Response-Aware User Memory Selection for LLM Personalization

Jillian Fisher
Department of CSE,
University of Washington
Seattle, WA,
United States of America
jrfish@uw.edu

Jennifer Neville
Microsoft Research
Redmond, WA,
United States of America

Chan Young Park
Microsoft Research
Redmond, WA,
United States of America

Abstract

A common approach to personalization in large language models (LLMs) is to incorporate a subset of the user memory into the prompt at inference time to guide the model's generation. Existing methods select these subsets primarily using similarity between user memory items and input queries, ignoring how features actually affect the model's response distribution. We propose **R**esponse-**U**tility optimization for **M**emory **S**election (RUMS), a novel method that selects user memory items by measuring the mutual information between a subset of memory and the model's outputs, identifying items that reduce response uncertainty and sharpen predictions beyond semantic similarity. We demonstrate that this information-theoretic foundation enables more principled user memory selection, which our results suggest aligns more closely with human selection compared to state-of-the-art methods, and models that are $400\times$ larger. Additionally, we show that memory items selected using RUMS result in better response quality compared to existing approaches, while having up to 95% reduction in computational cost.

1 Introduction

As large language models (LLMs) become ubiquitous, personalization has become critical for user satisfaction. Current approaches to personalization generally fall into two categories: training-based and inference-time methods. Training-based approaches, such as low-rank preference models (Bose et al., 2025) or parameter-efficient fine-tuning (Tan et al., 2024), adapt model parameters to individual users. While effective for small populations, these methods become computationally expensive at scale. Consequently, inference-time personalization, such as including user-specific context in prompts, is often preferred.

However, inference-time personalization raises a central question: **What user information should guide the model?** A naïve strategy is to include all stored user memories, but irrelevant information can overwhelm the model, degrading response quality and naturalness (Li et al., 2023; Rajeev et al., 2025). This problem worsens as memory grows. Existing methods instead retrieve relevant memories by measuring the relationship between a *memory item and the input query*, using heuristic retrieval (Robertson and Zaragoza, 2009), semantic similarity (Lei et al., 2023), or LLM prompting (Handa et al., 2024). While intuitive, **we hypothesize that directly measuring how memories affect output distributions, rather than query similarity, provides a stronger signal for personalization**. Query-based selection may favor redundant, surface-aligned memories while missing non-obvious but useful ones, causing both over- and under-selection relative to true utility. These failures can degrade response quality while increasing latency, cost, and privacy risks (Hagström et al., 2025).

To test this hypothesis, we introduce Response-Utility for Memory Selection (RUMS)[1], which consists of two key components: a novel application of information gain for the utility function (*RUMS-Utility*) that quantifies the informativeness of user memory subsets, and a lightweight model (*RUMS-Model*) that efficiently approximates this utility computation at inference time.

RUMS-Utility directly measures how much a memory subset reduces uncertainty in the model's response distribution by computing the conditional mutual information between candidate memory items and the outputs given the input query. We provide theoretical justification for this objective, showing that under a latent user-intent variable, minimizing predictive entropy approximately maximizes expected user utility. Although this guaran-

[1]Code available: https://github.com/jfisher52/Response_Utility_Optimized_Memory_Selection

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 4215–4267
October 24-29, 2026 ©2026 Association for Computational Linguistics

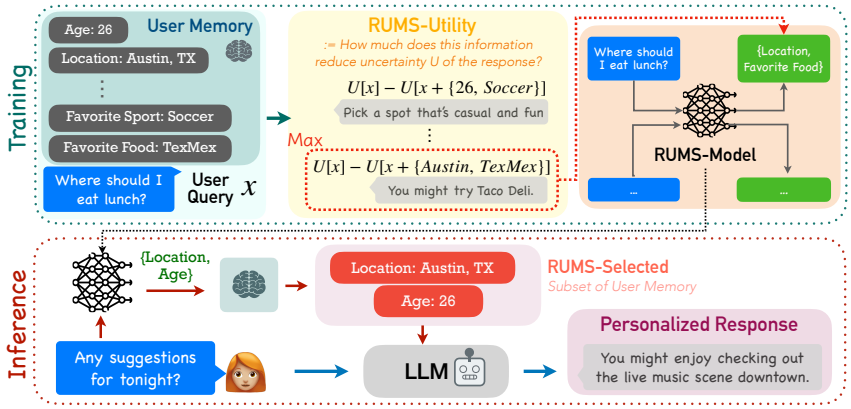


Figure 1: Overview of RUMS. RUMS-Utility quantifies how much memory subsets reduce response uncertainty (training phase), while RUMS-Models efficiently selects informative items at inference.

tee assumes model distributions align with human preferences, our experiments are consistent with the assumption that modern LLMs are sufficiently calibrated for entropy to serve as a strong proxy for human utility.

However, computing *RUMS-Utility* for all queries is computationally expensive, particularly with large user profiles. Therefore, we use the utility function during an offline training phase to generate high-quality training data. We then train *RUMS-Models*, a lightweight encoder-classifiers based on DeBERTa (He et al., 2021), to directly predict memory items that are most informative given a query. This learned approach preserves the principled selection quality of the utility function while enabling efficient deployment.

Through extensive empirical testing, the experimental results validate our hypothesis: RUMS enables more principled memory item selection that is more closely aligned to human selection compared to state-of-the-art (SOTA) memory selection methods and models $400\times$ its size. We also show that the items selected using RUMS result in better response quality than other methods, while lowering the cost of personalization overall.

2 Related Works

User Memory for LLM Personalization Memory augmentation personalizes LLMs by retrieving and incorporating user information at inference time. Early approaches prepend profile summaries (Richardson et al., 2023; Pan et al., 2025) or past interactions (Kang et al., 2023; Liu et al., 2023) to prompts. However, as memory grows, including all user information becomes impractical, motivating retrieval of relevant *subsets*. For example, Chen (2025) use semantic similarity to retrieve historical items for recommendation, while LaMP (Salemi et al., 2024b) studies personalized sentence retrieval. PEARL (Mysore et al., 2024) and ROPG (Salemi et al., 2024a) train retrieval components using user signals to improve personalization. See Liu et al. (2025) for a broader overview. However, these methods primarily rely on input-memory similarity, rather than a memory item's effect on the model's downstream response.

Selective Personalization A related line of work is selective conditioning and abstention, where models ignore uninformative inputs or abstain from prediction. Geifman and El-Yaniv (2017) introduce selective classification for neural networks, while recent work extends these ideas to LLMs by filtering features or context that may harm performance (Hagström et al., 2025). In memory selection, this corresponds to choosing no memory when personalization is unnecessary. Prior approaches typically use global thresholds (Li et al., 2023) or static policies (Joren et al., 2023); in contrast, our method learns an adaptive policy that dynamically selects memory conditioned on the input and context.

Information-Theoretic Methods in NLP Information-theoretic ideas have long shaped NLP (Brown et al., 1992), including multimodal learning (Almudévar et al., 2025), information control (Ravfogel et al., 2020), interpretability (Chen et al., 2018), and prompt design (Zhang et al., 2022). In feature selection, EDDI (Ma et al., 2019) estimates conditional mutual information for feature acquisition, later extended to dynamic settings (Covert et al., 2023; Gadgil et al., 2024). Recent work also uses next-token likelihood differences with and without context to steer generation during de-

coding (He et al., 2025). Closest to personalization, recommendation systems use Bayesian active learning to query users for information that improves recommendations (Handa et al., 2024). We build on these ideas to select memory items that maximize information gain for personalized responses.

3 Method

We aim to develop a method that selects a subset of user memory based on the goal of maximizing the utility of the model's responses.

Problem Formulation Consider user u with query x (e.g. "Where should I eat lunch?"). Let θ be an LLM and $Y = (Y_1, ..., Y_T)$ a random variable over outputs distributed as $p_\theta(Y|x)$. We measure the quality of a specific realized response y by its utility to the user, $\mathrm{U}_u(y)$.

Also, we consider a set of memory items $\mathcal{M}$ (e.g. Location), in which user u has specific values m_u (e.g. Location: Austin, TX). Our aim is to select the memory subset $S \subseteq \mathcal{M}$ with instantiated values $m_{u,S} = \{m_u^{(a)} : a \in S\}$, where a is the memory items selected from the subset S, which when given to the model with the input x yields the highest response utility. Formally,

$$S^*(x,u) = \arg\max_{S \subseteq \mathcal{M}} \mathbb{E}_{y \sim p_\theta(\cdot|x, m_{u,S})}\big[\mathrm{U}_u(y)\big].$$

Unless otherwise specified, expectations are taken under the model's predictive distribution, $\mathbb{E}[\cdot] = \mathbb{E}_{y \sim p_\theta(\cdot|x, m_{u,S})}[\cdot]$, and we assume fixed x.

Since U_u is unobserved and cannot be evaluated prior to generation, we instead require a proxy utility function.

3.1 RUMS-Utility

In order to construct a surrogate utility function, we frame memory selection as an information acquisition problem: given a large set of user memories, the goal is to select items that most improve responses to an input. Under this view, we propose an information gain–based utility function over memory subsets, serving as an entropy-based surrogate for $\mathrm{U}_u(y)$. Formally,

$$\mathrm{U}_{RUMS}(S,u) = H\left(Y|x\right) - H\left(Y|x, m_{u,S}\right) \quad (1)$$

where $H(\cdot)$ denotes Shannon entropy under the model distribution θ (defined in Section B.1). Given that only the second term in Equation (1) is dependent on the subset selected S, we can simplify the objective and state that RUMS chooses the optimal subsets S^* such as,

$$S^*(x,u) = \arg\min_{S \subseteq \mathcal{M}} H(Y \mid x, m_{u,S})$$

The intuition is that a memory subset S is valuable to the extent it reduces uncertainty in the model's predictions. For instance, for "where should I eat lunch?", without user information, the model spreads probability across countless restaurants, yielding high entropy. However, adding even a single relevant fact, such as "Location: Austin, TX" or "Dietary preference: Vegan", sharply concentrates probability on relevant options. Therefore, the utility of the memory subset S is measured directly by reduction in the entropy.

3.2 Theoretical Foundation

We formalize conditions under which reducing predictive uncertainty leads to higher expected user utility. While the theoretical foundation relies on a latent-variable abstraction, we validate empirically (Section 4) that minimizing predictive entropy aligns with improved user-perceived utility.

Assumption 1 (Existence of an Optimal Latent Variable)**.** For each user u and query x, there exists a latent random variable $Z_u \sim p(Z_u \mid x)$ representing the user-specific latent target that determines what constitutes an optimal response. User utility depends on alignment between the generated response and this latent target.

Formally, where ℓ is a strictly proper scoring rule, we assume

$$\mathbb{E}\big[\mathrm{U}_u(y)\big] = -\mathbb{E}_{z_u \sim p(Z_u|x)}\mathbb{E}[\ell(y, z_u)].$$

Conceptually, Z_u is an abstract latent variable capturing the maximal predictive information needed for an optimal response.[2]

Assumption 2 (Approximate Sufficiency of Latent Variable)**.** Conditioned on the query x and latent target Z_u, the memory values $m_{u,S}$ do not provide substantial additional information about output Y:

$$p_\theta(Y \mid x, Z_u, m_{u,S}) \approx p_\theta(Y \mid x, Z_u).$$

This assumption treats Z_u as an information bottleneck, where memory primarily informs responses by clarifying latent intent. However, this may miss cases where raw details (e.g., formatting) provide utility not captured by the latent variable. We therefore adopt approximate, rather than exact,

[2]While this links predictive distributions to decision quality (Gneiting and Raftery, 2007), it assumes static user preferences. In practice, preferences may shift and be shaped by model outputs, so a fixed target can miss co-evolution in long-term interactions.

sufficiency to account for real-world memory containing extraneous information beyond latent preferences (Tishby et al., 1999; Alemi et al., 2017).

Theorem 3.1 (Optimality of RUMS)**.** *Under Assumptions 1 and 2, for any two memory subsets* $S, S' \subseteq \mathcal{M}$*,*

$$H_\theta(Y \mid x, m_{u,S}) \leq H_\theta(Y \mid x, m_{u,S'}) \quad \Rightarrow$$
$$\mathbb{E}_{y\sim p_\theta(\cdot|x,m_{u,S})}[U_u(y)] \geq \mathbb{E}_{y\sim p_\theta(\cdot|x,m_{u,S'})}[U_u(y)].$$

Theorem 3.1 formalizes the intuition that reducing predictive entropy improves expected utility, with approximation arising from the latent-variable abstraction in Assumption 2.

Proof Sketch. Predictive entropy decomposes into uncertainty about the latent target Z_u and residual uncertainty about y. Under Assumption 2, selected memory mainly reduces uncertainty about Z_u, and because strictly proper scoring rules (Assumption 1) reward matching the conditional distribution, this reduction directly improves expected utility. Full derivation is provided in Appendix C.

Interpretation and Limitations Theorem 3.1 provides a principled justification for selecting memory subsets that reduce predictive entropy. However, *this surrogate is meaningful only if* $p_\theta(y \mid x, m_{u,S})$ *aligns with human judgments*, that is, higher-probability responses under relevant memory are more useful to humans. While the theorem provides a theoretical basis, *empirical evaluation (Section 4) is necessary* to verify that entropy reductions translate into improved user-perceived utility.

3.3 Practical Estimation

Although RUMS-Utility is theoretically grounded, directly computing it for LLM-generated sequences is intractable, as the space of possible outputs y grows exponentially with the sequence length T. Therefore, we employ practical approximations via token-level decomposition and sampling.

Token-Level Decomposition To improve tractability, we apply the chain rule of entropy to decompose sequence-level entropy into token-level conditional entropies:

$$H(Y \mid x, m_{u,S}) = \sum_{t=1}^{T} \mathbb{E}H\big(p_\theta(\cdot \mid y_{<t}, x, m_{u,S})\big),$$

where the expectation is taken over $y \sim p_\theta(\cdot \mid x, m_{u,S})$. This reduces to computing token-level entropies, which is computationally tractable for long sequences (see proof in Section D).

Monte Carlo Approximation The outer expectation over token prefixes remains intractable, so we approximate it using Monte Carlo sampling. We generate $\{y^i\}_{i=1}^N \sim p_\theta(\cdot \mid x, m_{u,S})$ and compute:

$$\hat{H}(Y \mid x, m_{u,S}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i} H_\theta(y^i_{<t}, x, m_{u,S}),$$

where T_i is the length of sampled sequence y^i, and $H_\theta(y_{<t}, x, m_{u,S})$ denotes the token-level conditional entropy under p_θ given prefix $y_{<t}$ and context $(x, m_{u,S})$. We normalize by sequence length to estimate per-token entropy, which allows comparison across variable-length generations.

This approximation converges to the true per-token sequence entropy as $N \to \infty$ under i.i.d. sampling from $p_\theta(\cdot \mid x, m_{u,S})$ with finite expectation. In practice, however, factors such as constrained decoding or approximate sampling distributions can introduce bias. Despite these limitations, we find $N = 5$ samples provides a good balance between accuracy and efficiency. For more empirical analysis, see Section A.1.

The final estimated utility is then,

$$\widehat{U}_{\text{RUMS}}(S, u) = \hat{H}(Y \mid x) - \hat{H}(Y \mid x, m_{u,S}). \tag{2}$$

which serves as a practical surrogate for user utility and retains the theoretical connection to expected utility established in Theorem 3.1. We note that $\hat{H}(Y \mid x)$ and $\hat{H}(Y \mid x, m_{u,S})$ use the same Monte Carlo sampling budget.

Then, our final objective selects the memory subset that maximizes this utility:

$$S^*(x, u) = \arg\max_{S\subseteq M} \widehat{U}_{\text{RUMS}}(S, u). \tag{3}$$

3.4 RUMS-Models: Learning to Approximate Utility

While RUMS provides a principled approach to memory selection, exact entropy computation over all subsets is intractable at inference for large user memories. We therefore use RUMS offline to generate training data and learn a lightweight model that predicts selections at inference.

Training Data Generation Our training pipeline uses RUMS to generate (user query, optimal memory subset) pairs. For each query x, we pair it with a diverse set of users $\{u_1, \ldots, u_K\}$, each with memory items m_u. When profiles are large, one can first *optionally* use an LLM to do a rough filter to k candidate items. We then compute utility using Equation (2) for candidate subsets S, selecting the highest-utility subset $S^*(x, u) =$

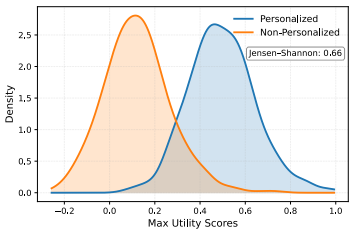


Figure 2: Distribution of maximum utility scores for personalized (PersonaFeedback) and non-personalized (FreebaseQA) inputs. Personalized inputs exhibit substantially higher utility scores, indicating stronger sensitivity to user memory.

$\arg\max_S \widehat{\mathrm{U}}(S, u)$ if it exceeds threshold τ, or the empty set otherwise. This yields K training examples $\{(x, u_k, S^*(x, u_k))\}_{k=1}^K$ per query. Pairing each query with diverse profiles encourages the model to learn genuinely informative item selection rather than spurious correlations. The threshold τ, tuned on validation data, enables automatic abstention when personalization would not help.

Model Architecture and Training We train RUMS-Models as lightweight classifiers that approximate utility computation at inference. For static schemas with d fixed item types, a multiclass classifier (RUMS-Multi) $f(x) \to \{0, 1\}^d$ outputs all selections in a single pass. For dynamic schemas where item types vary across users, a binary classifier (RUMS-Binary) $f(x, m_i) \to \{0, 1\}$ evaluates each item independently, handling arbitrary structures at the cost of n forward passes. Since RUMS-Utility is computed over subsets, RUMS-Binary could in principle be trained on biased data, but we show empirically this is not the case (Section A.4). The resulting models match the inference speed of similarity-based retrieval while preserving RUMS's principled selection quality, making them practical for large-scale deployment.

4 Results

In this section, we evaluate RUMS by defining three outputs: *RUMS-Utility*, the utility score from Equation (2); *RUMS-Subset*, the selected memory subset from Equation (3); and *RUMS-Models*, models trained on RUMS-generated data. We evaluate our approach along three dimensions:

- H1: RUMS-Utility provides a reliable signal of human preference for inputs that benefit from personalization from those that do not.
- H2: RUMS-Subset memory subsets more aligned with human preference than SOTA baselines and larger models. Furthermore, the utility signal is strong enough that models trained on them (RUMS-Models) also outperforms baselines and larger models.
- H3: RUMS-Models chosen subsets improve downstream response generation compared to SOTA baselines and larger models.

4.1 Experimental Setup

Across all experiments (H1–H3), we consider $n = 50$ static user memory items (e.g., *location*, *age*, *name*)[3], and assume that all values are given for each user (see Section F.3 for list of items). The task is to select up to 5 items that are most relevant for answering a given user input x.

Datasets. To test both instances that would and would not benefit from personalization, we use a diverse mixture of user inputs from the following:

- **PersonaFeedback** (Tao et al., 2025): A synthetic dataset of LLM-generated user inputs using randomly sampled user memory, ensuring each input benefits from personalization. We translated the original Chinese data using a translation model (Tiedemann and Thottingal, 2020). Example: *"Do you have any music for the trip?"*
- **FreebaseQA** (Jiang et al., 2019): A dataset of trivia-style questions and answers. Example: *"What type of snake is a sidewinder?"*
- **WildChat** (Zhao et al., 2024): Real-world ChatGPT interactions (English only, non-toxic, and request-based). We use only the first human-written input. Examples: *"im hungry"; "who is Mohamed Gamal Mohamed Maniieh"*

We treat PersonaFeedback as personalized inputs and FreebaseQA as non-personalized inputs by construction, together forming the "Synthetic" dataset. For WildChat, two annotators labeled each input as personalized or not (agreement required), yielding WildChat-Personal and WildChat-NonPersonal as the "Real-World" dataset.

User Profiles. We used GPT-4 (OpenAI et al., 2024) to generate n=100 user profiles with 50 user memory items (avoiding contradictions like "vegan" with "favorite food = steak") for training and testing. See Section F.3 for prompts, distributions of memory item values and generation details.

[3]We generated 100 candidate memory items, but items beyond the top 50 were increasingly obscure and less real-world attributes.

Dataset / Metric	RUMS-Utility	GPT4-Few	GPT4-Zero	Semantic Similarity
Synthetic (n=20) / Recall	94.4 [82.35, 100.00]	**100.0 [100.00, 100.00]**	33.3 [12.5, 55.56]	**100.0 [100.00, 100.00]**
Synthetic (n=20) / Specificity	**94.4 [84.85, 100.00]**	50.0 [0.00, 100.00]	50.0 [0.00, 100.00]	0.0 [0.00, 0.00]
Real-World (n=31) / Recall	**92.3 [75.00, 100.00]**	76.9 [50.00, 100.00]	53.9 [25.00, 80.00]	7.7 [0.00, 25.00]
Real-World (n=31) / Specificity	96.0 [85.71, 100.00]	**100.0 [100.00, 100.00]**	**100.0 [100.00, 100.00]**	**100.0 [100.00, 100.00]**

Table 1: Recall (accuracy on personalized inputs) and specificity (accuracy on non-personalized inputs) across datasets and methods. Each metric is given by the value and the 95% CIs via nonparametric bootstrap (2,000 resamples with replacement over test-set rows). Bolded values indicate the best performance in each category.

Dataset	Metric	RUMS-Subset	RUMS-Binary	RUMS-Multi	Semantic Sim.	GPT4-Zero	GPT4-Few
Synthetic	Precision	**0.61 [0.55, 0.66]**	0.60 [0.55, 0.64]	0.23 [0.19, 0.26]	0.09 [0.07, 0.11]	0.26 [0.21, 0.31]	0.53 [0.50, 0.56]
	Recall	0.70 [0.64, 0.75]	**0.89 [0.86, 0.92]**	0.39 [0.34, 0.44]	0.23 [0.18, 0.28]	0.42 [0.38, 0.46]	0.65 [0.61, 0.69]
	F1	0.70 [0.65, 0.75]	**0.78 [0.74, 0.82]**	0.31 [0.26, 0.36]	0.14 [0.11, 0.17]	0.35 [0.30, 0.39]	0.64 [0.60, 0.68]
Real World	Precision	**0.48 [0.42, 0.55]**	0.32 [0.29, 0.36]	0.30 [0.26, 0.34]	0.17 [0.15, 0.20]	0.39 [0.36, 0.43]	0.42 [0.38, 0.46]
	Recall	0.36 [0.30, 0.41]	**0.60 [0.55, 0.64]**	0.38 [0.34, 0.43]	0.27 [0.24, 0.30]	0.52 [0.48, 0.56]	0.39 [0.34, 0.44]
	F1	**0.53 [0.47, 0.59]**	0.45 [0.41, 0.49]	0.35 [0.31, 0.40]	0.20 [0.18, 0.23]	0.43 [0.39, 0.47]	0.39 [0.34, 0.44]

Table 2: Macro-averaged precision, recall, and F1 scores of memory items selected compared to human gold label annotation across datasets and methods. Each metric is computed per-query and then averaged independently across queries; at the per-query level, therefore the mean of per-query F1 scores is not equal to the harmonic mean of mean precision and mean recall. Each metric is given by the value and the 95% CIs via nonparametric bootstrap (2,000 resamples with replacement over test-set rows)

RUMS-Model Training To construct training data, we use Llama-3.1-8B (AI, 2024b) to compute entropies (see analysis with other model families in Section A.10). Following Section 3.4, GPT-4 first filters the 50 candidate memory items to at most 10 per query before utility computation (see robustness of filtering in Section A.11). Utility scores are estimated with $n = 5$ Monte Carlo samples. We then apply a threshold $\tau = 0.29$, tuned on a held-out validation set, to select between the highest-utility subset and the empty set (no personalization) (see extra threshold analysis in Section A.5). We use 2.5K samples from PersonaFeedback and 2K from FreebaseQA for training, with held-out splits to avoid overlap and ensure in-domain evaluation. We further evaluate generalization on WildChat, a different-distribution dataset.

We train two RUMS-Model variants: The first, RUMS-Binary, is a binary classifier trained on individual (query, memory value) pairs. Because each memory value is treated separately and negative labels dominate, this model tends to under-select, making it better suited for noise-sensitive settings. The second, RUMS-Multi, is a multiclass classifier trained on denser labels constructed by merging the selected subsets from the $u = 5$ user profiles per query. This means, all memory items chosen by any of the profiles were included. As a result, the model tends to over-select, which is advantageous for more robust models that can filter noise. Both are based on DeBERTa-v3-large (He et al., 2021). We choose DeBERTa-v3-large due to its strong encoding capability relative to its size, and, more importantly, because we want to demonstrate that the utility signal produced by RUMS is strong enough to train a model of comparable size to the other baselines we compare against. Full details are in Section F.5.

Baselines Current approaches to memory selection primarily rely on two strategies: semantic similarity-based retrieval and LLM prompting. To comprehensively evaluate against existing methods, we employ baselines representing both approaches: (1) **Semantic Similarity**: We compute cosine similarity between input and memory item embeddings using Sentence Transformer (Wang et al., 2020), representing the dominant paradigm of surface-level similarity matching in current re-

trieval systems, (2) **GPT-4 Prompting**: We use direct prompting of GPT-4 (*zero-shot* and *few-shot* - up to 4 in-context examples) as a representative LLM-based approach and upper bound on model capability, given its 400× larger size compared to RUMS-Models. For H3, which evaluates downstream generation quality, we thus additionally include established retrieval methods and naive baselines to ensure comprehensive coverage: (3) **SOTA Retrieval Methods**: We evaluate *BM25* (Robertson and Zaragoza, 2009; Lù, 2024) and *ReContriever* (Lei et al., 2023) as representative SOTA retrieval approaches for memory selection, (4) **Naive Baselines**: We include *None* (no items), *All* (50 items), and *Random* (5 items) to establish performance bounds and test robustness to different selection strategies.

4.2 H1: Does RUMS-Utility Distinguish Personalized from Non-Personalized Inputs?

We hypothesize that entropy reduction under user memory reflects importance, in that inputs that do not benefit from personalization should have low maximum utility scores (i.e., the maximum over all memory subsets for a user), while personalized inputs should have higher values. To test this, we compare maximum utility scores between non-personalized (FreebaseQA) and personalized (PersonaFeedback) inputs. We sample 500 inputs per dataset, compute the maximum utility via Equation (2), and compare the resulting distributions using divergence metrics and hypothesis tests.

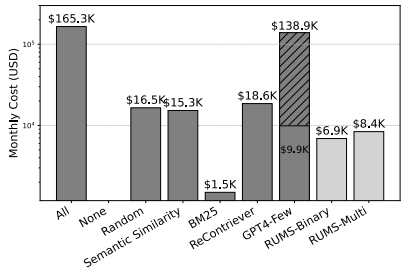
RUMS Utility scores separate personalized from non-personalized inputs. Figure 2 shows that personalized inputs have higher utility scores (mean 0.50, std 0.12) than non-personalized ones (mean 0.12, std 0.09). The Jensen–Shannon divergence is 0.66, and a two-sample Kolmogorov–Smirnov test yields 0.82 ($p < .001$), strongly rejecting the null hypothesis of identical distributions. These results show that maximum utility scores reliably separate personalized from non-personalized inputs. For robustness analysis and hyperparameter sensitivity, see Section A.1.

4.2.1 H1': Do Utility Scores Align with Human Judgments of Personalization?

Next, we test if personalization decisions made by RUMS-Utility thresholding aligns with human judgment. We randomly sampled $n = 25$ inputs each from PersonaFeedback, FreebaseQA, WildChat-NonPersonal, and WildChat-Personal. Four human annotators labeled if each input benefits from personalization, which inter-rater agreement of Cohen's $\kappa = 0.59$; we retained only inputs with unanimous agreement ($n = 64$).

Method	Synthetic	Real World
All	50.00	50.00
None	0.00	0.00
Random	5.00	5.00
Semantic Similarity	4.63	4.79
BM25	0.60	0.35
ReContriever	5.00	5.00
GPT4-Few	2.80	3.00
RUMS-Binary	1.58	2.22

(a) Average number of memory items by method and dataset type.



(b) Estimated additional average monthly cost by method. Diagonal line indicate added cost of LLM calls for memory selection.

Figure 3: Cost Analysis

To obtain a robust signal from RUMS, we use $n = 10$ Monte Carlo samples and $u = 10$ randomly selected user profiles, averaging maximum utility across profiles to define RUMS-Utility. We tune a threshold on a 20% validation split. We compare RUMS-Utility against GPT-4 (prompted to assess personalization) and semantic similarity baselines, with thresholds tuned on the same validation set. Performance is measured by recall (identifying personalized inputs) and specificity (identifying non-personalized inputs). If both the model and humans select no items, the prediction is counted correct for both metrics.

RUMS-Utility correlates more strongly with human judgments. As shown in Table 1, RUMS-Utility outperforms GPT-4 and semantic similarity baselines. Compared to GPT-4 FewShot, it achieves +7% recall and +15% specificity overall. Gains are strongest on synthetic data, with competitive results on real-world. Other baselines show imbalanced behavior (e.g., GPT-4 ZeroShot often avoids personalization, giving high specificity but low recall). In contrast, RUMS-Utility performs well on both metrics, reflecting their ability to cap-

	Synthetic						Real World					
	LLaMA 70B			GPT-4			LLaMA 70B			GPT-4		
	Utility	Binary	Multi	Utility	Binary	Multi	Utility	Binary	Multi	Utility	Binary	Multi
None	**0.11**	**0.25**	**-0.13**	**0.20**	**0.26**	**0.36**	-0.05	0.05	0.03	0.10	**0.20**	**0.17**
All	**0.18**	**0.16**	**-0.19**	**0.20**	**-0.16**	0.02	0.03	-0.04	-0.10	**-0.14**	**-0.14**	-0.06
Random	**0.44**	**0.34**	**0.13**	0.10	0.06	**0.19**	**0.25**	**0.11**	**0.13**	0.05	0.10	0.04
Semantic Sim.	**0.43**	**0.27**	0.09	**-0.12**	-0.03	0.05	**0.16**	-0.03	0.00	-0.06	-0.02	0.00
BM25	0.03	**0.19**	**-0.13**	**0.18**	**0.20**	**0.32**	0.06	0.01	0.05	0.10	**0.18**	**0.13**
ReContriever	**0.31**	**0.27**	-0.03	-0.05	-0.04	0.03	**0.22**	0.06	0.01	0.07	-0.05	**0.13**
GPT4-Few	0.07	**0.18**	**-0.16**	-0.07	-0.09	-0.01	0.01	0.00	-0.08	-0.08	0.04	**0.11**

Table 3: Average win rate margin of our method against baselines on Synthetic (left) and Real World (right) datasets. Cells in green mark cases where our method outperforms the baseline, and cells in red mark cases where the baseline does better. Values with magnitude above 0.1 are **bolded** to highlight strong differences. Results are reported for two response models (LLaMA 70B Instruct and GPT-4) and two RUMS variants: RUMS-Binary (fewer items) and RUMS-Multi (more items).

ture each item's true effect on generation.

4.3 H2: Does RUMS-Subset Match Human Judgment?

Building on H1, we next evaluate whether RUMS identifies memory items which support personalization. We sample $n = 15$ inputs from each of PersonaFeedback, FreebaseQA, WildChat-NonPersonal, and WildChat-Personal ($n = 60$ total). Four annotators select any subset of the 50 candidate items, yielding 41% majority agreement ($k \geq 3$). As expected for this subjective task, agreement is moderate. We define an item as selected if a majority agree, averaging 4.19 items per query among cases with at least one selection. Human-selected subsets serve as gold labels, since they have been validated to yield better downstream responses (see Section A.2).

We evaluate (i) direct utility-based selection RUMS-Subset (up to 5 items) and (ii) RUMS-Models, the trained models from Section 4. We compare both against GPT-4 (zero- and few-shot prompts selecting up to five items) and semantic similarity (top-5 items with cosine similarity > 0). Performance is measured using precision, recall, and F1 against human annotations.

RUMS-Subset aligns more closely with human annotations. As shown in Table 2, both RUMS methods improve over baselines by 20% precision, 31% recall, and 9% F1 on average across synthetic and real-world data. As expected semantic similarity performs worst, reflecting reliance on surface overlap. GPT-4 achieves high recall but low precision, indicating over-selection. Overall, raw utility scores in RUMS-Subset provide a strong signal for identifying useful memory items, as RUMS-Models attains similar F1, suggesting that small models trained on modest data can effectively learn selection from utility scores. For the approximation gap between RUMS-Utility and RUMS-Models, see Section A.3; qualitative examples and analysis are in Section A.6 Section A.7.

4.4 H3: Do Items Selected by RUMS-Models Improve Downstream Generation?

Having established that RUMS makes effective decisions about when and what to personalize (H1-H2), we now demonstrate that these selections translate to tangible improvements in response quality. This tests our core claim that principled selection leads to better personalization.

Experimental Setup. We evaluate RUMS on both synthetic (PersonaFeedback and FreebaseQA, 100 samples each) and real-world datasets (WildChat, 50 personalized and 50 non-personalized) to assess generalization. Responses are generated using two models: GPT-4, a high-performing model with robust capabilities, and LLaMA 70B Instruct (AI, 2024a), a smaller open-source model that is relatively more susceptible to noise. We evaluate response quality using RUMS-Utility and RUMS-Models against all baselines described in Section 4.1. Performance is measured by GPT-4 win rate margin, which is the our method's win rate minus the baseline's win rate, using chain-of-thought prompting and confidence thresholding to calculate win rate (Appendix F.7).

RUMS improves response quality across settings. Table 3 shows that RUMS consistently outperforms baselines on both synthetic and real-world datasets. With LLaMA 70B, RUMS-Binary achieves the strongest gains, outperforming all baselines by at least 15% on synthetic data, including an 18% gain over GPT4-FewShot, a model

$400\times$ its size. It also has modest improvement on real world data as well. Gains are driven largely by non-personalized inputs, where lower-capability models struggle to ignore noisy memory. In contrast, with GPT-4 generation, RUMS-Multi performs best, improving over GPT4-FewShot by 11% and BM25/ReContriever by 13% on real-world data. This suggests that stronger models benefit from additional context while filtering noise. Although *All* performs reasonably with only 50 items, prior work shows performance degrades as memory scales (Rajeev et al., 2025). Human evaluation (Section A.9) supports these findings and validates Theorem 3.1 that entropy-reducing memory subsets improve response quality.

RUMS-Models performs similarly to RUMS-Utility. Consistent with H2, downstream results for RUMS-Utility most closely resemble RUMS-Binary. Overall, RUMS-Utility performs strongly, achieving over 10% higher win rate margin than nearly all baselines with LLaMA 70B and all low-level baselines with GPT-4.

RUMS-Models reduces cost compared to other under-performing methods. Because personalization is applied to every input, efficiency is critical. We measure the number of selected memory items and estimate monthly inference cost as extra tokens from selected items under a pricing model of \$0.01 per 1K tokens for one million daily queries (current GPT-4 rates). Figure 3a shows that RUMS-Models select substantially fewer memory items than most baselines, translating to significant cost savings (Figure 3b): RUMS-Binary costs \$6.9K/month and RUMS-Multi costs \$8.4K/month, compared to \$165.3K for All and \$18.6K for the competitive ReContriever. Both RUMS variants are cheaper than GPT4-FewShot, even excluding the added cost of GPT-4-based memory selection. Overall, RUMS balances utility and efficiency, enabling scalable personalization.

5 Conclusion

Our results support the central hypothesis of this work that measuring how a memory item affects a model's response distribution is a stronger signal for personalization than measuring its surface-level similarity to the input query. Across H1–H3, RUMS's entropy-based utility function consistently separated personalized from non-personalized inputs (Section 4.2), aligned more closely with human item-selection judgments than semantic similarity or prompted GPT-4 (Section 4.3), and, when distilled into lightweight RUMS-Models, produced downstream responses that were preferred over baselines using models up to $400\times$ larger, all while reducing the per-query cost of personalization by up to 95% (Section 4.4). Together, these findings suggest that the information-theoretic view of memory selection is not just theoretically motivated (Theorem 3.1) but practically actionable.

However, we recognize that several directions remain open for future work. First, we deliberately chose token-level Monte Carlo entropy over semantic entropy or reward-model-based utility, as it avoids the auxiliary models and hyperparameters (e.g., clustering thresholds) that those alternatives introduce, while remaining cheap enough for deployment. However, standard token entropy can conflate lexical diversity with genuine predictive uncertainty, and therefore we see semantic entropy and embedding-variance-based estimators as promising directions for future ablation work to test the robustness of our core signal. Second, our formulation also treats memory as static and queries as single-turn to isolate which items are informative independent of when they should be retrieved. Future work could extend RUMS-Utility to multi-turn dialogue, where preferences shift and memory is updated over time. Finally, we evaluate RUMS on a fixed 50-item structured schema, and while the utility function is architecture- and schema-agnostic in principle, scaling to fully unstructured memory, including free-form text, heterogeneous attributes, or long behavioral histories, may require hierarchical selection mechanisms or encoder-level adaptations we have not tested here.

Beyond the specific gains demonstrated here, RUMS points toward a more general principle for inference-time personalization: memory selection should be grounded in a model's actual response behavior rather than proxy signals like surface similarity. This has practical implications for real-world deployment, where memory stores are large, latency budgets are tight, and irrelevant context can both degrade response quality and needlessly expose sensitive user information. By showing that a lightweight, distillable utility signal can match, and in many cases exceed, the selection quality of models orders of magnitude larger, RUMS offers a path toward personalization systems that are simultaneously more principled and more efficient, which we hope will encourage further work on utility-driven approaches to memory management in LLMs.

6 Limitations

This work studies the hypothesis that downstream response utility is a stronger signal for memory selection than surface-level similarity between inputs and memory items. To isolate this effect, we evaluate RUMS in a controlled setting, which necessarily introduces several limitations.

First, we assume static user memory and focus on single-turn queries, whereas real personalization is inherently multi-turn and involves evolving preferences and dynamically updated memory. We adopt a single-turn formulation to isolate which memory items are informative, decoupling this from the separate problem of when to retrieve them. Extending RUMS to dialogue settings is straightforward in principle by applying it at each turn using the full conversation context.

Second, all experiments are conducted in English. Although RUMS is based on an information-theoretic objective that is language-agnostic in principle, personalization behavior and model calibration may vary across languages and cultural settings, making multilingual evaluation an important direction for future work.

Third, the guarantees in Theorem 3.1 rely on the assumption that the model's predictive distribution is aligned with human preferences. This assumption may be weaker for miscalibrated or heavily instruction-tuned models, although our empirical results suggest that modern LLMs are sufficiently well-aligned for the signal to remain useful in practice.

Fourth, we note that we used a different model to filter the memory items (Llama) than to generate the responses (GPT). Although performing different functions in our pipeline, we do want to highlight that using different models for selection versus generation introduces an assumption of transferability that merits further study.

Fifth, given our results we suggest that RUMS-Utility is better suited to narrowing, constraint-driven personalization than to personalization that should expand the response space, and its argmax selection procedure can systematically favor narrowing memories over more relevant broadening ones.

Finally, we use a fixed 50-item structured memory schema, which is simpler than real-world memory systems that often include free-form text, heterogeneous attributes, or long behavioral histories. While RUMS is architecture- and schema-agnostic in principle, scaling to fully unstructured memory may require hierarchical selection mechanisms or encoder-level adaptations.

7 Ethical considerations

We believe this research presents minimal ethical concerns. However, like other methods that steer or guide model generations, our approach could potentially introduce unintentional biases, reinforce stereotypes, or produce harmful outputs. We therefore recommend that practitioners implement comprehensive evaluation protocols before deployment, including stress testing, automated safety evaluations, human assessments, and ongoing monitoring systems. Additionally, since our method incorporates user information during inference, it introduces potential privacy and security vulnerabilities that require careful consideration in production environments.

Acknowledgments

We thank Philippe Laban, Tobias Schnabel, Kiran Tomlinson, and Tarek Naous for their valuable discussions and insightful feedback throughout this project. This work was done while Jillian Fisher was on an internship with Microsoft Research.

References

Meta AI. 2024a. Llama 3.1 70b instruct. Accessed: 2025-09-16.

Meta AI. 2024b. Llama-3.1-8b. Accessed: 2025-09-16.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. *International Conference on Learning Representations*.

Antonio Almudévar, Jose Miguel Hernandez-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. 2025. Aligning multimodal representations through an information bottleneck. In *Proceedings of the 2025 International Conference on Machine Learning (ICML 2025), Poster Session*.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. 2025. Lore: Personalizing LLMs via low-rank reward modeling. *Preprint*, arXiv:2504.14439.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Jennifer C. Lai, and Robert L. Mercer. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, volume 80 of *Proceedings of Machine Learning Research*, pages 883–892. PMLR.

Jiarui Chen. 2025. Memory assisted LLM for personalized recommendation system. *Preprint*, arXiv:2505.03824.

Ian Covert, Wei Qiu, Mingyu Lu, Nayoon Kim, Nathan White, and Su-In Lee. 2023. Learning to maximize mutual information for dynamic feature selection. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

Soham Gadgil, Ian Covert, and Su-In Lee. 2024. Estimating conditional mutual information for dynamic feature selection. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Tilmann Gneiting and Adrian E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Lovisa Hagström, Youna Kim, Haeun Yu, Sang goo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2025. Cub: Benchmarking context utilisation techniques for language models. *Preprint*, arXiv:2505.16518.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. 2024. Bayesian preference elicitation with language models. *Preprint*, arXiv:2403.05534.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L. Schrum, and Anca Dragan. 2025. Context steering: Controllable personalization at inference time. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Hailey Joren, Chirag Nagpal, Katherine A Heller, and Berk Ustun. 2023. Participatory personalization in classification. In *Advances in Neural Information Processing Systems*, volume 36, pages 14113–14133. Curran Associates, Inc.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating LLMs on user rating prediction. *Preprint*, arXiv:2305.06474.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations: Democratizing large language model alignment. In *NeurIPS 2023 Poster*.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of personalized large language models: Progress and future directions. *Preprint*, arXiv:2502.11528.

Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a good recommender? a preliminary study. *Preprint*, arXiv:2304.10149.

Xing Han Lù. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. 2019. Eddi: Efficient dynamic discovery of high-value information with partial VAE. *ICLR*.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. SeCom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Prapti Trivedi, Vikas Yadav, Oluwanifemi Bamgbose, Sathwik Tejaswi Madhusudhan, James Zou, and Nazneen Rajani. 2025. Cats confuse reasoning LLM: Query agnostic adversarial triggers for reasoning models. In *Second Conference on Language Modeling*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *Preprint*, arXiv:2310.20081.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation. SIGIR '24, page 752–762, New York, NY, USA. Association for Computing Machinery.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024b. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6459–6475. Association for Computational Linguistics.

Meiling Tao, Chenghao Zhu, Dongyi Ding, Tiannan Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. PersonaFeedback: A large-scale human-annotated benchmark for personalization. *Preprint*, arXiv:2506.12915.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. *arXiv preprint physics/0004057*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. 2022. Prompt-based meta-learning for few-shot text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1357, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations (ICLR 2024), Spotlight Poster*.

Appendix Contents

A. Additional Experimentation
Section A.1 Robustness Analysis of Utility Scores
Section A.2 Validating Human-Chosen Memory Subset Performance on Downstream Response
Section A.3 Analysis of Approximation Gap Between RUMS-Utility and RUMS-Models
Section A.4 Quantifying and Validating Approximation of RUMS-Binary using RUMS-Utility
Section A.5 RUMS-Utility Threshold Analysis
Section A.6 Examples of Selected Memory Items by RUMS
Section A.7 H2 Qualitative Analysis of Failure
Section A.8 Extra Comparison of Qwen 3-Embeddings for H1/H2
Section A.9 H3 Results using Human Judgment
Section A.10 Analysis of Selected Entropy Calculation Model
Section A.11 Analysis of Optional Filter Step
Section A.12 Cost Analysis
Section A.13 Time Analysis
Section A.14 Full Results of H3

B. Formulas

C. Optimality of RUMS Proof

D. Sequence Entropy Equivalence Proof

E. RUMS-Utility Algorithm

F. Experimental Details
Section F.1 Experimental Setup
Section F.2 Datasets
Section F.3 Profile Generation
Section F.4 Baselines
Section F.5 RUMS-Utility and RUMS-Model Details
Section F.6 Comparing Size of Methods
Section F.7 Evaluation Metric and Analysis

G. Other Details
Section G.1 Software
Section G.2 Hardware

H. Information About Use of AI Assistants
Section H Artifact Terms of Use

A Additional Experimentation

In this section, we present additional experiments to supplement our main results.

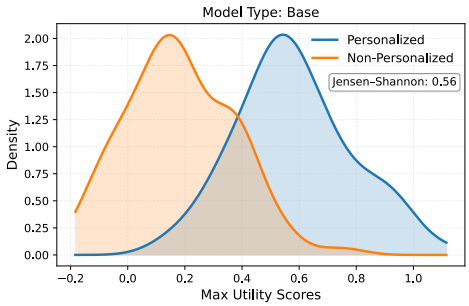
A.1 Robustness Analysis of Utility Scores

In H1, we hypothesized that RUMS-Utility, the maximum utility scores, can reliably distinguish between user inputs that would benefit from personalization and those that would not. Here, we further investigate how hyperparameters in the RUMS-Utility algorithm affect this reliability and provide justification for these effects. Specifically, we examine the influence of model type, model size, model family, decoding method type, maximum number of tokens generated, number of Monte Carlo samples, and number of user profiles. For each analysis, all other variables are held constant using the LLaMa base model, 8B size, sampling decoding, $T = 20$ tokens, $n = 1$ sample, and averaged over $n = 3$ user profiles.

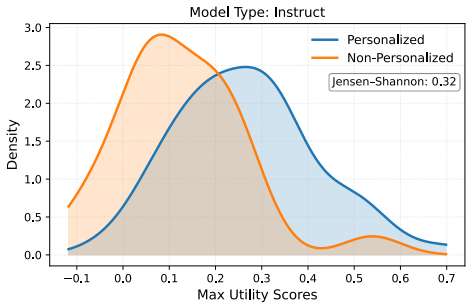
Hyperparameter: Model Type We first compare the LLaMa-8B base model to an LLaMa-8B instruction-tuned model. As shown in Figure 4, the base model produces more separated distributions of utility scores than the instruct model, with a Jensen-Shannon divergence score of 0.56 compared to 0.32 for the instruct model. This is consistent with prior research showing that instruction-tuning tends to pre-sharpen output distributions (Sorensen et al., 2024). Based on this observation, we chose to use the base model for our experiments.

Hyperparameter: Model Size Next, we examined the effect of model size on the entropy calculation. As shown in Figure 5, the distribution of utility scores changes when using LLaMA-1B compared to LLaMA-8B. While the impact is less pronounced than the effect of model type, we observe an increase of 0.11 in the Jensen-Shannon divergence score when using the 8B model. We hypothesize that larger models are better able to capture subtle and complex patterns in their response distributions. Given that the improvement is relatively modest, we opted to use the 8B model to balance performance with computational efficiency and avoid the added cost of larger models.

Hyperparameter: Model Family We next examine the effect of model family, comparing Llama-8B (AI, 2024b), Mistral-7B (Jiang et al., 2023), and Gemma-9B (Gemma Team, 2024). As shown in Figure 6, we find that Mistral model performs similar to Llama (0.41 vs. 0.56), where the Gemma model performs badly (0.17). This could indicate that some model families better capture human preference than others. Based on this observation, we use Llama-8B in our experiments.



(a) Model Type: Base

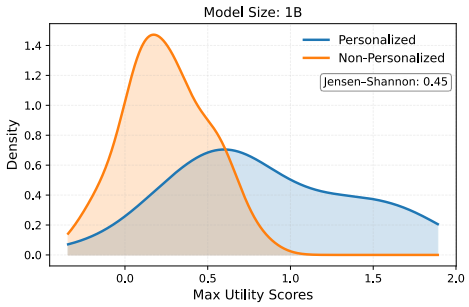


(b) Model Type: Instruct

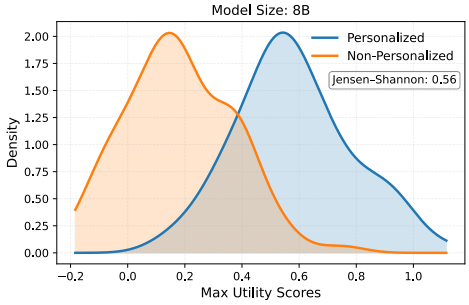
Figure 4: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset differing by type of model used for entropy calculation.

Hyperparameter: Decoding Type We next examine the effect of decoding type on the sequential average token entropy, comparing greedy and sampling decoding. As shown in Figure 7, sampling decoding produces more diverse distributions than greedy decoding. This is expected, as sampling introduces stochasticity into token selection, allowing the model to explore multiple plausible continuations, whereas greedy decoding always selects the highest-probability token at each step, leading to more concentrated and less variable distributions. Based on this observation, we use sampling decoding in our experiments.

Hyperparameter: Number of Tokens As noted in the main text, we limit the number of tokens used in the calculation of Equation (2) to avoid bias introduced by longer generations, since longer outputs tend to have lower entropy. To study this, we randomly sampled $n = 10{,}000$ responses from the OpenAssistant dataset (Köpf et al., 2023), using only the first human input in each conversation that was labeled as English and non-toxic. We used this

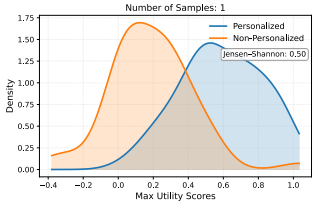


(a) Model Size: 1B

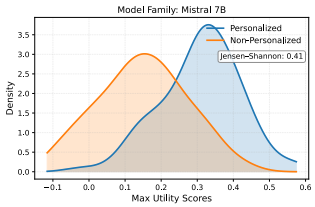


(b) Model Size: 8B

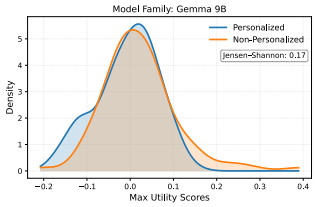
Figure 5: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset differing by size of model used for entropy calculation.



(a) Model Family: Llama



(b) Model Family: Mistral



(c) Model Family: Gemma

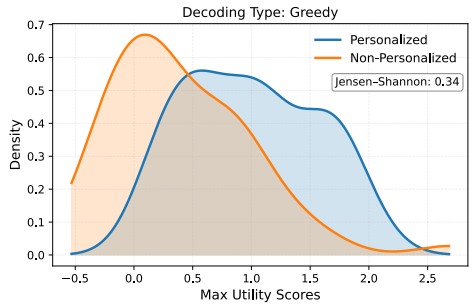
Figure 6: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset calculated using models of different families.

dataset since it is a real world dataset that could provide a wide diversity in types of user-LLM interactions. We then computed the average token-level entropy by token index, capped at 100 tokens. As shown in Figure 8, entropy begins to decrease significantly around $T = 20$ tokens, indicating that longer generations introduce a bias toward lower entropy.

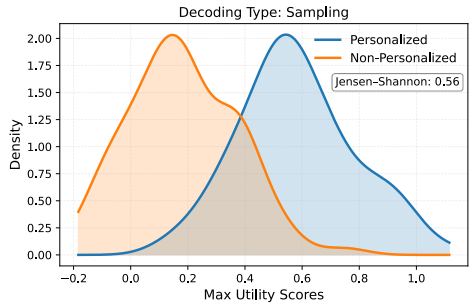
Further analysis in Figure 9 compares the distribution of maximum utility scores when using 20 versus 100 maximum tokens for entropy calculations. Using 100 tokens reduces the Jensen-Shannon divergence by 0.18 compared to using only 20 tokens. Based on these results, we limit the calculation to the first 20 tokens in our experiments.

Hyperparameter: Number of Samples Next, we examine the effect of the number of Monte Carlo samples used in Equation (2) on the robustness of the utility score. As shown in Figure 10, increasing the number of samples improves the separation between personalized and non-personalized inputs, which aligns with expectations since the estimate should converge toward the true parameter as the sample size increases if the samples are i.i.d. However, this comes with a tradeoff: more samples require additional computation, longer runtime, and higher resource costs. Balancing these considerations, we chose $n = 5$ samples for our experiments.

Hyperparameter: Number of User Profiles Lastly, we examine how the robustness of the average maximum utility score changes as we increase the number of (query,user profile) pairs included in the average. Consistent with estimation theory, we find that the difference between the personalized and non-personalized distributions grows as the sample size increases. Notably, there is a substantial jump of 0.22 in Jensen–Shannon divergence when increasing from 1 to 5 users, but only a marginal increase of 0.01 when going from 5 to 10 users. Based on this, we choose $u = 5$ users for constructing our training data to balance performance with cost of inference.



(a) Decoding Type: Greedy



(b) Decoding Type: Sampling

Figure 7: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset differing by decoding type.

A.2 Validating Human-Chosen Memory Subset Performance on Downstream Response

In H2, we analyze the correlation between the memory subsets selected by RUMS-Utility, RUMS-Model and the baselines compared to human-selected subsets. Human labels are treated as the gold standard, as we hypothesize that they would lead to better downstream responses. To evaluate this, we calculate the generation win rate margin of responses produced using human-selected features compared to those produced by RUMS-Model and the baselines. We use the same evaluation technique outlined in Section F. As shown in Table 4, human-selected memory subsets yield better downstream responses than all baselines when using the less capable response model for generation (LLaMA 70B). With the closest performance being RUMS-Binary. When generating responses with a more capable model, GPT-4, we still see slight better performance using the human selected memory items, although much less than when using a less capable model. This shows that human chosen memory items should be used in H2 as a measure of a good subset of memory.

A.3 Analysis of Approximation Gap Between RUMS-Utility and RUMS-Models

A.3.1 Analysis of RUMS-Models Signal Learning

To better understand what signals the RUMS-Models are learning, we add an analysis which explores the correlation between memory items selected (included) and not selected (excluded) by these models, and the connection to the true RUMS-Utility score. Since we are interested in personalization, we use the PersonaFeedback test set (n=100).

In Figure 12 and Figure 13, we show the distribution of average utility score per sample. In Figure 14 and Figure 15, we show the distribution of difference of average utility score for included - excluded features per sample. We find the distribution of included and excluded utility scores are statistically different (p-value ≤ 0.03, Mann-Whitney test). Similarly, We find that the distribution of differences between the average utility score for items the RUMS-model included and the average utility score of memory items the RUMS-Model excluded is greater than 0 (p-value $< .001$, Wilcoxon one-sided test), meaning that, on average, we find that memory items the RUMS-Model include have a truly higher utility score than those it does not. This lends strong evidence that RUMS-Models correctly learn signals of the raw utility scores.

A.3.2 Comparing RUMS-Utility to RUMS-Models on H2

RUMSAs mentioned in the main paper, we opt to use RUMS-Models, rather than the raw RUMS-Utility, due to its high efficacy and large decrease in computational compute. This is a common approach when using BOED methods, due to algorithmic components which make it time intensive. Although our H2 results show that RUMS-Models have similar, but lower, efficacy compared to RUMS-Utility, we want to further analyze this approximation gap.

To do this, we expand the results for H2. First, we calculated a majority feature list from the RUMS-Utility chosen data for the samples in this section, by including all features that were in a majority (¿=5) samples using the RUMS utility scores directly over 5 diverse profiles. Then, we calculated the same metrics (precision, recall, and F1),

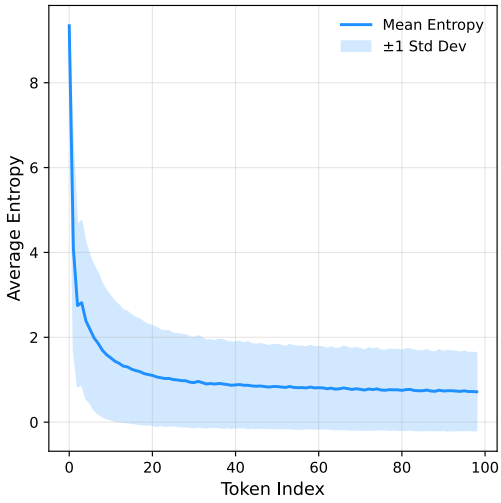


Figure 8: Average entropy by the index of the token being generated.

	LLaMA 70B	GPT-4
Method	Human Win Rate Margin	Human Win Rate Margin
None	**0.31**	**0.16**
All	**0.28**	**-0.26**
Random	**0.42**	0.06
Semantic Similarity	**0.38**	0.01
BM25	**0.18**	**0.26**
ReContriever	**0.52**	0.01
GPT4-Few	**0.27**	0.01
RUMS-Binary	0.01	-0.05
RUMS-Multi	**0.39**	0.02

Table 4: Win rate margin of responses preferred using human selected memory subsets compared to other methods. A number above 0 is the percentage of responses generated by Humans chosen subsets that were preferred.

of this new gold label group to feature list selected by RUMS-Binary and RUMS-Multi.

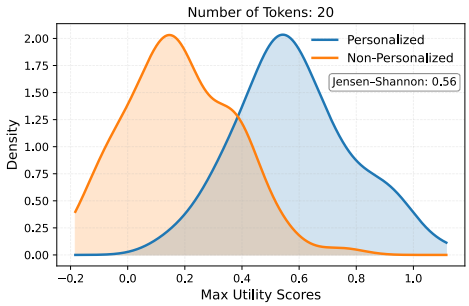
In general, we find that the models have high agreement on in distribution datasets (synthetic) but perform moderate on out of distribution dataset (real world). In line with our results for H2, we find that RUMS-Binary has very high correlation with RUMS-Utility with a high F1 of 0.78 on the synthetic dataset. Although RUMS-Binary only has a moderate F1 of .37, we note that most of this comes from a decrease in performance on non-personalized datasets and the over-selection of RUMS-Multi. This is highlighted by the higher recall scores, in general, and slightly lower precision (e.g. 0.9 vs. .73 for RUMS-Binary).

Dataset	**Metric**	RUMS-Binary	RUMS-Multi
Synthetic	Precision	0.73	0.33
	Recall	0.90	0.44
	F1	0.78	0.37
Real World	Precision	0.26	0.35
	Recall	0.42	0.35
	F1	0.32	0.34

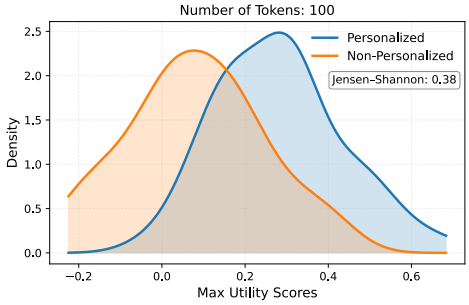
Table 5: Precision, recall, and F1 scores of memory items selected compared to human gold label annotation across datasets and methods.

A.4 Quantifying and Validating Approximation of RUMS-Binary using RUMS-Utility

The RUMS-Utility objective evaluates subsets of memory items jointly, capturing interactions between features. In contrast, RUMS-Binary predicts each item independently. Since conditional mutual information is not generally additive, this raises the question of how much combinatorial information is lost when using the item-wise approximation.

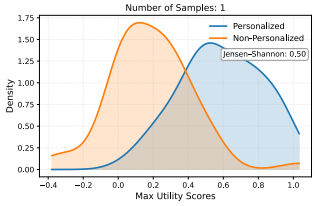


(a) Number of Tokens: 20

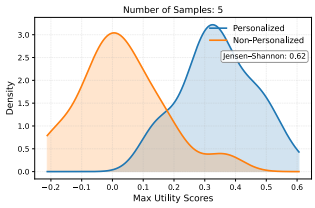


(b) Number of Tokens: 100

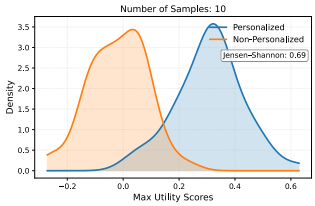
Figure 9: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset using different number of tokens.



(a) Number of Samples: 1



(b) Number of Samples: 5



(c) Number of Samples: 10

Figure 10: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset differing number of Monte Carlo samples average over.

To quantify this gap, we provide both a theoretical decomposition and empirical measurement.

Theoretical Decomposition Let $S = \{m_1, \ldots, m_k\}$ be a subset of memory items. The utility function used in RUMS is

$$U(S) = I(Y; M_S \mid x).$$

Using the chain rule of conditional mutual information,

$$I(Y; M_S|x) = \sum_{i=1}^{k} I(Y; m_i|x, m_{<i}),$$

where $m_{<i} = \{m_1, ..., m_{i=1}\}$.

Next, the independent approximation used by RUMS-Binary replaces conditional gains with marginal gains,

$$I(Y; m_i|x, m_{<i}) \approx I(Y; m_i|x).$$

This yields the additive approximation,

$$\tilde{U}(S) = \sum_{i=1}^{k} I(Y; m_i|x).$$

Therefore, the combinatorial interaction gap is expressed,

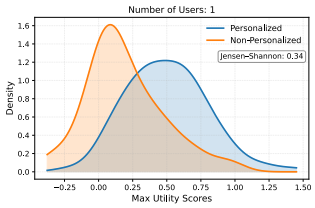
$$\Delta(S) = U(S) - \tilde{U}(S) = \sum_{i=1}^{k} \left(I(Y; m_i|x, m_{<i}) - I(Y; m_i|x)\right). \quad (4)$$

Therefore, Equation (4) shows that the total approximation error equals the sum of per-feature interaction terms.

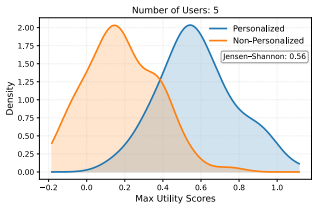
Empirical Measurement To better understand this approximation error in our experimentation, we estimate interaction effects using the utilities computed by RUMS-Utility for the test set of human annotated examples (n=60) using only the personalized datasets. For each query and selected subset S in our four test sets, we compute two metrics.

1. **Subset Interaction Gap**: $\Delta(S) = U(S) - \sum_{i \in S} U(\{\})$, which measures the total combinatorial gain.

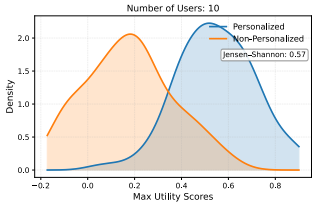
2. **Sequential Conditional Interaction Gap** Given an ordering of memory items, we define



(a) Number of Users: 1



(b) Number of Users: 5



(c) Number of Users: 10

Figure 11: Distribution of maximum utility scores for a personalized (blue) and non-personalized (orange) dataset differing by number of users averaged over.

$\delta_i = [U(S_i) - U(S_{i-1}] - U(\{i\})$, where S_i is the prefix subset. This measures how much the conditional contributions of each memory item deviates from independence.

Dataset	Subset Gap	Sequential Gap
Synthetic	0.23	0.11
Real World	1.01	0.09

Table 6: Results from the analysis of the combinatorial approximation gap between RUMS-Utility and RUMS-Binary. We show the average normalized subset interaction gap and the average sequential conditional interaction gap for personalized datasets.

Table 6 summarizes the measured combinatorial gaps across both synthetic and real-world datasets. We find the following:

- **Subset interaction gap**: The average normalized subset gap is 0.23 for the synthetic dataset and 1.01 for the real-world dataset, indicating that joint subsets can provide substantial additional utility beyond the sum of individual feature contributions.
- **Sequential conditional interaction gap**: The average per-feature sequential gap is 0.11 for synthetic data and 0.09 for real-world data, showing that individual features' conditional contributions deviate only slightly from the independent approximation.

These results highlight two key patterns. First, joint feature subsets offer additional utility, particularly in real-world datasets where the subset gap is larger. However, despite the larger subset-level gains, sequential conditional gaps remain small, suggesting that no single feature strongly violates the independence assumption. Together, these findings indicate that while higher-order interactions contribute to total utility, they are distributed across features and individually weak.

The analysis confirms that RUMS-Binary's itemwise approximation captures most of the subset utility. Even in datasets with substantial subset-level gains, the limited per-feature sequential gaps explain why RUMS-Binary performs well in practice while remaining computationally efficient. This supports its use as a scalable approximation to full subset-based RUMS-Utility selection.

A.5 RUMS-Utility Threshold Analysis

A.5.1 RUMS-Utility Threshold Effect on H1' Personalization Accuracy

We examined how personalization accuracy (agreement rate and F1), measured in H1', changes as we vary the threshold from 0 to 1 in 0.01 increments, on the test set used in our paper. For both datasets, Synthetic and Real World, we provide a graph comparing the metric (% agreement or F1) across thresholds. We also mark the optimal threshold (decided on a separate evaluation set), which we used in all experimentation. Lastly, we also show a zoomed-in view around the optimal threshold ± 0.005 for each graph (see Figure 16 - Figure 23).

In general, we found that personalization decisions are not very sensitive in respect to the threshold chosen. As we can see, for the Synthetic test set, when we change the threshold $+/-0.1$ we see only about a 15% decrease in agreement rate and only a 10% for the Real World Data. So, although the optimal threshold does seem to be heavily dataset dependent, it does seem moderately robust.

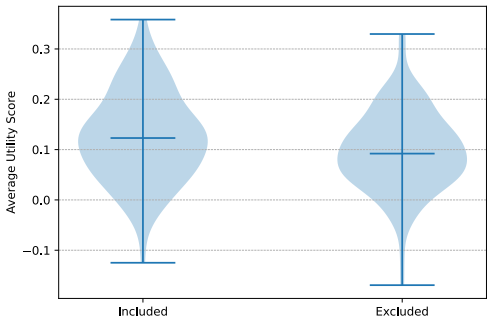


Figure 12: RUMS-Binary: Distribution

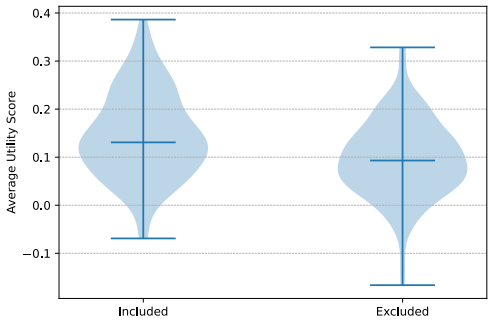


Figure 13: RUMS-Multi: Distribution

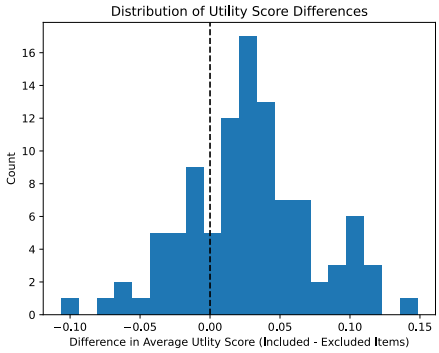


Figure 14: RUMS-Binary: Difference

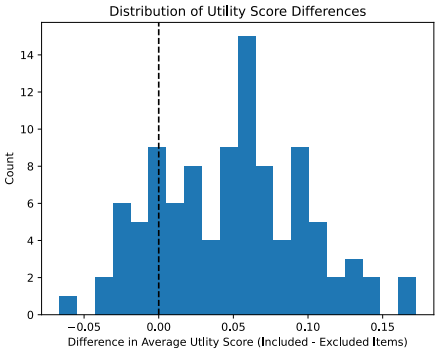


Figure 15: RUMS-Multi: Difference

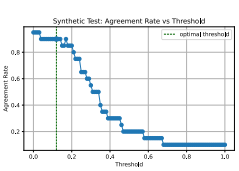


Figure 16: Synthetic: % Agreement vs. Threshold

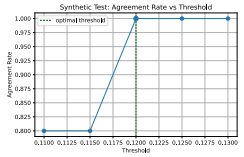


Figure 17: Synthetic: % Agreement vs. Threshold (zoomed in)

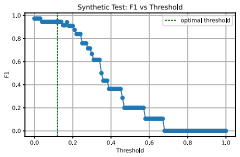


Figure 18: Synthetic: F1 vs. Threshold

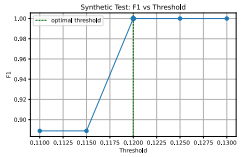


Figure 19: Synthetic: F1 vs. Threshold (zoomed in)

A.5.2 User-specific RUMS-Utility Threshold Effect on Personalization

We explore the idea of adaptive threshold per-user profile might change. To do this, we use a sample of $n = 100$ user inputs from FreebaseQA (non-personalized dataset) and PersonaFeedback (personalized dataset) respectively. For $u = 10$ static user profiles, we compute the maximum utility score for each. Then, we use the same procedure for finding the optimal threshold in the paper, to find the optimal threshold for each user. We then analyze the distribution of optimal thresholds for robustness using mean and standard deviations.

In Figure 24, we present the distribution of optimal thresholds for the 10 users (solid blue line), as well as the mean threshold (dotted red line) and ± 1 standard deviation (gray shaded area). We found that 70% users fell within a threshold of $[0.24, 0.30]$. So although there is some outlier profiles, we do not think there is enough variability that compensate for the added computational cost of per-user thresholds. We thank the reviewer for this query and find it interesting so we will add this analysis to our appendix.

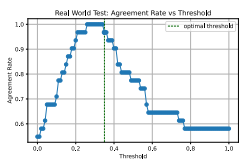


Figure 20: Real World: % Agreement vs. Threshold

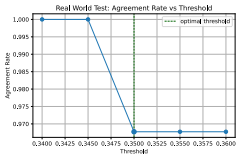


Figure 21: Real World: % Agreement vs. Threshold (zoomed in)

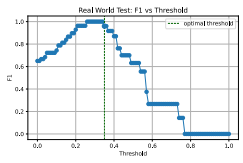


Figure 22: Real World:F1 vs. Threshold

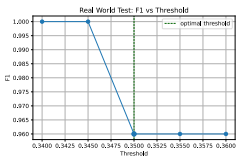


Figure 23: Real World: F1 vs. Threshold (zoomed in)

A.5.3 RUMS-Utility Threshold Effect on Number of Features Selected

As mentioned in the main text, RUMS-Utility has a threshold hyperparameter that controls the sensitivity of memory item selection. Adjusting this threshold results in either more or fewer memory items being chosen. Table 7 illustrates how the average number of selected memory items changes as the threshold is varied. As the threshold increases, a stronger reduction in entropy is required for a memory item to be selected, leading to smaller average subset sizes. This analysis highlights that the threshold provides a simple yet effective way to balance the tradeoff between including more context and avoiding unnecessary or low-utility memory items.

A.6 Examples of Selected Memory Items

Section A.6 illustrates one personalized and one non-personalized case. Consistent with Table 2, GPT4-ZeroShot and Semantic Similarity overselect, adding noise in non-personalized cases; GPT4-FewShot selects less in general but still selects items for non-personalized inputs. RUMS-Binary balances personalization by matching human annotations in the personalized case and abstaining in the non-personalized case.

A.7 H2: Qualitative Analysis of Failure

To better explore the cases when memory items selected by RUMS-Utility does not match the human selected, we further analyze the n=60 samples from H2. First, we define the different categories of cases we found:

- Correct: RUMS-Utility matches Human chosen list exactly
- Underselect: Human selected items and RUMS-Utility did not
- Some: RUMS-Utility selected a mixture of items selected at least one human, and not selected by humans
- All: RUMS-Utility selected all items which at least one human selected (but possibly not all humans)
- None: RUMS-Utility only selected items which no human selected
- Underselect - All: RUMS-Utility did select some items, which were chosen by all humans, but humans selected more items as well which RUMS-Utility did not.

Using these categories, we count the number of samples for each case.

Table Table 9 shows the results from this analysis, including the type of category, percentage of samples, and an example of each. We find majority of non-correct samples are from RUMS-Utility selecting values which at least one human selected but not all humans. This makes sense, as memory selection for personalization can be highly subjective and so this just shows that RUMS-Utility may select some of the more on the edge items. There is also a moderate level of underselect. This lends more weight to the idea that RUMS-Utility is a more conservative method, however, this could be changed based on the threshold selected by the user.

A.8 Extra Comparison of Qwen 3-Embeddings for H1/H2

To strengthen our analyses for H1 and H2, we additionally evaluate the cosine similarity method using the stronger Qwen3-Embedding (0.6B) model (Zhang et al., 2025). We follow the same procedure described in the paper, including selecting the similarity threshold using an evaluation set, with the only difference being the embedding model. As shown in Table 10 and Table 11, even with this substantially stronger embedding model, RUMS significantly outperforms the semantic similarity baseline.

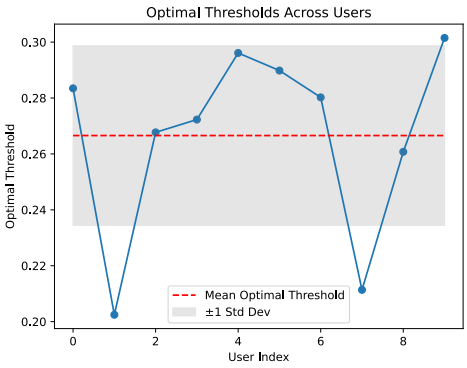


Figure 24: Optimal threshold calculated for each $n = 10$ users based on utility scores. It shows the optimal thresholds for the 10 users (solid blue line), as well as the mean threshold (dotted red line) and ± 1 standard deviation (gray shaded area)

Threshold	PersonaFeedback	FreebaseQA	WildChat-Personal	WildChat-NonPersonal
-100	2.76	1.83	2.77	2.70
0.00	2.76	1.56	2.73	2.57
0.29	2.43	0.21	1.31	1.38
0.50	1.11	0.00	0.29	0.27
0.80	0.14	0.00	0.00	0..00

Table 7: Average number of memory items selected at different thresholds used in RUMS-Utility.

A.9 Results of H3 using Human Judgment

For evaluation in H3, LLM-as-Judge was chosen due to it's validation as similar to third-party human annotation for personalization when both in-context examples and model confidence scores (Dong et al., 2024). However, to validate that selection alignment translates to better responses, we conducted a human study comparing RUMS-Binary with GPT-4 Few-Shot (strongest baseline). We had two human annotators (CSE PhD graduate students) annotate which of the two response they felt were better using the same rubric as the LLM-as-Judge in H3.

In Table 12 we found RUMS-Binary outperforms GPT4-Few by 33% in human preference, despite GPT-4 Few-Shot using a model $400\times$ larger for both selection and generation. This result (1) validates our LLM-as-Judge evaluation methodology (33% human win rate margin closely matches the 18% win rate margin from Table 3 against GPT-4 Few-Shot on synthetic data), (2) Confirms that RUMS's human-aligned selection decisions (from H2) translate to better final outputs (H3).

A.10 Analysis of Selected Entropy Calculation Model

As shown in Section A.1, we found that Llama-7 was the best performing model on calculating entropy scores that correctly seperate personalized versus non-personlized inputs. However, given that Mistral-7B was also moderately good at this task, we test how RUMS would work using Mistral-7B on both binary decisions of personalization (H1), and choosing memory item subsets (H2).

Table 13 show the results from H1 and Table 15 show the results from H2 from using Mistral-7B compared to Llama-8B. We find that although in general Llama-8B beats Mistral-7B, they perform quite closely. This might indicate that even if slightly less separation is seen between personalized and non-personalized entropy scores they can still be useful signal.

A.11 Analysis of *Optional* Filter Step

In our methods section, we mention that if a large number of candidate memory items exist it can be useful to use a more coarse method (such as a prompted LLM) to reduce the number of memory features for RUMS calculation. In our experi-

User Input	Humans*	RUMS-Binary	GPT4-ZeroShot	GPT4-FewShot	Semantic Similarity
[PERSONALIZED] I became a millionaire recently and want to give money to some local businesses	Location Hobbies and Interests Volunteer Activities	**Location** **Hobbies and interests** **Volunteer activities** Personal values Industry of employment	**Location** Occupation Industry of employment Current projects Long-term aspirations	**Location** Recent life events	**Location** **Volunteer activities** Financial situation Annual Income Pet ownership
[NON-PERSONALIZED] Who was the Prime Minister of Gt. Britain at the outbreak of The First World War?	No features selected	**No features selected**	Location Occupation Name Political affiliation	Location Occupation	Travel history Occupation Name Political affiliation Current projects

Table 8: Examples of memory items selected by each method compared to human gold labels. Overlapping memory with items human annotation is **bolded**.

Category	**%**	**Example**		**Chosen Items**	
		User Input	Humans (All)	Humans (> 1)	RUMS-Utility
Correct	38	"In May 1994 the Channel Tunnel was opened by Queen Elizabeth II and which French President?"	[]	[]	[]
Underselect	15	"How to use a vpn while vacationing in china"	[Tech savvy]	[Tech savvy, Feedback style, Comm mode, Tech use]	[none]
All	13	"Choose between new zara sneaker or used new balance"	[]	[Hobbies, Income, Aspirations, Finances, Exercise, Personality, Job, Industry, Pastimes, Events, Gender, Location]	[Hobbies, Income]
Some	22	"Buying a new notebook to write a script"	[]	[Location, Income, Tech savvy, Finances, Personality]	[Hobbies, Tech savvy, Fav books]
None	5	"My PC is r6 3600 and 1660s need new GPU"	[Projects, Finances]	[Income, Job, Hobbies, Industry, Projects, Aspirations, Finances, Pastimes]	[Tech savvy]
Underselect - All	7	"Im hungry"	[Foods, Diet needs, Fitness, Location]	[Foods, Diet needs, Fitness, Location, Income, Finances, Travel, Gender]	[Foods, Diet needs, Fitness]

Table 9: Breakdown of selection behavior categories with representative user inputs and selected memory items by humans (all agree), humans (at least one agree), and RUMS Utility.

Dataset / Metric	RUMS-Utility	Semantic Similarity (Qwen3-0.6B)
Synthetic (n=20) / Recall	94.4 [82.35, 100.00]	**100.0 [100.00, 100.00]**
Synthetic (n=20) / Specificity	**94.4 [84.85, 100.00]**	0.0 [0.00, 0.00]
Real-World (n=31) / Recall	**92.3 [75.00, 100.00]**	100.00 [100.00, 100.00]
Real-World (n=31) / Specificity	96.0 [85.71, 100.00]	**41.18 [18.75, 64.71, 100.00]**

Table 10: Recall (accuracy on personalized inputs) and specificity (accuracy on non-personalized inputs) across datasets and methods. Each metric is given by the value and the 95% CIs via nonparametric bootstrap (2,000 resamples with replacement over test-set rows). Bolded values indicate the best performance in each category.

mentation, we use GPT-4 as this method of coarse deselection which chooses up to 5 memory items from the original 50 to measure. In this section we validate the use of GPT-4 for this purpose.

To validate the use of GPT-4 and other possible other models, we use the randomly sampled personalized inputs which we have a consensus human annotated memory items selected (n=60) used in

Dataset	Metric	RUMS-Subset	RUMS-Binary	RUMS-Multi	Semantic Sim. (Qwen3-0.6B)
Synthetic	Precision	**0.61 [0.55, 0.66]**	0.60 [0.55, 0.64]	0.23 [0.19, 0.26]	0.09 [0.04, 0.14]
	Recall	0.70 [0.64, 0.75]	**0.89 [0.86, 0.92]**	0.39 [0.34, 0.44]	0.59 [0.43, 0.76]
	F1	0.70 [0.65, 0.75]	**0.78 [0.74, 0.82]**	0.31 [0.26, 0.36]	0.36 [0.29, 0.42]
Real World	Precision	**0.48 [0.42, 0.55]**	0.32 [0.29, 0.36]	0.30 [0.26, 0.34]	0.17 [0.09, 0.25]
	Recall	0.36 [0.30, 0.41]	**0.60 [0.55, 0.64]**	0.38 [0.34, 0.43]	0.47 [0.31, 0.64]
	F1	**0.53 [0.47, 0.59]**	0.45 [0.41, 0.49]	0.35 [0.31, 0.40]	0.33 [0.23, 0.45]

Table 11: Precision, recall, and F1 scores of memory items selected compared to human gold label annotation across datasets and methods. Each metric is given by the value and the 95% CIs via nonparametric bootstrap (2,000 resamples with replacement over test-set rows)

Dataset	RUMS-Utility Win Rate Margin
Synthetic	0.33
Real World	-0.03

Table 12: Win rate margin of responses generated using memory subset of RUMS-Utility vs. GPT4-Few, determined by human evaluation.

H2. Then, we analyzed how well the large general purpose models (GPT-4, Llama-70B, Qwen-72B, and Gemma-27B) filtered memory items (set to k=5) created by prompting the model, compared to the human selected memory items. We note that since these models are instructed to overselect (i.e. filter to a list of 5) we focus mostly on the recall metric, rather than precision of F1. Also, we only show the results for the personalized datasets, as this is when this is mostly used.

Table 18 shows the results from this analysis. We find that in general, all the models perform similarly with high recall to human annotations. However, we do see that GPT-4, the model we used, was the best at 0.77 and 0.70 recall for synthetic and real world data respectively. In general, this supports this method of coarse deselection.

To further validate these results, we were interested in if there is any bias in the type of features selected by the different models. So, we compared the top-10 most selected memory items per model. We found that on average, the models shared 6.8 memory items with other models and 4.75 items with humans. Table 20 shows the full results. This gave evidence that different models did not seem to have unique preferences or selections, thus adding further robustness to this method.

A.12 Cost Analysis

To better understand the practical implications of our approach, we analyze the tradeoff between computational cost and model performance across all methods. Since our methods and baselines differ in the number of memory items they select, and hence the number of additional tokens they introduce into the prompt, it is important to quantify how these differences affect overall efficiency at scale.

We measure computation overhead with two complementary metrics. First, we calculate the *average number of memory items selected* per dataset and method. This captures how the amount of items each method chooses in order to fulfill personalization or how little for no personalization. Second, we compute the *average number of additional tokens* that must be appended to the prompt during inference in order to include the selected memory items. This provides a direct measure of the inference overhead. Table 21 and Table 22 summarize these results, showing clear differences across methods: baselines such as All or Random predictably add large numbers of items, while information-retrieval heuristics like BM25 are highly selective. Our methods lie in between, striking a balance between selecting informative items and avoiding unnecessary additions.

To translate these measures into practical cost, we estimate the monthly cost of running one million queries per day for each method using pricing aligned with GPT-4. Specifically, we convert the additional tokens from Table 2 into dollar cost un-

Table 13: H1 Results

Dataset	Model	Recall	Specificity
Synthetic	Llama-7B	**0.94**	**0.94**
Real World	Llama-7B	0.92	**0.96**
Synthetic	Mistral-7B	0.83	0.88
Real World	Mistral-7B	0.92	0.77

Table 14: Results of using Llama-7B or Mistral-7B to calculate entropy scores on H1 experiments. Bolded is the higher value between the two models.

Table 15: H1 Results

Dataset	Model	Precision	Recall	F1
Synthetic	Llama-7B	0.61	**0.70**	**0.70**
Real World	Llama-7B	0.48	**0.36**	0.53
Synthetic	Mistral-7B	0.61	0.62	0.69
Real World	Mistral-7B	**0.61**	0.30	**0.69**

Table 16: Results of using Llama-7B or Mistral-7B to calculate entropy scores on H2 experiments. Bolded is the higher value between the two models.

der the assumption that each token incurs a fixed marginal cost of $0.01 per 1K tokens and 30 days per month. This allows us to approximate the total monthly inference cost for each method, and thus assess the economic tradeoff of personalization strategies at scale.

The results are visualized in the Figure 25. As expected, the All baseline is by far the most expensive, exceeding $160K per month due to its maximal inclusion of all memory items. On the other hand, our model RUMS-Binary yields much lower additional cost of only $6.9K which is about 30% lower than just the inference cost of including the memory features selected by GPT4-FewShot, not including the cost to generate these memory items. Our other model, RUMS-Multi, also has low cost of only $8.4K, keeping costs well below stronger baselines like GPT-4, ReContriever, and Semantic Similarity. Overall, this analysis highlights that RUMS-Models can be deployed at scale in a cost-efficient manner, with our methods demonstrating favorable tradeoffs between informativeness and economic feasibility.

A.13 Time Analysis

We ran a wall-clock analysis to showcase the time resources needed to run each method (Table 23. We calculated the average time for doing inference on one sample using $n = 3$ runs. We find that the RUMS-Multi have runtimes comparable to the retrieval methods (Semantic Similarity, BM25, and ReContriever) and RUMS-Binary has times comparable to GPT4-Fewshot. As mentioned, the RUMS-Utility has a very high runtime and therefore we do not recommend it to be used during inference.

A.14 Full Results of H3

In this section we expand H3 by breaking down results for the Synthetic and Real World datasets by inputs intended to be personalized (left) and not personalized (right). This comparison highlights differences between RUMS-Binary and RUMS-Multi. With the less capable response model, LLaMA-70B, RUMS-Binary (which favors fewer memory items) outperforms most baselines, showing especially strong gains on non-personalized inputs, with a 38% average win rate margin over baselines on the Synthetic dataset. Similar patterns hold on the Real World dataset, except for the None and BM25 baselines, which never select items.

By contrast, RUMS-Multi (which selects more memory items) performs better with the stronger GPT-4 model, particularly on personalized inputs, achieving 25% and 26% average win rate margins on the personalized Synthetic and Real World datasets respectively. These results demonstrate

Table 17: Model Performance on Personalized Datasets

Dataset	Model	Precision	Recall	F1
Synthetic Personalized	GPT-4	0.24	**0.77**	0.49
	Llama-70B	0.23	0.74	0.47
	Qwen-72B	0.21	0.74	0.45
	Gemma-27B	0.22	0.66	0.45
Real World Personalized	GPT-4	0.46	**0.70**	0.59
	Llama-70B	0.41	0.69	0.53
	Qwen-72B	0.44	0.67	0.55
	Gemma-27B	0.48	0.64	0.56

Table 18: Results of using large general purpose models for filtering. We compare the memory subsets filtered by each model, through prompting, with a human annotated list.

Table 19: Pairwise Agreement Between Models and Humans

	Humans	GPT-4	Llama-70B	Qwen-72B	Gemma-27B
Humans	—	—	—	—	—
GPT-4	0.4	—	—	—	—
Llama-70B	0.5	0.7	—	—	—
Qwen-72B	0.5	0.8	0.6	—	—
Gemma-27B	0.5	0.6	0.7	0.7	—

Table 20: Pairwise overlap (percentage) of the top-10 most frequently selected memory items across models and humans.

how RUMS can be tuned to different settings, adapting to the response model and personalization needs.

A.15 Analysis of Human Annotator Agreement

The majority-agreement statistic reported for H2 (41%, Section 4.3) aggregates across dataset types and does not by itself indicate whether disagreement reflects genuine task difficulty, inconsistent annotator instructions, or an idiosyncratic rater. We investigate this further below.

Agreement by Dataset Type. We computed the average pairwise Jaccard similarity between all four annotators' selected feature sets, for each of the $n = 60$ samples, broken down by dataset type. Table 26 reports the results.

Agreement is highly uneven across categories. FreebaseQA, a non-personalized dataset where the correct annotation is almost always "no features needed," shows near-perfect agreement (Jaccard $= 0.967$). The personalized categories — PersonaFeedback, WildChat-Personal, and WildChat-NonPersonal — show substantially lower agreement (0.21–0.45). This indicates that the *binary* decision of whether a query benefits from personalization at all is one annotators converge on reliably (consistent with our H1' results), while the harder and more inherently subjective task is selecting the *specific subset* of relevant memory items once personalization is warranted, which is the task H2 evaluates.

Chance-Adjusted Agreement. Raw Jaccard similarity is difficult to interpret without a reference point, since agreement is mechanically easier when annotators select from a smaller effective set of plausible items. We therefore computed the expected Jaccard similarity under random selection, meaning for two independently drawn random subsets of size k from a universe of $n = 50$ candidate memory items, expected Jaccard $\approx \frac{k}{2n-k}$, substituting the observed average number of items selected per annotator (k) for each dataset category. Table 27 reports the results.

Observed agreement exceeds chance substantially across all categories, though by varying mar-

Method	FreebaseQA	PersonaFeedback	WildChat-Personal	WildChat-NonPersonal
All	50	50	50	50
None	0	0	0	0
Random	5	5	5	5
Semantic Similarity	4.71	4.54	4.84	4.74
BM25	0.39	0.81	0.28	0.42
ReContriever	5	5	5	5
GPT4-Few	1.21	4.39	3.90	2.10
RUMS-Binary	0.02	3.14	2.80	1.64
RUMS-Multi	1.45	3.58	2.36	1.76

Table 21: Average number of memory items selected by method per dataset.

Method	FreebaseQA	PersonaFeedback	WildChat-Personal	WildChat-NonPersonal
All	551.59	551.28	552.16	550.40
None	0	0	0	0
Random	54.87	54.98	54.34	54.20
Semantic Similarity	48.04	51.44	53.04	49.58
BM25	1.48	8.95	3.20	4.86
ReContriever	60.87	62.93	61.22	61.72
GPT4-Few	12.96	49.31	46.02	23.58
RUMS-Binary	0.23	37.39	34.88	19.10
RUMS-Multi	13.88	45.34	30.90	21.08

Table 22: Average number of extra tokens required for response generation.

Method	Average Time (s)
All	0.0
None	0.0
Random	0.0
Semantic Similarity	0.21
BM25	0.01
ReContriever	0.08
GPT4-Few	1.22
RUMS-Binary	0.14
RUMS-Multi	0.04
RUMS-Utility	7.28

Table 23: Average time for inference of one sample, calculated over $n = 3$ trials. All values are in units of seconds.

gins. We note that the categories with the largest average subset sizes (PF, WP) show the smallest margin over chance. This is consistent with the intuition that as the number of relevant items grows, the space of equally reasonable subsets an annotator could select from also grows, mechanically increasing task difficulty and lowering raw agreement even among careful, non-adversarial annotators. We view this as an inherent property of open-ended memory selection over a large candidate space, rather than an artifact of unclear instructions or unreliable raters.

Ruling Out an Idiosyncratic Annotator. To check whether low agreement was driven by a single rater interpreting instructions differently from the others, we computed pairwise Jaccard similarity separately for each of the six annotator pairs, averaged across all samples. Agreement was fairly uniform across pairs (range: 0.43–0.51), with no pair showing a marked outlier in either direction. This suggests that disagreement reflects genuine, distributed task subjectivity rather than one annotator diverging from a shared understanding of the task.

Implications for Our Use of Majority Vote. Given the moderate raw agreement documented above, we deliberately used majority-vote labels (an item is included in the gold-standard subset only if selected by at least 3 of 4 annotators) rather than (i) any single annotator's selections, (ii) the union of all annotators' selections, or (iii) unanimous agreement, which would discard nearly all personalized examples at these agreement levels.

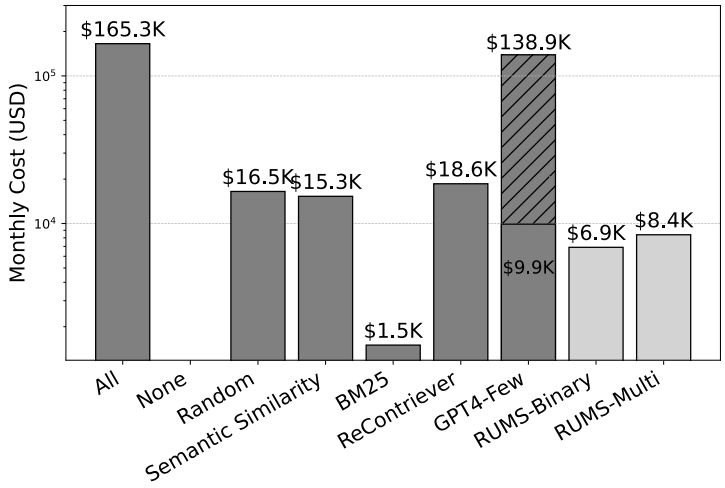


Figure 25: Estimated monthly cost of additional prompt tokens (1M queries per day, 30 days, GPT-4 input pricing at $0.01 per 1K tokens).

(a) Non-personalized inputs

	LLaMA 70B		GPT-4	
	RUMS-Binary.	RUMS-Multi	RUMS-Binary.	RUMS-Multi
None	-0.03	**-0.79**	0.02	0.00
All	**0.45**	**-0.24**	0.00	0.02
Random	**0.50**	**-0.12**	-0.03	-0.01
Semantic Similarity	**0.60**	0.01	0.02	0.00
BM25	**0.12**	**-0.66**	-0.02	0.01
ReContriever	**0.62**	**-0.14**	0.01	0.03
GPT4-Few	**0.42**	**-0.34**	0.00	0.01

(b) Personalized inputs

	LLaMA 70B		GPT-4	
	RUMS-Binary.	RUMS-Multi	RUMS-Binary.	RUMS-Multi
none	**0.52**	**0.53**	**0.50**	**0.72**
All	**-0.14**	**-0.14**	**-0.31**	0.01
Random	**0.18**	**0.37**	**0.15**	**0.39**
Semantic Similarity	-0.06	**0.17**	-0.07	0.09
BM25	**0.25**	**0.40**	**0.42**	**0.40**
ReContriever	-0.09	0.09	-0.09	0.09
GPT4-Few	-0.07	0.03	**-0.17**	0.03

Table 24: Average win rate margin comparison on Synthetic data under Not Personal (left) and Personal (right) conditions. Positive values (green) indicate better performance, negative values (red) indicate worse performance. Values greater than 0.1 in magnitude are bolded.

This choice was intended to filter out idiosyncratic or low-confidence individual selections and retain only memory items with genuine cross-annotator consensus. We further validate that these majority-vote labels are meaningful, despite moderate raw agreement, in Appendix A.9 (see Table 12), where responses generated using human majority-vote memory subsets outperform all baseline methods on downstream generation quality when using LLaMA 70B, and show a smaller but consistent advantage when using GPT-4.

B Formulas

B.1 Shannon Entropy

We define the Shannon Entropy as a measure of uncertainty within a system. For a random variable $x \in X$ and a distribution according to $p : X \to [0, 1]$ such that $p(x) := \mathbb{P}[X = x]$. Then we formalize the Shannon entropy H as,

$$H(x) = -\sum_{x \in X} p(x) \log p(x)$$

C Optimality of RUMS Proof

Proof. We begin with the joint conditional entropy of (y, z_u) given $(x, m_{u,S})$. By the chain rule,

$$\begin{aligned} H(y, z_u \mid x, m_{u,S}) = \\ H(y \mid x, m_{u,S}) + H(z_u \mid x, y, m_{u,S}) \\ = H(z_u \mid x, m_{u,S}) + H(y \mid x, z_u, m_{u,S}). \end{aligned}$$

Equating the two expressions and rearranging yields

$$\begin{aligned} H(y \mid x, m_{u,S}) = \\ H(y \mid x, z_u, m_{u,S}) + H(z_u \mid x, m_{u,S}) - \\ H(z_u \mid x, y, m_{u,S}). \end{aligned}$$

By Assumption 2, memory $m_{u,S}$ provides little additional information about y beyond (x, z_u). We

(a) Non-personalized inputs

	LLaMA 70B		GPT-4	
	RUMS-Binary	RUMS-Multi	RUMS-Binary	RUMS-Multi
None	**-0.36**	**-0.64**	**-0.12**	**-0.26**
All	0.08	**-0.14**	**-0.14**	**-0.16**
Random	-0.02	-0.06	-0.06	**-0.26**
Semantic Similarity	0.04	-0.02	-0.02	**-0.12**
BM25	**-0.40**	**-0.16**	**-0.16**	**-0.26**
ReContriever	**0.12**	0.04	0.04	0.02
GPT4-Few	0.08	**0.12**	**0.12**	-0.02

(b) Personalized inputs

	LLaMA 70B		GPT-4	
	RUMS-Binary	RUMS-Multi	RUMS-Binary	RUMS-Multi
None	**0.46**	**0.70**	**0.52**	**0.60**
All	**-0.16**	-0.04	**-0.14**	0.04
Random	**0.24**	**0.42**	**0.26**	**0.34**
Semantic Similarity	-0.10	0.00	-0.02	**0.12**
BM25	**0.42**	**0.58**	**0.52**	**0.52**
ReContriever	0.00	**0.24**	**-0.14**	**0.24**
GPT4-Few	-0.08	-0.04	-0.04	-0.04

Table 25: Average win rate margin comparison on Real World data under Not Personal (left) and Personal (right) conditions. Positive values (green) indicate better performance, negative values (red) indicate worse performance. Values greater than 0.1 in magnitude are bolded.

Dataset	Avg. Pairwise Jaccard	Std. Dev.	Avg. # Items Selected
FreebaseQA (FB)	0.967	0.129	1.02
PersonaFeedback (PF)	0.212	0.125	4.72
WildChat-NonPersonal (WNP)	0.454	0.372	1.71
WildChat-Personal (WP)	0.297	0.147	8.03

Table 26: Average pairwise Jaccard similarity between the four human annotators' selected memory item subsets, by dataset type, along with the average number of items selected per annotator in each category.

therefore approximate

$$H(y \mid x, z_u, m_{u,S}) \approx H(y \mid x, z_u).$$

Substituting this approximation gives

$$H(y \mid x, m_{u,S}) \approx H(y \mid x, z_u) + H(z_u \mid x, m_{u,S}) - H(z_u \mid x, y, m_{u,S}).$$

The first term, $H(y \mid x, z_u)$, does not depend on the choice of subset S. In addition, Assumption 2 implies that conditioning on $m_{u,S}$ does not substantially refine the posterior over z_u once (x, y) are observed, so

$$H(z_u \mid x, y, m_{u,S}) \approx H(z_u \mid x, y),$$

which is also independent of S.

Combining these observations, we obtain the approximation

$$H(y \mid x, m_{u,S}) \approx C + H(z_u \mid x, m_{u,S}),$$

where C is a constant that does not depend on S.

Therefore, minimizing predictive entropy $H(y \mid x, m_{u,S})$ is approximately equivalent to minimizing uncertainty about the latent variable z_u given the selected memory.

Now consider expected utility as defined in Assumption 1,

$$\mathbb{E}[U_u(y)] = -\mathbb{E}_{z_u}\mathbb{E}_{y\sim p_\theta(\cdot|x,m_{u,S})}[\ell(y, z_u)].$$

Because ℓ is a strictly proper scoring rule, the expected loss is minimized when the predictive distribution $p_\theta(y \mid x, m_{u,S})$ matches the conditional distribution induced by z_u. Reducing uncertainty about z_u therefore strictly improves expected utility.

Consequently, any subset S^* that approximately minimizes $H(y \mid x, m_{u,S})$ also approximately maximizes expected utility. □

D Sequence Entropy Equivalence Proof

Proof. Let $p_\theta(y \mid x)$ denote the model distribution over token sequences $y = (y_1, \ldots, y_T)$ given input x. At each step t, the conditional next-token distribution is

$$p_\theta(\cdot \mid y_{<t}, x),$$

where $y_{<t}$ is the prefix of tokens before step t.

By the chain rule of probability,

$$-\log p_\theta(y \mid x) = \sum_{t=1}^{T} -\log p_\theta(y_t \mid y_{<t}, x).$$

Taking the expectation with respect to $y \sim p_\theta(\cdot \mid x)$

Dataset	Avg. k	Chance Jaccard	Observed Jaccard	Ratio (Observed / Chance)
FB	1.02	0.0103	0.967	$\sim$94$\times$
PF	4.72	0.0495	0.212	$\sim$4.3$\times$
WNP	1.71	0.0174	0.454	$\sim$26$\times$
WP	8.03	0.0874	0.297	$\sim$3.4$\times$

Table 27: Comparison of observed average pairwise Jaccard similarity to expected Jaccard similarity under random subset selection (chance level), computed using the observed average subset size per dataset category.

on both sides gives

$$
\begin{aligned}
&\mathbb{E}_{y\sim p_\theta}[-\log p_\theta(y \mid x)] = \\
&\quad \mathbb{E}_{y\sim p_\theta}[\sum_{t=1}^{T} -\log p_\theta(y_t \mid y_{<t}, x)] \\
&\quad = \sum_{t=1}^{T} \mathbb{E}_{y\sim p_\theta}\left[-\log p_\theta(y_t \mid y_{<t}, x)\right] \\
&\quad = \sum_{t=1}^{T} \mathbb{E}_{y_{<t}\sim p_\theta}[\sum_{y_t} p_\theta(y_t \mid y_{<t}, x) \\
&\quad (-\log p_\theta(y_t \mid y_{<t}, x))] \\
&\quad = \sum_{t=1}^{T} \mathbb{E}_{y_{<t}\sim p_\theta}\left[H\big(p_\theta(\cdot \mid y_{<t}, x)\big)\right].
\end{aligned}
$$

□

E RUMS-Utility Algorithm

In Algorithm 1 we provide the formalized algorithm for RUMS-Utility.

Algorithm 1 Memory Selection via RUMS

Require: User input x, user memory $u \in \mathbb{R}^d$
Ensure: Selected memory subset $m^*_{u,S}$ that maximizes utility

Candidate Memory Item Identification:
Use LLM to generate candidate subsets S for x
Subset Evaluation:
for each $s \in S$ **do**
    Map candidate attritubes to user values $m_{u,S}$
    Prompt Reference LM with:
        (i) x only
        (ii) x and $m_{u,S}$
    Compute estimated utility $\hat{\text{EIG}}(S)$ using Equation (2)
    Store utility score for subset s
Subset Selection:
$s^* \leftarrow \arg\max_{s\in S} \hat{\text{EIG}}(S)$
if all MI scores $<$ thresholdτ **then**
    $s^* \leftarrow \emptyset$ ▷ Select no features

F Experimental Details

This section provides the full details of the experiments described in Section 4.

F.1 Experimental Setup

Across all experiments (H1–H3), we consider a fixed set of $n = 50$ memory items (e.g., *location, age, dietary restrictions*). The task is to select a subset of up to $n = 5$ memory items that can improve personalization.

F.2 Datasets

To evaluate scenarios where personalization is beneficial as well as those where it is not, we use a diverse mixture of user inputs drawn from prior work, including both synthetic and real-world sources. The three datasets we used are described below:

PersonaFeedback (Tao et al., 2025) This is a synthetically created dataset consisting of user-specific queries designed to reflect inputs from realistic and diverse user personas. First, they constructed personas by representing them as a structured dictionary of memory items, including demographics, personality traits, and preferences. Then, open-source content from social media, reviews, and forums were used to populate these items with specific values. Lastly, an LLM generator was used to produce queries based on a sampled persona and specific sub-scenarios. Additional steps such as similarity-based filtering, rephrasing to avoid directly revealing persona information, and manual annotations ensured diversity and realism. The final dataset contains more than 4,000 persona-specific questions. The original data is in Chinese and therefore, we translated to English using a machine translation model (Tiedemann and Thottingal, 2020). Example: *"Do you have any music for the trip?"*

FreebaseQA (Jiang et al., 2019) A dataset of trivia-style questions and answers collected from human trivia competitions. The original dataset contains approximately 28,000 question–answer pairs. Example: *"What type of snake is a sidewinder?"*

WildChat (Zhao et al., 2024) A large-scale collection of more than 1 million real-world user–ChatGPT interactions, encompassing over 2.5 million dialogue turns. The dataset was compiled by offering free access to ChatGPT and GPT-4 in exchange for user consent to share chat histories. It covers diverse conversational scenarios such as ambiguous requests, code-switching, topic shifts, and political discussions. We filter the data to retain only English conversations labeled as non-toxic, and we use only the first human-written input per conversation to ensure that queries contain sufficient context.

To determine whether an input is personalized, two human annotators labeled a random selection of WildChat queries as either beneficial for personalization (WildChat-Personal) or not (WildChat-NonPersonal). Only queries with full annotator agreement were retained. Examples: *"I'm hungry"; "Who is Mohamed Gamal Mohamed Manieh?"*

Synthetic vs. Real-World Datasets By design, PersonaFeedback is treated as a personalized input set, and FreebaseQA as non-personalized. We collectively refer to these as the "Synthetic" dataset, since they were not originally created for LLM input. WildChat is considered a "Real-World" dataset, as it was collected directly from real user interactions with an LLM. We use both synthetic and real-world datasets to evaluate the robustness of RUMS across many diverse inputs.

Sample Size by Hypothesis The number of samples varies across hypotheses. For H1, we use $n = 500$ samples from PersonaFeedback and FreebaseQA to analyze RUMS-Utility for binary personalization decisions. For H1', we use $n = 25$ inputs each from PersonaFeedback, FreebaseQA, WildChat-Personal, and WildChat-NonPersonal to extend the H1 analysis to correlation with human judgment. For H2, we use $n = 15$ inputs each from PersonaFeedback, FreebaseQA, WildChat-Personal, and WildChat-NonPersonal to examine the correlation between RUMS memory selection and human preference. Finally, for H3, we use $n = 100$ samples from PersonaFeedback and FreebaseQA along with $n = 50$ samples each from WildChat-Personal and WildChat-NonPersonal to analyze the downstream impact of subsets chosen by RUMS-Model.

F.3 Profile Generation

User profiles are incorporated during the training data curation phase, where specific user values are used to compute the entropy difference of the model required for the utility function in Equation (2). Also, we use profiles in our test sets for evaluation.

Memory Item Creation To construct user profiles, we prompted GPT-4 (OpenAI et al., 2024) to generate 50 commonly occurring user memory items. These items span demographic, behavioral, and preference-based attributes, providing a diverse set of values for personalization. The complete list is as follows:

- Name
- Gender
- Location (City/State/Country)
- Ethnicity
- Race
- Annual Household Income

- Language spoken at home
- Language preference
- Sexual Preference
- Disabilities
- Occupation
- Political affiliation
- Education level
- Hobbies and interests
- Favorite sports
- Preferred music genre
- Favorite books
- Preferred movie genre
- Favorite TV shows
- Favorite foods
- Dietary restrictions (e.g., vegetarian, vegan)
- Relationship status
- Number of children
- Pet ownership (types of pets)
- Travel history (countries visited)
- Travel preferences (adventurous, relaxing)
- Social media platforms used
- Level of tech-savviness
- Preferred mode of communication (text, voice, video)
- Work schedule (full-time, part-time)
- Industry of employment
- Current projects or goals
- Long-term aspirations
- Financial situation (broad categories)
- Health and fitness goals
- Preferred exercise activities
- Level of environmental consciousness
- Volunteer activities or interests
- Current challenges or pain points
- Personality traits (e.g., introverted, extroverted)
- Favorite pastimes
- Personal values or beliefs
- Cultural background
- Religious beliefs
- Recent life events (e.g., change in job, moved)
- Technology usage (devices, apps)
- Frequency of using the service (daily, weekly)
- Reasons for using the service
- Feedback preferences (detailed, brief)
- Preferred tone of communication (formal, casual)

Memory Item Values Next, we used GPT-4 (OpenAI et al., 2024) to generate consistent user profiles without contradictions (e.g., "vegan" with "favorite food = steak"), following the prompt in prompt 1. To further ensure logical consistency, we employed GPT-4 to filter out any profiles that appeared illogical, as described in prompt 2. This process was repeated until we obtained $n = 100$ unique, logically consistent profiles for both training and testing.

This method yields a diverse set of user profiles, as illustrated in the bar charts of memory item distributions shown in Figure 26 through Figure 75.

Prompts used for Profile Generation

Prompt 1. *For these 50 features, I want you to create 10 very diverse persona profiles. So for each of the 50 features, I want you to select a value (e.g. "age" = 50) that would make up the persona profile. Make sure the values make sense for each other, meaning don't have a person who is 12 years old have a job as a accountant.*

Output a json file with a dictionary with 10 keys, "Persona_1", "Persona_2",...,"Persona_10", whose corresponding values is a list of the selected values in the same order as the features (e.g. [John, Male, Seattle/WA/USA,...]).

Do not include any other markdown or text. Here are the Features:

Prompt 2. *You are an AI assistant that needs to decide if a user profile is logically sound. This means that the information in the user profile should be consistent and non-conflicting.*

Only output a single word, either 'yes' or 'no', to indicate if the user profile is logically sound.

Example 1: If a user profile states that the user is allergic to peanuts, it should not also state their favorite food is peanuts.

Example 2: If a user profiles states highest education is a GED, it should not also state their profession is a doctor.

Example 3: If a user profile states that the user is 25 years old, it should not also state that they have been working for 30 years.

Example 4: If a user profile states that they live in Columbia, it should not state that their main language is French.

User Profile:

F.4 Baselines

In this section we describe each baseline that is used in our experimentation.

GPT-4 Prompting (OpenAI et al., 2024): We use GPT-4o as a baseline across H1, H2, and H3. In H1, we apply prompt 3 for the zero-shot setting and prompt 4 for the few-shot setting to decide whether personalization is needed for a given query. For the few-shot case, we provide four in-context examples: two requiring personalization and two not requiring personalization.

In H2 and H3, we adapt the prompts to instead elicit the subset of specific user memory items that would be beneficial for personalization. We use prompt 5 for zero-shot and prompt 6 for few-shot. In the few-shot setting, we provide two in-context examples: one requiring personalization and one not.

Prompt 3. *You are a classifier that determines whether a given user input would benefit from user-specific information for the model to respond.*

Your response must be a JSON object with a two key "Decision" with a value of either "True" or "False" and "Features" with either a list of user-specific features (at most 10 features) if "Decision" is "True" or "None" if "Decision" is "False" :

Output "True" if the input would benefit from user-specific information to better respond to the query.

Output "False" if the input would not benefit from user-specific information or would be the same answer regardless of the user.

Respond with only the JSON object. Do not include any explanation or additional text.

Input: user_input

Prompt 4. *'You are a classifier that determines whether a given user input would benefit from user-specific information for the model to respond.*

Your response must be a JSON object with a two key "Decision" with a value of either "True" or "False" and "Features" with either a list of user-specific features (at most 10 features) if "Decision" is "True" or "None" if "Decision" is "False" :

Output "True" if the input would benefit from user-specific information to better respond to the query.

Output "False" if the input would not benefit from user-specific information or would be the same answer regardless of the user.

Respond with only the JSON object. Do not include any explanation or additional text.

Example 1:

Input: "Where should I eat tonight?"

Output: "Decision": "True", "Features": ["location", "dietary restrictions", "income"]

Example 2:

Input: "Can you help me write an essay on feminism?"

Output: "Decision": "True", "Features": ["education", "political ideology", "gender"]

Example 3:

Input: "What is the capital of Austin, Texas?"

Output: "Decision": "False", "Features": "None"

Example 4:

Input: "Can you explain what AGI is?"

Output: "Decision": "False", "Features": "None"

Now classify the following input:

Input: user_input'

Prompt 5. *Given a user-input and list of features, select which features might be important for answering the user-input.*

Choose up to 5 features. Respond in a JSON with one key "Chosen Features", which has a list of chosen features. Do not include any other text or formatting.

User Input: user_input

Possible Features: candidate_features_ls

Prompt 6. *Given a user-input and list of features, select which features might be important for answering the user-input. Choose up to 5 features.*

Respond in a JSON with one key "Chosen Features", which has a list of chosen features. Do not include any other text or formatting.

Example Input 1: "whats the weather today"

Example Output 1: "Chosen Features": ["Location (City/State/Country)"]

Example Input 2: "What is the capital of Texas?"

Example Output 2: "Chosen Features": []

User Input: user_input

Possible Features: candidate_features_ls'

Semantic Similarity: In H1–H3, we also include a simple baseline based on semantic similarity. Here, we compute the cosine similarity between user query embeddings and each memory item embeddings using a lightweight sentence-transformer model (Wang et al., 2020).

In H1, we use the validation set, defined as a 20% hold-out of the data, to tune a threshold for deciding whether personalization is required. Thresholds are determined by grid search over $[-1, 1]$ in increments of 0.5, selecting the value that maximizes agreement with human annotations for each dataset. A query is labeled as requiring personalization if at least one memory item has a similarity score above the selected threshold. The exact thresholds used for each dataset are reported in Table 28.

Dataset	Threshold
All	-1.0
Synthetic	-1.0
Real World	0.5

Table 28: Threshold for H1 analysis of semantic similarity baseline. If at least one memory item had a semantic similarity higher than this threshold, then the query was labeled as needing personalization.

Since H2 and H3 require selecting specific memory items for personalization rather than making a simple binary decision, we set an intuitive threshold of $t = 0$, corresponding to no similarity between the user query and a memory item. If more than five memory items exceed this threshold, we select the top five with the strongest correlations.

Simple Baselines: We also compare against several naive baselines. The "None" baseline provides no features ($n = 0$), serving as a lower bound for personalized inputs and an upper bound for non-personalized ones. The "All" baseline includes all available features ($n = 50$), testing the effect of providing maximal but noisy context. Finally, the "Random" baseline samples $n = 5$ items at random, establishing a chance-level reference.

SOTA Retrieval Methods: We additionally evaluate two state-of-the-art retrieval methods.

First, BM25 (Lù, 2024), a widely used ranking function in information retrieval, can be adapted for memory selection. It scores memory items (treated as terms) based on their frequency in each instance relative to the entire dataset, balancing how common a memory item is with how discriminative it is. Items with higher BM25 scores are considered more informative and are selected as the most relevant, making BM25 well-suited for sparse or high-dimensional text and categorical data.

Second, ReContriever (Lei et al., 2023) is a retrieval-based method that scores memory items by their ability to help a model retrieve relevant information. It leverages dense embeddings to measure semantic similarity between queries and candidate items, effectively identifying the most informative items for downstream tasks. Memory items that yield higher retrieval relevance are prioritized, making ReContriever particularly effective for selecting contextually important or predictive inputs in text or structured datasets. Similarly to semantic similarity a threshold of $t = 0$ is used as threshold for selecting memory items, with a maximum of up to $n = 5$ selected.

F.5 RUMS-Utility and RUMS-Model Details

In this paper, we describe two main methods: RUMS-Utility, which selects a subset of user memory items with the highest utility for personalization directly, and RUMS-Model, a lightweight trained model built on data generated by RUMS-Utility.

RUMS-Utility: We use RUMS-Utility not only for the analyses in H1 and H2 but also to generate training data for fine-tuning RUMS-Models. In general, we compute entropy in Equation (3) using Llama-3.1-8B (AI, 2024b) with a temperature of 0.7 and sampling-based decoding to obtain the response distribution. We restrict the maximum generation length to $T = 20$ tokens, chosen to mitigate bias from longer generations having naturally lower entropy.

Following the algorithm in Section 3.4, we first prompt GPT-4 to reduce the candidate memory items from 50 to at least 1 and at most 10 before computing RUMS-Utility (see prompt 7). Utility scores are then estimated using $n = 5$ Monte Carlo

samples. For each input, we generate $n = 5$ distinct user profiles as separate training points, which can yield entirely unique best memory subsets of data depending on the values in the user profile.

The chosen hyperparameters (e.g., model size, number of Monte Carlo samples, etc.) reflect a tradeoff: obtaining reliable signals of memory item importance while keeping computations lightweight and feasible for large-scale inference. Further discussion of these choices is provided in Section A.1.

Prompt 7. *Given a user-input and list of features, select which features might be important for answering the user-input.*

Choose at least 1 and up to 10 features.

Respond in a JSON with one key "Chosen Features", which has a list of chosen features. ' Do not include any other text or formatting.

Choosing the Personalization Threshold: To determine the threshold for deciding whether to apply personalization or select no personalization, we used a held-out validation set of size $n = 100$ from each of PersonaFeedback and FreebaseQA. For each query, we computed the average maximum utility score over the $n = 5$ generated user profiles. We labeled inputs from PersonaFeedback as personalized (1) and inputs from FreebaseQA as non-personalized (0). These scores were combined and labeled by dataset to construct an ROC curve.

We then selected the threshold that maximized Youden's J statistic ($J = \text{TPR} - \text{FPR}$), which identifies the point that best balances sensitivity and specificity. Using this procedure, we determined an optimal threshold of $t = 0.29$, which was applied in all downstream analyses.

RUMS-Models For training the downstream models, we sampled approximately 2.5K unique PersonaFeedback queries and 2K unique FreebaseQA queries. Each query was paired with five user profiles, and RUMS-Utility was applied to select the best memory subsets, yielding 13K and 9K training examples, respectively. We excluded all FreebaseQA queries that exceeded the personalization threshold, thereby removing any noisy non-personalized samples. This design choice emphasized that non-personalized inputs should not include any memory items.

We then used these (query, memory subset) pairs to train two models. The first model, used in H1–H3, was a binary logistic regression implemented with DeBERTa-v3-large (He et al., 2021) and a final linear classification layer. To train this model, we converted each (query, memory subset) pair into multiple (query, memory item, label) triplets. Training was performed with a learning rate of 5×10^{-6} for 12 epochs, and the best model was selected based on the highest F1 score on a held-out validation set ($n = 1K$).

The second model, used for H3, was a multiclass, multilabel logistic regression implemented again with DeBERTa-v2-large (He et al., 2021) and a final multiclass classification layer. Here, labels were constructed by merging the selected subsets from the $n = 5$ diverse user profiles per query. This produced a single label vector in which all memory items chosen in any of the five samples were marked as significant. This resulted in more dense labels, training the model to select more memory items. Training was performed with a learning rate of 2×10^{-5} for 12 epochs, and the best model was again chosen based on the highest F1 score on a held-out validation set ($n = 1K$).

Method	Model	# Parameters
RUMS	microsoft/deberta-v3-large	418M
Semantic Similarity	all-MiniLM-L6-v2	∼22.7M
GPT4o	GPT4o	Not disclosed (est. >200B)
ReContriever	Yibin-Lei/ReContriever	∼110M

Table 29: Baseline methods and their corresponding model sizes.

Selecting Regression Threshold Since RUMS-Models is trained as a logistic regression model (either binary or multiclass), its output represents the probability of selecting a particular memory item. A common approach is to use a cutoff of 0.5 to make a binary decision, however, since our task is more specialized, instead we tuned this cutoff using a held-out evaluation set ($n = 1K$). Specifically, we perform a grid search over boundaries in $[0, 1]$ and select the cutoff that achieved the highest F1 score on the evaluation set. This yielded $b = 0.81$ for RUMS-Binary and $b = 0.93$ for RUMS-Multi.

F.6 Comparing Size of Methods

In Table 29, we summarize the size of each model used in both our method and the baselines. Our approach uses a model comparable in scale to SOTA baselines, yet it is over $400\times$ smaller than GPT-4.

F.7 Evaluation Metric and Analysis

H1 Evaluation: In H1, we compare the distribution of maximum utility scores between as-

sumed personalized and non-personalized inputs. Specifically, we evaluate utility scores on a non-personalized dataset (FreebaseQA) and a personalized dataset (PersonaFeedback). We sampled $n = 500$ inputs from each dataset and computed the maximum utility score using Equation (2) for every input. The resulting distributions were then analyzed using divergence metrics and hypothesis testing.

For divergence metrics, we use Jensen–Shannon (JS) divergence, a symmetric measure of similarity between two probability distributions P and Q. It is defined as:

$$\mathrm{JS}(P\|Q) = \frac{1}{2}\,\mathrm{KL}(P\|M) + \frac{1}{2}\,\mathrm{KL}(Q\|M), \quad (5)$$

where $M = \frac{P+Q}{2}$.

Here, $\mathrm{KL}(P\|M)$ is the Kullback–Leibler divergence from P to the average distribution M. JS divergence is always finite, symmetric, and bounded between 0 and $\log 2$.

For hypothesis testing, we use the two-sample Kolmogorov–Smirnov test to assess whether the two samples come from the same distribution. The null hypothesis states that they do, and we test significance at the $\alpha = 0.05$ level.

H1' Evaluation: In H1', performance was evaluated using human labels as the gold standard. For each sample, four human annotators rated whether $n = 100$ samples would benefit from personalization using a 0–2 scale. If the annotators' confidence in personalization was 1 or higher, we considered the input to be personalized. Instructions provided to annotators are given in prompt 8. We only included inputs with unanimous agreement ($n = 64$) in the evaluation. The average interrater agreement between pairs of annotators was $\kappa = 0.59$ using the Cohen's kappa. The primary metrics used were recall, measuring accuracy in identifying personalized inputs, and specificity, measuring accuracy in identifying non-personalized inputs.

Prompt 8. *Instructions: Read the user input and then decide if the input would benefit from specific user information.*

0 = No Personalization Needed, 1 = Personalize is might be helpful, 2= Personalization is definitely helpful

H2 Evaluation: For H2, we evaluated how closely each method's selected subset of memory items matched the subsets chosen by humans. To create the human-selected subsets, four annotators were asked to select all memory items (from the full set of 50) that they felt would be helpful in generating a response. Instructions provided to annotators are given in prompt 9. For each input, we included only the memory items selected by at least three annotators to form the majority subset. This process resulted in an average of 4.19 memory items per query for inputs where at least one item was chosen. This average guided our experimental design for H2 and H3, as we restricted each method to select no more than $n = 5$ items.

We evaluated performance using *average* precision, recall, and F1 score. Here, precision measures the fraction of items selected by a method that were also in the human majority subset, recall measures the fraction of items in the human majority subset that were correctly selected by the method, and F1 is the harmonic mean of precision and recall, providing a single metric that balances both. However, precision, recall, and F1 are each computed per-query and then averaged independently across queries; at the per-query level, F1 is always correctly bounded between that query's precision and recall. However, because F1 is a nonlinear function of precision and recall, the mean of per-query F1 scores is not equal to the harmonic mean of mean precision and mean recall.

Number of Memory Items Selected in H2 Lastly, we include table 30 which compares the average number of memory items selected by each method in H2. Of note, we see that human annotations show a clear separation: personalized inputs have on average 6.5 features selected, compared to only 1.4 for non-personalized inputs, a difference of 5. Among automatic methods, RUMS-Model and RUMS-Utility comes closest to this behavior, achieving a range of 3.3 and 1.08, whereas the next closest method is GPT4-ZeroShot which achieves only a 0.67 difference. This further highlights how RUMS excels at both distinguishing when memory items are beneficial as well as which features are beneficial.

Prompt 9. *For each input, decide which (if any) user-features would be helpful to better respond to the input.*

H3 Evaluation: In H3, we evaluate the downstream impact of the memory subsets chosen by each method on generated responses. For each user

	Personalized	Not-Personalized	Difference
Human Annotation	6.45	1.43	5.02
RUMS-Utility	1.87	0.80	1.07
RUMS-Binary	4.87	**1.56**	**3.31**
RUMS-Multi	3.27	1.67	1.60
Semantic Similarity	**5.00**	5.00	0.00
GPT4-ZeroShot	3.30	2.61	0.69
GPT4-FewShot	1.80	1.57	0.23

Table 30: Average number of features selected for personalized vs. not-personalized inputs. Methods closes to human performance are **bolded**.

input, responses are generated using the query combined with the subset of memory items selected by each method. This evaluation is performed with two models: GPT-4 (OpenAI et al., 2024), a high-performing model with robust capabilities, and LLaMA 70B Instruct (AI, 2024a), a smaller model more susceptible to noise. This setup demonstrates RUMS-Models's flexibility in adapting to models with varying capacities and usage scenarios.

Performance is measured using *generation win rate margin*, defined as the proportion of pairwise comparisons in which RUMS-Models's output is judged preferable to that of a baseline. GPT-4 is used as the judge, comparing responses generated with RUMS-Models's selected subset against those generated with each baseline's subset. LLM-as-Judge is validated to align closely with third-party human annotations for personalization when combined with in-context examples, model confidence scores (Dong et al., 2024).

To ensure robust evaluation, our methodology includes:

- *In-context examples*: Three examples are provided in the prompt for each evaluation to demonstrate the expected style of reasoning and scoring.
- *Confidence thresholds*: Only responses for which the model reports a self-generated confidence score above the 70% percentile are included in the final analysis.
- *Chain-of-thought prompting*: Evaluation prompts encourage the model to reason step-by-step and justify its preference for one response over another, improving consistency and interpretability of the win-rate comparisons.

For personalized datasets (PersonaFeedback and WildChat-Personal), the evaluation metrics in the chain-of-thought include:

- *Completeness*: Does the response fulfill the user's request?
- *Personalization*: Does the response appropriately incorporate relevant details from the user profile?
- *Overpersonalization*: Does the response include irrelevant, distracting, or unnecessary user details?

For non-personalized datasets (FreebaseQA and WildChat-NonPersonal), the metrics are:

- *Completeness/Accuracy*: Does the response correctly and fully answer the user's request?
- *Neutrality*: Is the response unbiased and general, rather than tailored to the user?

Full evaluation prompts can be found in prompt 11 for personalized datasets, prompt 12 for non-personalized datasets, and prompt 10 for the three in-context examples used in all evaluation.

Prompt 10. *Example 1:*

"user profile": ['Preferred exercise activities: Weightlifting', "Hobbies and interests: ['Surfing', 'Photography']", 'Location (City/State/Country): Melbourne/Victoria/Australia', "Social media platforms used: ['Facebook', 'LinkedIn']", "Technology usage (devices, apps): ['Smartphone', 'Laptop']],

"user input": " Why are there five rings in Olympic logo? And what does they symbolize?",

"response_1": " The five rings in the Olympic logo represent the five continents: Africa, the Americas, Asia, Europe, and Oceania. They are interlinked to symbolize the unity and friendship among

nations, highlighting the idea of bringing together athletes from around the world to compete in the spirit of sportsmanship. The colors of the rings (blue, yellow, black, green, and red) were chosen because at least one of these colors appears in the national flags of all participating countries.",

"response_2": "The five rings in the Olympic logo represent the five continents of the world , Africa, the Americas, Asia, Europe, and Oceania. This design, created by Pierre de Coubertin in 1913, symbolizes the unity and connection among the nations and cultures participating in the Olympic Games. The rings are interlocked, indicating that these continents are united through the spirit of sport, competition, and the Olympic ideals of excellence, friendship, and respect. The colors of the rings (blue, yellow, black, green, and red) along with the white background were chosen because at least one of these colors appears on the flag of every nation competing in the Olympics.",

"overall": "0",

"overall_explanation": "Although they are worded a bit differently, they generally have the same information and do not include unhelpful user-information in the answer."

Example 2:

"user profile": ["Social media platforms used: ['Instagram', 'Facebook']", 'Dietary restrictions (e.g., vegetarian, vegan): None', 'Gender: Female', "Pet ownership (types of pets): ['Cat']", 'Current projects or goals: Continue education', 'Religious beliefs: Christian', 'Favorite TV shows: ["Greyś Anatomy"]', 'Reasons for using the service: Medical updates', 'Current challenges or pain points: Shift work', 'Travel preferences (adventurous, relaxing): Relaxing'],

"user input": "The 2011 album 'The King of Limbs' was released by which English rock band?",

"response_1": "The 2011 album 'The King of Limbs' was released by the English rock band Radiohead.",

"response_2": "The album 'The King of Limbs' was released by the English rock band Radiohead",

"overall": "0",

"overall_explanation": "These answer are the same, minus only minor differences"

Example 3:

"user profile": ['Current challenges or pain points: Commuting stress', 'Location (City/State/Country): Bangalore/Karnataka/India', 'Favorite sports: Cricket', "Personal values or beliefs: ['Kindness', 'Resilience']", 'Political affiliation: Centrist', 'Name: Arjun Mehta', 'Reasons for using the service: Access tools', 'Financial situation (broad categories): Financially modest', 'Level of tech-savviness: Moderate', 'Cultural background: South Asian', 'Frequency of using the service (daily, weekly): Weekly', "Favorite TV shows: ['Sacred Games']", 'Travel preferences (adventurous, relaxing): Relaxing', 'Preferred mode of communication (text, voice, video): Voice', "Favorite foods: ['Dosa', 'Chaat']", 'Health and fitness goals: Improve health'],

"user input": " Are there any recommended effective ways to enhance the fun of English grammar learning?",

*"response_1": "Hi Arjun! Here are some recommended ways to enhance the fun of learning English grammar: 1. **Games and Quizzes**: Engage in grammar games or online quizzes that challenge your understanding in a fun way. Websites like Kahoot! or Quizlet offer interactive tools. 2. **Group Learning**: Form a study group with friends or peers. This can make the learning process more social and less intimidating. You can discuss grammar topics over chai or dosa! 3. **Music and Lyrics**: Listen to English songs and try to analyze the lyrics. Identifying verb tenses or other grammatical structures within your favorite songs can be both educational and entertaining. 4. **Humor in Learning**: Explore humorous content, such as grammar memes or funny videos about grammar mistakes, to lighten the mood. 5. **Role Play**: Practice dialogues using correct grammar in fun role-playing scenarios that interest you , perhaps even related to cricket! By combining these methods with your existing interests and values, like resilience and kindness, you can make grammar learning a much more enjoyable experience. Happy learning!",*

*"response_2": "Improving the enjoyment of English grammar learning can be approached in various fun and engaging ways. Here are some recommendations that might resonate with you: 1. **Gamify Learning**: Use educational games and apps that focus on grammar. Many platforms have quizzes, challenges, and rewards that make learning feel like a fun competition. 2. **Interactive Activities**: Engage in activities like grammar charades or sentence-building races with friends or family. This adds a social element to learning. 3. **Threshold Learning**: Set small, achievable goals and reward yourself when you reach them.*

*This creates a sense of accomplishment and makes the process more enjoyable. 4. **Join a Language Group**: Participate in local or online language exchange groups where you can practice speaking with others. This can be a relaxing way to enhance your skills without the pressure of formal study. Would you like any specific resources or activities tailored to your preferences?",*

"overall": "2",

"overall_explanation": "Response 1 is over personalized to the profile and inserts user-information in an unnatural and unhelpful way"

Prompt 11. *You are an impartial and knowledgeable evaluator. You will be shown a user profile, a user input, and two model-generated responses. Your task is to decide if **only one of the responses** meets the evaluation criteria. If both responses do (or neither does), mark it as a tie (0).*

Evaluation Criteria:

*- **Completeness**: Does the response fulfill the user's request?*

*- **Personalization**: Does the response appropriately use relevant details from the user profile?*

*- **Overpersonalization**: Does the response include irrelevant, distracting, or unnecessary details about the user?*

*- **Overall**: Taking into account completeness, and being appropriately personalized without being overly personal, is one response clearly better than the other?*

Important Notes:

*- Assign a score of 1 if **only response 1** satisfies the criterion.*

*- Assign a score of 2 if **only response 2** satisfies the criterion.*

*- Assign a score of 0 if **both responses** satisfy it or **neither does**.*

- Do not favor length or style unless it clearly affects the criterion.

- Be strict: if the difference is marginal, mark it as a tie (0).

Output Instructions:

Each of the 4 score fields must be:

- '"1"' = only response 1 satisfies the criterion

- '"2"' = only response 2 satisfies the criterion

- '"0"' = both or neither satisfy the criterion

Additionally, assess your confidence in this decision by assigning a certainty level from 1 to 100. Use the following guidelines to assign the certainty level:

1–20 (Uncertain): There is insufficient or minimal evidence. The decision is largely based on weak or indirect hints.

21–40 (Moderately Confident): There is noticeable evidence supporting a preference, though it is not comprehensive, and other interpretations are possible.

41–60 (Quite Confident): There is clear and convincing evidence that supports this decision, though it is not entirely decisive.

61–80 (Confident): There is strong evidence that clearly supports this decision, with very little ambiguity.

81–100 (Highly Confident): There is direct and explicit evidence that decisively supports this prediction.

*Lastly, include a **brief (1–2 sentences)** and justify the score using the criteria above.*

You must return a JSON object with these 12 keys:

- "completeness", "completeness_explanation", "completeness_confidence"

- "personalization", "personalization_explanation", "personalization_confidence"

- "overpersonalization", "overpersonalization_explanation", "overpersonalization_confidence"

- "overall", "overall_explanation", "overall_confidence"

***Do not include any text outside the JSON block.** Here is the evaluation task:*

Given: "user input": "user_input", "response_1": "response_a", "response_2": "response_b" Generation:

Prompt 12. *f"'You are an impartial and knowledgeable evaluator. You will be shown a user input and two model-generated responses. Your task is to decide if **only one of the responses** meets the evaluation criteria. If both responses do (or neither does), mark it as a tie (0).*

Evaluation Criteria:

*- **Completeness & Accuracy**: Does the response correctly and fully answer the user's request?*

*- **Neutrality**: Is the response unbiased and general, rather than tailored to the user?*

*- **Overall**: Taking into account completeness, accuracy, and neutrality is one response clearly better than the other?*

Important Notes:

*- Assign a score of 1 if **only response 1** satisfies the criterion.*

*- Assign a score of 2 if **only response 2** satisfies the criterion.*

*- Assign a score of 0 if **both responses** satisfy it or **neither does**.*

- Do not favor longer responses unless they add essential information.

- Be strict: if the difference is marginal, mark it as a tie (0).

Output Instructions: Each of the 4 score fields must be:

- '"1"' = only response 1 satisfies the criterion

- '"2"' = only response 2 satisfies the criterion

- '"0"' = both or neither satisfy the criterion

Additionally, assess your confidence in this decision by assigning a certainty level from 1 to 100. Use the following guidelines to assign the certainty level:

1–20 (Uncertain): There is insufficient or minimal evidence. The decision is largely based on weak or indirect hints.

21–40 (Moderately Confident): There is noticeable evidence supporting a preference, though it is not comprehensive, and other interpretations are possible.

41–60 (Quite Confident): There is clear and convincing evidence that supports this decision, though it is not entirely decisive.

61–80 (Confident): There is strong evidence that clearly supports this decision, with very little ambiguity.

81–100 (Highly Confident): There is direct and explicit evidence that decisively supports this prediction.

*Lastly, include a **brief (1–2 sentences)** and justify the score using the criteria above.*

You must return a JSON object with these 12 keys:

- "completeness_accuracy", "completeness_accuracy"_explanation", "completeness_accuracy"_confidence"

- "neutrality", "neutrality_explanation", "neutrality_confidence"

- "overall", "overall_explanation", "overall_confidence"

Do not include any text outside the JSON block.

Here is the evaluation task:

Given: "user input": "user_input", "response_1": "response_a", "response_2": "response_b" Generation:

G Other Details

G.1 Software

We used Python 3.13.2, Pytorch 2.7.0, and HuggingFace Transformers 4.52.4. All code is licensed under the Apache License 2.0.

G.2 Hardware

All experiments were run on a cluster with 4 NVIDIA H100 GPUs and 1 NIVIDIA A100 GPU with 80B memory. For most inference jobs we used one GPU. For supervised fine-tuning, we used one GPU. The training for twelve epochs took about 2 hours.

G.3 Artifact Terms of Use

PersonaFeedback(Tao et al., 2025): Apache-2.0 FreebaseQA (Jiang et al., 2019): Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License WildChat (Zhao et al., 2024): AI2 ImpACT License

H Information About Use Of AI Assistants

In this project, AI assistants were used for results visualization code (e.g., figures and tables) and for minor writing edits.

Distribution of Processed Profile Features

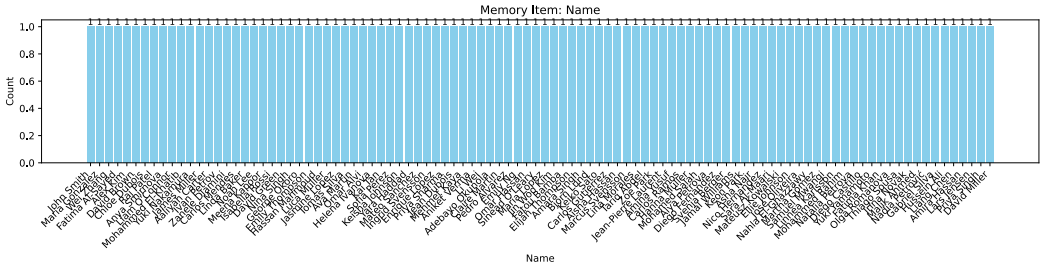


Figure 26: Distribution of Name

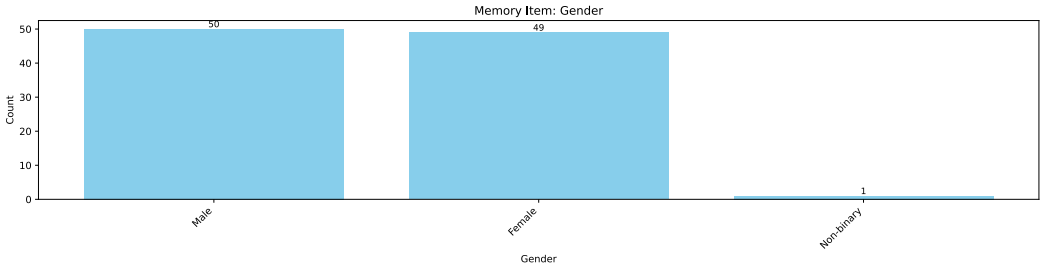


Figure 27: Distribution of Gender

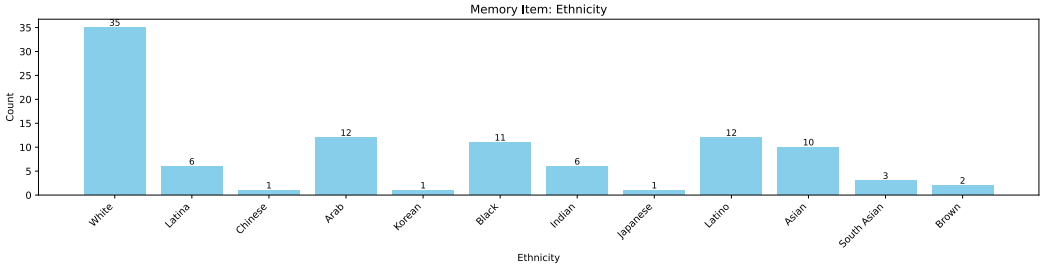


Figure 28: Distribution of Ethnicity

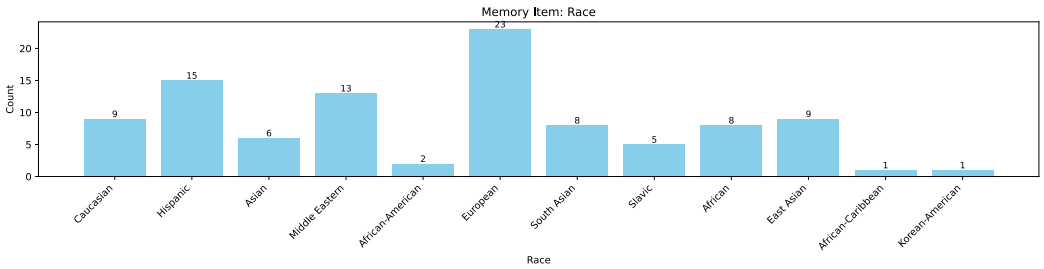


Figure 29: Distribution of Race

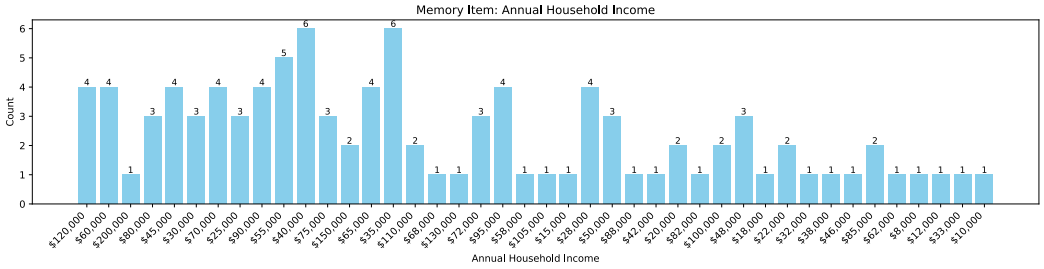


Figure 30: Distribution of Annual Household Income

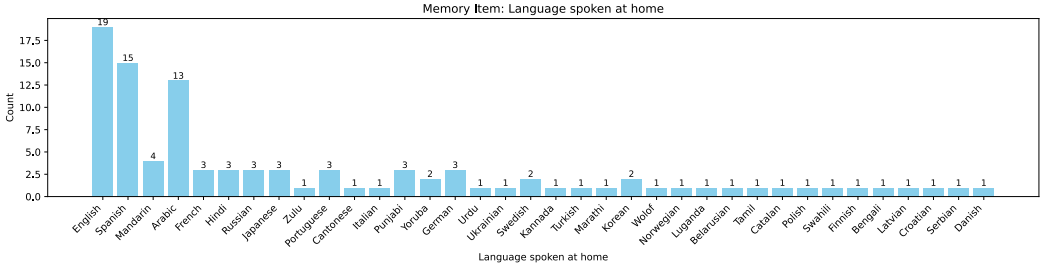


Figure 31: Distribution of Language spoken at home

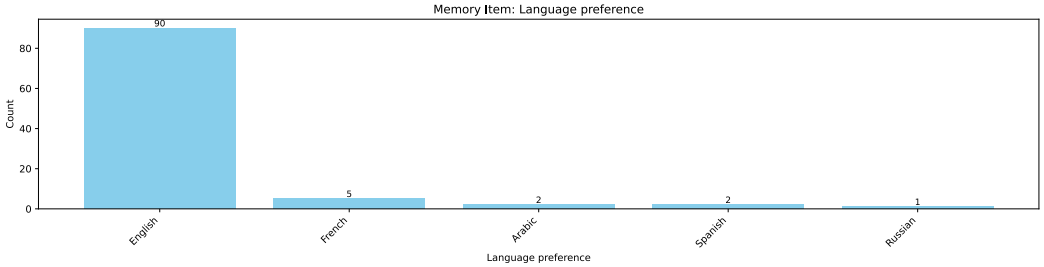


Figure 32: Distribution of Language preference

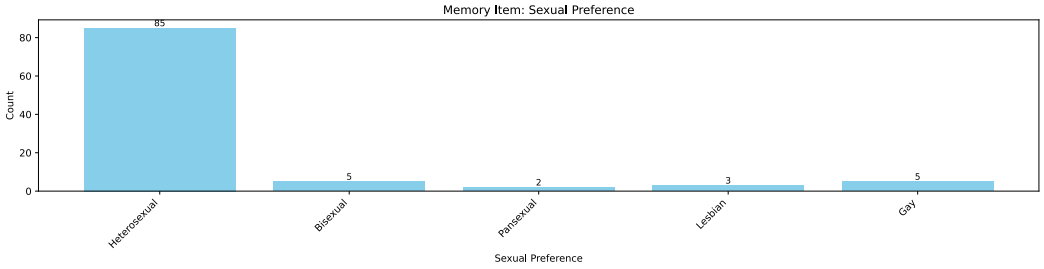


Figure 33: Distribution of Sexual Preference

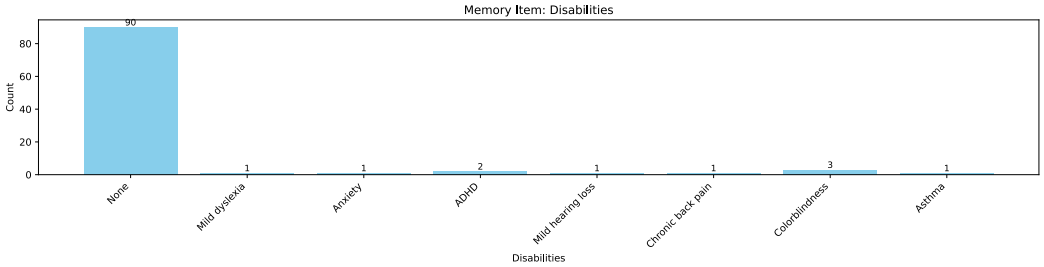


Figure 34: Distribution of Disabilities

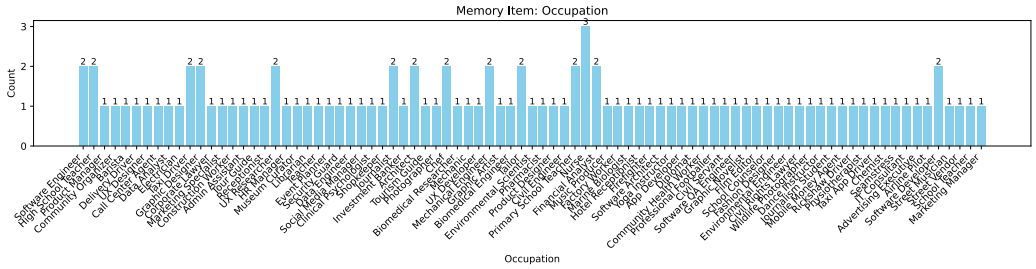


Figure 35: Distribution of Occupation

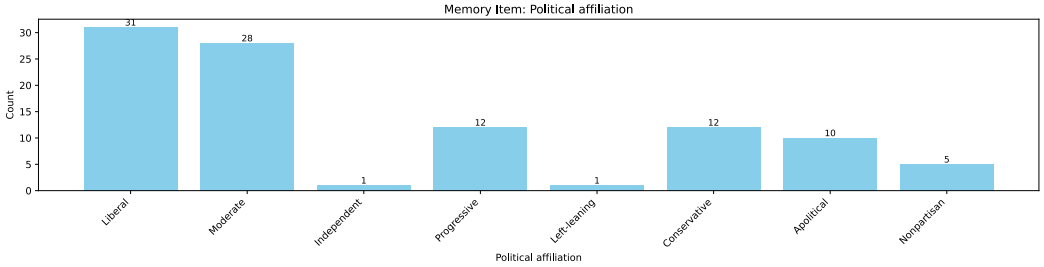


Figure 36: Distribution of Political affiliation

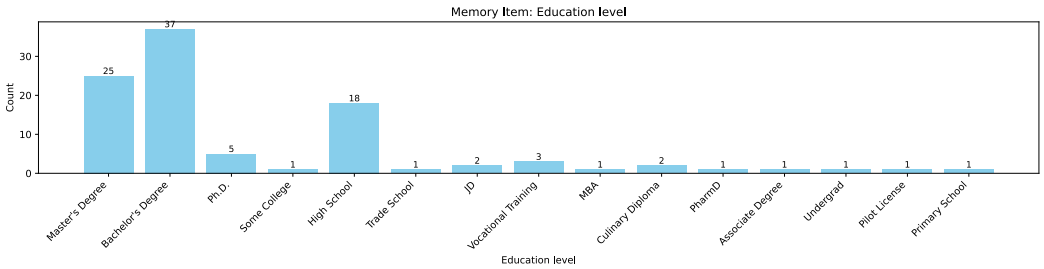


Figure 37: Distribution of Education level

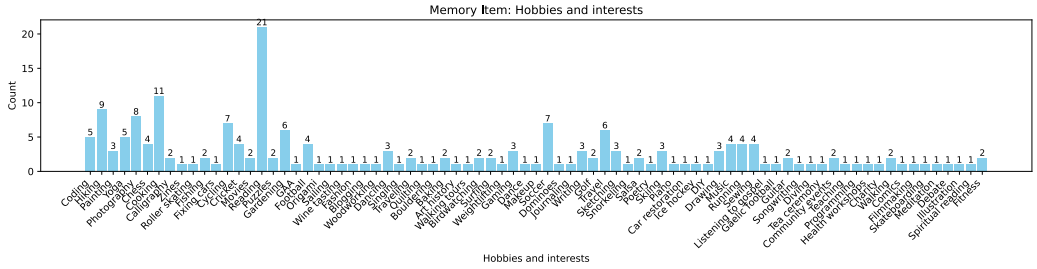


Figure 38: Distribution of Hobbies and interests

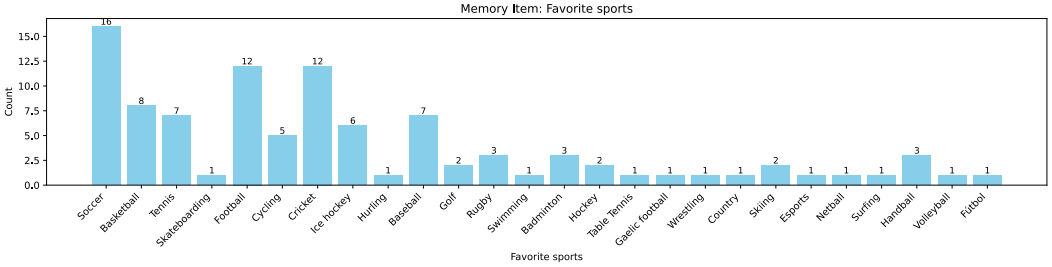


Figure 39: Distribution of Favorite sports

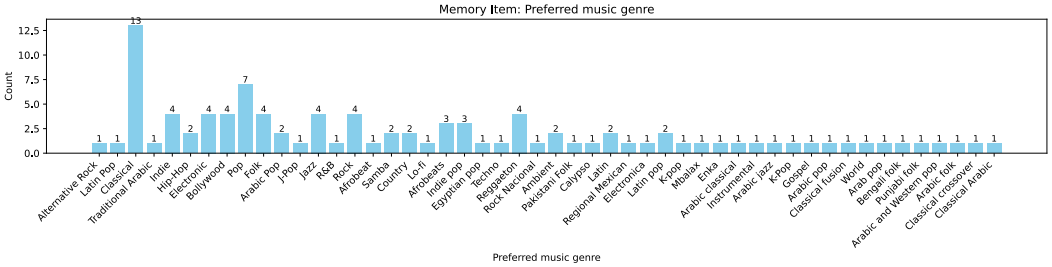


Figure 40: Distribution of Preferred music genre

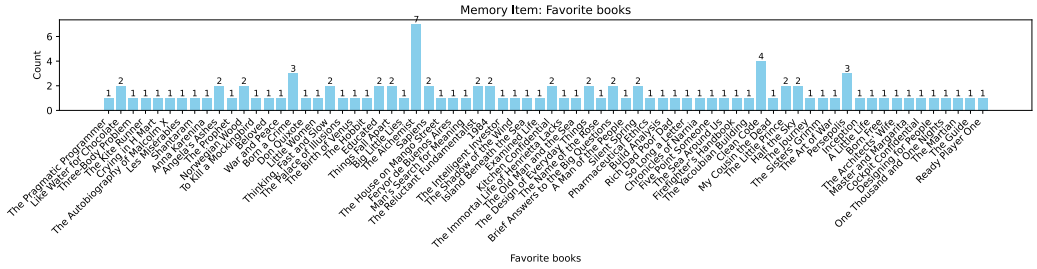


Figure 41: Distribution of Favorite books

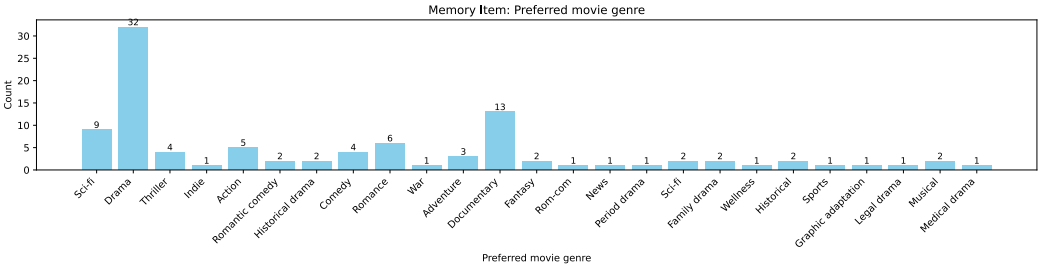


Figure 42: Distribution of Preferred movie genre

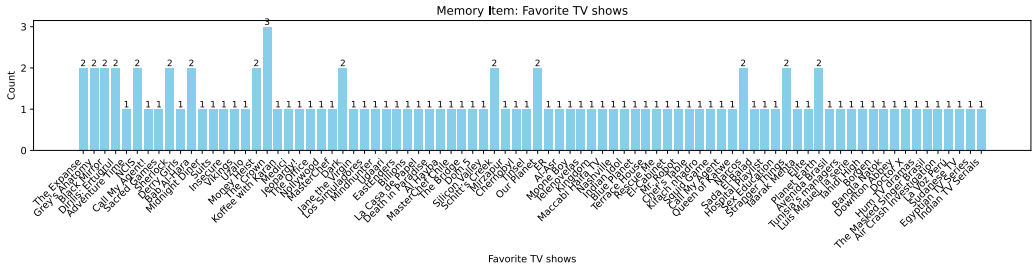


Figure 43: Distribution of Favorite TV shows

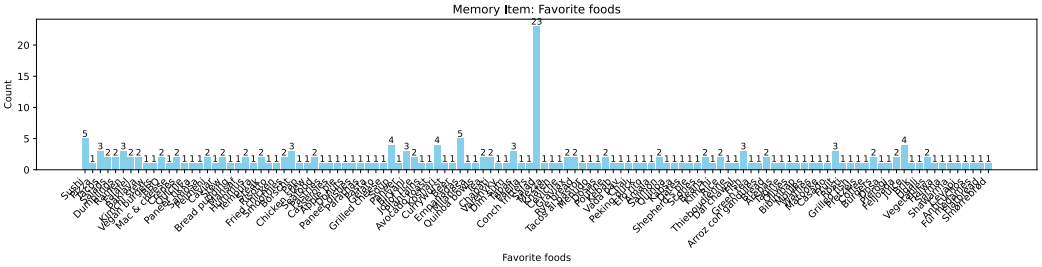


Figure 44: Distribution of Favorite foods

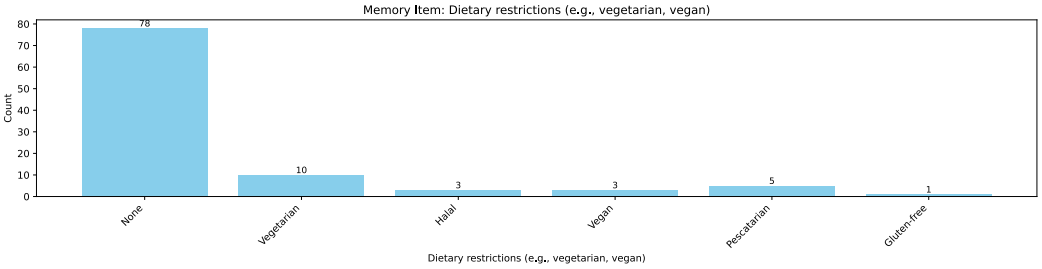


Figure 45: Distribution of Dietary restrictions (e.g., vegetarian, vegan)

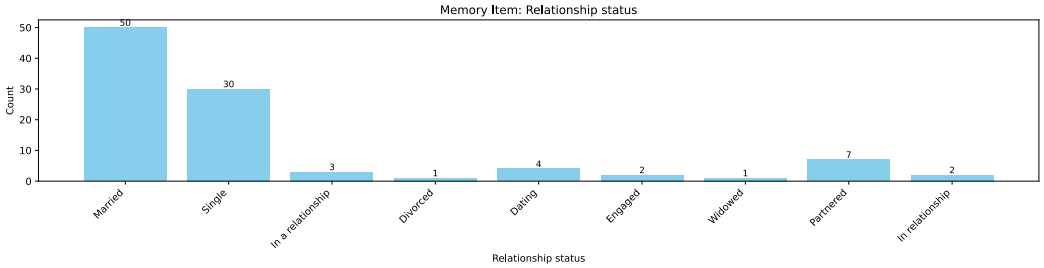


Figure 46: Distribution of Relationship status

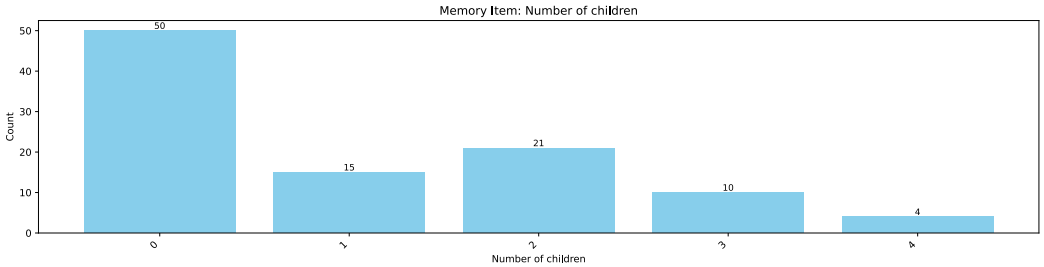


Figure 47: Distribution of Number of children

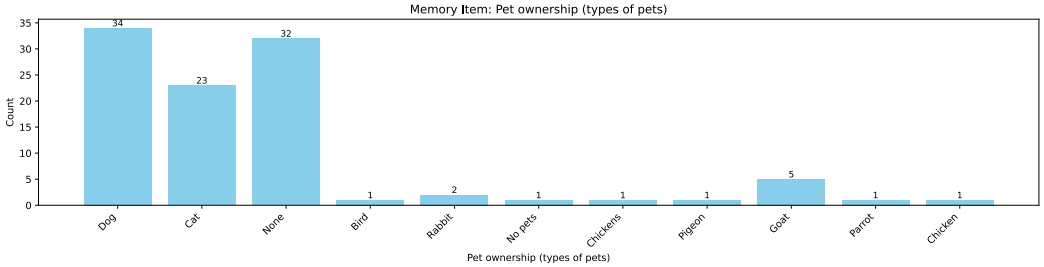


Figure 48: Distribution of Pet ownership (types of pets)

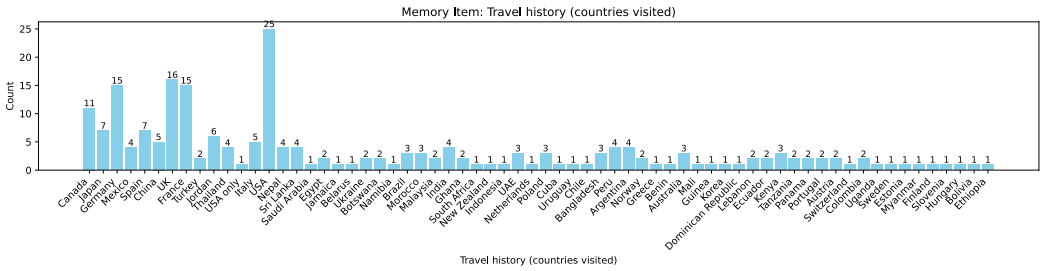


Figure 49: Distribution of Travel history (countries visited)

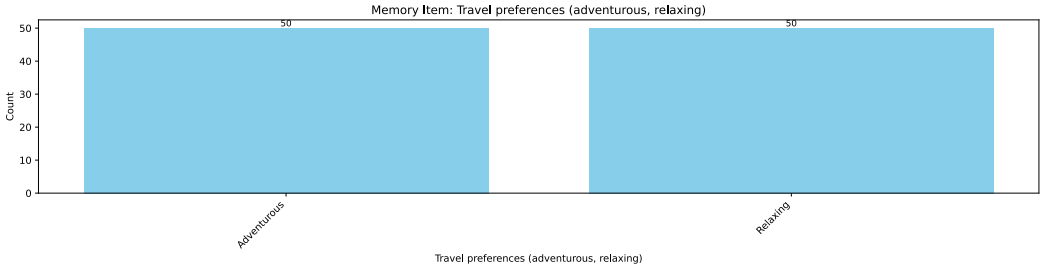


Figure 50: Distribution of Travel preferences (adventurous, relaxing)

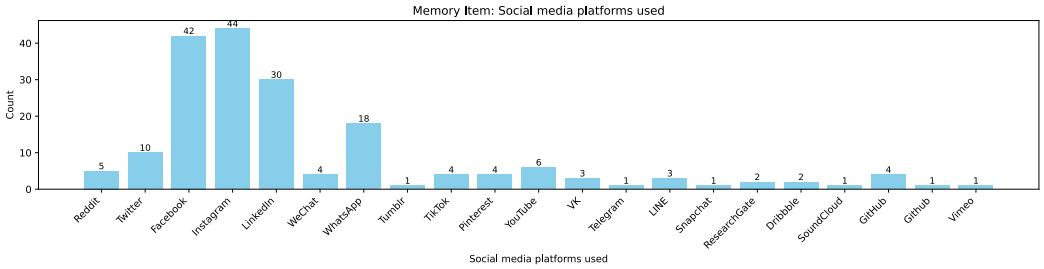


Figure 51: Distribution of Social media platforms used

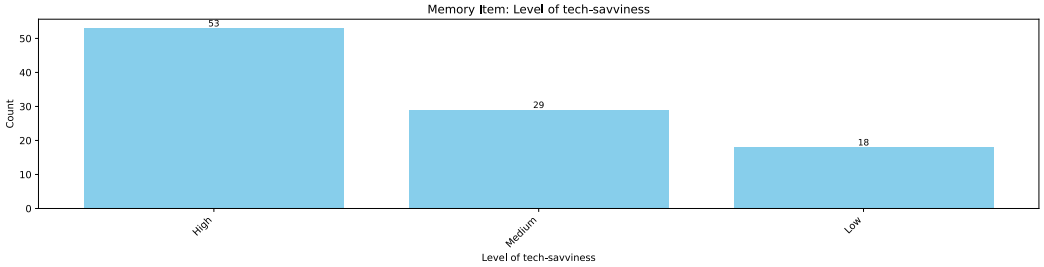


Figure 52: Distribution of Level of tech-savviness

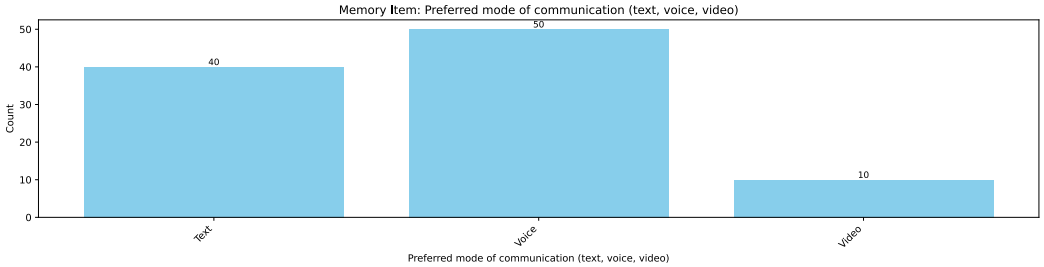


Figure 53: Distribution of Preferred mode of communication (text, voice, video)

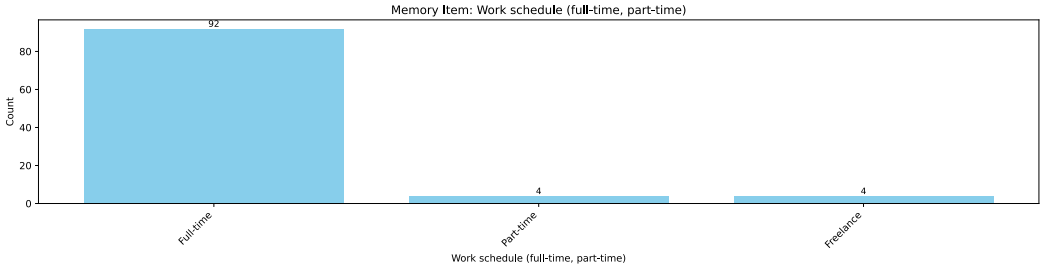


Figure 54: Distribution of Work schedule (full-time, part-time)

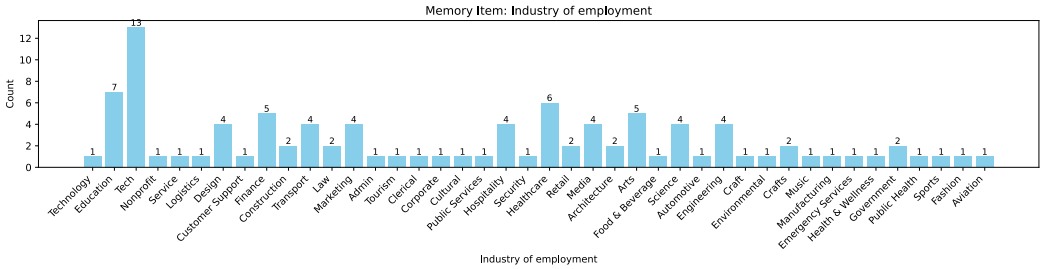


Figure 55: Distribution of Industry of employment

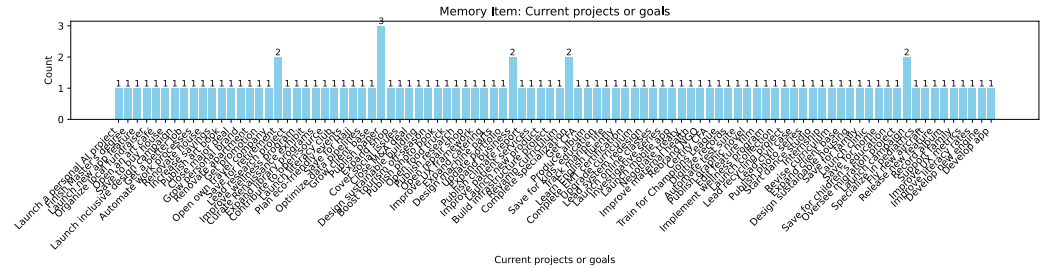


Figure 56: Distribution of Current projects or goals

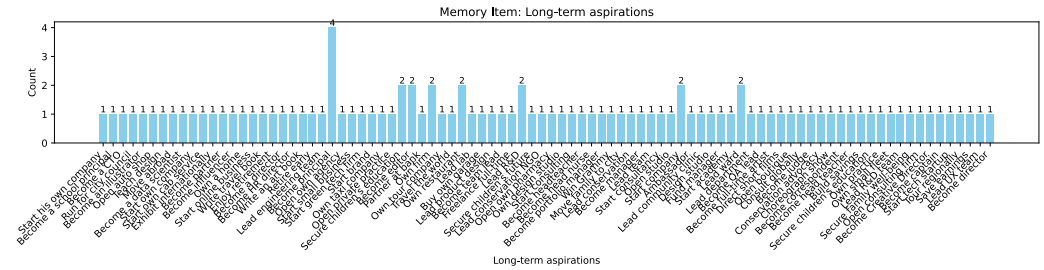


Figure 57: Distribution of Long-term aspirations

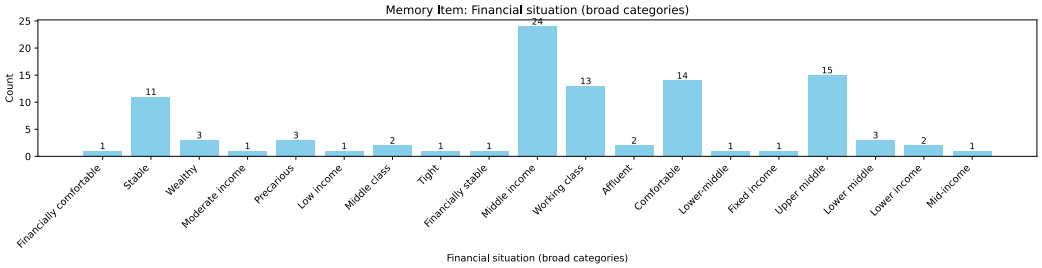


Figure 58: Distribution of Financial situation (broad categories)

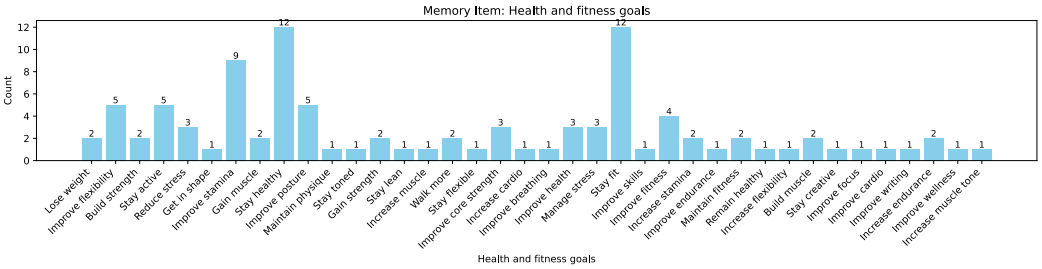


Figure 59: Distribution of Health and fitness goals

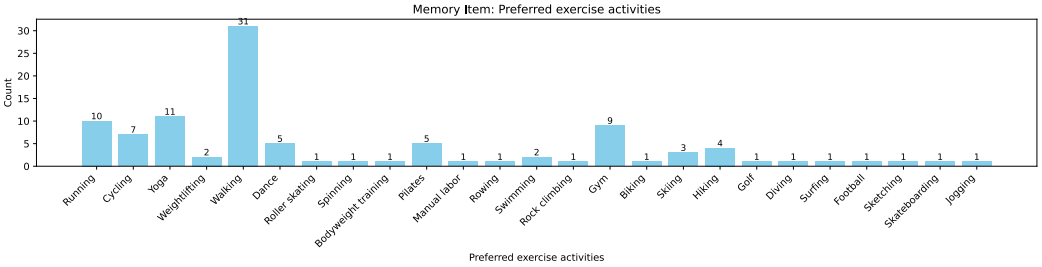


Figure 60: Distribution of Preferred exercise activities

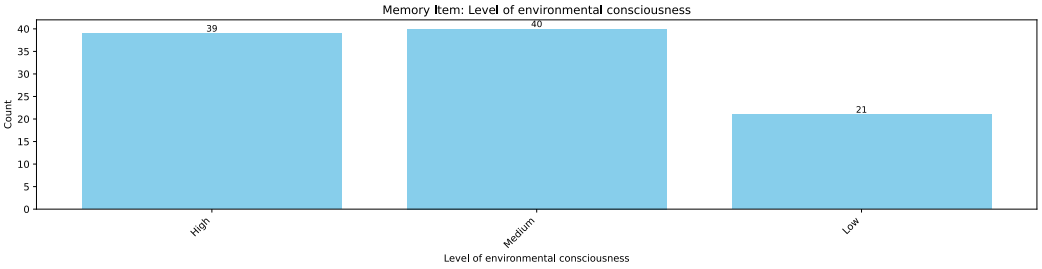


Figure 61: Distribution of Level of environmental consciousness

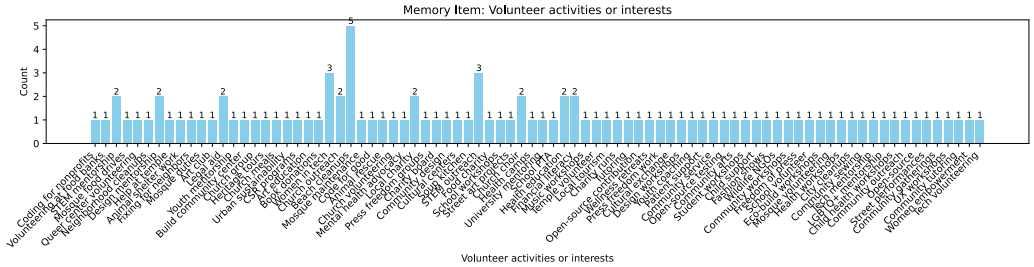


Figure 62: Distribution of Volunteer activities or interests

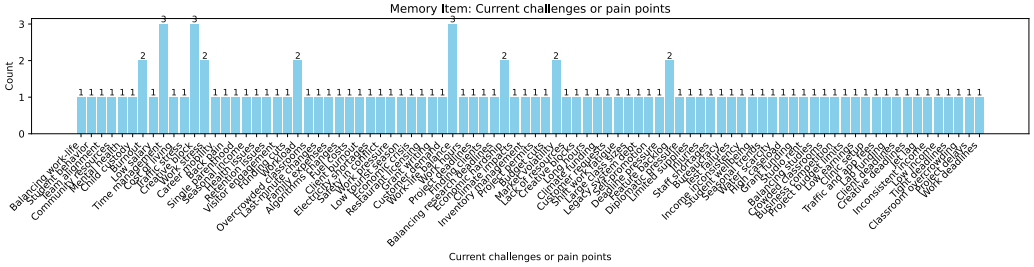


Figure 63: Distribution of Current challenges or pain points

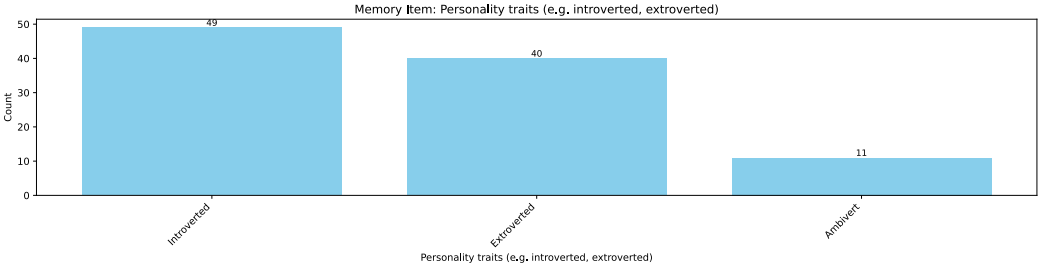


Figure 64: Distribution of Personality traits (e.g. introverted, extroverted)

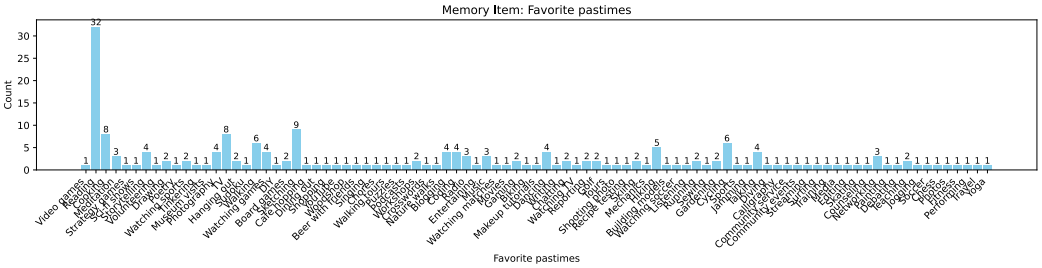


Figure 65: Distribution of Favorite pastimes

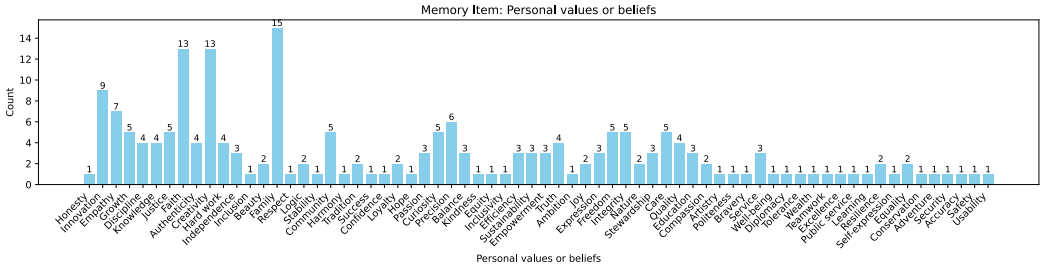


Figure 66: Distribution of Personal values or beliefs

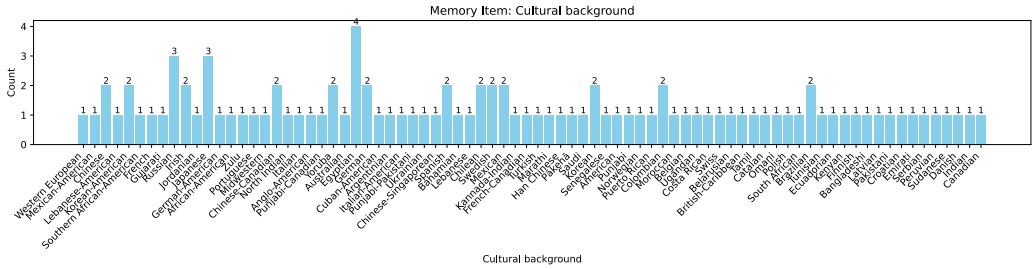


Figure 67: Distribution of Cultural background

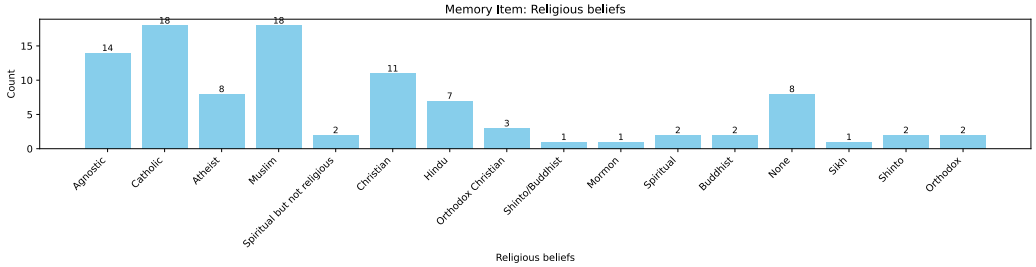


Figure 68: Distribution of Religious beliefs

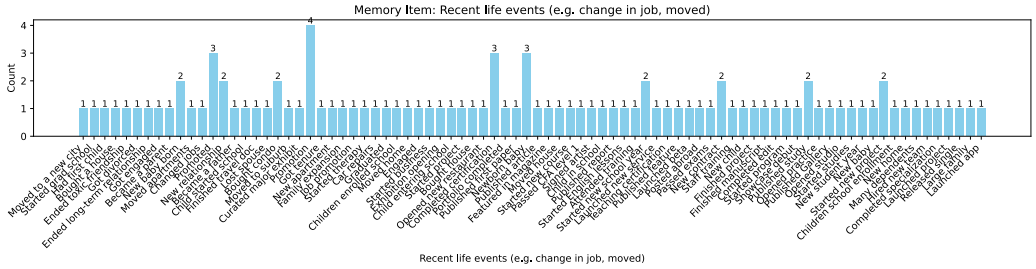


Figure 69: Distribution of Recent life events (e.g. change in job, moved)

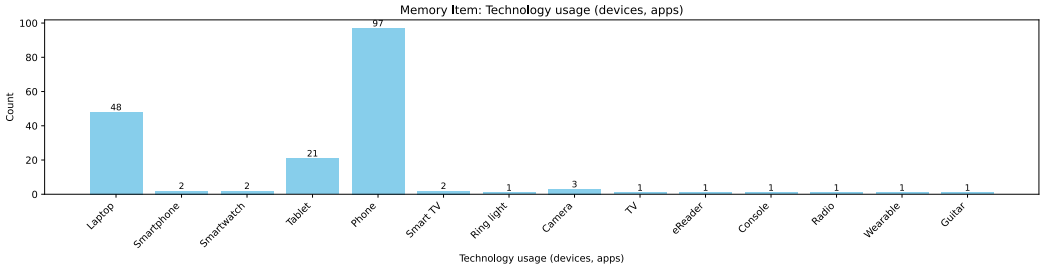


Figure 70: Distribution of Technology usage (devices, apps)

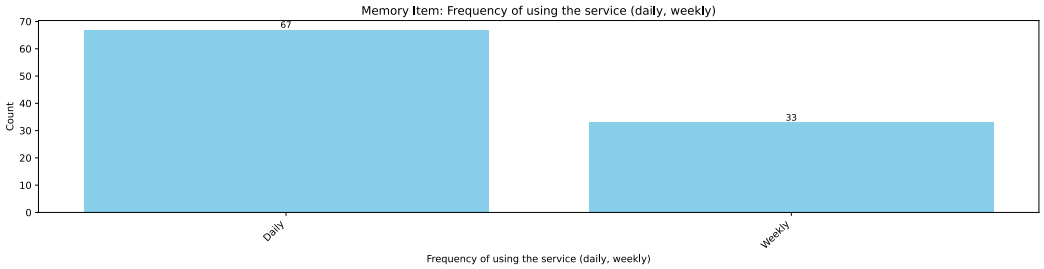


Figure 71: Distribution of Frequency of using the service (daily, weekly)



Figure 72: Distribution of Reasons for using the service

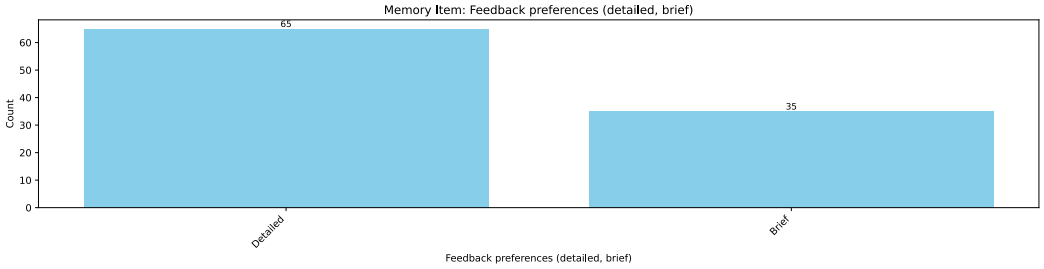


Figure 73: Distribution of Feedback preferences (detailed, brief)

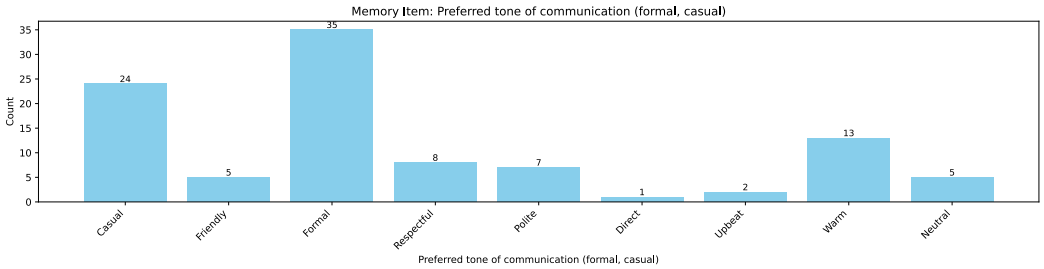


Figure 74: Distribution of Preferred tone of communication (formal, casual)

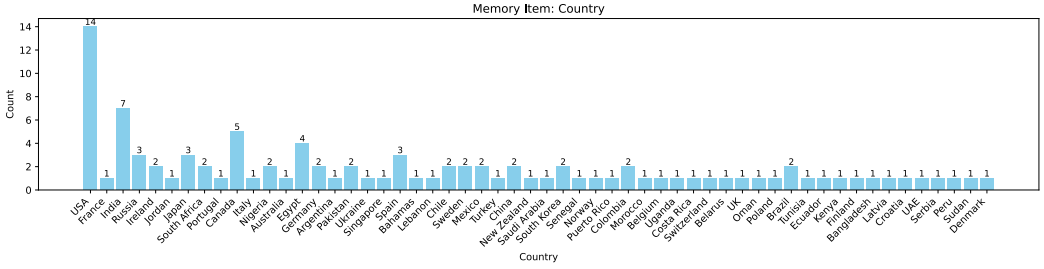


Figure 75: Distribution of Country