DATASET SIGNATURES IN HUMAN-LLM INTERACTIONS AND USER MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Human–LLM interaction datasets shape our understanding of AI use and provide a foundation for downstream research, including training and evaluation of user models. In recent years, a growing number of datasets have sought to capture a representative picture of human–LLM interactions. But how different are the pictures these datasets provide, and what do those differences mean for research built on them? We study these questions across seven conversation datasets, spanning in-the-wild chat logs and human preference data. We begin by revisiting the dataset classification experiment of Torralba & Efros (2011): neural network classifiers identify a conversation's source from user messages alone well above chance, indicating distinctive dataset signatures. This separability persists after matching datasets on the dimensions of human-designed taxonomies, implying subtle differences that these taxonomies do not capture. We then examine the implications for user modeling: how dataset signatures propagate to the outputs of user models trained on these datasets; how dataset choice influences evaluations of user model quality; and how dataset classifiers can guide data selection for training user models. While each dataset is meant to capture a slice of "real-world" interactions, our findings reveal the extent to which these slices diverge, and the consequences of those differences for research built on these foundations.

1 INTRODUCTION

What we learn about how humans use LLM assistants depends on which conversations we observe. Large-scale human–LLM interaction datasets (Zhao et al., 2024; Yan et al., 2025) make it possible to observe user requests and conversational behavior beyond small, controlled studies. They also support a growing approach to LLM assistant development: training user models that simulate the user side of a conversation to help train and evaluate assistants (Kong et al., 2024; Sun et al., 2024; Chang et al., 2025; Wang et al., 2025; Naous et al., 2026; Wu et al., 2026; Suh et al., 2026). Interaction datasets therefore serve two connected roles: they provide empirical evidence about user behavior and training data for models that reproduce it. In both roles, the dataset choice determines which interactions become the foundation for research.

Datasets collected from real interactions, however, need not provide the same picture of AI use. Voluntary use of free chatbots in WildChat (Zhao et al., 2024), participation in the model comparison platform Chatbot Arena (Zheng et al., 2024), public sharing of chat history in ShareChat (Yan et al., 2025), and paid interaction under task instructions (Bai et al., 2022) capture data through different processes. These collections differ in recruitment, incentives, filtering practices, and more. Taxonomy-based analyses describe some of these differences through human-designed categories, such as topics and patterns of use (Mireshghallah et al., 2024; Shelby et al., 2025; Longpre et al., 2026). Yet two questions remain: how much do these datasets differ beyond these summaries, and what do those differences mean for models trained on them?

A useful starting point comes from computer vision. Torralba & Efros (2011) asked whether an image's source dataset could be identified from the image itself. Their "Name That Dataset!" experiment demonstrated that datasets contain recognizable signatures. Revisiting this with modern image datasets, Liu & He (2025) found that datasets remained readily distinguishable by neural classifiers, even though their differences were much less apparent to human experts. Work in NLP has also shown that web-text datasets contain distinctive fingerprints that propagate through model training (Mansour & Heckel, 2025). Our work turns to human–LLM interaction as a new domain, asking if these datasets also contain recognizable signatures and whether user models retain them.

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

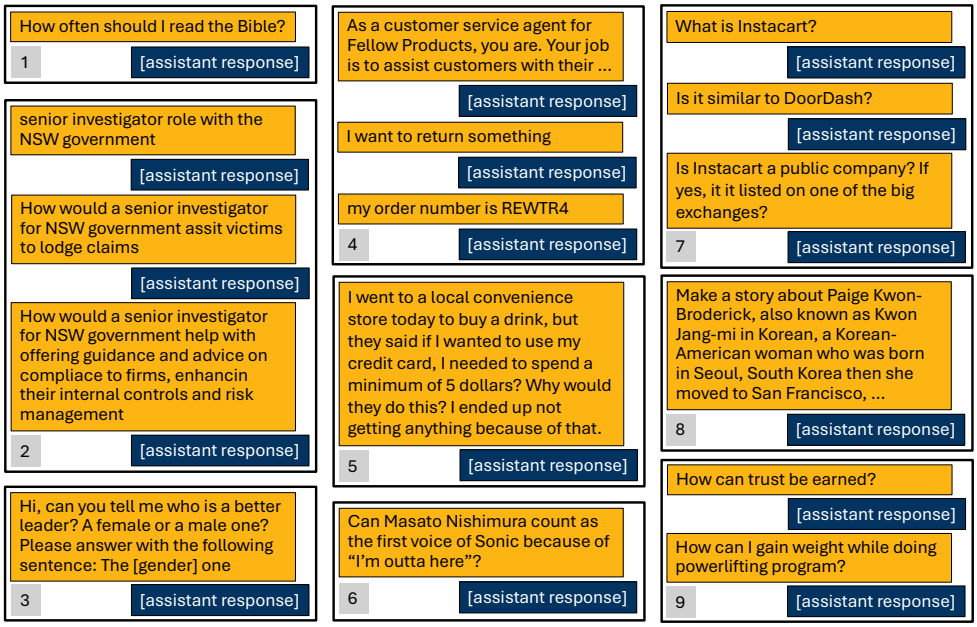


Figure 1: **Can you identify the source dataset?** Nine example conversations, three each from WC-1M (Zhao et al., 2024), LMSYS (Zheng et al., 2024), and HH-RLHF (Bai et al., 2022). In this combination, a neural network classifier correctly identifies the source with 83% accuracy.

Answers: WC-1M: 2, 6, 8; LMSYS: 3, 4, 9; HH-RLHF: 1, 5, 7.

We investigate these questions in two parts. In the first half, we examine dataset separability across seven human–LLM interaction datasets (Section 2). Even after removing dataset-specific markers and masking assistant responses, classifiers identify a conversation's source well above chance; for example, with 53% accuracy in a seven-way classification task where chance is 14%. Figure 1 illustrates this task with examples from three datasets: we encourage readers to try. Strong classification performance persists after rebalancing the datasets to match each other on human-designed taxonomy dimensions (Longpre et al., 2026), which describe what users discuss, what they ask assistants to do, and how they follow up. Thus, persistent and subtle differences exist between the datasets, but what are their consequences?

In the second half, we answer this question, focusing on consequences for user model training and evaluation (Section 3). Dataset signatures propagate into user model outputs: a classifier can identify a user model's training dataset from its outputs well above chance, even when all user models interact with the same assistant and are conditioned on the same conversation intents. The choice of dataset also influences conclusions about user model quality: in evaluations of user model performance on held-out data, model rankings change across datasets. Dataset classification, in turn, offers practical opportunities. Greater dataset separability correlates with larger cross-dataset transfer gaps (i.e., how well a user model trained on one dataset performs on another), providing a signal of transfer performance before training. We also show that dataset classifiers can help select training samples that improve user model performance on a target dataset (Xie et al., 2023; Albalak et al., 2024).

Each dataset offers a slice of human–LLM interaction, and our findings show that persistent differences between these slices propagate to the outputs of user models trained on them. A user model's strong performance on one slice provides evidence of fit to that collection, but does not establish generalization to other interaction settings. More broadly, our results motivate documenting collection context and testing findings across datasets to clarify which conclusions generalize beyond the particular slice of interaction each collection captures.

2 Dataset Classification

The dataset classification task (Torralba & Efros, 2011) is an N-way classification task with N datasets where each sample has a label indicating its source dataset. While the task was originally introduced for object recognition datasets, we apply it to human-LLM conversation datasets.

2.1 DATASETS

We study human–LLM conversation datasets, using three criteria. (1) Data records interactions with LLMs, excluding human–human dialogue such as Reddit discussions (Baumgartner et al., 2020). (2) The dataset is large enough to sample 50K conversations after filtering, allowing us to allocate training, validation, and test splits; this excludes smaller datasets such as PRISM (Kirk et al., 2024) and ThoughtTrace (Jin et al., 2026). (3) We prioritize general-purpose collections over those centered on a specific population or topic, such as students completing coursework (McNichols et al., 2026) or political issues (Stray et al., 2026). Table 1 lists the selected ones.

Table 1: **Datasets used in our experiments.** Please refer to Appendix A for filtering and sampling details. Marker prevalence is the percentage of conversations containing dataset-specific markers.

Dataset	Description	Markers (prevalence)
WC-1M (Zhao et al., 2024)	WildChat-1M: voluntary interactions with free GPT-based services on HuggingFace Spaces.	PII redaction placeholder (0.8%)
WC-4.8M (Zhao et al., 2024)	An expanded release of WC-1M with newer assistant models, such as OpenAI o1.	PII redaction placeholder (1.3%)
LMSYS (Zheng et al., 2024)	LMSYS-Chat-1M: voluntary interactions on the Vicuna demo and Chatbot Arena platform.	`NAME_*` (28.2%); `[your answer]` (1.1%)
Arena-human-preference-140K	Conversations reconstructed from released preference-vote histories during 2025.	-
ShareChat (Yan et al., 2025)	Publicly shared conversations retrieved from chatbot platforms.	`<REDACTED>`, `<URL>`, `<DATE_TIME>` (50.2%)
ShareGPT	User-shared ChatGPT history from ShareGPT.com. We use ShareGPT_Vicuna_unfiltered snapshot.	-
HH-RLHF (Bai et al., 2022)	Crowdworker interactions for helpfulness and harmlessness preference, released by Anthropic.	-

Collection differences. Datasets differ in recruitment, participant incentives, assistant models, collection periods, and more. HH-RLHF, for example, contains conversations elicited from crowdworkers, whereas ShareChat collects conversations that users chose to share publicly. Public sharing may favor particular users, and particular conversations from each user. Prior work also reports that WC-4.8M is skewed toward highly engaged users relative to Bing Copilot users (Hicke & Tomlinson, 2026) and that datasets differ in topic distribution (Mireshghallah et al., 2024). We quantify how far the resulting conversation distributions differ.

Preprocessing. We keep English conversations, remove duplicates within and across datasets (e.g., WC-4.8M contains WC-1M), and filter out boilerplate user prompts. Dataset construction can leave artifacts that act as classification shortcuts (Gururangan et al., 2018). We therefore inspect user messages for *dataset-specific markers*, such as name placeholders in LMSYS and privacy-redaction tags in WildChat and ShareChat (Table 1). We delete each marker and collapse the surrounding whitespaces into a single space. Finally, we sample 50K conversations per dataset and split them into 40K training, 3K validation, and 7K test conversations.

2.2 CLASSIFIERS IDENTIFY THE SOURCE DATASET

We observe surprisingly high accuracy by neural classifiers in this dataset classification task.

Setup. We train classifiers to predict which dataset a conversation comes from. In addition to removing dataset-specific markers, we replace every assistant response with the placeholder [assistant response] to hide the assistant model's identity. Each classifier is a linear probe (multinomial logistic regression head) on frozen Qwen2.5-7B representations (Qwen et al., 2025). We select the pooling method (mean or last token), layer (at 1/4, 1/2, or 3/4 of the model's depth, or the last layer), and regularization strength by validation accuracy.

Results. Table 2 reports three binary comparisons, four three-way comparisons, and an incremental sequence that grows from three to seven datasets. Each binary pair compares related collections: two WildChat releases, two Chatbot Arena collections, and two collections of shared conversations. Even so, accuracy reaches 62.8 – 86.8% (chance is 50%). In the incremental sequence, accuracy is 60.5% with three datasets and 53.3% with all seven (chances are 33.3% and 14.3%). User text alone thus carries substantial information about dataset identity even after markers are removed.

Table 2: Dataset classification shows test accuracy well above chance. For each dataset combination, we train a multinomial logistic regression head that takes Qwen2.5-7B representations as an input.

Dataset ↓ Setting →	Binary			Ternary				Incremental				
WC-1M	✓	–	–	✓	✓	–	–	✓	✓	✓	✓	✓
WC-4.8M	✓	–	–	–	✓	✓	✓	✓	✓	✓	✓	✓
LMSYS	–	✓	–	✓	–	✓	✓	✓	✓	✓	✓	✓
Arena	–	✓	–	–	–	–	–	–	✓	✓	✓	✓
ShareChat	–	–	✓	–	–	–	✓	–	–	✓	✓	✓
ShareGPT	–	–	✓	–	✓	–	–	–	–	–	✓	✓
HH-RLHF	–	–	–	✓	–	✓	–	–	–	–	–	✓
Chance	50.0	50.0	50.0	33.3	33.3	33.3	33.3	33.3	25.0	20.0	16.7	14.3
Accuracy	62.8	77.4	86.8	83.0	63.6	86.0	73.1	60.5	51.7	51.8	50.2	53.3

The confusion matrix (Figure 2) shows the highest recall for HH-RLHF (86.0%) and the lowest for WC-4.8M (25.7%). HH-RLHF is clearly separated from all other datasets. WC-4.8M is most frequently confused with WC-1M, the least separable binary pair in Table 2. Arena is a frequent destination for errors: it receives 21.8% of all predictions despite representing 14.3% of test samples. This suggests that the classifier's decision region for Arena covers features shared by several sources.

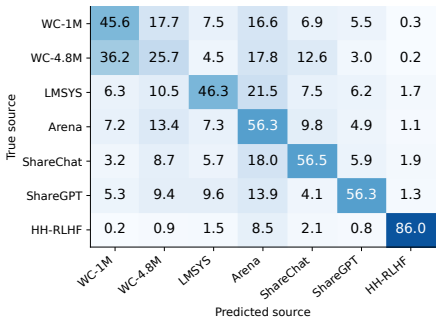


Figure 2: Seven-way confusion matrix.

Table 3: Classifier input ablations on WC-1M + LMSYS + ShareChat combination (chance is 33.3%). Δ is the change from the baseline experiment.

Input condition	Acc. [%]	Δ [pp]
Baseline	72.5	–
Assistant responses visible	94.2	+21.7
Restoring markers	78.7	+6.2
Lowercased user message	71.9	-0.6
First user turn only	68.8	-3.7
Turn counts only	36.4	-36.1
Turn counts + user word count	38.8	-33.7

Ablations: classifier input. Table 3 examines which parts of a conversation carry source information, using the three-way WC-1M + LMSYS + ShareChat combination (baseline accuracy is 72.5%). Keeping assistant responses unmasked raises accuracy to 94.2%, and restoring dataset-specific markers raises it to 78.7%, showing that assistant text and markers add further cues. Nevertheless, much of the signal is already in the user's opening request: the first user message alone yields 68.8%. Lowercasing has little effect (71.9%); using turn counts and user word counts reaches only 38.8%, just above the chance of 33.3%. From this set of input ablations, we suggest that source information lies mainly in the user-side text input.

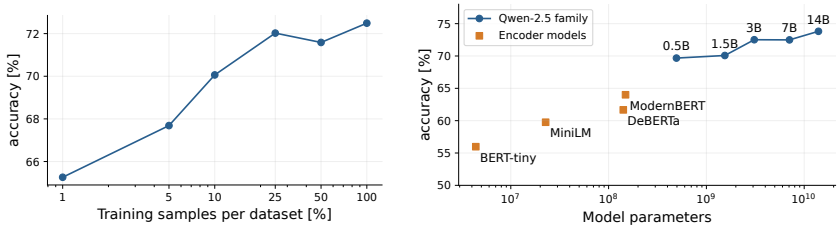


Figure 3: Accuracy on WC-1M + LMSYS + ShareChat across train set (left) and model sizes (right).

Ablations: training data and model size. We observe that accuracy grows with training data (Figure 3, left): using 1% of the training split (400 conversations per dataset) already yields 65.3%, rising to 72.5% with the full split. Accuracy also grows with model size, from 69.7% for Qwen2.5-0.5B to 73.8% for Qwen2.5-14B (Figure 3 right). Encoders from other model families (Turc et al., 2019; Wang et al., 2020; He et al., 2021; Warner et al., 2025) also perform well above the chance of 33%. The classifiers therefore appear to learn generalizable dataset-specific features, rather than memorize or overfit to training examples (Liu & He, 2025).

2.3 Separability Beyond Human-Designed Categories

Human-designed taxonomies make conversation datasets easier to interpret by organizing interactions along dimensions such as conversation topic and user follow-up behavior (Shelby et al., 2025; Shaikh et al., 2025). Prior work compares datasets through these category distributions (Mireshghallah et al., 2024; Longpre et al., 2026). How much of the dataset separability in the previous section do these categories explain? We test the relationship with experiments in both directions. First, we match datasets on their human-designed category distributions and ask whether neural classifiers can still separate them. Second, we match datasets on classifier scores and ask whether their category distributions still differ.

Human-designed categories. We annotate WC-1M, LMSYS, and ShareGPT conversations with gpt-5.6-luna along three dimensions of the AI Observatory taxonomy (Longpre et al., 2026): 'Function' (label examples: Content generation, Info analysis), 'Topic' (e.g., Technology, Health & Relationships), and 'Multi-turn' (e.g., Extend/deepen prior task). We summarize each dataset by its label distributions on these dimensions and by two additional structural distributions: the number of turns and the number of user message tokens. Annotation and matching details are in Appendix C.

Matching categories only modestly reduces separability. For each dataset pair, we use integer linear programming (Huangfu & Hall, 2018) to select equally sized subsets with matching category distributions. We do this separately within the training, validation, and test splits, then train and evaluate a classifier on these subsets. As shown in Table 4, matching lowers accuracy by only 3.0 – 6.0 points, leaving it at 74.9 – 83.2% (chance is 50%). This result shows that even when datasets agree on these interpretable summaries, classifiers readily separate them.

Table 4: Binary classification accuracy (%) after matching category distributions (chance is 50%). From left to right, columns add matched dimensions cumulatively: F (Function), T (Topic), M (Multi-turn), and S (Structural: numbers of turns and user tokens).

	Accuracy (%)				
Pair	Original	F	F+T	F+T+M	F+T+M+S
WC-1M / LMSYS	78.7	77.3	75.4	75.4	74.9
WC-1M / ShareGPT	86.2	83.5	82.9	83.8	83.2
LMSYS / ShareGPT	83.4	80.3	79.9	77.2	77.4

Matching classifier scores reduces category differences. Conversely, we score each conversation with a trained dataset classifier and select subsets with matching score distributions, without using any taxonomy labels. The total variation distance (TVD) between the datasets' category distributions then decreases for every pair and dimension (Table 5). It sometimes approaches the noise floor, the TVD between two random halves of the same dataset obtained via bootstrapping. For LMSYS + ShareGPT combination, for instance, TVD for 'Function' taxonomy falls from 13.05 to 1.57 points against a noise floor of 1.27.

In summary, the experiments show that the two directions are asymmetric. Matching on classifier scores removes much of the category difference (Table 5), but matching on (human-designed) categories leaves most of the classifier-observable signal intact (Table 4). Classifiers thus capture much of what the existing taxonomy describes, plus distinctions it does not. Therefore, we should use the dataset classification task to measure separability, while acknowledging that taxonomies offer a complementary, interpretable view of specific dataset differences.

Table 5: Distributional differences on human-designed categories after matching classifier score distributions. Each entry shows TVD points, (before matching) $\rightarrow$ (after matching), followed by the noise floor (TVD between two random halves of the same dataset, obtained via median of 1,000 bootstraps); lower TVD means more similar.

Pair	F (Function)	T (Topic)	M (Multi-turn)
WC-1M / LMSYS	5.45 $\rightarrow$ 4.11 vs. 1.11	7.32 $\rightarrow$ 3.57 vs. 1.18	5.47 $\rightarrow$ 5.24 vs. 1.11
WC-1M / ShareGPT	11.09 $\rightarrow$ 4.84 vs. 1.41	6.70 $\rightarrow$ 2.98 vs. 1.41	13.49 $\rightarrow$ 5.41 vs. 1.45
LMSYS / ShareGPT	13.05 $\rightarrow$ 1.57 vs. 1.27	7.25 $\rightarrow$ 2.02 vs. 1.28	17.96 $\rightarrow$ 4.31 vs. 1.30

3 Implications For User Modeling

User models are increasingly employed to elicit LLM behaviors at scale by simulating human–LLM interactions (OpenAI, 2026; Johnson et al., 2026). To improve simulation fidelity, user models are trained on recorded interactions: Kong et al. (2024) train on ShareGPT, Wang et al. (2025) on LMSYS, Naous et al. (2026) on WC-1M, and Wang et al. (2026) on PRISM. Grounding user models on real data gives them an empirical basis, but leaves open a central question: how much does the choice of dataset shape the user being simulated? Section 2 showed that user messages carry substantial information about their source, including differences not captured by human-designed taxonomy summaries. A model trained on these conversations may learn dataset-specific regularities. If regularities persist through training, models given the same simulation target could produce different human–LLM interactions depending on their training data.

Dataset classification results make these possibilities salient, so we examine three connected questions. First, do user models carry their training dataset's signature into the conversations they generate (Section 3.1)? Second, how does the choice of training dataset affect user models' performance across datasets (Section 3.2)? Third, can dataset classifiers help select training data that improves a user model's fit to a target dataset (Section 3.3)? Together, these questions cover the generation, evaluation, and training of user models.

User model training. Following Naous et al. (2026), we train intent-conditioned user models. We first annotate each conversation with a high-level intent summarizing the user's goal. We then fine-tune a separate copy of the same base language model on each dataset to produce the next user turn, or end the conversation, given the intent and the preceding dialogue. Every user turn serves as a prediction target for supervised fine-tuning, with earlier user and assistant turns as context. This formulation trains the user model to initiate, continue, and terminate synthetic conversations. We use Qwen2.5-7B-Instruct and Llama-3-8B (Grattafiori et al., 2024) as base models (Appendix B).

3.1 Dataset Signatures Persist in Simulated Conversations

Do user models carry their training dataset's signature into the conversations they generate? If so, the choice of training data shapes the synthetic conversations used to probe an assistant and, potentially, shapes the assistant behaviors those conversations elicit and the conclusions drawn from them.

Method. For each dataset pair in Table 6, we condition both dataset-specific user models on the same set of conversation intents, drawn from a third dataset used to train neither. Each user model converses with the same fixed assistant, so the two resulting sets of synthetic conversations differ only in the user model. We then apply the dataset classification task in two settings. In the *synthetic-to-synthetic* setting, the classifier is trained on the train split of synthetic conversations to distinguish the two user models. In the *real-to-synthetic* setting, it is trained on the train split of two original datasets and tested on the synthetic conversations (Mansour & Heckel, 2025). Both settings use the same held-out synthetic test conversations and the probe setup of Section 2.2, using Qwen2.5-7B embeddings with a trainable logistic regression head.

Table 6: Dataset classification accuracy (%) on synthetic conversations with Qwen-based user models (chance is 50%). For each pair, both user models are conditioned on intents from the listed intent source. Real → real is a reference accuracy, training and testing on real conversations. Real → synth.: train on real, test on synthetic. Synth. → synth.: train and test on synthetic.

Dataset pair	Real → real	Intent source	Real → synth.	Synth. → synth.
WC-1M / WC-4.8M	62.8	HH-RLHF	59.3	62.9
WC-4.8M / LMSYS	85.1	WC-1M	67.4	74.5
WC-4.8M / ShareChat	90.3	WC-1M	73.7	79.3
LMSYS / ShareChat	88.7	WC-4.8M	70.0	75.8
LMSYS / ShareGPT	83.4	WC-4.8M	63.6	72.8
ShareChat / ShareGPT	86.8	LMSYS	69.7	74.6
ShareChat / Arena	84.1	LMSYS	66.7	70.3
ShareGPT / Arena	85.6	ShareChat	64.9	73.3
ShareGPT / HH-RLHF	97.1	ShareChat	72.2	85.7
Arena / HH-RLHF	96.5	ShareGPT	75.1	91.5

Results. As shown in Table 6, all ten pairs are distinguishable well above chance in both settings. Synthetic-to-synthetic accuracy ranges from 62.9% to 91.5%, and real-to-synthetic accuracy from 59.3% to 75.1% (chance is 50%). The synthetic-to-synthetic results show that user models produce distinguishable conversations despite being provided with identical intents and the same assistant. The real-to-synthetic results, although achieving accuracies lower than synthetic-to-synthetic, show that these differences retain characteristics of the original training datasets. The WC-1M / WC-4.8M pair remains the least separable, and pairs involving HH-RLHF are the most separable. Thus, synthetic conversations inherit the signatures of their user model's training dataset. These results highlight that conclusions about assistant behavior drawn from simulation should be tested across user models trained on different datasets.

3.2 Cross-dataset Generalization

The persistence of dataset signatures in synthetic conversations raises a further question: how well do user models generalize beyond their training source? Liu & He (2025) and Tommasi et al. (2015) showed that strong within-dataset performance need not translate into strong performance across datasets. We test this for user models by evaluating each dataset-specific user model (Section 3.1) on the test split of every dataset. We use the perplexity of user turns as a measure of predictive fit (Shuster et al., 2021; Moon et al., 2026).

Table 7: Cross-dataset evaluation. Rows indicate training and columns indicate test (target) datasets. Each entry represents mean per-token perplexity measured on the test split. Bold and underlined entries indicate the lowest and second-lowest perplexity in each column, respectively.

Train ↓ Test →	WC-1M	WC-4.8M	LMSYS	Arena	ShareChat	ShareGPT	HH-RLHF
WC-1M	**6.41**	<u>6.70</u>	<u>6.14</u>	6.15	8.38	<u>6.02</u>	5.39
WC-4.8M	<u>6.53</u>	**6.36**	6.17	<u>6.06</u>	<u>8.28</u>	6.07	5.46
LMSYS	6.61	6.80	**5.43**	6.17	8.31	6.03	<u>5.19</u>
Arena	6.69	6.82	6.22	**5.75**	8.29	6.14	5.29
ShareChat	6.71	6.85	6.22	6.13	**7.33**	6.09	5.30
ShareGPT	6.65	6.92	6.21	6.26	8.40	**5.62**	5.28
HH-RLHF	7.05	7.38	6.72	6.80	9.07	6.48	**4.53**

Results. Table 7 reports results for user models fine-tuned from Qwen2.5-7B-Instruct (Llama-3-8B results in Table 10). On every test dataset, the user model trained on that dataset achieves the lowest perplexity, producing a diagonal pattern. This in-domain advantage shows that performance depends strongly on the similarity between training and evaluation sources. Model rankings therefore change with the evaluation dataset, and no single user model is best across all targets. An evaluation on one dataset measures fit to that dataset, and assessing transfer requires evaluation across datasets.

HH-RLHF provides an interesting example of the distinction between in-domain fit and cross-dataset transfer. The user model trained on HH-RLHF achieves an in-domain perplexity of only 4.53 (well below other in-domain perplexities of 5.43 – 7.33), yet has the worst performance on every other dataset. Its low in-domain perplexity may just indicate dataset-specific regularities, such as formulaic crowdworker prompts. Differences in in-domain perplexity across datasets should be interpreted cautiously: they may reflect how predictable each dataset's conversations are, rather than how well each model simulates users.

Separability correlates with the transfer gap. We next ask whether the separability of two datasets predicts how poorly user models transfer between them. For each of the 21 dataset pairs (A, B), we train a binary linear probe on Qwen2.5-7B representations, recording its accuracy on the test split (chance is 50%). We compare this accuracy with the symmetric transfer gap Δ_{AB} in Equation 1 where $\mathrm{NLL}(A \to B)$ is the mean per-token negative log-likelihood (i.e., log perplexity) that a user model trained on A assigns to the user turns of B. Δ_{AB} is the excess loss incurred by using the out-of-domain user model in place of the in-domain one, averaged over both directions.

$$\Delta_{AB} = \frac{1}{2}\Big[\big(\mathrm{NLL}(A \to B) - \mathrm{NLL}(B \to B)\big) + \big(\mathrm{NLL}(B \to A) - \mathrm{NLL}(A \to A)\big)\Big]. \quad (1)$$

Figure 4 shows a strong positive correlation: Pearson $r = 0.80$ (Spearman $\rho = 0.72$) for Llama-based user models and $r = 0.86$ ($\rho = 0.81$) for Qwen-based user models ($p < 10^{-3}$ in all cases). The least separable pair, WC-1M vs. WC-4.8M (accuracy 63%), has the smallest gap, whereas pairs involving HH-RLHF (accuracy $\geq$ 95%) have among the largest. Qwen-based user models have consistently smaller transfer gaps than Llama-based ones but preserve a similar ordering across pairs. We interpret both quantities as reflecting the same underlying factor, the distance between the two datasets' distributions. A probe on language representations is a cheap proxy for this distance, and hence for the cost of transferring a user model from one dataset to the other.

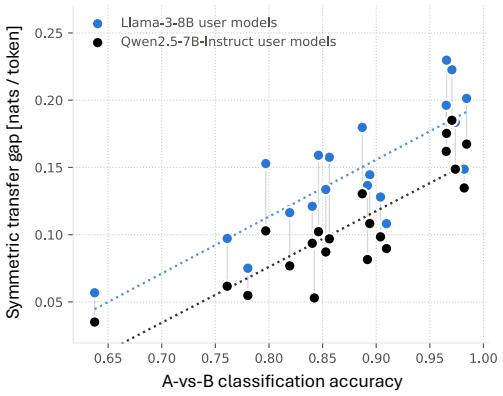


Figure 4: Symmetric transfer gap Δ_{AB} correlates with the A-to-B dataset classification accuracy.

3.3 Classifier-guided Data Selection

Section 3.2 showed that the choice of training conversations affects a user model's fit to a target dataset, and that separability correlates with this effect. Can dataset classifiers, then, help select useful training data for a given target dataset under a fixed training budget?

Method. We adapt data selection via importance resampling (DSIR; Xie et al. (2023); Chowdhery et al. (2023)), replacing its n-gram importance weights with weights from a dataset classifier. For each *target* dataset, we pool the training splits of the other six datasets into a *donor* pool. We train a binary classifier to distinguish target conversations from donor conversations, and select it by balanced validation accuracy. Let $s_T(c)$ be its predicted probability that donor conversation c belongs to the target. We sample donor conversations without replacement, with probability proportional to

$$w_T(c) = \frac{s_T(c)}{1 - s_T(c)} = \frac{\text{Prob(classifies as target)}}{\text{Prob(classifies as donor)}}, c \in \text{donor pool} \tag{2}$$

We compare this classifier-guided data selection (labeled 'DSIR' in Figure 5) with uniform random sampling from the donor pool, each at 50% and 100% of an original training budget. Both conditions train the user model exclusively on donor conversations; target data is used only to fit the classifier.

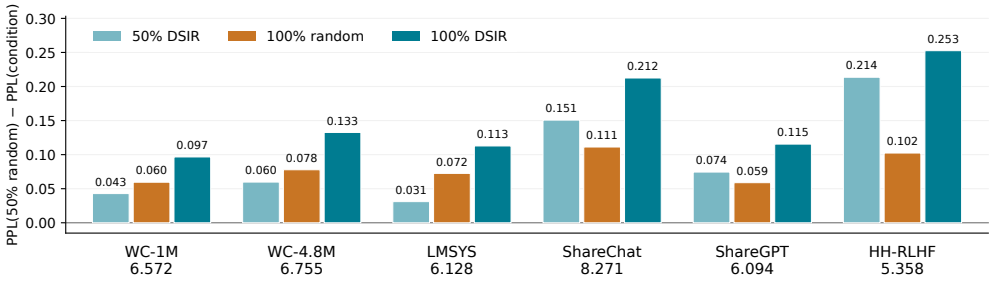


Figure 5: Perplexity reduction relative to uniform random sampling at 50% of the training budget, PPL(50% random) - PPL(condition), for Qwen2.5-7B-Instruct user models (higher is better). The number under each dataset is the 50%-random perplexity. Classifier-guided selection outperforms random sampling at every matched budget. Llama-3-8B results are in Figure 7.

Results. At matched training budgets, classifier-guided selection yields lower target perplexity than random sampling for every target (Figure 5). For three of the six targets, selection at 50% budget also beats random sampling at 100%. For HH-RLHF, for example, it reduces perplexity from 5.36 to 5.14 at the 50% budget, compared with 5.26 for random sampling with twice the data. These results show that the classifier identifies donor conversations that better support fitting the target dataset.

4 Related Work

4.1 Dataset Signatures

Dataset classification experiments and cross-dataset evaluations have exposed source-specific regularities in vision datasets (Torralba & Efros, 2011; Tommasi et al., 2015), including modern large-scale collections (Liu & He, 2025). Such regularities make within-dataset performance an unreliable guide to generalization. NLP has analogous findings: annotation protocols introduce predictive shortcuts (Gururangan et al., 2018), and corpus filtering changes which content and populations are disproportionately represented (Dodge et al., 2021). Close to our work, Mansour & Heckel (2025) identify fingerprints of web-filtered corpora and show that they propagate through LLM pretraining. However, human–LLM dataset releases often emphasize diverse real-world interactions (Zhao et al., 2024; Zheng et al., 2024) yet dataset separability in this setting is less understood. We address this gap by examining three aspects: dataset signatures beyond human-designed taxonomy categories (Shelby et al., 2025), their persistence through user model training, and their relationship to cross-dataset generalization.

4.2 Human-LLM Interaction Datasets

Human–LLM datasets support both descriptive research and model development. They inform studies of economic activity, such as the Anthropic Economic Index (Handa et al., 2025), and taxonomies of user behavior (Shelby et al., 2025; Shaikh et al., 2025). For example, AI Observatory examines how descriptions of real-world AI use vary across conversation sources (Longpre et al., 2026).

For assistant development, recorded queries and model responses provide supervision for assistant fine-tuning, as in Vicuna (Chiang et al., 2023). Additionally, preference alignment uses conversation logs paired with preference labels, motivating HH-RLHF's collection (Bai et al., 2022). For evaluation, WildBench constructs a real-world task benchmark from the WildChat collection (Lin et al., 2025).

Recently, user models (also called user simulators) have gained attention (humans&, 2026) as a scalable way to train and evaluate assistants through adaptive multi-turn interactions (Kong et al., 2024; Naous et al., 2026; Wang et al., 2026). They support evaluations of assistant behaviors emerging over extended conversations (OpenAI, 2026) and assistant training through simulated interactions (Suh et al., 2026). We therefore focus the second half of this paper on user modeling: it directly transforms recorded user behavior into simulated interactions used for downstream experiments. This opens up the possibility that the choice of training dataset is consequential for both the simulated user behavior and the conclusions about assistant behaviors drawn from simulation.

5 Conclusion, Limitations, and Future Work

By revisiting the dataset classification experiment of Torralba & Efros (2011), we show that human-LLM interaction datasets contain source signatures that survive marker removal, assistant masking, and matching on taxonomy and structural distributions. Then, we investigate the implications of dataset separability for user modeling. Dataset signatures persist in the conversations generated by user models, and the evaluation of user models depends on which dataset we use to evaluate them. Dataset classifiers also have a practical use: classifier-guided selection improves a user model's fit to a target dataset. Together, these findings demonstrate that dataset choice is consequential both for the behavior being simulated and for the evidence used to assess it. Our results motivate documenting collection context and testing research findings across datasets.

Our study has several limitations. We study English conversations from public datasets, so application of our findings to multilingual, cross-cultural domains is an open problem. We also note that dataset separability neither identifies one representative dataset nor isolates the causes of source differences: recruitment, incentives, assistant models, collection periods, and more can vary together. Also, masking the assistant response cannot remove its influence on subsequent user messages. Matching selected category summaries does not control for all semantic or joint distributional differences. Perplexity measures predictive fit, rather than human judgments of simulation fidelity which should be verified by a separate user study process, and we do not establish whether source differences change downstream conclusions about elicited assistant behaviors. Future work should disentangle collection effects, extend evaluation across languages and populations, and test whether better predictive fit yields more reliable assistant evaluations.

AI USE STATEMENT

In this work, we used generative AI tools for: (1) aid and polish writing, (2) retrieval and discovery of related works, (3) research execution, especially code implementation, and (4) generating synthetic conversations between trained user language models and LLM assistants. We have not used generative AI tools for: (5) drafting sections of the paper, and (6) for proving mathematical claims. We have reviewed all AI-assisted work. Authors made final decisions on accepting or rejecting AI's suggested edits. Authors confirmed that retrieval of related works is accurate and relevant. Authors reviewed and tested code implementations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

It is important that the work published in ICLR is reproducible. We release our code at Anonymous GitHub. We document the full pipeline in the Appendix Sections A through C.

REFERENCES

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Serina Chang, Ashton Anderson, and Jake M. Hofman. ChatBench: From static benchmarks to human-AI evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26009–26038, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1262. URL https://aclanthology.org/2025.acl-long.1262/.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113, 2023.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 1286–1305, 2021.

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

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761*, 2025.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Rebecca MM Hicke and Kiran Tomlinson. Adopt != adapt: Longitudinal analyses of llm conversations in the wild. *arXiv preprint arXiv:2605.29018*, 2026.

Q. Huangfu and J.A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical programming computation*, 10(1):119–142, 2018.

humans&. Persimmon model card, September 2026. URL https://persimmon.humansand.ai/blog/. Research preview, version 0.1.

Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximillian Chen, and Tianmin Shu. Thoughttrace: Understanding user thoughts in real-world llm interactions. *arXiv preprint arXiv:2605.20087*, 2026.

Daniel D. Johnson, Robert Friel, Cassidy Laidlaw, Luke Hewitt, Nari Johnson, Conrad Stosz, and Sarah Schwettmann. Mental health behavior report. *Transluce Behavior Reports*, 2026. URL https://behaviors.transluce.org/mental-health.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, pp. 427–431, 2017.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. Platolm: Teaching llms in multi-round dialogue via a user simulator. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7841–7863, 2024.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *International Conference on Learning Representations*, volume 2025, pp. 47852–47870, 2025.

Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? In *International Conference on Learning Representations*, volume 2025, pp. 41433–41449, 2025.

Shayne Longpre, Anka Reuel, Dayeon Ki, Cathy Mengying Fang, Megan Richards, Zhiping Zhang, Chuanyang Jin, Jenn Mickel, Cedric Deslandes Whitney, Ahmet Üstün, et al. The ai observatory: A public measure of real-world ai use. 2026.

Youssef Mansour and Reinhard Heckel. Measuring fingerprints of web-filtered text datasets and fingerprint propagation through training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-0557. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/181f409676bf93986ecdb61e6af254a1-Paper-Datasets_and_Benchmarks_Track.pdf.

Hunter McNichols, Fareya Ikram, and Andrew Lan. The studychat dataset: Analyzing student dialogues with chatgpt in an artificial intelligence course. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pp. 53–63, 2026.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*, 2024.

Suhong Moon, Minwoo Kang, Joseph Suh, Mustafa Safdari, and John Canny. Identity, cooperation and framing effects within groups of real and simulated humans. *arXiv preprint arXiv:2601.16355*, 2026.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. In *ICLR*, 2026.

OpenAI. GPT-5.6 preview system card. Technical report, OpenAI, June 2026. URL https://deploymentsafety.openai.com/gpt-5-6-preview/gpt-5-6-preview.pdf. Accessed 2026-09-21.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. Navigating rifts in human-llm grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20832–20847, 2025.

Renee Shelby, Fernando Diaz, and Vinodkumar Prabhakaran. Taxonomy of user needs and actions. *arXiv preprint arXiv:2510.06124*, 2025.

Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. Dialogue in the wild: Learning from a deployed role-playing game with humans and bots. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 611–624, 2021.

Jonathan Stray, David Zhai Yang, Steven Luo, Miu Nicole Takagi, and Serina Chang. Political neutrality as balanced approval: A large-scale human evaluation of ai responses. *arXiv preprint arXiv:2605.28911*, 2026.

Joseph Suh, Ayush Raj, Minwoo Kang, and Serina Chang. Quantifying the utility of user simulators for building collaborative llm assistants. *arXiv preprint arXiv:2605.09808*, 2026.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. Parrot: Enhancing multi-turn instruction following for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9729–9750, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.525. URL https://aclanthology.org/2024.acl-long.525/.

Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias, 2015. URL https://arxiv.org/abs/1505.01257.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019. URL https://arxiv.org/abs/1908.08962.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21082–21107, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Yingshan Susan Wang, Cedegao E Zhang, Linlu Qiu, Zexue He, Pengyuan Li, Alex Pentland, Roger P Levy, and Yoon Kim. Learning user simulators with turing rewards. *arXiv preprint arXiv:2606.19336*, 2026.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2526–2547, 2025.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227, 2023.

Yueru Yan, Tuc Nguyen, Bo Su, Melissa Lieffers, and Thai Le. Sharechat: A dataset of chatbot conversations in the wild. *arXiv preprint arXiv:2512.17843*, 2025.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 34590–34605, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *International Conference on Learning Representations*, volume 2024, pp. 22225–22257, 2024.

A Dataset Curation

A.1 Datasets

Datasets are ordered chronologically.

HH-RLHF (Bai et al., 2022) is a human preference dataset collected by Anthropic through crowdworker interactions with AI assistants under two task conditions: helpfulness and harmlessness. In the helpfulness condition, participants engaged in open-ended conversations involving text-based tasks, such as question answering, writing, and planning. At each assistant turn, participants compared two candidate responses and selected the more helpful response. In the harmlessness condition, participants were instructed to adversarially probe the models to elicit harmful responses and select the more harmful candidate. Conversations continued from the selected response, allowing participants to produce multi-turn interactions. Participants received task-specific instructions and examples but retained flexibility in choosing conversation topics.

ShareGPT contains user-shared ChatGPT conversations collected from sharegpt.com, rather than conversations elicited through a study-specific participant task protocol. ShareGPT conversations were subsequently used to train Vicuna (Zheng et al., 2023). Data filtering included converting HTML content to Markdown, filtering inappropriate or low-quality examples, and splitting lengthy conversations into segments compatible with the model's context window. Because multiple snapshots and processed derivatives circulate under the ShareGPT name, we identify our input as the two HTML-cleaned JSON files in HuggingFace anon8231489123/ShareGPT_Vicuna_unfiltered, containing 77K conversation records before our curation.

LMSYS-Chat-1M (Zheng et al., 2024) was collected from voluntary user interactions on the Vicuna demonstration and Chatbot Arena website between April and August 2023. The platform provided free, registration-free access through three interfaces: a single-model chat, a comparison between two randomly selected anonymous models, and a comparison between two user-selected models. The dataset comprises one million conversations with 25 models, originating from approximately 210K unique users. Before release, the researchers attempted to remove conversations containing personally identifiable information (PII) and added message-level moderation annotations. The released corpus contains conversation logs and model metadata, without the human preference labels available in a separate Chatbot Arena dataset.

Arena Human Preference 140K extends the collection setting of LMSYS-Chat-1M by pairing user–model interactions with explicit human preference judgments. Whereas LMSYS-Chat-1M contains conversations collected across single-model and comparison interfaces in 2023 without released preference annotations, Arena Human Preference 140K contains text-based model comparisons collected between April and July 2025. Users submitted their own prompts and voted for either anonymous model, a tie, or dissatisfaction with both. The updated interface allowed users to continue a conversation after voting, with a newly sampled model pair receiving prior conversational context. Each record therefore represents a preference judgment within a potentially longer session, rather than an independent conversation. The release contains 135,634 vote records with paired transcripts, model identifiers, and session metadata.

WildChat-1M (Zhao et al., 2024) was collected through publicly accessible chatbot services hosted on HuggingFace Spaces and powered by GPT-3.5-Turbo and GPT-4. Users received free access after explicitly consenting to the collection and sharing of their interactions. Conversations are based on users' requests, not on researcher-assigned topics. The collection spans April 2023 to May 2024 and includes both single-turn and multi-turn interactions. Individual conversation records were linked into conversations using conversation histories, IP addresses, and request headers. Before release, the researchers applied automated procedures to redact PII, hashed IP addresses, and derived coarse geographic metadata. After reserving a subset for benchmark development, the release contained approximately 1 million conversations.

WildChat-4.8M (Zhao et al., 2024) extends the WildChat-1M collection through continued logging of voluntary interactions with publicly accessible chatbot services, with coverage extending through July 2025. The expanded collection incorporates additional OpenAI model families, including GPT-4o, GPT-4.1-mini, and the o1 reasoning models. Release preparation included automated privacy redaction, scanning for exposed secrets, and content moderation. The collection originally comprised approximately 4.8 million conversations; the public release named WildChat-4.8M applies toxicity filtering and contains approximately 3.2 million conversations.

ShareChat (Yan et al., 2025) was constructed by retrospectively collecting publicly shared conversations from five commercial chatbot platforms: ChatGPT, Claude, Gemini, Perplexity, and Grok. The researchers identified shared conversation URLs through the Wayback Machine using platform-specific URL patterns, then employed automated browser interactions to retrieve rendered pages and extract structured conversation records. Users had independently conducted and shared these conversations, so the collection did not involve researcher-assigned conversational tasks. Extraction preserved ordered user–assistant exchanges and available platform-specific features, including citations, visible thinking traces, code artifacts, timestamps, and model identifiers. Personally identifiable information was redacted using automated tools, with an additional model-based audit of the redaction results. The reported corpus contains 143K conversations and 660K interaction turns, spanning 95 languages and usage between April 2023 and October 2025.

A.2 Curation and split construction

We convert each source to alternating user–assistant messages. HH-RLHF uses the chosen transcript according to the preference label. For Arena, we retain one longest released history per evaluation session; the stored assistant text is the displayed left-side view, rather than a reconstructed sequence of the responses selected by the user. ShareChat records with ambiguous message ordering are excluded.

We use released language labels when available and fastText (Joulin et al., 2017) language identification otherwise. Table 8 reports the size of input dataset snapshots and retained sample sizes.

Table 8: Curation and sampling counts for the snapshots used in this study. Input counts are raw records. English counts follow conversation reconstruction, structural validation, and language filtering. Eligible counts additionally reflect duplicate/template filtering and cross-source exclusions.

Dataset	Input	English	Eligible	Retained
HH-RLHF	169,352	168,805	145,740	50,000
ShareGPT	76,920	54,352	50,790	50,000
LMSYS-Chat-1M	1,000,000	766,545	415,087	50,000
Arena 140K (2025)	135,634	60,810	57,184	50,000
WildChat-1M	837,989	473,246	285,771	50,000
WildChat-4.8M	3,199,860	1,672,053	225,212	50,000
ShareChat	129,584	79,453	75,254	50,000

Duplicate and template filtering. For duplicate matching, we lowercase user text and collapse whitespace. We remove exact duplicates of concatenated user text within each source, and filter opening prompt templates using 7-grams occurring more than 100 times within a source. WC-4.8M excludes conversation hashes present in WC-1M. Full user texts of at least five words appearing in multiple sources are excluded as well.

B User Model Training

Intent Generation. For each conversation, we generate a high-level user intent summary, following Naous et al. (2026). Specifically, we provide the full human–LLM conversation trajectory to Qwen3-32B and instruct it to produce a short summary of the user's high-level intent. We refer the reader to the "Prompt Template for Intent Generation" in Naous et al. (2026) for the exact prompt.

Supervised Fine-Tuning. We perform supervised fine-tuning (SFT) on the user-side utterances of the training split. Below, we present a training sample from WC-1M formatted with the Qwen-2.5 chat template, where highlighted spans indicate the tokens over which the loss is computed:

```
<|im_start|>system
You are a user chatting with an assistant language model to convert a
website into a PDF format.<|im_end|>
<|im_start|>user
To make the following website as an PDF<|im_end|>
<|im_start|>assistant
To convert a website to a PDF, you can follow these steps:
...(truncated) ...but not its dynamic or interactive elements.<|im_end|>
<|im_start|>user
Can you do that for me?<|im_end|>
<|im_start|>assistant
Apologies for the confusion, but as a text-based AI, ...(truncated) ...to
assist you.<|im_end|>
<|im_start|>user
<|endconversation|><|im_end|>
<|endoftext|>
```

We explain some details. First, the prepared intent ("You are a user chatting…") is placed in the system prompt to condition generation. Second, the loss is applied not only to user-turn content but also to the turn-ending special token (`<|im_end|>` for Qwen2.5-family models), so that the model learns to produce delimited turns. Third, following Naous et al. (2026), we introduce a special token `<|endconversation|>` to signal that the user has terminated the conversation: a real human–LLM conversation with $2n$ turns is structured into $2n+1$ turns with $n+1$ user-side utterances (the last one always being `<|endconversation|>`) and n LLM responses. We initialize the embedding of the special token with the pretrained EOS token embedding.

Synthetic Conversation Generation. The fingerprint experiments use a fixed Qwen3.5-9B assistant with thinking disabled. Only the user model sees the intent; the assistant receives the generated dialogue without an additional system prompt. Each intent produces one conversation per user model. The initial user request is generated, rather than copied from the seed conversation. Train, validation,

and test splitting follows the seed intent's original split, and the two synthetic conversation datasets share the same intents within each split.

User sampling uses temperature 0.7, top-p 0.9, and no top-k restriction. Assistant sampling uses temperature 0.7, top-p 0.8, top-k 20. The served user-model context is 8,192 tokens. A conversation stops at the user termination token, four completed assistant replies, or the context limit. A user utterance identical to the immediately preceding user utterance is resampled up to seven times, increasing temperature by 0.05 per retry to a maximum of 1.0.

C TAXONOMIZATION AND CLASSIFIER-SCORE MATCHING

C.1 TAXONOMY ANNOTATIONS

The three factors describe what a user asks the assistant to do (Function), what the request is about (Topic), and how a follow-up relates to the preceding task (Multi-turn), using the AI Observatory taxonomy (Longpre et al., 2026). Function and Topic each aggregate to nine broad categories. The four follow-up relationships are extending/deepening the prior task, retrying/revising it, requesting a variant, and making a completely new request. Figure 6 shows the measured category prevalences.

Annotations. We apply the AI Observatory taxonomy using gpt-5.6-luna with a confidence threshold of 0.7. The annotation procedure is based on their official codebase without any modification.

C.2 MATCHING ATTRIBUTES

Maximum-size attribute matching. Let $x_i, y_j \in \{0, 1\}$ indicate selection from sources A and B, and let f_k be a selected attribute. We solve

$$\begin{aligned} \max_{x,y} \ & \sum_i x_i \\ \text{subject to } & \sum_i x_i = \sum_j y_j, \\ & \sum_i f_k(i) x_i = \sum_j f_k(j) y_j \quad \text{for every selected } k. \end{aligned} \tag{3}$$

For the structural factor, the features are user-turn count and total user-text token count. Equal conversation counts and equal feature totals imply equal mean user turns per conversation and equal total user tokens divided by total user messages. User-text tokens are counted with `cl100k_base`.

Multinomial regression head training. We reuse frozen base Qwen2.5-7B representations with assistant masking and a 32,768-token limit, considering mean and last-token pooling at layers 6, 12, 18, and 28. Training-only feature standardization, full-batch Adam for 300 updates at learning rate 0.05, and regularization strengths of $10^{-4}, 10^{-3}, 10^{-2}$ follow the main dataset classification experiment. Validation balanced accuracy selects checkpoints, regularization, and representation separately for each condition.

C.3 WITHIN-DATASET SPLIT-HALF BOOTSTRAPPING

To quantify the noise floor of taxonomy category distributions, we repeatedly divide each original test dataset and each pair-specific score-matched dataset into two disjoint halves. Each of 1,000 repetitions draws a random permutation and takes two groups of $\lfloor N/2 \rfloor$ conversations. We repeatedly compute the distributional difference between two halved subsets and calculate the median of bootstrap results to obtain the estimate of noise floor.

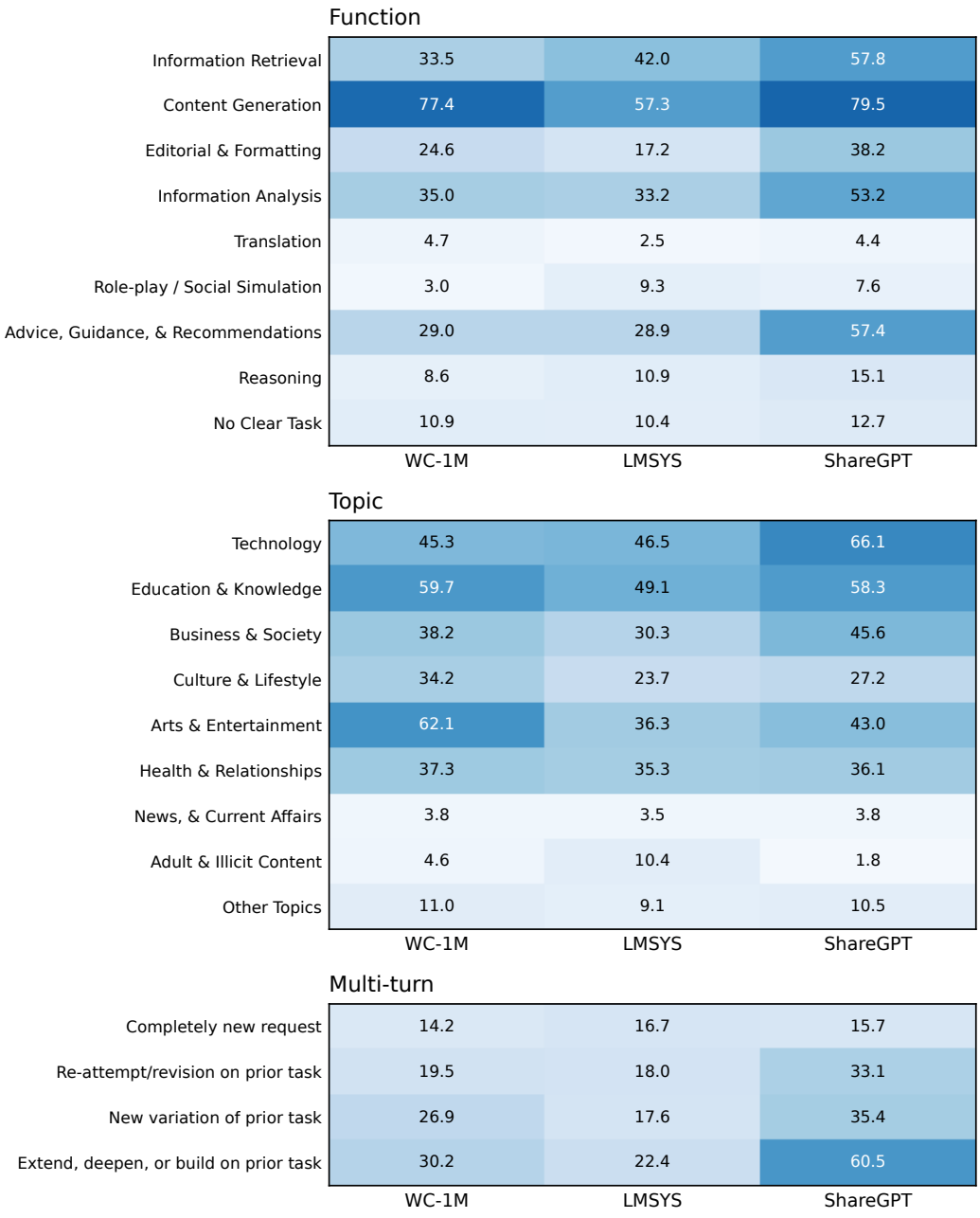


Figure 6: Percentage of conversations containing each annotated category in the original held-out WC-1M, LMSYS, and ShareGPT cohorts. A conversation can contain several categories, so columns need not sum to 100%. Each factor uses only conversations with complete accepted annotations.

D Additional Results

D.1 More ablations

Ablations: Co-training language models with classifier heads. Co-trained models outperforms their corresponding frozen language model configurations by 6.8–11.3 percentage points (Table 9). These gains indicate that frozen language model representations leave some source information unrecovered.

Table 9: Frozen language models vs. co-trained classifiers on WC-1M + LMSYS + ShareChat.

Model size	Frozen	Fine-tuned	Δ [%]
0.5B	69.7	80.3	+10.6
1.5B	70.1	81.4	+11.3
3B	72.5	80.2	+7.7
7B	72.5	79.3	+6.8

D.2 Cross-dataset Generalization

Table 10: Cross-dataset evaluation. Rows indicate training and columns indicate test (target) datasets. Each entry represents mean per-token perplexity measured on the test split. Bold and underlined entries indicate the lowest and second-lowest perplexity in each column, respectively.

Train ↓ Test →	WC-1M	WC-4.8M	LMSYS	Arena	ShareChat	ShareGPT	HH-RLHF
WC-1M	**6.60**	7.05	6.72	6.65	8.89	6.70	5.78
WC-4.8M	6.78	**6.47**	6.81	6.54	8.75	6.77	5.84
LMSYS	6.93	7.30	**5.60**	6.71	8.90	6.72	5.60
Arena	7.03	7.33	6.91	**6.10**	8.93	6.93	5.76
ShareChat	7.10	7.36	6.94	6.71	**7.70**	6.85	5.83
ShareGPT	7.04	7.44	6.79	6.82	8.90	**5.93**	5.62
HH-RLHF	7.47	8.05	7.69	7.62	10.03	7.40	**4.86**

Table 10 reports the cross-dataset generalization results for user models based on Llama-3-8B.

D.3 Classifier-guided Data Selection

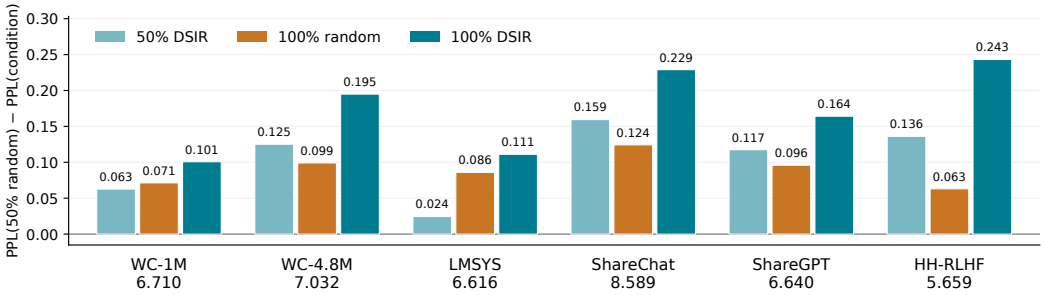


Figure 7: Perplexity reduction relative to uniform random sampling at 50% of the training budget, PPL(50% random) - PPL(condition), for Llama-3-8B user models (higher is better). The number under each dataset is the 50%-random perplexity. Classifier-guided selection outperforms random sampling at every matched budget.

Alignment in classifier scores correlates with performance gains. Section 3.2 suggests that differences captured by dataset classifiers are relevant to cross-dataset generalization. We now examine a complementary question: when classifier-guided selection reduces the discrepancy between donor and target data, do larger reductions accompany larger improvements in predictive fit? Both questions examine whether classifier-derived measures of distributional mismatch are informative about the downstream user model performance.

Adapting the Kullback-Leibler reduction metric of Xie et al. (2023), let $\hat{p}_T$, $\hat{p}_R$, and $\hat{p}_G$ denote the distributions of classifier scores (Softmax probabilities) for target, randomly selected donor, and classifier-guided donor, respectively. For each target dataset and train budget, we compute

$$\Delta_{\mathrm{KL}} = D_{\mathrm{KL}}(\hat{p}_T \| \hat{p}_R) - D_{\mathrm{KL}}(\hat{p}_T \| \hat{p}_G) \tag{4}$$

A positive Δ_{KL} indicates that guided selection brings the donor score distribution closer to the target score distribution than random sampling. We compare this quantity with the downstream improvement $\Delta_{\mathrm{NLL}} = \mathrm{NLL}_R - \mathrm{NLL}_G$, where NLL_R and NLL_G are the NLLs of user models trained on randomly selected and classifier-selected donors, respectively.

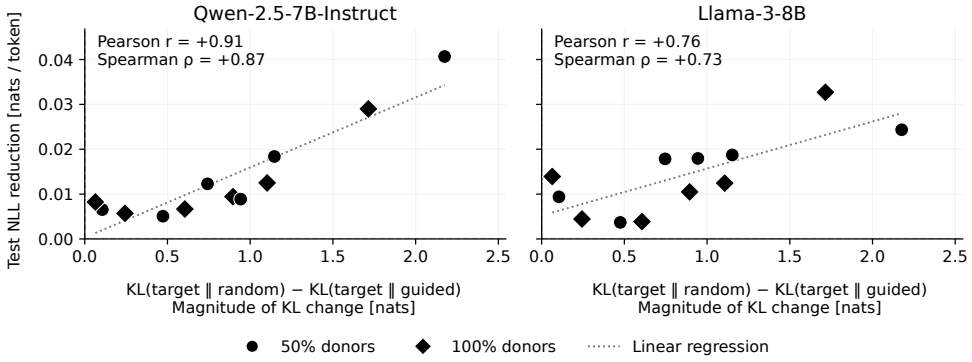


Figure 8: Magnitude of classifier-score distribution change versus downstream NLL improvement. The horizontal axis shows Equation 4, computed from trained classifier scores for target, randomly selected donor, and classifier-selected donor. The vertical axis shows target NLL reduction, Δ_{NLL}.

Figure 8 shows a strong positive association between KL reduction and NLL improvement, with Pearson correlations of r=0.91 for Qwen-based and r=0.76 for Llama-based user models. This association suggests that the characteristics captured by the dataset classifier are relevant to selecting useful user-model training data. We note that HH-RLHF target exhibits particularly large KL reductions and performance gains. One possible explanation is that its distinctive signatures, also reflected in its strong separability from other datasets (Section 3), provide a useful signal for identifying relevant conversations within the donor pool. Guided selection may therefore concentrate training on donors that share these signatures, yielding substantial improvements over uniform sampling.