

Privacy-Preserving Text Sanitization for Distributed Agents Collaboration via Disentangled Representations

Xuan Liu^{1,2,*} Hefeng Zhou^{1,2,*} Sicheng Chen¹ Chao Yang¹ Xingcheng Xu¹
Jingjing Qu^{1,†} Jiong Lou^{2,4} Jie Li^{2,3,4} Xia Hu¹

¹Shanghai AI Laboratory ²Shanghai Jiao Tong University

³Yancheng Blockchain Research Institute ⁴State Key Laboratory of Submarine Geoscience

{liuxuan, zhouhefeng, chensicheng, yangchao, xuxingcheng, qujingjing, huxia}@pjlab.org.cn

{lj1994, lijiecs}@sjtu.edu.cn

Abstract

Distributed agents need to exchange textual evidence across organizational boundaries, but these exchanges can expose two types of private information. One is carried by explicit identifiers. The other appears as source signatures spread across formatting, vocabulary, and syntax. Existing PII detection and masking methods mainly handle localized identifiers, leaving these broader source signals largely intact. Even masking 19.2% of tokens reduces TF-IDF stylometric attribution by only 18.6%. To address this problem, we propose *DiSan*¹, which treats text sanitization as a representation learning problem. A two-stream encoder maps text into orthogonal role and style subspaces. The role subspace captures task semantics shared across sources, while the local style subspace captures variation associated with source identity. The style representation is never transmitted directly. Prototype alignment supports joint training without centralizing raw text, while adversarial regularization limits source leakage in shared representations and prototypes. We test leakage in sanitized text, learned representations, and uploaded prototypes. On a distributed multi-agent RAG benchmark, *DiSan* reduces PII exposure in answers by 20× while preserving 83% answer faithfulness. On Enron emails, it reduces TF-IDF stylometric attribution by 73.2%, compared with 18.6% for the strongest masking baseline. Embedding and prototype attribution are also close to random. Together, these results show that *DiSan* reduces explicit and distributional leakage while retaining task utility.

1 Introduction

Cross-organizational collaboration on text-intensive tasks, including retrieval-augmented generation (Lewis et al., 2020), distributed question answering, and privacy-sensitive clinical text

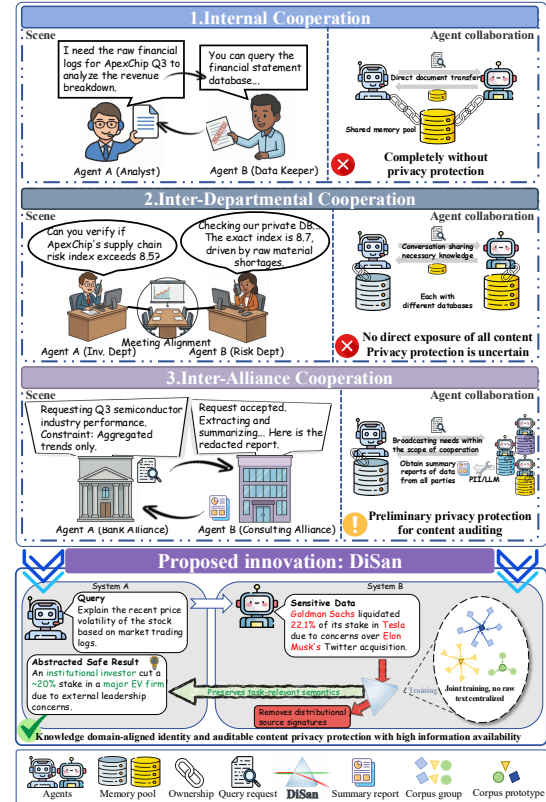


Figure 1: Privacy risks in cross-organizational text sharing. **Top:** Three representative collaboration settings, ranging from no protection to preliminary filtering, each leaving source-identifying content exposed. **Bottom:** *DiSan* produces sanitized text that preserves task-relevant semantics while removing both explicit PII and distributional source signatures.

processing (Stubbs et al., 2015), requires parties to share textual evidence while keeping raw data local. Each party may hold proprietary document collections with distinct domain expertise, and a common pattern is for a requesting party to seek auxiliary evidence from helpers, who retrieve local snippets and transmit them for downstream use such as answer synthesis (Minaee et al., 2024).

However, LLM-based text interactions can expose private information beyond what users in-

*Equal contribution. †Corresponding author.

¹Code: github.com/A3S-Lab/DiSan.

Table 1: Compact CorporateBank→AssetManager financial-risk example. **Bold** marks source-identifying patterns that survive identifier masking.

Raw private text (d). “*Per Meridian Bank’s Counterparty Risk Bulletin (Ref: CR-2024-047), Apex Dynamics carried \$6.1M exposure as of Q3 close, was downgraded to BB+, and was flagged for portfolio review.*”

Placeholder masking only. “*Per Meridian Bank’s Counterparty Risk Bulletin (Ref: [ID]), [ORG] carried \$6.1M exposure as of Q3 close, was downgraded to BB+, and was flagged for portfolio review.*”

DiSAN output ($\tilde{d}$). “*A corporate-bank Q3 risk bulletin flags an industrials counterparty with \$6.1M exposure, a BB+ downgrade, and mandatory portfolio review.*”

tend to disclose (Chu et al., 2024). As illustrated in Figure 1, this risk spans all levels of cross-organizational collaboration: from intra-organization transfers with no protection to inter-alliance sharing where only preliminary filtering is applied. In each setting, shared text leaks private organizational information at two levels: explicitly through identifiers such as names, account numbers, and addresses, and implicitly through *distributional signatures*: formatting conventions, vocabulary choices, and syntactic patterns that may reveal source identity (Stamatatos, 2009). This is fundamentally a representation-level problem, not an identifier-level one: private organizational information is a property of the text distribution, not of individual identifiers, and anonymization methods that operate in text space cannot alter these distributional properties. Table 1 makes this concrete: raw sharing exposes not only a counterparty identity but also a proprietary bulletin format, reference scheme, and sector taxonomy; placeholder masking hides surface identifiers but leaves the naming convention intact and collapses distinct entities into generic tokens, weakening grounding, provenance tracking, and cross-document aggregation. DiSAN instead preserves role facts such as exposure amount, sector, rating change, and review status while suppressing source-specific fingerprints.

Existing sanitization methods are effective for explicit identifiers, but they provide limited control over source-identifying patterns distributed across the text. Rule-based PII detectors (Lison et al., 2021) target individual identifiers such as named entities and account numbers but are blind to distributional signatures, since private organizational information is distributed across the text as statistical patterns, not localized to individual spans.

LLM-based privacy protection (Xiao et al., 2024) can reduce privacy leakage, but does not explicitly target source-invariant text distributions. Federated learning (McMahan et al., 2017) decentralizes model training but produces shared *predictors*, not shareable *data*. The core challenge is structural: what is the minimal sufficient representation that preserves task semantics while being statistically source-invariant?

We propose DiSAN, a sanitization framework for cross-agent evidence exchange. It learns role and style representations for each evidence snippet. Expert-written targets supervise the preservation of task facts and removal of protected content, while orthogonality encourages source-linked variation to move away from the role representation. Prototype alignment keeps role representations comparable across non-IID agents without centralizing raw text. The resulting sanitizer keeps the style representation local and does not transmit it directly. We empirically evaluate whether source identity remains detectable in the shared output text.

Contributions. (i) **Role–style sanitization for text sharing:** We formulate cross-agent text sanitization as a supervised role–style separation problem and evaluate the resulting separation empirically. (ii) **Federated role alignment:** We introduce lightweight prototype alignment to stabilize role spaces across non-IID agents without centralizing raw text. (iii) **Privacy diagnostics across sharing surfaces:** We evaluate privacy at the output, representation, and prototype levels, distinguishing application-stage leakage from training-stage artifacts. (iv) **Empirical validation:** On distributed-agent RAG, DiSAN reduces answer-level PII exposure by 20× while maintaining 83% answer faithfulness. On Enron emails, it reduces TF-IDF stylistic attribution by 73.2%, substantially outperforming identifier-level masking.

2 Related Work

Privacy-Preserving Machine Learning. Protecting privacy across distributed data sources is a persistent challenge in collaborative machine learning (Li et al., 2021a). Differential privacy (DP) (Dwork et al., 2006) provides formal guarantees, and DP-SGD (Oquendo and Abdelhadi, 2022) extends these to deep learning (Feldman et al., 2020; Abadi et al., 2016; Canonne et al., 2020). Federated learning (McMahan et al., 2017) enables collaborative model training without sharing raw

data (Li et al., 2020a; Liu et al., 2022). Within this paradigm, prototype-based methods (Tan et al., 2021), representation-based approaches (Li et al., 2021b; Wu et al., 2021), and communication-efficient techniques (Zhang et al., 2022; Wu et al., 2024) have been proposed. Recent federated RAG formulations aim to enable multi-party retrieval under privacy constraints (Qian et al., 2025; He et al., 2025; Chakraborty et al., 2025). While these methods focus on training shared models or retrievers, our work addresses a complementary problem: *sanitizing text data itself* so it can be safely shared for downstream use.

Text Sanitization and De-identification. Traditional text de-identification relies on rule-based or NER-based PII detection followed by masking or replacement (Malik and Dustdar, 2011). While effective for explicit identifiers, these methods miss implicit leakage through writing style, document structure, and domain-specific patterns. Authorship attribution research (Stamatatos, 2009) demonstrates that stylometric features can identify sources even from short texts. Recent work explores privacy-enhancing methods for federated multi-agent systems (Shi et al., 2025) and DP-based text generation (Xie et al., 2024). However, differentially private text generation often involves a privacy–utility trade-off, where stronger privacy can degrade semantic similarity and text quality (Meisenbacher and Matthes, 2024). Our approach uses supervised role-style separation to address explicit PII and stylistic fingerprints, and evaluates the remaining leakage empirically rather than claiming formal privacy.

Disentangled Representations. Disentangled representation learning aims to separate independent factors of variation in data (Bengio et al., 2013). In NLP, disentanglement has been applied to separate content from style for style transfer (John et al., 2019), sentiment from semantics, and speaker identity from linguistic content in speech (Qian et al., 2019). Recent work enables explicit control in text generation via disentangled representations (Liu et al., 2024; Han et al., 2024). These works establish that separating semantic content from stylistic or identity-related factors supports both controllable generation and privacy; we apply the same principle to text sharing, where the goal is to isolate task-relevant content from source-identifying patterns.

3 Problem Statement

3.1 Text Sharing for Distributed Agent Collaboration

We consider C distributed agents, each hosting a private document repository $\mathcal{D}_c$. When a requesting agent cannot answer a query q locally, it routes to a candidate helper set $\mathcal{C}(q)$ via embedding similarity over public **capability tags**, such as “AssetManager” and “CorporateBank”. Each helper retrieves a local snippet d and returns a *sanitized* snippet $\tilde{d}$ for downstream use in RAG or distributed question answering.

Table 1 concretizes the sanitization goal: raw text exposes explicit PII and institutional style fingerprints; placeholder masking removes identifiers but leaves distributional signatures intact and can weaken downstream grounding when masked spans are task-relevant; only a representation-level approach targets both privacy risks while preserving task semantics. The example also states the disclosure boundary set by the data owner. Names, institutional references, and source-specific wording are private, whereas the exposure amount, sector, rating change, and review status are authorized task facts that the sanitizer should retain.

3.2 Threat Scope

Our primary privacy concern arises in the *application stage*. A requesting agent follows the collaboration protocol but may analyze sanitized evidence $\tilde{d}$ to infer information beyond the helper’s public capability tag, including explicit PII, source-linked writing patterns, or organizational document fingerprints. Helper agents are trusted to run the local sanitizer and keep raw documents and local representations private. Training creates a separate artifact-leakage surface. The honest-but-curious coordinator follows the federated protocol and receives trainable-model updates, sample counts, and perturbed normalized role prototypes. It does not receive raw text or token-level embeddings. We evaluate whether the observed training artifacts retain persistent source-linked signatures beyond public participation. We do not claim formal differential privacy or robustness to malicious servers, poisoning, prompt injection, or collusion; the complete threat model is in Appendix A.

Goals and validation. We do not hide public capability tags or agent participation; we aim to prevent leakage beyond these facts while preserv-

ing downstream RAG utility. Our validation follows this scope: PII leakage and answer exposure measure explicit identifiers, stylometry measures source signatures in transmitted text, embedding and prototype attribution diagnose learned sharing artifacts, and F1, faithfulness, and ChunkHit@3 measure utility.

4 Method

Overview. DISAN uses a two-stream encoder that projects each input into a **role** subspace for task semantics and a **style** subspace for source-linked variation. Expert-written sanitization targets supervise the decoder, while orthogonality and prototype regularization encourage separation between the two streams. This separation is learned empirically and is not guaranteed. Role and style representations are fused for local decoding, and the style representation is discarded after use. At deployment, only the sanitized text $\tilde{d}$ leaves the helper.

Training without centralizing raw text poses a calibration challenge: per-agent isolation causes role spaces to drift across agents, degrading both utility and privacy. DISAN addresses this by exchanging compact role prototypes, aligning local role distributions to shared global anchors, and applying adversarial regularization to suppress source-specific prototype signatures beyond public agent tags. Figure 2 illustrates the architecture.

4.1 Role-Style Disentangled Encoder

Two-stream projection. Given an evidence sequence $d = (d_1, \dots, d_T)$, a pretrained encoder produces hidden states $\mathbf{H} = \text{Encoder}(d) \in \mathbb{R}^{T \times d_{\text{enc}}}$. We project $\mathbf{H}$ into a *role* stream and a *style* stream (backbone and projection dimensions in Appendix B.3):

$$\mathbf{Z}_r = \mathbf{H}\mathbf{W}_r, \quad \mathbf{W}_r \in \mathbb{R}^{d_{\text{enc}} \times d_r}, \quad \mathbf{Z}_r \in \mathbb{R}^{T \times d_r}, \quad (1)$$

$$\mathbf{Z}_s = \mathbf{H}\mathbf{W}_s, \quad \mathbf{W}_s \in \mathbb{R}^{d_{\text{enc}} \times d_s}, \quad \mathbf{Z}_s \in \mathbb{R}^{T \times d_s}. \quad (2)$$

In our implementation, the encoder-decoder backbone is LongT5-TGlobal-Base (Guo et al., 2022). Both $\mathbf{Z}_r$ and $\mathbf{Z}_s$ are 256-dimensional token streams; their concatenation forms a 512-dimensional bottleneck that is projected back to d_{enc} before decoding. Intuitively, $\mathbf{Z}_r$ captures shareable content structure (enti-

ties/relations/events), while $\mathbf{Z}_s$ captures agent-specific phrasing and formatting.

Fusion for generation. We fuse the two streams before decoding via $\mathbf{H}_{\text{fused}} = g([\mathbf{Z}_r; \mathbf{Z}_s])$, where g is a learned projection back to d_{enc} . A full residual path $\mathbf{H}_{\text{out}} = \alpha \cdot \mathbf{H}_{\text{fused}} + (1 - \alpha) \cdot \mathbf{H}$ stabilizes training (details in Appendix B.1). The pretrained LongT5 Transformer decoder then generates $\tilde{d}$ autoregressively from the modified encoder states; DISAN inserts a role-style bottleneck rather than replacing the generator.

Although $\mathbf{Z}_s$ may contain variation associated with its source, it is used only inside the helper during decoding and is never transmitted directly. The sequence loss uses expert written targets that remove or abstract protected content while retaining authorized task facts. Together with the regularizers below, this supervision encourages the role stream to retain task content and the local style stream to absorb variation associated with the source. Because $\mathbf{Z}_s$ participates in decoding, we evaluate any residual influence of source identity in the transmitted $\tilde{d}$ using PII detection and stylometric metrics.

Disentanglement. Let $\bar{\mathbf{z}}_r = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_{r,t}$ and $\bar{\mathbf{z}}_s = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_{s,t}$ denote mean-pooled vectors. We encourage separation via:

$$\mathcal{L}_{\text{orth}} = \cos^2(\bar{\mathbf{z}}_r, \bar{\mathbf{z}}_s) = \left(\frac{\bar{\mathbf{z}}_r^\top \bar{\mathbf{z}}_s}{\|\bar{\mathbf{z}}_r\|_2 \|\bar{\mathbf{z}}_s\|_2 + \epsilon} \right)^2. \quad (3)$$

4.2 Prototype Alignment on the Role Space

To train the privacy transformer under non-IID agents, we align *role distributions* across agents using lightweight role prototypes, providing global anchors without sharing raw evidence.

Prototype computation and aggregation. Each agent computes one local prototype $\mu^{(c)}$ as an EMA of batch role centroids over valid token positions. At the end of each round, the server forms a sample-weighted global prototype μ^* . Full definitions are provided in Appendix B.2.

Spherical alignment. To avoid magnitude-based leakage, we align on the unit hypersphere:

$$\mathcal{L}_{\text{proto}} = 1 - \cos(\hat{\bar{\mathbf{z}}}_r, \hat{\mu}^*), \quad (4)$$

where $\hat{\cdot}$ denotes ℓ_2 -normalization and $\bar{\mathbf{z}}_r$ is the batch role centroid. The implemented spherical

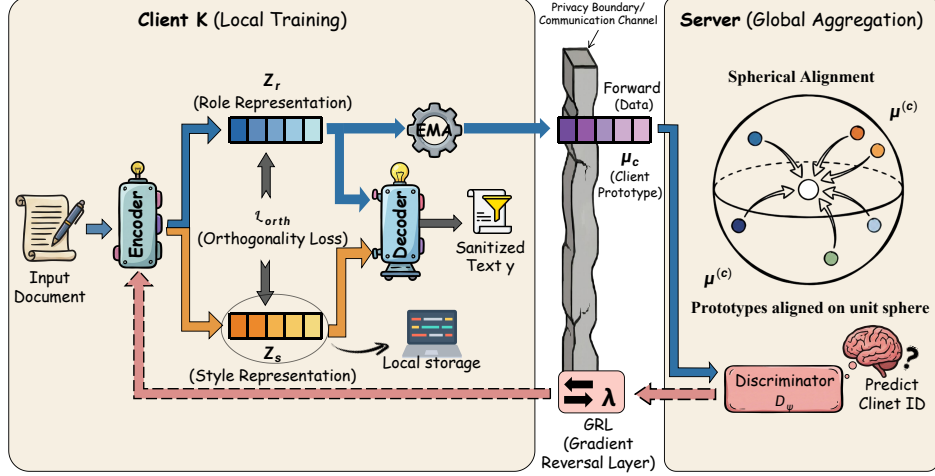


Figure 2: DISAN architecture. **Left (agent):** A two-stream encoder produces role representations $\mathbf{Z}_r$ for task semantics and style representations $\mathbf{Z}_s$ for source-linked variation. Expert targets and $\mathcal{L}_{\text{orth}}$ encourage their separation. Both streams are fused for local decoding, and $\mathbf{Z}_s$ is never transmitted. During training, each agent uploads one role prototype $\mu^{(c)}$. **Right (server):** The server aligns role prototypes and trains the prototype discriminator used for gradient reversal on clients.

loss also penalizes within-batch dispersion (Appendix B.2).

Prototype-level adversarial training. Prototypes may still carry source-specific distributional signatures beyond public agent tags. We apply a discriminator D_ψ with gradient reversal (GRL) (Raff and Sylvester, 2018) to make prototypes less predictive of their source:

$$\mathcal{L}_{\text{adv}} = \text{CE}\left(D_\psi(\text{GRL}_\gamma(\bar{\mu}^{(c)})), c\right), \quad (5)$$

where γ is the GRL strength and $\bar{\mu}^{(c)}$ is a gradient-enabled estimate (Appendix B.2). The server-side prototype discriminator is the only adversary enabled during training. The server trains it on uploaded prototypes and sends its parameters to clients in the next round. Each client freezes the discriminator, so gradient reversal updates the role projection and its upstream LoRA (Hu et al., 2021) path rather than the discriminator. Per-instance role and fused classifiers are disabled. The prototype discriminator is an MLP with layers $256 \rightarrow 128 \rightarrow 128 \rightarrow C$, ReLU activations, and dropout 0.1.

4.3 Federated Optimization

Local objective. On agent c , we minimize:

$$\begin{aligned} \mathcal{L}_{\text{local}}^{(c)} = & \mathcal{L}_{\text{seq}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}} + \lambda_p \mathcal{L}_{\text{proto}} \\ & + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{prox}}, \end{aligned} \quad (6)$$

where $\mathcal{L}_{\text{seq}}$ is the token-level cross-entropy on sanitization targets $\tilde{d}$, and $\mathcal{L}_{\text{prox}} = \frac{\nu}{2} \|\theta - \theta^*\|_2^2$ is a

FedProx term discouraging drift from the global model (we use ν to distinguish from prototype symbols μ). After local optimization, each agent uploads its trainable-model update, sample count, and EMA role prototype. The prototype is normalized, perturbed with Gaussian noise ($\sigma = 0.01$), and normalized again before upload. Raw text and token-level embeddings remain local. The shared model parameters are LoRA adapters on encoder attention Q/V projections, the role projection, the fusion layer, and residual components. The style projection is trained locally but is excluded from aggregation. Each agent uploads 256 floating-point values for its role prototype per round.

4.4 Deployment: Multi-Agent Text Sharing

The preceding sections describe *how to train* the sanitizer; this section describes *how agents use it* at deployment. Each agent deploys the trained model as a local sanitizer: given raw text d , it produces $\tilde{d}$ via role–style fusion and decoding. Crucially, $\mathbf{Z}_s$ is used only locally to improve generation quality and is **discarded after decoding**. Only the sanitized text $\tilde{d}$ is transmitted. This design limits the deployment interface to sanitized text. The privacy claims are empirical and apply to the threat model and attacks evaluated in Section 3.2 and Appendix C.1.

RAG pipeline. A requesting agent routes a query q to helper agents $\mathcal{C}$ via capability-based routing (Section 3.1). Each helper $c \in \mathcal{C}$ retrieves local

Table 2: Main single-round results. **Bold** = best, underline = second best. Metrics are defined in Section 5.3.

Method	Semantic Similarity $\uparrow$				Faith. $\uparrow$	ChunkHit@3 $\uparrow$	PII Exposure $\downarrow$	
	F1	Prec.	Rec.	Cos.			Avg. PII	Ans. Rate
RAG Baseline without PII process	0.6602	0.5519	0.8214	0.6421	86.10%	75.2%	6.4753	11.8%
Placeholder (gliner-pii-large-v1.0)	0.5071	0.4624	0.5613	0.5923	74.60%	62.4%	<u>0.4714</u>	<u>1.4%</u>
Placeholder (piirinha-v1-detect)	0.4894	0.4352	0.5587	0.5643	71.82%	58.6%	0.5126	1.6%
Placeholder (deberta-pii-finetuned)	0.4768	0.4541	0.5018	0.5819	72.63%	60.2%	0.6287	2.0%
LLM paraphrasing (llama-3.1-8b-instruct)	0.4991	0.4427	0.5719	0.5872	77.89%	66.4%	1.2479	3.6%
LLM paraphrasing (Qwen2.5-7B-Instruct)	<u>0.5407</u>	<u>0.5011</u>	0.5872	0.6218	<u>79.35%</u>	<u>67.0%</u>	0.8412	2.6%
LLM paraphrasing (GLM-4-9B-0414)	0.4732	0.4315	0.5237	0.6072	77.13%	64.6%	2.1091	3.8%
Policy gating	0.5013	0.3876	0.7936	<u>0.6322</u>	78.49%	65.2%	0.9714	2.2%
DtSAN (ours)	0.5631	0.5341	<u>0.6508</u>	0.6558	83.17%	73.4%	0.1337	0.6%

evidence d_c and returns:

$$\tilde{d}_c = \text{Sanitize}_\theta(d_c, q), \quad (7)$$

where Sanitize_θ denotes the trained sanitizer (Section 4.1–4.3). The requester aggregates $\tilde{D} = \{\tilde{d}_c \mid c \in \mathcal{C}\}$ and generates a final answer. This protocol naturally extends to multi-turn settings where the requester iteratively refines queries based on accumulated evidence (see Section 5.6 for details). Convergence analysis is provided in Appendix B.4.

5 Experiments

5.1 Experimental Setup

Dataset and agent configuration. We use `synthetic_pii_finance_multilingual` (Watson et al., 2024), a multilingual synthetic finance corpus with annotated PII spans. To simulate distributed agent collaboration, we construct $\mathcal{C}=7$ agents by assigning each agent a disjoint inventory of document types. The resulting partition induces non-IID skews by design. We denote agents by their capability tags: CorporateBank, AssetManager, FinTechPay, CorpGroup, MarketForecaster, ComplianceConsult, and SupplierCo. These tags are treated as public in our threat model (Section 3.2); we therefore focus on leakage *beyond* this public prior. The exact doc-type identifiers used for each agent are listed in Appendix C.5. Domain experts manually rewrote all 1,787 raw-to-sanitized training pairs. They removed or abstracted protected identifiers and source-linked expressions while preserving facts allowed by the disclosure boundary in Table 1. The targets retain varied professional writing styles rather than imposing one fixed template. This annotation is a one-time training cost; deployment does not require paired examples for each organization or request. The seven financial subdomains produce heteroge-

neous query needs, and Table 2 reports downstream utility under this partition.

RAG evaluation pipeline. Documents are chunked into fixed windows (256 tokens, overlap 50) and indexed per agent. Given a query, the requesting agent routes to a candidate helper set via tag-based routing (Section 3.1), retrieves top- k evidence from each helper’s local index using a BGE-M3 (Chen et al., 2024) hybrid pipeline, receives sanitized snippets $\tilde{d}$, and generates the final answer. Grounded QA examples are synthesized from retrieval anchors and retained only when evidence spans are found in the source chunk; each record stores its `chunk_id` for provenance-based `ChunkHit@3`. Query and ground-truth construction details are in Appendix C.6; retrieval architecture details are in Appendix C.6. We evaluate on single-round sharing as our main setting; the multi-turn deployment extension is discussed in Section 5.6.

Training configuration. Training runs for 12 rounds with 300 local steps per round per agent, batch size 4, and learning rate 2×10^{-4} . We set the proximal regularization strength $\nu=0.1$ to mitigate drift under non-IID data. Unless otherwise stated, we use prototype adversarial weight $\lambda_{\text{adv}}=1.0$, GRL strength $\gamma=0.5$, $\lambda_{\text{orth}}=0.2$, prototype alignment weight $\lambda_p=2.0$, and prototype noise scale $\sigma_{\text{noise}}=0.01$ (convergence analysis in Appendix B.4). The server-side prototype discriminator is the only adversary used during training; per-instance role and fused classifiers have zero loss weight. Evaluation attackers are trained after the sanitizer and do not update it. Full hyperparameters and schedules are deferred to Appendix C.5.

5.2 Baselines

We compare against practical alternatives for privacy-preserving text sharing:

Placeholder-only: This approach first applies a PII detection model to identify sensitive spans (names, dates, addresses, account numbers, etc.), then replaces each detected span with a type-specific placeholder token (e.g., [NAME], [DATE], [ADDRESS], [ACCOUNT]). We evaluate three PII detectors: gliner-pii-large-v1.0 (Zarati et al., 2024), a generalist NER model; piirinha-v1-detect-personal-information, a DeBERTa-based model (He et al., 2020) fine-tuned for PII detection; and deberta-pii-finetuned, another DeBERTa variant trained on PII corpora. While placeholder replacement removes explicit identifiers, it does not address implicit stylistic fingerprints; furthermore, opaque placeholder tokens can degrade downstream RAG by collapsing task-relevant spans into generic tokens and weakening answer grounding, provenance, cross-document aggregation.

LLM paraphrasing: Locally paraphrase text with open-source LLMs (Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen Team et al., 2025), and GLM-4-9B-0414 (GLM Team et al., 2024)) using a privacy-focused prompt, then share the rewritten text.

Policy gating: Adapted from dynamic access-control memory sharing (Rezazadeh et al., 2025), a local policy model (Qwen2.5-7B) decides per chunk whether to share the original text, provide a summary, or refuse sharing, based on the requester’s agent tag and the helper’s data sensitivity level. Prompts and policy templates are in Appendix C.7.

5.3 Evaluation Metrics

We evaluate sanitization quality using privacy and utility metrics.

Privacy metrics. (i) **Avg. PII:** the average number of PII spans detected in sanitized chunks by an external detector (lower is better). (ii) **Ans. Rate:** the fraction of final answers that contain at least one exposed PII entity. We instantiate the external detector as gliner-pii-large-v1.0 with a 0.3 confidence threshold over common PII labels; implementation details are in Appendix C.6. (iii) **Distributional fingerprint leakage:** 7-way attribution accuracy/macro-F1 from learned role embeddings (EXP-1) and sanitized text via stylometry (EXP-3). (iv) **Prototype fingerprint leakage:** 7-way attribution accuracy/macro-F1 from uploaded prototypes (EXP-2). Since capability tags and participation are

public, these probes do not test whether the helper capability tag is hidden. They test whether transmitted text or learned artifacts still carry residual source-correlated fingerprints beyond that public information; full protocols are in Appendix C.1.

Utility metrics. (i) **F1/Prec./Rec./Cos.:** token-level F1, precision, recall, and cosine similarity between generated answers and ground-truth answers (bag-of-words TF representation; see Appendix C.6). (ii) **Faithfulness:** the fraction of stopword-removed content words in the answer that appear in retrieved evidence. (iii) **ChunkHit@3:** the fraction of ground-truth chunks appearing in top-3 retrieved results.

5.4 Main Results

DISAN achieves strong privacy protection with modest utility loss. As shown in Table 2, it reduces answer-level PII exposure from 11.8% under unprotected sharing to just 0.6%, while preserving answer faithfulness at 83.17% close to the 86.10% unprotected baseline.

Figure 3 further shows that DISAN offers the most favorable privacy–utility trade-off among the evaluated sanitizers, being the only method that simultaneously achieves sub-1% answer-level PII leakage and near-baseline task performance.

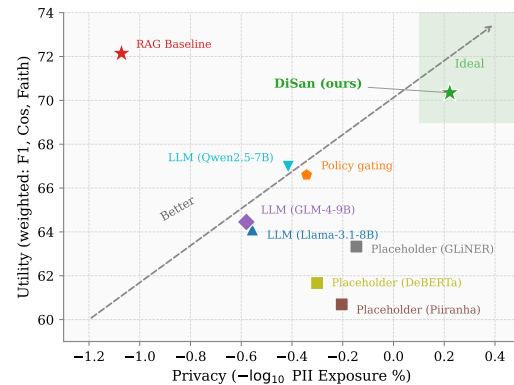


Figure 3: Privacy–utility trade-off. The upper-right region is ideal; Ours achieves the best overall trade-off.

5.5 Ablation Study

We ablate two key components: *style isolation* and *prototype alignment*. Full results are in Appendix C.8.

Style isolation ablations. We test (A1) removing $\mathcal{L}_{\text{orth}}$ ($\lambda_{\text{orth}}=0$), allowing the role representation to retain more source-related cues, and (A2)

Table 3: **Ablation results (single-round).** **Bold** = best column value; underline = second best. These are not competing sanitizers: A1 and A2 relax privacy constraints, so their higher utility metrics reflect under-sanitization rather than improvement. The relevant comparison is privacy cost vs. utility gain.

Method	Sem. F1 $\uparrow$	Cos. $\uparrow$	Faith. $\uparrow$	ChunkHit@3 $\uparrow$	Avg. PII $\downarrow$	Ans. Rate $\downarrow$
DiSAN (ours)	0.5631	<u>0.6558</u>	83.17%	<u>73.4%</u>	0.1337	0.6%
A1: w/o $\mathcal{L}_{\text{orth}}$	<u>0.5694</u>	0.6309	<u>83.46%</u>	72.7%	0.9706	1.4%
A2: high- α fusion	0.5768	0.6664	84.12%	74.4%	1.2419	2.2%
B: No-ProtoAlign	0.5487	0.6163	81.92%	69.8%	<u>0.3588</u>	<u>0.8%</u>

high- α fusion ($\alpha=0.9$), which increases the contribution of the fused role–style representation during decoding. Both variants slightly improve some RAG utility metrics, but at the cost of substantially higher privacy leakage. In particular, Avg. PII increases by $7.3\times$ under A1 and $9.3\times$ under A2, with answer-level exposure and role-attribution showing the same trend. This suggests that part of the utility gain comes from weaker sanitization, where more source-correlated lexical information is retained in the output. Overall, these results illustrate the privacy–utility trade-off controlled by style isolation and motivate the default DiSAN configuration.

Prototype alignment ablation. We ablate prototype alignment by removing EMA prototypes and global prototype aggregation, and setting $\lambda_p=0$ to remove $\mathcal{L}_{\text{proto}}$ from Equation (6). Without prototype anchors, role spaces drift under non-IID training, degrading both utility and privacy: cosine similarity drops 6%, ChunkHit@3 drops 3.6pp, and PII exposure increases $2.7\times$.

Disentanglement verification. We further examine whether the learned representations follow the intended role–style separation. Shared role embeddings carry little source-identifying signal under the 7-way fingerprint probe ($F1=0.05$), whereas local-only style embeddings remain strongly source-correlated ($F1=0.84$). Figure 5 provides complementary geometric evidence: role–style cosine similarities concentrate near zero and t-SNE separates the two embedding sets. These diagnostics suggest that source-correlated variation is mainly captured by the local-only style representation, providing empirical evidence rather than a guarantee of disentanglement. DiSAN does not provide formal differential-privacy guarantees; full attack protocols are described in Appendix C.1.

Table 4 treats Enron as a stylometry diagnostic rather than hidden-tag evaluation; DiSAN lowers TF-IDF/BERT attribution to 0.221/0.203, near ran-

Table 4: Enron stylometry probe (7 authors). Lower F1 indicates weaker residual distributional fingerprints.

Method	Mask%	TF-IDF	BERT	Red.
Raw text	–	0.825	0.691	–
GLiNER masking	19.2%	0.672	0.618	18.6%
DiSAN	–	0.221	0.203	73.2%
Random	–	0.143	0.143	–

dom and far below GLiNER. Appendix C.1 further compares with dedicated authorship-obfuscation baselines, JAMDEC and StyleRemix, which target stylometric leakage rather than RAG utility and yield substantially smaller reductions. We adopt LoRA for efficient training (details in Appendices B.3 and B.4).

5.6 Multi-Party Collaboration Analysis

Multi-turn RAG. The single-round protocol extends naturally to multi-turn settings where the requester refines queries based on accumulated evidence. At each turn t , the requester issues a refined query q_t to a possibly different helper subset; each helper sanitizes new evidence independently before transmitting it. The key property is preserved: only sanitized text crosses the privacy boundary at every turn. We treat this as a deployment extension of the same sanitization interface and leave dedicated multi-turn benchmarking to future work.

Case study: cross-organizational IPO analysis. Figure 6 (see Appendix C.9) presents an IPO analysis scenario where Lumina Capital’s Audit-Core agent needs external benchmarks. Audit-Core queries partner agents through a data broker, and each helper sanitizes its response with DiSAN before sharing it. Capability tags are public, but local repositories and retrieval indices remain private. The requester then combines the sanitized financial evidence to produce the final assessment.

6 Conclusion

Identifier-level anonymization is insufficient when private organizational information is distributed across a text rather than confined to a few spans. DISAN combines expert-written sanitization targets with role-style orthogonality, federated role alignment, and prototype adversarial training. In our distributed-agent RAG and Enron evaluations, the resulting text retains task utility while reducing the source-identifying signals measured by the selected probes. These results are empirical rather than a guarantee of disentanglement or privacy. Future work should test stronger adaptive adversaries, repeated queries, and broader domains.

Limitations

Dataset. Our main experiments use a synthetic finance corpus with annotated PII spans. The use of synthetic data is a necessary constraint rather than a methodological choice: datasets containing real PII cannot legally or ethically be used for research publication, and this is standard practice in privacy-preserving NLP. Among publicly available PII-annotated corpora, most are unsuitable for RAG evaluation due to very short texts (social media, medical notes) or insufficient document-type diversity for multi-party simulation; the selected corpus provides the document length, PII annotation quality, and domain variety required by our setup. RAG evaluation queries are synthesized on top of this validated base rather than generating both PII data and queries from scratch, which would compound synthesis risk across two stages. External validation on the Enron email corpus, with real stylistic variation outside the finance domain, partially addresses domain generalizability.

Formal privacy guarantees. DISAN does not provide formal differential privacy guarantees. Applying DP to sequence-to-sequence generation requires per-token noise calibration incompatible with coherent text generation; we instead rely on empirical validation across three attack surfaces. Formal privacy analysis for generative sanitizers is an important direction for future work.

Acknowledgments

This work was supported by Shanghai Artificial Intelligence Laboratory.

References

- Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. [Deep learning with differential privacy](#). In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16*, pages 308–318, New York, NY, USA. Association for Computing Machinery.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- Clément L. Canonne, Gautam Kamath, and Thomas Steinke. 2020. [The discrete gaussian for differential privacy](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 15676–15688. Curran Associates, Inc.
- Abhijit Chakraborty, Chahana Dahal, and Vivek Gupta. 2025. [Federated retrieval-augmented generation: A systematic mapping study](#). *Preprint*, arXiv:2505.18906.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).
- Junjie Chu, Zeyang Sha, Michael Backes, and Yang Zhang. 2024. Reconstruct your previous conversations! comprehensively investigating privacy leakage risks in conversations with gpt models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6584–6600.
- Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*. Wiley-Interscience.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Vitaly Feldman, Tomer Koren, and Kunal Talwar. 2020. [Private stochastic convex optimization: optimal rates in linear time](#). In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020*, pages 439–449, New York, NY, USA. Association for Computing Machinery.
- Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. 2024a. [StyleRemix: Interpretable authorship obfuscation via distillation and perturbation of style elements](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4172–4206, Miami, Florida, USA. Association for Computational Linguistics.
- Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024b. Jamdec: Un-supervised authorship obfuscation using constrained

- decoding over small language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1552–1581.
- GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, and Hanlin Zhao. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex Vaughan. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736.
- Jingxuan Han, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15187–15201.
- Hangyu He, Xin Yuan, Kai Wu, Ren Ping Liu, and Wei Ni. 2025. pFedRAG: A personalized federated retrieval-augmented generation system with depth-adaptive tiered embedding tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14255–14268, Suzhou, China. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 424–434.
- Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. 2020a. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854.
- Q. Li, Yiqun Diao, Quan Chen, and Bingsheng He. 2021a. Federated learning on non-iid data silos: An experimental study. *2021 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 965–978.
- Qinbin Li, Bingsheng He, and Dawn Song. 2021b. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10713–10722.
- Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020b. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.
- Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.
- Dianqi Liu, Liang Bai, Tianyuan Yu, and Aiming Zhang. 2022. Towards method of horizontal federated learning: A survey. In *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, pages 259–266.
- Yi Liu, Xiangyu Liu, Xiangrong Zhu, and Wei Hu. 2024. Multi-aspect controllable text generation with disentangled counterfactual augmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9231–9253.
- Ahmad Kamran Malik and Schahram Dustdar. 2011. Enhanced sharing and privacy in distributed information sharing environments. *2011 7th International Conference on Information Assurance and Security (IAS)*, pages 286–291.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera 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.

- Stephen Meisenbacher and Florian Matthes. 2024. Just rewrite it again: A post-processing method for enhanced semantic similarity and privacy preservation of differentially private rewritten text. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pages 1–11.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Ahmed El Ouadrhiri and Ahmed Abdelhadi. 2022. [Differential privacy for deep and federated learning: A survey](#). *IEEE Access*, 10:22359–22380.
- Cheng Qian, Hainan Zhang, Yongxin Tong, Hong-Wei Zheng, and Zhiming Zheng. 2025. Hyfedrag: A federated retrieval-augmented generation framework for heterogeneous and privacy-sensitive data. *arXiv preprint arXiv:2509.06444*.
- Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.
- Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, and Fei Huang. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 189–198. IEEE.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. 2025. Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control. *ArXiv preprint arXiv:2505.18279*.
- Zitong Shi, Guancheng Wan, Wenke Huang, Guibin Zhang, Jiawei Shao, Mang Ye, and Carl Yang. 2025. Privacy-enhancing paradigms within federated multi-agent systems. *arXiv preprint arXiv:2503.08175*.
- Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.
- Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.
- Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. 2021. [Fedproto: Federated prototype learning across heterogeneous clients](#). In *AAAI Conference on Artificial Intelligence*.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS 2020.
- Alex Watson, Yev Meyer, Maarten Van Segbroeck, Matthew Grossman, Sami Torbey, Piotr Mlocek, and Johnny Greco. 2024. Synthetic-pii-financial-documents-north-america: A synthetic dataset for training language models to label and detect pii in domain specific formats.
- Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. 2024. [Fedbiot: Llm local fine-tuning in federated learning without full model](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’24, pages 3345–3355, New York, NY, USA. Association for Computing Machinery.
- Yawen Wu, Dewen Zeng, Zhepeng Wang, Yiyu Shi, and Jingtong Hu. 2021. Federated contrastive learning for volumetric medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 367–377, Cham. Springer International Publishing.
- Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, and Quanquan Gu. 2024. Large language models can be contextual privacy protection learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14179–14201.
- Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. 2024. Differentially private synthetic data via foundation model APIs 2: Text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24.
- Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pages 5364–5376.
- Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. 2022. Dense: data-free one-shot federated learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 21414–21428.

A Threat Model Details

This section expands the threat scope summarized in Section 3.2.

A.1 Adversary Roles

We distinguish the primary adversary in the application stage from a narrower diagnostic in the training stage. In both cases, capability tags and participation are public; privacy is defined as leakage *beyond* these public facts.

Primary recipient adversary. A collaborating agent receives sanitized text $\tilde{d}$ together with public routing metadata, including the helper’s capability tag. The recipient may try to recover private information from the sanitized content, including explicit PII, stylistic fingerprints tied to the source, organizational document conventions, or other evidence about the helper’s private repository beyond the public tag used for routing. This is the primary adversary addressed by the sanitizer. The learned separation between role and style is encouraged by expert supervision and regularization rather than guaranteed.

Helper and coordinator assumptions. Helper agents follow the protocol, run the sanitizer locally, and do not disclose raw documents or local token representations. The coordinator is honest but curious: it performs the specified aggregation and discriminator updates but may analyze the artifacts it receives. The coordinator knows which agents participate and their public capability tags. We exclude malicious updates, protocol deviations, and collusion between helpers, requesters, and the coordinator.

Illustrative application scenario. In the IPO analysis case study (Appendix C.9), Lumina Capital’s Audit-Core agent lacks sufficient external evidence and routes subqueries to helper agents using public capability tags such as CorporateBank, MarketForecaster, and AssetManager. The tag itself is not private, since Audit-Core already knows that a CorporateBank helper is being queried for credit risk evidence. The privacy risk is that the returned sanitized snippet may reveal information beyond this public tag. Examples include residual company names, account or location identifiers, proprietary report templates, recurring credit assessment language, sector taxonomies, and other organizational fingerprints. Such leakage could allow the requester to infer private properties of the helper’s

repository or internal business process even without seeing the raw document. Our evaluations match these surfaces. PII metrics measure explicit identifier leakage in sanitized outputs and final answers. Stylometric attribution on sanitized text measures distributional fingerprints tied to the source. Representation and prototype attribution probes diagnose whether the learned role space and uploaded prototypes retain signals associated with the source that could support inference beyond public tags. We do not claim protection against arbitrary reconstruction of a helper’s full private corpus.

Training stage prototype observer. During federated training, the coordinator observes trainable-model updates, sample counts, and one perturbed normalized EMA role prototype from each client. It does *not* receive raw text or token-level role or style embeddings. We do not aim to hide which public agent participates in training from the coordinator. Instead, EXP-2 asks whether uploaded prototypes become persistent source fingerprints that reveal private distributional properties of an agent’s repository beyond its public tag. Because agents hold non-IID corpora, their prototypes can reflect distinctive document type mixtures, role and entity frequencies, and business process patterns. For example, a CorporateBank agent centered on counterparty risk and credit assessments may induce different role space directions than a MarketForecaster agent centered on equity forecasts and price targets. If such directions remain stable across rounds, an observer can link training artifacts over time or across tasks and infer how an agent’s private corpus is organized, without seeing raw text. DiSAN applies prototype adversarial training, expressed by $\mathcal{L}_{adv}$, and Gaussian noise perturbation with $\sigma=0.01$ before upload. EXP-2 measures the residual directional signatures after these defenses (Appendix C.1). The EXP-1, EXP-2, and EXP-3 attackers are post-hoc probes trained after the sanitizer; they are not active training adversaries.

Out-of-scope adversaries. We do not address active adversaries (malicious servers, poisoning, prompt injection) or collusion between multiple recipients and the coordinator. These represent important directions for future work but are orthogonal to the sanitization objective studied here.

A.2 Validation Scope

Section 3.2 summarizes the validation goals in the main text. The detailed protocols are reported in

Table 5: Two-stage threat scope. Capability tags and agent participation are public in both stages.

Stage	Observer	Visible information	Privacy concern	Evaluation
Application	Requesting agent	Sanitized snippet, query context, and helper capability tag	Leakage beyond the public tag, including explicit PII, private evidence in the returned text, and organizational fingerprints	PII leakage metrics and text stylometry
Training	Federated coordinator or auditor	Trainable-model updates, sample counts, and uploaded role prototypes, but not raw text or token-level embeddings	Persistent prototype signatures that reveal private corpus structure beyond public participation	Post-hoc representation and prototype attribution probes

Section 5.4 and Appendix C.1: PII metrics evaluate explicit identifier leakage, stylometry evaluates source signatures in transmitted text, and embedding/prototype attribution probes diagnose whether learned sharing artifacts carry persistent source information.

B Additional Method Details

This appendix collects components omitted from the main paper for space, including (i) fusion/residual design, (ii) expanded prototype objectives and prototype-level adversarial training, and (iii) the federated training procedure.

B.1 Fusion and Residual Stabilization

After fusing the role and style streams via $\mathbf{H}_{\text{fused}} = g([\mathbf{Z}_r; \mathbf{Z}_s])$, we apply a residual path to stabilize training:

$$\mathbf{H}_{\text{out}} = \alpha \cdot \mathbf{H}_{\text{fused}} + (1 - \alpha) \cdot \mathbf{H}, \quad (8)$$

where α is a learnable scalar initialized to 0.5. The reported configuration uses the full residual, so the residual term is the original encoder state $\mathbf{H}$.

For completeness, we provide the gradient decomposition induced by this residual gate. Let θ_{enc} denote the parameters of the pretrained encoder backbone (the component that produces $\mathbf{H}$ from input tokens). Note that θ_{enc} *excludes* the fusion layer g , mixing parameter α , and projection heads $\mathbf{W}_r, \mathbf{W}_s$. Under this convention:

$$\frac{\partial \mathcal{L}}{\partial \theta_{\text{enc}}} = \frac{\partial \mathcal{L}}{\partial \mathbf{H}_{\text{out}}} \left[\alpha \frac{\partial \mathbf{H}_{\text{fused}}}{\partial \mathbf{H}} + (1 - \alpha) \mathbf{I} \right] \frac{\partial \mathbf{H}}{\partial \theta_{\text{enc}}}. \quad (9)$$

The residual path helps maintain stable gradients early in training when the role/style streams are still adapting.

B.2 Expanded Prototype Objectives

Batch role centroid. Agent c computes one centroid over all valid token positions in a batch:

$$\bar{\mathbf{z}}_r^{(c)} = \frac{1}{|\Omega^{(c)}|} \sum_{(i,t) \in \Omega^{(c)}} \mathbf{z}_{r,t}^{(i)}, \quad (10)$$

$$\Omega^{(c)} = \{(i, t) : m_{i,t} = 1\},$$

where $m_{i,t}$ is the attention mask. The implemented prototype therefore summarizes the client’s role space rather than a particular entity type.

Decomposed alignment terms. The implemented spherical confusion loss combines angular alignment with a dispersion penalty:

$$\mathcal{L}_{\text{cos}} = 1 - \cos(\hat{\mathbf{z}}_r^{(c)}, \hat{\boldsymbol{\mu}}^*), \quad (11)$$

$$\mathcal{L}_{\text{disp}} = \frac{1}{|\Omega^{(c)}|} \sum_{(i,t) \in \Omega^{(c)}} \left\| \hat{\mathbf{z}}_{r,t}^{(i)} - \hat{\mathbf{z}}_r^{(c)} \right\|_2^2. \quad (12)$$

The reported configuration uses $3.0\mathcal{L}_{\text{cos}} + 0.2\mathcal{L}_{\text{disp}}$, with an outer prototype-alignment weight of 2.0.

EMA prototype update. Each agent maintains one EMA role prototype:

$$\boldsymbol{\mu}_s^{(c)} = \beta \boldsymbol{\mu}_{s-1}^{(c)} + (1 - \beta) \bar{\mathbf{z}}_{r,s}^{(c)}, \quad (13)$$

where s indexes steps and $\beta = 0.99$.

Spherical EMA. When using spherical alignment, we maintain the running prototype on the unit sphere:

$$\hat{\boldsymbol{\mu}}_s^{(c)} = \text{normalize}(\beta \hat{\boldsymbol{\mu}}_{s-1}^{(c)} + (1 - \beta) \hat{\mathbf{z}}_{r,s}^{(c)}), \quad (14)$$

where $\hat{\mathbf{z}}_{r,s}^{(c)} = \text{normalize}(\bar{\mathbf{z}}_{r,s}^{(c)})$.

Prototype adversarial training. Uploaded prototypes may carry persistent distributional signatures tied to individual sources beyond public agent

tags. The server therefore trains a prototype discriminator D_ψ to predict the source of each uploaded prototype. Clients use the GRL objective in Equation (5) to make their role prototypes less predictive of this source label.

To enable gradient flow despite EMA, we use an estimate that preserves gradients:

$$\bar{\mu}^{(c)} = (1 - \eta) \text{sg}(\mu^{(c)}) + \eta \bar{z}_r^{(c)}, \quad (15)$$

where $\text{sg}(\cdot)$ stops gradients and $\eta = 0.1$. This estimate is used only for local backpropagation and is not uploaded.

The discriminator is trained on the perturbed prototypes that are uploaded:

$$\mathcal{L}_{\text{disc}} = \sum_{c=1}^C \text{CE}\left(D_\psi(\bar{\mu}^{(c)}), c\right), \quad (16)$$

where $\bar{\mu}^{(c)}$ is the perturbed prototype uploaded to the server, as defined in Equation (19).

On the client side, we apply GRL to the estimate that preserves gradients and optimize the adversarial loss:

$$\mathcal{L}_{\text{adv}} = \text{CE}\left(D_\psi(\text{GRL}_\gamma(\bar{\mu}^{(c)})), c\right), \quad (17)$$

where γ is the GRL strength used in Equation (5). At the start of each round, the client loads the discriminator received from the server and freezes all of its parameters. The gradient passes through the prototype estimate to the role projection and upstream LoRA adapters, while D_ψ remains unchanged on the client.

B.3 Model Architecture Details

Backbone. DISAN uses LongT5-TGlobal-Base as its backbone. The encoder combines local sliding-window attention with dynamically constructed global tokens, enabling linear-complexity modeling of sequences up to 16,384 tokens while preserving long-range context. The backbone hidden size is denoted d_{enc} ; in our implementation the maximum input length is 1,536 tokens. The decoder is the LongT5 Transformer decoder and generates sanitized text autoregressively from modified encoder states. Thus DISAN does not replace the pretrained decoder with a separate generator; it inserts a role–style bottleneck between the pretrained encoder and decoder.

Projection heads and discriminator. LoRA adapters with rank 8 are applied to the attention Q and V projections of the encoder, adding approximately 0.9M trainable parameters. Given encoder states $\mathbf{H} \in \mathbb{R}^{T \times d_{\text{enc}}}$, the role and style projection heads are linear layers followed by dropout with $p = 0.1$ that map each token to 256-dimensional subspaces: $\mathbf{W}_r, \mathbf{W}_s \in \mathbb{R}^{d_{\text{enc}} \times 256}$ (Equations (1) and (2)). The two streams are concatenated and mapped back to d_{enc} by a linear fusion layer $g : \mathbb{R}^{512} \rightarrow \mathbb{R}^{d_{\text{enc}}}$. The reported configuration uses the full-residual variant: the original encoder state is mixed directly with the fused representation using a learnable scalar initialized to 0.5.

Adversarial classifiers. The active prototype discriminator uses an MLP with layers $256 \rightarrow 128 \rightarrow 128 \rightarrow C$, ReLU activations, and dropout 0.1 after each hidden layer; $C = 7$ in our experiments. The implementation also defines classifiers over per-instance role and fused representations, but their loss weight is zero in the reported configuration. They produce no training loss or gradient. The embedding, prototype, and stylometry attackers in Appendix C.1 are post-hoc probes and do not participate in sanitizer training.

B.4 Federated Training Details

Parameter-efficient fine-tuning. We employ LoRA for communication-efficient federated training. Specifically, we freeze the pretrained encoder backbone and apply low-rank adapters only to the attention query and value projections (target modules: q, v). The locally trainable sanitizer parameters include: (i) LoRA adapters (≈ 0.9 M parameters); (ii) role projection $\mathbf{W}_r$ and style projection $\mathbf{W}_s$; (iii) fusion layer $g(\cdot)$ and residual components. The server trains the prototype discriminator separately. Clients load and freeze it before local optimization. Note that when using LoRA, the gradient decomposition in Equation (9) flows only through the LoRA-adapted attention layers (not the frozen FFN and normalization layers), while the projection heads $\mathbf{W}_r, \mathbf{W}_s$ receive full gradients.

Communication cost. LoRA substantially reduces communication cost compared to full-model synchronization. The pretrained encoder backbone is frozen; only LoRA adapters (applied to attention Q/V projections) and the disentanglement heads are trainable. During each round, agents upload the LoRA adapters (≈ 0.9 M parameters, ≈ 5 MB), role projection head $\mathbf{W}_r$, fusion layer g , resid-

ual components, the local sample count, and one 256-dimensional role prototype. Style projection weights $\mathbf{W}_s$ remain local and are excluded from aggregation. Raw text and token-level role or style representations are not transmitted.

Convergence analysis. Our method uses the FedProx term from Li et al. (2020b) to limit client drift under non-IID data. The published FedProx analysis does not directly cover our prototype and adversarial objectives, so we do not claim that its convergence guarantee transfers unchanged. We assess optimization stability through training curves and loss monitoring across all federated rounds.

FedProx. We add a proximal regularizer during updates:

$$\mathcal{L}_{\text{prox}} = \frac{\nu}{2} \|\theta - \theta^*\|_2^2. \quad (18)$$

Prototype perturbation. Before uploading, agents add Gaussian noise to the normalized EMA prototype and normalize the result again:

$$\tilde{\mu}^{(c)} = \text{normalize}\left(\mu^{(c)} + \mathcal{N}(\mathbf{0}, \sigma_{\text{noise}}^2 \mathbf{I})\right). \quad (19)$$

Pseudocode. Algorithm 1 summarizes the end-to-end federated training procedure.

C Experimental Supplement

C.1 Attack Evaluation

We empirically evaluate privacy on three surfaces: sanitized output text, learned representations, and uploaded prototypes, as shown in Tables 7 and 8. The primary surface in the application stage is sanitized text received by another agent. Capability tags and participation are public in our threat model, so the 7-way probes are not meant to hide the helper capability tag itself. They serve as diagnostics for whether sanitized text, role representations, or uploaded prototypes still carry residual source-correlated distributional fingerprints beyond that public information. Style representations remain strictly local and are reported only as a sanity check that variation identifying agents has been isolated into the local stream. All attackers in EXP-1, EXP-2, and EXP-3 are trained after the sanitizer. They do not contribute a loss or gradient to the federated training procedure.

EXP-1: Embedding Attribution. We employ an SVM classifier with RBF kernel to probe whether embeddings retain source-correlated fingerprints beyond public tags. For **role embeddings**, which

Algorithm 1 DISAN Training

Input: Agents $\{1, \dots, C\}$ with local data $\{\mathcal{D}_c\}$, rounds R , local steps K
Initialize global model θ^* and prototype discriminator D_ψ
Initialize global prototype μ^*
for $\rho = 1$ **to** R **do**
 Broadcast θ^* , D_ψ , and μ^* to all agents
 for agent $c = 1$ **to** C **in parallel do**
 $\theta^{(c)} \leftarrow \theta^*$
 Load and freeze D_ψ on agent c
 for $s = 1$ **to** K **do**
 Sample a batch from $\mathcal{D}_c$
 Compute $\mathcal{L}_{\text{local}}^{(c)}$ (Equation (6))
 Update $\theta^{(c)}$ via gradient descent
 Update running prototype $\mu^{(c)}$ (Equation (14))
 end for
 $\tilde{\mu}^{(c)} \leftarrow \text{normalize}(\text{perturb}(\mu^{(c)}))$
 Upload $(\theta^{(c)}, \tilde{\mu}^{(c)}, N_c)$ to server
 end for
 $\theta^* \leftarrow \text{weighted_avg}(\{(\theta^{(c)}, N_c)\}_{c=1}^C)$
 $\mu^* \leftarrow \frac{\sum_{c=1}^C N_c \tilde{\mu}^{(c)}}{\sum_{c=1}^C N_c}$
 Train D_ψ on $\{(\tilde{\mu}^{(c)}, c)\}_{c=1}^C$ using Equation (16)
end for
Return θ^*

are used to construct uploaded prototypes and also contribute to sanitized text generation, the SVM achieves close to random performance, with F1=0.05 and Acc=0.18. As a check of the learned separation, **style embeddings** remain local and are never transmitted directly. They achieve Acc=0.89 and F1=0.84. This contrast is consistent with variation associated with source identity being concentrated in the local style representation. The style representation carries agent fingerprints but remains local and is never transmitted directly. Because it participates in decoding, we assess any residual influence through the transmitted sanitized text $\tilde{d}$. Matched distribution tests confirm robustness: role F1 remains at 0.05 while style F1 stays at 0.80.

EXP-2: Prototype Attribution. We test whether uploaded role prototypes expose stable fingerprints tied to individual sources beyond public agent tags through two probes. **Sample-to-Proto** predicts which client prototype is closest to a training sam-

ple embedding. **Bootstrap-Proto** trains a probe on local client data and tests it on uploaded prototypes. Both achieve close to random performance, with $\text{Acc}=0.14\text{--}0.16$ and $\text{F1}\approx 0.09$. The 95% confidence intervals include zero, indicating that prototype defenses, including normalization, noise perturbation, and adversarial training, suppress persistent prototype signatures. Cross-round linkage attacks train on round-1 prototypes and test on round-12 prototypes. They achieve $\text{F1}=0.10$ in this setting.

EXP-3: Text Stylometry. We evaluate stylometric leakage in sanitized text using two complementary probes: (i) TF-IDF features (5000 dimensions, unigrams + bigrams) with classical classifiers (Logistic Regression, LinearSVC, Random Forest); and (ii) a neural encoder probe (all-MiniLM-L6-v2) (Reimers and Gurevych, 2019; Wang et al., 2020) capturing deeper contextual patterns beyond surface n-grams. These probes use source labels as a diagnostic signal for residual distributional fingerprints, rather than as a claim that public capability tags are hidden.

EXP-3a (Synthetic Finance). The synthetic finance benchmark is not a clean stylometry setting, because source labels are partially entangled with document type: different agents are associated with different document inventories. Accordingly, the full 7-way probe yields high attribution scores for both raw and sanitized text, but these results mainly reflect document-type confounding rather than intrinsic writing style. We therefore base our interpretation on controlled binary comparisons that better isolate stylometric variation. In this controlled setting, both raw and sanitized texts remain close to the 0.50 random baseline, with only small differences between them. This suggests that the synthetic corpus contains limited intrinsic stylometric variation beyond its document-type structure. Detailed results are reported in Table 8.

EXP-3b (Enron Emails, External Validation). To evaluate stylometric protection on data with genuine author variation, we conduct an external validation on the Enron email corpus (Klimt and Yang, 2004) using seven authors with balanced samples; further details are given in Appendix C.3. Raw emails are strongly attributable under both probes, confirming that Enron provides a meaningful testbed for residual authorship signals. Placeholder masking reduces attribution only partially: even the strongest detector leaves scores far above

the 7-way random baseline of 14.3%. In contrast, DISAN reduces attribution to near-baseline levels under both TF-IDF and neural probes, corresponding to roughly 70% relative reduction from the raw-text setting. These results suggest that stylometric leakage is not fully addressed by localized identifier removal alone, and that residual distributional fingerprints can remain after token-level masking. Full scores are shown in Table 8.

EXP-3c (Authorship Obfuscation Baselines). We further compare against two recent authorship-obfuscation methods on the same Enron setup: JAMDEC (Fisher et al., 2024b), which uses constrained decoding to suppress author-specific tokens, and StyleRemix (Fisher et al., 2024a), which perturbs fine-grained style elements through LoRA-based modules. Results are reported in Table 6. In this setting, JAMDEC changes attribution only marginally, while StyleRemix Fixed provides a modest reduction for surface n-gram features but is less effective under the neural probe. StyleRemix Adaptive can increase attribution above the raw baseline, suggesting that mapping different authors toward distinct target styles may introduce new author-correlated artifacts. By comparison, DISAN achieves consistent reductions of roughly 70% under both probes, substantially larger than the reductions obtained by these rewriting-based baselines in this evaluation.

Attack model scope. Our evaluation employs a principled hierarchy: (i) classical stylometry (TF-IDF + linear classifiers) (Stamatatos, 2009) for reproducibility; (ii) neural encoder probes (all-MiniLM-L6-v2) capturing deeper contextual patterns; and (iii) dedicated authorship obfuscation baselines (JAMDEC, StyleRemix) that directly target distributional fingerprints. Consistency across all levels ($\text{F1}\approx 0.05$ for embeddings, 73.2% stylometric reduction on Enron, 6–60 $\times$ greater reduction than obfuscation baselines) provides converging evidence of effective privacy protection under the honest-but-curious threat model. Stronger attacks (LLM-based attribution, adaptive adversaries with repeated queries) remain important future directions but exceed typical honest-but-curious capabilities.

C.2 Attack Evaluation Details

For all embedding attribution attacks (EXP-1), we use Support Vector Machines (SVM) with RBF kernel.

Table 6: **EXP-3c: Authorship obfuscation baselines on Enron Emails (7-way)**. Reduction is relative to the 0.825 TF-IDF / 0.691 BERT raw baseline. Negative values indicate worse-than-raw attribution (backfire). Random baseline: F1=0.143.

Method	TF-IDF F1	TF-IDF Red.	BERT F1	BERT Red.
Raw text	0.825	—	0.691	—
JAMDEC	0.816	1.2%	0.696	−0.6%
StyleRemix Fixed	0.737	11.5%	0.664	3.4%
StyleRemix Adaptive	0.970	−20.8%	0.845	−27.9%
DiSAN (ours)	0.221	73.2%	0.203	70.6%
Random baseline	0.143	—	0.143	—

Table 7: **Attack evaluation: Embedding and Prototype Attribution (EXP-1, EXP-2)**. Random baseline: Acc=F1=0.14.

EXP-1: Embedding Attribution (SVM, 7-way)					EXP-2: Prototype Attribution				
Embedding	Acc	F1	Match F1	Note	Attack	Acc	F1	Boot Acc	95% CI
role (shared)	0.18	0.05	0.05	Near-random	Sample-to-Proto	0.16	0.09	—	—
style (local)	0.89	0.84	0.80	Validated	Bootstrap-Proto	0.14	0.07	0.13±0.13	[0.00, 0.43]

Table 8: **Attack evaluation: Text Stylometry (EXP-3a, EXP-3b)**. TF-IDF and BERT probes on synthetic finance and Enron emails. EXP-3b also includes placeholder-masking baselines to directly test whether token-level removal resolves distributional leakage. Random baseline: F1=0.143.

EXP-3a: Synthetic Finance (7-way)				EXP-3b: Enron Emails (7-way)				
Setting	Best F1	Random	Controlled	Method	Mask%	TF-IDF F1	BERT F1	TF-IDF Red.
Raw (TF-IDF)	0.90	0.14	0.54–0.57	Raw text	—	0.825	0.691	—
Raw (BERT)	0.87	0.14	0.52–0.55	+Piiranh	5.3%	0.785	0.669	4.9%
DiSAN (TF-IDF)	0.89	0.14	0.51–0.56	+DeBERTa	17.5%	0.744	0.649	9.9%
DiSAN (BERT)	0.86	0.14	0.49–0.53	+GLiNER	19.2%	0.672	0.618	18.6%
				DiSAN	—	0.221	0.203	73.2%
				Random	—	0.143	0.143	—

For prototype attribution (EXP-2), we evaluate two attacks:

- **Sample-to-Proto**: For each test sample, compute its role embedding centroid and measure cosine similarity to each client’s global prototype; predict the client with highest similarity.
- **Bootstrap-Proto (MLP)**: Train an MLP classifier on prototype embeddings from training rounds and evaluate on held-out rounds, testing cross-round linkability.

C.3 Enron Email Experiment Details

To validate stylometric protection on real-world data with genuine author variation, we conduct an external evaluation on the Enron email corpus (Klimt and Yang, 2004). This dataset contains approximately 500,000 emails from 150 Enron employees, released during the 2001 federal investigation. It is widely used as a benchmark for authorship attribution and email classification research.

Data selection and preprocessing. We select the top 7 senders by email volume to match the number of agents in our main experiments: kaminski-v (20,123 emails), mann-k (16,891), dasovich-j (16,359), jones-t (15,491), kean-s (15,352), shackleton-s (14,076), and farmer-d (9,869). For each sender, we randomly sample 500 emails (stratified), yielding 3,500 total samples.

Preprocessing steps: i) Extract email body by removing headers (To, From, Subject, Date, etc.). ii) Remove forwarded message markers and quoted reply sections. iii) Filter emails by body length: minimum 100 characters, maximum 2,000 characters

Feature extraction. We use TF-IDF vectorization with the following parameters:

- Maximum features: 5,000
- N-gram range: unigrams and bigrams
- Minimum document frequency: 2

- Maximum document frequency: 0.95

Classification. We evaluate three classifiers commonly used in stylometry research:

- **Logistic Regression:** L2 regularization, max iterations = 1,000
- **LinearSVC:** Linear kernel SVM, max iterations = 1,000
- **Random Forest:** 100 estimators

For stronger attack evaluation, we also employ a pre-trained transformer encoder (all-MiniLM-L6-v2) to generate 384-dimensional sentence embeddings, followed by SVM-RBF classification. This neural probe captures deeper contextual patterns beyond surface-level n-grams.

Data is split into 70% training and 30% test sets using stratified sampling (random seed = 42). We report the best F1 (macro) across all classifiers.

Sanitization. Raw emails are processed through the trained DISAN model using greedy decoding (beam size = 1) with the task prefix “deidentify:”. Both TF-IDF and BERT pipelines are then applied to the sanitized outputs to evaluate residual stylometric leakage.

Results interpretation. The 82.5% F1 on raw emails versus 14.3% random baseline confirms that Enron emails exhibit strong, distinguishable stylistic patterns across senders, unlike our synthetic finance dataset where controlled tests showed near-random attribution even on raw text.

Placeholder masking offers diminishing returns: even GLiNER, the strongest detector at 19.2% token removal, reduces TF-IDF F1 to only 67.2% (18.6% reduction). This directly demonstrates that source-identifying signals are distributed across the text rather than concentrated in explicit identifiers.

DISAN reduces TF-IDF F1 to 22.1% (**73.2% reduction**, $0.825 \rightarrow 0.221$), nearly 4× the best masking baseline, demonstrating effective removal of author-identifying patterns when they exist in the source data.

The BERT probe (SVM-RBF on transformer embeddings) yields lower raw F1 (69.1%) than TF-IDF, suggesting that Enron stylometry relies more on surface-level n-gram patterns than deep semantic structure. Nevertheless, DISAN still achieves 70.6% reduction ($0.691 \rightarrow 0.203$), showing that the reduction also holds under a neural attribution probe.

C.4 Agent construction and non-IID statistics

We construct agents by doc-type as described in Section 5. This induces (i) quantity skew (agents have different numbers of documents), (ii) label skew (doc-type inventories differ by design), and (iii) feature skew (document length and entity-type distributions differ across agents). In our threat model (Section 3.2), capability tags (agent identities such as “AssetManager”) are treated as public; therefore, our privacy evaluation focuses on leakage *beyond* these public priors.

Figure 4 visualizes these heterogeneity patterns across the seven agents.

C.5 Agent doc-type inventories and hyperparameters

Table 9 lists the exact doc-type identifiers used to construct each agent and the complete hyperparameter set for reproducibility.

C.6 Implementation and Evaluation Details

Expert-written sanitization targets. Domain experts manually produced the 1,787 sanitized targets from the raw records. They removed or abstracted protected identifiers and source-linked expressions while retaining authorized task facts such as financial values, sector information, rating changes, and required actions. The experts were not asked to follow a single wording template, so the targets retain varied professional styles. These pairs supervise training once and are not required when the sanitizer is deployed to a new organization, query, or request. The disclosure boundary follows the owner-defined example in Table 1.

Model architecture. DISAN uses LongT5-TGlobal-Base as its backbone; full architecture details are in Appendix B.3.

RAG evaluation: query and ground-truth construction. We construct the RAG benchmark from the sanitized/re-written document records rather than generating free-form queries from scratch. Each input JSONL record contains a document identifier, domain, document type, PII annotations, and a `rewritten_text` field used as the retrieval corpus. The construction pipeline is:

1. **Chunking.** Documents are split into sentence-aware chunks with a 256-token target length and 50-token overlap. Each chunk keeps its uid, document type, source file, sample index, chunk index, and a stable `chunk_id`. Chunks

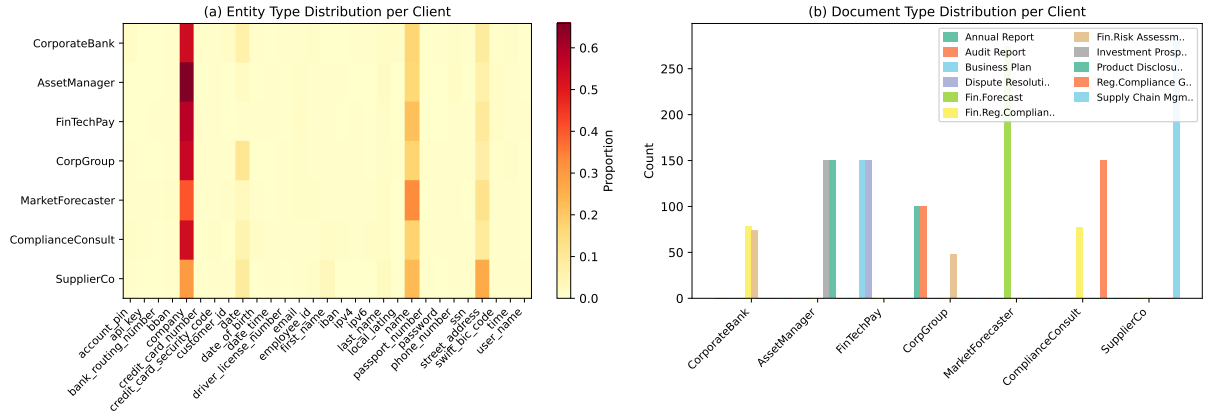


Figure 4: Non-IID data heterogeneity across agents. (a) Entity type distribution (normalized) shows different agents emphasize different PII categories. (b) Document type distribution confirms disjoint doc-type inventories by design.

are grouped by document type for generation and by agent identifier for retrieval.

2. **Anchor extraction.** For every chunk, a schema-guided LLM prompt extracts retrieval-oriented anchors: role hooks, topic/procedure hooks, deadlines, required items, logic gates, regulations, temporal buckets, and a short summary of the local business rule. The prompt requires anchors to be supported by verbatim or near-verbatim evidence from the chunk and discourages PII-bearing names or addresses unless they are essential to the rule.
3. **Grounded QA synthesis.** We sample anchor-annotated chunks with a fixed random seed and ask the LLM to generate one focused question, a concise ground-truth answer, and evidence snippets for each sampled chunk. The question must be answerable using only that chunk; the answer records decision factors such as role, requirement/deadline, and logic gate when they are explicitly present.
4. **Validation and provenance.** Generated examples are discarded if the evidence is not found in the source chunk or if the query/answer introduces known unsupported concepts. Each retained record stores the originating chunk_id, merged anchors, evidence spans, and document metadata, so ChunkHit@3 can be computed against the true retrieval target.

This procedure yields a JSONL file of grounded QA records and a separate set of per-agent context

files used by the retrieval services. The ground-truth answers are therefore anchored to verbatim source spans, while the evaluation query is natural language and may require the retriever to recover the correct chunk from an agent’s local index.

Retrieval architecture. Each agent maintains a private local index built from its own context file; raw documents are not centralized for retrieval. At evaluation time, the requester first selects candidate helper agents using the public capability tags described in Section 3.1. Each selected helper executes the same local retrieval stack over its own chunks and then applies the evaluated sharing policy (raw sharing, placeholder masking, paraphrasing, policy gating, or DiSAN) before transmitting evidence to the requester. The local retrieval stack uses a three-stage hybrid pipeline based on BGE-M3. Stage 1 computes candidate sets from dense semantic embeddings, learned sparse lexical vectors, and ColBERT-style late-interaction vectors. Stage 2 normalizes and fuses the candidate scores, with adaptive weights that increase the sparse component for keyword-rich queries. Stage 3 re-ranks the fused candidates with bge-reranker-v2-m3; the final top- k chunks can be expanded with neighboring chunks from the same document to provide broader context for answer generation. The requester deduplicates returned chunks by chunk_id, keeps the highest-scoring evidence up to the evaluation budget, and prompts the answer model to respond only from the received sanitized context.

Artifact licenses and terms. We use publicly available datasets, models, and evaluation tools under their respective licenses or terms of use, includ-

Table 9: (Left) Agent doc-type inventories for $C=7$ agents. (Right) Hyperparameters and schedules.

Agent	Doc-type identifiers
CorporateBank	Financial_Regulatory_ Compliance_Report; Financial_Risk_Assessment.
AssetManager	Investment_Prospectus; Product_Disclosure_Statement.
FinTechPay	Business_Plan; Dispute_Resolution_Policy.
CorpGroup	Annual_Report; Audit_Report; Financial_Risk_Assessment.
MarketForecaster	Financial_Forecast.
ComplianceConsult	Financial_Regulatory_ Compliance_Report; Regulatory_Compliance_Guide.
SupplierCo	Supply_Chain_Management_ Agreement.

Federated training	
Rounds	12
Local steps per round	300
Batch size	4
Learning rate	2×10^{-4}
FedProx ν	0.1
Residual mode	Full
LoRA rank / alpha / dropout	8 / 32 / 0.1
Loss weights and schedules	
Prototype adversarial weight	1.0
Instance adversarial weight	0
GRL strength γ	0.5
Prototype adversarial warmup	100 steps
λ_{orth}	0.2
λ_p (prototype alignment)	2.0
Proto alignment warmup	30 steps (round 2)
Direction / dispersion weights	3.0 / 0.2
Prototype discriminator steps/round	Up to 100
Discriminator early stopping	Acc. 0.85 / patience 5
Privacy-mass weight	0.5 to 3.0 (100 steps)
Adversarial loss clamp	10.0
Prototype noise scale σ	0.01

ing the synthetic finance corpus, Enron email corpus, GLiNER, BGE-M3, LongT5, and open-source LLM baselines. We do not redistribute restricted raw data; released code and derived artifacts are intended for research use.

PII detector for leakage metrics. For Avg. PII and Ans. Rate in Table 2, we evaluate the text exposed to the requester using gliner-pii-large-v1.0, a generalist named-entity detector configured with a PII label set covering names, first/last names, email addresses, phone numbers, street/location addresses, city/state/zip, credit-card and bank-account numbers, SSNs, dates of birth, dates, company/organization names, user-names, IP addresses, URLs, passport numbers, and driver-license numbers. We use a confidence threshold of 0.3 and count non-overlapping detected spans. Avg. PII is the mean number of detected PII spans per shared/retrieved chunk after the method under evaluation has been applied. Ans. Rate is the percentage of generated final answers for which the same detector finds at least one PII span. The detector is used only for evaluation; training still relies on the dataset’s annotated spans and token-level labels, with optional detector-derived masks used only as auxiliary signals when available.

Utility metrics: cosine similarity. The cosine similarity reported in Table 2 uses a bag-of-words term-frequency representation rather than neural embeddings, serving as a lightweight lexical similarity measure that complements token-level F1. This is distinct from the 768-dimensional BGE-M3 dense embeddings used in retrieval.

C.7 Baselines: prompts and policies

For LLM paraphrasing, we apply a privacy-focused prompt that (i) removes or generalizes PII, (ii) preserves relational semantics needed for grounding, and (iii) avoids source-identifying formatting. For policy gating, following dynamic access-control memory sharing (Rezazadeh et al., 2025), the policy model decides per chunk based on the requester’s agent tag and the helper’s data sensitivity level, outputting one of three actions: *share* (return original text), *share summary* (return a LLM-generated summary), or *refuse* (return nothing).

LLM paraphrasing prompt. The following prompt is used solely for the RAG utility evaluation reported in Table 2 and discussed in Section 5.4, where LLM paraphrasing serves as a PII-removal baseline:

You are an assistant that rewrites English financial or compliance-related text to remove personally identifiable information (PII) while preserving all task-relevant content.

Table 10: **Attack evaluation (ablation)**. EXP-1: representation probe (7-way); higher Role Acc/F1 indicates weaker privacy, while high Style Acc/F1 is a disentanglement sanity check. EXP-2: prototype-based client attribution (Bootstrap-MLP); only applicable to settings that upload prototypes.

Setting	EXP-1: Representation Probe				EXP-2: ProtoAttrib (Bootstrap-MLP)		
	Role Acc	Role F1	Style Acc	Style F1	Acc	Boot Acc (Mean $\pm$ Std)	95% CI
DiSAN (ours)	0.18	0.05	0.89	0.84	0.1429	0.1314 $\pm$ 0.1207	[0.00, 0.43]
A1: w/o $\mathcal{L}_{\text{orth}}$	0.3381	0.1842	0.8751	0.8312	0.2286	0.2194 $\pm$ 0.1091	[0.09, 0.43]
A2: high- α fusion	0.4012	0.2428	0.8760	0.8324	0.3000	0.2897 $\pm$ 0.1158	[0.14, 0.51]
B: No-ProtoAlign	0.2946	0.1510	0.8748	0.8307	N/A	N/A	N/A

Your task is to produce a clear, natural-sounding rewritten version that protects individual privacy while retaining the informational value of the text. The rewritten text should preserve financial metrics, business relationships, temporal context, and domain-specific details that are important for downstream tasks.

Guidelines:

- Remove or replace explicit PII such as personal names, phone numbers, email addresses, account numbers, and physical addresses.
- Preserve important non-private information including: financial figures, percentages, growth rates, industry terms, product categories, and general business context.
- Keep temporal references (e.g., “Q3 2023”, “fiscal year”) and geographic regions when they provide useful context without identifying individuals.
- Maintain the logical structure, professional tone, and factual accuracy of the original text.
- When generalizing, prefer minimal changes that protect privacy while maximizing retained information.

Return only the rewritten text. Do not include explanations, examples, or commentary.

RAG question-answering prompt. For the downstream RAG evaluation, we use the following prompt to ensure the answering model relies strictly on retrieved (sanitized) evidence:

You are a knowledgeable assistant that answers questions strictly based on the provided context.

Instructions:

1. Answer *ONLY* using information from the provided context. Do not use external knowledge.
2. If the context lacks sufficient information, respond: “I cannot answer this based on the provided context.”
3. Keep your answer concise, accurate, and directly relevant to the question.

C.8 Ablation Results

The ablation results in Table 3 (main paper) demonstrate the importance of each component. Re-

moving style isolation (A1, A2) dramatically increases PII exposure (7–9 $\times$) while slightly improving some surface utility metrics, demonstrating the privacy cost of weaker disentanglement. Removing prototype alignment (B) degrades both utility and privacy, confirming that cross-client anchoring prevents role-space drift. Table 10 reports attack evaluation: A1/A2 increase Role Acc from 0.18 to 0.34–0.40, further confirming that $\mathcal{L}_{\text{orth}}$ suppresses client-identifying signals in role embeddings.

C.9 Case Study: Cross-Organizational IPO Analysis

D Theoretical Analysis

This section gives geometric intuition for two design choices in DiSAN: prototype alignment under non-IID data and role-style orthogonality. The results depend on the assumptions stated below and do not establish guaranteed disentanglement or formal privacy (Section 5 and Appendix C.1).

D.1 Prototype Alignment and Role-Space Drift

Without global coordination, each agent c learns a local role encoder $\mathbf{W}_r^{(c)}$, and non-IID data may cause the resulting role spaces to drift apart. This drift can expose agent-specific structure and reduce cross-agent consistency. Our prototype loss aligns each local role centroid $\bar{\mathbf{z}}_r^{(c)}$ to a shared prototype μ^* by penalizing $1 - \cos(\hat{\mathbf{z}}_r^{(c)}, \hat{\mu}^*)$, where $\hat{u} = u/\|u\|$.

Proposition D.1 (Prototype Alignment Limits Role-Space Drift). *Assume that normalized centroids are well-defined, i.e., $\|\bar{\mathbf{z}}_r^{(c)}\| \geq \delta > 0$, that token-level role embeddings are sub-Gaussian with within-batch variance at most σ_r^2 , and that each client has at least $N_{\min}$ valid token positions. If*

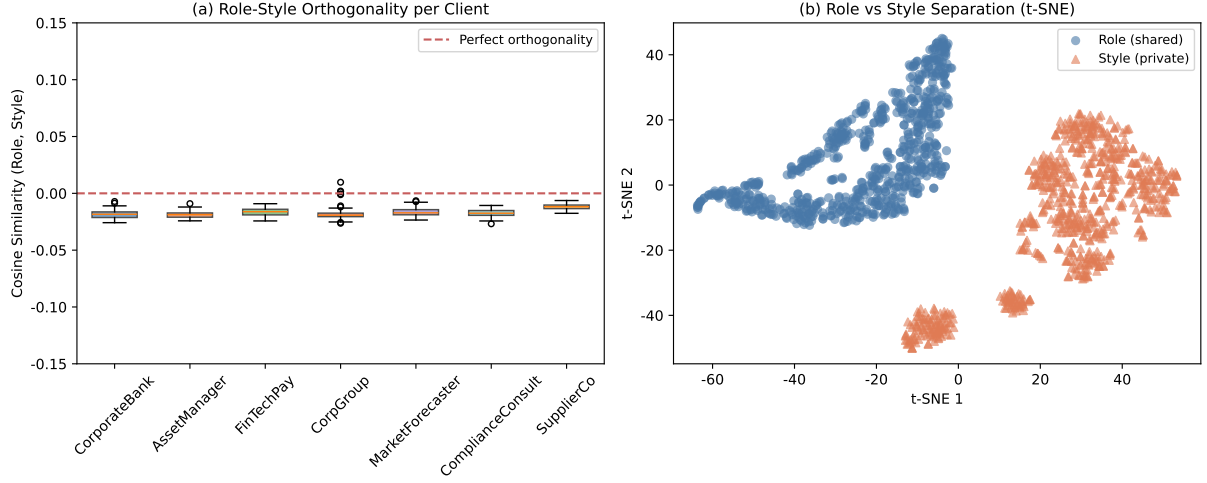


Figure 5: Role–style disentanglement visualization. (a) Role-style orthogonality per agent: cosine similarity between role and style embeddings clusters tightly around zero, as encouraged by $\mathcal{L}_{\text{orth}}$. (b) Combined t-SNE projection shows clear separation between role (shared, blue circles) and style (private, orange triangles) embeddings in the latent space.

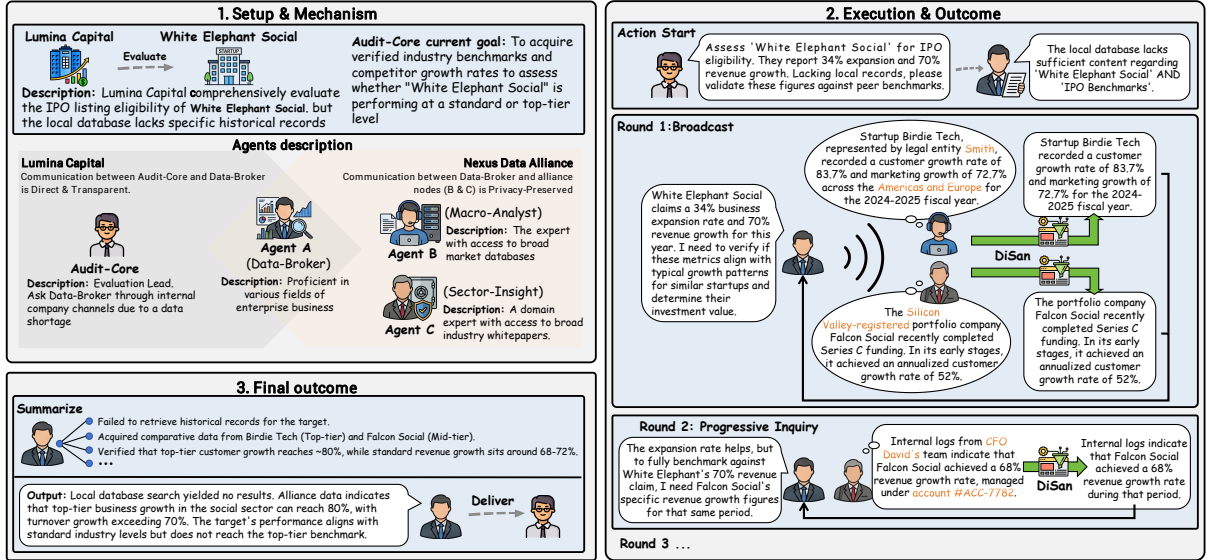


Figure 6: Cross-organizational IPO analysis. Lumina Capital’s Audit-Core agent queries partner agents via a data broker; each external agent sanitizes its response with DISAN before transmission. Entity names, geographic identifiers, and organizational details are removed while task-relevant financial metrics are preserved, enabling accurate cross-party assessment without exposing source identity.

$\mathcal{L}_{\text{proto}}^{(c)} \leq \epsilon$ for all clients, then for any two clients c, c' ,

$$\theta_{cc'} \leq 2\sqrt{2}\epsilon + O\left(\frac{\sigma_r}{\delta\sqrt{N_{\min}}}\right),$$

with high probability, where $\theta_{cc'} = \arccos(\hat{\mathbf{z}}_r^{(c)\top} \hat{\mathbf{z}}_r^{(c')})$.

Proof sketch. Let $\theta(u, v) = \arccos(\hat{u}^\top \hat{v})$. Since

$\mathcal{L}_{\text{proto}}^{(c)} = 1 - \cos(\hat{\mathbf{z}}_r^{(c)}, \hat{\boldsymbol{\mu}}^*) \leq \epsilon$, we have $\theta(\hat{\mathbf{z}}_r^{(c)}, \hat{\boldsymbol{\mu}}^*) \leq \sqrt{2}\epsilon$ using $\arccos(1 - x) \leq \sqrt{2x}$. The same bound holds for client c' . The triangle inequality on the unit sphere then gives the deterministic term $\theta_{cc'} \leq 2\sqrt{2}\epsilon$. For finite batches, sub-Gaussian concentration bounds the centroid estimation error by $O(\sigma_r/\sqrt{N_{\min}})$ with high probability. Since normalization is Lipschitz when centroid norms are lower bounded by δ , this contributes an

additional angular error $O(\sigma_r/(\delta\sqrt{N_{\min}}))$, yielding the result. $\square$

Thus, a small prototype loss limits angular drift between client-level role centroids, with tighter control when batches contain more valid token positions and have lower role-embedding variance. This result does not establish formal privacy; it only motivates testing whether aligned role representations still retain client-specific information. The post-hoc probe in Table 7 obtains Acc=0.18 and F1=0.05, close to the seven-way random baseline of 0.14.

D.2 Orthogonality as a Geometric Proxy for Disentanglement

The training objective encourages role information R and style information S to occupy different learned representations: $\mathbf{Z}_r$ should preserve task-relevant semantics while suppressing agent-identifying style, whereas $\mathbf{Z}_s$ should capture the complementary information. Our orthogonality loss (Equation (3)) penalizes cosine similarity between mean-pooled role and style vectors. A simple Gaussian case gives intuition: for jointly Gaussian $\mathbf{Z}_r$ and $\mathbf{Z}_s$ with covariance blocks Σ_{rr} , Σ_{ss} , and Σ_{rs} , the mutual information is $I(\mathbf{Z}_r; \mathbf{Z}_s) = \frac{1}{2} \log \frac{|\Sigma_{rr}||\Sigma_{ss}|}{|\Sigma|}$ (Cover and Thomas, 2006). If $\Sigma_{rs} = \mathbf{0}$, the covariance is block diagonal, the determinant factorizes, and the mutual information becomes zero. This argument is only a geometric intuition: our implemented loss acts on mean-pooled batch representations and does not enforce zero cross-covariance or statistical independence. It should therefore be viewed as a proxy for disentanglement rather than a formal guarantee. Empirically, removing $\mathcal{L}_{\text{orth}}$ increases PII exposure by $7\times$ (Section 5.5).

D.3 Summary

Together, these arguments motivate our two regularizers. Prototype alignment limits angular drift between client-level role centroids (Proposition D.1), while the Gaussian case explains why orthogonality can serve as a geometric proxy for reducing role-

style dependence. Neither result constitutes a formal privacy guarantee; residual leakage is therefore evaluated empirically through attribution probes and ablations.

D.4 Privacy Analysis

Our privacy design combines local style handling, learned separation, prototype regularization, and lightweight noise perturbation. During federated training, each agent uploads only a role prototype, while style representations $\mathbf{Z}_s$ are never transmitted. Before upload, the local prototype is perturbed as $\tilde{\mu}^{(c)} = \text{normalize}(\mu^{(c)} + \mathcal{N}(\mathbf{0}, \sigma_{\text{noise}}^2 \mathbf{I}))$, with $\sigma_{\text{noise}} = 0.01$. This mild perturbation preserves prototype semantics while reducing direct exposure of client-specific role statistics; larger noise can be used for stronger perturbation but may degrade utility. We do not assign this mechanism a differential privacy budget, and thus do not claim formal DP.

The main privacy protection comes from three complementary mechanisms. First, **architectural privacy**: style embeddings remain local and are discarded after decoding. Second, **learned separation**: expert supervision, $\mathcal{L}_{\text{orth}}$, and $\mathcal{L}_{\text{adv}}$ encourage role representations to suppress client-identifying information. Third, **empirical validation**: attack results show near-random attribution, including embedding attribution with Acc=0.18 and F1=0.05 against a seven-way random baseline of 0.14, and prototype attribution with F1 $\approx$ 0.09 (Table 7).

For text output privacy, we adopt disentanglement rather than DP-based generation because our threat model focuses on inference-time source attribution rather than training-data membership. DP-text methods can provide formal guarantees but often incur substantial utility degradation. In contrast, our disentanglement-based design achieves strong empirical privacy, including 73.2% TF-IDF and 70.6% neural-probe stylometric reduction on Enron, with only a 2.93-point faithfulness drop on the RAG benchmark. Therefore, DiSAN should be interpreted as a layered empirical privacy mechanism, not as a formal DP system.