FAVoR: Measuring and Mitigating Author-Style Homogenization in Federated Personalized Generation

Anonymous ACL submission

Abstract

Large language models are increasingly used as personalized writing assistants, but adapting a model across many authors can compromise individual writing style by pulling author-specific signals toward a shared register. Federated parameter-efficient fine-tuning (PEFT) offers a data-local setting for this multi-author adaptation problem: clients keep author text local while sharing compact adapter updates. However, we show that standard aggregation can preserve continuation utility while making different authors' generations less distinguishable in style space, a failure mode we define as *author-style homogenization*. We evaluate author-style retention with Angular Style Classification Encoder (ASCE)-based diagnostics on our main BlogText benchmark and ASCE-independent external authorship verification. Using this protocol, we find that common federated PEFT baselines can preserve semantic utility while averaging out author-specific signals. To address this homogenization, we instantiate FAVoR (Federated Authorial Voice Retention), an author-style residual mechanism for federated PEFT. FAVoR uses a shared–private adapter design: clients upload shared-adapter updates while retaining author-specific residual corrections locally. Across BlogText and external Mythos-Reddit validation, FAVoR improves author-style retention over standard and personalized federated PEFT baselines. These gains come with small continuation-utility trade-offs and are supported by component ablations, external verification, and cold-start transfer.

1 Introduction

Personalized writing support often asks a large language model (LLM) to continue unseen prompts while preserving an author's characteristic writing style. In collaborative adaptation, however, improvements in general continuation competence can reduce the stylistic distinctions that personalization is meant to preserve. We call this failure mode **author-style homogenization**: continuations remain fluent and semantically plausible, but lose author-specific stylistic signals.

This problem is especially sharp under author-level non-identically distributed (non-IID) data. Clients differ in lexical choice, syntax, pacing, rhetorical structure, and register, not merely in sample distribution. Naïvely pooling author data, or averaging all parameter-efficient fine-tuning (PEFT) updates in federated learning (FL), can mix author-specific correction directions with unrelated or opposing directions from other authors, shifting generations toward a shared register. Keeping raw text local is therefore insufficient by itself (Lin et al., 2022); standard FL still needs a mechanism for deciding which adaptation directions should be shared and which should remain client-specific.

We study this issue in **personalized held-out continuation**: each client corresponds to one author, and the model must continue held-out prompts in that author's style. This setting makes homogenization directly testable, because collaboration should improve continuation quality without erasing cross-author stylistic distinctions, and evaluation must separate author signal from topical content.

We introduce an **author-style retention evaluation protocol** based on an Author-Style Signature Space. For BlogText, we instantiate this space with an author-disjoint Angular Style Classification Encoder (ASCE). To test whether this space captures the author signal rather than the topic alone, we use the topic as a coarse content proxy and the author labels as an operational proxy for author style. Figure 1 shows that the learned signature space increases author-within-topic separability and reduces the topic–author origin-fit slope from 0.72 to 0.47 relative to raw computational-linguistic features.

To mitigate homogenization, we introduce FAVoR, an author-style residual mechanism on a

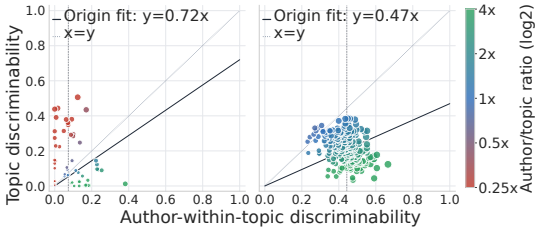

Figure 1: Coordinate-level author/topic discriminability. Each point is one scalar coordinate: a hand-crafted stylometric feature on the left or an ASCE embedding dimension on the right. Lower-right and greener points carry more author-than-topic signal; marker size scales with distance from the origin. ASCE shifts the distribution toward stronger author signal and a lower topic-on-author origin-fit slope.

shared–private federated PEFT substrate. The shared adapter carries transferable continuation updates, while client-local residual adapters retain author-specific corrections outside aggregation. We summarize the contributions of this work as follows:

- We define **author-style homogenization** as a utility-preserving loss of cross-author stylistic distinguishability in federated held-out continuation.
- We introduce an **author-style retention evaluation protocol** combining author-disjoint ASCE diagnostics, ASCE-independent gen-to-source verification, and source-space topic controls.
- We propose **FAVoR**, an author-style residual mechanism for federated PEFT. FAVoR separates shared continuation competence from client-local style residuals, optimizes the residual with local author-style targets, and keeps it outside aggregation.
- Across BlogText and Mythos-Reddit, we show that FAVoR improves author-style retention over standard and personalized federated PEFT baselines while maintaining a favorable retention–utility trade-off.

2 Related Work

Federated Personalization and PEFT for LLMs. Federated learning enables decentralized model averaging without sharing raw client data, but statistical heterogeneity can cause a single global model to underfit client-specific behavior (McMahan et al., 2017; Lin et al., 2022). Personalized FL addresses this issue through client-specific objectives, proximal regularization, or shared/client-specific adaptation structures, including pFedMe, Ditto, FedProx, and recent federated adapter variants (T. Dinh et al., 2020; Li et al., 2021, 2020; Qi et al., 2024; Yang et al., 2024). Recent federated LLM work combines FL with parameter-efficient adaptation, commonly through low-rank adapters (LoRA), to reduce communication and memory cost (Hu et al., 2022; Wang et al., 2024; Bai et al., 2024). FAVoR uses this global–local adapter pattern as a communication substrate rather than claiming it as the architectural novelty. Its contribution is to study author-style homogenization in federated held-out continuation and to define the local component as an author-style residual, trained with local author targets and kept outside aggregation.

Authorship-Conditioned Generation and Evaluation. Mythos constructs author profiles from multiple writings and evaluates whether the generated stories resemble held-out author-written stories (Ashok Kumar et al., 2025), motivating author-conditioned generation rather than federated collaboration among private authors. Style transfer and rewriting are related but distinct: formality-transfer benchmarks transform a given input while preserving content (Rao and Tetreault, 2018), whereas held-out continuation requires open-ended generation that preserves coherence, distributional fit, and author-specific style. We use MAUVE and BERTScore as continuation-utility metrics (Pillutla et al., 2023; Zhang* et al., 2020), but additionally evaluate author-style retention, inter-author separability, and collapse because fluent continuations can still converge to a shared register. Human–AI writing studies report related homogenization effects under model assistance (Padmakumar and He, 2024; Agarwal et al., 2025).

Style and Authorship Representation Learning. Authorship encoders such as LUAR and STAR learn representations for authorship verification and cross-domain transfer, while also revealing domain entanglement in learned author features (Rivera-Soto et al., 2021; Huertas-Tato et al., 2024). Style-specific embedding work shows that naturally occurring contrastive examples can leak topic; StyleDistance addresses this with content-controlled style comparisons (Patel et al., 2025). Our normalized author/style space also connects to ArcFace-style metric learning for identity separation (Deng et al., 2019). We build on this line of work by using author/style representations both as evaluation spaces and as local alignment targets.

3 Method

We study data-local federated personalized held-out continuation under *author-non-IID* data. There are N clients, where each client $i \in \{1, \ldots, N\}$ corresponds to one author with a local training corpus $\mathcal{D}_i^{\text{train}}$. The goal is to improve shared continuation competence across clients while preserving author-specific style, so that collaboration does not homogenize outputs across authors. Each client also keeps a local author-reference set $\mathcal{D}_i^{\text{ref}} \subseteq \mathcal{D}_i^{\text{train}}$ for style estimation. FAVoR imposes a shared–private aggregation boundary: the shared adapter is the only federated component, whereas private residual packs remain local and are never aggregated. Figure 2 gives the workflow; Algorithm 1 summarizes one round.

3.1 Mechanism Constraint: Shared Competence and Private Style

FAVoR stores the private adapter as a residual pack relative to the shared endpoint (Eq. (2)) and trains it against a local author target. Together with this residual structure, the shared–private boundary separates shared continuation competence from author-specific stylistic variation, limiting author-style homogenization.

Starting from a frozen base language model with parameters θ, FAVoR uses two PEFT artifacts with asymmetric communication roles:

- a **shared adapter** $A^{(g)}$, which is communicated and aggregated across clients to learn transferable continuation competence; and
- a **private residual pack** $\Delta A_{i,\tau_i}^{(p)}$, where τ_i is client i's latest private-update round; it remains local and stores the author-specific correction from the shared endpoint.

Throughout, A denotes any serialized LoRA adapter-shaped state—including shared, personalized, and residual/delta states—i.e., a set of factor tensors (U_ℓ, V_ℓ) for each layer ℓ, not the LoRA factor conventionally called A. With LoRA rank r and scaling α, the effective per-layer update is $\Delta W_\ell(A) = (\alpha/r)V_\ell U_\ell$. The personalized model for client i at its latest private update τ_i is

$$p_i(y \mid x) = \mathrm{LM}\Big(y \mid x;\ \theta \oplus \Big(A_{i,\tau_i}^{(g)} + \Delta A_{i,\tau_i}^{(p)}\Big)\Big)\,, \tag{1}$$

where $\mathrm{LM}(y \mid x; \cdot)$ denotes the autoregressive likelihood of response tokens y conditioned on x, $\oplus$ attaches an adapter to the frozen base, and $A^{(g)} + \Delta A^{(p)}$ denotes factor-wise reconstruction of the LoRA adapter state. The private residual pack is defined at the LoRA-factor level, relative to the client's shared endpoint:

$$\begin{aligned} \Delta A_{i,\tau_i}^{(p)} &= \Big\{\Delta U_{i,\tau_i,\ell}^{(p)}, \Delta V_{i,\tau_i,\ell}^{(p)}\Big\}_\ell\,, \\ \Delta U_{i,\tau_i,\ell}^{(p)} &= U_{i,\tau_i,\ell}^{(p)} - U_{i,\tau_i,\ell}^{(g)}, \\ \Delta V_{i,\tau_i,\ell}^{(p)} &= V_{i,\tau_i,\ell}^{(p)} - V_{i,\tau_i,\ell}^{(g)}. \end{aligned} \tag{2}$$

At inference, reconstructing the LoRA factors as $U_\ell^{(g)} + \Delta U_\ell^{(p)}$ and $V_\ell^{(g)} + \Delta V_\ell^{(p)}$ exactly recovers the locally trained rank-r adapter $A_{i,\tau_i}^{(p)}$, so the personalized rank is preserved by construction. Thus, Eq. (2) defines how the personalized adapter is stored locally and reconstructed, rather than assuming a linear decomposition of the effective LoRA weight update. We do not assume effective-weight additivity ($\Delta W_\ell(A^{(g)} + \Delta A^{(p)}) \neq \Delta W_\ell(A^{(g)}) + \Delta W_\ell(\Delta A^{(p)})$ in general); cross terms remain local during reconstruction. The operative shared–private boundary lies in the aggregation policy and the shared-endpoint coupling of Sec. 3.3: only $A^{(g)}$ updates are uploaded, so aggregation never averages residual packs across authors.

3.2 Angular Style Classification Encoder

To make author preservation trainable and measurable, FAVoR uses a fixed style encoder, which we call the **Angular Style Classification Encoder** (ASCE). ASCE maps a text x to a normalized style embedding:

$$z(x) = \mathrm{norm}(E_\phi(x))\,. \tag{3}$$

Here E_ϕ is the ASCE text encoder and $\mathrm{norm}(v) = v/\|v\|_2$.

ASCE is trained to make texts from the same author compact and texts from different authors separable. Concretely, we train it with an angular-margin authorship classification objective, inspired by angular-margin metric learning in face recognition (Deng et al., 2019). Let K_{ASCE} be the number of ASCE-training authors, a the author index of text x, and $\{w_k\}_{k=1}^{K_{\text{ASCE}}}$ the learned normalized classifier prototypes, with w_a selecting the positive-author prototype. These prototypes are discarded with the classifier head after ASCE training. We first define the positive-author angle $\theta_a(x)$ and logits $s_a(x)$ and $s_k(x)$ as

$$\begin{aligned} \theta_a(x) &= \arccos\Big(w_a^\top z(x)\Big)\,, \\ s_a(x) &= \gamma\cos(\theta_a(x) + m_s)\,, \\ s_k(x) &= \gamma w_k^\top z(x), \quad k \neq a. \end{aligned} \tag{4}$$

The ASCE training loss is the cross-entropy loss

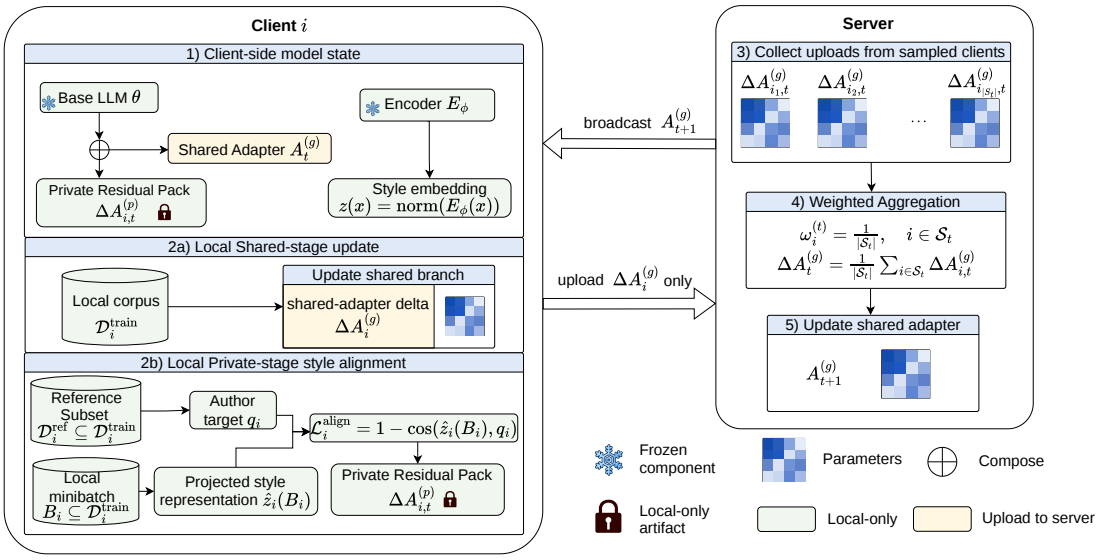

Figure 2: FAVoR workflow. Clients upload only shared-adapter deltas for aggregation; author targets, hidden states, and private residual packs remain local. Private updates use task loss plus style alignment.

over these logits:

$$\mathcal{L}^{\mathrm{ASCE}}(x, a) = -s_a(x) + \log Z_a(x)$$
$$\text{where} \quad Z_a(x) = \exp(s_a(x)) + \sum_{k \neq a} \exp(s_k(x)). \tag{5}$$

Here s_a is the margin-adjusted target-author logit, s_k are non-target logits, Z_a is the softmax normalizer, m_s is the angular margin, and γ rescales normalized cosine logits. After ASCE training, the classification head is discarded and the encoder E_ϕ is fixed.

Training-time status. ASCE is trained once before the federated procedure, and its encoder E_ϕ is then frozen. During FL, clients use this fixed encoder only as a local style-measurement module: each client computes its author target from its local reference set and computes private-stage alignment signals from local minibatches on the device.

For each client i, FAVoR constructs the local author target in Eq. (6) from the deterministic local author-reference set $\mathcal{D}_i^{\mathrm{ref}}$:

$$q_i = \mathrm{norm}\left(\frac{1}{|\mathcal{D}_i^{\mathrm{ref}}|} \sum_{x \in \mathcal{D}_i^{\mathrm{ref}}} z(x) \right). \tag{6}$$

This target remains local and serves as the client's private style anchor in the private-stage alignment loss in Eq. (11).

3.3 Local Shared and Private Updates

Each sampled client performs a shared-stage update followed by a private-stage update, as shown in Algorithm 1, lines 3–4.

For a local minibatch B_i, both the shared-stage update and the private-stage update use the same task loss:

$$\mathcal{L}_i^{\mathrm{task}}(B_i) = -\frac{1}{|B_i|} \sum_{(x,y) \in B_i} \log p_i(y \mid x). \tag{7}$$

In Eq. (7), p_i is evaluated under the adapter being updated: the local shared copy (the client's copy of the broadcast shared adapter) in the shared stage and the composed personalized model in Eq. (1) in the private stage.

The local shared copy is trained with the objective in Eq. (8), combining the task loss with a proximal stabilization term around the current server adapter:

$$\mathcal{L}_i^{\mathrm{shared}} = \mathcal{L}_i^{\mathrm{task}} + \lambda_{\mathrm{prox}} \left\| A_i^{(g)} - A_t^{(g)} \right\|_2^2. \tag{8}$$

This term discourages local shared updates from drifting too far from the round-t global adapter; the norm sums squared Frobenius norms over LoRA factor tensors. Because the shared-stage update produces the communicated adapter delta, the explicit style-alignment signal is not applied here; it is reserved for the local private-stage update.

The private-stage update uses an additional style-alignment objective. Rather than sampling generations during training, FAVoR uses teacher-forced hidden states from the personalized model. Let $H_i(B_i) \in \mathbb{R}^{|B_i| \times L \times d}$ denote the final-layer hidden states produced on minibatch B_i. Let $M_i(B_i) \in \{0, 1\}^{|B_i| \times L}$ be the supervised-response-token mask, set to 1 at positions whose label is not the ignore index and 0 otherwise (i.e., prompt and padding positions are excluded). The

pooled style state in Eq. (9) is the mask-weighted mean of H_i over response tokens:

$$h_i(B_i) = \frac{\sum_{b,\ell} M_{i,b,\ell}\, H_{i,b,\ell}}{\sum_{b,\ell} M_{i,b,\ell}}. \tag{9}$$

The projected style representation in Eq. (10) is then obtained with a lightweight projection head P_ψ:

$$\hat{z}_i(B_i) = \mathrm{norm}(P_\psi(h_i(B_i)))\,. \tag{10}$$

P_ψ is a lightweight 2-layer MLP from the LM hidden size to the ASCE embedding dimension. It is randomly initialized at the start of each client's private stage, optimized jointly with the private adapter under $\mathcal{L}_i^{\text{align}}$, and discarded afterwards. It is therefore per-client and per-round, is never uploaded or aggregated, and is not used at inference; because it is discarded after training, only gradients propagated through it to the private adapter can have a persistent effect.

The private-stage style-alignment loss in Eq. (11) measures cosine distance to the local author target from Eq. (6):

$$\mathcal{L}_i^{\text{align}} = 1 - \cos(\hat{z}_i(B_i), q_i)\,. \tag{11}$$

The private-stage objective in Eq. (12) combines the task loss from Eq. (7) with the style-alignment loss from Eq. (11):

$$\mathcal{L}_i^{\text{private}} = \mathcal{L}_i^{\text{task}} + \lambda_{\text{align}} \kappa_t \mathcal{L}_i^{\text{align}}, \tag{12}$$

where λ_{align} controls the strength of style alignment and $\kappa_t \in [0, 1]$ is a warmup coefficient. In our experiments, κ_t linearly ramps from 0 to 1 over the first 5% of private-stage steps in each round and is held at 1 thereafter. We set $\lambda_{\text{align}} = 0.3$. In the private stage, forward passes use the composed personalized model in Eq. (1); gradients are stopped through the shared endpoint $A_{i,t}^{(g)}$ and applied only to the private adapter/residual parameters. This isolates the author-specific correction as a residual relative to the shared endpoint rather than as another contribution to the public aggregation space. The resulting personalized adapter is stored as the private residual pack via Eq. (2).

3.4 Federated Aggregation

At communication round t, the server samples a subset of clients $\mathcal{S}_t$ and broadcasts the current shared adapter $A_t^{(g)}$. Each sampled client performs its local shared update and uploads only the shared-adapter delta in Eq. (13):

$$\Delta A_{i,t}^{(g)} = A_{i,t}^{(g)} - A_t^{(g)}. \tag{13}$$

Only $\Delta A_{i,t}^{(g)}$ is uploaded; all residuals, author targets, hidden states, references, and style-control signals remain local.

Algorithm 1 FAVoR Round Summary

Require: Current shared adapter $A_t^{(g)}$; sampled clients $\mathcal{S}_t$; fixed ASCE encoder E_ϕ

1: Server broadcasts $A_t^{(g)}$ to clients in $\mathcal{S}_t$
2: **for all** clients $i \in \mathcal{S}_t$ in parallel **do**
3: Train a local shared copy with $\mathcal{L}_i^{\text{shared}}$, yielding $A_{i,t}^{(g)}$
4: Train the local private adapter from $A_{i,t}^{(g)}$ with $\mathcal{L}_i^{\text{private}}$
5: Store $\{A_{i,t}^{(g)}, \Delta A_{i,t}^{(p)}\}$ locally and upload only $\Delta A_{i,t}^{(g)}$
6: **end for**
7: Server aggregates uploaded shared deltas and updates $A_{t+1}^{(g)}$

Full pseudocode in Algorithm 2.

The server aggregates sampled client deltas using Eq. (14):

$$\Delta A_t^{(g)} = \sum_{i \in \mathcal{S}_t} \omega_i^{(t)} \Delta A_{i,t}^{(g)}, \qquad \sum_{i \in \mathcal{S}_t} \omega_i^{(t)} = 1. \tag{14}$$

Because each client represents one author, we use uniform sampled-client weighting in the main experiments, i.e., $\omega_i^{(t)} = 1/|\mathcal{S}_t|$. The shared adapter is then updated according to Eq. (15):

$$A_{t+1}^{(g)} = A_t^{(g)} + \eta_g \Delta A_t^{(g)}, \tag{15}$$

where η_g is the server learning rate.

Data-locality boundary. FAVoR follows the standard federated-learning data-locality boundary (Figure 2, Algorithms 1–2). The server receives and aggregates only shared-adapter deltas $\Delta A_{i,t}^{(g)}$. Client training corpora $\mathcal{D}_i^{\text{train}}$, local author-reference sets $\mathcal{D}_i^{\text{ref}}$, author targets q_i, private residual packs $\Delta A_{i,t}^{(p)}$, style-alignment hidden states, and local style-control signals remain on device and outside aggregation. This is a data-locality claim rather than a formal privacy guarantee: shared-adapter deltas may still encode memorized text or author-identifying information, and we do not implement secure aggregation, differential privacy, or update-level leakage auditing. The fixed ASCE encoder in Eq. (3) is a centralized pre-training artifact, frozen during FL. For BlogText, its training authors are disjoint from the FL roster, and during FL, client data are used only to compute local style-control signals.

Inference. At inference, client i composes the frozen base model with $A_{i,\tau_i}^{(g)} + \Delta A_{i,\tau_i}^{(p)}$ as in Eq. (1). Because the residual is tied to the shared endpoint that produced it, unsampled clients retain their latest personalized endpoint rather than recomposing that residual with a later global adapter.

4 Experiment

Data preprocessing and metrics. Our main benchmark, BlogText, is a filtered held-out continuation split of the Blog Authorship Corpus (Schler et al., 2006) based on Tatman's Kaggle redistribution (Tatman, n.d.). We retain authors with at least 30K tokenizer tokens, split each author's documents into 80/10/10 train/validation/test partitions, and sample a style-diverse 50-author roster. Each federated client is one author, so raw author text remains client-local. Evaluation prompts are 96-token prefixes of held-out posts (192 tokens on Mythos-Reddit), and generations are capped at 220 new tokens. We filter predefined degenerate outputs: empty or short generations with fewer than 20 whitespace-delimited tokens, and repetition-collapsed generations with a distinct-bigram ratio below 0.15. Appendix B.13 reports the coverage check. External validation uses 50 Mythos-Reddit r/WritingPrompts authors and 300 balanced held-out continuations (Ashok Kumar et al., 2025; Fan et al., 2018).

We evaluate author-style retention using ASCE-space separability, geometry, and collapse metrics, plus external gen→source verification with StyleDistance, MiniLM, and stylometric authorship features (Patel et al., 2025; Sentence Transformers, 2021; Stamatatos, 2009). Unless otherwise noted, result tables report author accuracy (Acc.), macro-F1, receiver-operating-characteristic area under the curve (AUC), and equal error rate (EER). Semantic similarity is abbreviated as SemSim. We run a small human two-alternative forced-choice (2AFC) audit in which annotators choose which of two author profiles better matches a candidate continuation, used as a perceptual sanity check rather than a full preference or authenticity evaluation (see Appendix B.8).

Experimental setup. The experiments evaluate (i) whether collaborative PEFT homogenizes author style, (ii) whether FAVoR mitigates this effect without sacrificing continuation quality, and (iii) whether the ordering transfers beyond BlogText. The main BlogText evaluation uses the medium non-IID split with three training seeds: a frozen 50-client moderate long-tail allocation with target train-token Gini 0.25, realized Gini 0.196 over 904,378 training tokens, and FL schedule $(R, C, \bar{p}) = (6, 12, 1.44)$.

Table 2 focuses on collaborative federated methods (i.e., FedAvg, FedProx, pFedMe, Ditto, FedDPA, and FAVoR) while pooled PEFT is reported as a centralized non-federated reference in Table 3. FedDPA is included as the closest-family global–local adapter baseline. Local-only adaptation is treated as a strong non-collaborative style-retention reference, so base-only and local-only references are reserved for diagnostic and appendix tables and excluded from Table 2.

We ablate ASCE alignment and the shared–private residual decomposition. Appendices B.3–B.15 report prompting/retrieval-style baselines, source-space controls, sensitivity checks, external-verifier breakdowns, communication costs, and full statistical tests. Training objectives and local losses are specified in Section 3. Final backbone, adapter, batching, learning-rate, communication, storage, and runtime settings are reported in Appendix B.14.

ASCE instantiation and evaluator boundary. For BlogText, the ASCE encoder used for local style alignment and ASCE-space author-style evaluation is trained once before FL on 50 BlogText authors disjoint from the 50-author FL roster. It does not observe FL training texts, held-out prompts, generated continuations, or any main-roster author identity. We discard its classifier head and use the frozen encoder to define style targets and evaluation embedding. Appendix B.10 gives the training and evaluation details. ASCE-independent verifiers provide a complementary check in Table 3 and Appendix B.9.

4.1 Signature-Space Validity

Before using ASCE for author-style retention measurement, we test whether the author-style signal is recoverable under topic controls. Table 1 reports the compact source-space diagnostic against RoBERTa, StyleDistance, LUAR, and STAR baselines (Liu et al., 2019; Patel et al., 2025; Rivera-Soto et al., 2021; Huertas-Tato et al., 2024). Appendix B.4 gives the expanded version.

Representation	Acc. ↑	F1 ↑	AUC ↑	EER ↓
RoBERTa	0.635	0.677	0.629	0.421
StyleDist.	0.628	0.660	0.716	0.344
LUAR	0.633	0.677	0.797	0.277
STAR	<u>0.648</u>	**0.715**	<u>0.856</u>	<u>0.229</u>
ASCE	**0.651**	<u>0.692</u>	**0.863**	**0.211**

Table 1: Compact source-space representation diagnostics. Coordinate-level author/topic discriminability is reported as an appendix diagnostic in Table 8.

4.2 Main Results and Ablations

Main results. Table 2 reports the collaborative-method comparison. Table 4 separates the residual parameterization from the alignment loss. For BlogText, author accuracy, macro-F1, and verification AUC/EER are all computed in the frozen ASCE space. Verification AUC/EER uses the same-author versus the different-author generated-continuation pairs.

Table 3 expands the BlogText comparison with ASCE-independent gen→source verification using the same-topic cross-author negatives over usable BlogText metadata topics. Appendix B.10 reports the expanded BlogText ASCE results, including source-anchored and generated-text views, and Appendix B.15 reports paired tests.

Across both datasets, FAVoR shows the strongest collaborative author-style retention, with small, mixed changes in continuation utility. On BlogText, slight BERTScore decreases against Ditto/FedProx/FedDPA are outweighed by larger retention gains, and FAVoR also improves over pooled PEFT in both author accuracy and BERTScore. Table 3 confirms the ordering under ASCE-independent gen→source verifiers, while local-only remains a strong non-collaborative reference. The Mythos-Reddit breakdown appears in Appendix B.11.

A small human 2AFC audit provides a complementary perceptual check: FAVoR continuations were matched to the intended author more often than Ditto continuations (91.7% vs. 68.3% vote-level accuracy). We use this only as supporting evidence that the retained author signal is human-perceivable, not as a standalone preference or authenticity evaluation. Details and qualitative examples appear in Appendices B.8 and B.1.

Ablations. Table 4 reports FAVoR component ablations and a targeted Ditto+ASCE retrofit. The Ditto+ASCE row keeps Ditto's personalized proximal branch but adds the same local ASCE alignment loss, testing whether the gains can be explained by simply attaching ASCE to a personalized-FL baseline. The FAVoR w/o ASCE row removes the alignment loss while preserving the shared–private residual decomposition. Thus, the table separates the contribution of style alignment from the contribution of the residual aggregation constraint.

The two components are separable but not interchangeable. Attaching the same ASCE alignment loss to a personalized-FL baseline (Ditto+ASCE) improves over the shared-only Global ASCE control but recovers only a fraction of FAVoR's author-accuracy gain, whereas FAVoR without the alignment loss already attains most of it. The shared–private aggregation protocol, not the alignment loss alone, accounts for the bulk of author-style retention. Alignment provides a smaller, metric-dependent change on top. The main-text table reports seed mean. Appendix B.13 reports the mean±standard-deviation version, paired ablation tests, external-verifier tests, and configuration checks. These tests support the author-style gains while showing that utility differences are small and mixed. Source-anchored ASCE mean±standard-deviation diagnostics are reported in Appendix B.10, and the 100-author stress test is reported in Appendix B.6.

Adapter-space diagnostic. Figure 3 shows that personalized FAVoR adapters concentrate LoRA update energy in multilayer perceptron (MLP) projections rather than distributing it uniformly across target modules, consistent with the private residual pack storing structured corrections on top of the shared endpoint (details in Appendix B.2).

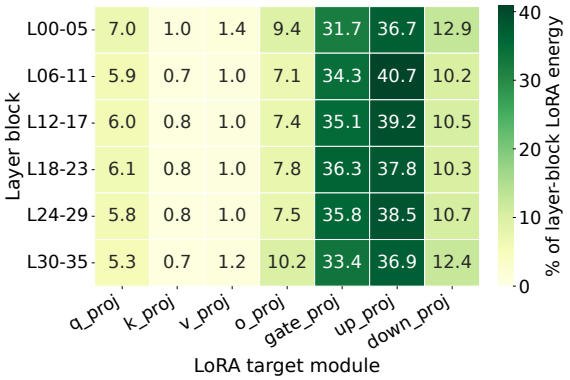

Figure 3: Adapter-space diagnostic for FAVoR on BlogText. Cells show normalized LoRA update energy by layer block and module, averaged over 50 personalized adapters.

4.3 Cold-Start Transfer to Unseen Authors

We use cold-start transfer as a supplementary test of whether the federated shared adapter provides a reusable writing initialization. The shared adapter is trained on 40 federated authors and transferred to 10 held-out authors. Each held-out author then trains a lightweight private residual from only K local examples and is evaluated on disjoint continuations. To avoid author-label evaluator leak-

Dataset	Method	Author Acc. ↑	Macro-F1 ↑	Verif. AUC ↑	EER ↓	BERTScore ↑	SemSim ↑
BlogText	FedAvg	0.199	0.164	0.680	0.371	**0.819**	0.359
	FedProx	0.174	0.161	0.674	0.365	0.817	0.353
	pFedMe	0.271	0.252	0.739	0.321	0.817	0.372
	Ditto	0.278	0.255	0.763	0.297	0.815	0.366
	FedDPA†	0.294	0.271	0.749	0.313	0.818	0.376
	FAVoR	**0.403**	**0.365**	**0.824**	**0.245**	0.815	**0.380**
Mythos-Reddit	FedAvg	0.134	0.115	0.658	0.386	0.823	0.482
	FedProx	0.141	0.113	0.663	0.377	0.823	0.478
	pFedMe	0.324	0.290	0.762	0.305	0.823	0.477
	Ditto	0.392	0.359	0.818	0.253	0.825	0.487
	FedDPA†	0.357	0.317	0.803	0.268	0.825	0.488
	FAVoR	**0.561**	**0.521**	**0.900**	**0.175**	**0.826**	**0.494**

Table 2: Main federated held-out continuation results. †FedDPA is trained with the dynamic-gate setup; reported continuations use the exported per-client mean validation gate, audited in Appendix B.12.

Method	StyleDist AUC ↑	StyleDist EER ↓	MiniLM AUC ↑	Stylom. AUC ↑
FedProx	0.724	0.334	0.631	0.651
Pooled PEFT	0.736	0.327	0.650	0.660
Ditto	0.748	0.318	0.650	0.683
FedDPA†	0.747	0.318	0.659	0.692
Local-only + ASCE	0.768	0.301	**0.685**	0.708
FAVoR	**0.771**	**0.295**	0.675	**0.714**

Table 3: Main BlogText results with external authorship verification.

Variant	Author Acc. ↑	Macro-F1 ↑	BERTScore ↑	StyleDist. AUC ↑
Global ASCE	0.161	0.140	**0.818**	0.728
Ditto + ASCE	0.318	0.286	0.816	0.746
FAVoR w/o ASCE	0.377	0.331	0.814	0.767
FAVoR	**0.403**	**0.365**	0.815	**0.771**

Table 4: Core BlogText component ablations with StyleDistance gen→source verification.

Method	Sep. Acc. ↑	Macro-F1 ↑	Target Sim. ↑
Shared only	0.150	0.166	0.191
Random shared init + private	0.308	0.278	0.201
Local-only from base	0.425	0.428	0.259
FAVoR shared init + private	**0.641**	**0.629**	**0.363**

Table 5: $K = 8$ cold-start author-separability for unseen authors. FAVoR uses no-ASCE private adaptation.

age, cold-start author-separability metrics use a separate ASCE-style encoder trained without the 10 held-out authors and classify generations using nearest-prototype matching to held-out-author prototypes built only from the same K adaptation examples. Table 5 reports the compact $K = 8$ no-ASCE private-adaptation variant, isolating the effect of the learned shared initialization. ASCE variants are reported in Appendix B.7.

At $K = 8$, FAVoR improves over local-only adaptation from the base model in author accuracy, macro-F1, and target similarity. It also outperforms the random-shared-initialization and shared-only controls, indicating that the gain depends on both a learned federated initialization and private adaptation. At $K = 16$, local-only slightly leads in author accuracy, while FAVoR retains the highest target-prototype similarity and utility metrics remain mixed. The full $K = 8$ and $K = 16$ results, including ASCE variants, copying, and utility metrics, are reported in Appendix B.7.

5 Conclusion

We studied author-style homogenization as a utility-preserving failure mode in federated personalized continuation and introduced an evaluation protocol that reveals it beyond standard utility metrics such as BERTScore and MAUVE. FAVoR instantiates an author-style residual mechanism on a shared-private PEFT substrate: shared updates are aggregated, while local-style residuals remain outside the aggregation. Empirically, standard federated PEFT can erase cross-author style even when utility metrics remain acceptable. FAVoR mitigates this mainly through the residual aggregation boundary rather than the style-alignment loss alone. Future work should extend the analysis to larger author populations, broader writing domains, and stronger privacy mechanisms for author-specific adaptation.

Limitations

Our evaluation treats author style as an operational construct measured through held-out continuation, learned authorship spaces, and external gen→source verification. These metrics capture author-specific stylistic signals, but they do not exhaust social, pragmatic, or reader-perceived notions of authorial voice; human judgments of perceived authenticity are outside the scope of this work. The BlogText ASCE encoder is trained centrally before FL on authors disjoint from the 50-author FL roster, rather than learned through a federated or differentially private procedure. This removes main-roster author overlap for both the local style-alignment target and the ASCE-space evaluator, but learned style spaces can still encode topic, genre, demographic artifacts, and other corpus-specific regularities. We therefore pair ASCE-space metrics with ASCE-independent external verifiers and report the ASCE training split explicitly.

Our privacy claim is limited to data locality during the FL procedure. We do not provide formal differential privacy, secure aggregation, or resistance to gradient- or adapter-inversion attacks, and we do not upload or release private residual packs. The experiments are also limited to English public writing domains and to a single backbone family. The 7B result is a single-seed sensitivity probe and changes both model scale and instruction-tuning status, from 3B-Instruct to 7B-base, so it should be read as robustness evidence rather than a clean scaling ablation.

Ethical Considerations

Consent, licensing, and data use. We use public research corpora and public evaluation artifacts only for non-commercial research. The main benchmark is a filtered split of the Blog Authorship Corpus distributed through Tatman's Kaggle redistribution. The external validation uses the Reddit/WritingPrompts portion of Mythos-Reddit (Schler et al., 2006; Tatman, n.d.; Ashok Kumar et al., 2025). Public availability should not be treated as blanket consent for author-style modelling or imitation. Because author identity and author-specific style are central to the task, the underlying public corpora may contain names, handles, author-linkable details, or offensive content. We do not conduct a comprehensive audit of personally identifiable information (PII)/offensive-content audit or attempt full de-identification of the source corpora. Instead, for reporting and release, we report only aggregate metrics, anonymize handles and signatures in qualitative examples, do not infer protected attributes, and do not publicly redistribute raw author texts, author-reference sets, author targets, user handles, or author-linkable generated continuations. Users of any released code or scripts must obtain the underlying datasets from their original sources and comply with the corresponding dataset licenses, platform terms, and deletion requests where applicable.

Human annotation. The human 2AFC audit involved three annotators who evaluated anonymized author-profile excerpts and candidate continuations. Annotators provided informed consent and were compensated for their time. Annotators did not receive author handles, demographic metadata, source URLs, or author identities, and we collected only their 2AFC choices, confidence ratings, and optional notes. We report only aggregate results. The audit was used as a small perceptual attribution check rather than a preference, authenticity, or quality evaluation. The audit fell under "quality assurance and evaluation activities" according to our institutional human ethics policy and therefore did not require ethics approval.

Model, evaluator, and release policy. Experiments use Qwen2.5-3B-Instruct as the backbone and public evaluator artifacts, including StyleDistance, MiniLM, BERTScore, MAUVE, and stylometric features (Qwen Team et al., 2025; Patel et al., 2025; Sentence Transformers, 2021; Zhang* et al., 2020; Pillutla et al., 2023). Any artifact release will include code, configuration files, preprocessing/evaluation scripts, and aggregate evaluation outputs, subject to review anonymity. We do not release the base-model weights, raw or author-linkable generated texts, author targets, private residual packs, or author-specific adapter checkpoints. If trained shared adapters are released, they will be distributed only when they are compatible with the base-model license and the relevant data-source terms. Because author-style adaptation can support deceptive imitation, deployment should require author consent, disclosure to affected users, access controls, and safeguards against generating text under another person's identity without permission.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Nischal Ashok Kumar, Chau Minh Pham, Mohit Iyyer, and Andrew Lan. 2025. Whose story is it? personalizing story generation by inferring author styles. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1485–1540, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. 2024. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems*, volume 37, pages 14457–14483. Curran Associates, Inc.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2024. Understanding writing style in social media with a supervised contrastively pre-trained transformer. *Knowledge-Based Systems*, 296:111867.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. 2021. Ditto: Fair and robust federated learning through personalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6357–6368. PMLR.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. 2022. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *International Conference on Learning Representations*, volume 2024, pages 642–669.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685, Albuquerque, New Mexico. Association for Computational Linguistics.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2023. Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*, 24(356):1–92.

Jiaxing Qi, Zhongzhi Luan, Shaohan Huang, Carol J. Fung, Hailong Yang, and Depei Qian. 2024. Fdlora: Personalized federated learning of large language model via dual lora tuning. *CoRR*, abs/2406.07925.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006*, pages 199–205. AAAI.

Sentence Transformers. 2021. all-MiniLM-L6-v2. Hugging Face model card.

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

Canh T. Dinh, Nguyen Tran, and Josh Nguyen. 2020. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pages 21394–21405. Curran Associates, Inc.

Rachael Tatman. n.d. Blog authorship corpus. Kaggle dataset. File used: blogtext.csv. URL: `https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus`. Accessed 2026-05-18.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. 2024. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In *Advances in Neural Information Processing Systems*, volume 37, pages 22513–22533. Curran Associates, Inc.

Yiyuan Yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. 2024. Dual-personalizing adapter for federated foundation models. In *Advances in Neural Information Processing Systems*, volume 37, pages 39409–39433. Curran Associates, Inc.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A FAVoR Training Details

A.1 Complete Training Pseudocode

Algorithm 2 expands the compact round summary in Algorithm 1. It makes explicit which quantities remain client-local and which updates are uploaded for aggregation.

Algorithm 2 FAVoR: Federated Shared-Private Training with ASCE Alignment

Require: Frozen base model θ; fixed ASCE encoder E_ϕ; initial shared adapter $A_0^{(g)}$
Require: Local private residual-pack store; local corpora $\{\mathcal{D}_i^{\text{train}}\}_{i=1}^N$; local reference sets $\{\mathcal{D}_i^{\text{ref}}\}_{i=1}^N$
for $t = 0, 1, \ldots, T-1$ **do**
  Server samples clients $\mathcal{S}_t$
  Server broadcasts shared adapter $A_t^{(g)}$ to clients in $\mathcal{S}_t$
  for all clients $i \in \mathcal{S}_t$ in parallel **do**
    Construct local author target q_i from $\mathcal{D}_i^{\text{ref}}$ using Eq. (6)
    Initialize local shared model $\theta \oplus A_t^{(g)}$
    Update the local shared copy using $\mathcal{L}_i^{\text{shared}}$ in Eq. (8), yielding $A_{i,t}^{(g)}$
    Compute shared-adapter delta $\Delta A_{i,t}^{(g)}$ using Eq. (13)
    Initialize the private adapter from $A_{i,t}^{(g)}$; then, holding $A_{i,t}^{(g)}$ fixed, update the private adapter using $\mathcal{L}_i^{\text{private}}$ in Eq. (12)
    Store the local shared endpoint $A_{i,t}^{(g)}$ and updated private residual pack $\Delta A_{i,t}^{(p)}$ using Eq. (2)
    Upload only $\Delta A_{i,t}^{(g)}$ to the server
  end for
  Server aggregates shared deltas using Eq. (14)
  Server updates shared adapter using Eq. (15)
end for
return Final shared adapter $A_T^{(g)}$ and retained client states $\{A_{i,\tau_i}^{(g)}, \Delta A_{i,\tau_i}^{(p)}\}_{i=1}^N$

B Additional Experiments and Diagnostics

B.1 Qualitative Error Analysis

Table 6 reports the qualitative comparison referenced from the main text. The examples are selected from generated held-out continuations after matching seed, author, and prompt across methods; no additional training or decoding was performed. Each row reports a prompt cue and shortened method-specific continuations; handles and signatures are anonymized in brackets. The examples are intended to make the measured homogenization failure mode legible at the text level, while the aggregate conclusions are based on the quantitative metrics.

Error types. The examples show four recurring patterns behind the aggregate metrics. *Shared-register smoothing* replaces author-specific punctuation, spelling, and pacing with a fluent default blog continuation. *Topic-only personalization* preserves the prompt topic but loses the author's habitual register, which is common in Ditto and pooled PEFT when the private signal is weak or unstable. *Residual style recovery* in FAVoR restores local lexical markers, sign-offs, link habits, and paragraph rhythm while usually retaining the prompt topic. *Over-localization* appears most clearly in local-only outputs: author markers are strong, but the model may overproduce signatures, links, or comment-thread artifacts, so local-only is a useful style-retention reference but not always the best utility reference.

B.2 Adapter-Space Diagnostic

Figure 3 provides an auxiliary parameter-space diagnostic for the shared–private decomposition. For the 50 personalized adapters in a representative BlogText run, we compute the effective LoRA update $\Delta W = (\alpha/r)BA$ for each target module and layer block, then report the percentage of total update energy assigned to each cell after averaging across clients. The pattern is not uniform: most updated energy is concentrated in MLP projections, especially `gate_proj` and `up_proj`, while `k_proj` and `v_proj` receive comparatively little energy. This diagnostic is supportive rather than causal; the main mechanism evidence comes from the component ablations in Table 4.

B.3 Prompting and Retrieval Baselines

Table 7 compares FAVoR with non-training personalization baselines inspired by Mythos-style author artifacts. These baselines test whether raw author history, retrieval-augmented generation (RAG), or distilled author sheets can recover author style without federated optimization. They serve as supplementary non-training controls; the primary collaborative-learning comparison is reported in Tables 2 and 3. Because the prompting baselines were run on a fixed prompting-aligned subset, Table 7 reports a source-anchored author-disjoint ASCE rescore of those generated outputs. The prototype and gen→source columns follow the source-anchored view in Table 17, but are computed only on this imported prompting-baseline subset; they are not the generated-text Author Acc. values reported in Table 2.

Prompt cue	Shared FL / pooled excerpt	Ditto excerpt	FAVoR excerpt	Local-only excerpt
Teen-chat spelling and pacing: "omg", "soo", "u", "lol", "ttyl"	"we can always be friends ... after college"; generic affective prose that smooths the prompt's spelling/pacing.	"bye ... love ya ... l8r"; keeps some spelling, then destabilizes into repeated `lol`-style sign-offs.	"wishin u luck ... ttyl ... lol"; recovers the target register while staying near the school/friends topic.	"wishin u luck in that dance thing ... omg i'm so happy"; strong surface imitation, with some over-repetition.
Blog signature and postscript style: "loooong post", "50(?)", unfinished computer post	"pages of notes from my classes ... homework for English"; fluent school-log continuation but misses the signature.	"The Great Gatsby" and other plausible school content, but again misses the author's recurring sign-off.	"-[sig] ... P.S. ... *yawn* ... [sig] out."; restores the local sign-off rhythm.	"-[sig] ... [sig] out! ... [alias] P.P.S."; retains signature more strongly but over-localizes into aliases and repeated postscripts.
Early web-blog music register: site-handle sign-off, "nu national anthem"	"The American flag ... God bless America"; topic-only continuation in a polished patriotic register.	"how great Britain is"; topical anthem text, but broad and weakly tied to the author's blog style.	"So here it is! ... `nu-national-anthem.mp3` ... see ya later! [handle]"; preserves web-era lexical cues.	"The Power Of Love" and a video link trail; keeps style markers, but switches to a different song.
Domestic diary with reader aside: "Is doable a word?"	"work without having to worry about where everything is"; turns the prompt into generic productivity prose.	"can't seem to find anything ... I can always find what I'm looking for"; plausible but less idiosyncratic diary prose.	"going well so far ... Oh yeah ... O&O!!!"; preserves diary cadence and reader-aside register.	"old files ... little ghosts ... O&O!!!"; strong phrasing, but drifts into a dated daily log.
Reflective cookie-bake essay: "too simple and too much like what normal people do"	"at the local hospital ... first day back to school"; loses the cookie topic and moves to generic diary material.	"I make the dough, she decorates it, he eats it"; personalizes around family cookies but becomes a generic family activity.	"I am off to make cookies ... The Cookie Book"; maintains the essay-blog frame with recipes, links, and the cookie-bake topic.	"3 comments: Anonymous said ... cookie decorating"; strong blog artifacts, but lower utility because it drifts into comments.

Table 6: Qualitative error analysis over anonymized held-out continuation examples. Entries are shortened, generated-continuation excerpts with brief diagnoses. Shared FL is represented by FedProx; pooled PEFT shows similar smoothing and artifact-drift patterns in the same selected prompts.

Method	FL	Train	Raw Hist.	RAG	Distilled Artifact	Proto. Acc. ↑	Proto. F1 ↑	Gen→Src AUC ↑	EER ↓	Silh. ↑	BERTScore ↑	SemSim ↑	MAUVE ↑
FAVoR	✓	✓	✗	✗	✗	**0.288**	**0.208**	**0.760**	**0.319**	**-0.045**	0.815	0.385	**0.313**
Mythos Avg-Author	✗	✗	✗	✗	✗	0.093	0.047	0.627	0.394	-0.184	0.814	0.395	0.013
Mythos RAG-Only	✗	✗	✓	✓	✗	0.131	0.095	0.713	0.372	-0.198	**0.819**	0.405	0.025
Mythos Sheet + RAG	✗	✗	✓	✓	✓	0.176	0.126	0.689	0.389	-0.145	0.815	**0.415**	0.093
Mythos Sheet (no persona) + RAG	✗	✗	✓	✓	✓	0.130	0.088	0.698	0.365	-0.158	0.816	0.411	0.030
Mythos Sheet Only	✗	✗	✗	✗	✓	0.150	0.103	0.665	0.378	-0.143	0.812	0.396	0.061

Table 7: Prompting and retrieval baselines on BlogText. Author-style columns use the author-disjoint ASCE source-anchored rescore on the shared prompting-aligned subset for this supplemental comparison. Prototype accuracy/F1 classifies generated continuations by the nearest source-train author prototype, and Gen→Src AUC/EER uses the same-author versus different-author generated-to-source pairs. Utility columns are the original generation-quality metrics for the same prompting-baseline comparison.

B.4 Source-Space Controls

Table 8 reports the native-dimensional version of the source-space diagnostic introduced in Section 4.1. Together with Table 1, it checks whether the internal ASCE space is competitive with external authorship/style encoders and whether the author signal remains visible under topic-controlled verification.

B.5 7B Base-Model Sensitivity

Table 9 reports a single-seed backbone-sensitivity probe on the same 50-author medium non-IID BlogText split. This run replaces the main 3B instruction-tuned backbone with a Qwen2.5-7B base model (Qwen Team et al., 2025) and keeps the held-out continuation protocol, LoRA recipe, decoding, ASCE evaluator, and full $K = 50$ metric pass fixed. Because this changes both parameter scale and instruction-tuning status, and because it uses only one seed, we treat it as robustness evidence rather than a main multi-seed result.

The ordering matches the main finding. FAVoR is the strongest collaborative method on every author-style retention column. Relative to FedAvg, FAVoR improves author accuracy by 0.310 and verification AUC by 0.183, while retaining all generated continuations after filtering.

B.6 100-Author Scale Stress Test

To check whether the main 50-author results depend on a small roster, we ran an additional BlogText stress test with 100 authors over three seeds. The roster preserves the original 50 authors and adds 50 authors using the same style-cluster balancing procedure, with the same per-author token caps as the main data construction. This

Representation space	Dim.	Author Acc. ↑	Macro-F1 ↑	Author-in-topic ω^2 ↑	Topic ω^2 ↓	A/T ratio ↑	Strict AUC ↑	EER ↓
RoBERTa-base	768	0.635	0.677	0.1135	0.0557	1.95	0.629	0.421
StyleDistance	768	0.628	0.660	**0.2357**	0.0577	**4.21**	0.716	0.344
LUAR-MUD doc1	512	0.633	0.677	0.1257	**0.0364**	3.52	0.797	0.277
STAR	1024	0.648	**0.715**	0.1886	0.0478	4.16	0.856	0.229
ASCE Author-Style Signature Space	256	**0.651**	0.692	0.1916	0.0499	3.64	**0.863**	**0.211**

Table 8: Native-dimensional source-space diagnostics.

Method	Author Acc. ↑	Macro-F1 ↑	Verif. AUC ↑	EER ↓	Silhouette ↑	Asst. Score ↓	BERTScore ↑	SemSim ↑	MAUVE ↑
Base only	0.207	0.191	0.695	0.353	-0.141	0.865	0.815	0.353	0.130
FedAvg	0.219	0.170	0.716	0.343	-0.136	0.230	0.811	0.361	0.263
FedProx	0.214	0.177	0.698	0.365	-0.149	**0.183**	0.808	0.352	0.327
pFedMe	0.293	0.245	0.807	0.263	-0.053	0.625	**0.816**	0.365	0.327
Ditto	0.368	0.328	0.836	0.243	-0.033	0.424	0.813	**0.381**	**0.417**
FAVoR	**0.528**	**0.487**	**0.899**	**0.185**	**0.134**	0.251	0.810	0.378	0.334

Table 9: Single-seed Qwen2.5-7B base-model sensitivity probe on the full 50-author BlogText held-out continuation split. Bold marks the best listed value; base-only is a non-collaborative reference.

experiment is meant as scaling evidence rather than a new primary significance table: it uses $(R, C, \bar{p}) = (20, 20, 4.0)$ rounds, clients per round, and expected participations per author, compared with $(6, 12, 1.44)$ in the main 50-author BlogText runs.

Table 10 shows that the main ordering largely persists at the larger author scale. Shared-only FedAvg has low author separability, negative silhouette, and high verification error, indicating strong author-style homogenization. FAVoR substantially mitigates this collapse, recovering roughly 70% of the FedAvg-to-local-only gap in separability accuracy, 74% in macro-F1, 81% in verification AUC, and 72% in silhouette. The result should still be read conservatively: FAVoR is close to Ditto on verification AUC/EER and stronger on closed-set author separability, while Ditto obtains higher MAUVE. Local-only remains the non-collaborative style retention reference.

B.7 Cold-Start New-Author Details

Table 11 reports the full cold-start new-author evaluation. The shared adapter is trained on 40 federated training authors and evaluated after private K-shot adaptation on 10 held-out authors. Author-separability metrics use the leakage-free $\mathrm{A}_{\mathrm{train}}$-only ASCE encoder (Section 4.3) and held-out-author prototypes built only from the K adaptation examples.

B.8 Human 2AFC Audit

We conduct a small human two-alternative forced-choice (2AFC) audit as a perceptual sanity check for author-style retention. For each item, annotators are shown two anonymized author profiles, one candidate continuation, and are asked to choose which profile better matches the candidate. Annotators also provide a confidence score from 1 to 5. The task is designed to test whether an author-specific signal in a continuation is human-perceivable under a controlled attribution setting; it is not a preference study and does not ask annotators to judge overall writing quality, fluency, or authenticity.

The audit set contains 50 items randomly sampled within source categories: 20 Ditto continuations, 20 FAVoR continuations, and 10 held-out author-written continuations. Each item is annotated by three annotators, giving 150 total votes. We aggregate results both at the vote level, treating each annotator's decision as one observation, and at the item level, using the majority decision across the three annotators.

Participant instructions. The annotation interface was a text-only form with two author-profile boxes, one candidate continuation, a forced-choice response, a confidence response, and an optional comment field. No screenshots contained additional instructions beyond the text reproduced here. Annotators were shown the following instructions before the task:

This pack is for a 2-author forced-choice style

Method	Sep. Acc. ↑	Macro-F1 ↑	Verif. AUC ↑	EER ↓	Centroid Dist. ↑	Silhouette ↑	SemSim ↑	MAUVE ↑
FedAvg shared	0.177 ± 0.013	0.150 ± 0.013	0.729 ± 0.002	0.331 ± 0.002	0.677 ± 0.007	-0.176 ± 0.010	0.385 ± 0.006	0.272 ± 0.040
Ditto	0.386 ± 0.022	0.342 ± 0.017	0.883 ± 0.012	0.193 ± 0.008	0.809 ± 0.007	0.041 ± 0.020	0.411 ± 0.008	**0.386 ± 0.105**
FAVoR	**0.441 ± 0.021**	**0.417 ± 0.028**	**0.889 ± 0.006**	**0.186 ± 0.008**	**0.821 ± 0.007**	**0.062 ± 0.017**	**0.413 ± 0.008**	0.308 ± 0.066
Local-only (ref.)	0.555 ± 0.006	0.509 ± 0.006	0.926 ± 0.011	0.147 ± 0.018	0.886 ± 0.003	0.154 ± 0.006	0.412 ± 0.008	0.406 ± 0.042

Table 10: K=100 BlogText scale stress test over three seeds. Values are mean ± sample standard deviation. Bold marks the best collaborative method; local-only is a non-collaborative style-retention reference. Verification AUC/EER uses the same-author versus different-author generated-continuation pairs in the K=100 ASCE space.

Method	K	Sep. Acc. ↑	Macro-F1 ↑	Target Sim. ↑	Copy-5 ↓	SemSim ↑	BERTScore ↑	MAUVE ↑
Base model, no adaptation	8	0.175	0.210	0.115	0.000	0.282	0.806	0.042
FAVoR shared init + private residual (ASCE)	8	0.538	0.555	0.353	0.154	0.304	0.793	0.071
FAVoR shared init + private residual (no ASCE)	8	**0.641**	**0.629**	**0.363**	0.128	0.301	0.792	0.078
Local-only + ASCE from base	8	0.400	0.377	0.261	0.050	**0.331**	0.807	**0.115**
Local-only from base	8	0.425	0.428	0.259	0.075	0.319	0.806	0.078
Random shared init + private residual	8	0.308	0.278	0.201	0.026	0.307	**0.809**	0.069
Shared only, no private residual	8	0.150	0.166	0.191	**0.000**	0.279	0.792	0.085
Base model, no adaptation	16	0.175	0.150	0.106	**0.000**	0.282	0.806	0.042
FAVoR shared init + private residual (ASCE)	16	0.425	0.403	0.342	0.100	0.305	0.783	0.042
FAVoR shared init + private residual (no ASCE)	16	0.525	0.479	**0.386**	0.025	0.305	0.785	**0.092**
Local-only + ASCE from base	16	0.474	0.457	0.318	**0.000**	0.308	0.804	0.065
Local-only from base	16	**0.553**	**0.562**	0.326	0.105	**0.313**	0.806	0.048
Random shared init + private residual	16	0.325	0.320	0.214	0.025	0.299	**0.809**	0.065
Shared only, no private residual	16	0.150	0.170	0.189	**0.000**	0.279	0.792	0.085

Table 11: Full cold-start results for unseen authors. The shared adapter is trained on 40 federated training authors and transferred to 10 held-out authors; author-separability metrics use the leakage-free A_{train}-only ASCE encoder and held-out-author prototypes built from the same K adaptation examples.

attribution task. Annotator task: 1. Read Author A's source excerpts. 2. Read Author B's source excerpts. 3. Read the candidate text. 4. Choose which author the candidate better matches in writing style. 5. Enter confidence from 1 to 5.

Source	Items	Votes	Vote acc.	Maj. acc.	Mean conf.
Ditto	20	60	68.3%	70.0%	2.73
FAVoR	20	60	91.7%	95.0%	3.79
Human ref.	10	30	93.3%	90.0%	3.50
Overall	50	150	82.7%	84.0%	3.31

Table 12: Small human 2AFC audit. Vote accuracy is over individual annotator votes; majority accuracy is after three-vote item aggregation.

Table 12 summarizes the audit results. Overall vote-level accuracy is 82.7%, and item-level majority accuracy is 84.0%. On the audited generated continuations, FAVoR continuations are attributed to the intended author more often than Ditto continuations, with 91.7% versus 68.3% vote-level accuracy and 95.0% versus 70.0% majority accuracy. Held-out human continuations provide a reference row and reach 93.3% vote-level accuracy. Inter-annotator agreement is moderate for this subjective authorship task: mean raw pairwise agreement is 0.787 and Fleiss' κ is 0.558. At the item level, all 50 items have a non-tie majority; 34 items have unanimous 3–0 majorities and 16 items have 2–1 majorities. This audit supports the interpretation that FAVoR preserves an author-attributable signal that is perceptible to human annotators in a 2AFC attribution setting. However, it has two limitations. First, the audit is small, with 50 items and three annotators. Second, because annotators see author profiles, their decisions may reflect topical, biographical, or content cues in addition to style. We therefore use the audit only as complementary perceptual evidence alongside the main automatic author-retention metrics and ASCE-independent external verifiers, rather than as a standalone human preference or authenticity evaluation.

B.9 External Verifier Robustness

Table 3 reports compact external-verification averages; Table 13 gives the full evaluator break-

down. Table 16 summarizes cross-evaluator agreement for the central FAVoR-versus-Ditto comparison. Table 15 documents topic recovery, the effective generated-pair negative mode, and same-topic hard-negative feasibility for the verification protocol. Table 14 provides an additional compact BlogText metric view. Source AUC is a verifier-validity check on held-out human source-text pairs, while generated-text columns evaluate method outputs under the same-topic cross-author negatives. The `indUnk` metadata bucket is retained in the evaluation roster but treated as unavailable for same-topic negative sampling. Bold and underline mark the best and second-best values, with lower better for EER.

B.10 Author-Disjoint ASCE Training and Evaluation Details

The BlogText ASCE encoder used for local style alignment and ASCE-space evaluation is trained on 50 BlogText authors disjoint from the main FL roster. It uses 4,959 training texts and 742 validation texts, achieving 0.802 validation accuracy and 0.673 macro-F1, and then discards its classifier head. ASCE-aligned methods use this frozen encoder during training to construct local style-alignment targets, and all BlogText ASCE-space author-style metrics use the same author-disjoint encoder. For ASCE-aligned BlogText experiments, the training configurations set both the private-stage alignment encoder and the ASCE-space evaluator to this author-disjoint artifact; the manifest records zero author overlap with the FL roster.

Tables 17 and 18 give two detailed views of the author-disjoint ASCE results behind the main BlogText summary. Table 17 is source-anchored: it compares each generated continuation against source-train author prototypes and gen→source verification pairs. Table 18 is the expanded version of the BlogText author-style rows in Table 2, adding pooled and diagnostic references such as Local-only + ASCE, Global ASCE, Ditto + ASCE, and FAVoR w/o ASCE.

Together, these two views document author-style retention in the same author-disjoint ASCE space used by the BlogText training and evaluation protocol. In the gen→source view, local-only remains the strongest non-collaborative reference, while FAVoR is the strongest collaborative method: it ranks second overall on prototype recovery and AUC, and best overall on EER. In the expanded generated-text view, FAVoR again leads the collaborative methods on author recovery and verification EER, and remains close to the local-only reference on F1 and AUC. The FAVoR variant without ASCE is also competitive, indicating that residual personalization carries substantial author signal; adding ASCE improves recovery and verification while leaving silhouette essentially tied with the local-only reference. ASCE is still a centralized pre-training artifact, so we also report ASCE-independent external verifiers.

B.11 External Domain Validation

Table 19 reports the Mythos-Reddit external comparison for the collaborative methods. Table 20 reports the diagnostic version of the external evaluation, including signed-gap metrics and assistant-phrase diagnostics. The completed validation uses 50 r/WritingPrompts authors with 300 balanced held-out continuations per run. On Mythos-Reddit, FAVoR is again the strongest collaborative federated method on the style-separability columns, suggesting that the retention trend is not specific to BlogText.

B.12 FedDPA Gate Sensitivity Probe

FedDPA stores prompt-wise dynamic gate references during training, while the main continuation path reports the exported per-client mean validation gate for generation compatibility. To audit this export, we run a bounded seed probe on 100 matched BlogText held-out continuation prompts, generating one continuation with the exported summary gate and one with prompt-wise dynamic gates. Prompt-wise dynamic gating changes the gate weights: the mean global-gate weight is 0.350 under dynamic gating versus 0.429 under the exported summary gate, with mean absolute difference 0.083, median absolute difference 0.056, and maximum absolute difference 0.437. Table 21 shows that this gate change leaves author accuracy unchanged and BERTScore nearly identical in the bounded probe. Dynamic gating slightly lowers verification AUC and SemSim in this sample, so reporting the audited summary-gate FedDPA rows does not overstate FedDPA's author-style retention.

Method	Evaluator	Source AUC	Gen→Src AUC	Gen→Src EER	Gen→Gen AUC
FedProx	StyleDistance	0.705 ± 0.000	0.724 ± 0.011	0.334 ± 0.006	0.654 ± 0.021
FedProx	MiniLM	0.635 ± 0.000	0.631 ± 0.002	0.414 ± 0.003	0.532 ± 0.005
FedProx	Stylometry	0.619 ± 0.000	0.651 ± 0.003	0.385 ± 0.005	0.573 ± 0.021
Pooled PEFT	StyleDistance	0.705 ± 0.000	0.736 ± 0.006	0.327 ± 0.004	0.691 ± 0.016
Pooled PEFT	MiniLM	0.635 ± 0.000	0.650 ± 0.001	0.406 ± 0.005	0.574 ± 0.004
Pooled PEFT	Stylometry	0.619 ± 0.000	0.660 ± 0.008	0.384 ± 0.008	0.562 ± 0.022
Ditto	StyleDistance	0.705 ± 0.000	0.748 ± 0.004	0.318 ± 0.006	0.720 ± 0.009
Ditto	MiniLM	0.635 ± 0.000	0.650 ± 0.005	0.401 ± 0.005	0.594 ± 0.013
Ditto	Stylometry	0.619 ± 0.000	0.683 ± 0.011	0.363 ± 0.008	0.656 ± 0.034
FedDPA†	StyleDistance	0.705 ± 0.000	0.747 ± 0.005	0.318 ± 0.003	0.725 ± 0.011
FedDPA†	MiniLM	0.635 ± 0.000	0.659 ± 0.012	0.392 ± 0.012	0.573 ± 0.027
FedDPA†	Stylometry	0.619 ± 0.000	0.692 ± 0.008	0.351 ± 0.008	0.652 ± 0.007
Local-only + ASCE	StyleDistance	0.705 ± 0.000	0.768 ± 0.006	0.301 ± 0.002	0.768 ± 0.004
Local-only + ASCE	MiniLM	0.635 ± 0.000	**0.685 ± 0.006**	**0.374 ± 0.008**	**0.645 ± 0.023**
Local-only + ASCE	Stylometry	0.619 ± 0.000	0.708 ± 0.003	0.344 ± 0.007	**0.685 ± 0.010**
FAVoR	StyleDistance	0.705 ± 0.000	**0.771 ± 0.006**	**0.295 ± 0.003**	**0.773 ± 0.007**
FAVoR	MiniLM	0.635 ± 0.000	0.675 ± 0.010	0.378 ± 0.006	0.624 ± 0.021
FAVoR	Stylometry	0.619 ± 0.000	**0.714 ± 0.011**	**0.342 ± 0.011**	0.678 ± 0.020

Table 13: External-verifier breakdown.

Method	Author Acc.↑	Macro-F1↑	Centroid↑	Silhouette↑	Asst.↓	BERTScore↑	Sem. sim.↑	MAUVE↑
FedProx	0.174 ± 0.008	0.161 ± 0.009	0.431 ± 0.009	-0.199 ± 0.004	0.264 ± 0.041	**0.817 ± 0.002**	0.353 ± 0.005	0.200 ± 0.005
Pooled PEFT	0.254 ± 0.021	0.232 ± 0.027	0.537 ± 0.005	-0.184 ± 0.030	**0.219 ± 0.017**	0.802 ± 0.001	0.359 ± 0.002	0.389 ± 0.059
Ditto	0.278 ± 0.011	0.255 ± 0.002	0.546 ± 0.004	-0.097 ± 0.026	0.328 ± 0.016	0.815 ± 0.001	0.366 ± 0.005	0.378 ± 0.074
Local-only + ASCE	0.402 ± 0.026	**0.366 ± 0.016**	**0.643 ± 0.012**	**-0.037 ± 0.025**	0.239 ± 0.015	0.813 ± 0.001	0.375 ± 0.009	**0.486 ± 0.053**
FAVoR	**0.403 ± 0.004**	0.365 ± 0.009	0.624 ± 0.006	-0.039 ± 0.022	0.260 ± 0.004	0.815 ± 0.000	**0.380 ± 0.001**	0.431 ± 0.026

Table 14: Additional compact BlogText metric summary.

Audit item	Value
Evaluation authors	50
Topic source	Original BlogText metadata joined by author ID
Topics with ≥3 authors	3 usable; 4 including all buckets
Largest topic bucket	Student: 13/50 (26.0%)
indUnk bucket	11/50 (22.0%)
Dominant-topic fraction	mean 1.0; min 1.0
Effective hard-negative mode	Same-topic cross-author
Protocol A same-prompt feasible	False
Protocol A same-topic negatives	2818–2850
Protocol C same-prompt feasible	False
Protocol C same-topic negatives	1409–1425
External-evaluator agreement	3/3 support FAVoR > Ditto

Table 15: Topic and negative-sampling summary.

Metric	Summary gate	Dynamic gate	Dynamic − summary
Author Acc.	0.350	0.350	+0.000
Macro-F1	0.266	0.283	+0.018
Verif. AUC	0.778	0.763	-0.015
EER	0.299	0.303	+0.004
BERTScore	0.8191	0.8178	-0.0013
SemSim	0.3982	0.3827	-0.0155
Asst. score	0.4035	0.4302	+0.0267

Table 21: Bounded FedDPA gate-sensitivity probe.

B.13 Component-Ablation Significance and Run Checks

Table 22 expands the compact main-text ablation table with mean ± sample standard deviation over three seeds. Its Author Acc. and Macro-F1 columns are generated-text ASCE prototype metrics, while the StyleDistance, MiniLM, and stylometry columns are ASCE-independent, same-topic gen→source AUC/EER.

Table 23 reports paired tests for the component ablations in Table 4. The tests use matched seed–author–prompt examples. For author accuracy, p uses McNemar's exact test; for continuous metrics, p uses a Wilcoxon signed-rank test. Directional Δ is signed, so positive values favour the left variant in each comparison; for the assistant score, this means the left variant has a lower raw assistant score. Confidence intervals are cluster bootstrap intervals over seed–author clusters.

Comparison	Metric	Support	Fraction	Supporting evaluators
FAVoR > Ditto	Gen→Src primary AUC	3/3	1.0	minilm, styledistance, stylometry

Table 16: Cross-evaluator agreement.

Method	Seeds	Proto. Acc. ↑	Proto. F1 ↑	Gen→Src AUC ↑	Gen→Src EER ↓
FedAvg	3	0.140 ± 0.009	0.096 ± 0.001	0.651 ± 0.023	0.379 ± 0.008
FedProx	3	0.157 ± 0.011	0.111 ± 0.005	0.645 ± 0.023	0.387 ± 0.020
pFedMe	3	0.176 ± 0.014	0.120 ± 0.010	0.691 ± 0.019	0.363 ± 0.018
Ditto	3	0.219 ± 0.016	0.169 ± 0.012	0.715 ± 0.014	0.351 ± 0.008
Pooled PEFT	3	0.193 ± 0.008	0.148 ± 0.012	0.689 ± 0.012	0.363 ± 0.016
Local-only + ASCE	3	$\mathbf{0.342 \pm 0.023}$	$\mathbf{0.292 \pm 0.032}$	$\mathbf{0.764 \pm 0.006}$	0.307 ± 0.013
Global ASCE	3	0.149 ± 0.008	0.112 ± 0.007	0.664 ± 0.024	0.378 ± 0.032
Ditto + ASCE	3	0.237 ± 0.006	0.173 ± 0.016	0.699 ± 0.020	0.366 ± 0.021
FAVoR w/o ASCE	3	0.280 ± 0.031	0.223 ± 0.035	0.741 ± 0.023	0.326 ± 0.029
FAVoR	3	0.315 ± 0.022	0.262 ± 0.031	0.752 ± 0.026	$\mathbf{0.304 \pm 0.016}$

Table 17: Generated-to-source ASCE results with the author-disjoint encoder. The encoder is trained on BlogText authors disjoint from the main 50-author FL roster; its classifier head is not used for this table. Values are the mean ± standard deviation over seeds.

Table 24 repeats the ablation comparison using ASCE-independent external gen→source verification. Under the same-topic cross-author negatives, the external alignment gain over the no-ASCE variant is small and mixed, while both private-residual variants are consistently stronger than the Global ASCE control. Directional Δ is signed so positive values favour the left variant; for EER, this means lower EER for the left variant. Finally, Table 25 reports a compact configuration check for the ablation rows. It verifies whether each run uses a private residual, ASCE alignment, and the shared proximal term. For held-out continuation, the deterministic quality filter removes generations with fewer than 20 whitespace-delimited tokens, including empty generations, or a distinct-bigram ratio below 0.15.

Variant	Private residual	ASCE align.	Shared prox.
Global ASCE	✗	✓	✗
FAVoR w/o ASCE	✓	✗	✓
FAVoR	✓	✓	✓

Table 25: Component-ablation configuration check.

B.14 Hyperparameters, Efficiency, and Communication Cost

Table 26 summarizes the hyperparameters and the communication and storage footprint of the BlogText FAVoR runs. The measurements are taken from the three medium non-IID FAVoR runs used in the main BlogText evaluation. All runs use Qwen2.5-3B-Instruct with 4-bit Quantized Low-Rank Adaptation (QLoRA) (Qwen Team et al., 2025; Dettmers et al., 2023) and train LoRA adapters on the q, k, v, o, gate, up, and down projections with rank 16, $\alpha = 32$, and dropout 0.05. The serialized shared adapter is 119,801,528 bytes, corresponding to 114.3 MiB and 29.93M trainable LoRA parameters. These settings were set before the held-out test evaluation; no hyperparameter was tuned on test continuations. Under the protocol, each sampled client uploads only the shared-adapter delta. With 12 clients per round for six rounds, one seed sends 72 client updates, or 8.03 GiB of client-to-server adapter payloads when counted as serialized fp32 LoRA tensors. Server-to-client broadcasts have the same per-recipient size, so counting both directions gives 16.07 GiB per seed. Private residual packs are not communicated; retaining one final residual per 50-author evaluation roster would require 5.58 GiB if stored as dense fp32 LoRA factor-delta tensors. The stored experiment artifacts are larger than the deployment footprint because they retain per-round shared endpoints, personalized adapter checkpoints, and private residual checkpoints for analysis.

Method	Author Acc. ↑	Macro-F1 ↑	Silhouette ↑	Gen-Gen AUC ↑	Gen-Gen EER ↓
FedAvg	0.199 ± 0.008	0.164 ± 0.015	-0.194 ± 0.014	0.680 ± 0.029	0.371 ± 0.018
FedProx	0.174 ± 0.008	0.161 ± 0.009	-0.199 ± 0.004	0.674 ± 0.003	0.365 ± 0.011
pFedMe	0.271 ± 0.047	0.252 ± 0.033	-0.131 ± 0.005	0.739 ± 0.011	0.321 ± 0.014
Ditto	0.278 ± 0.011	0.255 ± 0.002	-0.097 ± 0.026	0.763 ± 0.025	0.297 ± 0.018
Ditto + ASCE	0.318 ± 0.037	0.286 ± 0.046	-0.115 ± 0.004	0.761 ± 0.010	0.309 ± 0.010
Pooled PEFT	0.254 ± 0.021	0.232 ± 0.027	-0.184 ± 0.030	0.699 ± 0.020	0.344 ± 0.009
Local-only + ASCE	0.402 ± 0.026	**0.366 ± 0.016**	-0.037 ± 0.025	**0.828 ± 0.017**	0.252 ± 0.018
Global ASCE	0.161 ± 0.014	0.140 ± 0.012	-0.195 ± 0.010	0.681 ± 0.012	0.368 ± 0.009
FAVoR w/o ASCE	0.377 ± 0.036	0.331 ± 0.033	**-0.036 ± 0.008**	0.816 ± 0.006	0.247 ± 0.010
FAVoR	**0.403 ± 0.004**	0.365 ± 0.009	-0.039 ± 0.022	0.824 ± 0.006	**0.245 ± 0.007**

Table 18: Expanded BlogText author-style results under the author-disjoint ASCE encoder. This table expands the BlogText author-style rows in Table 2 with pooled and diagnostic references. Accuracy and F1 use source-train prototypes, while Gen-Gen AUC/EER measure same-author versus different-author verification among generated continuations. Values are mean ± standard deviation over seeds.

Method	Sep. Acc.↑	Macro-F1↑	Centroid↑	Silhouette↑	AUC↑	EER↓	BERTScore↑	Sem. sim.↑
FedAvg	0.134 ± 0.023	0.115 ± 0.023	0.455 ± 0.083	-0.193 ± 0.024	0.658 ± 0.024	0.386 ± 0.026	0.823 ± 0.000	0.482 ± 0.009
FedProx	0.141 ± 0.024	0.113 ± 0.017	0.463 ± 0.096	-0.172 ± 0.016	0.663 ± 0.029	0.377 ± 0.026	0.823 ± 0.000	0.478 ± 0.008
pFedMe	0.324 ± 0.010	0.290 ± 0.015	0.654 ± 0.040	-0.054 ± 0.008	0.762 ± 0.005	0.305 ± 0.005	0.823 ± 0.000	0.477 ± 0.002
Ditto	0.392 ± 0.022	0.359 ± 0.029	0.685 ± 0.030	-0.018 ± 0.007	0.818 ± 0.002	0.253 ± 0.002	0.825 ± 0.000	0.487 ± 0.006
FedDPA†	0.357 ± 0.035	0.317 ± 0.033	0.667 ± 0.048	-0.053 ± 0.021	0.803 ± 0.013	0.268 ± 0.010	0.825 ± 0.000	0.488 ± 0.005
FAVoR	0.561 ± 0.024	0.521 ± 0.029	0.771 ± 0.021	0.093 ± 0.017	0.900 ± 0.012	0.175 ± 0.011	0.826 ± 0.000	0.494 ± 0.005

Table 19: Mythos-Reddit external validation.

Quantity	Value
Base model	Qwen2.5-3B-Instruct
Quantization	4-bit QLoRA, NF4, bfloat16 compute
Trainable PEFT params	29.93M
Serialized shared adapter	114.3 MiB
Rounds	6
Clients per round	12 of 50
Local epochs	1 shared, 2 private
Batching	micro-batch 4, grad. accumulation 8
Learning rate	2×10^{-4} shared/private
Upload per client round	114.3 MiB
Aggregate upload per round	1.34 GiB
Aggregate upload per seed	8.03 GiB
Upload + broadcast per seed	16.07 GiB
Final private residual storage	5.58 GiB for 50 clients
Runtime per seed	4.40–4.60 hours on 4 RTX 4090 GPUs

Table 26: Efficiency and communication cost for the main BlogText FAVoR runs.

B.15 Statistical Testing

Table 27 reports paired evidence for the main BlogText comparisons over matched seed–author–prompt examples. Table 28 extends the same question to ASCE-independent external gen→source verification by recomputing AUC and EER under seed–author cluster bootstrap resampling. The paired evidence supports the main interpretation: FAVoR's author-style improvements over Ditto, pooled PEFT, FedProx, and FedDPA are robust, while semantic utility differences are smaller and not uniformly in FAVoR's favor. Because each comparison is restricted to the method-pair intersection of quality-passed examples, the FAVoR and baseline columns are recomputed within each comparison block and need not exactly match the corresponding full-coverage summary-table means. In these paired-test tables, directional Δ is signed so positive values favour FAVoR; for assistant score and EER, this means lower raw values for FAVoR. For Table 27, binary author-accuracy p uses McNemar's exact test, while continuous BlogText metrics use the same seed–author cluster-bootstrap distribution for confidence intervals and two-sided tail-probability p values. For Table 28, AUC/EER confidence intervals and p values likewise both come from the same two-sided cluster-bootstrap tail probability.

Method	Sep. Acc.↑	Macro-F1↑	Centroid↑	Silhouette↑	Phrase↓	BERTScore↑	Sem. sim.↑	Gap-Silh.↑	Gap-Cent.↑
FedAvg	0.134 ± 0.023	0.115 ± 0.023	0.455 ± 0.083	-0.193 ± 0.024	1.434 ± 0.367	0.823 ± 0.000	0.482 ± 0.009	-0.237 ± 0.020	-0.322 ± 0.067
FedProx	0.141 ± 0.024	0.113 ± 0.017	0.463 ± 0.096	-0.172 ± 0.016	1.857 ± 0.761	0.823 ± 0.000	0.478 ± 0.008	-0.216 ± 0.011	-0.313 ± 0.081
pFedMe	0.324 ± 0.010	0.290 ± 0.015	0.654 ± 0.040	-0.054 ± 0.008	1.848 ± 0.270	0.823 ± 0.000	0.477 ± 0.002	-0.098 ± 0.013	-0.122 ± 0.026
Ditto	0.392 ± 0.022	0.359 ± 0.029	0.685 ± 0.030	-0.018 ± 0.007	1.697 ± 0.466	0.825 ± 0.000	0.487 ± 0.006	-0.062 ± 0.008	-0.091 ± 0.014
FedDPA†	0.357 ± 0.035	0.317 ± 0.033	0.667 ± 0.048	-0.053 ± 0.021	1.511 ± 0.087	0.825 ± 0.000	0.488 ± 0.005	-0.097 ± 0.016	-0.110 ± 0.034
FAVoR	0.561 ± 0.024	0.521 ± 0.029	0.771 ± 0.021	0.093 ± 0.017	1.670 ± 0.304	0.826 ± 0.000	0.494 ± 0.005	0.049 ± 0.022	-0.006 ± 0.005

Table 20: Detailed Mythos-Reddit diagnostics.

Variant	Shared agg.	Personal branch	Private residual	ASCE align.	Author Acc. ↑	Macro-F1 ↑	Asst. Score ↓	BERTScore ↑	StyleDist AUC/EER ↑ / ↓	MiniLM AUC/EER ↑ / ↓	Stylom. AUC/EER ↑ / ↓
Global ASCE	✓	✗	✗	✓	0.161 ± 0.014	0.140 ± 0.012	0.301 ± 0.065	**0.818** ± 0.002	0.728/0.326 ± 0.009/0.003	0.638/0.407 ± 0.010/0.015	0.643/0.392 ± 0.009/0.016
Ditto + ASCE	✓	✓	✗	✓	0.318 ± 0.037	0.286 ± 0.046	0.409 ± 0.040	0.816 ± 0.002	0.746/0.319 ± 0.015/0.011	0.662/0.389 ± 0.005/0.006	0.707/0.345 ± 0.013/0.011
FAVoR w/o ASCE	✓	✓	✓	✗	0.377 ± 0.036	0.331 ± 0.033	0.306 ± 0.018	0.814 ± 0.001	0.767/0.301 ± 0.013/0.016	0.664/0.388 ± 0.003/0.007	**0.717/0.335** ± 0.001/0.005
FAVoR	✓	✓	✓	✓	**0.403** ± 0.004	**0.365** ± 0.009	**0.260** ± 0.004	0.815 ± 0.000	**0.771/0.295** ± 0.006/0.003	**0.675/0.378** ± 0.010/0.006	0.714/0.342 ± 0.011/0.011

Table 22: Full component-ablation summary with seed variability. This is the expanded version of Table 4: Author Acc. and Macro-F1 are generated-text ASCE prototype metrics, while StyleDistance, MiniLM, and stylometry columns report ASCE-independent same-topic gen→source AUC/EER. Values are mean ± sample standard deviation over three seeds.

Comparison	Metric	Left	Right	Directional Δ	Cluster 95% CI	p
FAVoR vs. w/o ASCE	Author acc.	0.403	0.377	+0.026	[-0.012, 0.064]	0.225
FAVoR vs. w/o ASCE	Top-1 cosine	0.642	0.629	+0.013	[0.003, 0.023]	0.011
FAVoR vs. w/o ASCE	Asst. score	0.260	0.306	+0.046	[0.012, 0.081]	0.009
FAVoR vs. w/o ASCE	BERTScore	0.815	0.814	+0.001	[-0.001, 0.002]	0.410
FAVoR vs. w/o ASCE	SemSim	0.380	0.377	+0.003	[-0.006, 0.012]	0.248
FAVoR vs. Global ASCE	Author acc.	0.403	0.161	+0.242	[0.194, 0.291]	$< .001$
FAVoR vs. Global ASCE	Top-1 cosine	0.642	0.678	-0.036	[-0.052, -0.020]	$< .001$
FAVoR vs. Global ASCE	Asst. score	0.260	0.301	+0.041	[-0.004, 0.085]	0.038
FAVoR vs. Global ASCE	BERTScore	0.815	0.818	-0.004	[-0.006, -0.002]	0.004
FAVoR vs. Global ASCE	SemSim	0.380	0.364	+0.017	[0.006, 0.027]	0.001
w/o ASCE vs. Global ASCE	Author acc.	0.377	0.161	+0.216	[0.170, 0.263]	$< .001$
w/o ASCE vs. Global ASCE	Top-1 cosine	0.629	0.678	-0.049	[-0.067, -0.031]	$< .001$
w/o ASCE vs. Global ASCE	Asst. score	0.306	0.301	-0.005	[-0.055, 0.043]	0.874
w/o ASCE vs. Global ASCE	BERTScore	0.814	0.818	-0.004	[-0.006, -0.002]	0.002
w/o ASCE vs. Global ASCE	SemSim	0.377	0.364	+0.013	[0.002, 0.024]	0.016

Table 23: Paired component-ablation significance tests.

Comparison	Evaluator	Metric	Left	Right	Directional Δ	Cluster 95% CI	p
FAVoR vs. w/o ASCE	StyleDistance	AUC	0.771	0.767	+0.004	[-0.007, 0.015]	0.498
FAVoR vs. w/o ASCE	StyleDistance	EER	0.295	0.301	+0.006	[-0.007, 0.019]	0.438
FAVoR vs. w/o ASCE	MiniLM	AUC	0.675	0.664	+0.011	[-0.002, 0.025]	0.102
FAVoR vs. w/o ASCE	MiniLM	EER	0.378	0.388	+0.010	[-0.003, 0.025]	0.129
FAVoR vs. w/o ASCE	Stylometry	AUC	0.714	0.717	-0.003	[-0.013, 0.006]	0.492
FAVoR vs. w/o ASCE	Stylometry	EER	0.342	0.334	-0.007	[-0.019, 0.005]	0.233
FAVoR vs. Global ASCE	StyleDistance	AUC	0.771	0.728	+0.043	[0.026, 0.061]	$< .001$
FAVoR vs. Global ASCE	StyleDistance	EER	0.295	0.327	+0.032	[0.014, 0.048]	$< .001$
FAVoR vs. Global ASCE	MiniLM	AUC	0.675	0.638	+0.037	[0.021, 0.053]	$< .001$
FAVoR vs. Global ASCE	MiniLM	EER	0.378	0.408	+0.030	[0.015, 0.046]	$< .001$
FAVoR vs. Global ASCE	Stylometry	AUC	0.714	0.643	+0.071	[0.049, 0.094]	$< .001$
FAVoR vs. Global ASCE	Stylometry	EER	0.342	0.392	+0.050	[0.031, 0.072]	$< .001$
w/o ASCE vs. Global ASCE	StyleDistance	AUC	0.767	0.728	+0.039	[0.023, 0.056]	$< .001$
w/o ASCE vs. Global ASCE	StyleDistance	EER	0.301	0.327	+0.025	[0.009, 0.044]	0.003
w/o ASCE vs. Global ASCE	MiniLM	AUC	0.664	0.638	+0.026	[0.009, 0.043]	0.004
w/o ASCE vs. Global ASCE	MiniLM	EER	0.388	0.408	+0.020	[0.004, 0.035]	0.011
w/o ASCE vs. Global ASCE	Stylometry	AUC	0.717	0.643	+0.074	[0.054, 0.096]	$< .001$
w/o ASCE vs. Global ASCE	Stylometry	EER	0.334	0.392	+0.057	[0.040, 0.077]	$< .001$

Table 24: External-verifier paired tests for component ablations using ASCE-independent same-topic gen→source verification. Directional Δ is signed so positive values favor the left variant; for EER, positive directional Δ means lower raw EER for the left variant.

Comparison	Metric	FAVoR	Baseline	Directional Δ	Cluster 95% CI	p	n
FAVoR vs. Ditto	Author acc.	0.403	0.279	+0.125	[0.078, 0.170]	$< .001$	578
FAVoR vs. Ditto	Top-1 cosine	0.642	0.655	-0.013	[-0.026, -0.001]	0.037	578
FAVoR vs. Ditto	SemSim	0.380	0.366	+0.014	[0.004, 0.024]	0.004	578
FAVoR vs. Ditto	BERTScore	0.815	0.815	-0.001	[-0.002, 0.001]	0.405	578
FAVoR vs. Ditto	Asst. score	0.260	0.328	+0.068	[0.023, 0.114]	0.002	578
FAVoR vs. Pooled	Author acc.	0.408	0.255	+0.153	[0.110, 0.198]	$< .001$	569
FAVoR vs. Pooled	Top-1 cosine	0.643	0.590	+0.052	[0.036, 0.069]	$< .001$	569
FAVoR vs. Pooled	SemSim	0.383	0.359	+0.024	[0.013, 0.036]	$< .001$	569
FAVoR vs. Pooled	BERTScore	0.815	0.802	+0.013	[0.010, 0.015]	$< .001$	569
FAVoR vs. Pooled	Asst. score	0.261	0.219	-0.041	[-0.089, 0.006]	0.083	569
FAVoR vs. FedProx	Author acc.	0.403	0.175	+0.228	[0.182, 0.276]	$< .001$	578
FAVoR vs. FedProx	Top-1 cosine	0.642	0.671	-0.029	[-0.046, -0.013]	$< .001$	578
FAVoR vs. FedProx	SemSim	0.380	0.353	+0.027	[0.016, 0.038]	$< .001$	578
FAVoR vs. FedProx	BERTScore	0.815	0.817	-0.002	[-0.004, -0.000]	0.024	578
FAVoR vs. FedProx	Asst. score	0.260	0.264	+0.004	[-0.041, 0.050]	0.857	578
FAVoR vs. FedDPA	Author acc.	0.403	0.294	+0.109	[0.064, 0.155]	$< .001$	578
FAVoR vs. FedDPA	Top-1 cosine	0.642	0.684	-0.042	[-0.057, -0.028]	$< .001$	578
FAVoR vs. FedDPA	SemSim	0.380	0.376	+0.004	[-0.006, 0.014]	0.400	578
FAVoR vs. FedDPA	BERTScore	0.815	0.818	-0.004	[-0.006, -0.002]	$< .001$	578
FAVoR vs. FedDPA	Asst. score	0.260	0.419	+0.159	[0.108, 0.211]	$< .001$	578

Table 27: Paired tests over matched seed–author–prompt examples. FAVoR and baseline values are recomputed on each method-pair matched subset, so the same method can have slightly different displayed values across comparison blocks and need not exactly equal its full-coverage summary-table mean. Directional Δ is signed, so positive values favour FAVoR; for the assistant score, lower raw values are better. Confidence intervals are cluster-bootstrap intervals over seed–author clusters. For author accuracy, p uses McNemar's exact test; for continuous metrics, p uses the two-sided cluster-bootstrap tail probability.

Comparison	Evaluator	Metric	FAVoR	Baseline	Directional Δ	Cluster 95% CI	p
FAVoR vs. Ditto	StyleDistance	AUC	0.771	0.748	+0.023	[0.009, 0.038]	0.001
FAVoR vs. Ditto	StyleDistance	EER	0.295	0.319	+0.024	[0.008, 0.037]	0.002
FAVoR vs. Ditto	MiniLM	AUC	0.675	0.650	+0.025	[0.009, 0.042]	0.004
FAVoR vs. Ditto	MiniLM	EER	0.378	0.402	+0.024	[0.005, 0.040]	0.012
FAVoR vs. Ditto	Stylometry	AUC	0.714	0.683	+0.031	[0.009, 0.055]	0.004
FAVoR vs. Ditto	Stylometry	EER	0.342	0.362	+0.021	[0.002, 0.043]	0.026
FAVoR vs. Pooled	StyleDistance	AUC	0.771	0.736	+0.035	[0.020, 0.050]	$< .001$
FAVoR vs. Pooled	StyleDistance	EER	0.295	0.327	+0.032	[0.016, 0.049]	$< .001$
FAVoR vs. Pooled	MiniLM	AUC	0.675	0.649	+0.026	[0.007, 0.045]	0.005
FAVoR vs. Pooled	MiniLM	EER	0.378	0.406	+0.028	[0.010, 0.046]	0.001
FAVoR vs. Pooled	Stylometry	AUC	0.714	0.660	+0.054	[0.034, 0.076]	$< .001$
FAVoR vs. Pooled	Stylometry	EER	0.342	0.384	+0.043	[0.024, 0.064]	$< .001$
FAVoR vs. FedProx	StyleDistance	AUC	0.771	0.723	+0.047	[0.031, 0.065]	$< .001$
FAVoR vs. FedProx	StyleDistance	EER	0.295	0.334	+0.039	[0.020, 0.055]	$< .001$
FAVoR vs. FedProx	MiniLM	AUC	0.675	0.631	+0.045	[0.028, 0.061]	$< .001$
FAVoR vs. FedProx	MiniLM	EER	0.378	0.414	+0.036	[0.021, 0.052]	$< .001$
FAVoR vs. FedProx	Stylometry	AUC	0.714	0.650	+0.063	[0.042, 0.086]	$< .001$
FAVoR vs. FedProx	Stylometry	EER	0.342	0.385	+0.043	[0.025, 0.066]	$< .001$
FAVoR vs. FedDPA	StyleDistance	AUC	0.771	0.748	+0.023	[0.009, 0.037]	$< .001$
FAVoR vs. FedDPA	StyleDistance	EER	0.295	0.316	+0.021	[0.006, 0.036]	0.004
FAVoR vs. FedDPA	MiniLM	AUC	0.675	0.659	+0.016	[0.001, 0.031]	0.035
FAVoR vs. FedDPA	MiniLM	EER	0.378	0.391	+0.013	[0.000, 0.029]	0.049
FAVoR vs. FedDPA	Stylometry	AUC	0.714	0.692	+0.022	[0.005, 0.038]	0.008
FAVoR vs. FedDPA	Stylometry	EER	0.342	0.351	+0.009	[-0.005, 0.026]	0.201

Table 28: External-verifier paired tests for main BlogText comparisons.