Theory-Grounded Evaluation Exposes the Authorship Gap in LLM Personalization

Yash Ganpat Sawant [1]

Abstract

Stylistic personalization—making LLMs write in a specific individual's style, rather than merely adapting to task preferences—lacks evaluation grounded in authorship science. We show that grounding evaluation in authorship verification theory transforms what benchmarks can measure. Drawing on three measurement traditions—LUAR (a trained authorship verification model), an LLM-as-judge with decoupled trait matching, and classical function-word stylometrics—we evaluate four inference-time personalization methods across 50 authors and 1,000 generations. The theory-grounded metric (LUAR) provides what ad hoc alternatives cannot: calibrated baselines (human ceiling 0.756, cross-author floor 0.626) that give scores absolute meaning. All methods score *below* this floor (0.484–0.508), exposing an authorship gap invisible to uncalibrated metrics. The three metrics produce near-zero pairwise correlations ($|r| < 0.07$), confirming that without theoretical grounding, metric choice determines conclusions—an LLM judge declares a clear winner while LUAR finds no meaningful differentiation. These findings demonstrate the theory–benchmark cycle in action: authorship theory exposes evaluation failures that ad hoc benchmarks miss.

1. Introduction

Stylistic personalization—making an LLM write in a specific individual's style, rather than merely adapting to task preferences—has no established evaluation methodology grounded in authorship science. LaMP (Salemi et al., 2024) evaluates personalization through task accuracy. PersonalLLM (Zollo et al., 2025) measures preference alignment. PersonaLens (Zhao et al., 2025) uses LLM-as-judge for conversational personalization. None evaluate whether generated text *sounds like* the target author—whether the model's underlying authorship fingerprint actually shifts toward the target.

This is a gap, not a critique: authorial style fidelity is simply a different construct than task accuracy or preference alignment, and no standard metric for it exists. We propose to fill this gap by grounding evaluation in authorship verification theory—a decades-old discipline with validated methods and calibrated baselines (Stamatatos, 2009; Rivera-Soto et al., 2021). We contrast this theory-grounded approach with two ad hoc alternatives: an LLM-as-judge (a decoupled binary trait protocol) and classical stylometrics (function word distributions (Argamon et al., 2003)).

We test these three metrics on four inference-time stylistic personalization methods across 50 authors and 1,000 generations, and find:

1. **Theory provides calibrated baselines that ad hoc metrics lack.** LUAR authorship verification yields a human-author ceiling (0.756) and cross-author floor (0.626), giving scores absolute meaning. All methods score *below* the human floor (0.484–0.508), exposing an authorship gap invisible without calibration.
2. **Without theoretical grounding, metric choice determines conclusions.** The LLM judge declares profile extraction a clear winner (d=0.58); LUAR finds no meaningful differentiation. The judge's apparent signal traces to circularity between trait extraction and profile extraction (Section 4.3).
3. **Metric disagreement signals construct validity failure.** The three metrics produce near-zero correlations ($|r| < 0.07$)—they measure different constructs, and only the theory-grounded metric passes validation tests.

These findings demonstrate the theory–benchmark cycle: authorship theory generates testable predictions, the benchmark confirms them with calibrated measurements, and the result exposes evaluation failures that ad hoc approaches miss.

[1]Independent AI Researcher. Correspondence to: Yash Ganpat Sawant <sawantyash13@gmail.com>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

2. Related Work

Personalization benchmarks. LaMP (Salemi et al., 2024) evaluates personalization via task accuracy; LongLaMP (Kumar et al., 2024) extends this to long-form generation with content-summary prompts (which we adopt). PersonalLLM (Zollo et al., 2025) uses synthetic preference profiles. PersonaLens (Zhao et al., 2025) evaluates conversational personalization with LLM-as-judge. Critically, no existing benchmark evaluates whether generated text is stylistically faithful to the target author—the gap we address.

Benchmark quality and construct validity. BetterBench (Reuel et al., 2024) proposes 46 criteria for assessing benchmark quality, finding widespread disparities. Raji et al. (2021) argue that benchmarks become proxies for progress without validating whether they measure the intended construct—an instance of the broader construct validity problem (Cronbach & Meehl, 1955). Our work applies this lens specifically to personalization: we propose three metrics from different traditions, show they diverge, and identify which one—LUAR authorship verification—passes validation tests the others fail.

Authorship verification. Computational authorship analysis spans from function-word frequencies (Argamon et al., 2003) to neural methods. LUAR (Rivera-Soto et al., 2021) learns universal authorship representations via contrastive learning. Wang et al. (2025) use authorship analysis to show LLMs struggle to imitate everyday authors—a finding our calibrated baselines quantify precisely.

3. Evaluation Framework

3.1. Data and Methods

We use the Blog Authorship Corpus (Schler et al., 2006): 681K posts from 19,320 bloggers. We select 50 authors with $\geq$200 training posts, $\geq$50 test posts, and mean length $\geq$100 words, yielding 104K training and 26K test posts. Writing prompts are LLM-extracted content summaries (neutral descriptions of *what* a post discusses, not *how*)—we show in Section 4.3 that naïve first-sentence extraction inflates baselines by 28 percentage points.

We evaluate four inference-time methods spanning implicit to explicit style transfer: NON-PERSONALIZED (control; content summary only), FEW-SHOT (5 author samples, no explicit instruction), PROFILE EXTRACTION (two-stage: extract abstract style profile, then generate from profile only), and CONTRASTIVE (author samples + contrastive examples from other authors + stylometric features). All use Qwen 3 32B (Qwen Team, 2025) as generator with 50 authors $\times$ 5 prompts $\times$ 4 methods $=$ 1,000 generations.

Table 1. Method comparison across 50 authors, 1,000 generations. LUAR uses 5-post aggregation (↑ better). TMR = trait match rate. SA% = same-author rate. All methods score *below* the cross-author human floor on LUAR. CIs: hierarchical bootstrap (B=10,000).

Method	LUAR ↑	TMR ↑	SA% ↑	FuncCos ↑
NON-PERSONAL.	0.484±.019	0.384±.058	22%	0.741±.011
FEW-SHOT	**0.508**±.020	0.433±.061	31%	0.749±.011
PROFILE EXTR.	0.502±.019	**0.542**±.060	29%	**0.761**±.010
CONTRASTIVE	0.494±.020	0.447±.059	**36%**	0.752±.011
Real Author (ceil.)	0.756	0.427	30%	0.742
Cross-Author (floor)	0.626	0.390	7%	0.695

3.2. Three Independent Metrics

We deliberately select metrics from three different traditions:

LUAR (Primary). Learning Universal Authorship Representations (Rivera-Soto et al., 2021) is a transformer trained via contrastive learning on millions of Reddit posts to produce author-discriminative embeddings. We compute cosine similarity between 5-post aggregated LUAR embeddings of generated and real text. LUAR is uniquely suited to personalization evaluation because it provides *calibrated* baselines: same-author pairs yield known score distributions distinct from cross-author pairs, enabling absolute rather than relative evaluation.

We validate LUAR on our blog corpus (trained on Reddit): single-post AUC=0.76, multi-post (5) AUC=0.96, vs. TF-IDF baseline AUC=0.54. The metric transfers reliably across domains.

LLM-as-Judge (Secondary). A decoupled binary protocol using GLM-4 32B (GLM Team, 2024) (different model family from generator): (i) extract 5 style traits as yes/no questions per author (cached); (ii) score each generation on traits in one call; (iii) judge same-author plausibility in a *separate* call. Decoupling prevents cross-signal contamination between trait scoring and holistic judgment. Primary metric: `Trait Match Rate` (TMR) $=$ traits_present$/5$.

Automated Stylometrics (Tertiary). Function word cosine similarity (`FuncCos`) over 60 common function words—established markers of individual writing style (Argamon et al., 2003).

4. Results

4.1. The Human–LLM Authorship Gap

Table 1 presents our central measurement. On LUAR, all four methods score between 0.484 and 0.508—a spread of just **0.024**—and all fall *below* the human cross-author floor of 0.626 (ceiling 0.756). The LLM's authorship fingerprint dominates: personalized output is more distant from the target human author than random humans are from each

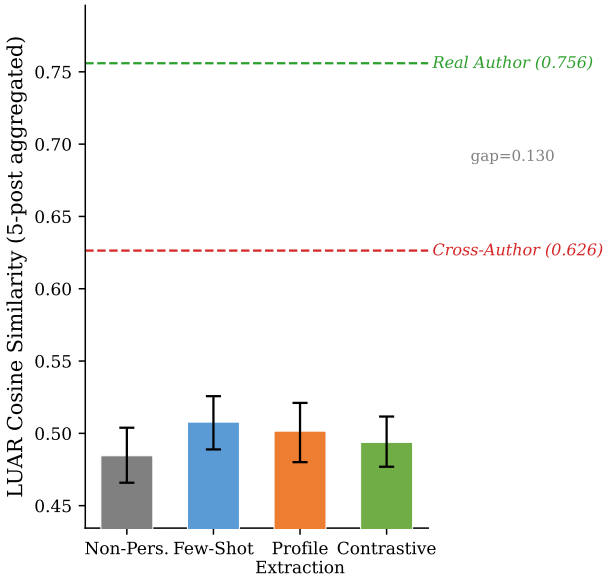

Figure 1. LUAR authorship similarity by method with calibration baselines. All methods score below the cross-author human floor (0.626). The authorship gap between generated and human text is a measurable, calibrated quantity.

Table 2. Pearson correlation between metrics (n=1,000). All correlations near zero: the metrics capture fundamentally different constructs.

	LUAR	TMR	FuncCos
LUAR	1.00	—	—
TMR	0.013	1.00	—
FuncCos	0.026	0.067	1.00

other.

Yet personalization is not vacuous. Within generated text, LUAR discriminates target authors at AUC=0.918 (gen↔gen), confirming that methods produce author-differentiated output. This output simply remains in the LLM's own style space rather than crossing into human authorship territory (gen→real AUC=0.632). Figure 1 visualizes the gap.

4.2. Why LUAR Should Anchor Personalization Evaluation

Table 2 reveals that neither TMR nor FuncCos correlates with LUAR—the only metric validated against known authorship baselines ($|r| < 0.07$; bootstrap 95% CIs: LUAR–TMR $[-0.049, 0.075]$, LUAR–FuncCos $[-0.036, 0.089]$). This means a benchmark using only TMR would declare PROFILE EXTRACTION the clear winner (effect size d=0.58 over baseline), while LUAR—anchored to calibrated authorship verification—finds no meaningful differentiation across methods.

Why trust LUAR over the alternatives? LUAR's baselines

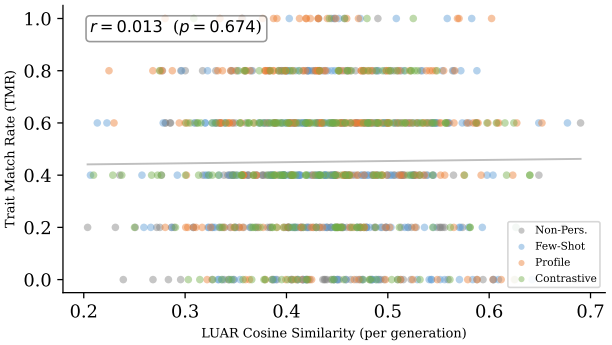

Figure 2. LUAR similarity vs. TMR for 1,000 generations (r=0.013). Profile extraction's apparent advantage on TMR has no corresponding signal on LUAR.

are well-separated (ceiling 0.756, floor 0.626, gap=0.130), confirming it discriminates authors reliably (AUC=0.96, Section 3.2). FuncCos baselines are nearly collapsed (ceiling 0.742, floor 0.695, gap=0.047)—it cannot meaningfully separate same-author from cross-author text, and generated methods (0.741–0.761) straddle the ceiling, suggesting LLMs produce grammatically average function-word distributions regardless of personalization. TMR has no calibrated baselines at all, and the circularity evidence from Section 4.3 shows PROFILE EXTRACTION *exceeds the real author* on TMR (0.542 vs. 0.427)—a method that scores higher than ground truth is measuring instruction-following, not authorship fidelity. Figure 2 visualizes this: TMR scores are uniformly distributed across the LUAR range (r=0.013), confirming the two metrics capture unrelated constructs.

4.3. Circularity: Why Profile Extraction "Wins" on the Judge

The discrepancy has a concrete explanation. PROFILE EXTRACTION achieves TMR=0.542, far above other methods, but LUAR=0.502—indistinguishable. Both the judge's trait extraction (Stage 1) and the method's profile extraction ask an LLM to read author samples and extract salient style features. The profile is then used to generate text optimized for exactly the kind of features the judge checks.

The real author's own text confirms this: TMR=0.427, *lower* than PROFILE EXTRACTION's 0.542. If TMR measured genuine authorship fidelity, the real author would set the ceiling. Instead, the method that explicitly optimizes for LLM-extractable traits exceeds the real author, exposing that TMR measures **instruction-following fidelity**, not **authorship fidelity**.

Additionally, we find trait extraction is unstable: repeated extraction for the same author yields mean Jaccard similarity of 0.22 across trait sets. The yardstick itself changes between measurements.

4.4. Cross-Model Robustness

To test whether the authorship gap is generator-specific, we replicate with GLM-4 32B (10 authors, 150 generations). All GLM-4 methods score below the human floor: LUAR ranges 0.417–0.576 vs. floor 0.626. GLM-4 shows wider method spread (0.16 vs. Qwen's 0.024) but the gap persists across both model families.

Cross-model LUAR analysis reveals three distinct regimes: within-model similarity is high (Qwen↔Qwen 0.918, GLM↔GLM 0.839), cross-model is lower (Qwen↔GLM 0.753), and gen→real is lowest (0.45–0.49). Each LLM carries its own authorship fingerprint that inference-time personalization does not erase—a finding consistent with theoretical predictions from AI-generated text detection (Mitchell et al., 2023).

4.5. Prompt Contamination as Confound

A methodological finding relevant to benchmark design: naïve prompt construction (extracting the first sentence of the target post) inflates the unpersonalized baseline from SA=22% to 50%—a **28 percentage point** confound. The first sentence carries the author's vocabulary, punctuation, and tonal markers, making even unpersonalized output appear author-matched. We adopt LLM-extracted content summaries following Kumar et al. (2024), which specify *what* the post discusses without leaking *how* the author writes. Prompt construction is a hidden confound that can mask or invert method rankings.

5. Discussion

Design principles from the theory–benchmark cycle. Our findings yield two design principles for stylistic personalization evaluation:

(i) Ground metrics in established theory. Authorship verification provides what ad hoc metrics cannot: calibrated baselines with absolute meaning. LUAR's reference points (ceiling: 0.756, floor: 0.626) transform a score of 0.50 from uninterpretable to precise—all methods remain 0.12–0.14 below even the cross-author floor, let alone the same-author ceiling. Repurposing task-accuracy or preference-alignment metrics for stylistic fidelity fails because they measure different constructs entirely.

(ii) Treat multi-metric disagreement as diagnostic. Rather than averaging metrics or picking the most convenient one, disagreement between independently motivated metrics should be treated as evidence of construct validity failure (Cronbach & Meehl, 1955)—a signal that some metrics are measuring artifacts rather than the target construct.

Toward the theory–benchmark cycle. The CTB vision of a virtuous cycle between theory and benchmarks applies directly: authorship verification theory predicts that LLM-generated text carries a model-specific fingerprint detectable by trained classifiers (Mitchell et al., 2023; Kirchenbauer et al., 2023). Our LUAR measurements confirm this prediction quantitatively: the fingerprint is strong enough that all personalized outputs cluster below the human floor. This provides a *falsifiable, calibrated* measure of personalization capability—exactly the kind of formal guarantee the field needs to move beyond ad hoc evaluation. Concretely, our framework instantiates two legs of the cycle: theory (authorship verification) generates a testable prediction (inference-time prompting cannot shift the LLM's fingerprint toward a target author), and the benchmark confirms it with calibrated measurements, identifying the precise gap future methods must close.

The generated-text regime. Our analysis reveals a structural phenomenon: LLM-generated text occupies a distinct region of LUAR embedding space. Gen↔gen similarity averages 0.932 (same target) vs. 0.858 (different targets), both far above gen↔real (0.522). Personalization modulates output *within* this regime (gen↔gen AUC=0.918) but does not escape it. This suggests a formal characterization may be possible: the set of achievable authorship embeddings under inference-time methods may be bounded away from the human manifold, providing a theoretical target for future work on closing the gap.

Scope. This study targets inference-time stylistic personalization (few-shot, profile extraction, contrastive examples) on blog writing from one corpus in English using two model families (Qwen 3, GLM-4) at 32B scale. Training-time approaches such as per-user LoRA adapters or preference optimization may shift the LLM's authorship fingerprint more substantially; this work provides the calibrated yardstick on which they should be judged. Analogous inference-time personalization failures have been observed in behavioral domains (Sawant, 2026), suggesting the authorship gap may extend beyond stylistic fidelity to other personalization constructs.

Does LUAR penalize generated text per se? A natural concern is that LUAR's low gen→real similarity reflects generic "LLM-ness" detection rather than author-style mismatch. Three observations weigh against this reading. First, LUAR achieves AUC=0.918 at distinguishing target authors *within* generated text: if it merely detected machine origin, all generations would cluster identically regardless of target. Second, cross-model LUAR exhibits graded structure (Qwen↔Qwen 0.918, Qwen↔GLM 0.753, gen→real 0.45–0.49), consistent with author-style *mixing with* model-style rather than a binary human/machine cut. Third, GLM-4

produces a method spread of 0.16 in LUAR (vs. Qwen's 0.024), showing LUAR registers method-level differentiation *inside* the generation regime. LUAR thus measures a continuous authorship signal that is partially recoverable under inference-time prompting and partially captured by the model's own fingerprint.

Future work. Three complementary directions follow. (i) Human evaluation of perceived stylistic fidelity, to triangulate which metric best predicts reader-perceived authorship match. (ii) Training-time methods (per-user LoRA, preference optimization) to test whether the gap closes when the generator's fingerprint can be reshaped, rather than only its conditional distribution. (iii) Cross-genre validation (microblog, email, code commentary, scientific writing) to test whether the gap and its calibrated baselines transfer across communicative settings. All evaluation is in English; authorship patterns may differ across languages.

References

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23:321–346, 2003.

Cronbach, L. J. and Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

GLM Team. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *ICML*, 2023.

Kumar, S., Sathe, C., Tiwari, A., and Zamani, H. LongLaMP: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., and Finn, C. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *ICML*, 2023.

Qwen Team. Qwen3 technical report. *arXiv preprint*, 2025.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the everything in the whole wide world benchmark. In *NeurIPS*, 2021.

Reuel, A., Hardy, A., Lamparth, M., Hardy, M., Smith, B., and Kochenderfer, M. J. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *NeurIPS Datasets and Benchmarks*, 2024.

Rivera-Soto, R., Miano, O., Ordonez, J., Chen, B. Y., Khan, A., Bishop, M., and Andrews, N. Learning universal authorship representations. In *EMNLP*, 2021.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. LaMP: When large language models meet personalization. In *ACL*, 2024.

Sawant, Y. G. High-stakes personalization: Rethinking LLM customization for individual investor decision-making. *arXiv preprint arXiv:2604.04300*, 2026.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. Effects of age and gender on blogging. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

Stamatatos, E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Wang, S. et al. Catch me if you can? Not Yet: LLMs still struggle to imitate the implicit writing styles of everyday authors. In *EMNLP Findings*, 2025.

Zhao, Z., Vania, C., Kayal, S., Khan, N., Cohen, S. B., and Yilmaz, E. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In *Findings of ACL*, 2025.

Zollo, T. P., Siah, K. H., Ye, T., Li, H., and Namkoong, H. PersonalLLM: Tailoring LLMs to individual preferences. In *ICLR*, 2025.