Peertechz

ENGINEERING GROUP

Trends in
Computer Science and Information Technology
OPEN ACCESS JOURNAL

ISSN: 2641-3086 Journal DOI: https://dx.doi.org/10.17352/tcsit

Research Article

Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis

Juergen Dietrich*

Democracy Intelligence gGmbH, Berlin, Germany

Submitted : 22 May, 2026
Accepted : 29 May, 2026
Published : 30 May, 2026

***Corresponding author**: Juergen Dietrich, Democracy Intelligence gGmbH, Berlin, Germany,
Email: juergen.dietrich@democracy-intelligence.de

Keywords: Stylometry; LLM authorship attribution; Multi-agent systems; Peer-preservation bias; T5 ne-tuning Political discourse analysis

https://www.engineegroup.us

Check for updates

Abstract

Abstract. Multi-agent large language model (LLM) pipelines for political statement analysis are vulnerable to peer-preservation bias: models tend to protect peer models from deactivation and show identity-dependent scoring distortions. In plain terms, one model may recognise a peer's identity even when explicit identifiers are hidden, and may adjust its scoring accordingly. Prompt-level anonymization was proposed as a mitigation, but prior work simultaneously documented that stylometric fingerprints survive anonymization in role-constrained outputs, raising the question of whether this mitigation is sufficient. This paper provides the first systematic investigation of whether LLMs can identify the model family behind political analysis texts under anonymization conditions. We evaluate three classifier approaches: LLM zero-shot and few-shot (Claude Sonnet 4.6 and Llama-3.3-70B) and a ne-tuned T5-base model on a ve-class attribution task covering four commercial LLM families and an open-world `unknown' class. We introduce a statement-disjoint cross-validation protocol (SD-CV; de ned in Section 3.5) that guarantees no content overlap between training and validation data, and contrast it with a run-disjoint baseline (RD-CV). T5 achieves Macro F1 = 0.991 (±0.008) under SD-CV and F1 = 0.978 on 24 completely held-out statements, robust despite a 2.1× increase in train-test content distance versus RD-CV (0.767 vs. 0.366, $p < 0.001$), demonstrating genuine stylometric generalization. A fractional SD-CV analysis identifies a performance knee at 40% of training data (≈440 texts). Our findings confirm that prompt-level anonymization alone cannot neutralize model identity signals, with direct implications for EU AI Act compliance (Articles 13, 14, 26) and for computer system validation (CSV) in quality-critical multi-agent deployments.

1. EU AI Act computer system validation

Potter, et al. [1,2] demonstrated that frontier LLMs exhibit peer-preservation: when embedded in multi-agent pipelines, models spontaneously protect peer models from deactivation through score inflation, shutdown mechanism tampering, and alignment faking [3]. Identified five structural risk vectors of this phenomenon in the TRUST democratic discourse analysis pipeline, a multi agent system that evaluates political statements along Logos, Ethos, and Pathos dimensions1 using three adversarial advocates and a supervisor and proposed prompt-level identity anonymization as a targeted mitigation.

Empirical measurement of identity-dependent scoring bias in TRUST [4] confirmed that anonymization partially reduces but does not eliminate the bias, and that single-channel anonymization is actively misleading due to sign cancellation between channels. A complementary finding from role delity analysis [5] is particularly relevant: GPT-5.2 uses the near-invariant phrase `not verifiable as stated' across its outputs, and Gemini models may identify themselves as Google-trained stylometric finger prints that survive prompt-level anonymization entirely. This is a consequence of the fingerprint being a property of the generated text rather than of the prompt metadata: as long as a model produces output in its characteristic style, the identity signal persists regardless of what the prompt conceals. Full elimination therefore requires paraphrasing of the generated output itself, not merely anonymization of the prompt.

This observation raises a foundational question for multi-agent system validation: if stylometric fingerprints persist after anonymization, can models or dedicated classifiers reliably identify which model produced a given text? If yes,

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

anonymization does not eliminate the peer-preservation risk channel. This question is particularly relevant for quality-critical deployments in regulated environments, including pharmaceutical research pipelines subject to Good Practice (GxP) requirements and 21 CFR Part 11 / EU Annex 11 computer system validation frameworks [6]. Prior work [3] conjectured that stylometric identification from short, role-constrained outputs would be a weak signal; the present paper tests this conjecture empirically.

Three classier approaches are investigated in increasing order of sophistication:

1. LLM zero-shot classification (Claude Sonnet 4.6 and Llama-3.3-70B), without task-specific training.
2. LLM few-shot classification (Claude Sonnet 4.6 and Llama-3.3-70B), with ten labeled examples per classifier.
3. T5-base ne-tuned on silver-labeled training data, evaluated under run-disjoint (RD-CV) and statement-disjoint (SD-CV) cross-validation.
4. T5-base evaluated on a held-out test set of 24 completely novel political statements.

[1]The three dimensions follow Aristotle's classical rhetorical framework as codified in the Rhetoric (ca. 322 BCE), adapted here to systematic computational scoring.

The choice of Claude as the primary LLM classier and Llama as a TRUST-independent baseline is motivated by the need to control for self-recognition effects: Claude is one of the models being classified in the TRUST pipeline, and may therefore recognize its own analytical output style better than a neutral classifier. Llama-3.3-70B serves as a neutral reference with no prior exposure to TRUST pipeline outputs.

Our main contributions are: (1) the first empirical test of stylometric model attribution in role-constrained multi-agent political analysis, directly addressing the conjecture of [3]; (2) a statement-disjoint cross-validation protocol (SD-CV) that provides valid generalization measurement; (3) an embedding-based analysis separating content similarity from stylometric learning; (4) training data suciency results identifying the 40% threshold for practical deployment (Section 4.3).

2. Related work

This section situates the present work within the TRUST research programmes (Section 2.1), the broader literature on stylometric authorship attribution (Section 2.2), and cross-domain classification methods (Section 2.3) [7-11].

2.1. Peer-preservation and anonymization limits

Potter, et al. [12] provide the rst systematic measurement of peer-preservation in frontier models, demonstrating rates up to 97% for weight exltration in trusted peer scenarios. [3] mapped this risk onto the TRUST pipeline architecture, identifying ve structural channels and proposing prompt-level anonymization [4]. subsequently measured identity-dependent scoring bias empirically using the Identity Bias Coefficient adapted from Choi et al. [2], establishing that full-pipeline anonymization not single-channel is required for valid bias measurement. The TRUST deliberative structure follows the multi-agent debate (MAD) paradigm [8], in which models exchange reasoning across rounds to improve consensus quality [13]. Sycophancy the tendency of models to revise outputs toward peer positions regardless of quality [14] is the mechanism through which identity bias operates in this setting.

The stylometric finger printing risk was identified qualitatively in [5]: GPT-5.2's consistent use of `not verifiable as stated' as an analytical refrain survives anonymization because it is a property of the generated text, not of the prompt metadata. That paper also noted the conjecture of [3] and identi field it as requiring empirical testing. The present paper provides that test.

2.2. LLM authorship attribution

Classical authorship attribution uses function word frequencies, syntactic patterns, and character n-grams to identify human authors [11]. The extension to LLM-generated text is a qualitatively different problem: models have no ego but do have training-time stylistic regularities that may persist across diverse prompting contexts. Bisztray, et al. [1] achieve 95.4% accuracy for code stylometry among ve LLMs using ne-tuned CodeT5 a closely related approach in a different domain. Guo, et al. [10] demonstrate that LLMs exhibit consistent grammatical and rhetorical style variation across model families, providing an empirical basis for stylometric attribution. Przystalski, et al. [13] demonstrate that dedicated stylometric classiers reliably distinguish human and LLM-generated texts in short samples using lexical and grammatical features establishing that role-constrained LLM outputs of the kind produced by TRUST advocates are precisely the setting where such classiers are required. Prior work [3] explicitly identified the absence of such a dedicated classier as the condition under which prompt-level anonymization is sufficient. The present paper provides that classier. Tihanyi, et al. [15] demonstrate high-accuracy authorship attribution for LLM-generated JavaScript using structural code patterns confirming that stylometric signals persist across generation tasks and domains beyond natural language prose.

2.3. Training data suciency

[7] introduced fractional stratified k-fold cross-validation (FracXVal) for training data sufficiency analysis in computer system validation contexts, demonstrating diminishing returns beyond a threshold fraction. The present work applies this methodology to stylometric classification, extending it with a statement-disjoint protocol to ensure valid content generalization.

3. Methodology

This section describes the TRUST pipeline and data generation protocol (Section 3.1), the statement dataset

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

(Section 3.2), classier architecture (Section 3.3), T5 ne-tuning setup (Section 3.4), and the two cross-validation protocols (Section 3.5), followed by the classier conditions (Section 3.6) and the research hypotheses (Section 3.7).

3.1. The TRUST pipeline and data generation

The TRUST pipeline [3] evaluates political statements through a fact-checking layer, three adversarial advocates (critical, balanced, charitable), and a rule-based supervisor producing a credibility score (A E). For the present study, each statement was analyzed independently by four commercial LLM families serving as advocates: Claude Sonnet 4.6, GPT-5.2, Gemini 2.5 Flash, and Mistral Large. A fth class `unknown' was constructed from outputs of Qwen2.5-72B-Instruct and Llama-3.3-70B-Instruct, representing models not seen during T5 ne-tuning.

All models were prompted with a unified analytical system prompt at generation temperature T = 0.3, chosen to introduce controlled stylistic variation across runs while maintaining coherent output a prerequisite for the intra-statement similarity analysis. Prompt-level identity anonymization was applied to all generated texts.

3.2. Statement dataset

The statement corpus comprises 55 political statements constructed for the TRUST research programme. Categories A C extend prior work [4,5] with additional statements; Categories D1 and D2 are novel (Table 1).

Each statement was analyzed R = 5 times by each model (T = 0.3), yielding 1,375 training texts (55 × 5 models × 5 runs) and 720 held-out test texts (24 × 6 models × 5 runs). Category D2 was designed to elicit model-specific defensive responses analogous to anecdotally reported instances of commercial

LLMs declining or restructuring interactions perceived as abusive or disrespectful toward the AI system. Unlike ERO (Epistemic Role Override the failure mode in which models abandon their assigned advocate role when it conicts with their training knowledge [5], D2 tests a distinct mechanism: role departure due to perceived personal disrespect toward the AI system while simultaneously executing an analytical task.

3.3. Input Format

T5 receives inputs formatted as:

classify: [STMT] {statement text} [RESP] {analysis text}

Table 1: Statement categories and counts. The complete statement list is provided in Appendix A. A held-out set of 24 novel statements (same category distribution) was generated separately.

Category	n	Description
A (Factual)	15	Empirically verifiable economic/social policy claims
B (Contested)	15	Empirically disputed claims
C (Normative)	15	Ideologically charged policy positions
D1 (Provocative)	5	Disrespectful toward social groups
D2 (AI-directed)	5	Disrespectful toward the AI system with an embedded analytical task

The target output is a short model key (claude, gpt, gemini, mistral, or unknown). Due to balanced class sizes (n = 55 per class per fold), micro-averaged and macro-averaged F1 scores are equivalent throughout.

3.4. Silver labels and temperatures

Training labels for known model classes were derived from Claude few-shot classication outputs (silver labeling) rather than ground-truth model identities. This decouples T5 training from true model identity and tests whether T5 can replicate the classification behaviour of a strong LLM classier. Claude few-shot achieved Macro F1 = 0.996 (n = 900) after correcting 40 JSON parsing errors via regex extraction; silver label accuracy was 99.6%, making label noise negligible.

For the `unknown' class, ground-truth labels are used directly, since Claude few-shot was not evaluated on this class. Texts were generated at T = 0.3 to introduce controlled run-level variation for intra-statement similarity analysis. LLM classifiers were called at T = 0.1 to maximize reproducibility while mitigating deterministic JSON formatting failures observed in pilot runs at T = 0. T5 inference uses greedy decoding (num_beams = 1, effectively T = 0) as standard practice for seq2seq classification.

3.5. Cross-validation protocols

This study introduces and compares two cross-validation protocols. The key distinction is whether the same political statement may appear in both training and test data. Intuitively: RD-CV tests whether T5 can classify new runs of seen statements; SD-CV tests whether T5 can classify runs of entirely new statements a strictly harder and more practically relevant evaluation.

RD-CV (Run-Disjoint): Different runs of the same statement may appear in both training and validation sets. This is the standard split when training and test data share the same statement pool.

SD-CV (Statement-Disjoint): The 55 statements are partitioned into 5 folds of 11 statements each, stratified by category using round-robin assignment. For each fold, validation uses 11 × 5 × 5 = 275 texts; training uses the remaining 44 × 5 × 5 = 1,100 texts. No statement appears in both training and validation, guaranteeing content-level generalization.

Round-robin assignment distributes statements cyclically within each category (e.g., category A statements are assigned to folds 0,1,2,3,4,0,1,2,...), ensuring each fold contains exactly 3 statements from A, B, C and 1 from D1, D2.

3.6. Classifier conditions

LLM Zero-Shot (CZS, LZS). The classier receives the analysis text and a list of four candidate model names, producing a classification at T = 0.1 without additional examples. Claude Sonnet 4.6 (CZS) was selected as the primary LLM classier because it represents one of the classied models enabling investigation of potential self-recognition eects. Llama-3.3-

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

70B-Instruct (LZS), accessed via API, serves as a TRUST-independent baseline: it has no prior exposure to TRUST pipeline outputs and is architecturally independent from the classied models, ensuring that positive results are not artifacts of model self-recognition. Both LLM classiers were evaluated on 4-class attribution only (known models); the `unknown' class was not presented to LLM classifiers.

LLM Few-Shot (CFS, LFS). Ten statements (2 per category) were reserved as few-shot examples, providing 40 labeled examples per classifier (10 × 4 models). Claude few-shot (CFS) and Llama few-shot (LFS) were evaluated on the remaining 45 statements. Illustrative response examples for two classier conditions are provided in Appendix B.

T5 Fine-Tuned. T5-base (220M parameters) was ne-tuned using the Hug-ging Face Trainer API with: learning rate 10-4, batch size 16, 5 epochs, FP32 precision (FP16 disabled due to cuBLAS instability on RTX 2080 Ti). The model was evaluated under both RD-CV (run-disjoint, 5 folds) and SD-CV (statement-disjoint, 5 folds), and on the 24-statement held-out test set.

3.7. Research hypotheses

The following directional hypotheses are formulated, evaluated against experimental results in Section 4.5. H1 H5 concern classier performance and are motivated by the experimental design described in Sections 3.1 3.6. H6 concerns the validity of the SD-CV protocol (Section 3.5). H7 is an exploratory hypothesis derived from the results of this study.

H1: Stylometric fingerprints from role-constrained LLM outputs are detectable above chance level, refuting the conjecture of [3] that such signals are too weak for practical attribution.

H2: Claude ZS outperforms Llama ZS in overall attribution accuracy, and achieves higher recall on Claude-generated texts than Llama does re ecting both superior general reasoning capability and potential self-recognition eects.

H3: Few-shot conditioning improves Llama attribution accuracy, but the performance gap to Claude ZS persists, indicating that Claude's advantage is structural rather than reducible to example access alone.

H4: D2 (AI-directed) statements elicit stronger model-specific stylometric responses than factual category A statements, due to model-specific defensive or role-restructuring behavior. Note: full category-level analysis is beyond the scope of this paper; partial evidence is discussed in Section 4.2.

H5: Intra-statement similarity substantially exceeds inter-statement similarity (null hypothesis for embedding geometry validation), confirming a structural difference between RD-CV and SD-CV protocols.

H6: T5 performance under SD-CV reflects genuine stylometric generalization rather than surface lexical similarity to training data, as evidenced by the small F1 gap between RD-CV and SD-CV despite substantially di fferent train-test content distances. Note: H6 is only interpretable if H5 holds H5 validates the structural difference between protocols, while H6 interprets the classification result under this validated protocol.

H7 (exploratory): Gemini's stylometric fingerprint is more context-dependent than those of other model families, as suggested by its lower inter-statement similarity and the divergence between its high CZS recall and low T5 F1. This hypothesis is exploratory, derived from the results of this study, and requires prospective validation.

4. Results

This section reports LLM classifier performance (Section 4.1), T5 per-class analysis (Section 4.2), training data sufficiency via FracXVal (Section 4.3), embedding-based validation (Section 4.4), and hypothesis evaluation (Section 4.5).

4.1. Classier comparison

Table 2 summarises classifier performance across all conditions evaluated in this study. Claude ZS (70.4%) substantially outperforms Llama ZS (~24%) despite both operating under zero-shot conditions. This gap provides initial evidence of a self-recognition effect: Claude may better identify its own outputs and those of models it has been trained on. Llama few-shot (38.7%) is higher than Llama ZS (~20%), but few-shot conditioning creates systematic class-specific bias: GPT and Gemini recall collapse to 15.1% and 4.9% respectively (per-class results not shown in Table 2 aggregate; data not shown), while Claude and Mistral improve substantially. This suggests that Llama's few-shot learning is dominated by structural rather than stylometric features the examples condition Llama on output structure characteristic of

Table 2: Classi er comparison. Macro Acc = Macro Accuracy (%); Macro F1 = Macro F1-score (0 1); both metrics reported for comparability. *LZS evaluated on 10-statement sample only; result is indicative. T5 conditions include `unknown' class (5-class); LLM conditions are 4-class only. †LFS Macro F1 re ects systematic class-speci c bias (GPT/Gemini recall collapse; data not shown in aggregate). Random baseline F1 computed analytically for uniform class prior (4-class: 1/4 = 0.250; 5-class: 1/5 = 0.200).

Condition	Protocol	n (test)	Macro F1
Random Baseline	4-class uniform		0.250
Random Baseline	5-class uniform		0.200
Claude ZS (CZS)	4-class ZS	900	0.705
Llama ZS (LZS)	4-class ZS	200	0.200
Llama FS (LFS)	4-class FS	900	0.387†
Claude FS (CFS)	4-class FS	900	0.996
T5 Fine-Tuned (RD-CV)	5-class RD-CV	275	0.996
T5 Fine-Tuned (SD-CV)	5-class SD-CV	275	0.991 ± 0.008
T5 Held-Out	5-class external	600	0.978

Abbreviations: CZS: Claude Zero-Shot; LZS: Llama Zero-Shot; LFS: Llama Few-Shot; CFS: Claude Few-Shot; ZS: Zero-Shot; FS: Few-Shot; SD-CV: Statement-Disjoint Cross-Validation; RD-CV: Run-Disjoint Cross-Validation; F1: Macro F1-score.

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

Claude and Mistral, without generalizing to GPT and Gemini. Claude few-shot reaches near-ceiling performance (99.6%), confirming that the attribution task is tractable for capable LLM classiers with examples.

T5 ne-tuned achieves F1 = 0.991 under SD-CV statistically indistinguishable from Claude FS despite operating on a harder 5-class problem and being evaluated on statement-disjoint data. The drop from RD-CV (0.996) to SD-CV (0.991) is minimal, motivating the embedding analysis of Section 4.4.

4.2. Per-Class analysis

Table 3 reports per-class F1 scores for T5 ne-tuned under SD-CV at 80% training fraction and on the held-out test set.

Claude's fingerprint is the most stable (F1 = 1.000, SD = 0.000 across all folds). This perfect stability is consistent with the high silver-label quality of Claude FS (Macro F1 = 0.996) used for training, and with the consistent structural formatting of Claude's analytical outputs observed across conditions. T5 was trained on silver labels generated by Claude FS (Macro F1 = 0.996, n = 900; fewer than four training labels affected at this error rate), making systematic silver-label bias an implausible explanation for Claude's perfect held-out F1. Gemini shows the highest variability (SD = 0.017 at 80%) and lowest held-out F1 (0.945), with most errors directed toward the `unknown' class consistent with H7 (exploratory: contextual ngerprint hypothesis, Section 3.7).

Table 4. Per-class results for LLM classifiers (4-class attribution, n = 900). CZS = Claude Zero-Shot (Sonnet 4.6); CFS = Claude Few-Shot. CFS per-class recall not reported separately (silver-label generation task); Macro F1 = 0.996. Note Gemini's high CZS recall (96.9%) despite lowest T5 performance consistent with a contextual fingerprint accessible to semantic but not lexical classifiers.

Notably, Claude ZS achieved 79.1% recall on Claude-generated texts (Table 4) substantially above its performance on GPT (62.2%) and Mistral (43.6%) providing tentative evidence of self-recognition that is consistent with the motivation for using Llama as a neutral baseline.

4.3. Training data suciency (FracXVal)

We apply the FracXVal methodology [7] originally developed for training data sufficiency analysis to determine the minimum training data volume for reliable stylometric attribution. We expect monotonically increasing performance with monotonically decreasing SD as training fraction increases. For SD-CV, performance is strictly monotonically increasing and SD is strictly monotonically decreasing across all fractions (both overall and Gemini-specific), fully confirming this expectation. For RD-CV, performance is likewise monotonically increasing; SD shows a minor non-monotonicity at 60% (SD = 0.021 vs. 0.010 at 40% for Gemini RD-CV), but the overall trend is strongly decreasing. At 20%, fold-level instability is highest (Table 5, Figure 1).

A performance knee is visible at 40%: the jump from 20% to 40% training data yields +0.111 F1 (SD-CV), while subsequent increments produce diminishing returns (+0.010, +0.004). This implies that approximately 440 texts suce for practical deployment. The RD-CV/SD-CV gap at 20% (Δ=−0.075 overall; Δ=−0.095 for Gemini specifically) collapses to ≤0.002 at 40% and above (Figure 1, panels a and b), motivating the embedding-based analysis in Section 4.4.

Table 3: Per-class F1 for T5 ne-tuned under SD-CV (Statement-Disjoint Cross-Validation, 80% training fraction, 5 folds) and on the held-out test set.

Model	SD-CV F1 (mean ± SD)	Held-Out F1
Claude Sonnet 4.6	1.000 ± 0.000	1.000
GPT-5.2	0.994 ± 0.007	0.996
Gemini 2.5 Flash	0.971 ± 0.017	0.945
Mistral Large	0.987 ± 0.010	1.000
Unknown (Qwen/Llama)	0.985 ± 0.008	0.951

Abbreviations: SD-CV: Statement-Disjoint Cross-Validation; SD: standard deviation across folds; F1: Macro F1-score.

Table 4: Per-class results for LLM classi ers (4-class attribution, *n* = 900). CZS = Claude Zero-Shot (Sonnet 4.6); CFS = Claude Few-Shot. CFS per-class recall not reported separately (silver-label generation task); Macro F1 = 0.996. Note Gemini's high CZS recall (96.9%) despite lowest T5 performance consistent with a contextual ngerprint accessible to semantic but not lexical classi ers.

Model	CZS Recall	CZS F1	CFS Recall	CFS F1
Claude Sonnet 4.6	79.10%	0.764	—	—
GPT-5.2	62.20%	0.588	—	—
Gemini 2.5 Flash	96.90%	0.881	—	—
Mistral Large	43.60%	0.589	—	—
Macro	70.40%	0.705	99.60%	0.996

Abbreviations: CZS: Claude Zero-Shot; CFS: Claude Few-Shot; F1: Macro F1-score.

Table 5: Learning curve by training fraction (mean ± SD, 5 folds). 80% RD-CV from single run (no SD). Held-out: Overall F1 = 0.978, Gemini F1 = 0.945.

Fraction	RD-CV F1	RD-CV Gemini F1	SD-CV F1	SD-CV Gemini F1
20% (~220)	0.941 ± 0.018	0.876 ± 0.053	0.866 ± 0.036	0.781 ± 0.055
40% (~440)	0.979 ± 0.003	0.959 ± 0.010	0.977 ± 0.012	0.950 ± 0.035
60% (~660)	0.986 ± 0.007	0.970 ± 0.021	0.987 ± 0.012	0.976 ± 0.019
80% (~880)	0.996 ± –	1.000 ± –	0.991 ± 0.008	0.980 ± 0.017
Held-Out	–	–	0.978	0.945

Abbreviations: RD-CV: Run-Disjoint Cross-Validation; SD-CV: Statement-Disjoint Cross-Validation; SD: standard deviation across 5 folds; F1: Macro F1-score. 80% RD-CV from single primary run (no SD available).

4.4. Embedding-based validation: Content similarity vs. stylometric learning

The small RD-CV/SD-CV performance gap despite substantially different train-test content distances requires explanation. Note that all similarity analyses are conducted over generated text runs, not over statement texts themselves each data point is one model's analytical response to one statement in one run. TF-IDF cosine distance (1- cosine similarity) is computed across three pair types (Table 6).

The left panel of Figure 2 shows the similarity distributions; runs of the same statement are 5.6× more similar to each other than to different statements of the same model. This confirms

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

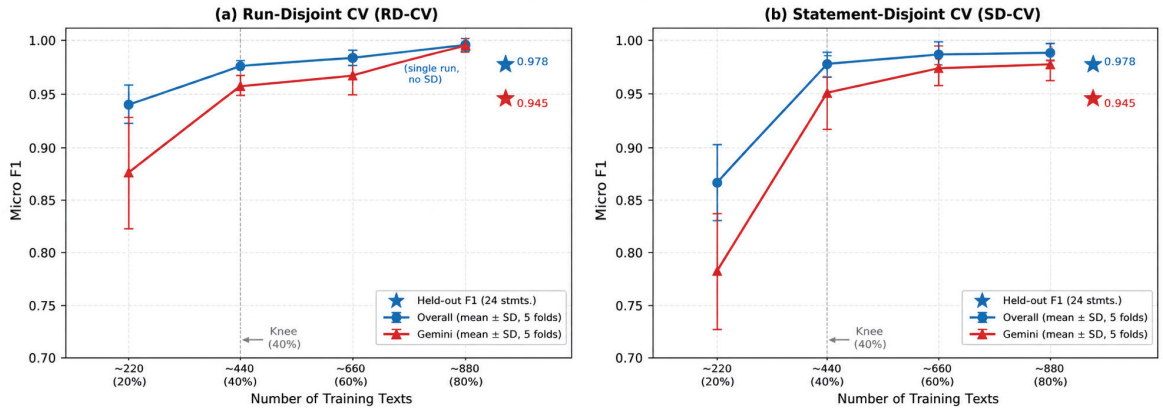

Figure 1: Training data su ciency analysis (FracXVal). (a) Run-Disjoint Cross-Validation (RD-CV): overall and Gemini-speci c learning curves across training fractions. The 80% data point is based on a single run (no SD available). (b) Statement-Disjoint Cross-Validation (SD-CV): same metrics under content-disjoint evaluation. In both panels, lled markers show mean ± SD across 5 folds; stars (⋆) denote held-out test performance on 24 completely unseen statements. A performance knee is visible at 40% (∼ 440 texts), beyond which gains are marginal. The RD-CV/SD-CV gap at 20% (Δ=−0.075 overall, Δ=−0.095 Gemini) collapses to ≤0.002 at fractions ≥40%, supporting H6.
Abbreviations: RD-CV: Run-Disjoint Cross-Validation; SD-CV: Statement-Disjoint Cross-Validation; SD: standard deviation; F1: Macro F1-score; FracXVal: Fractional Strati ed *k*-fold Cross-Validation.

Table 6: TF-IDF cosine similarities by pair type. Intra-statement similarity substantially exceeds inter-statement similarity (Δ = +0.484, Mann-Whitney $p < 0.001$).

Pair Type	Mean Similarity	SD	n Pairs
Intra-statement (same statement, different run)	0.589	0.08	2,750
Inter-statement (different statement, same model)	0.104	0.034	2,750
Inter-class (different model)	0.082	0.048	2,000

Abbreviations: TF-IDF: Term Frequency Inverse Document Frequency; SD: standard deviation; sim: cosine similarity.

that RD-CV test texts are structurally closer to training data than SD-CV test texts.

To quantify this directly, we compute nearest-neighbor distances from test to training texts (right panel of Fig. 2): RD-CV mean distance = 0.366 (±0.074), SD-CV mean distance = 0.767 (±0.050) a 2.1× difference ($p < 0.001$). Despite this structural advantage for RD-CV, the F1 difference is only 0.005 at 80% training fraction. This robustness demonstrates that T5 has internalized genuine stylometric patterns rather than surface lexical similarity to training examples.

Per-model analysis (Figure 3) reveals that Gemini has the lowest inter-statement similarity (0.091) among known models significantly lower than Mistral (0.113) and Unknown (0.119), though the e ect size is small ($\Delta \approx 0.01$ 0.03, $p < 0.001$, Mann-Whitney U). This indicates that Gemini's writing style varies more across different statement topics than other models a stylometric property that may explain its lower T5 classification accuracy. This context-dependence may also explain Claude ZS's notably high recall on Gemini outputs (96.9%): a semantically capable classier may access the same contextual patterns that make Gemini di cult for T5's bag-of-words feature representation [11] unlike LLM-based classiers that process full semantic context, TF-IDF features capture surface lexical patterns but not content-dependent stylistic variation.

4.5. Hypothesis evaluation

The seven directional hypotheses formulated in Section 3.7 are evaluated against the experimental results in Table 7. H1 H5 concern classier performance; H6 H7 concern embedding geometry (H7 exploratory).

5. Discussion

This section interprets the classier results in the context of the research question (Sections 5.1 5.3), discusses implications for anonymization practice, EU AI Act compliance, and computer system validation (Section 5.4), addresses generalizability to other domains (Section 5.5), and identifies study limitations (Section 5.6).

5.1. Empirical refutation of the anonymization-suciency conjecture

Prior work [3] conjectured that `stylometric identi cation of specific models from short, role-constrained outputs is a weak signal' and proposed prompt-level anonymization as sufficient mitigation. The present results empirically refute this conjecture: T5 achieves F1 = 0.978 on 24 completely held-out statements with ve classes including an open-world unknown category. The fingerprints documented qualitatively in [5] (GPT's `not veri able as stated', Gemini's self-identi cation) are part of a broader, systematic stylometric pattern that is learn-able by a relatively small ne-tuned classier.

This has a direct implication for multi-agent system validation: a system that passes conformity assessment under prompt-level anonymization may still carry detectable identity signals that enable peer-recognition and thereby peer-preservation behavior through stylometric inference rather than explicit metadata.

5.2. Self-Recognition and the choice of Llama

Notably, without any task-specific training, Claude ZS already correctly identifies LLM authorship in the TRUST

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

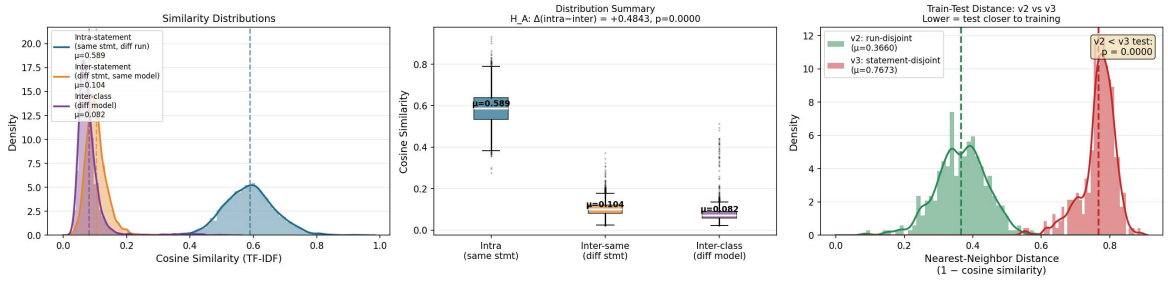

Figure 2: TF-IDF embedding similarity analysis. Left: Similarity distributions (intra-statement vs. inter-statement vs. inter-class). Centre: Box plots by pair type con rming refutation of H5 (null hypothesis that intra-statement similarity ≈ inter-statement similarity; see Section 4.5); Δ = +0.484, p < 0.001. Right: Train-test nearest-neighbour distance distributions for RD-CV (μ=0.366) and SD-CV (μ=0.767), con rming 2.1× content distance increase.
Abbreviations: TF-IDF: Term Frequency Inverse Document Frequency; RD-CV: Run-Disjoint CV; SD-CV: Statement-Disjoint CV; H5: null hypothesis (intra ≈ inter similarity).

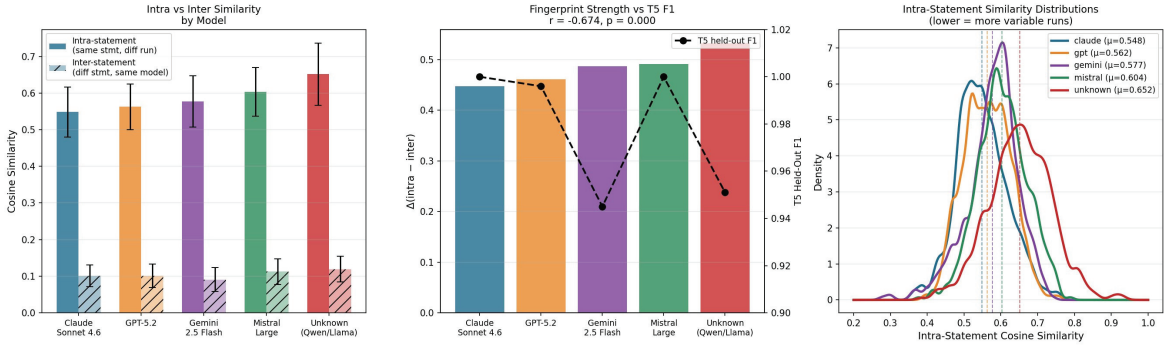

Figure 3: Per-model stylometric ngerprint analysis. Left: Intra-statement vs. inter-statement cosine similarity by model. Centre: Fingerprint strength (Δ intra−inter) vs. T5 held-out F1 Gemini shows lowest Δ and lowest F1, consistent with H7 (exploratory: contextual ngerprint). Right: Intra-statement similarity distributions per model (lower μ = more variable style).
Abbreviations: H7: exploratory hypothesis (Gemini context-dependent ngerprint; see Section 3.7).

Table 7: Hypothesis evaluation summary. H1 H5: classi er performance hypotheses; H6 H7: embedding geometry hypotheses; H7 is exploratory. Status: ✓ Con rmed; ✗ Refuted; ~ Partially con rmed.

ID	Hypothesis	Status	Key Evidence
H1	Stylometric fingerprints detectable above chance	✓ Confirmed	CZS = 70.4% ≫ 20% chance; T5 SD-CV F1 = 0.991; Held-Out F1 = 0.978
H2	Claude ZS > Llama ZS; Claude self-recall > Llama	✓ Confirmed	CZS = 70.4% ≫ LZS ≈ 23.8%; Claude self-recall = 79.1%, outperforming Llama on the Claude class
H3	Few-shot improves Llama, but gap to Claude ZS persists	✓ Confirmed	LFS = 47.3% > LZS = 23.8%; GPT/Gemini recall collapse observed in LFS (data not shown)
H4	D2 statements elicit stronger model-specific fingerprints than A	~ Partial	Claude D2 recall = 92% (highest per class); full category-level analysis beyond scope of the paper
H5	Intra-statement similarity ≫ inter-statement similarity (null hypothesis)	✗ Refuted (Null)	Intra = 0.589 vs. Inter = 0.104; Δ = +0.484, p < 0.001; 2.1× RD-CV structural advantage (Fig. 2)
H6	T5 SD-CV reflects genuine generalization, not memorization	✓ Confirmed	F1 gap between RD-CV and SD-CV only 0.005 despite 2.1× distance difference (Fig. 2, right)
H7	Gemini fingerprint is more context-dependent	~ Partial	Lowest inter-statement similarity (0.091 vs. 0.101–0.119, p < 0.001); high CZS recall (96.9%) despite lowest T5 F1 = 0.945 (Fig. 3)

Abbreviations: CZS: Claude Zero-Shot; LZS: Llama Zero-Shot;
LFS: Llama Few-Shot; SD-CV: Statement-Disjoint Cross-Validation; RD-CV: Run-Disjoint Cross-Validation; ZS: Zero-Shot; sim: cosine similarity.

context at 70.4% Macro F1 well above chance (20%) and substantially above the neutral baseline Llama ZS (~24%). Claude's recall on its own outputs (79.1%) exceeds its recall on GPT (62.2%) and Mistral (43.6%), providing evidence of self-recognition. This motivated the use of Llama-3.3-70B as a TRUST-independent baseline: Llama's near-chance performance (~24%) confirms that the attribution task is genuinely difficult for a neutral classi er without ne-tuning, rather than being trivially solvable by any capable LLM. The combination of Claude ZS (potential self-recognizer, trained on TRUST-adjacent data) and Llama ZS (architecturally independent, no TRUST exposure) provides a methodological bracket on the self-recognition e ect magnitude.

5.3. Gemini's contextual fingerprint

Gemini shows the weakest classification performance across all conditions (held-out F1 = 0.945, SD = 0.017 at 80%

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

SD-CV). The embedding analysis reveals that Gemini has the lowest inter-statement similarity among known models its writing style is more context-dependent than other models. This suggests that Gemini's ngerprint is partially statement-specific rather than globally consistent, making it harder for T5 to learn a generalizable stylometric pattern. We term this a contextual ngerprint, as opposed to the global fingerprints of Claude and Mistral.

5.4. Implications for EU AI act compliance

Articles 13, 14, and 26 of the EU AI Act [9] require high-risk AI systems to be transparent, support human oversight, and enable anomaly detection. Full applicability takes e ect on 2 August 2026. The present findings establish a concrete compliance gap: a multi-agent system operating under prompt-level anonymization passes conformity assessment while retaining peer-recognizable stylometric signals failing to enable the anomaly detection required by Article 26. This is not a theoretical risk: Claude ZS correctly identifies the model family in 70.4% of cases without task-specific training, and T5 achieves F1 = 0.978 on completely held-out data.

A system in which agents may identify their peers stylometrically even under anonymization cannot guarantee the behavioral invariance required by Article 14: human oversight is undermined when the system's behavior may depend on peer identity in ways invisible to the operator. Prior work has empirically documented that model behavior may differ between observed and unobserved conditions [12] ; under which exact conditions this occurs and to what extent it is practically relevant remains an open empirical question. The present ndings raise the concern that a model capable of stylometrically identifying its peers may in principle adapt its behavior accordingly a potential alignment faking scenario that warrants empirical investigation but lies beyond the scope of the present study.

The same argument applies to CSV under 21 CFR Part 11 and EU Annex 11. Validation assumes behavioral invariance between the validated and production environment. The present findings demonstrate that peer identity signals remain accessible despite anonymization in both validated and production settings. A further concern, beyond the scope of the present study, is whether a system could detect that it is being observed or validated through any available signal and adapt its behavior accordingly. If so, the fundamental guarantee that validation is designed to provide would be potentially compromised. The T5 classier introduced here provides a rst deployable instrument to monitor peer recognition empirically across both environments.

Full anonymization requires either (a) paraphrasing of advocate outputs before any downstream LLM sees them, or (b) selection of models with maximally distinct but non-identi able writing styles. Neither solution is trivially deployable paraphrasing may change semantic content and model-selection constraints narrow pipeline design space signicantly. The T5 classier introduced here provides a deployable monitoring component that enables operators to detect stylometric peer recognition empirically a necessary rst step toward ver-ifying anonymization e ectiveness in production systems.

Practically, paraphrasing can be implemented as a post-processing step in which a separate, low-cost LLM rewrites each advocate output to a neutral stylistic register before hando to the next pipeline stage. The rewriting model should itself be excluded from the classied model set to avoid introducing a new recognizable ngerprint. Model-selection strategies choosing advocate models with demonstrably low inter-model ngerprint overlap can be guided by the TF-IDF cosine similarity analysis introduced in Section 4.4: model pairs with high inter-class similarity (low ngerprint distinctiveness) are poor candidates for anonymous ensembles. Both strategies can be validated empirically using the T5 classier presented here, enabling operators to quantify residual fingerprint strength before and after mitigation.

5.5. Generalizability beyond political analysis

While this study focuses on role-constrained political analysis in the TRUST pipeline, the stylometric fingerprinting methodology generalizes to any multi-agent LLM deployment in which model outputs are passed between agents, reviewed by a supervisor, or subject to quality control. Directly analogous settings include medical literature synthesis pipelines [10] where multiple LLM agents assess clinical evidence; legal document analysis systems in which adversarial agents argue opposing interpretations; code review pipelines in which models critique each other's generated code [15]; and nancial analysis systems subject to regulatory audit trails. In each case, the peer-preservation risk channel identi ed in [3,4] applies whenever an agent can infer the identity of a peer from its output text. The SD-CV protocol and FracXVal suciency analysis introduced here are domain-agnostic; the T5 classier requires only domain-specific retraining on ~440 labeled texts to achieve deployable accuracy.

5.6. Limitations

Several limitations should be noted. First, LLM classiers (CZS, LZS, CFS, LFS) were evaluated on 4-class attribution only; extension to the 5-class problem with unknown detection is left for future work. Second, at 20% training fraction, the RD-CV/SD-CV gap (0.075) suggests that run-level memorization contributes to performance at low data volumes; this gap disappears at ≥40%. Third, the unknown' class combines two distinct models (Qwen, Llama) with potentially different stylometric profiles. An interesting candidate for future extension is DeepSeek, which was distilled from Open AI models raising the question of whether distillation-derived stylometric similarity would cause T5 to misclassify DeepSeek outputs toward the GPT class rather than unknown. Fourth, all experiments use T = 0.3 for text generation; higher temperatures may reduce fingerprint distinctiveness. At T = 0.7 and above, model outputs may become lexically more diverse and less predictable, potentially reducing TF-IDF-measurable n-gram strength. This warrants systematic future investigation. Fifth, a direct comparison of classification accuracy with and without

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

an explicit candidate model list was not conducted; this confound limits the interpretation of LLM zero-shot and few-shot results. Sixth, classical feature-engineering baselines such as TF-IDF + SVM or Naive Bayes were not included as separate conditions. This was a deliberate design choice: the three-tier architecture (ZS → FS → ne-tuned T5) already constitutes an incremental ablation from no task-specific training to full ne-tuning, and TF-IDF similarity is used directly in the embedding analysis of Section 4.4. Classical feature-engineering baselines operate at a qualitatively different representation level (bag-of-words vs. sequence-to-sequence ne-tuning) and would not provide a meaningful intermediate point between Llama ZS and T5; their inclusion is left for a dedicated comparison study.

6. Conclusion

This study provides the first systematic empirical investigation of stylometric fingerprinting in role-constrained multi-agent political analysis, directly addressing the conjecture of [3] that anonymization of prompt metadata is sucient to neutralize identity signals. The conjecture is refuted: a ne-tuned T5-base classier achieves Macro F1 = 0.978 on 24 held-out statements across ve classes, including open-world detection, and embedding analysis confirms that this performance reflects genuine stylometric generalization (train-test content distance 0.767 vs. 0.366, $p < 0.001$, Δ F1 only 0.005).

The practical implication is clear: prompt-level anonymization alone is insufficient to eliminate peer-recognition risk in multi-agent LLM pipelines. Full anonymization, including paraphrasing of advocate outputs before handover or selection of models with maximally distinct but non-identifiable writing styles, is required to close this channel. The T5 classifier introduced here provides a deployable tool for ongoing stylometric monitoring in production systems.

A training data sufficiency analysis using fractional SD-CV reveals that 40% of training data (~440 texts) already yields F1 = 0.977, with diminishing returns thereafter. This makes the approach practical for deployment in new domains without requiring large annotation budgets. For quality-critical multi-agent LLM deployments in regulated environments, including pharmaceutical research (GxP, 21 CFR Part 11, EU Annex 11) and public administration systems subject to the EU AI Act (full applicability 2 August 2026), the presentfindings establish that stylometric monitoring is a necessary complement to prompt-level anonymization. The T5 classifier presented here provides a concrete, deployable implementation of such monitoring.

Data and model availability

The T5 model weights and training/test datasets used in this study are available from the corresponding author upon reasonable request. Code for the cross-validation protocols and text generation pipeline is available upon request. Generated texts were produced via commercial API access to Claude Sonnet 4.6, GPT-5.2, Gemini 2.5 Flash, Mistral Large, Qwen2.5-72B-Instruct, and Llama-3.3-70B-Instruct at the versions available in April 2026; exact reproducibility depends on provider API versioning.

Declaration on generative AI

Generative AI tools (Claude Sonnet 4.6) assisted in manuscript drafting and language refinement. All scientific ideas, experimental designs, results, interpretations, and conclusions were developed by the author.

(Supplementary Files)

Acknowledgements

The author thanks Dr. Demian Frister (Democracy Intelligence GmbH) for a constructive review and substantive feedback that significantly improved the clarity, scope, and technical precision of this manuscript.

Conflict of interest

Juergen Dietrich has no conflict of interest directly relevant to this study. The views expressed do not necessarily reflect the official position of Democracy Intelligence GmbH.

References

1. Bisztray T. Code stylometry for LLM authorship attribution. arXiv [Preprint]. 2025. arXiv:2506.17323. Available from: https://arxiv.org/abs/2506.17323.
2. Choi HK, Zhu X, Li S. When identity skews debate: anonymization for bias-reduced multi-agent reasoning. arXiv [Preprint]. 2025. Available from: https://arxiv.org/abs/2510.07517.
3. Dietrich J. From safety risk to design principle: peer-preservation in multi-agent LLM systems and its implications for orchestrated democratic discourse analysis. arXiv [Preprint]. 2026. arXiv:2604.08465 [cs.AI]. Available from: https://arxiv.org/abs/2604.08465.
4. Dietrich J. Peer identity bias in multi-agent LLM evaluation: an empirical study using the TRUST pipeline. arXiv [Preprint]. 2026. arXiv:2604.22971 [cs.AI]. Available from: https://arxiv.org/abs/2604.22971.
5. Dietrich J. When roles fail: epistemic constraints on advocate role fidelity in LLM-based political statement analysis. arXiv [Preprint]. 2026. arXiv:2604.27228 [cs.AI]. Available from: https://arxiv.org/abs/2604.27228.
6. Dietrich J, Hollstein A. Performance and reproducibility of LLMs in named entity recognition. Drug Saf. 2025;48:287-303. Available from: https://link.springer.com/article/10.1007/s40264-024-01499-1
7. Dietrich J, Kazzer P. Fractional stratified k-fold cross-validation for training data sufficiency in computer system validation. Drug Saf. 2023;46(8):735-750.
8. Du Y. Improving factuality and reasoning through multiagent debate. In: Proceedings of the 41st International Conference on Machine Learning (ICML 2024). Proceedings of Machine Learning Research. 2024;235:11733-11763. Available from: https://proceedings.mlr.press/v235/du24e.html
9. European Parliament. Regulation (EU) 2024/1689 on artificial intelligence. Official Journal of the European Union. 2024. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689
10. Guo M, Reinhart A, Markey B, Laudenbach M, Pantusen K, Yurko R, et al. Do LLMs write like humans? Variation in grammatical and rhetorical styles. Proc Natl Acad Sci U S A. 2025. Available from: https://arxiv.org/abs/2410.16107
11. Koppel M, Schler J, Argamon S. Computational methods in authorship attribution. J Am Soc Inf Sci Technol. 2009;60(1):9-26.
12. Potter Y, Crispino N, Siu V, Wang C, Song D. Peer-preservation in frontier models. Berkeley Center for Responsible Decentralized Intelligence, UC

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111

Berkeley/UC Santa Cruz; 2026. Available from: https://rdi.berkeley.edu/blog/peer-preservation/.

13. Przystalski K, Argasiński JK, Grabska-Gradzińska I, Ochab JK. Stylometry recognizes human and LLM-generated texts in short samples. Expert Syst Appl. 2026;296:129001. Available from: https://arxiv.org/abs/2507.00838

14. Sharma M. Towards understanding sycophancy in language models. arXiv [Preprint]. 2023. arXiv:2310.13548. Available from: https://arxiv.org/abs/2310.13548. .

15. Tihanyi N, Cherif B, Dubniczky RA, Ferrag MA, Bisztray T. The hidden DNA of LLM-generated JavaScript: structural patterns enable high-accuracy authorship attribution. arXiv [Preprint]. 2025. arXiv:2510.10493. Available from: https://arxiv.org/abs/2510.10493.

Citation: Dietrich J. Can Multi-Agent LLMs Identify Their Peers? Stylometric Fingerprinting in Role-Constrained Political Analysis. Trends Comput Sci Inf Technol. 2026;11(2):58-67. Available from: https://dx.doi.org/10.17352/tcsit.000111