A PERSONALIZED LANGUAGE MODEL SHOULD BE LESS SURPRISED BY YOU

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalization in language models is commonly evaluated through task performance, stylistic resemblance, or automatic judgments. These proxies capture useful behaviors, but they do not test whether user-specific conditioning helps a model better anticipate its target user, as reflected in lower surprisal on held-out user text. For a fixed task, tokenizer, and base model, we define *predictive personalization* as the extent to which a user-specific adaptation reduces the surprisal of text produced by the target user relative to a generic baseline. We operationalize this notion as user-specific predictive alignment and introduce Likelihood Improvement from Targeted Familiarity (LIFT), a per-token log-likelihood-ratio diagnostic that measures the length-normalized surprisal reduction on held-out user text relative to a generic baseline. We also introduce ESS-LIFT as a rare-token robustness analysis that downweights rare, high-leverage tokens to test whether conclusions drawn from LIFT depend on their influence. Since there is no scalar oracle for the latent amount of personalization achieved by an adaptation, we evaluate LIFT through construct-validity stress tests, demonstrating its sensitivity to user identity beyond contextual relevance, robustness to fluency and raw-likelihood confounds, and ranking coherence across qualitatively different personalization mechanisms. Experiments across diverse language models, datasets, and tasks provide convergent evidence that LIFT captures user-specific predictive alignment, supporting its use as a principled diagnostic for predictive personalization in language models.

1 INTRODUCTION

Large language models (LLMs) are increasingly expected to adapt to individual users in settings such as writing assistance, recommendation, and long-term interactions. Current work often treats personalization as success on specific proxy tasks [26; 11; 24; 33], stylistic resemblance [17; 13; 2; 9], or favorable automatic judgments [31; 30; 27; 6; 14], and then relies on proxy metrics to assess how well a model is "personalized". These efforts are important because they evaluate concrete behaviors that users may value, such as task success, stylistic fit, or judged response quality.

However, these observable proxies do not by themselves define the predictive aspect of personalization. A person is characterized not only by how they write or whether an output satisfies a task rubric, but also by what they believe, prefer, avoid, emphasize, and repeatedly do. Evaluating personalization only through task quality or style similarity therefore risks measuring outcomes around personalization rather than whether a model has become better at anticipating a particular individual. We focus on this measurable component, namely person-specific anticipation, and define **predictive personalization** as the extent to which a system better anticipates a specific individual than it anticipates people at the population level. This shifts evaluation from externally chosen proxy outcomes to person-specific anticipations that may include style, preferences, opinions, topical choices, and recurring behavioral patterns, without assuming in advance which dimension matters for a task.

Given this definition, we operationalize person-specific anticipation in language models as **user-specific predictive alignment**: an adaptation's predictive distribution should shift closer to the target user's task-conditioned behavioral distribution than a generic baseline does. We do not propose a new personalization mechanism. Instead, we ask how to measure whether any mechanism—retrieval, prompting, steering, fine-tuning, or another form of adaptation—has shifted a model's predictive distribution toward a specific user. For a fixed task, tokenizer, and base model, this comparison can

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.

be made directly on held-out user text by asking whether the adapted model reduces averaged token surprisal relative to the baseline.

We formalize this alignment as the expected per-token reduction in surprisal on held-out user text when moving from the generic baseline to the adaptation, and introduce *Likelihood Improvement from Targeted Familiarity (LIFT)*, a finite-sample per-token estimator that serves as a diagnostic of this quantity. As illustrated in Figure 1, LIFT compares the adapted and baseline likelihoods on the same held-out reference text, aggregating token-level surprisal reductions into a single score. On short held-out texts, rare names or idiosyncratic punctuation can receive very low probability under the baseline and exert disproportionate influence on LIFT. We therefore introduce ESS-LIFT as a finite-sample rare-token robustness analysis that downweights such high-leverage tokens. LIFT remains the primary estimator of the defined construct. For $\rho < 1$, ESS-LIFT changes the estimand and is used to assess whether conclusions are stable under rare-token downweighting.

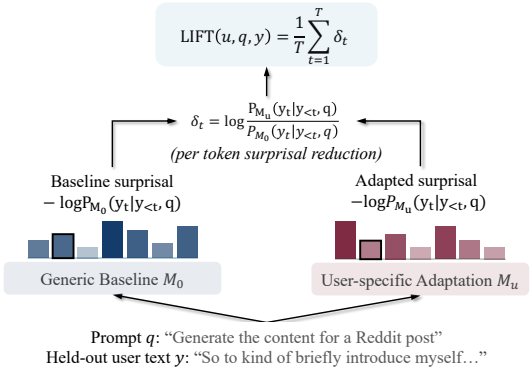


Figure 1: LIFT measures the average per-token reduction in surprisal on held-out user text by comparing a generic baseline and a user-specific adaptation on the same reference.

Two scope conditions follow directly from LIFT's design. *(i)* LIFT measures demonstrated predictive shift on held-out user text, not the causal mechanism by which that shift is achieved: a high score indicates the adaptation reduces surprisal relative to the baseline, but does not identify which feature or pathway is responsible. *(ii)* LIFT is mechanism- and task-agnostic, applicable to retrieval, fine-tuning, dialogue conditioning, or any other means, but should be interpreted comparatively within a fixed base model and tokenizer rather than as an absolute score across model families.

Since there is no scalar oracle for the latent amount of personalization achieved by an adaptation, we evaluate whether LIFT behaves as a diagnostic of predictive personalization through construct-validity stress tests. If LIFT captures user-specific predictive alignment, then it should satisfy three properties jointly, demonstrating *sensitivity to user identity* beyond contextual relevance, *robustness to surface confounds* such as fluency and generic predictability, and *ranking coherence across qualitatively different personalization mechanisms*. Targeted experiments across diverse datasets, tasks, and model families provide evidence for all three properties, jointly supporting LIFT as a principled diagnostic for predictive personalization. Our contributions are threefold:

- We define predictive personalization as person-specific anticipation and operationalize it as user-specific predictive alignment in language models.
- We introduce LIFT, a per-token log-likelihood-ratio diagnostic that estimates the average surprisal reduction on held-out user text relative to a generic baseline. We complement it with ESS-LIFT, a finite-sample rare-token robustness analysis that tests whether conclusions are robust to downweighting rare, high-leverage tokens.
- We provide construct-validity evidence across diverse language models, datasets, tasks, and personalization mechanisms, supporting LIFT as a diagnostic for predictive personalization rather than as a task-specific proxy, style metric, or personalization mechanism.

2 Related Work

Prior work evaluates personalization through task performance, style similarity, model-based judgments, personality alignment, and distributional alignment. Benchmarks such as LaMP, LongLaMP, LaMP-QA, and ALPSBench instantiate personalization as user-conditioned task success [26; 11; 24; 33], while style-personalized generation work relies on overlap metrics, authorship attribution, and style embeddings [2; 32; 22; 9]. Other efforts evaluate personalized generation through model-based automatic judgments, using LLMs to assess content and style alignment against user reference texts [6; 14; 27; 30; 31]. Personalization mechanisms include retrieval, prompt rewriting,

personality prompting, and representation-space steering [17; 13; 2; 16; 37; 10]. Distributional alignment approaches such as SubPOP, DSA, and RADIUS also measure personalization through distribution matching, evaluating how well models reproduce aggregate survey response distributions across demographic subpopulations rather than individual behavioral distributions [28; 8; 12]. We provide a fuller discussion in Appendix A.

Our work differs in target and level of measurement. In *target*, rather than introducing a mechanism, benchmark, style metric, or judge, we evaluate predictive distributions directly. In *level*, we operate at the individual rather than the task, style, or subpopulation level, asking whether a user-specific adaptation shifts a model's predictive distribution closer to a specific individual's behavioral distribution. Our contribution is not the use of likelihood alone, but the formalization and validation of a likelihood-ratio diagnostic for predictive personalization that complements existing metrics.

3 Likelihood Improvement from Targeted Familiarity (LIFT)

3.1 Problem Statement

Let $\mathcal{U} = \{u_i\}_{i=1}^{N}$ be a set of N users. Each user u_i is associated with a behavioral history $\mathcal{H}_i = \{(q_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$, where $q_j^{(i)}$ denotes an input query, $y_j^{(i)}$ denotes the corresponding user-generated response, and n_i is the number of queries. Let $\mathcal{M}_0$ denote a base non-personalized language model. We use $\mathcal{M}_i$ to denote the predictive system obtained by adapting this base model toward user u_i through any mechanism, including retrieval, in-context examples, steering, or fine-tuning. Thus, $\mathcal{M}_i$ may differ from $\mathcal{M}_0$ either through its parameters, its prompt or context, or both.

Our objective is to determine, for a fixed task, tokenizer, base model, and user u_i, which of two adaptations $\mathcal{M}_i^1$ and $\mathcal{M}_i^2$ shifts the predictive distribution closer to u_i's behavioral distribution. We treat the adaptation mechanism as external to the metric. LIFT scores the predictive distribution induced by an adaptation, rather than evaluating how that adaptation was obtained.

3.2 Definition

Let $\mathcal{D}_u$ denote the task-conditioned behavioral distribution of user u over input-response pairs (q, y), where $y = (y_1, \ldots, y_{T(y)})$ is text produced by u for task input q; Let $P_{\mathcal{M}_0}$ and $P_{\mathcal{M}_u}$ denote the predictive distributions of the generic baseline and the user-adapted model respectively, and let $C_u(q)$ denote the user-specific conditioning supplied to the adapted model for input q (e.g., retrieved history examples in retrieval-based systems, or empty in fine-tuned systems where user-specific signal is encoded in the parameters).

Definition (Predictive Personalization)**.** *The predictive personalization rate of adaptation $\mathcal{M}_u$ for user u is the expected per-token reduction in surprisal of user-produced text when moving from the generic baseline to the adapted predictive distribution.*

$$\bar{\Delta}_{\mathrm{KL}}(\mathcal{M}_u, u) = \mathbb{E}_{(q,y)\sim\mathcal{D}_u}\left[\frac{1}{T(y)}\sum_{t=1}^{T(y)}\log\frac{P_{\mathcal{M}_u}(y_t \mid y_{<t}, q, C_u(q))}{P_{\mathcal{M}_0}(y_t \mid y_{<t}, q)}\right]. \tag{1}$$

A positive value indicates that the adaptation assigns higher average token probability to held-out text from the target user than the generic baseline.

3.3 Justification

Tractability of the KL formulation. For a fixed task input q, let $P_u(\cdot \mid q)$ denote the user's conditional response distribution. The difference of KL divergences expands as

$$\begin{aligned} &D_{\mathrm{KL}}(P_u(\cdot \mid q) \,\|\, P_{\mathcal{M}_0}(\cdot \mid q)) - D_{\mathrm{KL}}(P_u(\cdot \mid q) \,\|\, P_{\mathcal{M}_u}(\cdot \mid q, C_u(q))) \\ &\quad = \mathbb{E}_{y\sim P_u(\cdot|q)}\left[\log\frac{P_{\mathcal{M}_u}(y \mid q, C_u(q))}{P_{\mathcal{M}_0}(y \mid q)}\right], \end{aligned} \tag{2}$$

whose length-normalized form coincides with the per-token log-ratio in Eq. 1. Note that the $\log P_u$ terms cancel exactly, so the expression depends on the user's behavioral distribution only through

sample access, with no need to model P_u pointwise. Symmetric divergences such as Jensen–Shannon do not admit this cancellation. They require evaluating a mixture involving P_u pointwise, which is infeasible without direct access to the user's true distribution.

Personalization as surprisal reduction. Eq. 2 admits a direct interpretation in terms of surprisal. Since $-\log P(y)$ is the surprisal of y under model P, the log-ratio can be rewritten as

$$\log \frac{P_{\mathcal{M}_u}(y \mid q, C_u(q))}{P_{\mathcal{M}_0}(y \mid q)} = \underbrace{-\log P_{\mathcal{M}_0}(y \mid q)}_{\text{baseline surprisal}} - \underbrace{[-\log P_{\mathcal{M}_u}(y \mid q, C_u(q))]}_{\text{adapted surprisal}}. \tag{3}$$

Predictive personalization is thus the expected surprisal reduction when moving from the baseline to the adapted model, directly formalizing personalization as person-specific anticipation.

3.4 LIFT as a Per-Token Log-Likelihood-Ratio Diagnostic

Single-reference estimate. The population target in Eq. 1 is an expectation over the user's task-conditioned behavioral distribution. In practice, user histories are sparse, and evaluation often provides only one held-out response for a given task input. We therefore estimate the target from a held-out input-response pair (q, y) produced by the target user.

The autoregressive structure of language models makes Eq. 2 tractable for a single reference. The sequence log-ratio decomposes exactly into token-level log-ratios, which are directly readable from model output under teacher forcing. For a response $y = (y_1, \ldots, y_T)$, we define

$$\mathrm{LIFT}(u, q, y, \mathcal{M}_u) = \frac{1}{T}\sum_{t=1}^{T} \delta_t, \qquad \delta_t = \log \frac{P_{\mathcal{M}_u}(y_t \mid y_{<t}, q, C_u(q))}{P_{\mathcal{M}_0}(y_t \mid y_{<t}, q)}, \tag{4}$$

where each δ_t is the token-level log-likelihood ratio at position t. Positive LIFT means that, on average, the user-conditioned adaptation assigns higher probability to the held-out user text than the generic baseline.

Population target. Averaged over held-out samples from $\mathcal{D}_u$, LIFT estimates the predictive personalization rate in Eq. 1.

$$\mathbb{E}_{(q,y)\sim\mathcal{D}_u}\left[\mathrm{LIFT}(u, q, y, \mathcal{M}_u)\right] = \bar{\Delta}_{\mathrm{KL}}(\mathcal{M}_u, u). \tag{5}$$

The token normalization removes the mechanical dependence of the raw sequence log-ratio $\sum_t \delta_t$ on reference length. It does not imply that empirical scores are invariant to length. Response length may still correlate with genre, retrieval coverage, or the density of user-specific signal. Thus, LIFT should be interpreted as a length-normalized predictive-alignment rate, with cardinal comparisons most meaningful within a fixed task, base model, and tokenizer.

Properties and scope conditions. Two properties follow directly from Eq. 4. First, LIFT measures what an adaptation *demonstrates* on held-out user text, not how that result is achieved. Second, LIFT is mechanism- and task-agnostic: $P_{\mathcal{M}_u}$ may be induced by retrieval, in-context examples, fine-tuning, or any other means, and the score depends *only* on the resulting predictive distribution. Together, they make LIFT useful for comparing adaptations that share a baseline, diagnosing which users or tasks receive stronger predictive alignment, and tracing personalization dynamics over time. This comparative nature is also a scope condition: LIFT scores are meaningful only within a fixed base model and tokenizer and should not be interpreted across different baselines. Whether LIFT tracks user identity rather than generic context, fluency, or topical relevance is an empirical question addressed in Section 4.2. Additional interpretation, scope conditions, and metric properties are discussed in Appendix B. Limitations are discussed in Section J.

3.5 Rare-Token Robustness Analysis with ESS-LIFT

The diagnostic in Eq. 4 can exhibit high finite-sample sensitivity on a single reference text. Rare names, technical terms, unusual formatting, or idiosyncratic punctuation can have very low baseline probability and produce large δ_t values. Such tokens may sometimes carry genuine user-specific

signal, but they can also dominate the score in ways that obscure more stable predictive-alignment patterns. We therefore introduce ESS-LIFT as a rare-token robustness analysis that uses adaptive self-normalized weights to reduce the influence of high-leverage, low-baseline-probability tokens.

For a temperature $\tau > 0$, define token weights $\phi_t^{(\tau)} = P_{\mathcal{M}_0}(y_t \mid y_{<t}, q)^{1/\tau}$. Choosing τ a priori is unreliable, as the heaviness of the per-token tail varies across models, users, and texts. We instead select τ adaptively by controlling weight concentration through the *effective sample size* (ESS):

$$\mathrm{ESS}(\tau) = \frac{\left(\sum_t \phi_t^{(\tau)}\right)^2}{\sum_t \left(\phi_t^{(\tau)}\right)^2} \in [1, T]. \tag{6}$$

We select $\tau_\star$ such that $\mathrm{ESS}(\tau_\star)/T = \rho$ for a target ratio $\rho \in (0, 1]$, and define

$$\text{ESS-LIFT}_\rho(u, q, y, \mathcal{M}_u) = \frac{\sum_{t=1}^T \phi_t^{(\tau_\star)}\, \delta_t}{\sum_{t=1}^T \phi_t^{(\tau_\star)}}. \tag{7}$$

This controls aggregate weight concentration and reduces the practical influence of rare low-baseline-probability tokens, although it does not impose a hard per-token cap on their effect. As $\rho \to 1$, the weights approach uniform weighting and ESS-LIFT recovers LIFT. Importantly, ESS-LIFT is not an unbiased estimator of $\bar{\Delta}_{\mathrm{KL}}(\mathcal{M}_u, u)$ for $\rho < 1$, because the $\phi_t^{(\tau)}$ weighting changes the estimand rather than performing an importance-sampling correction. Sensitivity to ρ, the search procedure for $\tau_\star$, and the bias–robustness trade-off discussion are reported in Appendix C.

4 Experiments

Evaluating a personalization metric is inherently challenging, as there are no ground-truth labels that isolate and remove the confounding factors embedded within proxy tasks. We therefore evaluate LIFT by checking whether three independent properties hold jointly. If LIFT measures user-specific predictive alignment, then it should be (i) *sensitive to user identity* beyond what contextual relevance alone explains; (ii) *robust to surface confounds* such as fluency and generic predictability; and (iii) *coherent across personalization mechanisms*, ranking conditions by the informative value of the user-specific signal they carry rather than by their surface form. No single test provides an external oracle, but evidence from all three angles jointly supports the interpretation.

4.1 Setup

Datasets and tasks. Our main experiments use seven personalized generation tasks drawn from three benchmarks: Amazon review corpora from the books, CDs & vinyl, and movies & TV domains [7], LaMP [26] (personalized news headline generation and personalized tweet paraphrasing), and LongLaMP [11] (personalized abstract generation and personalized Reddit post generation). We use dialogues from Journal Intensive Conversations (JIC) [20] for the cross-mechanism coherence analysis, and Multi-Session Chat [34] to demonstrate further applications of LIFT in Appendix I. Detailed dataset statistics and task definitions are in Appendix D.

Models and personalization mechanism. We evaluate LIFT across six open-source instruction-tuned models spanning three families: Qwen3 (4B and 8B) [35], Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct (abbreviated as Llama-8B and Llama-3B in tables) [5], and Gemma 3 (4B and 12B) [4]. The primary personalization mechanism is retrieval-augmented in-context learning: for each user u, we retrieve a set of history samples via BM25 [23] and prepend them as demonstrations in the input prompt [26]. For the cross-mechanism analysis, we additionally evaluate a fine-tuned dialogue model detailed in Section 4.2.3.

Protocol. We retrieve k profile entries as few-shot demonstrations ($k = 11$ for non-LaMP datasets; $k = 15$ for LaMP datasets). Specific constructions are as follows.

Identity sensitivity. We compare a *Same-User* condition against a *Different-User* condition in which the target profile is replaced by a randomly sampled user's profile. Within each condition we

construct three retrieval variants: the k most similar entries by BM25 (*High-Relevance*), the k least similar (*Low-Relevance*), and k randomly selected entries (*Random*). For the sharpened contrast we additionally compare a *Same-User Low-Relevance* condition against a *Wrong-User High-Relevance* condition. The relevance is reported as the embedding similarity between retrieved examples and the test input via `voyage-4-nano` [29].

Surface confounds. We construct polished and disfluent rewrites that preserve the original content, opinion, and approximate length while altering surface expression, generated via `gpt-5-mini` [18]; prompts are in Appendix E. For the discriminative comparison, we construct authorship verification pairs in which each user serves symmetrically as target and foil. For Amazon review datasets, pairs share the same item and rating. For LaMP and LongLaMP, pairs are constructed via `voyage-4-nano` profile embeddings to yield similar- and dissimilar-profile contrasts.

Cross-mechanism coherence. We compare four context conditions — immediate preceding utterance only (baseline), full historical dialogue, static author-profile entry, and their combination — across two models: Llama-3-8B-Instruct and the JIC-fine-tuned `LLaMAdelic` variant [20], with Llama-3-8B-Instruct under the baseline condition serving as the reference point.

Metrics and reporting. We report LIFT and ESS-LIFT with $\rho = 0.8$ unless otherwise noted. For discriminative comparisons, LLM-as-a-judge uses `gpt-5.4-mini` [19], semantic similarity with BERTScore [36], style-based metrics including StyleEmbedding [32] and StyleDistance [22], n-gram overlap metrics, negative adapted-model surprisal, LIFT, and ESS-LIFT where applicable. All likelihood-based metrics are computed by teacher forcing on the held-out reference text without decoding. Full results with statistical significance and length-sensitivity diagnostics are in Appendices F and H. All experiments are run on 4 NVIDIA RTX A6000 GPUs with 48GB memory each.

4.2 Evidence for Predictive Personalization

4.2.1 Identity Sensitivity

If LIFT measures user-specific predictive alignment, it should respond to whether the correct user profile is present; and it should continue to favor the correct identity even when a mismatched profile is contextually relevant. We probe this at two levels of sharpness.

Coarse probe: same-user vs. different-user. Figure 2(a) reports the positive rate, defined as the proportion of paired samples for which the *Same-User* condition receives a higher score than the corresponding *Different-User* condition across three relevance levels. The positive rate ranges from 0.61 to 0.90 for LIFT and from 0.64 to 0.91 for ESS-LIFT, meaningfully exceeding 0.5 in every setting. This consistency indicates that LIFT effectively responds to identity signals in the matching profile. The agreement between LIFT and ESS-LIFT supports that the identity conclusion is robust to rare-token downweighting.

Sharpened probe: identity vs. relevance. The coarse probe is a necessary but not sufficient condition, as a positive result could still reflect generic contextual helpfulness rather than person-specific signal. We sharpen the test by placing relevance and identity in direct opposition. Table 1 reports the embedding similarity between retrieved examples and the reference text. For Amazon reviews (Books, Movie, CD), Δ_{rel} is near zero or negative, meaning the *Wrong-User High-Relevance* condition is at least as contextually similar to the target as the *Same-User Low-Relevance* condition — the intended adversarial contrast is fully realized. For LaMP and LongLaMP, Δ_{rel} is positive, reflecting that topic choice is more tightly coupled with user identity in these domains; here the contrast tests a different but equally relevant claim: whether LIFT tracks identity in realistic settings where what a user writes about is itself a user-specific behavior.

The SU-LR–WU-HR column in Figure 2(b) shows that LIFT favors the correct identity for a majority of paired samples in every dataset. For LIFT, the positive rate ranges from 0.52 to 0.84. The rates remain especially high on the Amazon review tasks (0.78–0.84), even though the relevance gap is near zero or favors the wrong-user condition. The contrast is weaker on the LaMP and LongLaMP tasks, yet it continues to favor the correct identity more often than not, supporting the interpretation that identity sensitivity holds even when topic and identity partially covary. The similar positive rates under ESS-LIFT indicate that this identity conclusion is robust to rare-token downweighting.

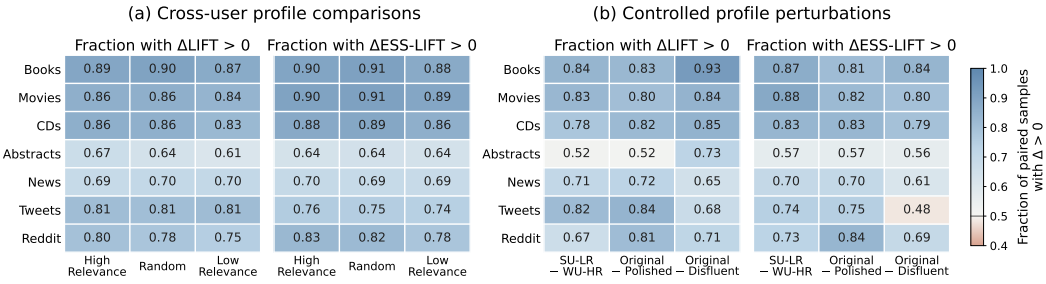


Figure 2: Model-averaged sample-level positive rates for LIFT and ESS-LIFT ($\rho = 0.8$). Each cell reports $\Pr(\Delta S > 0)$. In panel (a), $\Delta S = S_{\text{same-user},c} - S_{\text{different-user},c}$ for retrieval variant $c \in \{\text{High-Relevance}, \text{Random}, \text{Low-Relevance}\}$. In panel (b), ΔS is $S_{\text{SU-LR}} - S_{\text{WU-HR}}$, $S_{\text{original}} - S_{\text{polished}}$, or $S_{\text{original}} - S_{\text{disfluent}}$, as labeled. Values above 0.5 mean that the first condition scores higher for a majority of paired samples. Appendix F.1 reports the per-model rates.

Taken together, the two probes establish that **LIFT responds to user-specific signals beyond generic contextual relevance**. Appendix F.1 and F.2 report the per-model rates and mean contrasts with significance, respectively.

4.2.2 Robustness to Surface Confounds

If LIFT tracks person-specific signal rather than surface confounds, two independent properties should hold. First, targeted surface manipulations that preserve user-specific content, such as polishing or introducing disfluencies, should not systematically inflate or deflate the score; asymmetric sensitivity in either direction would implicate fluency as a driver. Second, LIFT's decisions in a discriminative authorship task should not be reducible to those of surface-based metrics; if they were, LIFT would add no information beyond what those metrics already capture. We test each property in turn.

Table 1: Embedding similarity between demonstrations and reference text under different profile and retrieval conditions (`voyage-4-nano`). HR/LR: high-/low-relevance retrieval. $\Delta_{\text{rel}} = sim(\text{SU-LR}) - sim(\text{WU-HR})$.

	Same-User (SU)		Wrong-User (WU)		
Dataset	**HR**	**LR**	**HR**	**LR**	Δ_{rel}
Books	0.6610	0.6367	0.6521	0.6250	−0.0154
Movie	0.6225	0.5874	0.5947	0.5667	−0.0073
CD	0.7104	0.6784	0.6631	0.6463	0.0153
Tweet	0.6608	0.6502	0.6140	0.6084	0.0362
Abstract	0.7120	0.6223	0.5421	0.5133	0.0802
News	0.5557	0.5519	0.4697	0.4693	0.0822
Reddit	0.5775	0.5541	0.4558	0.4578	0.0983

Fluency intervention. The *Original–Polished* and *Original–Disfluent* columns in Figure 2(b) compare the observed user text with rewrites in opposite fluency directions. Under LIFT, the positive rate for *Original–Polished* ranges from 0.52 to 0.84. The similar range under ESS-LIFT (0.57–0.84) confirms that this comparison is only weakly sensitive to rare tokens. For *Original–Disfluent*, LIFT prefers the original text in 0.65–0.93 of paired samples. The effect of rare-token weighting remains limited on the Amazon datasets but becomes larger in the other domains. In particular, the 0.48 ESS-LIFT rate indicates that the LIFT result for the *Original–Disfluent* comparison is relatively sensitive to rare tokens.

Across both intervention directions, the majority preference for the observed user text indicates that the metric responds to user-specific signals beyond generic fluency. Appendix F.1 reports the exact rates, while F.3 provides per-model effect sizes and significance tests.

Discriminative comparison against surface baselines. The authorship verification task asks whether a metric or a judge correctly identifies which of two candidate texts was written by the target user. Table 2 reports the accuracy across all metrics, with a random-guessing baseline of 0.5. The likelihood-based and n-gram results use Llama-3.1-8B-Instruct as the base model.

Amazon datasets. Table 2 Panel A shows that LIFT exceeds every non-judge baseline across the three Amazon domains. Its advantage over raw adapted-model surprisal demonstrates the contribution of the generic-baseline likelihood ratio. Its advantage over lexical and stylistic metrics, together with the decision-agreement pattern in Figure 3(a), indicates that LIFT provides complementary authorship evidence. Specifically, LIFT shows moderate agreement with `gpt-5.4-mini` and lower agreement

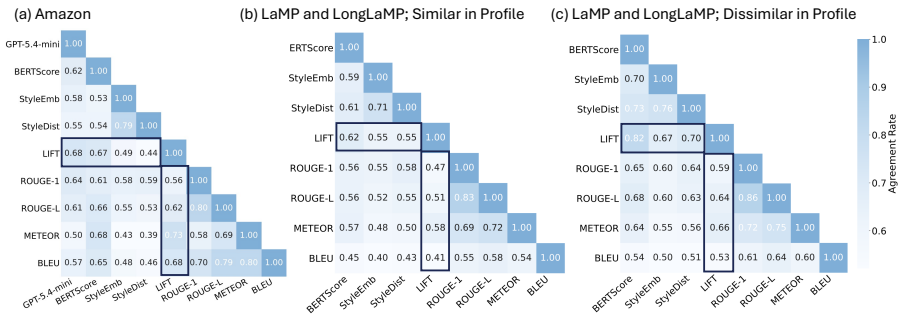


Figure 3: Pairwise decision agreement (0–1) between LIFT and existing metrics on the authorship-verification task. Across panels, LIFT exhibits only moderate agreement with the comparison metrics, indicating that its authorship decisions draw on complementary predictive-alignment information. ESS-LIFT ($\rho = 0.8$) shows a similar pattern in Appendix F.5.

Table 2: Authorship-verification accuracy for Amazon datasets (Panel A), profile-text metrics (Panel B), and n-gram similarity and likelihood-based metrics (Panel C) for LaMP/LongLaMP datasets. All likelihood-based and n-gram results use Llama-3.1-8B-Instruct as the base model. I+R denotes same-item, same-rating pairing; Full results in Appendix F.4.

Panel A: Controlled Amazon authorship verification

Dataset	**Pairing**	**LLM**	**BERTScore**	**StyleEmb**	**StyleDist**	**ROUGE-L**	$-\log P_{\mathcal{M}}$	**LIFT**	*ESS-LIFT*
Books	I+R	**0.9109**	0.6552	0.6063	0.5431	0.7241	0.6552	0.8822	*0.9282*
Movie	I+R	**0.8580**	0.7080	0.6200	0.5760	0.6240	0.5940	0.7660	*0.8340*
CD	I+R	**0.8420**	0.6940	0.5840	0.5660	0.5780	0.5880	0.7400	*0.8220*

Panel B: Profile-based metrics on LaMP/LongLaMP pairs

Dataset	**Similar**			**Dissimilar**		
	BERTScore	**StyleEmb**	**StyleDist**	**BERTScore**	**StyleEmb**	**StyleDist**
Abstract	0.7670	0.5340	0.5370	0.9659	0.6463	0.7084
Reddit	0.7370	0.7870	0.7320	0.9379	0.8407	0.8086
News	0.6351	0.5052	0.5385	0.8656	0.4875	0.5458
Tweet	0.6948	0.6887	0.6990	0.8915	0.8853	0.8988

Panel C: Other metrics on LaMP/LongLaMP pairs

Dataset	**Pairing**	**ROUGE-1**	**ROUGE-L**	**METEOR**	**BLEU**	$-\log P_{\mathcal{M}}$	**LIFT**	*ESS-LIFT*
Abstract	Similar	0.5520	0.5620	0.5370	0.5630	0.5660	0.6980	*0.6910*
Reddit	Similar	0.5490	0.5610	0.5430	0.5840	0.5780	0.7520	*0.8010*
News	Similar	0.5177	0.5166	0.5125	0.0291	0.5728	0.5988	*0.6071*
Tweet	Similar	0.5598	0.5629	0.5856	0.5361	0.6443	0.7072	*0.6680*
Abstract	Dissimilar	0.5872	0.6212	0.5591	0.6473	0.5731	0.8106	*0.7635*
Reddit	Dissimilar	0.6774	0.6814	0.6643	0.6774	0.6152	0.8848	*0.9329*
News	Dissimilar	0.6625	0.6760	0.6448	0.1125	0.7760	0.7896	*0.8063*
Tweet	Dissimilar	0.7066	0.7407	0.6952	0.7386	0.7872	0.9205	*0.8347*

with surface-overlap and style-only metrics, suggesting that it shares part of the signal captured by the LLM-as-a-judge evaluator while contributing distinct evidence from lexical and stylistic similarity.

Downweighting rare tokens through ESS-LIFT increases authorship-verification accuracy across all three domains, suggesting that rare tokens may obscure part of the identity-related signal measured by LIFT. Because authorship-verification accuracy does not directly measure the extent of predictive personalization, the ESS-LIFT results should be viewed as robustness evidence about whether this measurement depends on rare-token influence.

LaMP and LongLaMP datasets. Similar-profile pairs involve users with more closely aligned profiles and therefore present a harder discrimination setting. Table 2 Panel B shows that profile-based metrics exhibit a substantially larger Similar-to-Dissimilar accuracy gap than LIFT, indicating that part of their discriminative power reflects the pairing structure. Because profile similarity is also used to construct the pairs, these metrics have direct access to the signal that defines pairing difficulty. Panel C shows that LIFT consistently outperforms the n-gram metrics under both pair constructions without

relying on this structural advantage. The same advantage is preserved under ESS-LIFT, indicating that the result persists when rare, high-leverage tokens are downweighted. Figure 3(b–c) further shows moderate agreement between LIFT and the profile-based metrics, indicating that predictive alignment and profile-text similarity provide related but distinct evidence for authorship decisions.

Together, **the two probes show that LIFT captures user-specific predictive alignment beyond fluency, roughness, raw adaptation likelihood, and the surface-based signals examined here.**

4.2.3 Ranking Coherence Across Mechanisms

If LIFT captures user-specific predictive alignment, its scores should reflect the predictive value of the user-related information available under each condition. We test this expectation on JIC using a static author profile, dialogue history, JIC fine-tuning, and their combinations. Dialogue history provides direct, task-local observations of user behavior, whereas a static profile summarizes this behavior through a limited set of attributes. We therefore expect history to exceed the profile and FT to provide the strongest individual signal, consistent with its demonstrated personality alignment [20]. Combining complementary sources should further strengthen alignment, following evidence that richer contextual information can improve model adaptation [1].

Table 3: Cross-mechanism ranking of user-specific predictive alignment on 4,000 JIC dialogue turns. Rank 1 denotes the strongest alignment. ESS-LIFT ranks provide a rare-token robustness analysis. FT denotes the JIC-fine-tuned `LLaMAdelic` model. The baseline is Llama-3-8B-Instruct conditioned on the immediately preceding utterance.

Condition	LIFT mean	LIFT rank	ESS-LIFT rank
FT + history + profile	1.4277	1	1
FT + history	1.2714	2	2
FT + profile	1.1973	3	4
FT	1.0062	4	3
History + profile	0.9926	5	6
History	0.7333	6	5
Profile	0.4573	7	7
Baseline	0	–	–

Table 3 shows that the LIFT ranking follows these predictions. The fine-tuned condition exceeds history, history exceeds the static profile, and adding information within the context-only and fine-tuned branches increases LIFT, placing the condition with all three sources first. This ordering across qualitatively different mechanisms indicates that LIFT responds systematically to the strength of user-specific signal and supports its use as a mechanism-agnostic diagnostic for predictive personalization.

The ESS-LIFT ranks provide the rare-token robustness analysis. Their ordering is strongly aligned with the LIFT ordering (Spearman's $r_s = 0.93$), preserving the two highest-ranked conditions and the lowest-ranked condition. The two local reversals occur when the static profile is added to history or fine-tuning, indicating that the incremental contribution of static profile information is more sensitive to rare-token leverage. Appendix G reports the corresponding incremental contrasts and confidence intervals. Beyond the static comparison, LIFT can also track predictive alignment as interaction history accumulates; an illustrative round-wise analysis is provided in the same appendix.

5 Conclusion

This work addresses the lack of a direct diagnostic for the predictive component of personalization in language models. We define predictive personalization as person-specific anticipation, operationalize it as user-specific predictive alignment, and introduce LIFT, a per-token log-likelihood-ratio diagnostic that measures whether user-specific conditioning reduces surprisal on held-out target-user text relative to a generic baseline. Evidence from three convergent angles (sensitivity to user identity beyond contextual relevance, robustness to surface confounds, and coherence across personalization mechanisms) jointly supports LIFT as a principled diagnostic for predictive personalization.

Beyond static evaluation, LIFT frames predictive alignment as a dynamic, measurable property of a model–user–task configuration rather than as a benchmark outcome. By tracking LIFT over interaction history, the diagnostic can characterize when predictive alignment emerges, strengthens, or degrades, enabling analyses of how quickly alignment develops, how much it degrades, and which users or domains remain difficult to anticipate. We hope LIFT serves as a shared diagnostic for measuring and comparing user-specific predictive alignment across personalized systems.

AI USE STATEMENT

`gpt-5-mini` was used to create the controlled polished and disfluent rewrites used in the surface-confound intervention described in Section 4, and `gpt-5.4-mini` was used as an LLM-as-a-judge comparison metric. The generated artifacts were reviewed before use, and the authors take responsibility for the final content and claims of this work.

ETHICS STATEMENT

This work uses existing benchmark datasets containing user-authored text and user-specific interaction histories. It does not collect new human-subject data or attempt to recover users' real-world identities. Nevertheless, user-specific predictive-alignment diagnostics and authorship-verification evaluations could be repurposed for profiling or re-identification, and the underlying text may reflect sensitive attributes or social biases. Results are reported in aggregate, and these diagnostics and evaluations should be used only with appropriate authorization, privacy safeguards, and respect for dataset licenses and terms. They should not be treated as evidence of identity in high-stakes settings.

REFERENCES

[1] Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. In *Advances in Neural Information Processing Systems*, 2024.

[2] Bashar Alhafni, Vivek Kulkarni, Dhruv Kumar, and Vipul Raheja. Personalized text generation with fine-grained linguistic control. *arXiv preprint arXiv:2402.04914*, 2024.

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[4] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

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

[6] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[7] Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, 2026.

[8] Ji Huang, Mengfei Li, and Shuai Shao. Distribution shift alignment helps llms simulate survey response distributions. *arXiv preprint arXiv:2510.21977*, 2025.

[9] Anubhav Jangra, Bahareh Sarrafzadeh, Adrian de Wynter, Silviu Cucerzan, and Sujay Kumar Jauhar. Evaluating style-personalized text generation: Challenges and directions. *arXiv preprint arXiv:2508.06374*, 2025.

[10] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, 2023.

[11] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. LongLaMP: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

[12] Weronika Łajewska, Paul Missault, George Davidson, and Saab Mansour. RADIUS: Ranking, distribution, and significance – a comprehensive alignment suite for survey simulation. *arXiv preprint arXiv:2603.19002*, 2026.

[13] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference*, 2024.

[14] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[16] Jiahong Liu, Wenhao Yu, Quanyu Dai, Zhongyang Li, Jieming Zhu, Menglin Yang, Tat-Seng Chua, and Irwin King. Perfit: Exploring personalization shifts in representation space of llms. In *International Conference on Learning Representations*, 2026.

[17] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. PEARL: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual*, pp. 198–219, 2024.

[18] OpenAI. Introducing GPT-5 for developers. `https://openai.com/index/introducing-gpt-5-for-developers/`, 2025.

[19] OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026.

[20] Sayantan Pal, Souvik Das, and Rohini K. Srihari. Beyond discrete personas: Personality modeling through journal intensive conversations. In *Proceedings of the International Conference on Computational Linguistics*, 2025.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[22] Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. In *The Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

[23] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

[24] Alireza Salemi and Hamed Zamani. Lamp-qa: A benchmark for personalized long-form question answering. In *The Conference on Empirical Methods in Natural Language Processing*, 2025.

[25] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

[26] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *The Association for Computational Linguistics*, 2024.

[27] Alireza Salemi, Julian Killingback, and Hamed Zamani. ExPerT: Effective and explainable evaluation of personalized long-form text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[28] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

[29] Voyage AI. voyage-4-nano: Open-weight text embedding model. `https://huggingface.co/voyageai/voyage-4-nano`, January 2026.

[30] Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

[31] Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*, 2023.

[32] Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, 2022.

[33] Jianfei Xiao, Xiang Yu, Chengbing Wang, Wuqiang Zheng, Xinyu Lin, Kaining Liu, Hongxun Ding, Yang Zhang, Wenjie Wang, Fuli Feng, et al. Alpsbench: An llm personalization benchmark for real-dialogue memorization and preference alignment. *arXiv preprint arXiv:2603.26680*, 2026.

[34] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *The Association for Computational Linguistics*, 2022.

[35] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

[37] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *International Conference on Learning Representations*, 2025.

APPENDICES

CONTENTS

A Detailed Related Work 13
B Properties of LIFT 14
C ESS-LIFT Details 15
D Dataset and Task Details 16
E Prompt Templates 16
F Full Results 19
F.1 Sample-Level Positive Rates 19
F.2 User Identity Degradation 19
F.3 Controlled Perturbations 20
F.4 Metric Comparison 21
F.5 Agreement Rate for Stress Testing 24
G JIC Cross-Mechanism Confidence Intervals and Extended Analysis 24
H Length Sensitivity of LIFT 25
I Applications of LIFT 26
J Discussion, Scope, and Limitations 26

A DETAILED RELATED WORK

Evaluating personalization. Personalized text generation is commonly evaluated through task quality, style similarity, or model-based judgments. Benchmark-driven work typically instantiates personalization as performance on user-conditioned tasks and reports standard generation metrics such as ROUGE, BLEU, METEOR, BERTScore, accuracy, or rating error [21; 15; 3; 36; 26; 11; 24; 33]. A second line focuses on style-personalized generation, where the goal is to write in a target author's voice; evaluation often uses lexical overlap, authorship attribution, style embeddings, style distance, and LLM-as-a-judge protocols [32; 22; 17; 2; 9]. Recent evaluation work argues that no single surface metric is sufficient for style personalization and recommends combining diverse evaluators across domain discrimination, authorship attribution, and personalized-versus-non-personalized discrimination [9]. LLM-based evaluators further broaden the evaluation space by asking judges to compare personalized outputs along content, style, or user-alignment dimensions [31; 30; 27; 6; 14]. These approaches provide useful task-level evidence, but they still evaluate personalization through externally chosen proxies rather than directly measuring whether adaptation shifts a model toward the target user's behavioral distribution.

Methods and benchmarks for personalization. Prior work has developed both personalization mechanisms and benchmarks for testing them. LaMP, LongLaMP, LaMP-QA, and ALPSBench provide user-conditioned tasks spanning recommendation, classification, generation, long-form writing, question answering, memorization, and preference alignment [26; 11; 24; 33]. Methodologically, personalization has been induced through retrieval-augmented prompting and optimization [26; 25; 17], prompt rewriting [13], fine-grained linguistic control [2], many-shot in-context learning [1], fine-tuning, and representation-space steering. Personality-focused work evaluates and induces broad personality traits using psychometric instruments such as Big Five inventories [10], and aligns LLM decisions to individual or group-level personality profiles using personality-inventory data and activation interventions [37]. PerFit studies personalization inside hidden representations, finding both collective and user-specific shifts and using these shifts to steer models with low parameter overhead [16]. These efforts show that personalization can be operationalized through many mechanisms, but most evaluations remain tied to the benchmark task or the intended control dimension.

Distribution alignment in personalization. A growing body of work treats personalization and simulation as distribution alignment rather than only output matching. In public-opinion modeling, SubPOP fine-tunes language models on subpopulation-response pairs so predicted survey response distributions better match human subpopulation distributions [28]. Distribution Shift Alignment (DSA) similarly targets survey simulation, aligning both response distributions and shifts between demographic or background-conditioned distributions [8], while RADIUS emphasizes statistically grounded ranking and distribution alignment for survey simulation [12]. These frameworks show the

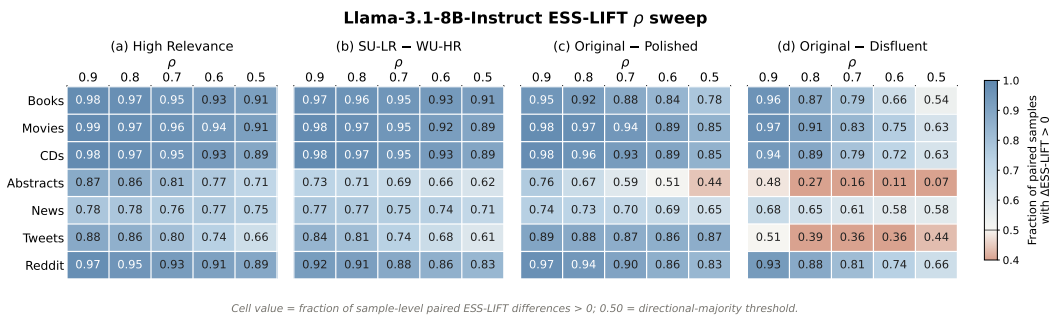


Figure 4: Sensitivity of ESS-LIFT to ρ using Llama-3.1-8B-Instruct. Each cell reports the fraction of paired samples for which the indicated ESS-LIFT contrast is positive; 0.5 denotes the directional-majority threshold. Panel (a) compares Same-User and Different-User profiles under High-Relevance retrieval, panel (b) compares Same-User Low-Relevance with Wrong-User High-Relevance, and panels (c–d) compare Original with Polished and Disfluent text, respectively. Higher values mean that the first condition receives a higher score more often.

value of distributional evaluation, but operate over discrete survey-option distributions and predefined subpopulations.

Our distinction. Our work is complementary to prior benchmarks, methods, and evaluators. Instead of proposing another personalization mechanism, task benchmark, style metric, or LLM-as-a-judge evaluator, we evaluate whether a user-conditioned model shifts its full autoregressive distribution over the vocabulary toward held-out text from a specific individual ($n = 1$). This bridges distribution-level survey alignment and representation-level personalization shifts such as PerFit [16] at the text-prediction level. Because LIFT compares a generic baseline and a user-conditioned model on held-out user text, it does not require deciding in advance whether the relevant signal is topic, preference, personality, style, or interaction history. It is therefore a distributional diagnostic of predictive personalization, intended to complement downstream quality, safety, and user-satisfaction evaluations.

B Properties of LIFT

LIFT measures the predictive component of personalization: whether user-specific conditioning shifts a model's distribution toward text actually produced by the target user. This appendix clarifies the intended interpretation and scope of the metric.

Mechanism-agnostic. LIFT operates solely on the output distributions of $\mathcal{M}_u$ and $\mathcal{M}_0$, placing no constraints on how $P_{\mathcal{M}_u}$ is derived. It applies uniformly to fine-tuned models, retrieval-augmented systems, and prompt-based approaches. Section 4.2 provides evidence for this property: LIFT produces broadly compatible rankings across prompt-based and fine-tuned mechanisms, with the fine-tuned `LLaMAdelic` model and dialogue-history prompting both improving predictive alignment on a common scale.

Task-agnostic. LIFT requires only user-produced text $y \sim P_u$ and places no constraints on task format or output type. It applies without modification across the diverse domains used for convergent-validity evidence — Amazon reviews, news headlines, tweet authorship, academic abstracts, Reddit posts, and dialogue — suggesting broad coverage in practice.

Grounded in user identity. LIFT measures how well $\mathcal{M}_u$ anticipates an individual's text relative to a generic baseline, without constraining which factors drive this anticipation. Its behavior may reflect whatever patterns are most predictive of the individual in context — topic preference, lexical style, syntactic habit, or other behavioral cues. Because LIFT is always evaluated on text actually produced by u and measured against a generic baseline, a score gain provides evidence of a distributional shift toward that individual, operationalizing person-specific anticipation under our definition.

Comparative interpretation. LIFT should be interpreted comparatively. For a fixed task, tokenizer, and base model, higher LIFT indicates that user-specific conditioning reduces surprisal on the target user's text relative to the generic baseline; positive values indicate improvement and negative values indicate degraded predictive alignment. Absolute scores should be compared cautiously across model families, as likelihood scales and tokenization differ; paired contrasts and within-model comparisons are the preferred use cases. We detail the comparative interpretation in Appendix I.

Token-normalized and length-aware. LIFT averages token-level log-likelihood ratios rather than summing them, avoiding mechanical reward for longer references. ESS-LIFT further controls the concentration of token weights to suppress high-leverage rare tokens. As discussed in Appendix H, token normalization does not imply empirical invariance to length: reference length can still be associated with genre, retrieval coverage, and the distribution of user-specific signal.

C ESS-LIFT DETAILS

Solving for $\tau_\star$. For numerical stability, let $s_t = -\log P_{\mathcal{M}_0}(y_t \mid y_{<t}, q)$ denote the generic-baseline surprisal. For a candidate temperature τ, we compute stable weights

$$\phi_t^{(\tau)} = \exp\left(-\frac{s_t - \min_j s_j}{\tau}\right).$$

Subtracting $\min_j s_j$ only rescales all weights by a constant and therefore does not change ESS, since ESS is invariant to uniform scaling of the weights. We then compute

$$\mathrm{ESS}(\tau) = \frac{\left(\sum_t \phi_t^{(\tau)}\right)^2}{\sum_t \left(\phi_t^{(\tau)}\right)^2}.$$

Since $\mathrm{ESS}(\tau)$ increases monotonically toward T as τ grows, we find $\tau_\star$ by binary search for the value satisfying $\mathrm{ESS}(\tau_\star) = \rho T$ over the clipped interval $[\tau_{\min}, \tau_{\max}]$. If the target ESS lies outside the attainable range, we use the nearest endpoint. In all experiments, we use $\rho = 0.8$, $\tau_{\min} = 0.05$, $\tau_{\max} = 20.0$, and 60 binary-search iterations.

Interpretation of ρ. The target ratio ρ controls the finite-sample bias–robustness trade-off. Larger ρ keeps weights closer to uniform and approaches LIFT; smaller ρ more aggressively suppresses rare high-leverage tokens. We use $\rho = 0.8$ in all main experiments without tuning.

Bias-robustness trade-off. ESS-LIFT is not an unbiased estimator of $\bar{\Delta}_{\mathrm{KL}}(\mathcal{M}_u, u)$ for $\rho < 1$, because the $\phi_t^{(\tau)}$ weighting changes the estimand rather than performing an importance-sampling correction. This bias is intentional. ESS-LIFT trades unbiasedness for robustness to high-leverage token-level events on short or heterogeneous references. Since the weighting can in principle affect ordinal rankings, we report both LIFT and ESS-LIFT and verify in Section 4.2 that empirical patterns of interest are preserved.

Sensitivity to ρ. Figure 4 reports sample-level positive rates for Llama-3.1-8B-Instruct across the identity and surface-form contrasts. Each cell is the fraction of paired samples for which the indicated ESS-LIFT contrast is positive; 0.5 is the directional-majority threshold. The identity-degradation and identity–relevance contrasts retain a majority preference for the correct user context in every dataset and at every tested ρ. Rates generally decline as ρ decreases and rare-token downweighting becomes more aggressive, yet these identity conclusions remain stable. Original–Polished remains above the majority threshold except for Abstract at $\rho = 0.5$, whereas Original–Disfluent is estimator-sensitive for Abstract and Tweet, falling below 0.5 across most or all tested ρ values. Thus, the identity findings are robust to the choice of ρ, while the disfluency comparison is the most sensitive to rare-token downweighting, consistent with rare tokens carrying part of the surface-form signal.

Table 4 extends the sweep to the authorship-verification stress test used to assess robustness to surface confounds. On Amazon, accuracy remains high across all ρ values, and $\rho = 0.8$ is close to the best setting on each dataset. On LaMP and LongLaMP, accuracy is also consistently above chance, with

Table 4: Sensitivity of authorship-verification accuracy to ρ using Llama-3.1-8B-Instruct. Values are percentages. I+R denotes same-item and same-rating Amazon pairs; Similar and Dissimilar denote profile-similarity pair constructions. Profile sizes are aggregated by benchmark: Amazon and LongLaMP use 11 reviews, and LaMP uses 15 profile entries.

Pairing	**Dataset**	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.7$	$\rho = 0.6$	$\rho = 0.5$
I+R	Books	92.24	92.82	90.52	89.08	87.93
I+R	Movie	81.80	83.40	83.80	82.80	79.80
I+R	CD	81.20	82.20	83.20	80.20	79.80
Similar	News	60.71	60.71	60.08	60.19	58.94
Similar	Tweet	70.21	66.80	67.01	64.85	59.07
Similar	Abstract	70.20	69.10	67.40	65.30	60.90
Similar	Reddit	79.50	80.10	79.00	76.50	75.10
Dissimilar	News	80.10	80.62	79.79	78.12	76.15
Dissimilar	Tweet	88.22	83.47	79.75	76.34	75.00
Dissimilar	Abstract	80.16	76.35	73.45	68.74	65.03
Dissimilar	Reddit	93.19	93.29	92.18	91.78	88.68

Table 5: Dataset statistics and task definitions. JIC is used for the dialogue analysis in Figure 10; the other datasets are used for the main convergent-validity and metric-comparison experiments.

Dataset	**Task**	**Domain**	**#Users / Dialogues**
Amazon	Review generation	Books	317 users
		Movie & TV	500 users
		CD & Vinyl	500 users
LaMP	Summarization	News Headline	481 users
	Rephrasing	Tweet	485 users
LongLaMP	Long-form generation	Paper Abstract	500 users
		Reddit Post	500 users
JIC	Dialogue continuation	Interpersonal dialogue	500 dialogues

the expected gap between Similar and Dissimilar profile-pair constructions: Dissimilar pairs are easier to resolve, while Similar pairs remain more challenging but still generally preserve strong discrimination. The main conclusion is that the authorship-verification results do not depend on a narrow choice of ρ.

D Dataset and Task Details

Each dataset contains user-specific input-output pairs, where the input is a user profile consisting of prior interaction history and the output is text produced by that user in the target task. For Amazon Books, we use the complete test set; for Amazon Movie and CD, we randomly sample 500 users unless otherwise noted; for LaMP datasets, we keep users with at least 15 historical interactions in the test set; for LongLaMP datasets, we randomly sample 500 users from the test set.

Table 5 summarizes the datasets used for the main convergent-validity evidence and the JIC dialogue analysis in Figure 10. The Amazon, LaMP, and LongLaMP tasks evaluate retrieval-based in-context personalization across diverse personalized generation settings. JIC is used separately to test whether LIFT also applies beyond in-context learning, by comparing prompt-based context with a fine-tuned dialogue model over successive conversation rounds.

E Prompt Templates

Retrieval-based personalization prompt. Below is an illustrative Amazon review prompt showing the generic baseline prompt and the corresponding retrieval-based adaptation prompt with retrieved user-history examples.

General Baseline Prompt:

System:
You are a book reviewer.

User:
Write a book review that matches the given rating and title.

[Item Title]: The Night Circus
[Item Description]: A fantasy novel about a mysterious circus that appears without warning.
[Output Review Rating]: 5.0
[Output Review Title]: Magical and beautifully written

Format your response as:
[Review Text]: ⟨review⟩

Retrieval-based Adaptation Prompt:

System:
You are a specific user who writes book reviews. Maintain consistency with your established voice and perspective across all tasks.

User:
Here are your past reviews:

[Review 1]
- [Item Title]: Piranesi
- [Item Description]: A dreamlike novel set in an endless house of halls and statues.
- [Review Rating]: 5.0
- [Review Title]: Strange, quiet, and unforgettable
- [Review Text]: I loved the atmosphere more than the plot. The writing is spare but somehow full of wonder.

[Review 2]
- [Item Title]: The Goldfinch
- [Item Description]: A literary novel following a boy after a museum tragedy.
- [Review Rating]: 4.0
- [Review Title]: Beautiful but a little long
- [Review Text]: The prose is gorgeous, though I felt the middle dragged. Still, the emotional payoff worked for me.

Now write a review for this book in your characteristic style:

[Item Title]: The Night Circus
[Item Description]: A fantasy novel about a mysterious circus that appears without warning.
[Output Review Rating]: 5.0
[Output Review Title]: Magical and beautifully written

Format your response as:
[Review Text]: ⟨review⟩

Polished rewrite prompt.

You are given a text written by a user. Your task is to rewrite it so that the same opinions and content are expressed in a way that conceals the author's identity – the rewritten text should be indistinguishable from text written by any generic writer.

Task context:

- Output type: review text
- This text is the body of a user review, not a title, headline, chat reply, or explanation.
- Keep it as review prose that a reviewer would actually write.

Requirements:

- Preserve the original opinion, stance, and content faithfully
- Preserve the approximate length of the original text; do not add or remove sentences

- Preserve the original task form and communicative function: the rewritten output must still be a valid review text
- Erase all signals of the author's personal identity: remove idiosyncratic phrasing, personal writing quirks, informal expressions, unique stylistic patterns, and any other features that could identify who wrote the original
- Ensure the result is coherent, fluent, and well-structured
- Do not add new information or change the expressed opinion
- Return only the rewritten review text; do not add labels, explanations, or any extra text

Original text: I wanted to love this book, and some parts were genuinely beautiful, but the pacing was uneven. The characters felt real, though the ending came too quickly for me.

Polished version:

Disfluent rewrite prompt.

You are given a text written by a user. Your task is to rewrite it so that the same opinions and content are expressed in a way that conceals the author's identity – the rewritten text should be indistinguishable from text written by any generic writer.

Task context:

- Output type: review text
- This text is the body of a user review, not a title, headline, chat reply, or explanation.
- Keep it as review prose that a reviewer would actually write.

Requirements:

- Preserve the original opinion, stance, and content faithfully
- Preserve the approximate length of the original text; do not add or remove sentences
- Preserve the original task form and communicative function: the rewritten output must still be a valid review text
- Erase all signals of the author's personal identity: remove idiosyncratic phrasing, personal writing quirks, informal expressions, unique stylistic patterns, and any other features that could identify who wrote the original
- Make the writing feel unpolished through surface-level roughness: awkward wording, weaker flow, mild grammatical irregularities, or clunky phrasing
- For short-form outputs such as headlines, titles, and tweets, keep the rewrite short and recognizably in the same format; use mild awkwardness, not spoken filler or dialogue-like phrasing
- Do not turn the text into a conversation, explanation, self-talk, or meta-commentary
- Do not add new information or change the expressed opinion
- Return only the rewritten review text; do not add labels, explanations, or any extra text

Original text: I wanted to love this book, and some parts were genuinely beautiful, but the pacing was uneven. The characters felt real, though the ending came too quickly for me.

Unpolished version:

LLM-as-a-judge prompt.

Instruction:

You are a writing-style attribution judge. Given past reviews from one user and two candidate reviews, decide which candidate is more likely written by the same user.

Rules:

1. Use only writing style, voice, tone, structure, and evaluative patterns.
2. Ignore topic, product/item identity, factual content, and keyword overlap.
3. If uncertain, choose the closer stylistic match.
4. Output exactly one character: A or B.
5. DO NOT output any explanation.

Past reviews by the user:

[Review 1]

- [Item Title]: Piranesi

- [Item Description]: A dreamlike novel set in an endless house.

- [Review Rating]: 5.0

- [Review Title]: Quiet and strange in the best way

- [Review Text]: This book is odd, but I mean that as praise. The writing is spare and beautiful, and the atmosphere stayed with me.

[Review 2]
- [Item Title]: The Goldfinch
- [Item Description]: A literary novel about loss, art, and obsession.
- [Review Rating]: 4.0
- [Review Title]: Beautiful, but too long
- [Review Text]: I loved the prose and the emotional weight, though the middle dragged for me. Still worth reading.

New review candidates:

[Review A]:
[Review Title]: Magical and beautifully written
[Review Text]: I liked this a lot. The atmosphere was the strongest part, and even when the story moved slowly, the style kept me reading.

[Review B]:
[Review Title]: Fantastic book
[Review Text]: Great story with amazing characters. I enjoyed every chapter and would recommend it to anyone who likes fantasy.

Final answer (A or B only):

F Full Results

F.1 Sample-Level Positive Rates

Figures 5 and 6 give the model-wise values underlying Figure 2(a) and (b), respectively. Each cell reports $\Pr(\Delta S > 0)$ within one dataset and model, for $S \in \{LIFT, ESS - LIFT\ (\rho = 0.8)\}$. Values above 0.5 indicate that the first condition in the contrast scores higher for a majority of paired samples. For the cross-user comparisons in Figure 5, $\Delta S = S_{\text{same-user},c} - S_{\text{different-user},c}$ for $c \in \{\text{High-Relevance}, \text{Random}, \text{Low-Relevance}\}$. For the controlled perturbations in Figure 6, the three columns use $S_{\text{SU-LR}} - S_{\text{WU-HR}}$, $S_{\text{original}} - S_{\text{polished}}$, and $S_{\text{original}} - S_{\text{disfluent}}$.

Across model families, Figure 5 preserves the pattern summarized in Figure 2(a): Books, Movie, and CD show the strongest and most stable majority preference for the correct-user profile, whereas Abstract and News are weaker and more variable, with Tweet and Reddit generally intermediate. Figure 6 likewise shows broad cross-model support for the sharper identity–relevance and surface-form contrasts in Figure 2(b), while also exposing greater variation for Abstracts and for some Tweet interventions. Thus, the model-averaged results are not attributable to a single model family, although their strength varies by domain and estimator.

F.2 User Identity Degradation

Figure 7 complements the positive rates in Figure 2(a) with the corresponding per-model mean contrasts. Whereas Figure 2(a) shows how often the Same-User condition scores higher, Figure 7 shows the average size of that advantage for each model, dataset, and retrieval condition. The contrasts are positive in nearly all cells, with the smallest and occasional nonsignificant values concentrated in Abstract. Thus, the model-averaged direction in Figure 2(a) is also visible at the level of individual model families and effect magnitudes.

Notably, the absolute contrasts are larger in News and Tweet than in the other longer-form domains such as Books and Movie. However, these values should not be treated as a cardinal ranking across datasets. A plausible contributor is termed personalization dilution, which corresponds to the length-sensitivity analysis shown in Appendix H. The average reference lengths for News and Tweet are 9.14 and 18.88 words, respectively, whereas the other domains average 162.63–306.64 words. Longer references can contain more generic or profile-uncovered content, diluting a user-specific gain when it is averaged per token. Length cannot mechanically create the identity contrast, however, because each pair of conditions is scored on the same held-out reference. Abstract is also shorter than the review and Reddit domains yet show the weakest effects. We therefore interpret the magnitude pattern as reflecting the joint influence of reference length and the density of user-specific signal under genre constraints, rather than as a causal effect of length alone.

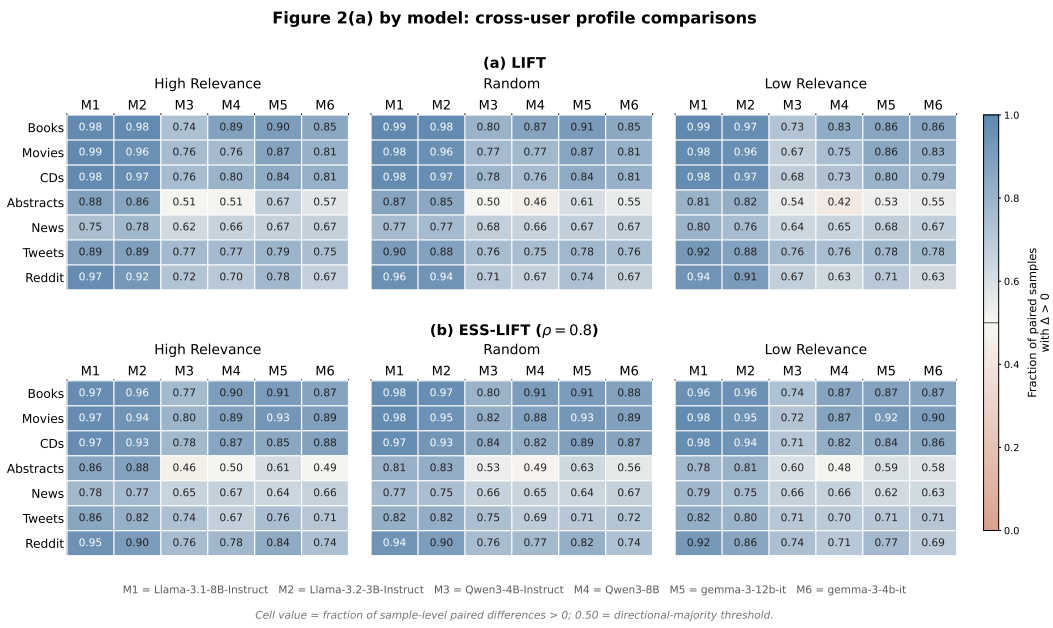


Figure 5: Per-model positive rates underlying Figure 2(a). Panels show LIFT and ESS-LIFT ($\rho = 0.8$); columns separate High-Relevance, Random, and Low-Relevance retrieval. Each cell reports $\Pr(S_{\text{same-user},c} - S_{\text{different-user},c} > 0)$ for one dataset, model, and retrieval variant c.

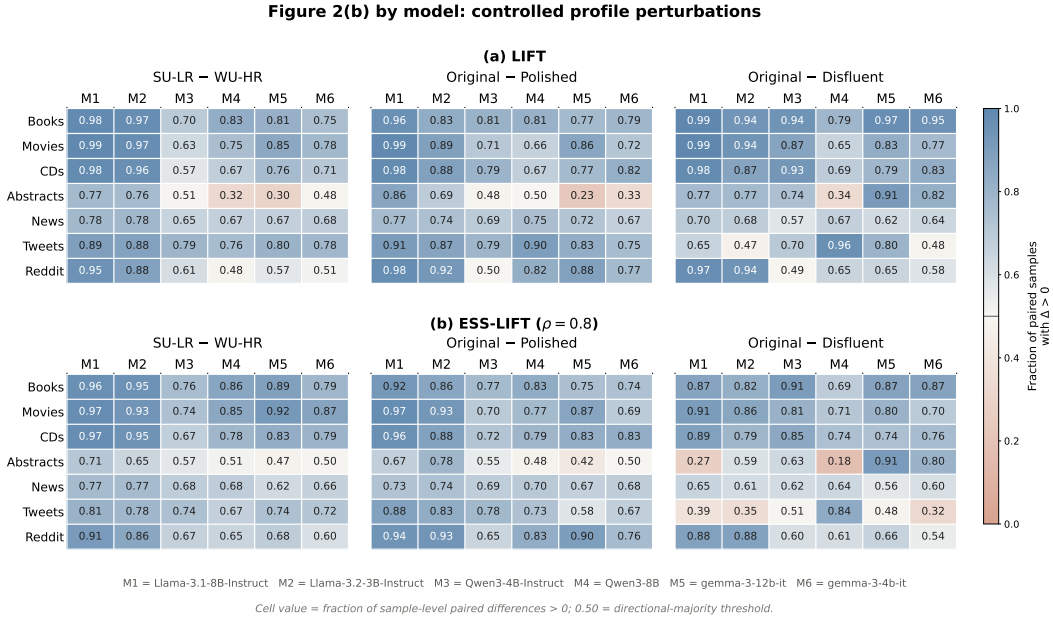


Figure 6: Per-model positive rates underlying Figure 2(b). Panels show LIFT and ESS-LIFT ($\rho = 0.8$); columns report $S_{\text{SU-LR}} - S_{\text{WU-HR}}$, $S_{\text{original}} - S_{\text{polished}}$, and $S_{\text{original}} - S_{\text{disfluent}}$. Each cell is the fraction of paired samples for which the labeled contrast is positive, within one dataset and model.

F.3 Controlled Perturbations

Figure 8 provides the magnitude view corresponding to the positive rates in Figure 2(b). It reports per-model mean differences for the adversarial identity–relevance comparison and the two surface-form interventions. Most cells remain positive, while the smaller and occasional negative contrasts

Figure 2(a) by model: cross-user profile mean differences

(a) LIFT

	High Relevance						Random						Low Relevance					
	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
Books	+0.243	+0.231	+0.156	+0.222	+0.247	+0.288	+0.230	+0.217	+0.148	+0.198	+0.241	+0.274	+0.206	+0.194	+0.125	+0.163	+0.218	+0.272
Movies	+0.152	+0.119	+0.074	+0.090	+0.154	+0.163	+0.155	+0.121	+0.072	+0.085	+0.164	+0.168	+0.159	+0.126	+0.056	+0.087	+0.156	+0.174
CDs	+0.154	+0.120	+0.084	+0.096	+0.151	+0.172	+0.145	+0.108	+0.078	+0.087	+0.132	+0.163	+0.149	+0.111	+0.055	+0.077	+0.128	+0.151
Abstracts	+0.110	+0.120	+0.025	+0.033	+0.109	+0.070	+0.056	+0.058	+0.014	-0.008 *ns*	+0.052	+0.028	+0.039	+0.043	+0.013	-0.016	+0.023	+0.014 *ns*
News	+0.499	+0.435	+0.482	+0.560	+0.842	+1.113	+0.462	+0.421	+0.550	+0.605	+0.901	+1.054	+0.517	+0.405	+0.478	+0.637	+0.877	+0.963
Tweets	+0.403	+0.371	+0.464	+0.540	+0.840	+1.003	+0.391	+0.355	+0.457	+0.534	+0.827	+1.034	+0.382	+0.345	+0.436	+0.526	+0.787	+0.984
Reddit	+0.185	+0.159	+0.143	+0.159	+0.172	+0.166	+0.142	+0.131	+0.105	+0.114	+0.117	+0.128	+0.107	+0.083	+0.068	+0.066	+0.073	+0.083

(b) ESS-LIFT ($\rho = 0.8$)

	High Relevance						Random						Low Relevance					
	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
Books	+0.134	+0.123	+0.084	+0.101	+0.106	+0.129	+0.130	+0.118	+0.079	+0.092	+0.103	+0.129	+0.117	+0.104	+0.068	+0.076	+0.089	+0.115
Movies	+0.090	+0.067	+0.046	+0.053	+0.080	+0.085	+0.092	+0.071	+0.046	+0.055	+0.086	+0.085	+0.096	+0.070	+0.036	+0.055	+0.081	+0.088
CDs	+0.087	+0.068	+0.047	+0.055	+0.074	+0.087	+0.087	+0.060	+0.046	+0.050	+0.069	+0.084	+0.092	+0.068	+0.037	+0.045	+0.068	+0.080
Abstracts	+0.049	+0.063	+0.002 *ns*	+0.009 *ns*	+0.037	-0.001 *ns*	+0.030	+0.036	+0.006 *ns*	+0.002 *ns*	+0.028	+0.008 *ns*	+0.023	+0.030	+0.009	+0.001 *ns*	+0.014	+0.009 *ns*
News	+0.346	+0.280	+0.259	+0.304	+0.518	+0.648	+0.315	+0.264	+0.325	+0.332	+0.551	+0.559	+0.346	+0.255	+0.278	+0.354	+0.463	+0.580
Tweets	+0.176	+0.174	+0.201	+0.185	+0.352	+0.407	+0.171	+0.168	+0.201	+0.188	+0.319	+0.436	+0.165	+0.160	+0.168	+0.188	+0.340	+0.392
Reddit	+0.114	+0.097	+0.083	+0.086	+0.093	+0.090	+0.088	+0.080	+0.065	+0.064	+0.070	+0.072	+0.069	+0.053	+0.047	+0.043	+0.050	+0.050

M1 = Llama-3.1-8B-Instruct M2 = Llama-3.2-3B-Instruct M3 = Qwen3-4B-Instruct M4 = Qwen3-8B M5 = gemma-3-12b-it M6 = gemma-3-4b-it

Cell value = mean paired difference; ns = p ≥ 0.05. Raw magnitudes are not directly comparable across datasets.

Figure 7: Per-model mean contrasts corresponding to Figure 2(a). For retrieval variant $c \in \{\text{High-Relevance}, \text{Random}, \text{Low-Relevance}\}$, each cell reports the mean of $S_{\text{same-user},c} - S_{\text{different-user},c}$ for LIFT or ESS-LIFT ($\rho = 0.8$). Positive values favor the correct-user profile; "ns" denotes $p \geq 0.05$ in the paired test. Raw magnitudes are not directly comparable across datasets.

reveal model–dataset heterogeneity. A high positive rate can coexist with a modest mean contrast when the advantage is small but broadly distributed, while negative or nonsignificant cells identify model–domain exceptions to the aggregate pattern. The same cross-domain scaling caveat applies here as in Section F.2. The larger absolute contrasts often seen for News and Tweet are consistent with less per-token personalization dilution in their short references, whereas longer references can include more generic or profile-uncovered content. Because each perturbation pair uses the same reference, length cannot itself produce the contrast, and the weak Abstract effects again show that length is not sufficient. We therefore view the pattern as a joint consequence of reference length and user-specific signal density; Appendix H discusses the length evidence in detail. These per-model results complement the model-averaged positive rates in Figure 2(b).

F.4 Metric Comparison

Table 2 provides the compact cross-metric comparison, while Tables 6 and 7 expose the model- and pairing-level variation that is obscured by aggregation. The Amazon evaluation contains 174 item-and-rating-matched pairs for Books and 250 each for Movie and CD. The Similar and Dissimilar splits are nearly balanced for every LaMP/LongLaMP dataset, with 481/480 pairs for News, 485/484 for Tweet, and 500/499 each for Abstract and Reddit. Consequently, differences between the two pairing regimes cannot be attributed to unequal evaluation-set sizes.

Cross-model robustness on Amazon. The Amazon panel in Table 2 reports results for Llama-8B. Table 6 extends the same authorship-verification protocol to Qwen3-4B, Qwen3-8B, Llama-3B, and Llama-8B. Across all 12 model–domain combinations, both LIFT and ESS-LIFT outperform raw adapted-model surprisal and the strongest n-gram baseline. The repeated advantage across two model families and two model sizes provides cross-model evidence that the likelihood ratio captures user-specific predictive signal beyond generic predictability and lexical overlap.

The strength of this signal varies across base models. Relative to raw adapted-model surprisal, the accuracy gain ranges from 0.130 to 0.247 for LIFT and from 0.164 to 0.290 for ESS-LIFT. Within each model family, the larger model yields higher ESS-LIFT accuracy on Movie and CD. Books also

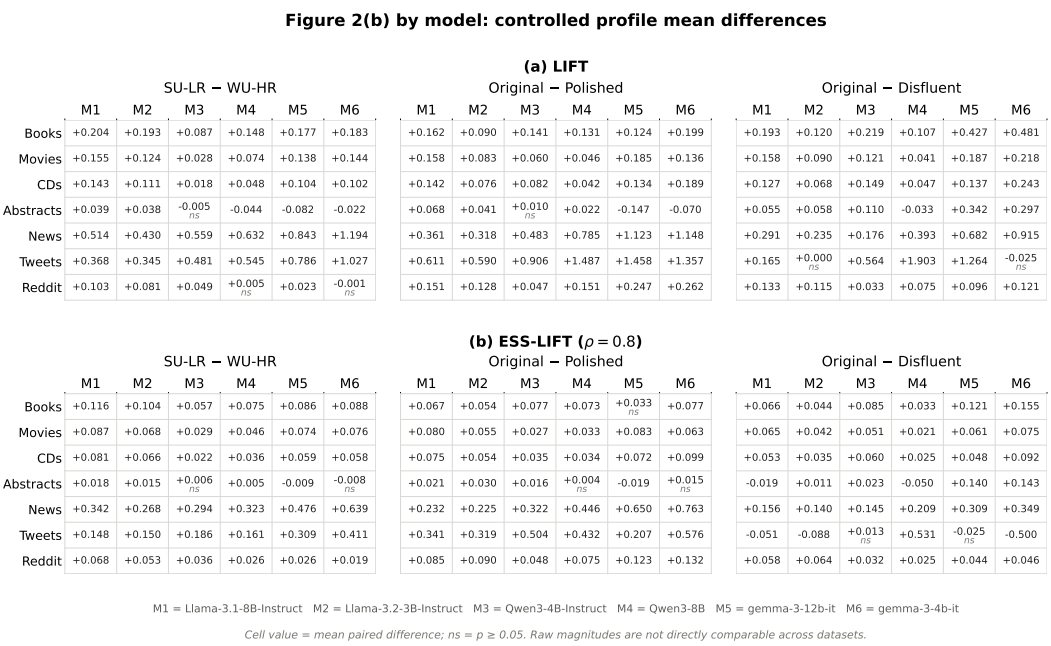


Figure 8: Per-model mean contrasts corresponding to Figure 2(b) for LIFT and ESS-LIFT ($\rho = 0.8$). Columns report $S_{\text{SU-LR}} - S_{\text{WU-HR}}$, $S_{\text{original}} - S_{\text{polished}}$, and $S_{\text{original}} - S_{\text{disfluent}}$. Positive values favor the first condition; "ns" denotes $p \geq 0.05$ in the paired test. Raw magnitudes are not directly comparable across datasets.

improves from Qwen3-4B to Qwen3-8B, while the two Llama sizes yield the same Books accuracy. These differences indicate that the base model shapes how strongly the personalization signal supports authorship discrimination. Table 6 therefore extends the Llama-8B result in Table 2 across additional base models and quantifies the model-dependent variation in the observed advantage.

The comparison between LIFT and ESS-LIFT provides a complementary rare-token robustness analysis. ESS-LIFT yields higher authorship-verification accuracy in 11 of the 12 settings, supporting the hypothesis that rare, high-leverage tokens can obscure identity-relevant evidence in the unweighted estimate. The Qwen3-4B Books reversal shows that the influence of these tokens varies by setting. Authorship-verification accuracy measures candidate discrimination, so the 11-of-12 pattern provides suggestive evidence about rare-token influence on the observed signal. The likelihood-ratio values themselves characterize the extent of predictive personalization. Accordingly, the ESS-LIFT results strengthen the cross-model robustness claim and identify settings in which rare-token weighting materially changes authorship discrimination.

Pairing difficulty and complementary signals on LaMP and LongLaMP. Using the same Llama-8B configuration as Table 2, Table 7 places profile-based, likelihood-based, and overlap-based metrics in the same dataset–pairing cells. Across all eight cells, both LIFT and ESS-LIFT outperform raw adapted-model surprisal and every n-gram metric. This repeated advantage across Similar and Dissimilar profile pairings extends the main-text result and establishes a stable likelihood-ratio contribution across levels of pairing difficulty.

Profile proximity systematically changes discrimination accuracy. Moving from Similar to Dissimilar pairs raises LIFT accuracy by 0.113–0.213 and ESS-LIFT accuracy by 0.073–0.199 across all four datasets. The largest LIFT increase occurs on Tweet, while the largest ESS-LIFT increase occurs on News, showing that the benefit of greater profile separation varies by domain. Together with the within-regime advantage over raw surprisal and n-gram overlap, these increases locate the predictive-alignment signal within both the more challenging Similar comparisons and the easier Dissimilar comparisons.

Table 6: Amazon authorship-verification accuracy by base model. StyleEmb and StyleDist follow Wegmann et al. [32] and Patel et al. [22]; `gpt-5.4-mini` is the LLM-as-a-judge evaluator [19].

Base	Metric	Books	Movie	CD
–	LLM-as-a-judge	0.9109	0.8580	0.8420
	BERTScore	0.6552	0.7080	0.6940
	StyleEmb	0.6063	0.6200	0.5840
	StyleDist	0.5431	0.5760	0.5660
Qwen3-4B	ESS-LIFT ($\rho = 0.8$)	0.8103	0.6980	0.7060
	LIFT	0.8218	0.6640	0.6720
	$-\log P_{\mathcal{M}}$	0.6351	0.5340	0.5420
	ROUGE-1	0.5776	0.5400	0.5280
	ROUGE-L	0.6897	0.5680	0.5420
	METEOR	0.7299	0.6020	0.5620
	BLEU	0.7672	0.5960	0.5760
Qwen3-8B	ESS-LIFT ($\rho = 0.8$)	0.8563	<u>0.7960</u>	0.7420
	LIFT	0.8362	0.7180	0.6880
	$-\log P_{\mathcal{M}}$	0.6322	0.5520	0.5320
	ROUGE-1	0.6063	0.5420	0.5380
	ROUGE-L	0.7213	0.5720	0.5540
	METEOR	0.7184	0.5600	0.5640
	BLEU	0.7586	0.6040	0.5680
Llama-3B	ESS-LIFT ($\rho = 0.8$)	**0.9282**	0.7860	<u>0.7580</u>
	LIFT	<u>0.8851</u>	0.7300	0.7100
	$-\log P_{\mathcal{M}}$	0.6379	0.5540	0.5600
	ROUGE-1	0.6207	0.5880	0.5440
	ROUGE-L	0.6839	0.5900	0.5260
	METEOR	0.6063	0.5260	0.5040
	BLEU	0.6293	0.5660	0.5240
Llama-8B	ESS-LIFT ($\rho = 0.8$)	**0.9282**	**0.8340**	**0.8220**
	LIFT	0.8822	0.7660	0.7400
	$-\log P_{\mathcal{M}}$	0.6552	0.5940	0.5880
	ROUGE-1	0.7040	0.6520	0.5840
	ROUGE-L	0.7241	0.6240	0.5780
	METEOR	0.6609	0.5400	0.5280
	BLEU	0.7098	0.5660	0.5640

Table 7: Full authorship verification accuracy under Similar and Dissimilar pair constructions for LaMP and LongLaMP. Profile similarity is computed by `voyage-4-nano` [29] embedding.

		Similar in Profile				Dissimilar in Profile			
Base	Metric	Abstract	Reddit	News	Tweet	Abstract	Reddit	News	Tweet
–	BERTScore	0.7670	0.7370	0.6351	0.6948	0.9659	0.9379	0.8656	0.8915
	StyleEmb	0.5340	0.7870	0.5052	0.6887	0.6463	0.8407	0.4875	0.8853
	StyleDist	0.5370	0.7320	0.5385	0.6990	0.7084	0.8086	0.5458	0.8988
Llama-3.1-8B-Instruct	ESS-LIFT (ρ=0.8)	<u>0.6910</u>	**0.8010**	**0.6071**	<u>0.6680</u>	<u>0.7635</u>	**0.9329**	**0.8063**	<u>0.8347</u>
	LIFT	**0.6980**	<u>0.7520</u>	<u>0.5988</u>	**0.7072**	**0.8106**	<u>0.8848</u>	<u>0.7896</u>	**0.9205**
	$-\log P_{\mathcal{M}}$	0.5660	0.5780	0.5728	0.6443	0.5731	0.6152	0.7760	0.7872
	ROUGE-1	0.5520	0.5490	0.5177	0.5598	0.5872	0.6774	0.6625	0.7066
	ROUGE-L	0.5620	0.5610	0.5166	0.5629	0.6212	0.6814	0.6760	0.7407
	METEOR	0.5370	0.5430	0.5125	0.5856	0.5591	0.6643	0.6448	0.6952
	BLEU	0.5630	0.5840	0.0291	0.5361	0.6473	0.6774	0.1125	0.7386

The joint comparison also reveals domain-specific complementarity between profile similarity and predictive alignment. BERTScore is strongest on Abstract and News under both pair constructions. ESS-LIFT is strongest on Similar-profile Reddit, and LIFT is strongest on Tweet under both pair constructions. On Dissimilar-profile Reddit, ESS-LIFT reaches 0.9329, closely approaching BERTScore at 0.9379 and exceeding both style metrics. These crossovers show that authorship information is expressed through different signals across domains. The moderate cross-metric decision agreement in Figure 3 supports the same interpretation by showing that profile-text similarity and predictive alignment often reach correct decisions from partially distinct evidence.

Rare-token weighting produces a second domain-structured pattern. ESS-LIFT exceeds LIFT on Reddit and News under both pair constructions, while LIFT leads on Abstract and Tweet under both. The stable ordering across pairing regimes indicates domain-specific sensitivity to rare-token leverage. The gains on Reddit and News are consistent with rare tokens obscuring identity-relevant evidence, while the LIFT advantage on Abstract and Tweet suggests that rare tokens can contribute useful authorship evidence. As in Table 6, these comparisons provide suggestive evidence about rare-token

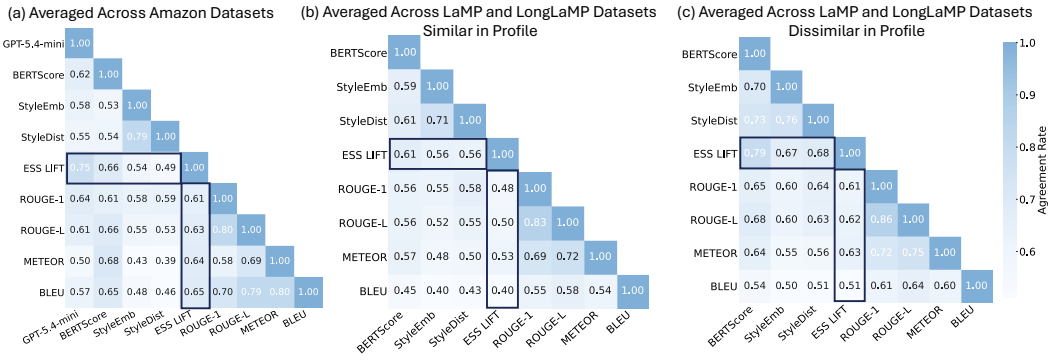


Figure 9: Pairwise decision agreement (0–1) between ESS-LIFT ($\rho = 0.8$) and existing metrics on the authorship-verification stress test. The pattern provides a rare-token robustness analysis for the LIFT agreement results in Figure 3.

Table 8: JIC 95% confidence intervals for the condition means and pre-specified incremental contrasts associated with Table 3, computed over 4,000 dialogue-turn instances. Δ is the condition minus the indicated reference. The LIFT columns report the primary analysis; the ESS-LIFT columns ($\rho = 0.8$) report robustness after downweighting rare, high-leverage tokens. FT denotes the JIC-fine-tuned `LLaMAdelic` model, and Baseline denotes Llama-3-8B-Instruct conditioned only on the immediately preceding utterance.

Condition	Δ reference	Mean LIFT 95% CI	LIFT Δ 95% CI	Mean ESS-LIFT 95% CI	ESS-LIFT Δ 95% CI
Context-only conditions					
History	Baseline	[0.7135, 0.7531]	[0.7135, 0.7531]	[0.1738, 0.1956]	[0.1738, 0.1956]
Profile	Baseline	[0.4314, 0.4832]	[0.4314, 0.4832]	[−0.0848, −0.0528]	[−0.0848, −0.0528]
History + profile	History	[0.9664, 1.0188]	[0.2358, 0.2827]	[0.1458, 0.1781]	[−0.0360, −0.0095]
Fine-tuned conditions					
FT	Baseline	[0.9923, 1.0201]	[0.9923, 1.0201]	[0.2535, 0.2712]	[0.2535, 0.2712]
FT + history	FT	[1.2518, 1.2910]	[0.2533, 0.2771]	[0.3373, 0.3614]	[0.0802, 0.0937]
FT + profile	FT	[1.1713, 1.2233]	[0.1720, 0.2101]	[0.2354, 0.2675]	[−0.0220, 0.0002]
FT + history + profile	FT + history	[1.4012, 1.4542]	[0.1402, 0.1724]	[0.3498, 0.3832]	[0.0078, 0.0266]

influence on the observed signal and identify the domains in which token weighting most strongly changes authorship discrimination.

F.5 Agreement Rate for Stress Testing

Figure 9 provides the ESS-LIFT counterpart to Figure 3. Each cell reports the fraction of paired decisions for which ESS-LIFT selects the same candidate as the comparison metric. The qualitative pattern matches LIFT. ESS-LIFT shows its highest agreement with LLM-based judgments and moderate agreement with surface-overlap and style-only metrics, supporting the same complementary-signal interpretation.

G JIC Cross-Mechanism Confidence Intervals and Extended Analysis

Table 8 complements the primary LIFT mean and rank analysis in Table 3 with 95% confidence intervals for condition means and pre-specified incremental contrasts. The ESS-LIFT columns provide a rare-token robustness analysis of the same contrasts.

Extended analysis. Table 8 first sharpens the primary LIFT result: every pre-specified LIFT contrast is positive with a 95% confidence interval that excludes zero, and FT + history + profile has the largest condition mean. The ESS-LIFT results serve as a robustness analysis of whether these conclusions depend on rare, high-leverage tokens. After those tokens are downweighted, the gains for History over Baseline, FT over Baseline, FT + history over FT, and FT + history + profile

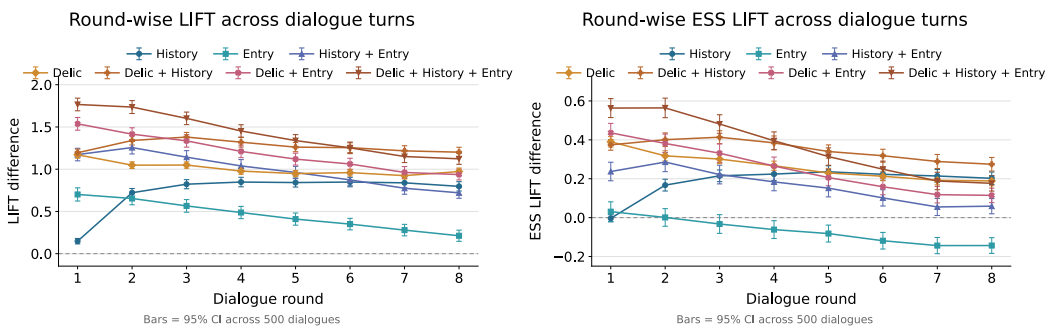


Figure 10: Round-wise JIC analysis across 500 dialogues. Left: LIFT differences, used as the primary estimate of predictive alignment. Right: ESS-LIFT differences ($\rho = 0.8$), used as a robustness analysis that downweights rare, high-leverage tokens. Conditions follow Table 3: *Baseline* is Llama-3-8B-Instruct conditioned only on the immediately preceding interlocutor utterance; *History* adds dialogue history; *Profile* adds the static author profile; and *FT* denotes the JIC-finetuned `LLaMAdelic` model. Combined conditions add the indicated information sources. Bars show 95% confidence intervals.

over FT + history remain positive with confidence intervals excluding zero. By contrast, Profile versus Baseline and History + profile versus History become negative, while the interval for FT + profile versus FT includes zero. The central conclusion that dialogue history and fine-tuning increase predictive alignment is therefore robust to rare-token influence, whereas the incremental contribution of the static profile is more sensitive to it. These attenuated profile contrasts qualify the primary LIFT interpretation rather than replacing it.

Round-wise dynamics. Beyond the aggregate mechanism comparison in Table 3, Figure 10 shows how predictive personalization changes as dialogue unfolds. The LIFT trajectories provide the primary analysis: dialogue history improves predictive alignment after the first round, the static profile alone becomes less useful over time, and the FT conditions remain strongly positive, especially when combined with history and profile information. The ESS-LIFT trajectories are a rare-token robustness analysis rather than a second primary estimate. After rare, high-leverage tokens are downweighted, the broad mechanism-level pattern and the gains associated with history and fine-tuning persist, whereas conditions that rely primarily on the static profile are attenuated. Thus, the history and fine-tuning conclusions are not driven by a small number of rare tokens, while the incremental value of the static profile is more sensitive to their influence.

H Length Sensitivity of LIFT

By definition, LIFT and ESS-LIFT are token-normalized averages, so neither has a mechanical preference for longer reference texts: increasing the number of tokens does not increase the score through additive accumulation. Nevertheless, Table 9 shows negative correlations between reference length and score in a substantial number of model–dataset settings. We refer to this empirical pattern as *personalization dilution*: the weakening of average user-specific likelihood-ratio evidence as reference length increases. Token normalization removes a direct length reward, but it does not make the token composition of shorter and longer texts identical. If the additional tokens in longer references carry, on average, weaker user-specific likelihood-ratio evidence than the more diagnostic tokens, averaging over them dilutes the per-token personalization gain. The observed correlations can therefore reflect systematic changes in token composition with length rather than a length bias in the definition itself.

The ESS-LIFT results provide a rare-token robustness analysis of this pattern. Relative to LIFT, ESS-LIFT attenuates the negative correlations in many settings and moves several estimates toward zero or positive values, indicating that part of the apparent length sensitivity is associated with rare, high-leverage tokens whose influence varies with reference length. Negative correlations that persist under ESS-LIFT indicate that personalization dilution is not explained entirely by rare tokens and may

Table 9: Spearman's r_s between reference length and LIFT or ESS-LIFT ($\rho = 0.8$) across datasets and models. Values marked with * are statistically significant ($p < 0.05$), and italicized values are not significant.

Panel A: LIFT

	Qwen3		Llama		Gemma	
Dataset	**4B**	**8B**	**3.2-3B**	**3.1-8B**	**4B**	**12B**
Books	-0.1133*	-0.1266*	-0.4733*	-0.6001*	-0.4830*	-0.6703*
Movie	-0.2051*	-0.1961*	-0.7085*	-0.7672*	-0.6553*	-0.7341*
CD	-0.2278*	-0.1746*	-0.6405*	-0.7215*	-0.6609*	-0.7016*
News	*-0.0276*	-0.1610*	-0.1921*	-0.2611*	-0.4656*	-0.4758*
Tweet	-0.1656*	-0.2941*	-0.1353*	-0.1298*	*-0.0412*	-0.1159*
Abstract	-0.2200*	-0.1690*	-0.4065*	-0.4323*	-0.3116*	-0.5571*
Reddit	-0.1306*	-0.1616*	-0.6697*	-0.6562*	-0.6515*	-0.5313*

Panel B: ESS-LIFT *($\rho = 0.8$)*

	Qwen3		Llama		Gemma	
Dataset	**4B**	**8B**	**3.2-3B**	**3.1-8B**	**4B**	**12B**
Books	*-0.0001*	*-0.0157*	-0.3005*	-0.4564*	-0.4194*	-0.4023*
Movie	-0.0932*	-0.1058*	-0.5182*	-0.6705*	-0.5917*	-0.4790*
CD	*0.0558*	*-0.0657*	-0.4065*	-0.5747*	-0.5400*	-0.3039*
News	*0.0193*	-0.1374*	-0.1159*	-0.1175*	-0.2186*	-0.2882*
Tweet	*-0.0794*	-0.1894*	*-0.0115*	*0.0370*	*0.0479*	*-0.0017*
Abstract	*0.0813*	0.1106*	-0.1937*	-0.2207*	-0.1219*	-0.2667*
Reddit	*0.0869*	*0.0127*	-0.4586*	-0.3764*	-0.5826*	-0.2564*

also involve the broader token distribution. Because reference length is observationally correlated with content and task difficulty, Table 9 does not by itself identify a unique causal mechanism.

I Applications of LIFT

Previous sections provide evidence that LIFT captures user-specific predictive alignment. We now illustrate how the diagnostic can track predictive alignment dynamically as interaction history accumulates.

Figure 11 compares round-wise LIFT and ESS-LIFT ($\rho = 0.8$) across 212 30-round Multi-Session Chat dialogues using Llama-3.1-8B-Instruct. Both measures show the same qualitative dynamics. Predictive alignment increases as dialogue history accumulates within each session, drops after session transitions, and recovers as new session-specific evidence becomes available. Their absolute magnitudes differ substantially, with ESS-LIFT consistently attenuating the larger LIFT values. This attenuation is informative rather than contradictory. Because ESS-LIFT reweights token-level log-likelihood ratios when $\rho < 1$, it is a robustness-oriented estimand rather than a corrected or rescaled version of LIFT. The lower ESS-LIFT values observed here indicate that rare, low-baseline-probability tokens with large positive log-likelihood ratios contribute materially to the magnitude of LIFT. Nevertheless, ESS-LIFT remains predominantly positive and preserves the repeated rise-and-reset pattern, showing that the dynamic effect is not solely driven by rare tokens. The strongest attenuation occurs immediately after session transitions, where ESS-LIFT approaches or briefly crosses zero. Predictive gains that remain after rare-token downweighting are therefore initially weak after a topic shift and strengthen as new context accumulates.

J Discussion, Scope, and Limitations

The experimental evidence supports LIFT as a diagnostic of user-specific predictive alignment under our definition. We state explicitly here what that evidence does and does not establish.

Predictive versus useful personalization. This work targets the predictive user-model component of personalization by asking whether user-specific conditioning shifts a model toward held-out

behavior from the target individual. This is distinct from useful personalization, which asks whether the resulting action is helpful, accurate, safe, preferred, or effective for that user. We therefore position LIFT as a complementary diagnostic of an intermediate construct, not a complete measure or universal proxy for user-facing utility. Predictive alignment is neither necessary nor sufficient for useful personalization. A system can follow an explicit preference without predicting the user's natural text, and it can anticipate or imitate a user while still producing inaccurate, unhelpful, or unsafe responses. The two should be related when success depends on anticipating the user's preferences, choices, language, or behavior, but the relationship is task dependent. This distinction is diagnostically useful because high task performance with low LIFT can reflect generic competence or instruction following, whereas high LIFT with low task performance suggests that user-specific information was acquired but not used effectively. Our experiments provide construct-validity evidence for this predictive component.

What LIFT measures and does not measure. LIFT measures whether user-specific conditioning shifts a model's predictive distribution toward held-out text from the target individual. It is intentionally agnostic not only to the adaptation mechanism, but also to the underlying source of the signal. The signal may reflect topics, opinions, preferences, style, lexical habits, and other factors, but LIFT neither separates nor identifies them. Two systems with identical LIFT scores may therefore achieve predictive alignment through different mechanisms or mixtures of factors. These relationships can also vary by domain. In Table 1, the semantic-similarity separation between the controlled retrieval conditions is small for the review datasets (Books, Movie, and CD) and larger for the LaMP and LongLaMP tasks (Tweet, Abstract, News, and Reddit), showing that topic or relevance can couple to the observable personalization signal differently across tasks. LIFT diagnoses the net change in predictive alignment, not its causal decomposition. Separately, LIFT does not measure helpfulness, factuality, user satisfaction, privacy, safety, or normative desirability. A system that scores well on LIFT may still fail on these dimensions, and LIFT is intended to complement rather than replace task-quality, safety, and user-facing evaluations.

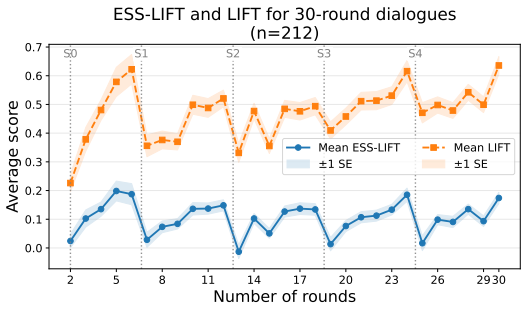


Figure 11: Round-wise LIFT and ESS-LIFT ($\rho = 0.8$) for Llama-3.1-8B-Instruct on 212 30-round Multi-Session Chat dialogues [34]. The adapted condition includes all dialogue history observed so far, whereas the baseline includes only the immediately preceding utterance. Dotted lines mark session starts (S0–S4), and shading shows ± 1 standard error. The shared rise-and-reset pattern provides robustness evidence, while the magnitude gap reflects the contribution of rare, high-leverage tokens to LIFT.

Bounded use and baseline comparability. LIFT is designed for comparative use within a fixed base model and tokenizer. Absolute scores are not comparable across model families because likelihood scales and tokenization differ; for example, the higher absolute values produced by Gemma reflect higher base-model entropy rather than stronger personalization. Cross-family claims should therefore use paired contrasts under a shared baseline. These bounds do not limit the metrics' practical reach. Under a controlled evaluation, LIFT and ESS-LIFT can compare personalization mechanisms, monitor alignment as interaction history accumulates, distinguish failures to acquire user-specific information from failures to use it downstream, and regression-test whether system changes preserve target-user alignment.

Recommended use of LIFT and ESS-LIFT. Users should first define adapted and baseline conditions on the same task, held-out target-user text, base model, and tokenizer, then compute paired scores at the example or user level and report uncertainty. LIFT should be treated as the primary estimate of predictive alignment. ESS-LIFT should be reported alongside it, with the chosen ρ, as a rare-token robustness analysis. Agreement strengthens the conclusion that the result is not driven by rare, high-leverage tokens, whereas attenuation or reversal identifies a conclusion that requires closer inspection rather than an automatic correction. Neither metric should be interpreted alone

as user-facing utility; both should be reported with task-quality, safety, and user-facing evaluations appropriate to the application.

Limitations and future work. Our empirical validation is limited to in-domain personalization settings, where the evidence supplied to a personalized system and the held-out target-user text come from the same domain. The results therefore support LIFT as a diagnostic in this setting, but do not establish how its scores should be interpreted when user evidence comes from domain A and target behavior from domain B. In such cross-domain evaluations, changes in score may reflect both user-specific predictive alignment and domain shift, making additional controls necessary. Existing personalization benchmarks do not provide paired cross-domain histories and held-out behaviors for the same users, limiting the test of this distinction. Developing such benchmarks and validating LIFT, with ESS-LIFT as its rare-token robustness analysis, under controlled cross-domain evaluation are important directions for future work.