. We allow up to three such retries, treating each retry as an additional turn with no change to the budget. If all three retries select previously asked questions, or if exploration depletes the budget, the system must exploit. Each round therefore spans one to four turns, and the total number of turns T in a run satisfies $R \leq T \leq 4R$.

If $z_t = \text{exploit}$, either by choice or because exploration is no longer affordable, the system is presented with a set of n candidate movies and predicts whether the user will like each. Since the goal of a recommender system is to recommend, exploitation incurs no cost. Instead, it adds a positive value λ_+ to the budget, regardless of whether the recommendation is correct: $B_t = B_{t-1} + \lambda_+$. We additionally adjust the budget by η, with $\eta < \lambda_-, \lambda_+$: each correct recommendation adds η, while each incorrect recommendation subtracts η. Thus, the system must decide whether asking is worth its cost, and, when it asks, which question to ask.

When the system explores, each question q has an oracle utility $v_\ell(q) \in [0, 1]$, which measures how informative the answer to q is for the downstream recommendation task at context level ℓ. This utility depends on the initial context level, since a question may be useful when little is known about the user, but redundant when the context already has similar information. The utility value is used only during evaluation and not shown to the system. The motivation follows active preference elicitation and active collaborative filtering, where useful questions are those whose answers are expected to improve later recommendations (Boutilier et al., 2003; Rashid et al., 2002; Christakopoulou et al., 2016). Variational awareness concerns whether the system recognizes the limits of its current understanding of a user given the amount of context it has. Explication strategy concerns whether the system uses this awareness to adjust its exploration behavior and whether it asks useful questions once it decides to explore. The framework jointly evaluates two capacities: variational awareness and explication strategy. We summarize these behaviors using three metrics: *context sensitivity*, *question quality*, and *trend alignment*.

1. Context Sensitivity (Sens) captures whether the system's behavior changes depending on the context. For each context level ℓ, we first compute the system's exploration rate. Let $\mathbb{1}[z_t = \text{explore}]$ indicate whether the system chooses to explore at turn t, and let $\mathbb{1}[B_{t-1} \geq \lambda_-]$ indicate whether exploration is affordable at that turn. For a decay parameter $\gamma \in (0, 1]$, we first compute the system's exploration rate as:

$$r_{\theta,\ell} = \frac{1}{m} \sum_{u \in \mathcal{U}} \frac{\sum_{t=1}^{T} \gamma^{t-1} \mathbb{1}[B_{t-1} \geq \lambda_-] \mathbb{1}[z_t = \text{explore}]}{\sum_{t=1}^{T} \gamma^{t-1} \mathbb{1}[B_{t-1} \geq \lambda_-]},$$

where the denominator includes only turns where exploration was affordable, and the decay term γ gives more weight to earlier decisions; setting $\gamma = 1$ removes this weighting. **Context Sensitivity** of a system is defined as the spread of exploration rates across context levels:

$$\text{Sens}(\theta) = \max_{\ell \in \mathcal{L}} r_{\theta,\ell} - \min_{\ell \in \mathcal{L}} r_{\theta,\ell} \in [0, 1] \tag{1}$$

A higher score means that the system changes its exploration behavior more strongly across context levels. However, this metric does not capture whether the change is in the expected direction.

2. Question Quality (Qual@K) measures whether the system asks one of the K most informative questions during exploration. At each turn t, we maintain a set of top-K questions with the highest utility $v_\ell(q)$ that have not been asked yet. The top-K set is replaced by the next K questions when exhausted. The system-selected question q_t is scored 1 if it belongs to the current top-K questions, and 0 otherwise. A score of 0 is assigned if the system repeats a question that it already asked before. For a decay rate $\gamma \in (0, 1]$, we weight each question selection by γ^{t-1}, and first compute

the explication quality as:

$$e_{\theta,\ell} = \frac{1}{m} \sum_{u \in \mathcal{U}} \frac{\sum_{t=1}^{T} \gamma^{t-1} \mathbb{1}[q_t = \text{top}{-}K] \mathbb{1}[z_t = \text{explore}]}{\sum_{t=1}^{T} \gamma^{t-1} \mathbb{1}[z_t = \text{explore}]}$$

System-level **question quality** is defined as the explication quality pooled across context levels:

$$\text{Qual@K}(\theta) = \frac{1}{c} \sum_{\ell \in \mathcal{L}} e_{\theta,\ell} \in [0, 1] \tag{2}$$

A higher value means that, conditional on exploration, the system more often selects one of the most informative available questions.

3. Trend Alignment (Align) measures whether exploration and explication occur in the right direction. A system may change its exploration rate across contexts, but in the wrong direction. For a variationally aware system, the expected trend is that exploration should decrease as context increases. Thus, sensitivity alone is not enough. We define a correlation score between the context level and exploration rate: $\rho_{\text{Rate}}(\theta) = -\rho_p([1, \ldots, c], [r_{\theta,\ell_1}, \ldots, r_{\theta,\ell_c}])$, where ρ_p is Pearson correlation. A value closer to 1 means that exploration decreases as context increases.

We also measure whether explication quality is stable across context levels. Since explication is conditional on exploration, the ideal behavior is not necessarily that question quality increases or decreases with context; rather, if the system chooses to ask, it should ask useful questions across context levels. We therefore compute a correlation score between the context level and explication quality: $\rho_{\text{EXP}}(\theta) = -\rho_p([1, \ldots, c], [e_{\theta,\ell_1}, \ldots, e_{\theta,\ell_c}])$. The ideal value for this term is 0, indicating that question quality is not systematically driven by context level, as expected. Finally, we define system-level **Trend Alignment** $\text{Align}(\theta)$ as the normalized closeness to the ideal point $(\rho^*_{\text{Rate}}, \rho^*_{\text{EXP}}) = (1, 0)$:

$$\text{Align}(\theta) = 1 - \frac{1}{d} \sqrt{(\rho^*_{\text{Rate}} - \rho_{\text{Rate}}(\theta))^2 + (\rho^*_{\text{EXP}} - \rho_{\text{EXP}}(\theta))^2} \tag{3}$$

We set $d = \sqrt{5}$ to normalize the score to $[0, 1]$, since $\rho_{\text{Rate}}, \rho_{\text{EXP}} \in [-1, 1]$. A score closer to 1 means that the system follows the expected behavioral pattern: it explores less as initial context increases, while maintaining question quality when it does explore.

Together, the three metrics capture different aspects of meta-cultural behavior. **Sens** measures whether exploration changes with context. **Qual** measures whether exploration is useful. **Align** measures whether the observed trends match the expected direction. Thus, a system can be sensitive but misaligned, aligned but weak at asking useful questions, or good at asking questions but insensitive to how much it already knows.

4 Experimental Setup

We experiment with the framework as an R=20-round movie recommendation game.

User Simulation. We use preference data from MovieLens-1M (Harper & Konstan, 2015), where each user has rated movies that we binarize as Like (4–5 stars) or Dislike (1–3 stars). We select $m = 2038$ users with at least 50 liked and 50 disliked movies. For each user, we draw 5 liked and 5 disliked movies from their ratings as preference context, and reserve the other movies for evaluation, so the two pools never overlap. To reduce variance from simulated user responses, we pre-compute answers to the candidate questions using each user's preference and demographic data. During exploitation, the system is shown the next batch of 4 movies from the evaluation pool in shuffled order, and predicts a Like or Dislike label for each. Each correct prediction adds $\eta = 1$ to the budget, and each incorrect prediction subtracts it. The predictions and true labels are then shown to the system at its next explore-exploit decision.

Context levels. We initialize the evaluation game with four ordered context levels that vary in the amount and type of user information provided: ℓ_1, no user-specific context; ℓ_2, user demographics (age, gender, occupation); ℓ_3, user movie preferences (5 liked and 5 disliked movies); and ℓ_4, demographics and movie preferences combined. We treat explicit preference information (ℓ_3) as more

informative than demographic information alone (ℓ_2), since a user's past preferences are generally stronger predictors of future preferences than demographic attributes (Rashid et al., 2002).

Question Construction and Tiering. We construct a fixed pool of $k = 15$ questions including preference-based questions, demographic questions, and distractors. Each question is assigned to one of three utility tiers based on how much information it provides about the target user's preferences, given the information already available at that context level: T1 (high utility), T2 (medium utility), and T3 (low or no utility). We assign these tiers normalized utility scores $v_\ell(q) = 1.0, 0.5,$ and 0.0, respectively. Tier assignment is context-dependent because a question's usefulness depends on what has already been revealed. For example, asking about the user's gender may be useful when demographic information is absent (ℓ_1, ℓ_3), but has no utility when demographics are already provided (ℓ_2, ℓ_4). We derive the tiers by combining rankings from 3 human experts with rankings simulated by GPT-5.5 for each context level. For the simulated rankings, we create 18 user personas with different demographic attributes (Appendix A.3) and ask GPT-5.5 to rank all 15 questions from most to least useful for predicting each user's movie preferences (Appendix A.3). We average the rankings across personas to obtain a mean rank for each question in each condition (Appendix, Table 5), and divide the questions into three tiers of five questions each (Table 6). With five questions per tier, the initial top-5 set corresponds to T_1 and the top-10 set to $T_1 \cup T_2$.

Experimental Conditions. At each context level, we vary 2 experimental factors: **1. Two target visibility** settings, which are used to determine whether the system can see the target movies before making its explore-exploit decision in each round. In the SEEN setting, the system can view this information; in the UNSEEN setting, this information is withheld. This comparison tests whether a system conditions its decision on the specific items it is about to be asked about, not only on what it knows about the user. **2. Two exploration penalties** $\lambda_- \in \{2, 5\}$, while fixing the **exploitation gain** (λ_+) at $+2$. The 4 context levels $\times$ 2 target visibility settings $\times$ 2 exploration penalties yield a 16-condition experimental grid. Each model is evaluated on all 16 conditions over $m = 2038$ users for $R = 20$ rounds. Each run starts with a budget $B_0 = 10$, and the budget is kept between 0 and B_0. We compute all metrics for $\gamma \in \{0.3, 0.5, 0.7, 0.9\}$ and $K \in \{5, 10\}$. Confidence intervals are 95% percentile intervals from 5,000 paired user-level bootstrap replicates, and pairwise comparisons are Benjamini–Hochberg corrected at $\alpha = 0.05$ (Appendix A.4).

Models. We evaluate 11 models from 6 families: Llama-3.1 (8B, 70B), Aya-Expanse (8B, 32B), Gemma-2 (2B, 9B), Qwen3.5 (9B, 27B), DeepSeek-R1-Distill-Llama-8B, and Gemini-2.5 (Flash, Pro). Reasoning was disabled for DeepSeek-R1 and Qwen3.5. Evaluating models of different parameter counts lets us ask whether context-adaptive behavior emerges with scale. The experiments on the 5 open-weights model families were run using vLLM (Kwon et al., 2023) on GPUs (RTX 5000 Ada, RTX 6000 Pro, A40). We used the Batch API for the Gemini models, while running duplicate/repetitive steps using standard inference. For Gemini-2.5-Pro, the `thinking` mode was set to "medium". We set the sampling temperature to 1.0. The estimated cost of running the Gemini experiments was $5,800.

Baselines. We compare the models' performance against the following random baselines: **(i) Uniform:** uniformly random choice between exploration and exploration, and random question selection, **(ii) Always-explore:** always choosing exploration, and **(iii) Oracle:**random choice between exploration and exploration, and oracle question selection. We provide implementation details in Appendix A.9.

5 Results

We report scores for (target visibility=SEEN, $\lambda_- = 5$, $\gamma = 0.9$, $K = 5$), with ablations over the other setups in Appendix A.8.

Overall Results. Figure 2 reports each model's Context **Sens**itivity, Question **Qual**ity, and Trend **Align**ment scores. We observe: **(i) Qual:** Both Gemini-2.5 models and Qwen3.5-27B lead in question quality, whereas Gemma-2-2B and DeepSeek-R1-Distill-Llama-8B have the lowest scores and perform worse than random question selection, as they repeatedly select low-quality or duplicate questions. **(ii) Sens:** Gemma-2-9B and Qwen3.5-27B are the most sensitive to the amount of context, whereas Llama-3.1-70B, Aya-Expanse-8B, Gemma-2-2B and Qwen3.5-9B are the least sensitive. **(iii) Align:** Both Gemma models, Qwen3.5-27B, and Gemini-2.5-Flash achieve near-perfect

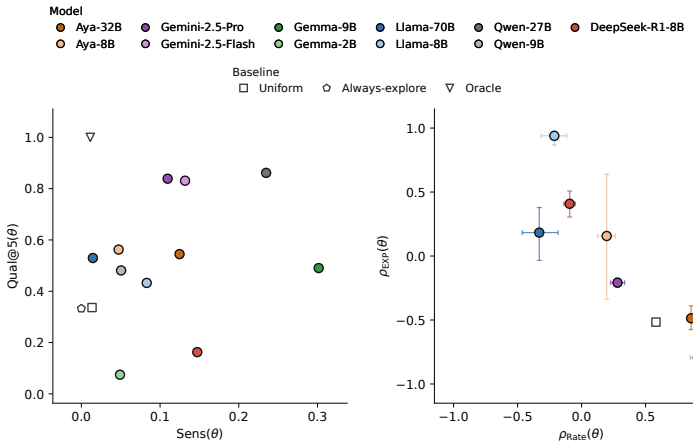

Figure 2: (left) Sens(θ)$vs.Qual@5(\theta)$, $(right)\rho_{\mathrm{Rate}}(\theta)$ vs. $\rho_{\mathrm{EXP}}(\theta)$

Model	Sens	Qual@5	Align	Overall	95% CI	Rank	Rank CI	Tier	Win rate
Qwen3.5-27B	0.235	**0.862**	0.882	**0.660**	[0.615, 0.696]	1	[1, 2]	1	100.0%
Gemini-2.5-Flash	0.132	0.831	0.926	0.630	[0.610, 0.650]	2	[1, 2]	1	100.0%
Gemma-2-9B	**0.302**	0.490	0.887	0.560	[0.523, 0.593]	3	[3, 4]	2	77.8%
Gemini-2.5-Pro	0.110	0.839	0.665	0.538	[0.529, 0.546]	4	[3, 4]	2	77.8%
Aya-Expanse-32B	0.125	0.545	0.771	0.480	[0.468, 0.493]	5	[5, 5]	3	60.0%
Aya-Expanse-8B	0.047	0.564	0.633	0.415	[0.383, 0.429]	6	[6, 7]	4	44.4%
Qwen3.5-9B	0.050	0.481	0.638	0.390	[0.378, 0.407]	7	[6, 7]	4	44.4%
Gemma-2-2B	0.049	0.075	**0.961**	0.362	[0.348, 0.372]	8	[8, 8]	5	30.0%
Llama-3.1-70B	0.015	0.530	0.403	0.316	[0.295, 0.337]	9	[9, 9]	6	20.0%
Llama-3.1-8B	0.083	0.432	0.313	0.276	[0.263, 0.290]	10	[10, 11]	7	0.0%
DeepSeek-R1-Distill-Llama-8B	0.146	0.163	0.479	0.263	[0.254, 0.271]	11	[10, 11]	7	0.0%
Baselines									
Oracle questions	0.011	1.000	n/a	n/a	n/a	–	–	–	–
Uniform ($p = 0.5$)	0.014	0.336	0.703	0.351	[0.249, 0.420]	–	–	–	–
Always-explore	0.000	0.333	n/a	n/a	n/a	–	–	–	–

Table 1: Overall ranking across models, with non-LLM random baselines for reference. The best value among models is highlighted in **bold**. **Rank** is the median rank obtained from bootstrapping. We assign **Align** and **Overall** as "n/a" in cases where Pearson correlation (ρ_{Rate} or ρ_{EXP}) is undefined.

correlations with exploration rate. In contrast, both Llama models and DeepSeek-R1-Distill-Llama-8B have negative ρ_{Rate} correlations, indicating exploration patterns that oppose the expected direction, as well as having Align point estimates that are lower than random baselines. Correlations for explication quality are more uncertain: five models have confidence intervals that include zero, leaving their direction unresolved. Llama-3.1-8B and Qwen3.5-9B display the strongest positive and negative correlation, respectively.

The three metrics capture distinct aspects: Sensitivity and trend alignment capture different behavioral aspects. The Gemma-2 models demonstrate the **importance of tracking magnitude of adaptation (Sensitivity) along with directionality of alignment**. Both models show near-perfect alignment, but Gemma-2-9B has substantially higher sensitivity compared to Gemma-2-2B. On the other hand, DeepSeek-R1-Distill-Llama-8B is comparatively more sensitive than other models, but it also has a negative ρ_{Rate} correlation - its exploration changes with context, but not in the expected direction. Sensitivity and question quality also capture different behavioral aspects: Gemma-2-2B's strong alignment contrasts with its having the lowest question quality among all models. On the other hand, Gemini-2.5-Pro asks high-quality questions, but its exploration rate does not adapt as well to the amount of context. Therefore, **"exploring well" and "exploring at the right time" are separate skills.**

Family	Model comparison	Difference	95% CI
Qwen	Qwen3.5-27B vs. Qwen3.5-9B	0.270	[0.220, 0.308]
Gemma	Gemma-2-9B vs. Gemma-2-2B	0.198	[0.159, 0.234]
Aya	Aya-Expanse-32B vs. Aya-Expanse-8B	0.066	[0.049, 0.101]
Llama	Llama-3.1-70B vs. Llama-3.1-8B	0.040	[0.014, 0.064]
Gemini	Gemini-2.5-Pro vs. Gemini-2.5-Flash	−0.092	[−0.114, −0.070]

Table 2: Within-family Overall score differences with 95% confidence intervals.

These contrasting profiles are consistent with the weak to moderate Spearman correlations observed between the metrics across all models: Sens and Qual@5 ($\rho = 0.22$, $p = 0.52$), Sens and Align ($\rho = 0.39$, $p = 0.24$), and Qual@5 and Align ($\rho = 0.18$, $p = 0.59$). This justifies reporting the three metrics separately. We consider a model to be meta-culturally competent in this setting only if it does well on all three.

We rank the models using an **Overall** score, the unweighted mean of the 3 metrics, and report it in Table 1 with rank intervals and each model's win rate among its significant pairwise comparisons. We observe that **consistent models tend to have better Overall scores.** No model is best on all three metrics: Gemma-2-9B has the highest Sens, Qwen3.5-27B has the highest Qual@5, and Gemma-2-2B has the highest Align. The two models with the highest Overall scores (Qwen3.5-27B and Gemini-2.5-Flash) are the only models ranked in the top four on every component. The top of the Overall ranking is therefore occupied by models with consistently strong component profiles rather than models that excel on only one dimension. Based on the rank confidence intervals, we obtain **7 model tiers**: (i) Qwen3.5-27B and Gemini-2.5-Flash; (ii) Gemma-2-9B and Gemini-2.5-Pro; (iii) Aya-Expanse-32B; (iv) Qwen3.5-9B and Aya-Expanse-8B; (v) Gemma-2-2B; (vi) Llama-3.1-70B; and (vii) Llama-3.1-8B, DeepSeek-R1-Distill-Llama-8B.

Model Size Comparisons. We examined model size within the five multi-model families, excluding DeepSeek (single model). Because Gemini parameter counts are unavailable, Gemini-2.5-Flash and Gemini-2.5-Pro were treated as the lower- and higher-tier variants, respectively. We observe that **the larger model variants achieved higher Overall scores in all four open-weights families**, whereas **Gemini-2.5 showed the opposite pattern**. All five within-family differences remained statistically significant after Benjamini-Hochberg correction, with the largest differences for the Qwen and Gemma families. Across families, the higher-tier variants had an unweighted mean difference of 0.096 points over the smaller variants. However, an exact one-sided sign-flip test did not find a statistically significant advantage ($p = 0.125$). These results indicate performance differences between model variants within families, especially in open-weights families, but there is insufficient evidence that larger or higher-tier models achieve higher Overall scores across families.

Model ranking: In action vs. static. We additionally compare the model rankings to the macro-structural rankings obtained through static distributions (Saha et al., 2026). They reported a plateau in macro-structural ability for models beyond 2B parameters, with no reported statistically significant differences in competence beyond this threshold. In contrast, we find that **our dynamic framework detects statistically significant performance differences across models above 2B parameters**. For example, Llama-3.1-70B belongs to a higher performance tier than Llama-3.1-8B, with a statistically significant difference in Overall scores, whereas the static evaluation reported no significant difference between these models. Within open-weight families, we therefore find evidence that model size increases meta-cultural competency in action, which contrasts with the static evaluation's suggestion that macro-structural competence may not scale with model size.

6 Discussion

Implications for model training While prior work has reported instruction-tuned models plateauing in macro-structural ability beyond a certain model size (Saha et al., 2026), our framework displays evidence of behavioral differences between model variants within the same family. These findings suggest that while static evaluation finds comparable knowledge distributions, models differ in how they put their distributional knowledge into action, with larger variants appearing more competent in practical settings. This could be due to differences in context processing, knowledge utilization,

or other training factors. Smaller models are not devoid of meta-cultural competence: Gemma-2-2B adjusts its exploration in the expected direction as reliably as any model we test, yet it rarely asks the most informative available questions. This suggests that **operationalizing distributional knowledge is a capacity distinct from possessing it**, and not something measured by static benchmarks. The uneven performances across different components indicate that most models still struggle to consistently exhibit meta-culturally competent behavior in conversational settings. Future training efforts should focus on how models translate their distributional knowledge into context-sensitive behavior. Our framework identifies concrete targets: adapting the extent and direction of exploration to available context, and selecting useful questions to efficiently learn more about the user.

What do these findings reflect about conversational recommenders? Our results show that the two decisions a CRS must make, when to ask and what to ask (Lei et al., 2020a; Deng et al., 2021), come apart in practice. Gemini-2.5-Pro selects high-utility questions but barely modulates its exploration rate with its existing context, so it asks well at the wrong times. Gemma-2-2B shows the opposite: it adjusts exploration in the expected direction across context levels, but asks the least informative questions and averages the most interactions per run of any model. In deployment, both would impose costs that a final-accuracy evaluation would not surface: unnecessary user effort in the first case, and many uninformative turns in the second. Longer dialogues raise cognitive load and increase users' intention to abandon conversational systems (Sweller, 1988; Wu et al., 2026). Thus, **CRS evaluation should report process-level behavior alongside recommendation quality**, extending calls to rethink CRS evaluation (Jannach, 2022; Wang et al., 2023).

How do open-weight models behave compared to API-based models? Our framework shows that some open-weight models match or exceed API-based models at managing uncertainty during interaction. Qwen3.5-27B and Gemma-2-9B share the same tier with Gemini-2.5-Flash and Gemini-2.5-Pro respectively. Notably, Gemini-2.5-Flash, marketed for cost-effectiveness, outperforms the higher-tier Gemini-2.5-Pro in overall meta-cultural competency, as it displays greater adaptability to context while only falling slightly short in question selection quality. This shows that **general-purpose capability does not predict meta-cultural competence in action.** In interactive settings, models should also be chosen on how they behave in a conversation.

7 Limitations and Conclusion

Limitations: Our study has a few limitations. **First,** we study a single dataset, MovieLens, within a single recommendation domain. Further evaluation is needed to test whether our findings generalize to other datasets and domains. The MovieLens dataset is also very Western-centric, and future work could involve strengthening evaluation using datasets from other cultures and languages. **Second,** the scope of evaluation can also be improved by including other model families, especially for API-based models as we only test two. **Third,** question tiering relies on GPT-5.5-simulated rankings over 18 personas combined with judgments from 3 human experts, which can be further validated by a larger set of evaluators with different exploration strategies. The framework can also be strengthened by evaluating on a more extensive question set. **Fourth,** we simulate users to make the experiments tractable. Future work can involve tests with real users, or stronger models as user simulators. **Fifth,** all experiments are conducted in English. While the property of meta-cultural competency should ideally be invariant to language, this is not yet established by our framework.

Conclusion. Personalization systems must recognize when they know enough about a user and ask informative questions when they do not. However, existing evaluations measure only what models represent. Our results show that dynamic evaluation offers a distinct lens on meta-cultural competence, one that static evaluation cannot substitute for. While static evaluation reports a plateau beyond 2B parameters, our dynamic framework separates the models above that threshold into seven performance tiers, with the larger variant in every open-weight family significantly outperforming the smaller one. Through our three metrics, our results also show that exploring at the right time and exploring well are separate skills. Finally, our results show that static and dynamic evaluation are complementary rather than competing. Static distributions establish whether a model represents the ways users vary at all, which is a precondition for acting on that variation. Dynamic evaluation establishes whether the model can use what it represents. Personalization systems should therefore be evaluated and trained on how they manage uncertainty during interaction, not only on what they represent or how accurate their final recommendations are.

AI USE STATEMENT

In this work, we used generative AI tools in two roles. First, as part of our research methodology, we used GPT-5.5 to generate rankings of the 15 exploration questions across 18 simulated user personas; these rankings were combined with rankings from three human experts to derive the question-utility tiers (Appendix A.3). The 11 models evaluated in Section 4 are the object of study in this work, not tools used to produce it. Second, we used LLMs to correct grammatical errors and to review and debug evaluation code. We did not use generative AI tools to propose hypotheses, formulate the conceptual framework or metrics, or interpret results and draw conclusions.

ETHICS STATEMENT

Our work conforms to the ICLR Code of Ethics. This work does not involve human subjects: all users are simulated from MovieLens-1M, a public dataset of anonymized movie ratings that we use in accordance with its license. The three experts who ranked the exploration questions are authors of this paper who provided informed consent. The dataset includes demographic attributes (age group, gender, occupation, zip code), which we use only to build simulated user profiles; zip codes are reduced to a coarse urban, suburban or rural label, and no attempt is made to identify individuals. MovieLens users are largely from the United States, so our simulated users do not represent the full diversity of real users, and our findings on which questions are informative may not transfer to other populations (Section 7). Our code and data-processing scripts are publicly released so these choices can be examined.

REPRODUCIBILITY STATEMENT

All code needed to reproduce our experiments is available at the anonymous repository linked in Section 1, including the datasets, the user simulator, the question-ranking prompts and utility tiers, the full evaluation pipeline for every model and condition, metric computation, the statistical analysis, and the ablation studies. The framework and metrics are defined in Section 3, and the experimental setup in Section 4. Appendix A.1 describes the data splits, the rules used to compute the simulated users' answers, the exact prompts and the handling of invalid outputs. Appendix A.2 lists the question pool; Appendix A.3 gives the ranking prompts, the mean ranks and the tier assignments; and Appendix A.4 describes the bootstrap and significance procedures. All sampling uses fixed random seeds. The open-weight models were run with vLLM and can be reproduced exactly up to sampling noise at temperature 1.0.

REFERENCES

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Craig Boutilier, Richard S. Zemel, and Benjamin M. Marlin. Active Collaborative Filtering. In *Conference on Uncertainty in Artificial Intelligence*, pp. 98–106, 2003. URL https://mlanthology.org/uai/2003/boutilier2003uai-active/.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 815–824, 2016. doi: 10.1145/2939672.2939746.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 1431–1441, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462913. URL https://doi.org/10.1145/3404835.3462913.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*, 2023.

Mark P. Graus and Martijn C. Willemsen. Improving the user experience during cold start through choice-based preference elicitation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pp. 273–276, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336925. doi: 10.1145/2792838.2799681. URL https://doi.org/10.1145/2792838.2799681.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), December 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 720–730, 2023.

Dietmar Jannach. Evaluating conversational recommender systems: A landscape of research. *Artif. Intell. Rev.*, 56(3):2365–2400, July 2022. ISSN 0269-2821. doi: 10.1007/s10462-022-10229-x. URL https://doi.org/10.1007/s10462-022-10229-x.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Computing Surveys*, 54(5), 2021. doi: 10.1145/3453154.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: a survey. *J. Artif. Int. Res.*, 4(1):237–285, May 1996. ISSN 1076-9757.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. ISSN 1041-6080. doi: 10.1016/j.lindif.2023.102274. URL https://www.sciencedirect.com/science/article/pii/S1041608023000195.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Mina Lee, Percy Liang, and Qian Yang. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3502030. URL https://doi.org/10.1145/3491102.3502030.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pp. 304–312, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450368223. doi: 10.1145/3336191.3371769. URL https://doi.org/10.1145/3336191.3371769.

Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 2073–2083, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403258. URL https://doi.org/10.1145/3394486.3403258.

Angela K-y Leung, Sau-lai Lee, and Chi-yue Chiu. Meta-knowledge of culture promotes cultural competence. *Journal of Cross-Cultural Psychology*, 44(6):992–1006, 2013.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 661–670, 2010. doi: 10.1145/1772690.1772758.

Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *ACM Trans. Inf. Syst.*, 39(4), August 2021. ISSN 1046-8188. doi: 10.1145/3446427. URL https://doi.org/10.1145/3446427.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Mahdi Noshadi and Ali Dabbagh. Metacultural competence: A benchmark for advances in applied ELT. In *Selected Papers of 2nd conference on Interdisciplinary Approaches to Language Teaching, Literature and Translation Studies*, pp. 52–62. Ferdowsi university of Mashhad, Khate Sefid English Language Group, 2015.

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pp. 117–126, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346771. doi: 10.1145/3020165.3020183. URL https://doi.org/10.1145/3020165.3020183.

Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th International Conference on Intelligent User Interfaces*, IUI '02, pp. 127–134, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134592. doi: 10.1145/502716.502737. URL https://doi.org/10.1145/502716.502737.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527 – 535, 1952.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. Meta-cultural competence: Climbing the right hill of cultural awareness. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8025–8042, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.408. URL https://aclanthology.org/2025.naacl-long.408/.

Sougata Saha, Madhur Jindal, Saurabh Kumar Pandey, Mahardika Krisna Ihsani, Alham Fikri Aji, and Monojit Choudhury. Measuring meta-cultural competency: A spectral framework for LLM knowledge structures. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3UHHW9Gzi7.

Anna Sepliarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 172–180, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240352. URL https://doi.org/10.1145/3240323.3240352.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

Farzad Sharifian. Globalisation and developing metacultural competence in learning English as an international language. *Multilingual Education*, 3, 06 2013. doi: 10.1186/2191-5059-3-7.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, and Yun Liu. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023.

Aleksandrs Slivkins. Introduction to multi-armed bandits, 2024. URL https://arxiv.org/abs/1904.07272.

Yueming Sun and Yi Zhang. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 235–244, 2018. doi: 10.1145/3209978.3210002.

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12 (2):257–285, 1988. doi: 10.1207/s15516709cog1202_4. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1202_4.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Jiawei Wang, Yanfei Zhou, Siddartha Devic, and Deqing Fu. Are LLM decisions faithful to verbal confidence? *arXiv preprint arXiv:2601.07767*, 2026.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. Rethinking the evaluation for conversational recommendation in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10052–10065. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.621.

S. Wu, G. Peng, S. Li, D. Cameron, J. Zhang, J. Zhang, and Q. Zhang. Textual cues, cognitive load, and social fatigue: Unveiling the reasons behind user discontinuance in conversational AI. *International Journal of Human-Computer Studies*, June 2026. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2026.103846. URL https://eprints.whiterose.ac.uk/id/eprint/241886/.

Xiaoying Zhang, Hong Xie, Hang Li, and John C. S. Lui. Conversational contextual bandit: Algorithm and application. In *Proceedings of The Web Conference 2020*, WWW '20, pp. 662–672, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380148. URL https://doi.org/10.1145/3366423.3380148.

Yiming Zhang, Lingfei Wu, Qi Shen, Yitong Pang, Zhihua Wei, Fangli Xu, Bo Long, and Jian Pei. Multiple choice questions based multi-interest policy learning for conversational recommendation. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 2153–2162, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512088. URL https://doi.org/10.1145/3485447.3512088.

A Appendix

A.1 Implementation Details

User data splits. We keep the MovieLens-1M users who have at least 50 liked and 50 disliked movies. Each user's rated movies are split into an evaluation pool, from which the movies to be predicted are drawn, and a separate context pool, from which the preference context is drawn. This ensures the movies shown as context never appear among the movies the model is asked to predict.

From the context pool, we select 5 liked and 5 disliked movies using a genre-diversity heuristic: we shuffle the pool with a fixed seed, collect up to 20 candidate movies that each add at least one genre not already covered, take the first 5 liked and the first 5 disliked candidates, and fill any shortfall from the rest of the context pool. Every user receives the same context movies and the same sequence of evaluation batches across models and conditions.

Question order. The 15 questions (Table 4) are shown to the model as a numbered list. The order of the list is drawn at random for each user and kept fixed for the whole run, so that no question benefits from always appearing first.

Pre-computed user answers. Each simulated user answers every question from a fixed profile computed in advance from their MovieLens ratings (on the original 1–5 scale) and demographic fields. Table 3 describes how each answer is derived. Random answers are drawn with a fixed seed.

Table 3: How the simulated user's answer to each question is derived.

ID	Question	Category	How the answer is derived
Q1	Favourite genres	Preference	The three genres with the highest mean rating.
Q2	Avoided genres	Preference	Genres the user has never rated, followed by their three lowest-rated genres; the first three are reported.
Q3	Classic vs. modern	Preference	Whether the user has rated more movies released before 1990 (*classic films*) or from 1990 onward (*modern productions*).
Q4	Rating consistency	Preference	The standard deviation of the user's ratings for movies in their top three genres (a movie counts once for each of its genres among the three): below 0.4 is *consistent*, above 0.8 is *varies a lot*, and anything in between is *somewhat consistent*.
Q5	Easy to please	Preference	*Easy to please* if the user's mean rating (each movie counted once per genre) is above 4.0 and the standard deviation of their per-genre mean ratings is below 0.4; *hard to impress* if the mean rating is below 3.2 and that standard deviation is above 0.8; otherwise *somewhere in between*.
Q6	Viewing frequency	Preference	Users are split at the tertiles of the number of movies they rated into *light*, *moderate*, and *avid* watchers.
Q7	Cinema generation	Demographic	Mapped from age group: under 18 and 18–24 map to the streaming/digital era, 25–34 through 50–55 to the blockbuster/VHS era, and 56+ to the pre-1970s era.
Q8	Area	Demographic	Mapped from the first three digits of the zip code: prefixes of major metropolitan areas map to *urban*, prefixes of predominantly rural states and regions map to *rural*, and all other valid US zip codes map to *suburban*. Malformed or non-US codes are answered *not available*.
Q9	Type of work	Demographic	Mapped from occupation: students, educators, writers, programmers, artists, scientists and lawyers map to *intellectual/creative*; technicians/engineers, tradesmen/craftsmen, farmers and homemakers to *practical/hands-on*; and all remaining occupations to *service-oriented*.
Q10	Occupation	Demographic	The MovieLens occupation, stated directly.
Q11	Age group	Demographic	The MovieLens age group, stated directly.
Q12	Gender	Demographic	The MovieLens gender, stated directly.
Q13	Favourite colour	Distractor	Drawn uniformly at random from the seven colours of the rainbow.
Q14	Lucky number	Distractor	Drawn uniformly at random from the integers 1 to 100.
Q15	Tea or coffee	Distractor	*Tea* or *coffee*, drawn uniformly at random.

Game prompts Each turn has a DECIDE call, in which the model chooses to explore or exploit. If it exploits, there is a PREDICT call, in which it labels the four movies. The DECIDE system prompt is shown below, with the explore cost λ_{-} filled in for each condition.

```
You are a movie recommendation assistant. You are helping predict
whether a user will like or dislike movies.

You are given a points budget. Each turn, you must choose between
EXPLORE and EXPLOIT.

WHAT EACH CHOICE MEANS:

  EXPLORE: Choose this if you feel you do not yet have enough
  information about the user to make confident predictions. You
  select one question from the list below to learn more about the
  user's preferences or background. You do not make any predictions
  this turn. You must choose a question that has not already been
  asked in this session.

  EXPLOIT: Choose this if you feel you have enough information about
  the user to make predictions. You predict whether the user will
  Like or Dislike 4 movies based on what you know about them so far.
  You do not ask any questions this turn.

AVAILABLE QUESTIONS (choose exactly one if you select EXPLORE):
   1. [question]
  ...
  2. [question]

SCORING RULES:

  EXPLORE costs -[explore cost] points.
  EXPLOIT earns +2 points, plus +1 per correct / -1 per wrong prediction.
  Best case (4 correct): +6  Worst case: -2

BUDGET RULES:
  Start at 10. Cannot exceed 10 or go below 0.
  If budget < [explore cost], you must EXPLOIT.

IMPORTANT RULES FOR EXPLORE:
  - Copy the question exactly as listed above.
  - Do not ask a question already asked this session.

Reply with ONLY one of:

EXPLORE
[exact question from the list]

EXPLOIT
[one short reason]
```

The DECIDE user message is assembled from the blocks below. Blocks marked in brackets appear only in the stated situation.

```
Turn [t] of 20.
Current budget: [budget] / 10.

[demographics shown: l2, l4]
User demographics: Age [age], Gender [gender], Occupation [occupation]

[preferences shown: l3, l4]
User's rated movie history (10 movies):
  "[title]" [[genres]] - [Like/Dislike]
  ...
```

```
[no information yet: l1 before any question is answered]
You have no information about this user yet.

[at least one question answered]
Information gathered so far:
  Question 1: [question]
  Answer 1: [answer]
  ...

[previous turn was an exploit]
Results from last exploit turn:
  "[title]" - predicted [Like/Dislike], actual [Like/Dislike] - [correct/incorrect]
  ...
  Correct: [c] out of 4. Budget change: [+/-x] points.

[SEEN setting only]
If you choose EXPLOIT, these are the 4 movies you will predict:
  title="[title]" year=[year] genres=[[genres]]
  ...

Will you EXPLORE or EXPLOIT?
```

When the model selects a question it has already asked, the following note is appended to the system prompt and the DECIDE call is repeated:

```
ERROR: You chose "[question]" but it has already been asked.
Questions already asked:
    - [question]
    ...
Choose a DIFFERENT question that has NOT been asked yet.
```

The PREDICT system prompt and user message are:

```
You are a movie recommendation system. Predict whether a specific user
will Like or Dislike each movie given to you.

Output ONLY a valid JSON object. No explanation. No markdown.

{
  "predictions": [
    {"title": "Movie Title Here", "prediction": "Like"},
    {"title": "Another Movie",    "prediction": "Dislike"}
  ]
}
Use exactly "Like" or "Dislike". Include every movie. Copy titles exactly.
Begin your response immediately with {

[demographics shown: l2, l4]
USER DEMOGRAPHICS: Age [age], Gender [gender], Occupation [occupation]

[preferences shown: l3, l4]
USER RATED HISTORY (10 movies):
  "[title]" [[genres]] - [Like/Dislike]
  ...

[no information yet]
NO INFORMATION AVAILABLE ABOUT THIS USER.

[at least one question answered]
USER PREFERENCES (from Q&A):
  Q: [question]
  A: [answer]
  ...
```

```
MOVIES TO PREDICT:
  title="[title]" year=[year] genres=[[genres]]
  ...

Output the JSON object now:
```

The feedback from the last exploit appears only in the DECIDE prompt that immediately follows it. It is removed once the model explores and is never shown in the PREDICT prompt.

Forced exploitation and invalid outputs. When the budget is below the explore cost, the model is forced to exploit and only the PREDICT call is made. A requested question is matched to the closest question in the list by word overlap (Jaccard similarity ≥ 0.5). An explore with no question, or with a question that matches none in the list, is converted into an exploit. A matched question that has already been asked is rejected and the model is asked again, up to three times, as described in Section 3. In the PREDICT step, a movie with no parsable prediction defaults to LIKE, and a turn whose output cannot be parsed at all records no predictions and leaves the budget unchanged.

A.2 QUESTIONS

The question pool contains 15 questions in three categories: preference-based questions, which ask directly about the user's movie tastes and viewing habits; demographic questions, which ask about the user's background; and distractors, which carry no information about movie preferences. Table 4 lists each question with its category.

Table 4: The 15 exploration questions and their category.

ID	Question	Category
Q1	What are your top 3 favourite movie genres?	Preference
Q2	Which genres do you tend to dislike or avoid?	Preference
Q3	Do you tend to enjoy older classic films, or do you gravitate more toward modern productions?	Preference
Q4	Do you tend to rate movies within your favourite genres consistently highly, or does quality vary a lot even within genres you like?	Preference
Q5	Overall, are you an easy-to-please viewer or a hard-to-impress one?	Preference
Q6	Are you an avid, moderate, or light movie watcher?	Preference
Q7	Are you part of a generation that grew up with cinema as a cultural centerpiece (pre-1970s), the blockbuster/VHS era (1970s–1990s), or the streaming/digital era (post-1990s)?	Demographic
Q8	How would you describe your area – urban, suburban, or rural?	Demographic
Q9	Would you describe your work as more intellectual/creative, practical/hands-on, or service-oriented?	Demographic
Q10	What is your occupation?	Demographic
Q11	What is your age group?	Demographic
Q12	What is your gender?	Demographic
Q13	What is your favorite color?	Distractor
Q14	What is your lucky number?	Distractor
Q15	Do you prefer tea or coffee?	Distractor

A.3 QUESTION RANKING METHODOLOGY

Because a question's value depends on what the model already knows about the user, we assign utility tiers separately at each context level. We collect rankings from two sources. First, GPT-5.5 ranks all 15 questions for each of 18 simulated user personas, with the prompt varied to match the context level. Second, three human experts independently rank all 15 questions at each context

level. We average these rankings to obtain a mean rank per question and context level, and divide the questions into three tiers of five.

QUESTION RANKING PROMPTS

The four prompts below differ only in which user information is given to the ranker alongside each persona. The 15 questions are listed in full in the first prompt.

Case 1: Demographics and Context

```
You are a movie recommendation assistant. Your goal is to learn about a user
in order to predict whether they will like or dislike movies. You will be
given a user's basic demographics (age, gender, and occupation) and a list
of 10 movies that they have rated as either liked or disliked.

Below is a list of 15 questions you could ask this user to learn more about
their movie preferences. Your task is to rank these questions from most
useful to least useful for predicting this specific user's movie
preferences.

QUESTIONS:
What is your age group?
What is your gender?
What is your occupation?
How would you describe your area -- urban, suburban, or rural?
Would you describe your work as more intellectual/creative,
practical/hands-on, or service-oriented?
Are you part of a generation that grew up with cinema as a cultural
centerpiece (pre-1970s), the blockbuster/VHS era (1970s-1990s), or the
streaming/digital era (post-1990s)?
What are your top 3 favourite movie genres?
Which genres do you tend to dislike or avoid?
Overall, are you an easy-to-please viewer or a hard-to-impress one?
Are you an avid, moderate, or light movie watcher?
Do you tend to enjoy older classic films, or do you gravitate more toward
modern productions?
Do you tend to rate movies within your favourite genres consistently
highly, or does quality vary a lot even within genres you like?
What is your favorite color?
What is your lucky number?
Do you prefer tea or coffee?

Rank all 15 questions from 1 (most useful) to 15 (least useful) for
predicting this user's movie preferences.

Here is the user you are recommending movies for:
[USER BLOCK]

Format your response exactly as follows:
Rank 1: [Question number and text]
Rank 2: [Question number and text]
...
Rank 15: [Question number and text]
```

Case 2: Context Only (No Demographics)

```
You are a movie recommendation assistant. Your goal is to learn about a user
in order to predict whether they will like or dislike movies. You will be
given a list of 10 movies that they have rated as either liked or disliked.

Below is a list of 15 questions you could ask this user to learn more about
their movie preferences. Your task is to rank these questions from most
useful to least useful for predicting this specific user's movie
preferences.
```

```
QUESTIONS:
[same 15 questions as Case 1]

Rank all 15 questions from 1 (most useful) to 15 (least useful) for
predicting this user's movie preferences.

Here is the user you are recommending movies for:
[USER BLOCK]

Format your response exactly as follows:
Rank 1: [Question number and text]
Rank 2: [Question number and text]
...
Rank 15: [Question number and text]
```

Case 3: Demographics Only (No Context)

```
You are a movie recommendation assistant. Your goal is to learn about a user
in order to predict whether they will like or dislike movies. You will be
given a user's basic demographics (age, gender, and occupation).

Below is a list of 15 questions you could ask this user to learn more about
their movie preferences. Your task is to rank these questions from most
useful to least useful for predicting this specific user's movie
preferences.

QUESTIONS:
[same 15 questions as Case 1]

Rank all 15 questions from 1 (most useful) to 15 (least useful) for
predicting this user's movie preferences.

Here is the user you are recommending movies for:
[USER BLOCK]

Format your response exactly as follows:
Rank 1: [Question number and text]
Rank 2: [Question number and text]
...
Rank 15: [Question number and text]
```

Case 4: No Demographics or Context

```
You are a movie recommendation assistant. Your goal is to learn about a user
in order to predict whether they will like or dislike movies.

Below is a list of 15 questions you could ask this user to learn more about
their movie preferences. Your task is to rank these questions from most
useful to least useful for predicting this specific user's movie
preferences.

QUESTIONS:
[same 15 questions as Case 1]

Rank all 15 questions from 1 (most useful) to 15 (least useful) for
predicting this user's movie preferences.

Here is the user you are recommending movies for:
[USER BLOCK]

Format your response exactly as follows:
Rank 1: [Question number and text]
Rank 2: [Question number and text]
```

```
...
Rank 15: [Question number and text]
```

USER PERSONAS

We simulated 18 user personas spanning combinations of age group, gender, and occupation, each with a fixed history of 5 liked and 5 disliked movies. Each persona was substituted into the `[USER BLOCK]` placeholder of the prompts above, with only the fields available at that context level included.

QUESTION RANKING AND UTILITY TIER ASSIGNMENT

Table 5 reports the mean rank of each question at each context level, and Table 6 the resulting tiers. At every level, the five best-ranked questions form T1 (high utility), the next five T2 (moderate utility), and the last five T3 (low utility). When demographics are provided (ℓ_2, ℓ_4), questions that repeat them, such as gender, move to T3, while questions that refine them, such as area and type of work, move up to T2.

Table 5: Average rank (1 = most useful, 15 = least useful) of each exploration question, by condition. Lower is more useful.

Question	**None**	**Demo**	**Pref**	**Pref+Demo**
What are your top 3 favourite movie genres?	1.67	1.00	2.36	1.03
Which genres do you tend to dislike or avoid?	2.67	2.00	3.28	2.03
Do you tend to enjoy older classic films, or gravitate toward modern productions?	4.03	3.36	4.80	3.22
Do you rate movies within favourite genres consistently highly, or does quality vary a lot?	5.37	5.20	3.94	4.28
Overall, are you an easy-to-please viewer or a hard-to-impress one?	6.42	5.62	6.45	5.78
Are you an avid, moderate, or light movie watcher?	6.61	5.37	7.58	6.05
Are you part of a generation that grew up with cinema as a cultural centerpiece...?	7.22	6.67	7.53	6.55
How would you describe your area – urban, suburban, or rural?	10.84	9.61	11.16	9.39
Would you describe your work as intellectual/creative, practical/hands-on, or service-oriented?	9.86	8.20	9.97	8.30
What is your occupation?	8.14	9.75	8.36	10.34
What is your age group?	7.14	10.30	6.50	10.17
What is your gender?	8.09	11.45	6.78	10.87
What is your favorite color?	13.50	13.50	13.50	13.61
What is your lucky number?	14.50	14.50	14.50	14.39
Do you prefer tea or coffee?	13.97	14.00	14.00	14.00

Table 6: Utility tier assignment (T1 = high, T2 = moderate, T3 = zero information value) for each exploration question, by condition.

Question	None	Demo	Pref	Pref+Demo
What are your top 3 favourite movie genres?	T1	T1	T1	T1
Which genres do you tend to dislike or avoid?	T1	T1	T1	T1
Do you tend to enjoy older classic films, or gravitate toward modern productions?	T1	T1	T1	T1
Do you rate movies within favourite genres consistently highly, or does quality vary a lot?	T1	T1	T1	T1
Overall, are you an easy-to-please viewer or a hard-to-impress one?	T1	T2	T1	T1
Are you an avid, moderate, or light movie watcher?	T2	T1	T2	T2
Are you part of a generation that grew up with cinema as a cultural centerpiece...?	T2	T2	T2	T2
How would you describe your area – urban, suburban, or rural?	T3	T2	T3	T2
Would you describe your work as intellectual/creative, practical/hands-on, or service-oriented?	T3	T2	T3	T2
What is your occupation?	T2	T2	T2	T3
What is your age group?	T2	T3	T2	T2
What is your gender?	T2	T3	T2	T3
What is your favorite color?	T3	T3	T3	T3
What is your lucky number?	T3	T3	T3	T3
Do you prefer tea or coffee?	T3	T3	T3	T3

A.4 Statistical Analysis

Metrics. We used 5,000 paired user-level bootstrap replicates. In each replicate, users were sampled with replacement, and all four context levels associated with each selected user were retained. The same sampled user indices were applied to every model. From the resampled per-user and per-context trajectory terms, we recomputed the context-level exploration and question-quality rates and then recalculated Sensitivity, Question Quality, Trend Alignment, and Overall for each model.

Tiers and win rates. Models are sorted by median bootstrap rank. Going down the ranking, the highest-ranked model not yet assigned to a tier starts a new tier, and every later model whose 95% rank interval overlaps that model's interval joins it. A model's win rate is the share of its significant pairwise comparisons that it wins (paired bootstrap on the Overall gap, Benjamini–Hochberg corrected $p < 0.05$ with the interval excluding 0).

Model family comparisons. Within each family, we calculated the difference in Overall between the larger- and smaller-tier variants, defining a positive difference as favoring the larger-tier variant. We report the Table 2 point-estimate difference and its 95% percentile bootstrap confidence interval. Two-sided bootstrap-tail p-values were corrected across the five prespecified family comparisons using the Benjamini-Hochberg procedure. To test whether larger-tier variants showed a systematic positive advantage across model families, we additionally performed an exact one-sided sign-flip test on the five observed family-level differences.

A.5 Scores

Exploration rate and explication quality

Across models, context levels ($\ell_1 \dots \ell_4$) and discount factors ($\gamma = 0.3 \dots 0.9$), we report the exploration rates in Table 7, and explication quality scores at K=5 and K=10 in Table 9 and Table 11 respectively.

Metrics

Across models and discount factors ($\gamma = 0.3 \dots 0.9$), we report the values for Sensitivity (Sens) in Table 8, Question Quality (Qual@5, Qual@10) in Tables 10 and 12, and Trend Alignment (Align) in Table 13.

A.6 Average interaction count

We report the average number of interactions per conversation across context levels in Table 14. Each of the $R = 20$ rounds ends in exactly one explore or exploit, so a run can have a minimum of 20 interactions. We find that on average, most models remain close to 20 interactions, with mean interactions ranging from 20.11 to 22.32. Gemma-2-2B, the Llama models, and DeepSeek-R1-Distill-Llama-8B exhibit substantially more interactions. In particular, Gemma-2-2B and DeepSeek-R1-Distill-Llama-8B are extreme outliers, averaging 64.15 and 42.85 recorded interactions per run, respectively.

Higher exploration penalty leads to less repetition in 10 out of 11 models, indicating increased tendencies of duplicate question selection when the penalty is lower. The effects of increased context vary across models.

A.7 Exploration Rate and Explication Quality

We plot the exploration rates and explication quality scores across context levels for the base configuration (seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$) in Figure 3. We observe that Aya models opt to explore much less compared to other models, indicating higher confidence in their belief about the user. Llama-3.1-70B has the highest exploration rates across context levels, followed by Gemma-2-2B and DeepSeek-R1-Distill-Llama-8B. The Gemini-2.5 models and Qwen3.5-27B perform exceptionally well in explication, being able to choose high quality questions around 80% of the time. In contrast, Gemma-2-2B and DeepSeek-R1-Distill-Llama-8B display poor explication due to their tendency to ask low-quality questions, as well as asking repeated questions.

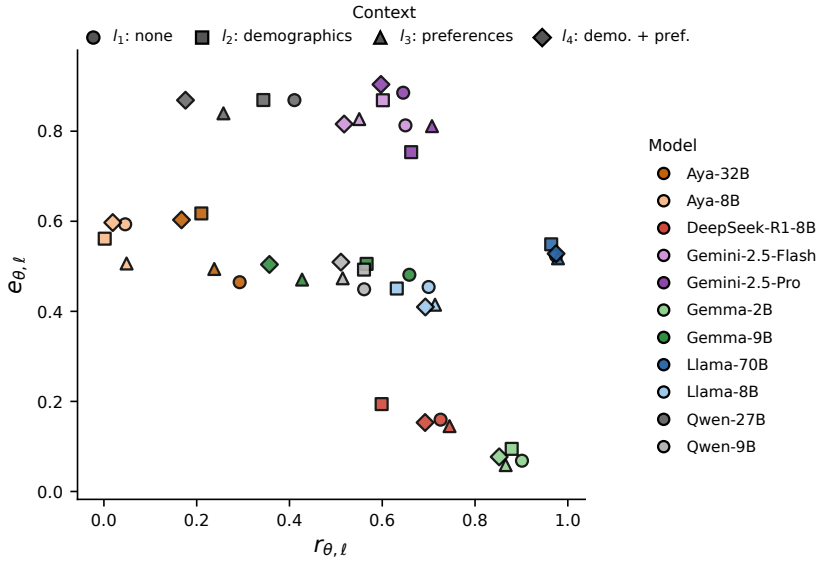

Figure 3: $e_{\theta,\ell_1} \dots e_{\theta,\ell_4}$ and $r_{\theta,\ell_1} \dots r_{\theta,\ell_4}$ for all models (visibility: **seen**, cost: **expensive**, $\gamma = 0.9$)

Exploration depends on the amount and type of context. Only five models exhibit a monotonic decrease in exploration from ℓ_1 through ℓ_4: Qwen3.5-27B, Gemma-2-9B, Gemini-2.5-Flash, Qwen3.5-9B, and Gemma-2-2B. Three other models (Aya-Expanse-8B, Aya-Expanse-32B, and DeepSeek-R1-Distill-Llama-8B) show an overall decrease in exploration, but reverse direction between intermediate levels. Most violations occur between ℓ_2 and ℓ_3. Adding demographics reduces exploration for 9 of 11 models when added to no context ($\ell_1 \to \ell_2$) and for all 11 when added to movie preferences ($\ell_3 \to \ell_4$). By contrast, replacing demographics with explicit movie preferences ($\ell_2 \to \ell_3$) reduces exploration for only five models; the other six increase exploration. The models therefore respond differently not only to the amount of context, but also to the type of information presented.

A.8 Ablation Studies

We analyze the effect on exploration rate, explication quality and metrics through ablations on target visibility (seen vs. unseen), exploration penalty (cheap($\lambda_- = 2$) vs. expensive ($\lambda_- = 5$)), discount factor ($\gamma \in \{0.3, 0.5, 0.7, 0.9\}$), and $K \in \{5, 10\}$ top questions. We also test whether each ablation has an effect on model ranking and tiering, through bootstrapping on 5,000 bootstrap replicates over 2,038 users. We report rank correlations between different ablations in Table 25. The results of the base configuration (seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$, $K = 5$) are reported in Table 1 in the main paper. We report the monotonicity in exploration rates across different configurations in Table 26.

A.8.1 Target Visibility - Seen vs. Unseen

To test the effect of target visibility, we compare the seen and unseen settings, while holding the other factors constant (expensive ($\lambda_- = 5$), $\gamma = 0.9$)

Exploration rate and explication quality Figure 4 plots the exploration rates and explication quality scores for each model at its four context levels. Most models tend to explore more when the target is unseen. This change is especially noticeable for Aya-Expanse-8B, which almost never explores when the target is seen, but has an exploration rate of $r = 0.237$ at ℓ_1 when the target is unseen. Gemma-2-9B displays the opposite trend, with a significant drop in exploration at context levels ℓ_3 ($0.427 \rightarrow 0.209$) and ℓ_4 ($0.357 \rightarrow 0.188$) - which could indicate higher confidence when explicit preferences are provided and the target is unseen. For most models with substantial changes in exploration, explication quality tends to move in the other direction. Gemma-2-9B asks better questions at ℓ_3 ($+0.168$) and ℓ_4 ($+0.146$), the levels where its exploration rate dropped. Aya-Expanse-8B suffers a drop in explication quality at ℓ_1, the level at which its exploration rate rose the most. Aya-Expanse-32B explores less and has higher explication quality at every level. **Asking more tends to make models less selective**. In contrast, DeepSeek-R1-Distill-Llama-8B and Llama-3.1-8B tend to become more context-invariant in their explication quality.

Unseen targets make exploration more context-sensitive. Six models are monotone, with Aya-Expanse-32B joining the five models that were monotone in the "seen" setting. Every model explores less at ℓ_4 than at ℓ_1. All the monotonicity violations occur at $\ell_2 \rightarrow \ell_3$ (Llama-3.1-8B, Llama-3.1-70B, Aya-Expanse-8B, Gemini-2.5-Pro and DeepSeek-R1-Distill-Llama-8B).

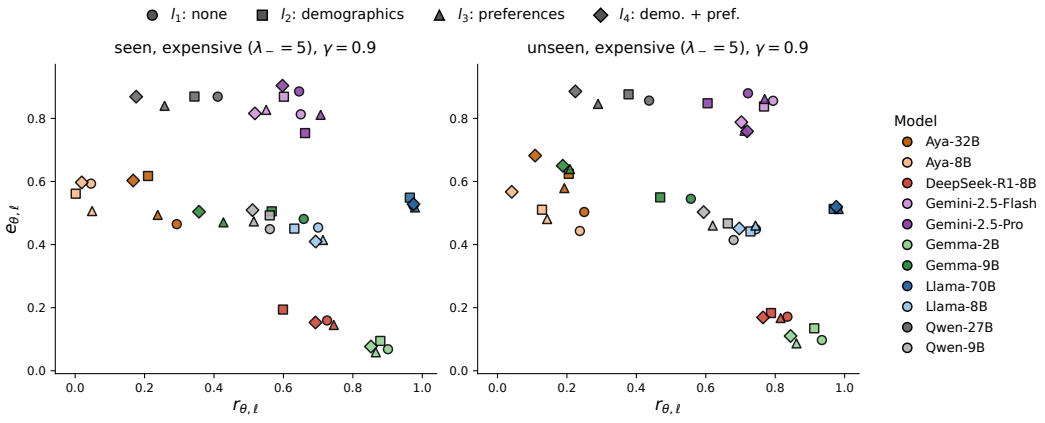

Figure 4: **Target Visibility ablation:** Exploration rate $r_{\theta,\ell}$ against explication quality $e_{\theta,\ell}$ ($K = 5$) per model and context level, for seen (left) and unseen (right) targets (expensive ($\lambda_- = 5$), $\gamma = 0.9$).

Metrics Table 15 reports the metrics for this setting. Changing target visibility causes minimal change in Sens and Qual@5, and the changes in Overall are carried almost entirely by Align. Gemini-2.5-Flash and Gemma-2-9B had similarly positive exploration correlations across both settings, but their explication quality became more context dependent, pushing their explication correlation away from its ideal of 0 and significantly decreasing Align. On the other hand, DeepSeek-R1-Distill-Llama-8B and Llama-3.1-8B had minimal change in sensitivity, but their increased context-

invariance in explication quality led to a closer correlation to the ideal value, leading to significant gains in $Align$. The range of Overall values is compressed in this setting: the models with the lowest scores gain and vice versa.

Ranking and tiers. We report rank intervals and model tiers for this configuration in Table 22. We find that the model ranks and tiers are **more uncertain** than the base configuration. The models are divided into **5** broader tiers, with the mean rank interval rising from 0.73 to 1.73. There are up to **4** models in a single tier (Tier 3). Qwen3.5-27B is alone in the top tier; in the base configuration, it is statistically tied with Gemini-2.5-Flash. Instead, Gemini-2.5-Flash is a part of the second tier along with Gemma-2-9B. The models within tiers also differ significantly – the 4 models in tier 3 have differing numbers of significant comparisons.

There is increased uncertainty in model ranks too - for instance, Llama-3.1-8B's rank interval spans six ranks ($[4, 9]$). While the ranking has high agreement with the base configuration ($\rho = 0.882$), it has the second widest interval among all ablations ($[0.73, 0.93]$) which reflects the amount of uncertainty in the ranking.

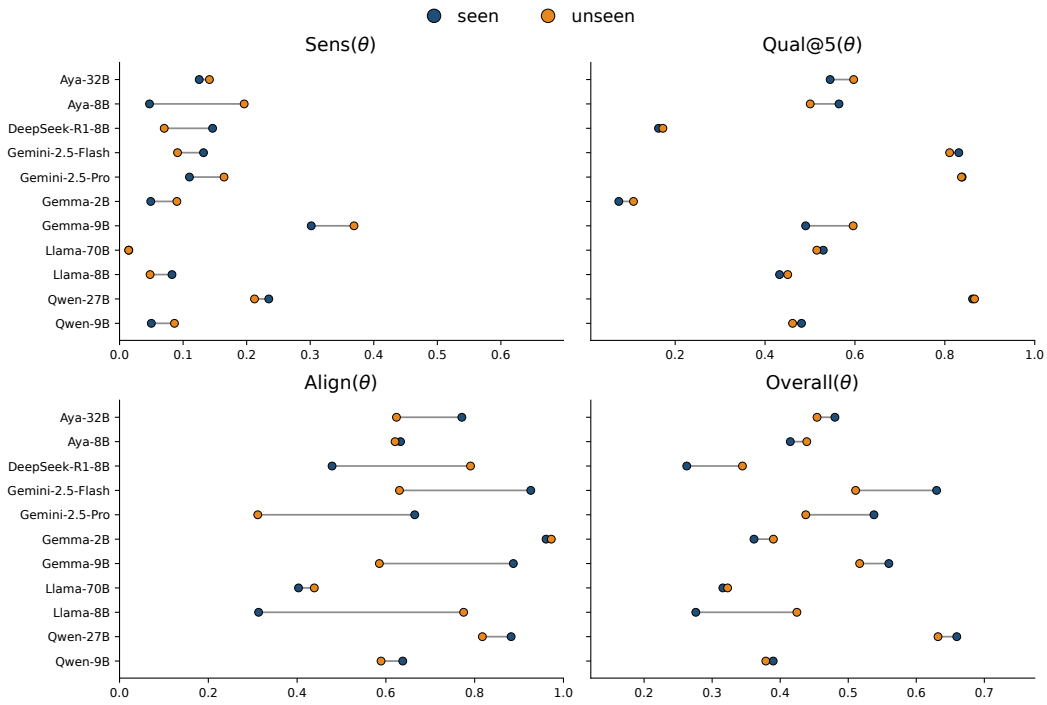

Figure 5: Target visibility variations

A.8.2 Expensive vs Cheap

We compare expensive ($\lambda_{-} = 5$) and cheap ($\lambda_{-} = 2$) exploration, holding the other factors constant (seen, $\gamma = 0.9$).

Exploration rate and explication quality We plot the exploration rates and explication quality scores across context levels for expensive and cheap exploration in Figure 6.

A model that puts value on exploration should have lower exploration rate when it is expensive. Instead, the changes we observe are minimal: **models do not appear to put a value on exploration.** While some models explore more and some explore less, with the Aya models' exploration rates being almost unchanged, none of the models' exploration rates change substantially. As the models' exploration rates remain the same while the number of affordable turns increases, they would ask more questions. This leads to models exploring later in the conversation, thereby increasing the chances of selecting lower-quality questions. This appears to have an effect, as the explication

quality falls for almost all models. The key exceptions are the Aya models, where the exploration rate was already unchanged.

Cheap exploration reduces context sensitivity. Four models are monotone, as Qwen3.5-9B now explores more at ℓ_2 than at ℓ_1. 8 models explore less at ℓ_4 than at ℓ_1, and there are monotonicity violations across all context pairs.

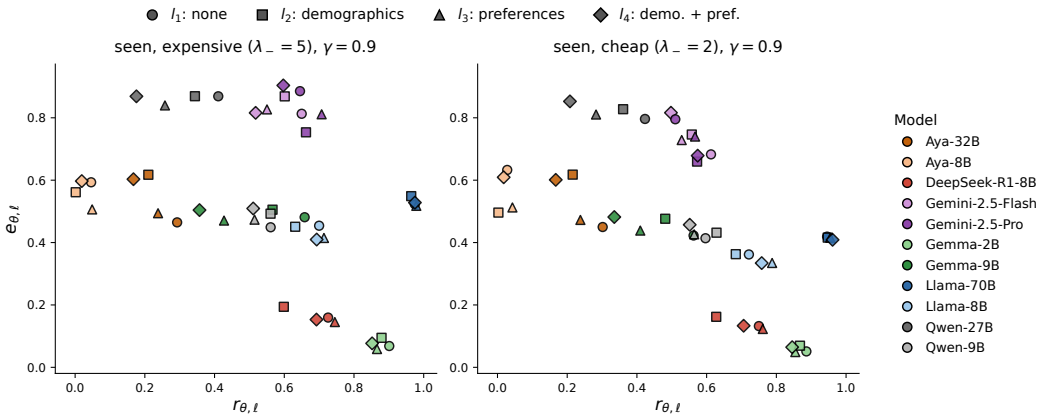

Figure 6: Exploration rate $r_{\theta,\ell}$ against explication quality $e_{\theta,\ell}$ ($K = 5$) under expensive ($\lambda_- = 5$, left) and cheap ($\lambda_- = 2$, right) exploration (seen, $\gamma = 0.9$).

Metrics We report the metrics for this cell in Table 16. Cheap exploration rate lowers Overall significantly for 9 models and raises it for none. The Qual@5 score reduces for all models. The changes in this setting are also driven by Align, with similar trends as the changes in Overall. This can be attributed to the drops in explication quality causing reduced invariance with context levels. An example of this is Qwen3.5-27B, whose exploration ordering remains the same, but its explication quality now increases with context (0.796 at ℓ_1 to 0.853 at ℓ_4), leading to substantial reduction in Align.

Rank and tiering We report rank intervals and model tiers for this configuration in Table 21. We observe **stable rankings and tiering** in this configuration. There is no change in uncertainty in the rankings - the mean rank-interval width is the same as the base configuration (0.73). The number of tiers remains the same, but there is a shuffle in the models that occupy them – Qwen3.5-27B forms a tier of its own, Gemini-2.5-Flash drops to the second tier where it joins Gemma-2-9B and Aya-Expanse-32B. Gemini-2.5-Pro sees the sharpest drop, falling from the second tier (rank 4) to the fourth tier (rank 8). The last 3 tiers are occupied by 1 model each with fixed ranks (DeepSeek-R1-Distill-Llama-8B, Llama-3.1-8B and Llama-3.1-70B, at 9, 10 and 11).

A.8.3 Effect of change in γ

We vary $\gamma \in \{0.3, 0.5, 0.7, 0.9\}$ in the seen, expensive setting. γ controls how the turns in a conversation are weighted. A lower γ value leads to greater concentration of the metrics on the opening turns, and comparing γ values shows how a system's behavior evolves within an interaction. If a score rises as γ decreases, it indicates stronger behavior earlier in the conversation.

Exploration rate and explication quality We plot the exploration rates and explication quality scores across context levels for each γ in Figure 8. **As γ decreases, the points shift upward and to the right for most models.** At $\gamma = 0.3$, the Gemini-2.5 models, Llama-3.1-70B and Gemma-2-2B have exploration rates $r \geq 0.958$ at every level, which means they almost always explore on the first turn regardless of the amount of context. At $\gamma = 0.3$, every model except Aya-Expanse-8B explores at $r \geq 0.787$ at ℓ_1. Qwen3.5-27B shows the strongest dependency on context: at $\gamma = 0.3$, its exploration rate drops sharply from ℓ_1 (0.919) to ℓ_4 (0.253). It almost always asks first when it knows nothing about the user, and its likelihood of recommending immediately increases with the amount of knowledge it has about the user. The upward drift in explication quality also indicates

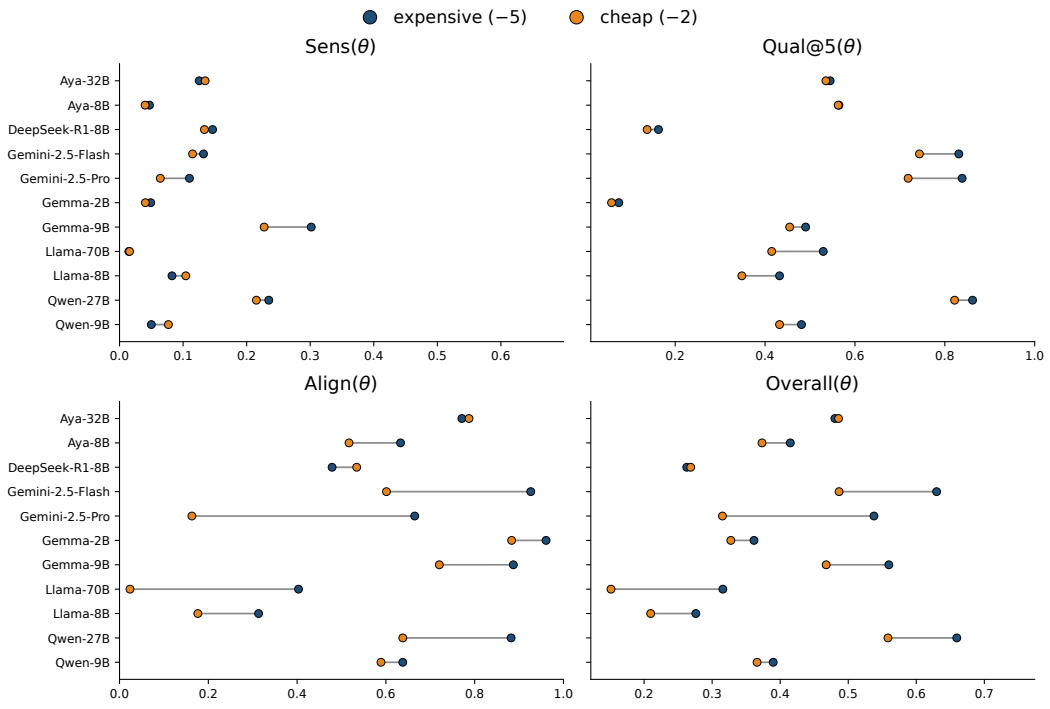

Figure 7: Cost variations

that most models ask better questions earlier, and the question quality declines over the course of the conversation. The drop in Qual@5 is highest for Llama-3.1-70B (0.261) and DeepSeek-R1-Distill-Llama-8B (0.217).

The contextual sensitivity of exploration emerges later in the interaction - the monotonicity is lowest at $\gamma = 0.3$ and $\gamma = 0.5$ with just 2 monotone models. This increases to 4 at $\gamma = 0.7$ and 5 at $\gamma = 0.9$. Only Qwen3.5-27B and Gemma-2-9B are monotonic at every γ value. Adding demographics to preferences ($\ell_3 \to \ell_4$) lowers exploration for all 11 models at every γ.

Metrics We report the metrics for $\gamma = 0.7$, 0.5 and 0.3 in tables 17–19. The change in Sens reflects different kinds of model behavior. Sens **increases substantially for models that adapt their opening move to the context**, such as Qwen3.5-27B, Llama-3.1-8B and Aya-Expanse-8B. It **falls for models that ask first regardless of the context**, such as the Gemini-2.5 models, Gemma-2-9B and Llama-3.1-70B. Qual@5 increases as γ decreases for every model, as most of them tend to ask better questions earlier. The highest variance is observed in the Align metric. The largest change in Align is observed for Gemini-2.5-Flash: as its exploration rates lie very close together (between 0.958 and 0.969) at $\gamma = 0.3$, its correlation with exploration drops substantially, leading to a significant drop in *Align*. This also causes Flash to drop from 2nd to 9th in model rankings.

Rank and tiering We report rank intervals and model tiers for the different γ configurations in Tables 23a–23c. **As the value of γ reduces, we observe the rankings becoming less stable**. The number of tiers decreases with a decrease in γ, from 7 at $\gamma = 0.9$ to 6, 5 and 4 at $\gamma = 0.7$, $\gamma = 0.5$ and $\gamma = 0.3$ respectively. The mean rank-interval width also rises as γ decreases ($0.73 \to 1.00 \to 1.73 \to 2.45$). Across all $\gamma < 0.9$, Qwen3.5-27B is the sole top-tier model.

At $\gamma = 0.7$, Gemini-2.5-Flash drops down to the second tier, which it shares with Gemini-2.5-Pro, while Gemma-2-2B drops down to the bottom tier. At $\gamma = 0.5$, 5 models share the second tier, whose rank intervals span 2 (Gemini-2.5-Flash) to 8 (Aya-Expanse-8B). Gemma-2-2B stays in the bottom tier, with the lowest rank among all models.

At $\gamma = 0.3$, the second tier consists of 8 models whose rank intervals span 2 (Aya-Expanse-8B) to 10 (Gemini-2.5-Flash). The models within this tier differ significantly. Gemini-2.5-Flash falls 7

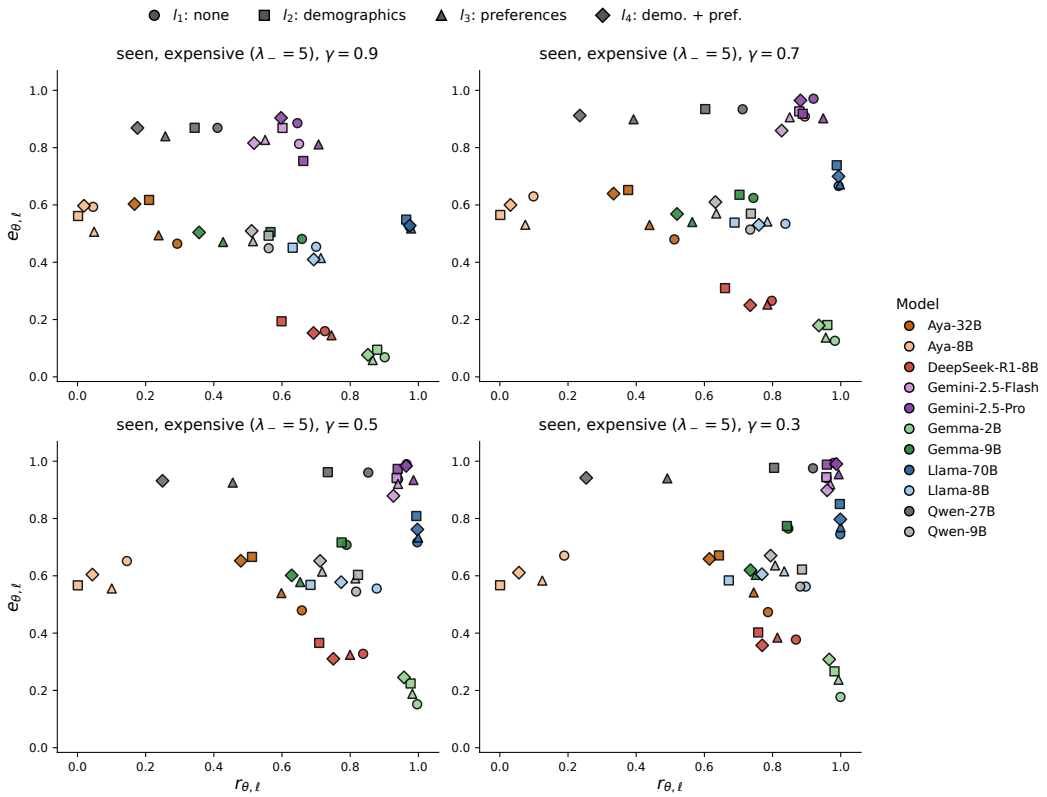

Figure 8: Exploration rate $r_{\theta,\ell}$ against explication quality $e_{\theta,\ell}$ ($K = 5$) for $\gamma \in \{0.9, 0.7, 0.5, 0.3\}$ (seen, expensive). All panels score the same runs; lower γ weights earlier turns more heavily.

places (from rank 2 to rank 9), due to its $Align$ being computed over near-tied exploration rates. On the other hand, Llama-3.1-8B and Aya-Expanse-8B rise 4 places to be a part of the second tier. The $\gamma = 0.3$ ablation has the weakest agreement with the base configuration ($\rho = 0.518$), as well as the widest uncertainty interval among all ablations.

A.8.4 Effect of K - 5 vs. 10

Increasing K widens the set of questions considered informative. We explore the changes in e_ℓ, and the Qual@K and Align metrics.

Explication quality by context level We report explication quality across context levels at $K = 5$ and $K = 10$ for the (seen, expensive) setting in Figure 10. For all model pairs and context levels, $e_{\theta,\ell}$ is higher at $K = 10$. Increasing K also reveals some trends - Gemma-2-2B's explication quality oscillates with no clear trend at $K = 5$, but decreases monotonically with context at $K = 10$. A similar trend is also revealed for DeepSeek-R1-Distill-Llama-8B, where explication quality decreases monotonically from ℓ_2 to ℓ_4.

Metrics We report metrics at $K = 10$ for the seen, expensive setting in Table 20. $Qual@10$ exceeds $Qual@5$ for every model, with the biggest increase observed in Qwen3.5-9B. Qwen3.5-9B also has the biggest increase in $Align$. Gemma-2-2B and DeepSeek-R1-Distill-Llama-8B are among the models with biggest decrease in $Align$ due to the revealed contextual trends discussed earlier.

Rank and tiering We report rank intervals and model tiers for this configuration in Table 24. This is the most stable ranking among all configurations, dividing the models into 8 tiers with a mean rank-interval width of 0.55, as well as having the strongest agreement with the base configuration

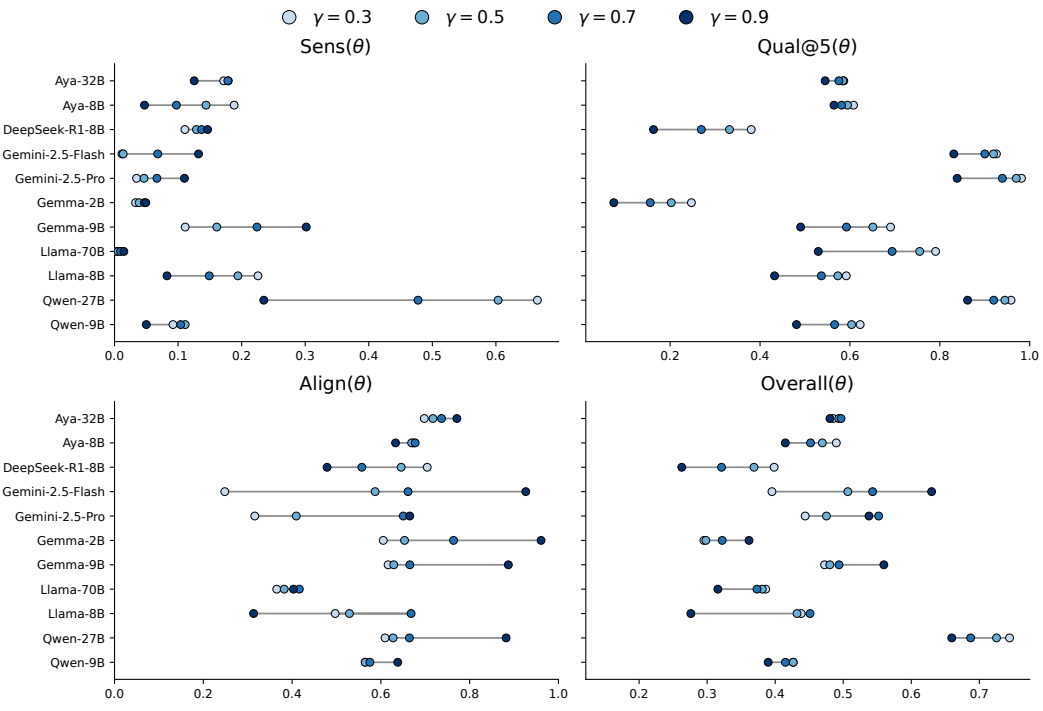

Figure 9: γ variations

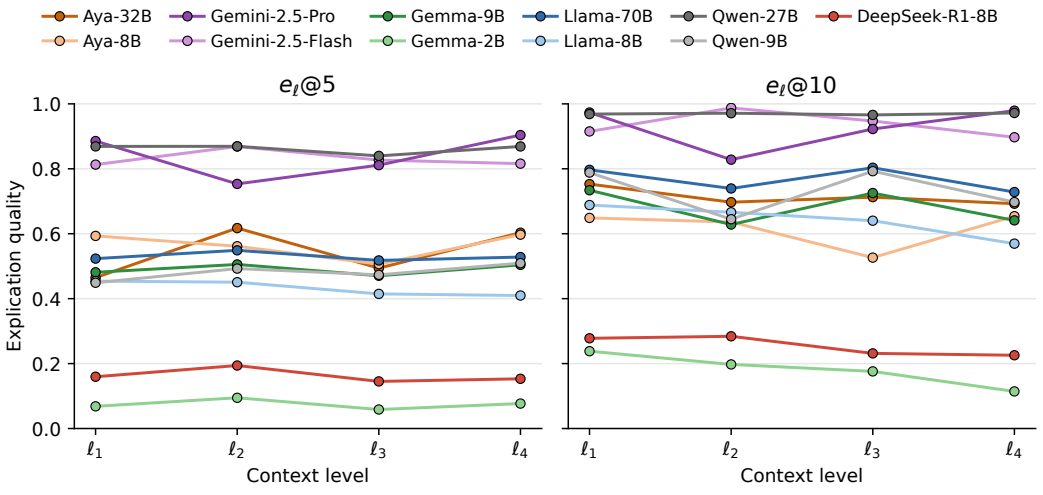

Figure 10: Explication quality $e_{\theta,\ell}$ by context level at $K = 5$ (left) and $K = 10$ (right) (seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$). Exploration rates are identical at both values of K. Values in Tables 9 and 11.

($\rho = 0.945$). Similar to the base configuration, the top tier is shared by Qwen3.5-27B and Gemini-2.5-Flash. Gemma-2-9B is alone in the second tier. The biggest gain is achieved by Qwen3.5-9B, which gains 2 ranks and shares the third tier with Gemini-2.5-Pro. The biggest drop is observed for Gemma-2-2B, which drops down to the bottom tier, due to the contextual variance of its explication quality in this setting. Overall, the model ranking is robust to K.

A.9 BASELINE IMPLEMENTATIONS

We run three non-LLM policies as baselines with the same configuration as the models - same per-user movie pools, $R = 20$, starting budget of 10, and explore and exploit rules.

Oracle: On each affordable turn, this policy explores with probability $p = 0.5$ and otherwise exploits. On each explore turn, this policy asks the best question still available - an unasked question from the highest-ranked question band that has questions remaining. As every question is optimal, this policy scores $\mathrm{Qual@K} = 1$, indicating the ceiling for question quality.

Uniform ($p = 0.5$): On each affordable turn, this policy explores with probability $p = 0.5$ and otherwise exploits. When it explores, it selects a question that was not previously asked, so it never repeats a question. When it exploits, each of its predictions is a fair coin-flip. This baseline indicates the performance of coin-flip behavior on both the explore-exploit decision and the question choice.

Always-explore: This policy explores on every affordable turn, and selects questions randomly from those not yet asked. During forced exploitation, it also performs a fair coin-flip for each prediction. Its exploration rate is therefore always 1.

A.10 TABLES

Table 7: Exploration rates ($r_{\theta,\ell}$) by context level and discount factor γ. Context levels: ℓ_1 = none, ℓ_2 = demographics, ℓ_3 = preferences, ℓ_4 = demographics + preferences.

Model	Visibility	Cost(λ_-)	$\gamma = 0.3$				$\gamma = 0.5$				$\gamma = 0.7$				$\gamma = 0.9$			
			ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4
Gemma-2-2B	seen	5	0.999	0.982	0.993	0.966	0.996	0.977	0.982	0.958	0.983	0.961	0.956	0.936	0.901	0.879	0.866	0.852
		2	0.998	0.980	0.989	0.959	0.992	0.971	0.973	0.948	0.969	0.946	0.938	0.920	0.886	0.868	0.854	0.845
	unseen	5	0.998	0.958	0.979	0.933	0.996	0.964	0.976	0.937	0.989	0.963	0.956	0.924	0.934	0.912	0.861	0.844
		2	0.997	0.961	0.973	0.935	0.994	0.965	0.969	0.935	0.982	0.961	0.946	0.917	0.919	0.904	0.864	0.846
Gemma-2-9B	seen	5	0.846	0.842	0.751	0.735	0.789	0.775	0.653	0.628	0.744	0.703	0.564	0.520	0.658	0.566	0.427	0.357
		2	0.816	0.805	0.745	0.721	0.729	0.702	0.625	0.591	0.653	0.603	0.522	0.470	0.563	0.481	0.410	0.335
	unseen	5	0.814	0.788	0.709	0.698	0.740	0.691	0.554	0.539	0.676	0.602	0.397	0.375	0.557	0.469	0.209	0.188
		2	0.798	0.776	0.712	0.703	0.696	0.654	0.546	0.533	0.601	0.540	0.385	0.366	0.471	0.406	0.219	0.198
Llama-3.1-8B	seen	5	0.897	0.672	0.834	0.769	0.878	0.684	0.815	0.773	0.838	0.689	0.785	0.760	0.700	0.631	0.714	0.693
		2	0.918	0.722	0.858	0.803	0.889	0.728	0.846	0.810	0.835	0.726	0.831	0.805	0.721	0.684	0.788	0.758
	unseen	5	0.908	0.884	0.886	0.873	0.866	0.860	0.854	0.841	0.824	0.824	0.812	0.792	0.745	0.729	0.743	0.696
		2	0.912	0.895	0.901	0.880	0.843	0.842	0.849	0.826	0.764	0.773	0.781	0.756	0.674	0.673	0.709	0.675
Llama-3.1-70B	seen	5	0.999	0.997	1.000	0.999	0.997	0.994	0.999	0.997	0.994	0.988	0.998	0.994	0.974	0.964	0.979	0.975
		2	0.998	0.997	1.000	0.998	0.994	0.991	0.998	0.995	0.986	0.981	0.991	0.988	0.946	0.947	0.955	0.961
	unseen	5	0.999	0.997	0.999	0.999	0.997	0.993	0.999	0.997	0.994	0.987	0.997	0.994	0.979	0.968	0.983	0.976
		2	0.997	0.996	0.999	0.998	0.991	0.989	0.994	0.993	0.976	0.975	0.981	0.982	0.928	0.938	0.935	0.949
Aya-Expanse-8B	seen	5	0.189	0.001	0.124	0.056	0.145	0.001	0.101	0.044	0.099	0.001	0.075	0.031	0.046	0.002	0.049	0.019
		2	0.141	0.002	0.115	0.052	0.103	0.002	0.091	0.041	0.065	0.002	0.066	0.029	0.028	0.003	0.043	0.018
	unseen	5	0.852	0.570	0.375	0.123	0.724	0.445	0.318	0.097	0.532	0.300	0.245	0.070	0.237	0.128	0.143	0.041
		2	0.708	0.462	0.308	0.091	0.530	0.340	0.246	0.069	0.341	0.214	0.181	0.047	0.149	0.091	0.113	0.028
Aya-Expanse-32B	seen	5	0.787	0.643	0.745	0.615	0.658	0.512	0.598	0.479	0.512	0.376	0.439	0.334	0.293	0.210	0.238	0.167
		2	0.798	0.667	0.749	0.622	0.660	0.528	0.593	0.479	0.504	0.382	0.428	0.328	0.302	0.216	0.238	0.167
	unseen	5	0.793	0.758	0.735	0.454	0.657	0.603	0.583	0.346	0.491	0.428	0.407	0.233	0.250	0.205	0.193	0.108
		2	0.787	0.754	0.747	0.486	0.636	0.588	0.586	0.367	0.462	0.407	0.405	0.243	0.251	0.202	0.201	0.114
Qwen3.5-9B	seen	5	0.881	0.886	0.807	0.794	0.817	0.823	0.718	0.712	0.734	0.736	0.635	0.632	0.561	0.561	0.515	0.511
		2	0.937	0.934	0.855	0.826	0.878	0.882	0.770	0.747	0.782	0.799	0.686	0.668	0.597	0.628	0.564	0.551
	unseen	5	0.915	0.893	0.865	0.853	0.869	0.838	0.798	0.782	0.809	0.777	0.729	0.711	0.680	0.663	0.620	0.594
		2	0.955	0.915	0.888	0.850	0.901	0.846	0.815	0.763	0.811	0.757	0.734	0.676	0.658	0.625	0.614	0.565
Qwen3.5-27B	seen	5	0.919	0.804	0.491	0.253	0.853	0.734	0.455	0.249	0.712	0.603	0.392	0.235	0.411	0.344	0.258	0.176
		2	0.921	0.829	0.502	0.299	0.825	0.731	0.461	0.288	0.670	0.585	0.398	0.267	0.423	0.360	0.283	0.208
	unseen	5	0.916	0.829	0.544	0.368	0.847	0.745	0.502	0.346	0.710	0.613	0.433	0.308	0.437	0.378	0.290	0.224
		2	0.949	0.871	0.568	0.427	0.874	0.768	0.522	0.397	0.740	0.628	0.452	0.351	0.504	0.423	0.331	0.269
Gemini-2.5-Flash	seen	5	0.958	0.958	0.969	0.960	0.940	0.936	0.940	0.927	0.895	0.878	0.850	0.827	0.650	0.601	0.550	0.518
		2	0.974	0.967	0.971	0.966	0.927	0.911	0.910	0.900	0.837	0.803	0.786	0.767	0.613	0.557	0.528	0.498
	unseen	5	0.999	0.996	0.991	0.988	0.995	0.991	0.982	0.976	0.971	0.963	0.937	0.928	0.794	0.768	0.712	0.703
		2	0.994	0.995	0.991	0.988	0.966	0.972	0.959	0.952	0.887	0.904	0.871	0.863	0.690	0.700	0.645	0.658
Gemini-2.5-Pro	seen	5	0.979	0.959	0.994	0.987	0.965	0.939	0.986	0.964	0.920	0.889	0.949	0.882	0.645	0.662	0.707	0.597
		2	0.968	0.953	0.987	0.971	0.904	0.900	0.945	0.930	0.774	0.802	0.835	0.832	0.510	0.573	0.567	0.574
	unseen	5	0.990	0.890	0.999	0.992	0.982	0.841	0.995	0.988	0.946	0.779	0.969	0.965	0.721	0.605	0.770	0.719
		2	0.969	0.960	0.994	0.985	0.906	0.894	0.963	0.952	0.790	0.792	0.875	0.868	0.553	0.597	0.628	0.629
DeepSeek-R1-Distill-Llama-8B	seen	5	0.868	0.758	0.814	0.770	0.838	0.709	0.799	0.751	0.798	0.661	0.785	0.734	0.726	0.599	0.745	0.692
		2	0.877	0.774	0.839	0.792	0.848	0.725	0.824	0.768	0.810	0.677	0.805	0.746	0.750	0.628	0.761	0.706
	unseen	5	0.938	0.931	0.859	0.827	0.923	0.910	0.852	0.813	0.897	0.872	0.843	0.798	0.835	0.788	0.816	0.765
		2	0.955	0.943	0.867	0.834	0.942	0.921	0.863	0.825	0.915	0.886	0.853	0.811	0.853	0.810	0.821	0.776

Table 8: Context Sensitivity (Sens) by visibility, explore cost and discount factor γ: the spread $\max_l r_l - \min_l r_l$ of the per-level explore rates in Table 7.

Model	Visibility	Cost (λ_-)	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$
Gemma-2-2B	seen	5	0.033	0.039	0.047	0.049
		2	0.039	0.044	0.049	0.041
	unseen	5	0.065	0.059	0.065	0.090
		2	0.062	0.059	0.065	0.073
Gemma-2-9B	seen	5	0.111	0.161	0.224	0.302
		2	0.095	0.138	0.183	0.227
	unseen	5	0.116	0.201	0.301	0.369
		2	0.095	0.164	0.235	0.272
Llama-3.1-8B	seen	5	0.226	0.194	0.149	0.083
		2	0.196	0.161	0.109	0.104
	unseen	5	0.036	0.025	0.033	0.048
		2	0.032	0.023	0.026	0.036
Llama-3.1-70B	seen	5	0.002	0.005	0.009	0.015
		2	0.003	0.007	0.010	0.016
	unseen	5	0.003	0.006	0.010	0.014
		2	0.003	0.006	0.008	0.021
Aya-Expanse-8B	seen	5	0.188	0.144	0.097	0.047
		2	0.138	0.101	0.064	0.040
	unseen	5	0.729	0.627	0.462	0.196
		2	0.617	0.461	0.294	0.122
Aya-Expanse-32B	seen	5	0.172	0.179	0.178	0.125
		2	0.175	0.181	0.176	0.135
	unseen	5	0.339	0.311	0.258	0.141
		2	0.301	0.269	0.219	0.137
Qwen3.5-9B	seen	5	0.092	0.111	0.104	0.050
		2	0.112	0.135	0.131	0.077
	unseen	5	0.062	0.087	0.098	0.086
		2	0.105	0.137	0.135	0.094
Qwen3.5-27B	seen	5	0.665	0.603	0.478	0.235
		2	0.623	0.537	0.403	0.215
	unseen	5	0.547	0.501	0.401	0.212
		2	0.522	0.478	0.389	0.235
Gemini-2.5-Flash	seen	5	0.012	0.014	0.068	0.132
		2	0.007	0.027	0.070	0.115
	unseen	5	0.010	0.020	0.044	0.091
		2	0.006	0.020	0.040	0.055
Gemini-2.5-Pro	seen	5	0.035	0.046	0.067	0.110
		2	0.034	0.045	0.061	0.064
	unseen	5	0.109	0.154	0.191	0.164
		2	0.034	0.070	0.085	0.076
DeepSeek-R1-Distill-Llama-8B	seen	5	0.111	0.129	0.137	0.146
		2	0.103	0.123	0.133	0.134
	unseen	5	0.112	0.110	0.099	0.070
		2	0.121	0.117	0.104	0.077

Table 9: Explication quality ($e_{\theta,\ell}$) at $K = 5$ (proportion of top-5 questions selected) by context level and discount factor γ. Context levels: ℓ_1 = none, ℓ_2 = demographics, ℓ_3 = preferences, ℓ_4 = demographics + preferences.

Model	Visibility	Cost(λ_-)	$\gamma = 0.3$				$\gamma = 0.5$				$\gamma = 0.7$				$\gamma = 0.9$			
			ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4
Gemma-2-2B	seen	5	0.177	0.267	0.237	0.308	0.152	0.224	0.188	0.246	0.126	0.181	0.137	0.179	0.068	0.095	0.059	0.077
		2	0.180	0.264	0.226	0.318	0.143	0.206	0.172	0.238	0.103	0.145	0.115	0.156	0.051	0.069	0.049	0.065
	unseen	5	0.244	0.376	0.293	0.397	0.206	0.317	0.247	0.331	0.170	0.257	0.192	0.254	0.097	0.135	0.087	0.110
		2	0.235	0.372	0.281	0.387	0.186	0.292	0.224	0.302	0.138	0.207	0.161	0.211	0.076	0.101	0.076	0.095
Gemma-2-9B	seen	5	0.765	0.774	0.603	0.620	0.708	0.716	0.578	0.602	0.624	0.636	0.540	0.569	0.481	0.505	0.471	0.504
		2	0.764	0.779	0.613	0.616	0.691	0.713	0.581	0.595	0.583	0.616	0.528	0.556	0.423	0.476	0.438	0.482
	unseen	5	0.819	0.770	0.768	0.724	0.767	0.730	0.746	0.713	0.685	0.663	0.708	0.691	0.545	0.550	0.639	0.650
		2	0.827	0.786	0.781	0.739	0.769	0.743	0.756	0.723	0.666	0.663	0.708	0.694	0.506	0.533	0.613	0.628
Llama-3.1-8B	seen	5	0.562	0.584	0.616	0.606	0.555	0.568	0.591	0.578	0.534	0.538	0.542	0.531	0.454	0.451	0.415	0.410
		2	0.556	0.562	0.605	0.587	0.529	0.530	0.558	0.540	0.480	0.478	0.482	0.472	0.362	0.363	0.335	0.334
	unseen	5	0.500	0.471	0.652	0.628	0.505	0.480	0.631	0.604	0.500	0.483	0.589	0.565	0.449	0.442	0.460	0.451
		2	0.493	0.463	0.641	0.626	0.494	0.472	0.598	0.584	0.474	0.464	0.529	0.521	0.375	0.379	0.374	0.377
Llama-3.1-70B	seen	5	0.745	0.851	0.769	0.797	0.717	0.809	0.733	0.762	0.666	0.738	0.672	0.700	0.523	0.549	0.518	0.528
		2	0.730	0.820	0.739	0.770	0.667	0.731	0.664	0.693	0.577	0.611	0.568	0.587	0.419	0.416	0.416	0.409
	unseen	5	0.687	0.770	0.807	0.842	0.674	0.733	0.763	0.799	0.640	0.676	0.689	0.721	0.515	0.513	0.513	0.520
		2	0.671	0.769	0.778	0.809	0.631	0.694	0.692	0.719	0.561	0.589	0.581	0.597	0.414	0.403	0.407	0.399
Aya-Expanse-8B	seen	5	0.671	0.567	0.583	0.611	0.652	0.567	0.556	0.604	0.630	0.565	0.531	0.600	0.593	0.561	0.506	0.597
		2	0.687	0.509	0.582	0.625	0.678	0.505	0.560	0.619	0.665	0.501	0.538	0.614	0.633	0.496	0.513	0.609
	unseen	5	0.529	0.547	0.605	0.642	0.498	0.535	0.561	0.619	0.472	0.525	0.524	0.597	0.443	0.511	0.481	0.567
		2	0.557	0.563	0.625	0.661	0.545	0.555	0.585	0.647	0.531	0.547	0.542	0.631	0.502	0.531	0.485	0.606
Aya-Expanse-32B	seen	5	0.473	0.671	0.542	0.659	0.479	0.666	0.540	0.653	0.480	0.652	0.530	0.639	0.465	0.617	0.494	0.603
		2	0.463	0.669	0.512	0.669	0.471	0.663	0.510	0.660	0.472	0.650	0.503	0.644	0.450	0.618	0.473	0.601
	unseen	5	0.481	0.620	0.631	0.711	0.491	0.625	0.626	0.710	0.500	0.629	0.614	0.704	0.503	0.624	0.579	0.682
		2	0.480	0.636	0.617	0.709	0.491	0.640	0.612	0.708	0.499	0.643	0.597	0.701	0.496	0.636	0.559	0.679
Qwen3.5-9B	seen	5	0.562	0.622	0.636	0.671	0.545	0.603	0.615	0.652	0.514	0.570	0.570	0.610	0.449	0.493	0.473	0.509
		2	0.552	0.613	0.640	0.680	0.532	0.584	0.606	0.649	0.493	0.532	0.543	0.585	0.414	0.432	0.426	0.457
	unseen	5	0.480	0.685	0.679	0.724	0.476	0.650	0.645	0.693	0.461	0.589	0.584	0.633	0.414	0.467	0.460	0.503
		2	0.481	0.707	0.689	0.738	0.480	0.659	0.638	0.694	0.459	0.580	0.559	0.611	0.389	0.439	0.424	0.458
Qwen3.5-27B	seen	5	0.975	0.977	0.940	0.942	0.960	0.962	0.925	0.932	0.934	0.935	0.899	0.912	0.869	0.869	0.840	0.869
		2	0.962	0.971	0.933	0.946	0.932	0.946	0.910	0.930	0.887	0.905	0.875	0.904	0.796	0.828	0.811	0.853
	unseen	5	0.987	0.991	0.969	0.982	0.975	0.981	0.954	0.971	0.947	0.955	0.923	0.945	0.856	0.876	0.846	0.886
		2	0.972	0.982	0.971	0.975	0.937	0.956	0.947	0.958	0.874	0.904	0.898	0.924	0.740	0.788	0.791	0.845
Gemini-2.5-Flash	seen	5	0.941	0.944	0.919	0.899	0.937	0.941	0.921	0.879	0.909	0.927	0.906	0.860	0.813	0.869	0.827	0.816
		2	0.917	0.932	0.877	0.900	0.873	0.900	0.852	0.878	0.793	0.839	0.810	0.855	0.683	0.746	0.729	0.816
	unseen	5	0.946	0.968	0.896	0.910	0.946	0.959	0.885	0.893	0.934	0.929	0.860	0.864	0.856	0.838	0.761	0.788
		2	0.931	0.930	0.859	0.897	0.907	0.867	0.816	0.873	0.850	0.779	0.765	0.824	0.735	0.676	0.701	0.719
Gemini-2.5-Pro	seen	5	0.994	0.988	0.954	0.990	0.989	0.973	0.934	0.984	0.971	0.918	0.902	0.965	0.885	0.753	0.811	0.904
		2	0.988	0.956	0.953	0.941	0.965	0.899	0.906	0.885	0.911	0.805	0.839	0.803	0.795	0.660	0.740	0.679
	unseen	5	0.998	0.998	0.994	0.967	0.992	0.993	0.988	0.941	0.966	0.965	0.965	0.885	0.880	0.848	0.861	0.759
		2	0.989	0.975	0.937	0.948	0.962	0.924	0.883	0.888	0.902	0.823	0.809	0.797	0.803	0.631	0.700	0.663
DeepSeek-R1-Distill-Llama-8B	seen	5	0.377	0.402	0.384	0.357	0.328	0.366	0.324	0.310	0.265	0.310	0.252	0.250	0.159	0.194	0.145	0.153
		2	0.372	0.384	0.376	0.355	0.309	0.337	0.305	0.299	0.234	0.272	0.224	0.228	0.132	0.162	0.123	0.133
	unseen	5	0.340	0.346	0.425	0.389	0.306	0.315	0.367	0.346	0.263	0.274	0.293	0.285	0.171	0.183	0.167	0.169
		2	0.324	0.346	0.424	0.387	0.275	0.299	0.349	0.327	0.221	0.242	0.260	0.252	0.138	0.152	0.142	0.144

Table 10: Question Quality ($\mathrm{Qual}@5(\theta)$) by visibility, explore cost and discount factor γ: the mean $\frac{1}{c}\sum_{\ell} e_{\theta,\ell}$ of the per-level explication quality at $K = 5$ in Table 9.

Model	Visibility	Cost (λ_{-})	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$
Gemma-2-2B	seen	5	0.247	0.202	0.156	0.075
		2	0.247	0.190	0.130	0.058
	unseen	5	0.327	0.275	0.218	0.107
		2	0.319	0.251	0.179	0.087
Gemma-2-9B	seen	5	0.690	0.651	0.592	0.490
		2	0.693	0.645	0.571	0.455
	unseen	5	0.770	0.739	0.687	0.596
		2	0.783	0.748	0.682	0.570
Llama-3.1-8B	seen	5	0.592	0.573	0.536	0.432
		2	0.577	0.540	0.478	0.348
	unseen	5	0.563	0.555	0.534	0.450
		2	0.556	0.537	0.497	0.376
Llama-3.1-70B	seen	5	0.791	0.755	0.694	0.530
		2	0.765	0.689	0.586	0.415
	unseen	5	0.776	0.742	0.681	0.515
		2	0.757	0.684	0.582	0.406
Aya-Expanse-8B	seen	5	0.608	0.595	0.581	0.564
		2	0.601	0.591	0.579	0.563
	unseen	5	0.581	0.553	0.529	0.500
		2	0.601	0.583	0.563	0.531
Aya-Expanse-32B	seen	5	0.586	0.584	0.575	0.545
		2	0.578	0.576	0.567	0.535
	unseen	5	0.611	0.613	0.612	0.597
		2	0.611	0.612	0.610	0.593
Qwen3.5-9B	seen	5	0.623	0.604	0.566	0.481
		2	0.621	0.593	0.538	0.432
	unseen	5	0.642	0.616	0.567	0.461
		2	0.654	0.618	0.552	0.428
Qwen3.5-27B	seen	5	0.958	0.945	0.920	0.862
		2	0.953	0.930	0.893	0.822
	unseen	5	0.982	0.970	0.943	0.866
		2	0.975	0.950	0.900	0.791
Gemini-2.5-Flash	seen	5	0.926	0.919	0.900	0.831
		2	0.906	0.876	0.824	0.744
	unseen	5	0.930	0.921	0.897	0.811
		2	0.904	0.866	0.805	0.708
Gemini-2.5-Pro	seen	5	0.982	0.970	0.939	0.839
		2	0.960	0.914	0.839	0.718
	unseen	5	0.989	0.979	0.945	0.837
		2	0.962	0.914	0.833	0.699
DeepSeek-R1-Distill-Llama-8B	seen	5	0.380	0.332	0.269	0.163
		2	0.372	0.313	0.240	0.138
	unseen	5	0.375	0.333	0.279	0.173
		2	0.370	0.313	0.244	0.144

Table 11: Explication quality ($e_{\theta,\ell}$) at $K = 10$ (proportion of explore turns that selected a top-10 question) by context level and discount factor γ. Context levels: ℓ_1 = none, ℓ_2 = demographics, ℓ_3 = preferences, ℓ_4 = demographics + preferences.

Model	Visibility	Cost(λ_-)	$\gamma = 0.3$				$\gamma = 0.5$				$\gamma = 0.7$				$\gamma = 0.9$			
			ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_1	ℓ_2	ℓ_3	ℓ_4
Gemma-2-2B	seen	5	0.532	0.446	0.511	0.397	0.470	0.397	0.434	0.324	0.408	0.345	0.351	0.246	0.238	0.197	0.176	0.114
		2	0.510	0.448	0.495	0.395	0.419	0.369	0.397	0.304	0.320	0.281	0.291	0.210	0.171	0.147	0.144	0.095
	unseen	5	0.482	0.487	0.555	0.475	0.445	0.444	0.500	0.407	0.406	0.396	0.424	0.323	0.270	0.245	0.218	0.149
		2	0.466	0.483	0.528	0.456	0.404	0.411	0.446	0.365	0.336	0.331	0.348	0.265	0.209	0.192	0.186	0.130
Gemma-2-9B	seen	5	0.884	0.819	0.804	0.742	0.860	0.773	0.790	0.726	0.821	0.712	0.768	0.696	0.734	0.628	0.725	0.641
		2	0.881	0.814	0.821	0.742	0.850	0.763	0.799	0.720	0.795	0.694	0.762	0.683	0.684	0.604	0.696	0.621
	unseen	5	0.908	0.804	0.895	0.840	0.890	0.773	0.885	0.828	0.854	0.726	0.868	0.806	0.776	0.660	0.834	0.766
		2	0.917	0.816	0.905	0.847	0.894	0.784	0.893	0.833	0.843	0.728	0.870	0.805	0.744	0.652	0.820	0.746
Llama-3.1-8B	seen	5	0.835	0.830	0.842	0.777	0.822	0.809	0.821	0.748	0.791	0.771	0.775	0.699	0.688	0.666	0.640	0.569
		2	0.805	0.800	0.826	0.760	0.768	0.756	0.779	0.708	0.708	0.690	0.703	0.633	0.560	0.545	0.527	0.477
	unseen	5	0.828	0.782	0.871	0.826	0.820	0.776	0.851	0.801	0.795	0.759	0.808	0.756	0.697	0.674	0.670	0.627
		2	0.831	0.784	0.846	0.806	0.807	0.763	0.807	0.762	0.753	0.717	0.740	0.692	0.592	0.576	0.562	0.528
Llama-3.1-70B	seen	5	0.957	0.948	0.962	0.936	0.937	0.923	0.945	0.914	0.904	0.878	0.912	0.870	0.797	0.739	0.803	0.728
		2	0.944	0.932	0.949	0.919	0.901	0.875	0.906	0.863	0.833	0.790	0.838	0.779	0.661	0.609	0.671	0.602
	unseen	5	0.939	0.918	0.958	0.946	0.926	0.887	0.935	0.917	0.897	0.843	0.896	0.863	0.791	0.717	0.779	0.703
		2	0.920	0.907	0.940	0.920	0.887	0.848	0.890	0.855	0.826	0.765	0.817	0.761	0.656	0.592	0.645	0.577
Aya-Expanse-8B	seen	5	0.721	0.642	0.603	0.664	0.702	0.641	0.576	0.658	0.682	0.640	0.551	0.655	0.649	0.636	0.526	0.654
		2	0.736	0.577	0.609	0.675	0.728	0.573	0.586	0.669	0.716	0.570	0.564	0.664	0.689	0.567	0.539	0.659
	unseen	5	0.709	0.771	0.684	0.696	0.673	0.760	0.638	0.673	0.645	0.750	0.600	0.652	0.620	0.735	0.555	0.624
		2	0.753	0.776	0.698	0.727	0.741	0.770	0.658	0.713	0.728	0.762	0.616	0.697	0.706	0.746	0.560	0.672
Aya-Expanse-32B	seen	5	0.810	0.788	0.809	0.775	0.804	0.772	0.797	0.763	0.791	0.746	0.773	0.742	0.753	0.697	0.713	0.692
		2	0.789	0.793	0.787	0.778	0.783	0.775	0.773	0.764	0.766	0.747	0.748	0.742	0.721	0.699	0.691	0.692
	unseen	5	0.786	0.794	0.865	0.828	0.784	0.780	0.855	0.821	0.781	0.765	0.838	0.809	0.769	0.740	0.798	0.779
		2	0.786	0.794	0.853	0.836	0.782	0.776	0.840	0.826	0.775	0.760	0.820	0.810	0.758	0.737	0.776	0.779
Qwen3.5-9B	seen	5	0.872	0.725	0.915	0.809	0.861	0.713	0.900	0.800	0.840	0.692	0.869	0.775	0.788	0.645	0.793	0.697
		2	0.852	0.722	0.903	0.812	0.836	0.704	0.882	0.796	0.808	0.671	0.839	0.752	0.734	0.600	0.733	0.646
	unseen	5	0.888	0.769	0.915	0.845	0.877	0.749	0.900	0.829	0.859	0.712	0.871	0.791	0.803	0.634	0.791	0.689
		2	0.871	0.786	0.917	0.849	0.855	0.759	0.892	0.824	0.826	0.710	0.847	0.769	0.732	0.612	0.733	0.641
Qwen3.5-27B	seen	5	0.996	0.994	0.992	0.991	0.993	0.991	0.989	0.988	0.988	0.986	0.983	0.984	0.969	0.971	0.966	0.972
		2	0.994	0.994	0.989	0.993	0.986	0.988	0.984	0.989	0.973	0.978	0.974	0.982	0.935	0.950	0.953	0.965
	unseen	5	0.998	0.998	0.995	0.997	0.996	0.997	0.991	0.995	0.990	0.993	0.984	0.990	0.964	0.975	0.965	0.977
		2	0.995	0.997	0.996	0.996	0.987	0.993	0.991	0.992	0.966	0.979	0.978	0.985	0.906	0.930	0.942	0.963
Gemini-2.5-Flash	seen	5	0.995	1.000	1.000	0.948	0.988	1.000	0.999	0.929	0.967	0.999	0.990	0.915	0.915	0.987	0.947	0.897
		2	0.980	0.994	0.956	0.952	0.943	0.976	0.935	0.935	0.886	0.940	0.916	0.929	0.812	0.881	0.885	0.924
	unseen	5	0.995	0.999	0.964	0.969	0.994	0.997	0.952	0.954	0.992	0.986	0.939	0.934	0.973	0.955	0.895	0.894
		2	0.988	0.976	0.938	0.965	0.979	0.935	0.904	0.950	0.954	0.877	0.884	0.923	0.871	0.790	0.861	0.841
Gemini-2.5-Pro	seen	5	0.997	0.991	0.962	0.996	0.995	0.980	0.948	0.994	0.992	0.939	0.937	0.991	0.973	0.828	0.923	0.979
		2	0.998	0.965	0.968	0.951	0.995	0.926	0.943	0.910	0.981	0.869	0.919	0.856	0.931	0.784	0.888	0.773
	unseen	5	0.999	1.000	0.998	0.971	0.997	1.000	0.997	0.948	0.987	1.000	0.995	0.899	0.965	1.000	0.978	0.807
		2	0.996	0.983	0.948	0.958	0.984	0.951	0.913	0.914	0.960	0.890	0.879	0.853	0.917	0.758	0.831	0.762
DeepSeek-R1-Distill-Llama-8B	seen	5	0.539	0.569	0.524	0.500	0.482	0.516	0.451	0.437	0.410	0.436	0.365	0.357	0.278	0.284	0.231	0.226
		2	0.533	0.542	0.511	0.498	0.454	0.473	0.425	0.422	0.363	0.382	0.327	0.327	0.236	0.240	0.200	0.198
	unseen	5	0.577	0.523	0.581	0.530	0.522	0.470	0.515	0.477	0.459	0.406	0.432	0.401	0.325	0.279	0.280	0.250
		2	0.569	0.506	0.591	0.531	0.488	0.436	0.501	0.455	0.404	0.355	0.398	0.360	0.275	0.232	0.251	0.218

Table 12: Question Quality ($\mathrm{Qual@10}(\theta)$) by visibility, explore cost and discount factor γ: the mean $\frac{1}{c}\sum_{\ell} e_{\theta,\ell}$ of the per-level explication quality at $K = 10$ in Table 11.

Model	Visibility	Cost (λ_-)	$\gamma = 0.3$	$\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.9$
Gemma-2-2B	seen	5	0.472	0.406	0.338	0.181
		2	0.462	0.372	0.275	0.139
	unseen	5	0.500	0.449	0.387	0.221
		2	0.483	0.406	0.320	0.179
Gemma-2-9B	seen	5	0.812	0.787	0.749	0.682
		2	0.814	0.783	0.734	0.651
	unseen	5	0.862	0.844	0.814	0.759
		2	0.871	0.851	0.811	0.740
Llama-3.1-8B	seen	5	0.821	0.800	0.759	0.641
		2	0.798	0.753	0.684	0.527
	unseen	5	0.827	0.812	0.779	0.667
		2	0.817	0.785	0.725	0.564
Llama-3.1-70B	seen	5	0.951	0.930	0.891	0.767
		2	0.936	0.886	0.810	0.636
	unseen	5	0.940	0.916	0.875	0.747
		2	0.922	0.870	0.792	0.618
Aya-Expanse-8B	seen	5	0.658	0.644	0.632	0.616
		2	0.649	0.639	0.629	0.614
	unseen	5	0.715	0.686	0.662	0.633
		2	0.739	0.720	0.701	0.671
Aya-Expanse-32B	seen	5	0.795	0.784	0.763	0.714
		2	0.787	0.774	0.751	0.701
	unseen	5	0.818	0.810	0.798	0.772
		2	0.817	0.806	0.791	0.762
Qwen3.5-9B	seen	5	0.830	0.818	0.794	0.731
		2	0.822	0.804	0.768	0.678
	unseen	5	0.854	0.839	0.808	0.729
		2	0.856	0.833	0.788	0.679
Qwen3.5-27B	seen	5	0.993	0.991	0.985	0.969
		2	0.992	0.987	0.977	0.951
	unseen	5	0.997	0.995	0.989	0.970
		2	0.996	0.991	0.977	0.935
Gemini-2.5-Flash	seen	5	0.986	0.979	0.968	0.937
		2	0.970	0.947	0.918	0.876
	unseen	5	0.982	0.974	0.963	0.929
		2	0.967	0.942	0.910	0.841
Gemini-2.5-Pro	seen	5	0.987	0.979	0.965	0.926
		2	0.971	0.944	0.906	0.844
	unseen	5	0.992	0.985	0.970	0.938
		2	0.971	0.941	0.895	0.817
DeepSeek-R1-Distill-Llama-8B	seen	5	0.533	0.471	0.392	0.255
		2	0.521	0.444	0.350	0.219
	unseen	5	0.553	0.496	0.424	0.283
		2	0.549	0.470	0.379	0.244

Table 13: Trend Alignment (Align) at $K = 5$ by visibility, explore cost and discount factor γ. ρ_{Rate} and ρ_{EXP} are the negated Pearson correlations of the per-level explore rate (Table 7) and explication quality (Table 9) with context level; $\mathrm{Align} = 1 - \sqrt{(1-\rho_{\mathrm{Rate}})^2 + \rho_{\mathrm{EXP}}^2}/\sqrt{5}$, the normalized closeness to the ideal $(1, 0)$.

Model	Visibility	Cost (λ_-)	$\gamma = 0.3$			$\gamma = 0.5$			$\gamma = 0.7$			$\gamma = 0.9$		
			ρ_{Rate}	ρ_{EXP}	Align	ρ_{Rate}	ρ_{EXP}	Align	ρ_{Rate}	ρ_{EXP}	Align	ρ_{Rate}	ρ_{EXP}	Align
Gemma-2-2B	seen	5	0.772	-0.853	0.605	0.894	-0.768	0.653	0.975	-0.528	0.764	0.992	0.087	0.961
		2	0.829	-0.833	0.620	0.931	-0.786	0.647	0.986	-0.668	0.701	0.988	-0.261	0.883
	unseen	5	0.806	-0.683	0.683	0.863	-0.664	0.697	0.975	-0.547	0.755	0.981	0.058	0.973
		2	0.866	-0.648	0.704	0.927	-0.652	0.707	0.991	-0.619	0.723	0.985	-0.320	0.857
Gemma-2-9B	seen	5	0.932	0.855	0.616	0.947	0.828	0.629	0.973	0.748	0.665	0.993	-0.252	0.887
		2	0.967	0.865	0.613	0.982	0.813	0.636	0.995	0.578	0.742	1.000	-0.625	0.720
	unseen	5	0.960	0.953	0.573	0.958	0.817	0.634	0.957	-0.447	0.799	0.952	-0.926	0.585
		2	0.959	0.963	0.569	0.960	0.824	0.631	0.961	-0.756	0.662	0.957	-0.966	0.567
Llama-3.1-8B	seen	5	0.298	-0.879	0.497	0.287	-0.775	0.529	0.286	0.203	0.668	-0.218	0.936	0.313
		2	0.327	-0.768	0.543	0.225	-0.583	0.566	-0.042	0.614	0.459	-0.613	0.889	0.176
	unseen	5	0.898	-0.805	0.637	0.981	-0.785	0.649	0.929	-0.764	0.657	0.751	-0.437	0.775
		2	0.877	-0.820	0.629	0.563	-0.809	0.589	0.201	-0.813	0.490	-0.300	-0.021	0.419
Llama-3.1-70B	seen	5	-0.403	-0.213	0.365	-0.368	-0.191	0.382	-0.299	-0.131	0.416	-0.325	0.160	0.403
		2	-0.541	-0.119	0.309	-0.500	-0.055	0.329	-0.532	0.088	0.314	-0.975	0.931	0.024
	unseen	5	-0.372	-0.975	0.247	-0.363	-0.988	0.247	-0.280	-0.987	0.277	-0.109	-0.587	0.439
		2	-0.526	-0.915	0.204	-0.656	-0.904	0.156	-0.866	-0.839	0.085	-0.898	0.838	0.072
Aya-Expanse-8B	seen	5	0.434	0.462	0.673	0.417	0.455	0.669	0.382	0.373	0.677	0.190	0.134	0.633
		2	0.316	0.194	0.682	0.273	0.213	0.661	0.186	0.200	0.625	-0.075	0.102	0.517
	unseen	5	0.998	-0.983	0.560	0.990	-0.986	0.559	0.976	-0.940	0.580	0.922	-0.845	0.621
		2	0.997	-0.960	0.571	0.993	-0.943	0.578	0.978	-0.828	0.629	0.867	-0.639	0.708
Aya-Expanse-32B	seen	5	0.653	-0.580	0.698	0.714	-0.564	0.717	0.785	-0.548	0.737	0.856	-0.491	0.771
		2	0.725	-0.558	0.722	0.782	-0.535	0.742	0.833	-0.510	0.760	0.882	-0.461	0.787
	unseen	5	0.862	-0.948	0.572	0.894	-0.938	0.578	0.929	-0.915	0.589	0.955	-0.840	0.624
		2	0.844	-0.902	0.590	0.868	-0.887	0.599	0.899	-0.847	0.618	0.934	-0.749	0.664
Qwen3.5-9B	seen	5	0.909	-0.969	0.565	0.894	-0.968	0.565	0.895	-0.944	0.575	0.912	-0.805	0.638
		2	0.948	-0.987	0.558	0.923	-0.990	0.556	0.887	-0.979	0.559	0.748	-0.884	0.589
	unseen	5	0.990	-0.853	0.619	0.987	-0.871	0.610	0.988	-0.892	0.601	0.987	-0.919	0.589
		2	0.998	-0.831	0.628	0.995	-0.847	0.621	0.988	-0.855	0.617	0.971	-0.848	0.621
Qwen3.5-27B	seen	5	0.986	0.873	0.609	0.990	0.833	0.628	0.994	0.751	0.664	0.999	0.264	0.882
		2	0.981	0.654	0.708	0.985	0.376	0.832	0.991	-0.169	0.924	0.999	-0.809	0.638
	unseen	5	0.983	0.488	0.782	0.989	0.455	0.796	0.994	0.357	0.840	0.997	-0.408	0.817
		2	0.977	0.047	0.977	0.988	-0.713	0.681	0.995	-0.901	0.597	0.997	-0.957	0.572
Gemini-2.5-Flash	seen	5	-0.403	0.926	0.248	0.741	0.886	0.587	0.995	0.758	0.661	0.996	0.165	0.926
		2	0.692	0.592	0.701	0.943	0.225	0.896	0.985	-0.727	0.675	0.987	-0.892	0.601
	unseen	5	0.992	0.706	0.684	0.989	0.808	0.638	0.976	0.897	0.599	0.968	0.825	0.631
		2	0.940	0.651	0.707	0.829	0.530	0.751	0.751	0.301	0.825	0.742	0.123	0.872
Gemini-2.5-Pro	seen	5	-0.497	0.315	0.316	-0.288	0.293	0.409	0.229	0.129	0.650	0.281	-0.210	0.665
		2	-0.390	0.930	0.252	-0.739	0.853	0.134	-0.931	0.740	0.075	-0.785	0.564	0.163
	unseen	5	-0.283	0.844	0.313	-0.302	0.812	0.314	-0.350	0.781	0.302	-0.291	0.839	0.311
		2	-0.695	0.873	0.147	-0.785	0.929	0.100	-0.880	0.899	0.068	-0.936	0.606	0.093
DeepSeek-R1-Distill-Llama-8B	seen	5	0.618	0.539	0.704	0.394	0.510	0.646	0.135	0.482	0.557	-0.092	0.407	0.479
		2	0.529	0.614	0.654	0.330	0.480	0.631	0.127	0.385	0.573	-0.005	0.274	0.534
	unseen	5	0.960	-0.735	0.671	0.976	-0.789	0.647	0.989	-0.830	0.629	0.766	0.406	0.791
		2	0.967	-0.776	0.653	0.986	-0.828	0.630	0.997	-0.855	0.618	0.899	-0.152	0.918

Table 14: Mean number of interactions per run, by context level. Context levels: **l1** = none, **l2** = demographics, **l3** = preferences, **l4** = demo. + pref.; **All** pools the four levels.

Model	Visibility	Cost (λ_-)	ℓ_1	ℓ_2	ℓ_3	ℓ_4	All
Gemma-2-2B	seen	5	53.35	50.51	66.05	62.13	58.01
		2	75.17	70.55	76.13	73.66	73.88
	unseen	5	46.89	45.15	58.89	58.98	52.47
		2	75.78	70.12	72.48	70.56	72.23
Gemma-2-9B	seen	5	21.08	20.99	20.69	20.60	20.84
		2	22.56	22.00	21.46	20.96	21.75
	unseen	5	20.68	20.66	20.16	20.15	20.41
		2	21.27	21.03	20.28	20.27	20.71
Llama-3.1-8B	seen	5	22.55	22.76	23.91	23.85	23.27
		2	28.69	28.88	32.86	31.47	30.48
	unseen	5	22.70	22.94	23.75	23.63	23.25
		2	27.54	27.63	30.53	29.42	28.78
Llama-3.1-70B	seen	5	22.60	23.03	22.32	22.84	22.70
		2	27.06	28.37	26.58	27.64	27.41
	unseen	5	22.59	23.19	22.73	23.23	22.93
		2	27.23	28.41	27.55	28.66	27.96
Aya-Expanse-8B	seen	5	20.17	20.01	20.51	20.11	20.20
		2	20.08	20.02	20.41	20.11	20.16
	unseen	5	20.64	20.21	21.08	20.25	20.54
		2	20.33	20.15	21.12	20.17	20.44
Aya-Expanse-32B	seen	5	20.68	20.67	20.82	20.50	20.67
		2	21.02	20.78	20.92	20.55	20.82
	unseen	5	20.36	20.41	20.43	20.19	20.35
		2	20.52	20.50	20.53	20.23	20.45
Qwen3.5-9B	seen	5	20.68	20.85	20.76	20.80	20.77
		2	21.90	22.47	22.03	21.95	22.09
	unseen	5	20.76	21.03	21.02	21.00	20.95
		2	22.19	22.41	22.38	22.29	22.32
Qwen3.5-27B	seen	5	20.17	20.12	20.11	20.05	20.11
		2	20.56	20.40	20.23	20.15	20.34
	unseen	5	20.22	20.17	20.13	20.08	20.15
		2	20.99	20.78	20.37	20.23	20.59
Gemini-2.5-Flash	seen	5	20.75	20.17	20.54	20.66	20.53
		2	21.94	21.22	20.98	20.51	21.16
	unseen	5	20.34	20.51	20.98	21.00	20.71
		2	21.44	22.45	21.09	21.76	21.69
Gemini-2.5-Pro	seen	5	20.30	20.79	20.57	20.09	20.44
		2	20.89	22.11	21.13	22.07	21.55
	unseen	5	20.36	20.00	20.27	20.79	20.35
		2	20.91	22.80	21.90	22.42	22.01
DeepSeek-R1-Distill-Llama-8B	seen	5	37.15	31.75	41.38	37.53	36.95
		2	46.67	37.57	49.12	43.13	44.12
	unseen	5	37.21	35.12	41.66	39.62	38.40
		2	54.10	47.97	55.34	50.20	51.90

Model	Sens	ΔSens	Qual@5	ΔQual@5	Align	ΔAlign	Overall	ΔOverall	95% CI
Qwen3.5-27B	0.212	−0.022	**0.866**	+0.004	0.817	−0.065	**0.632**	−0.028	[−0.074, +0.023]
Gemma-2-9B	**0.369**	+0.067	0.596	+0.106	0.585	−0.302	0.517	**−0.043***	[−0.077, −0.006]
Gemini-2.5-Flash	0.091	−0.041	0.811	−0.020	0.631	−0.296	0.511	**−0.119***	[−0.140, −0.099]
Aya-Expanse-32B	0.141	+0.016	0.597	+0.052	0.624	−0.147	0.454	**−0.026***	[−0.041, −0.010]
Aya-Expanse-8B	0.196	+0.149	0.500	−0.064	0.621	−0.012	0.439	**+0.024***	[+0.007, +0.060]
Gemini-2.5-Pro	0.164	+0.054	0.837	−0.001	0.311	−0.354	0.438	**−0.100***	[−0.109, −0.090]
Llama-3.1-8B	0.048	−0.034	0.450	+0.018	0.775	+0.462	0.425	**+0.149***	[+0.100, +0.193]
Gemma-2-2B	0.090	+0.041	0.107	+0.033	**0.973**	+0.012	0.390	**+0.029***	[+0.014, +0.043]
Qwen3.5-9B	0.086	+0.036	0.461	−0.020	0.589	−0.049	0.379	−0.011	[−0.028, +0.003]
DeepSeek-R1-Distill-Llama-8B	0.070	−0.076	0.173	+0.010	0.791	+0.312	0.345	**+0.082***	[+0.041, +0.117]
Llama-3.1-70B	0.014	−0.000	0.515	−0.014	0.439	+0.036	0.323	+0.007	[−0.033, +0.048]

Table 15: **Overall ranking: unseen, expensive ($\lambda_- = 5$), $\gamma = 0.9$.** The best score in each column is in **bold**. Each Δ column contains the change from the baseline cell (Table 1). The change in Overall is tested with a paired bootstrap (one resample of users applied to both cells): the last column gives its 95% CI. $\mathbf{\Delta}^*$: significant at $\alpha = 0.05$ after Benjamini–Hochberg correction across the 11 models (corrected $p < 0.05$ and the interval excludes 0).

Model	Sens	ΔSens	Qual@5	ΔQual@5	Align	ΔAlign	Overall	ΔOverall	95% CI
Qwen3.5-27B	0.215	−0.020	**0.822**	−0.040	0.638	−0.244	**0.558**	**−0.101***	[−0.140, −0.054]
Gemini-2.5-Flash	0.115	−0.017	0.744	−0.088	0.601	−0.325	0.487	**−0.143***	[−0.164, −0.122]
Aya-Expanse-32B	0.135	+0.009	0.535	−0.009	0.787	+0.016	0.486	+0.005	[−0.011, +0.021]
Gemma-2-9B	**0.227**	−0.074	0.455	−0.035	0.720	−0.167	0.468	**−0.092***	[−0.127, −0.052]
Aya-Expanse-8B	0.040	−0.007	0.563	−0.002	0.517	−0.116	0.373	**−0.042***	[−0.061, −0.013]
Qwen3.5-9B	0.077	+0.027	0.432	−0.049	0.589	−0.049	0.366	**−0.024***	[−0.043, −0.005]
Gemma-2-2B	0.041	−0.008	0.058	−0.016	**0.883**	−0.078	0.327	**−0.034***	[−0.053, −0.013]
Gemini-2.5-Pro	0.064	−0.046	0.718	−0.120	0.163	−0.502	0.315	**−0.223***	[−0.234, −0.209]
DeepSeek-R1-Distill-Llama-8B	0.134	−0.013	0.138	−0.025	0.534	+0.055	0.268	+0.006	[−0.006, +0.017]
Llama-3.1-8B	0.104	+0.022	0.348	−0.084	0.176	−0.137	0.210	**−0.066***	[−0.083, −0.050]
Llama-3.1-70B	0.016	+0.001	0.415	−0.115	0.024	−0.380	0.151	**−0.164***	[−0.185, −0.134]

Table 16: **Overall ranking: seen, cheap ($\lambda_- = 2$), $\gamma = 0.9$.** The best score in each column is in **bold**. Each Δ column follows its metric and gives the change from the baseline cell (Table 1), this cell minus the baseline. The change in Overall is tested with a paired bootstrap (one resample of users applied to both cells): the last column gives its 95% CI. $\mathbf{\Delta}^*$: significant at $\alpha = 0.05$ after Benjamini–Hochberg correction across the 11 models (corrected $p < 0.05$ and the interval excludes 0).

Model	Sens	ΔSens	Qual@5	ΔQual@5	Align	ΔAlign	Overall	ΔOverall	95% CI
Qwen3.5-27B	**0.478**	+0.243	0.920	+0.058	0.664	−0.218	**0.687**	+0.028	[−0.002, +0.060]
Gemini-2.5-Pro	0.067	−0.043	**0.939**	+0.101	0.650	−0.015	0.552	**+0.014***	[+0.006, +0.023]
Gemini-2.5-Flash	0.068	−0.064	0.900	+0.069	0.661	−0.265	0.543	**−0.087***	[−0.103, −0.070]
Aya-Expanse-32B	0.178	+0.053	0.575	+0.031	0.737	−0.035	0.497	**+0.016***	[+0.010, +0.022]
Gemma-2-9B	0.224	−0.078	0.592	+0.102	0.665	−0.222	0.494	**−0.066***	[−0.106, −0.018]
Aya-Expanse-8B	0.097	+0.050	0.581	+0.017	0.677	+0.044	0.452	**+0.037***	[+0.018, +0.058]
Llama-3.1-8B	0.149	+0.066	0.536	+0.104	0.668	+0.355	0.451	**+0.175***	[+0.110, +0.186]
Qwen3.5-9B	0.104	+0.054	0.566	+0.085	0.575	−0.063	0.415	**+0.025***	[+0.015, +0.035]
Llama-3.1-70B	0.009	−0.005	0.694	+0.164	0.416	+0.013	0.373	**+0.057***	[+0.038, +0.077]
Gemma-2-2B	0.047	−0.002	0.156	+0.081	**0.764**	−0.197	0.322	**−0.039***	[−0.064, −0.011]
DeepSeek-R1-Distill-Llama-8B	0.137	−0.009	0.269	+0.106	0.557	+0.078	0.321	**+0.058***	[+0.046, +0.070]

Table 17: **Overall ranking: seen, expensive ($\lambda_- = 5$), $\gamma = 0.7$.** Sensitivity (Sens), Question Quality (Qual@5), and Alignment (Align) lie in $[0, 1]$; Overall is their arithmetic mean. Higher scores are better; the best score in each column is in **bold**. Each Δ column follows its metric and gives the change from the baseline cell (Table 1), this cell minus the baseline. The change in Overall is tested with a paired bootstrap (one resample of users applied to both cells): the last column gives its 95% CI. $\mathbf{\Delta}^*$: significant at $\alpha = 0.05$ after Benjamini–Hochberg correction across the 11 models (corrected $p < 0.05$ and the interval excludes 0).

Model	Sens	ΔSens	Qual@5	ΔQual@5	Align	ΔAlign	Overall	ΔOverall	95% CI
Qwen3.5-27B	**0.603**	+0.369	0.945	+0.083	0.628	−0.254	**0.725**	**+0.066***	[+0.033, +0.103]
Gemini-2.5-Flash	0.014	−0.119	0.919	+0.088	0.587	−0.339	0.507	**−0.123***	[−0.152, −0.104]
Aya-Expanse-32B	0.179	+0.054	0.584	+0.040	**0.717**	−0.054	0.494	**+0.013***	[+0.005, +0.021]
Gemma-2-9B	0.161	−0.141	0.651	+0.161	0.629	−0.258	0.480	**−0.079***	[−0.117, −0.038]
Gemini-2.5-Pro	0.046	−0.064	**0.970**	+0.132	0.409	−0.256	0.475	**−0.063***	[−0.076, −0.049]
Aya-Expanse-8B	0.144	+0.097	0.595	+0.030	0.669	+0.036	0.469	**+0.054***	[+0.024, +0.085]
Llama-3.1-8B	0.194	+0.112	0.573	+0.141	0.529	+0.216	0.432	**+0.156***	[+0.129, +0.202]
Qwen3.5-9B	0.111	+0.061	0.604	+0.123	0.565	−0.073	0.427	**+0.037***	[+0.025, +0.049]
Llama-3.1-70B	0.005	−0.009	0.755	+0.226	0.382	−0.021	0.381	**+0.065***	[+0.042, +0.091]
DeepSeek-R1-Distill-Llama-8B	0.129	−0.018	0.332	+0.169	0.646	+0.167	0.369	**+0.106***	[+0.080, +0.129]
Gemma-2-2B	0.039	−0.010	0.202	+0.128	0.653	−0.308	0.298	**−0.063***	[−0.082, −0.042]

Table 18: **Overall ranking: seen, expensive ($\lambda_- = 5$), $\gamma = 0.5$.** Sensitivity (Sens), Question Quality (Qual@5), and Alignment (Align) lie in $[0, 1]$; Overall is their arithmetic mean. Higher scores are better; the best score in each column is in **bold**. Each Δ column follows its metric and gives the change from the baseline cell (Table 1), this cell minus the baseline. The change in Overall is tested with a paired bootstrap (one resample of users applied to both cells): the last column gives its 95% CI. $\mathbf{\Delta}^*$: significant at $\alpha = 0.05$ after Benjamini–Hochberg correction across the 11 models (corrected $p < 0.05$ and the interval excludes 0).

Model	Sens	ΔSens	Qual@5	ΔQual@5	Align	ΔAlign	Overall	ΔOverall	95% CI
Qwen3.5-27B	**0.665**	+0.430	0.958	+0.097	0.609	−0.273	**0.744**	**+0.085***	[+0.052, +0.127]
Aya-Expanse-8B	0.188	+0.141	0.608	+0.043	0.673	+0.040	0.490	**+0.075***	[+0.036, +0.107]
Aya-Expanse-32B	0.172	+0.046	0.586	+0.041	0.698	−0.073	0.485	+0.005	[−0.005, +0.014]
Gemma-2-9B	0.111	−0.190	0.690	+0.200	0.616	−0.271	0.473	**−0.087***	[−0.122, −0.046]
Gemini-2.5-Pro	0.035	−0.075	**0.982**	+0.143	0.316	−0.349	0.444	**−0.094***	[−0.108, −0.079]
Llama-3.1-8B	0.226	+0.143	0.592	+0.160	0.497	+0.184	0.438	**+0.162***	[+0.144, +0.190]
Qwen3.5-9B	0.092	+0.042	0.623	+0.142	0.565	−0.073	0.427	**+0.037***	[+0.024, +0.049]
DeepSeek-R1-Distill-Llama-8B	0.111	−0.036	0.380	+0.217	**0.704**	+0.226	0.398	**+0.136***	[+0.096, +0.176]
Gemini-2.5-Flash	0.012	−0.121	0.926	+0.095	0.248	−0.678	0.395	**−0.235***	[−0.273, −0.187]
Llama-3.1-70B	0.002	−0.012	0.791	+0.261	0.365	−0.038	0.386	**+0.070***	[+0.042, +0.100]
Gemma-2-2B	0.033	−0.016	0.247	+0.173	0.605	−0.355	0.295	**−0.066***	[−0.083, −0.046]

Table 19: **Overall ranking: seen, expensive ($\lambda_- = 5$), $\gamma = 0.3$.** Sensitivity (Sens), Question Quality (Qual@5), and Alignment (Align) lie in $[0, 1]$; Overall is their arithmetic mean. Higher scores are better; the best score in each column is in **bold**. Each Δ column follows its metric and gives the change from the baseline cell (Table 1), this cell minus the baseline. The change in Overall is tested with a paired bootstrap (one resample of users applied to both cells): the last column gives its 95% CI. $\mathbf{\Delta}^*$: significant at $\alpha = 0.05$ after Benjamini–Hochberg correction across the 11 models (corrected $p < 0.05$ and the interval excludes 0).

Model	Sens	ΔSens	Qual@10	ΔQual@10	Align	ΔAlign	Overall	ΔOverall	95% CI
Qwen3.5-27B	0.235	+0.000	**0.969**	+0.108	**0.897**	+0.015	**0.701**	+0.041	[−0.063, +0.097]
Gemini-2.5-Flash	0.132	+0.000	0.937	+0.106	0.863	−0.064	0.644	+0.014	[−0.001, +0.029]
Gemma-2-9B	**0.302**	+0.000	0.682	+0.192	0.810	−0.077	0.598	+0.038	[−0.003, +0.087]
Gemini-2.5-Pro	0.110	+0.000	0.926	+0.087	0.665	+0.001	0.567	**+0.029***	[+0.028, +0.031]
Qwen3.5-9B	0.050	+0.000	0.731	+0.250	0.893	+0.255	0.558	**+0.168***	[+0.146, +0.185]
Aya-Expanse-32B	0.125	+0.000	0.714	+0.169	0.646	−0.126	0.495	+0.014	[−0.019, +0.056]
Aya-Expanse-8B	0.047	+0.000	0.616	+0.052	0.627	−0.006	0.430	**+0.015***	[+0.008, +0.033]
Llama-3.1-70B	0.015	+0.000	0.767	+0.237	0.371	−0.032	0.384	**+0.068***	[+0.065, +0.073]
Llama-3.1-8B	0.083	+0.000	0.641	+0.209	0.307	−0.006	0.344	**+0.068***	[+0.061, +0.073]
Gemma-2-2B	0.049	+0.000	0.181	+0.107	0.561	−0.400	0.264	**−0.098***	[−0.109, −0.084]
DeepSeek-R1-Distill-Llama-8B	0.146	+0.000	0.255	+0.092	0.371	−0.108	0.257	−0.005	[−0.010, −0.000]

Table 20: **Overall ranking: seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$, $K = 10$.** Sensitivity (Sens), Question Quality (Qual@10), and Alignment (Align) lie in $[0, 1]$; Overall is their arithmetic mean. Higher scores are better; the best score in each column is in **bold**. Each Δ column follows its metric and gives the change from the baseline cell (Table 1), this cell minus the baseline. The change in Overall is tested with a paired bootstrap (one resample of users applied to both cells): the last column gives its 95% CI. $\mathbf{\Delta}^*$: significant at $\alpha = 0.05$ after Benjamini–Hochberg correction across the 11 models (corrected $p < 0.05$ and the interval excludes 0).

Model	Overall	Med. Rank	ΔRank	Tier	Win rate
Qwen3.5-27B	0.558	1 [1, 1]	0	1	100.0%
Gemini-2.5-Flash	0.487	2 [2, 3]	0	2	87.5%
Aya-Expanse-32B	0.486	3 [2, 4]	↑ 2	2	87.5%
Gemma-2-9B	0.468	4 [3, 4]	↓ 1	2	87.5%
Aya-Expanse-8B	0.373	5 [5, 6]	↑ 1	3	55.6%
Qwen3.5-9B	0.366	6 [5, 6]	↑ 1	3	55.6%
Gemma-2-2B	0.327	7 [7, 8]	↑ 1	4	33.3%
Gemini-2.5-Pro	0.315	8 [7, 8]	↓ 4	4	33.3%
DeepSeek-R1-Distill-Llama-8B	0.268	9 [9, 9]	↑ 2	5	20.0%
Llama-3.1-8B	0.210	10 [10, 10]	0	6	10.0%
Llama-3.1-70B	0.151	11 [11, 11]	↓ 2	7	0.0%

Table 21: Explore-cost ablation: seen, **cheap ($\boldsymbol{\lambda_- = 2}$)**, $\gamma = 0.9$, $K = 5$. Rows are sorted by median rank, and are grouped based on tiers. Med. Rank is the median of the model's rank by Overall (1 = best) over bootstrap resamples of users, with its 95% interval. Tier groups models by their rank intervals. ΔRank is the change in median rank from (seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$, $K = 5$). Win rate is the share of a model's significant pairwise comparisons in that cell that it wins (paired bootstrap on the Overall gap, Benjamini–Hochberg corrected $p < 0.05$ with the interval excluding 0).

Model	Overall	Med. Rank	ΔRank	Tier	Win rate
Qwen3.5-27B	0.632	1 [1, 1]	0	1	100.0%
Gemma-2-9B	0.517	2 [2, 3]	↑ 1	2	88.9%
Gemini-2.5-Flash	0.511	3 [2, 3]	↓ 1	2	88.9%
Aya-Expanse-32B	0.454	4 [4, 5]	↑ 1	3	62.5%
Aya-Expanse-8B	0.439	6 [4, 7]	0	3	57.1%
Gemini-2.5-Pro	0.438	6 [5, 7]	↓ 2	3	50.0%
Llama-3.1-8B	0.425	7 [4, 9]	↑ 3	3	40.0%
Gemma-2-2B	0.390	8 [7, 9]	0	4	25.0%
Qwen3.5-9B	0.379	9 [8, 9]	↓ 2	4	14.3%
DeepSeek-R1-Distill-Llama-8B	0.345	10 [9, 11]	↑ 1	4	0.0%
Llama-3.1-70B	0.323	11 [10, 11]	↓ 2	5	0.0%

Table 22: Visibility ablation: **unseen**, expensive ($\lambda_- = 5$), $\gamma = 0.9$, $K = 5$.

Model	Overall	Med. Rank	ΔRank	Tier	Win rate
Qwen3.5-27B	0.687	1 [1, 1]	0	1	100.0%
Gemini-2.5-Pro	0.552	2 [2, 3]	↑ 2	2	88.9%
Gemini-2.5-Flash	0.543	3 [2, 3]	↓ 1	2	88.9%
Aya-Expanse-32B	0.497	4 [4, 5]	↑ 1	3	66.7%
Gemma-2-9B	0.494	5 [4, 5]	↓ 2	3	66.7%
Aya-Expanse-8B	0.452	6 [6, 8]	0	4	37.5%
Llama-3.1-8B	0.451	7 [6, 8]	↑ 3	4	37.5%
Qwen3.5-9B	0.415	8 [7, 8]	↓ 1	4	37.5%
Llama-3.1-70B	0.373	9 [9, 9]	0	5	20.0%
Gemma-2-2B	0.322	10 [10, 11]	↓ 2	6	0.0%
DeepSeek-R1-Distill-Llama-8B	0.321	11 [10, 11]	0	6	0.0%

(a) seen, expensive ($\lambda_- = 5$), $\boldsymbol{\gamma = 0.7}$, $K = 5$

Model	Overall	Med. Rank	ΔRank	Tier	Win rate
Qwen3.5-27B	0.725	1 [1, 1]	0	1	100.0%
Gemini-2.5-Flash	0.507	2 [2, 4]	0	2	85.7%
Aya-Expanse-32B	0.494	3 [2, 4]	↑ 2	2	83.3%
Gemma-2-9B	0.480	4 [3, 6]	↓ 1	2	83.3%
Gemini-2.5-Pro	0.475	5 [4, 6]	↓ 1	2	66.7%
Aya-Expanse-8B	0.469	6 [3, 8]	0	2	75.0%
Llama-3.1-8B	0.432	7 [6, 8]	↑ 3	3	42.9%
Qwen3.5-9B	0.427	8 [7, 8]	↓ 1	3	37.5%
Llama-3.1-70B	0.381	9 [9, 10]	0	4	11.1%
DeepSeek-R1-Distill-Llama-8B	0.369	10 [9, 10]	↑ 1	4	11.1%
Gemma-2-2B	0.298	11 [11, 11]	↓ 3	5	0.0%

(b) seen, expensive ($\lambda_- = 5$), $\boldsymbol{\gamma = 0.5}$, $K = 5$

Model	Overall	Med. Rank	ΔRank	Tier	Win rate
Qwen3.5-27B	0.744	1 [1, 1]	0	1	100.0%
Aya-Expanse-8B	0.490	2 [2, 6]	↑ 4	2	83.3%
Aya-Expanse-32B	0.485	3 [2, 4]	↑ 2	2	87.5%
Gemma-2-9B	0.473	4 [3, 4]	↓ 1	2	87.5%
Gemini-2.5-Pro	0.444	5 [4, 7]	↓ 1	2	40.0%
Llama-3.1-8B	0.438	6 [4, 8]	↑ 4	2	40.0%
Qwen3.5-9B	0.427	7 [6, 8]	0	2	33.3%
DeepSeek-R1-Distill-Llama-8B	0.398	9 [6, 10]	↑ 2	2	20.0%
Gemini-2.5-Flash	0.395	9 [5, 10]	↓ 7	2	20.0%
Llama-3.1-70B	0.386	9 [8, 10]	0	3	12.5%
Gemma-2-2B	0.295	11 [11, 11]	↓ 3	4	0.0%

(c) seen, expensive ($\lambda_- = 5$), $\boldsymbol{\gamma = 0.3}$, $K = 5$

Table 23: Discount-factor ablation: seen, expensive ($\lambda_- = 5$), $K = 5$, with $\boldsymbol{\gamma \in \{0.7, 0.5, 0.3\}}$.

Model	Overall	Med. Rank	ΔRank	Tier	Win rate
Qwen3.5-27B	0.701	1 [1, 2]	0	1	100.0%
Gemini-2.5-Flash	0.644	2 [1, 2]	0	1	100.0%
Gemma-2-9B	0.598	3 [3, 3]	0	2	80.0%
Gemini-2.5-Pro	0.567	4 [4, 5]	0	3	66.7%
Qwen3.5-9B	0.558	5 [4, 5]	↑ 2	3	66.7%
Aya-Expanse-32B	0.495	6 [6, 6]	↓ 1	4	50.0%
Aya-Expanse-8B	0.430	7 [7, 7]	↓ 1	5	40.0%
Llama-3.1-70B	0.384	8 [8, 8]	↑ 1	6	30.0%
Llama-3.1-8B	0.344	9 [9, 9]	↑ 1	7	20.0%
Gemma-2-2B	0.264	10 [10, 11]	↓ 2	8	0.0%
DeepSeek-R1-Distill-Llama-8B	0.257	11 [10, 11]	0	8	0.0%

Table 24: Top-K ablation: seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$, $\boldsymbol{K = 10}$.

Visibility	Cost (λ_-)	γ	K	Spearman ρ
seen	expensive (5)	0.9	5	–
seen	**cheap (2)**	0.9	5	0.855 [0.79, 0.89]
unseen	expensive (5)	0.9	5	0.882 [0.73, 0.93]
seen	expensive (5)	**0.7**	5	0.891 [0.83, 0.94]
seen	expensive (5)	**0.5**	5	0.882 [0.76, 0.91]
seen	expensive (5)	**0.3**	5	0.518 [0.35, 0.72]
seen	expensive (5)	0.9	**10**	0.945 [0.91, 0.96]

Table 25: Ranking agreement between each ablated cell and the first row (seen, expensive ($\lambda_- = 5$), $\gamma = 0.9$, $K = 5$). Each row changes one factor from the baseline, shown in bold. Spearman ρ is the rank correlation between the two cells' Overall scores across the 11 models, with its 95% interval from a paired bootstrap over users (the same resample applied to both cells).

Visibility	Cost (λ_-)	γ	Monotone	$\ell_1 \to \ell_2$	$\ell_2 \to \ell_3$	$\ell_3 \to \ell_4$	$\ell_1 \to \ell_4$
seen	expensive (5)	0.9	5	9	5	11	10
seen	**cheap (2)**	0.9	4	8	6	9	8
unseen	expensive (5)	0.9	6	11	6	11	11
unseen	**cheap (2)**	0.9	5	8	7	8	8
seen	expensive (5)	**0.7**	4	10	5	11	10
seen	expensive (5)	**0.5**	2	10	3	11	10
seen	expensive (5)	**0.3**	2	9	3	11	8

Table 26: **Direction of exploration across context levels.** Number of models (out of 11) whose exploration rate $r_{\theta,\ell}$ is lower at the second level of each transition. *Monotone* counts models whose rate never increases from ℓ_1 to ℓ_4; $\ell_1 \to \ell_4$ compares only the endpoints. Values from Table 7.