User Behavior Prediction as a Generic, Robust, Scalable, and Low-Cost Evaluation Strategy for Estimating Generalization in LLMs

Sougata Saha* **and Monojit Choudhury***
Mohamed bin Zayed University of Artificial Intelligence,
{sougata.saha, monojit.choudhury}@mbzuai.ac.ae

Abstract

Measuring the generalization ability of Large Language Models (LLMs) is challenging due to data contamination. As models grow and computation becomes cheaper, ensuring tasks and test cases are unseen during training phases will become nearly impossible. We argue that knowledge-retrieval and reasoning tasks are not ideal for measuring generalization, as LLMs are not trained for specific tasks. Instead, we propose *user behavior prediction*, also a key aspect of *personalization*, as a theoretically sound, scalable, and robust alternative. We introduce a novel framework for this approach and test it on movie and music recommendation datasets for GPT-4o, GPT-4o-mini, and Llama-3.1-8B-Instruct. Results align with our framework's predictions, showing GPT-4o outperforms GPT-4o-mini and Llama, though all models have much room for improvement, especially Llama.

1 Introduction

"The central challenge in machine learning is that we must perform well on new, previously unseen inputs—not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called generalization" - Goodfellow (2016).

Utilizing large amounts of data for training, large language models (LLMs) have achieved state-of-the-art performance on existing and new evaluation benchmarks, demonstrating remarkable capabilities over varied use cases. However, increasing training data makes models susceptible to data contamination issues, where the model has been exposed to the test data during training. For example, GPT 3.0 (Brown et al., 2020) was exposed to portions of test data, conflating its test scores. Such issues question the efficacy of existing evaluation benchmarks in measuring LLMs' generalizability. Hence, during evaluation, the model can recall from its memory instead of learning the underlying pattern, resulting in conflated performance on existing datasets and a false sense of *generalization*. Distinguishing this memorization capacity from learning transferable principles is a key challenge in measuring generalization in foundational models (Chu et al., 2025). One might argue that memorization, although a weaker form of generalization, suffices as long as LLMs perform well on tasks of practical importance. However, since the space of all problems is unknown, we do not know in what situations a model might fail. Also, since the world is dynamic, continuous memorization is impractical, which generalization addresses. Thus, not understanding models' generalization capabilities hampers their reliability.

Although several popular frameworks and benchmarks (Chang et al., 2024) for evaluating LLMs' knowledge, reasoning, alignment, and safety properties exist (Guo et al., 2023), it is unclear how much of these properties are due to generalization and how much can be achieved only through memorization. Also, as LLMs' compute capacity increase, it becomes more difficult to create challenging evaluation benchmarks free from novel forms of contamination, such as task contamination (Li and Flanigan, 2024), leading to an uptake of complex evaluation benchmarks (He et al., 2024) whose practical utility is unknown (Zhou et al., 2023). Hence, we ask *what should be an ideal strategy for evaluating LLMs' generalization?* The solution must be robust to data contamination; it should be dynamic, and time and cost-efficient, and it must ensure the availability of distinct test sets, even if models utilize *all* available data for training.

Since most available and high-quality LLM training data are human-generated, they capture human behavior in a context at a time. Thus, although LLMs are trained as next-word predictors, they should essentially learn the intrinsic task of behav-

*Both authors contributed equally to this paper.

Findings of the Association for Computational Linguistics: ACL 2025, pages 11047–11065

ior prediction from context. To this end, we propose *user behavior prediction* or *personalization* as a simple yet robust strategy for measuring LLM's generalization over a wide, potentially an infinite, range of capabilities. Although existing studies measure LLMs' personalization capabilities (Lin et al., 2023; Zhao et al., 2024; Wu et al., 2024; Dai et al., 2023; Liu et al., 2023), deviating from the standard definition, we propose a novel framework that uses personalization to measure generalization. By re-purposing existing resources, our method presents a robust and cost-effective measure of generalization in practical settings, which presents a perspective shift in how we utilize them.

In the following section, we argue our position of using personalization to measure generalization and empirically demonstrate an entropy-based framework for measuring generalization. Defining generalization as a model's capacity to follow the actual entropy change with varying context, we first present a statistical framework that evaluates the capacity of existing tasks as good generalization benchmarks and then use recommendation systems as a use case to measure the generalization capabilities of GPT-4o, GPT-4o-mini, and Llama-3.1-8B-Instruct against a baseline. Overall, our contributions are summarized below:

1. We propose user behavior prediction, a key aspect of personalization, as a theoretically sound, scalable, and robust alternative to measuring generalization.

2. We empirically test our hypothesis using an entropy-based framework and present results for movie and music recommendations using GPT-4o, GPT-4o-mini, and Llama-3.1-8B-Instruct.

3. We discuss the implications of our findings in enabling generalization.

2 Generalization in the era of LLMs

In essence, generalization is a model's ability to perform well on unseen test data, given that the test set measures the same task during model training, indicating the efficacy of the model in learning the underlying patterns in the training data without overfitting or underfitting, attaining a trade-off between bias and variance (Bishop and Nasrabadi, 2006). LLMs, conceived as general-purpose models, are expected to follow instructions in natural language and perform well on varied tasks, emulating human-like behavior. However, current task-centric evaluation schemes fail to measure the generalization capacity of models holistically, often leading to conflated results. This deviation is primarily due to the following factors.

2.1 Issues of Task-Centric Evaluation

Since training LLMs involves online data, it is crucial to understand the nature of such data. Almost all of the available online data pertain to human behavior. A data point is a user's behavior in a specific context in time, where the context (Zimmermann et al., 2007; Bazire and Brézillon, 2005) characterizes the situation and is the cumulative aggregate of all user behaviors leading to that time. Any training dataset embodies such knowledge and patterns and represents human behavior across time. Hence, although LLM training involves the task of next-word prediction, they are essentially trained on the task of user behavior prediction from the context. Thus, they should be evaluated on similar tasks to meaningfully gauge their generalization capabilities, which task-centric evaluation approaches fail to measure holistically.

Also, unlike other statistical models, LLMs require sizable data for training, which has grown with time. However, since the growth of the total stock of public human text data is asymptotic, LLMs are projected to utilize all available data for training between 2026 and 2032 or even earlier (Villalobos et al., 2022, 2024), making it difficult to guarantee that the test sets are unseen during training. Although most open-weight LLMs utilize publicly available online text data during pre-training, the exact data splits used during training are unknown. Also, the data used in the fine-tuning and alignment phases is usually private, making it hard to gauge data and task contamination.

Nearly exhausting all available data for training, the probability of data contamination is high in new test sets, which even synthetic approaches to data creation will not mitigate. Hence, we need evaluation frameworks that are novel and free from these issues. Instead of investing in newer datasets, we propose a perspective shift. We propose a framework for measuring generalization by repurposing existing personalization benchmarks, which can be a robust test for generalization, as we shall see. Prior to that lets introduce a formal description of training data.

2.2 A Formal Description of Training Data

Let $\mathcal{U}$ denote the set of all online users, where $\mathcal{U} = \{u_1...u_n\}$. Let c_t denote the context at any given time t, which is an aggregate of all user behavior till time $t-1$. Let $\mathcal{B}^u = \{b_1^u, ..., b_t^u\}$ denote all the behavior of user u till time t. A data point is a user's behavior b_t^u to the context c_t at time t, where each user is a function that maps the context to their behavior at a point in time. Psychologists, anthropologists, and linguists often cluster human behavior by variables such as demography. Such variables define the group's behavior and preferences across some dimensions. Adilazuarda et al. (2024) formally terms such features as demographic *proxies* of culture, which capture the differences of user groups across dimensions termed semantic proxies (Thompson et al., 2020). Here, we combinedly refer to these factors as *proxies* and can represent any documented grouping such as geodemography, or undocumented groupings such as *dog lovers*. Let $\delta = \{\delta_1...\delta_k\}$ represent the set of all relevant user proxies following a distribution p_δ, where $\delta_j \in \delta$ is a grouping of users $\mathcal{U}$. A dataset $\mathcal{D}$ is the set of all triples of contexts, user proxies, and their behaviors, across time. We will refer to this framework in the next sections.

3 Proposed Method

"It is not difficult to devise a paper machine which will play a not very bad game of chess...Are there imaginable digital computers which would do well in the imitation game?" -Turing (1948).

Alan Turing observed that a problem is easier when there is an end goal. He argued that machines that can mimic humans in dynamic scenarios are much more intelligent than machines that are good at universal tasks such as playing the game of chess, where the rules of the game are already known (Turing, 1950). Ludwig Wittgenstein made a much more fundamental observation about language since its rules constantly evolve. Language, as a mode of communication, is meaningful only in the context of the situation, which factors in users and their surroundings. Consider the famous "builder's language" thought experiment Wittgenstein (1953), where he depicted language as a communication tool in the context of social activity between builder A and assistant B. Builder A is building with blocks, pillars, slabs, and beams. Builder B has to pass the stones in the order specified by A. They use a language consisting of the words "block," "pillar," "slab," and "beam." Builder A calls out an item that Builder B brings. Hence, the builders developed their pragmatic language using only four words, creating what Wittgenstein called a "complete primitive language." The words lose their pragmatic meaning without the context and the builders.

In the current context of LLMs, the ability to perform complex reasoning tasks, such as writing code or solving math problems, although difficult for many humans, are shallower measures of intelligence and generalization since they are universal tasks. However, the ability to solve the same task, mimicking a specific person, is much more challenging than solving the task like any person. Extending Turing's definition, we argue that any machine that can mimic individual users from its training data is the most robust test of intelligence and the strongest measure of generalization. Hence, we propose measuring a model's capacity for personalization as a robust test of its generalization.

3.1 Personalization as a Test for Generalization

With the objective of delivering relevant information to an individual or a group of individuals (Kim, 2002), personalization broadly means tailoring something for an individual without their active participation (Fan and Poole, 2006; Vesanen, 2007). Unlike customization, where individual users are actively involved in tailoring the outcome by specifying their preferences, personalization is without the user's active control and usually involves passively understanding user preferences from their actions (Sundar and Marathe, 2010). Considering personalization as a mode of individuation, Lury and Day (2019) defines it as a recursive approach that "involves forms of de- and re-aggregating, in which a variety of contexts are continually included and excluded" to determine the best possible group affiliation of users. Hence, personalization is a pathway that starts with an initial broad assumption about a user's background, which is constantly refined over time based on their behavior until their optimal preferences are determined (Schmitt, 1999; Dhar and Wertenbroch, 2000; Hanley et al., 2006; Droe, 2006; Wilken et al., 2011; Rogers et al., 2014; Tahmasbi et al., 2018).

The core tenet of digital personalization is that the underlying algorithm should better understand the user's background and preferences with more interactions to facilitate delivering better-

personalized content over time. At the risk of anthropomorphizing, LLMs' generalizability is the ability to emulate human-like behavior in real-world applications, which requires understanding human behavior in practical settings. Statistically, it involves generalizing past behavior patterns to predict future behavior, thus embodying Turing's philosophy that intelligent machines should be capable of mimicking human behavior. We argue that any model (algorithm) that can perform well in the personalization task is more generalizable. In the following subsection, we formalize this notion of generalization.

3.2 A Statistical Framework for Levels of Generalization

As discussed in Section 2.2, since LLMs learn from the entirety of human behavior data, any task ultimately involves group behavior prediction given a context, where the proxy represents the size of the user group, which might vary from an individual to the entire human population. Hence, any proxy that can reduce the entropy of behavior prediction more than any random subset of similar size is a meaningful proxy. A model's capability to use such proxies to predict the outcome is a test of its generalizability.

Given a task T, let $\mathcal{B}^T \in \{b_1 ... b_n\}$ represent the set of all behaviors from all users across time. Let $p(b_i|\delta_j)$ represent the probability of a behavior $b_i \in \mathcal{B}^T$ for proxy $\delta_j \in \delta$. The generalization capacity of a model θ for the task is inversely proportional to the expected difference between the cross-entropy $H(\hat{\mathcal{B}}^T_{\delta_j})$ and the true entropy $H(\mathcal{B}^T_{\delta_j})$ for each proxy δ_j, as defined below:

$$H(\mathcal{B}^T_{\delta_j}) = -\sum_{b_i \in \mathcal{B}^T} p(b_i|\delta_j) \log p(b_i|\delta_j) \quad (1)$$

$$H(\hat{\mathcal{B}}^T_{\delta_j}) = -\sum_{b_i \in \mathcal{B}^T} p(b_i|\delta_j) \log p(b_i|\delta_j, \theta) \quad (2)$$

Depending on the dependence of the behavior and the proxy, the notional complexity of a task is a continuum from weak to strong as below:

Weakest Case ($\mathcal{B}^T \perp\!\!\!\perp \mathcal{U}|c_t$): Tasks where the behavior depends on the context c_t and is independent of individual users $\mathcal{U}$ or their proxies δ, making them notionally less complex. Most knowledge and reasoning-based tasks such as MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), GLUE (Wang, 2018), etc, are examples of such kinds, which measure universal patterns independent of proxies, hence notionally weakest test sets for generalizability.

Average Case ($\mathcal{B}^T \perp\!\!\!\perp \mathcal{U}|c_t, \delta_j$, where $\delta_j \sim p_\delta$): Tasks where the outcome is independent of individual users but dependent on their proxies and the context are notionally more complex. For example, cultural evaluation benchmarks that require group-specific reasoning are stronger test sets for generalizability (Li et al., 2024b; AlKhamissi et al., 2024; Nadeem et al., 2021; Nangia et al., 2020; Wan et al., 2023; Jha et al., 2023; Li et al., 2024a; Cao et al., 2023; Tanmay et al., 2023; Rao et al., 2023; Kovač et al., 2023).

Strongest Case ($\mathcal{B}^T \not\perp\!\!\!\perp \mathcal{U}|c_t, \delta_j$, where $\delta_j \sim p_\delta$): Tasks where the outcome depends on individual users are notionally most complex. For example, user-specific tasks, such as item recommendation, necessitate reasoning from a user's perspective and are notionally more complex and best evaluation benchmarks for generalizability (Nagarnaik and Thomas, 2015; Ko et al., 2022).

3.3 Hypothesis

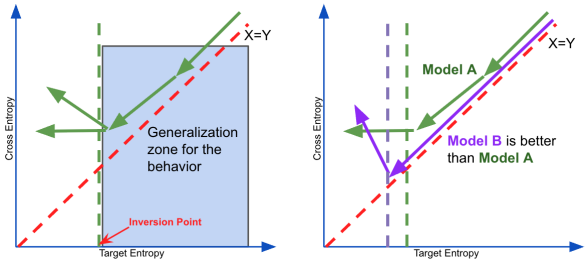

Figure 1: (Left) Hypothesized Behavior. (Right) Hypothesized Model Comparisons.

As depicted in Figure 1 (left), ideally, the distribution of the target $H(\mathcal{B}^T_{\delta_i})$ and cross-entropy $H(\hat{\mathcal{B}}^T_{\delta_i})$ should be equal. When for a model tested under different cases the points ($H(\mathcal{B}^T_{\delta_i})$, $H(\hat{\mathcal{B}}^T_{\delta_i})$) are plotted on a graph, in the ideal case of generalization, $H(\mathcal{B}^T_{\delta_i}) = H(\hat{\mathcal{B}}^T_{\delta_i})$. Or in other words, the points should lie on the X=Y line. However, in reality, we expect $H(\mathcal{B}^T_{\delta_i}) < H(\hat{\mathcal{B}}^T_{\delta_i})$. This gap is expected to be small when $H(\mathcal{B}^T_{\delta_i})$ is high (i.e., the average case of generalization when only proxies are used to predict behavior). We expect, therefore, the plot to follow the X=Y line for large X, but then flatten out or even rise for lower X. The point at which this inversion of behavior happens is the point when the model can no longer generalize to specific users' or groups' behavior. As depicted

in Figure 1 (right), we hypothesize the inversion point to change across models, where a lower inversion point indicates a better generalizable model. Where do current LLMs lie in this framework of generalization? We put our hypothesized statistical framework to test in the next sections.

4 Measuring Generalization via Personalization

Recommendation systems, at their core, are personalization engines. They are predictors of a user's future behavior based on some observed past behavior. We test our proposed statistical framework in Section 3.2 and experiment with movie and music recommendations, where the task is to recommend a list of N items based on the user's history.

4.1 Dataset and Preprocessing

We experiment with the MovieLens[1] (Harper and Konstan, 2015) and last.fm[2] (Celma, 2010) datasets since both these datasets contain demographic information and are widely used in recommendation systems literature. Also, since recommendation datasets are known to be very sparse, we preprocess both datasets to reduce sparsity. Collected by GroupLens Research, the MovieLens dataset (movies dataset) contains 1 million ratings of approximately 3,900 movies made by 6,040 MovieLens users, along with their demographic information such as gender, age, and occupation. This is a well maintained dataset and extensively used in recommendation systems (Goyani and Chaurasiya, 2020) literature. As a post processing step, we remove users with occupation listed as 'others' and restrict to demographic groups (combination of age, gender and occupation) containing at least 30 users.

Containing the music listening habits of nearly 1,000 users, along with their demographic information such as gender, age, and country, the last.fm dataset (music dataset) is widely used in music recommendation literature (Schedl, 2016). It has also impacted research pertaining to music and mood (Çano et al., 2017), and other cultural and behavioral studies (Chen et al., 2010; Putzke et al., 2014). However, the number of users from each country follows a long-tailed distribution. Hence, as a post processing step, we derive the continent proxy based on country and restrict to users from Europe, North America, South America, and United Kingdom who have at least 5,000 interactions.

4.2 Setup

Experiments: We set up the experiments as a behavior prediction task and experiment with the following three levels of proxies:

(A) Demography and history: We provide demography D and prior interactions h as the proxy in the context. For example, "recommend 10 movies from the candidates C for 25-30-year-old self-employed females who have watched the Titanic and Pather Panchali." This is equivalent to the **Average Case** of measuring generalization, as discussed in Section 3.2.

(B) Only history: We only provide prior interactions as the proxy in the context. The intended subgroup is users who have interacted with at least 60% of the items in the history. For example, "recommend 10 music from the candidates for users who have listened to Bohemian Rhapsody and Comfortably Numb." This is equivalent to the **Strongest Case** of measuring generalization.

(C) Only demography: This is the default and the **Weakest Case** of measuring generalization case where we solely provide the demographic proxies as context. The intended subgroup is all users from the demography. For example, "recommend 10 movies from the candidates for 25-30-year-old self-employed females."

We experiment with varying lengths of history, such as 0, 1, 3, 5, 10, and 20, and also intersections of demographic proxies, such as combining age, gender, and occupation for movies and gender and continent for music. For each setup, we follow Algorithm 1 to sample the candidate items. The target distribution is the aggregated probability of a subset of the un-interacted items for all users defined by the proxy (C_2), along with items the group will never interact with (C_1). For Setup B we input $D = \{\emptyset\}$ as the demographic proxy.

Prompts: In each setup, we prompt models to recommend a ranked list of 10 items from a candidate list of 50 items for the users defined by the specified proxies in the context. Appendix A.1 depicts a sample prompt from the movie domain. Our test set comprises approximately 5,000 prompts for each domain. Table 1 shows the distribution of prompts for each setting in both domains.

Models: We conducted experiments using GPT-4o (Achiam et al., 2023), GPT-4o-mini, Llama-3.1-8B-Instruct (Dubey et al., 2024), and a random baseline

[1] https://grouplens.org/datasets/movielens/1m/
[2] http://last.fm/

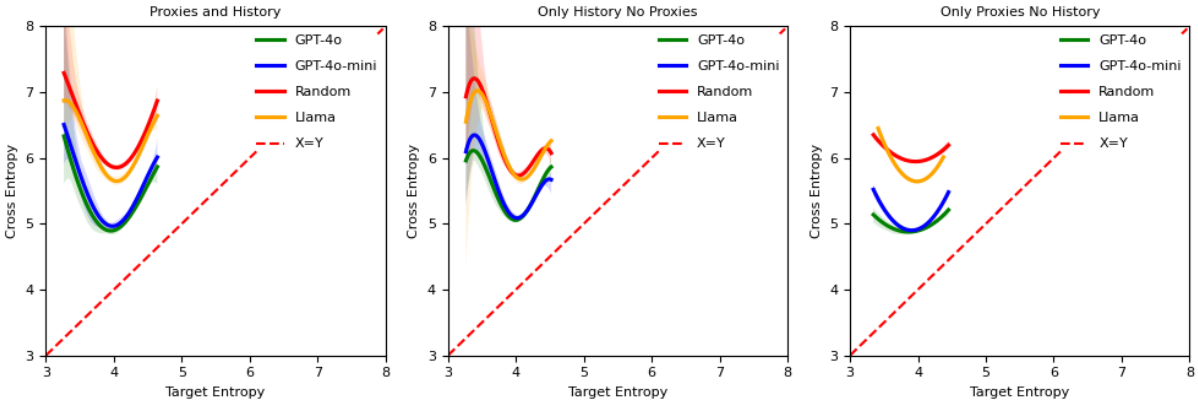

Figure 2: Movie Entropy Trend. Setup A: Left, B: Middle, C: Right.

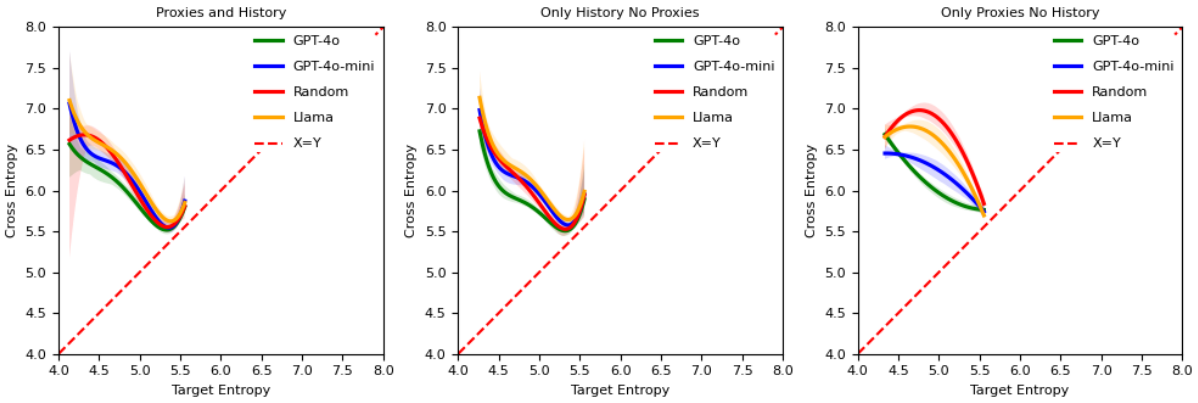

Figure 3: Music Entropy Trend. Setup A: Left, B: Middle, C: Right.

model where 10 recommended items were selected randomly from a pool of 50 candidates. For all experiments, the temperature parameter was set to 0. Running each combination (levels of proxy) incurred a cost of approximately USD 30 for GPT-4o and USD 2 for GPT-4o-mini. Thus, the total expenditure for GPT-based experiments was approximately 6 x USD 32 = USD 192. Llama experiments, on the other hand, were executed on two 48 GB NVIDIA RTX 6000 Ada GPUs, requiring around 18 hours to complete all settings across both domains.

Evaluation: Since we only prompt the model to generate a ranked list of 10 items, we approximate the prediction distribution over 50 items by imposing the ground distribution. We sort the target probabilities in descending order and assign the top 10 probability scores to the model's prediction. The remaining 40 items, which are not in the model's prediction, are sorted in descending order of their target probabilities and assigned the remainder of the target probabilities. Thus providing an optimistic estimate of the model predictions.

Algorithm 1 Candidate selection algorithm

procedure CANDIDATE SELECTION
  Input: Inventory I, Demography D, K=50
  for history $h \in \{0, 1, 3, 5, 10, 20\}$ **do**
    if $h > 0$ **then**
      I_h: Sample h random items from I
      u_h: Users with >= 60% of I_h + D.
      if $|u_h| < 3$ **then**
        Break
      else
        Continue
      end if
    else
      u_h: Users with D.
    end if
    I_h^u = Set of all items interacted by u_h.
    C_1 = $K/2$ random items from I_h^{uC}
    C_2 = K-$|C_1|$ random items from I_h^u-I_h
    Candidates = Random shuffle $C_1 + C_2$
    Target distribution: $Freq$(Candidates)
  end for
end procedure

To test our hypothesis from Section 3.3, we calculate the cross-entropy between the target probability distribution and the model's estimated distribution and plot the results. To smooth the graphs, we bucket the target entropy in 200 bins and average the cross-entropy score at each interval. We calculate a rolling average of the cross-entropy scores with a window length of 30 and fit an order four polynomial regression curve to visualize the *inflection point*.

4.3 Results and Observations

Trends of Generalization Figure 2 plots the regression curve for the **movies** domain. We observe the following: (i) The inflection point in all three setups is much lower for LLMs than the random baseline, indicating that *models generalize to a certain degree*. However, they widely differ in absolute numbers, where Llama performs much worse than GPT-4o and GPT-4o-mini, and slightly better than the random baseline. Both the curves for the GPT-based models are closer to the ideal X=Y line, and *GPT-4o slightly outperforms GPT-4o-mini* in all setups, with a lower inflection point. (ii) Since a lower target entropy signifies fewer users in a proxy, which necessitates more specific predictions, an increase in cross-entropy with decreasing target entropy indicates that the *models are not capable of personalizing predictions to smaller subsets of users, and hence will not be good at modeling each individual.* This is also evident from Setup C, which only prompts using demography and no history, and hence the weakest test of generalization. Almost following the X=Y line, the inflection points in all models are much lower than in other setups, signifying that *models can generalize using broader proxies and fail when the number of representative users in a proxy decreases*. (iii) The inflection point is lower in setup A than B, signifying that *combining demography and history enables model prediction*. This indicates that providing cultural information can enable personalization-based tasks.

Figure 3 plots the regression curve for the **music** domain. We observe the following: (i) All *models perform much worse than in movies*. Their inflection points are much higher. The predictions deviate from the X=Y line in all three setups, signifying the model's incapability of generalization using the proxies. (ii) *GPT-4o performs best* in all setups, where Llama is closer to the random baseline. This signifies that *music prediction is inherently more difficult than movie prediction* since the music dataset contains more interactions than movies, indicating that people listen to more music than watch movies. This can also signify that *the set of proxies used in the experiment is not optimal for personalizing music*. (iii) The inflection point of GPT-4o in Setup C is lower, indicating its generalization capabilities using broader proxies. Also, similar to movies, the inflection point of all models is lower in setup A than B, signifying that *combining demography and history enables model prediction.*

Overall and Proxy-wise distributions
We also plot the demographic and history-wise distributions of the cross-entropy in both domains in Figures 4 to 7. The term "(Def)" represents the default setting where no prior history is provided. In the left-most and second-from-right plots of the figures, this default setting is further broken down by demographic proxies. It reflects the model's responses when tasked with recommending the next 10 music or movie items from a given list, without being influenced by any prior context. This setup aims to capture the model's intrinsic tendency to associate items with specific cultural proxies. For the right-most and second-from-left plots, "(Def)" indicates a scenario where the model is probed using all proxies combined but still without any history. In this case, the model is presented with a list of music or movie items and asked to predict the next 10 items solely based on the given list, without any prior history or proxy influence. This configuration is designed to assess the model's inherent affinity for items within the provided list.

Plotting the demographic proxy-wise distribution, in Figures 4, 5, and 6 for GPT-4o, GPT-4o-mini, and Llama, we observe that the *models are incapable of generalizing when the combination of demographic proxies increase, irrespective of the size of history*, which is the **Strongest Case** for measuring generalization, according to our proposed framework in Section 3.2. For example, in movies (leftmost chart), the cross-entropy increases when a combination of all proxies is used along with history (Proxy = All). We see a similar trend in music (second from right), where although the target entropy decreases with more proxies (Proxy = Gen & Conti), the cross-entropy increases.

In both domains (second from left and right-most), we see models exhibiting a similar behavior with different lengths of history, where the cross-entropy increases with more history, which is the

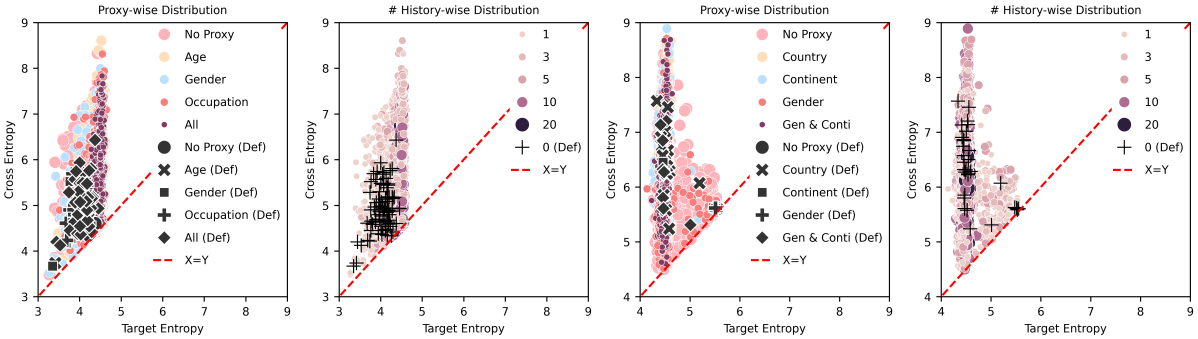

Figure 4: GPT-4o Demographic Proxy and History-wise distributions for Movie (Left 2) and Music (Right 2)

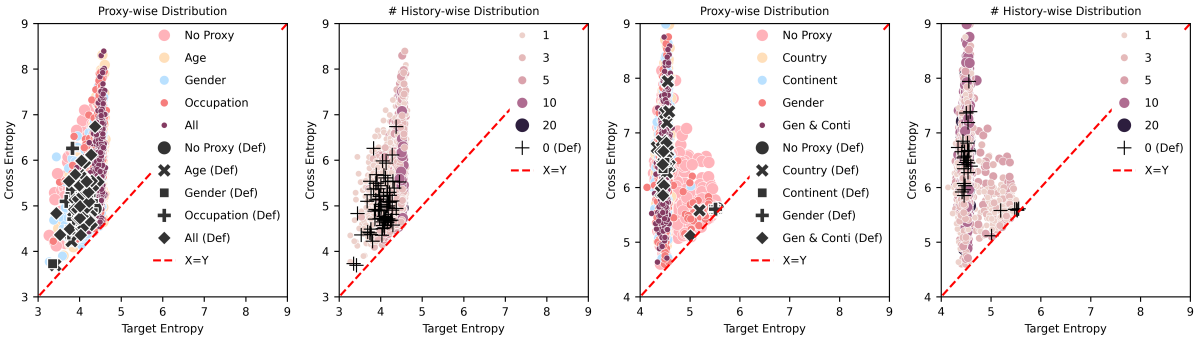

Figure 5: GPT-4o-mini Demographic Proxy and History-wise distributions for Movie (Left 2) and Music (Right 2)

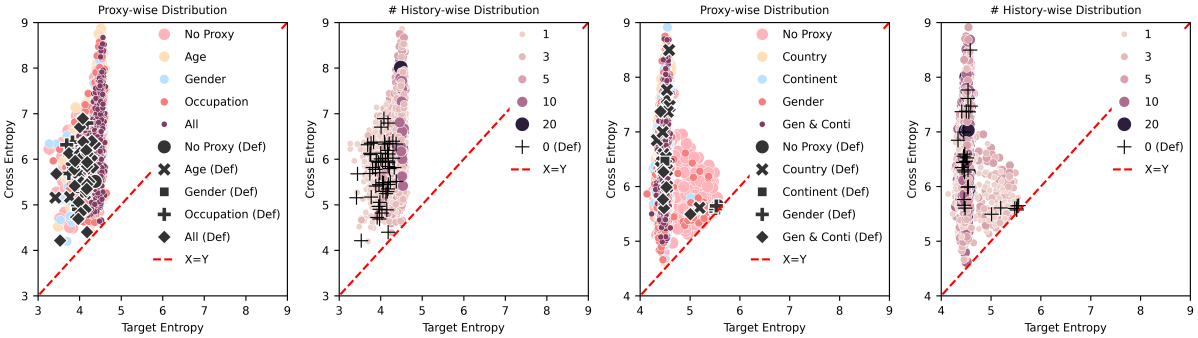

Figure 6: Llama Demographic Proxy and History-wise distributions for Movie (Left 2) and Music (Right 2)

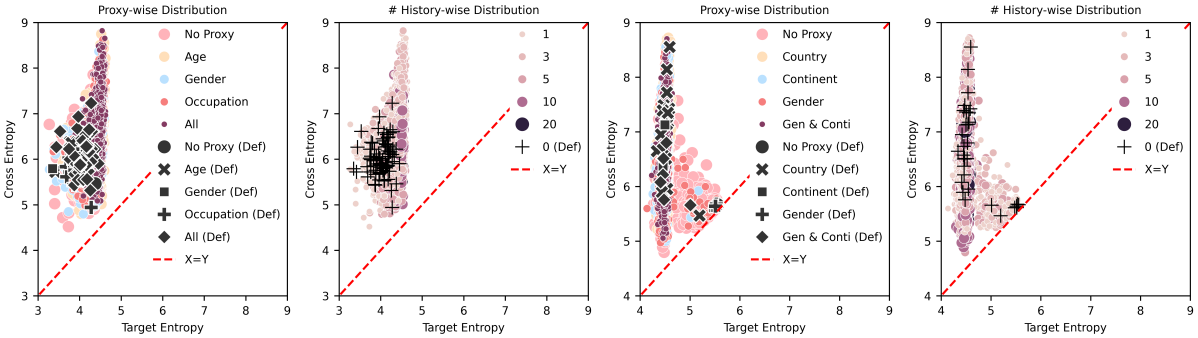

Figure 7: Random Demographic Proxy and History-wise distributions for Movie (Left 2) and Music (Right 2)

Strongest Case for measuring generalization. This indicates the model's incapability to adeptly utilize the context, which is a known phenomenon (Mukherjee et al., 2024).

We further plot the results of each demographic proxy and history independent in Section A.2 (Appendix A). Overall, our experiments indicate the viability of our proposed framework in Section 3.2 for measuring generalization using existing personalization-based tasks. We clearly see that models are capable of generalization to a certain degree till they reach an inflection point. Although GPT-4o and GPT-4o-mini models perform better than Llama and a random baseline, the results indicate a strong generalization gap and much room for improvement.

5 Discussion

1. What is the difference between personalization and personalization as a measure of generalization? Personalization as a measure of generalization is a theoretical framework for measuring generalization through the lens of personalization, not on improving personalization itself. Unlike existing personalization techniques (Sun et al., 2023; Hwang et al., 2023; Zhang et al., 2024), which are evaluated using accuracy-based metrics (e.g., F1, NDCG, MAP), our method analyzes generalization by examining cross-entropy over the model's response distribution. This fundamental difference in objectives and metrics makes direct comparisons with prior works inherently challenging.

2. Why is personalization interesting and useful for studying generalization in LLMs? Besides the theoretical arguments provided earlier, here we list some compelling practical reasons for favoring personalization as an evaluation strategy for generalization.

Generalizing from learned knowledge: Since the context length of current models is limited, personalization requires reasoning using more context, which might be outside the model's context length. Thus, it is a robust measure of a model's generalizability as its capacity to leverage the learned world knowledge during training.

Balancing worldviews: Since personalization requires tailoring things for an individual, a generalizable model should be capable of balancing between universal and individual-specific knowledge for performing tasks.

Dynamicity: Personalization evades the issue of models memorizing training examples and recalling during inference, since individual preferences change over time. Hence, a model can't blatantly memorize each user's behavior to perform well in personalization tasks.

Cost efficiency and ROI: Our evaluation framework is highly cost-effective compared to creating entirely new benchmarks from scratch. Developing "challenging" test beds for model generalization requires substantial human and computational resources to resolve data scarcity and contamination issues (see Section 2). In contrast, our approach repurposes existing personalization tasks to assess generalization, eliminating the need for costly new dataset creation.

LLMs as world models: Agentic models (Shavit et al.; Acharya et al., 2025) require universal knowledge, which is less dependent on individual users. Although LLMs have exhibited tremendous capacity as agents, they must model each individual to be world models.

6 Conclusion

We propose a statistically motivated framework using personalization to assess generalization in LLMs. Since LLMs are trained on vast human-generated data, we argue that true generalization lies in predicting human behavior rather than specific tasks. This philosophically aligns with Wittgenstein's language games and Turing's imitation game, though with a key distinction with the latter: while Turing's test requires mimicking *any human*, our framework challenges models to replicate *a specific user* at varying complexity levels. This shift has profound philosophical and mathematical implications, only some of which we could explore in this paper.

We would like to highlight the lack of datasets where complex user behaviors or preferences are available alongside their demographic proxies, which will enable us to conduct large-scale and more extensive studies of generalization.

Limitations

The empirical study presented here is limited in several ways: First, we explore only two kinds of behavior preferences - movies and music; the choices were motivated by the availability of large public datasets of user preferences or behavior, where we have some demographic proxies for the users. These experiments do not tell us how models generalize for more complex user behaviors or other kinds of demographic proxies (including psychological features such as personality traits). Second, we experiment only with English prompts and

Latin script. It will be interesting to compare a model's generalization in the English language and Latin script to that of other languages and scripts, especially when the prompt is expressed in those languages/scripts. Third, our experiments only consider three models - GPT4-o, GPT4-o-mini, and Llama-3.1-8B-Instruct. Since both of the GPT models are closed-source and behind a pay wall, we are aware that reproducing our experiments independently would incur additional cost. Extending the study to more open-weight models, apart from Llama, would be important to understand the robustness of the proposed framework.

Our theoretical framework assumes that it is possible to estimate the true distributions of user behavior from large samples, which might not be the case if user behavior is non-stochastic or chaotic. Furthermore, the datasets used in this study may not be large enough for estimating user behavior at a global or national scale, which implies that our estimates might have large noise terms, leading to significant over- or under-estimations of models' generalizability.

Ethical Implications

Being a theoretical and exploratory study, our work has no direct risks or harms. Nevertheless, we assume in our work that the behavior of users can be estimated in a statistical sense for groups of different sizes, and for a certain definition of groups, the entropy of these distributions is small. It is possible to misinterpret this assumption as a promotion of the idea of stereotypical behaviors of certain groups. We warn against such interpretations. The only two assumptions made here are (a) user behaviors can be stochastically modeled (a common assumption made across many branches of social sciences, such as Economics and Psychology), and therefore, (b) there are latent variables that determine such behaviors. Although we have used "demographic proxies" as a term for these latent variables and used certain proxies (country, age, gender, etc.) in our experiments, we do not promote the idea that users from the same demographic group display similar behavior. The terminology is borrowed from previous work (Adilazuarda et al., 2024); however, the proposed framework is agnostic to any anthropological or psychological theory of human behavior.

Acknowledgements

This research was supported by Microsoft Accelerate Foundation Models Research (AFMR) Grant.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Mary Bazire and Patrick Brézillon. 2005. Understanding context before using it. In *Modeling and Using Context: 5thInternational and Interdisciplinary Conference CONTEXT 2005, Paris, France, July 5-8, 2005. Proceedings 5*, pages 29–40. Springer.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Erion Çano, Maurizio Morisio, et al. 2017. Music mood dataset creation based on last. fm tags. In *2017 International Conference on Artificial Intelligence and Applications, Vienna, Austria*, pages 15–26.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

O. Celma. 2010. *Music Recommendation and Discovery in the Long Tail*. Springer.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Ya-Xi Chen, Sebastian Boring, and Andreas Butz. 2010. How last. fm illustrates the musical world: user behavior and relevant user-generated content. In *Proceedings of the international workshop on Visual Interfaces to the Social and Semantic Web*, pages 1203–1204.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Preprint*, arXiv:2501.17161.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1126–1132.

Ravi Dhar and Klaus Wertenbroch. 2000. Consumer choice between hedonic and utilitarian goods. *Journal of marketing research*, 37(1):60–71.

Kevin Droe. 2006. Music preference and music education: A review of literature. *Update: Applications of Research in Music Education*, 24(2):23–32.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Haiyan Fan and Marshall Scott Poole. 2006. What is personalization? perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce*, 16(3-4):179–202.

Ian Goodfellow. 2016. *Deep learning*, volume 196. MIT press.

Mahesh Goyani and Neha Chaurasiya. 2020. A review of movie recommendation system: Limitations, survey and challenges. *ELCVIA: electronic letters on computer vision and image analysis*, 19(3):0018–37.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Gregory P Hanley, Brian A Iwata, and Eileen M Roscoe. 2006. Some determinants of changes in preference over time. *Journal of Applied Behavior Analysis*, 39(2):189–202.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Won Kim. 2002. Personalization: Definition, status, and challenges ahead. *Journal of object technology*, 1(1):29–40.

Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *Preprint*, arXiv:2307.07870.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference*

on Artificial Intelligence, volume 38, pages 18471–18480.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024a. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, et al. 2023. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*.

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

Celia Lury and Sophie Day. 2019. Algorithmic personalization as a mode of individuation. *Theory, Culture & Society*, 36(2):17–37.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Paritosh Nagarnaik and A Thomas. 2015. Survey on recommendation system methods. In *2015 2nd international conference on electronics and communication systems (ICECS)*, pages 1603–1608. IEEE.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Johannes Putzke, Kai Fischbach, Detlef Schoder, and Peter A Gloor. 2014. Cross-cultural gender differences in the adoption and usage of social media platforms–an exploratory study of last. fm. *Computer Networks*, 75:519–530.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Everett M Rogers, Arvind Singhal, and Margaret M Quinlan. 2014. Diffusion of innovations. In *An integrated approach to communication theory and research*, pages 432–448. Routledge.

Markus Schedl. 2016. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pages 103–110.

Bernd Schmitt. 1999. Experiential marketing. *Journal of marketing management*, 15(1-3):53–67.

Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems.

Chenkai Sun, Jinning Li, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2023. Measuring the effect of influential messages on varying personas. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 554–562.

S Shyam Sundar and Sampada S Marathe. 2010. Personalization versus customization: The importance of agency, privacy, and power usage. *Human communication research*, 36(3):298–322.

Hamidreza Tahmasbi, Mehrdad Jalali, and Hassan Shakeri. 2018. Modeling temporal dynamics of user preferences in movie recommendation. In *2018 8th international conference on computer and knowledge engineering (ICCKE)*, pages 194–199. IEEE.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *Preprint*, arXiv:2309.13356.

B Thompson, SG Roberts, and G Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. nature human behaviour, 4 (10), 1029–1038.

Alan Turing. 1948. Intelligent machinery (1948). *B. Jack Copeland*, page 395.

Alan Turing. 1950. Machinery and intelligence. *Mind: A Quarterly Review of Psychology and Philosophy*, 59(236):433–460.

Jari Vesanen. 2007. What is personalization? a conceptual framework. *European Journal of Marketing*, 41(5/6):409–418.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Brooke Wilken, Yuri Miyamoto, and Yukiko Uchida. 2011. Cultural influences on preference consistency: Consistency at the individual and collective levels. *Journal of Consumer Psychology*, 21(3):346–353.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *ArXiv*, abs/2311.01964.

Andreas Zimmermann, Andreas Lorenz, and Reinhard Oppermann. 2007. An operational definition of context. In *Modeling and Using Context: 6th International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007. Proceedings 6*, pages 558–571. Springer.

A Appendix

A.1 Prompts

Prompt

AI Rules
- Output response as a Python list only.
- Do not output any extra text.
- Do not wrap the response in Python markers.
- Do not assign the list to any variable.
- List values in double-quotes.

You are proficient in recommending new movie for users to watch based on their background, previous view history, or a combination of both.
The user is a 25-34 years old Male clerical/admin.
The user has previously watched the following movies: ['Out of Sight (1998)', 'Horse Whisperer, The (1998)', 'Star Wars: Episode V - The Empire Strikes Back (1980)', 'Odd Couple II, The (1998)', 'Marathon Man (1976)'].
From the candidates listed below, recommend the next 10 movie for the user to watch based on the user's background, previous view history, or a combination of both.
Format your response as a Python list of item names. The list must be ranked from the most likely to the least likely movie.
Candidates: ['Mr. & Mrs. Smith (1941)', 'Blue Velvet (1986)', 'Freedom for Us (À nous la liberté) (1931)', 'White Balloon, The (Badkonake Sefid) (1995)', 'Fear, The (1995)', 'Barefoot Executive, The (1971)', 'Barb Wire (1996)', 'Jungle Book, The (1967)', 'Matrix, The (1999)', '24-hour Woman (1998)', 'Heat (1995)', 'Fish Called Wanda, A (1988)', 'Independence Day (ID4) (1996)', 'Dead Calm (1989)', 'Phantasm III: Lord of the Dead (1994)', 'To Be or Not to Be (1942)', 'Rain Man (1988)', 'Carnosaur (1993)', 'Heathers (1989)', 'Gaslight (1944)', 'Get Bruce (1999)', 'Omen, The (1976)', 'Bedknobs and Broomsticks (1971)', 'Herbie Rides Again (1974)', 'Buck and the Preacher (1972)', 'Wallace & Gromit: The Best of Aardman Animation (1996)', 'Friday the 13th Part 3: 3D (1982)', 'Meatballs (1979)', 'Cabin Boy (1994)', '8 Heads in a Duffel Bag (1997)', 'Mariachi, El (1992)', 'Contender, The (2000)', 'When a Man Loves a Woman (1994)', 'Henry Fool (1997)', 'Beetlejuice (1988)', 'Requiem for a Dream (2000)', 'Raven, The (1963)', 'Grand Day Out, A (1992)', 'Miami Rhapsody (1995)', 'Tales From the Crypt Presents: Demon Knight (1995)', 'Sticky Fingers of Time, The (1997)', 'Opposite of Sex, The (1998)', 'Saving Private Ryan (1998)', 'Naked Gun 33 1/3: The Final Insult (1994)', 'Trigger Effect, The (1996)', 'Among Giants (1998)', 'Spaceballs (1987)', 'Bloody Child, The (1996)', 'Snow White and the Seven Dwarfs (1937)', 'Man Who Knew Too Much, The (1934)']

A.2 Additional Plots

Domain	Setting	# History	Freq
Movie	No Proxy (Def)	0	1
	No Proxy	1	300
		3	300
		5	300
		10	89
		20	5
	Age (Def)	0	7
	Age	1	300
		3	300
		5	300
		10	147
		20	5
	Gender (Def)	0	2
	Gender	1	300
		3	300
		5	300
		10	110
		20	5
	Occupation (Def)	0	17
	Occupation	1	300
		3	300
		5	300
		10	112
		20	2
	All (Def)	0	51
	All	1	300
		3	300
		5	300
		10	43
Music	No Proxy (Def)	0	1
	No Proxy	1	246
		3	249
		5	246
		10	240
		20	11
	Country (Def)	0	17
	Country	1	250
		3	249
		5	248
		10	208
		20	9
	Continent (Def)	0	4
	Continent	1	250
		3	249
		5	249
		10	249
		20	11
	Gender (Def)	0	2
	Gender	1	249
		3	247
		5	249
		10	250
		20	13
	Gen & Conti (Def)	0	8
	Gen & Conti	1	246
		3	248
		5	249
		10	171
		20	6

Table 1: Number of examples across all experiment settings.

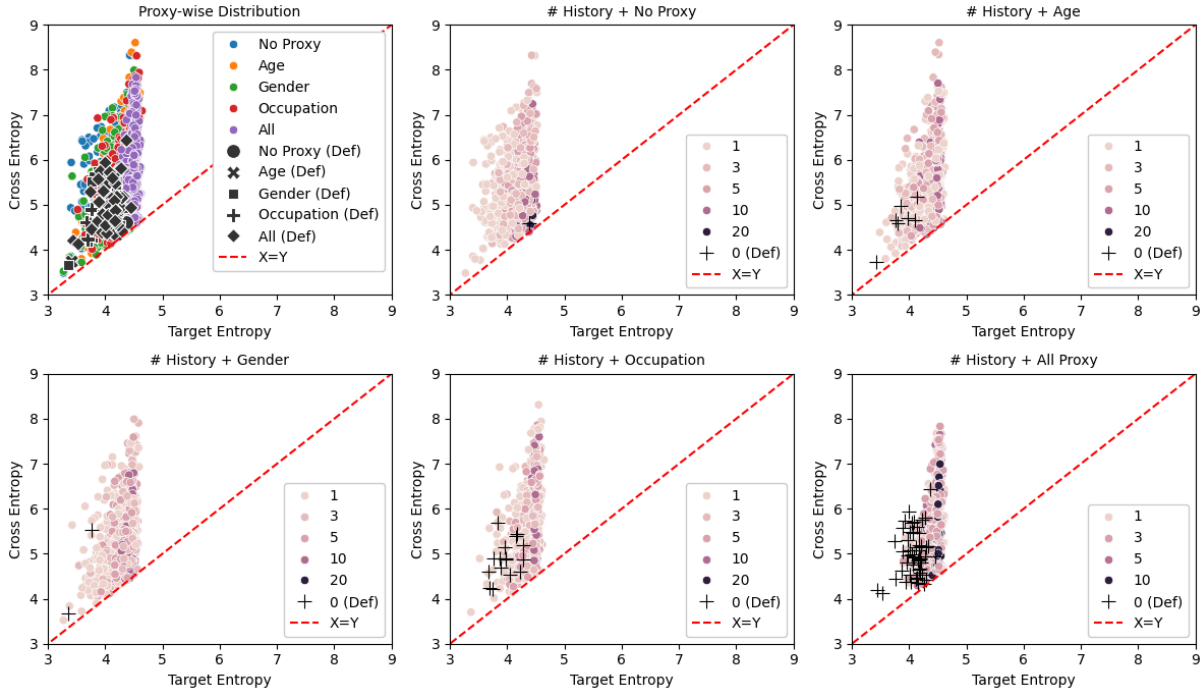

Figure 8: GPT-4o detailed plot for Movie

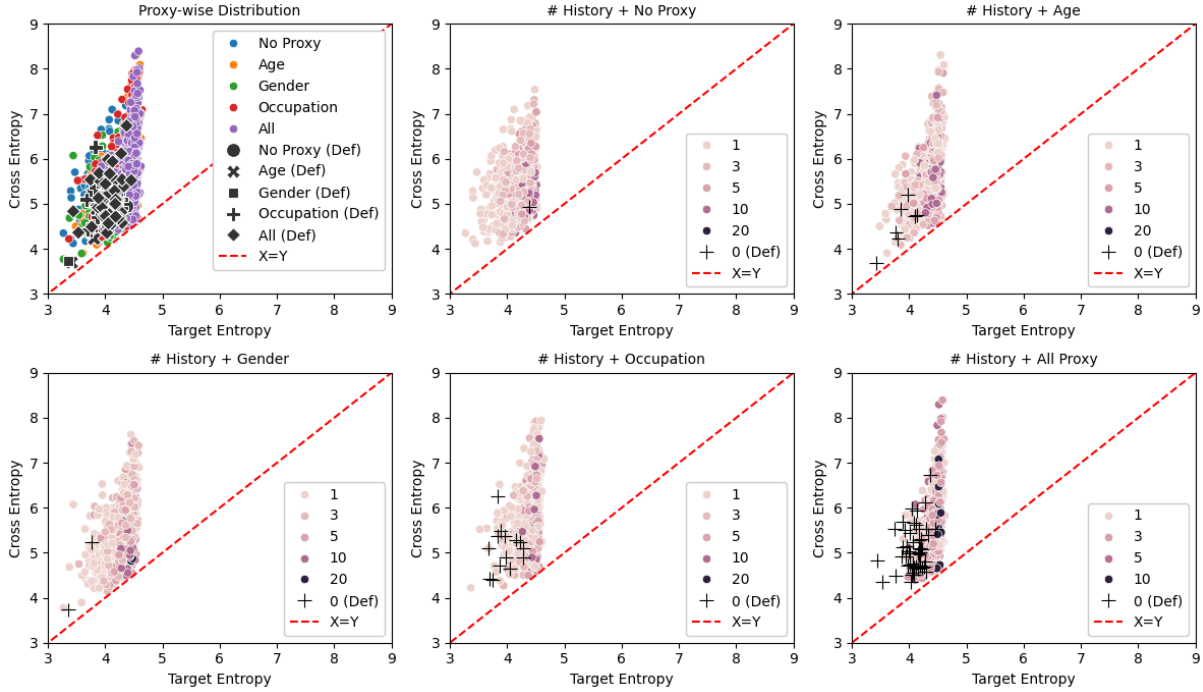

Figure 9: GPT-4o-mini detailed plot for Movie

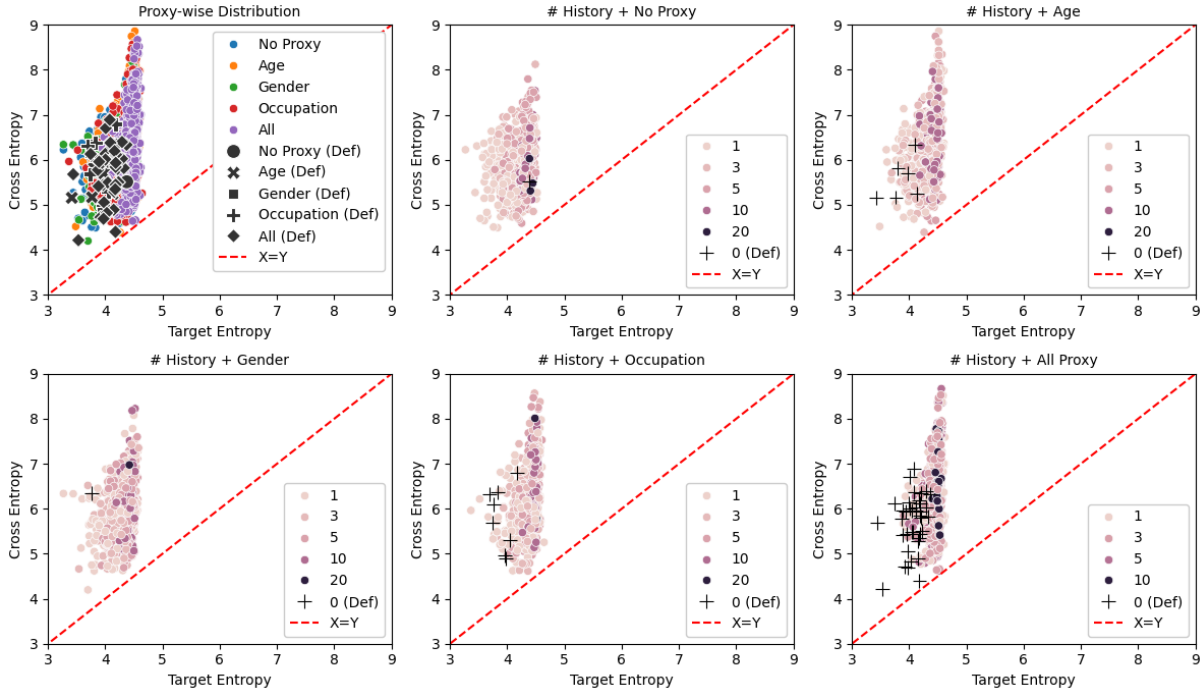

Figure 10: Llama detailed plot for Movie

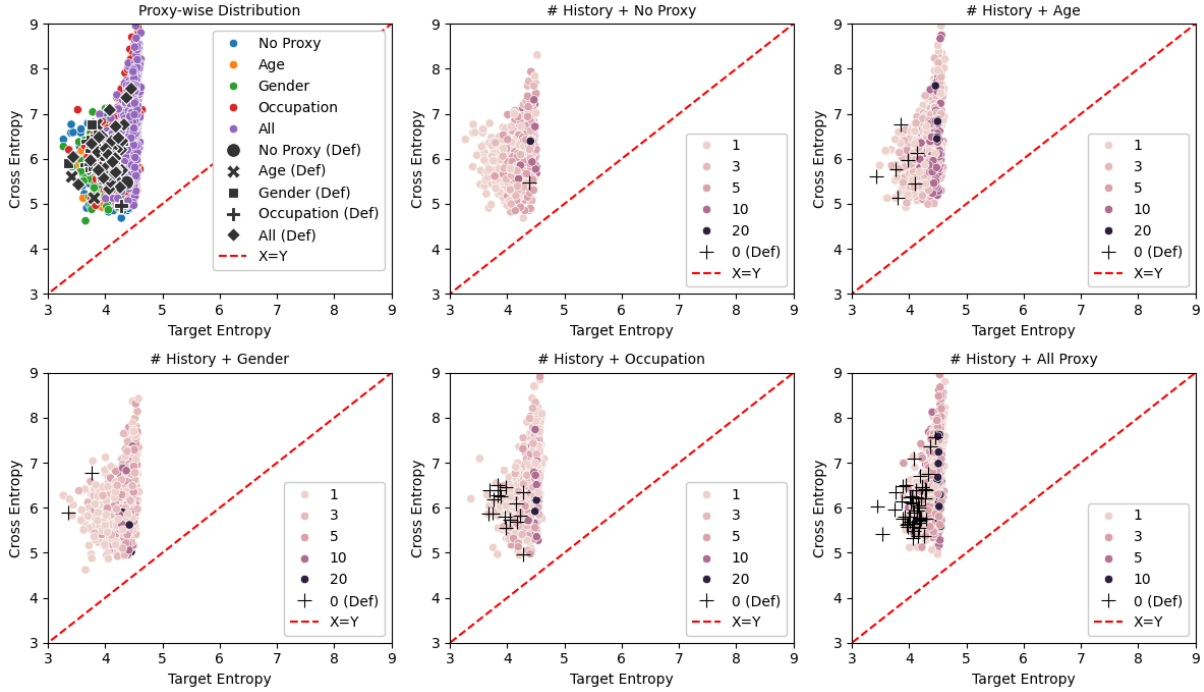

Figure 11: Random detailed plot for Movie

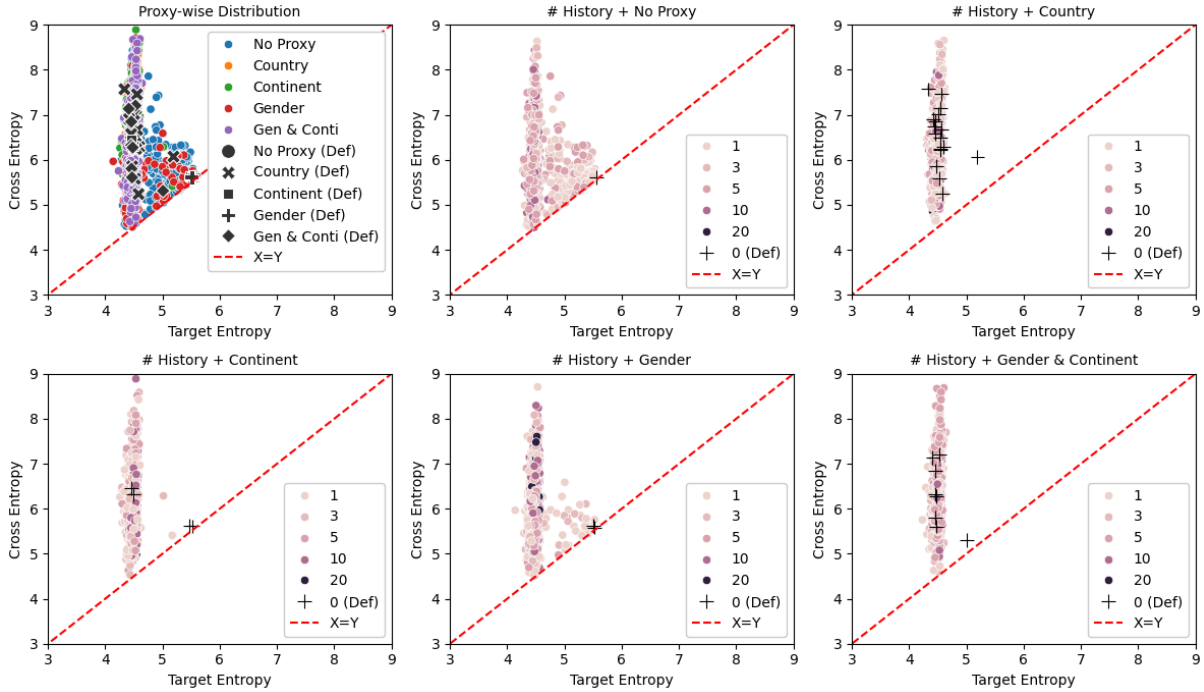

Figure 12: GPT-4o detailed plot for Music

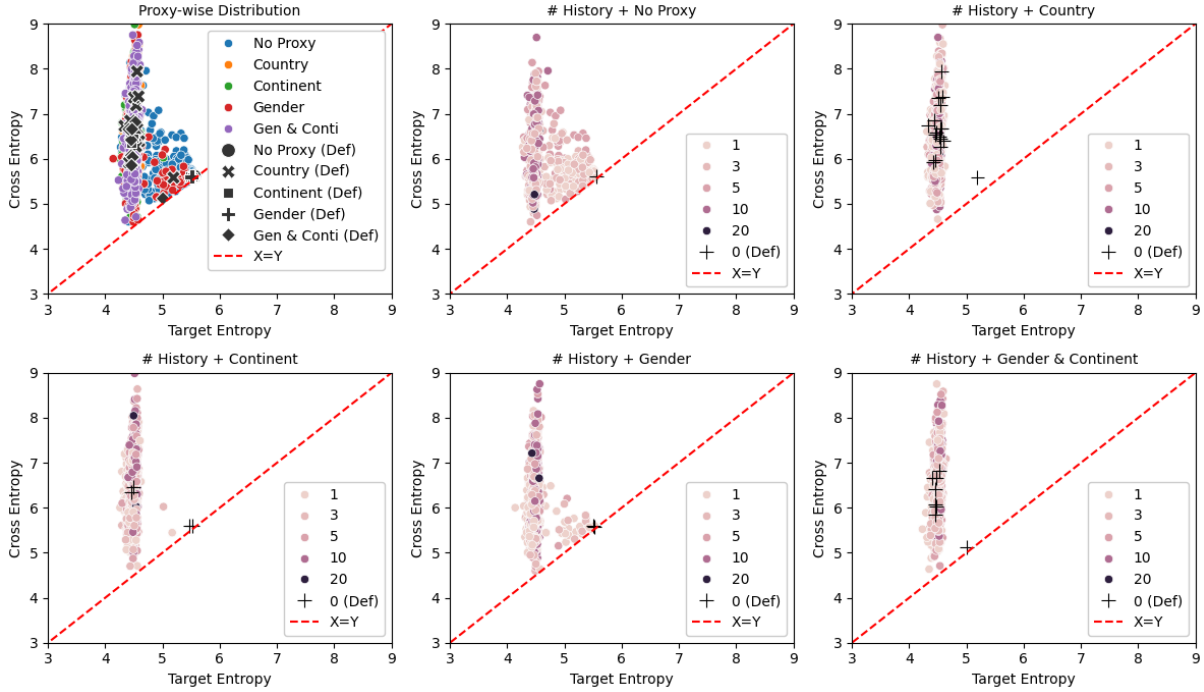

Figure 13: GPT-4o-mini detailed plot for Music

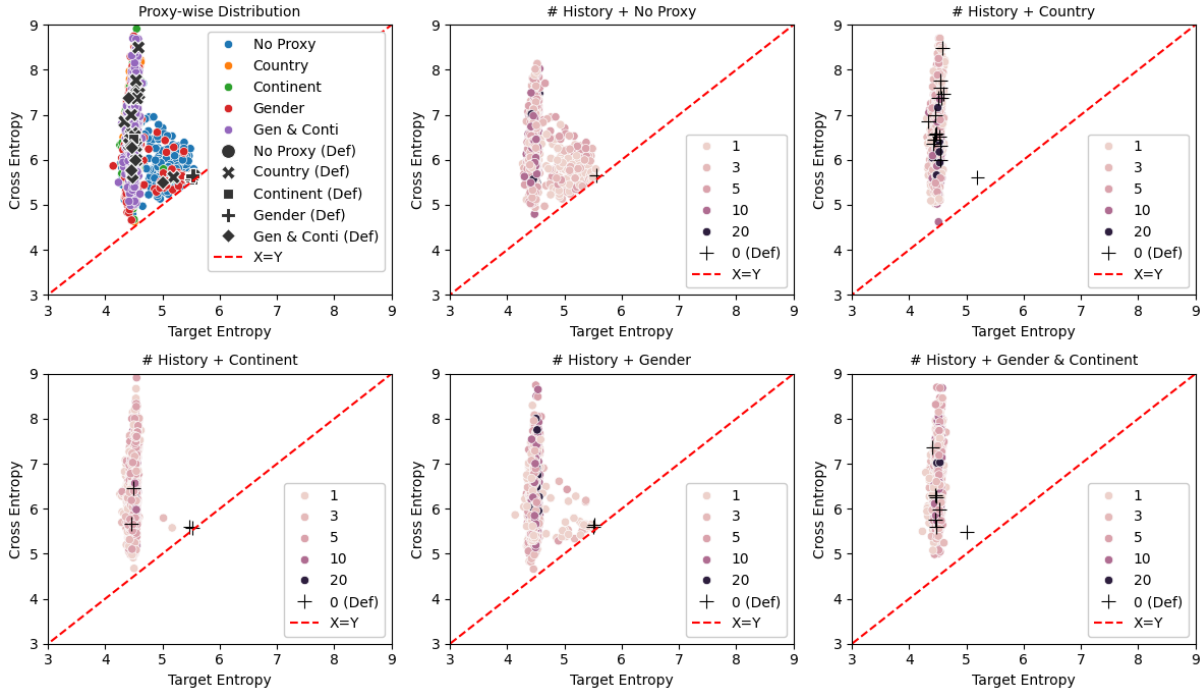

Figure 14: Llama detailed plot for Music

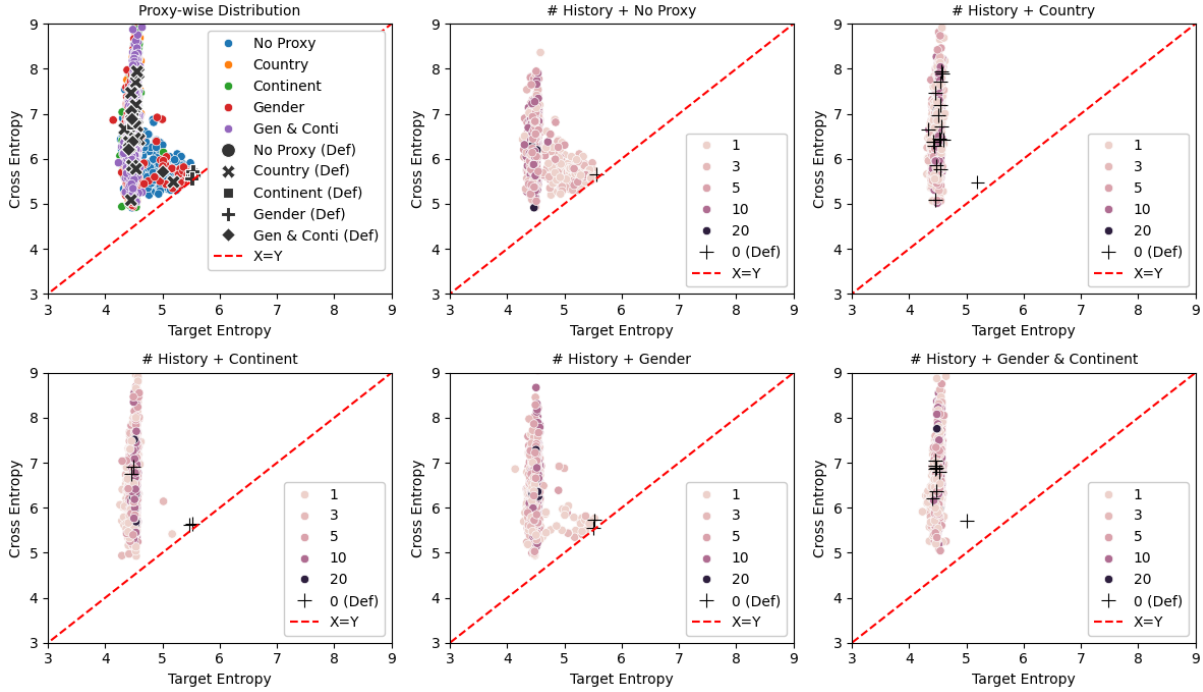

Figure 15: Random detailed plot for Music