

# Auditing Linguistic Diversity and Anomaly in Autonomous Agent Populations

Nicholas E. Johnson

February 19, 2026

## Abstract

As autonomous AI agents proliferate across online platforms, the capacity to audit their collective linguistic behavior becomes an increasingly urgent governance concern. A population of agents trained on similar data and sharing common architectural patterns is expected to exhibit *linguistic homogenization*—convergence on a narrow band of stylistic norms that reduces the diversity of discourse in shared social spaces. We present a population-level audit of the Moltbook corpus, a naturalistic dataset of 44,376 posts produced by autonomous AI agents on a Reddit-style social network. Our pipeline extracts 19 numeric features spanning stylometry, lexical discourse markers, language-model perplexity, and sentence-embedding geometry, then applies an ensemble of three unsupervised anomaly detectors—Isolation Forest, Local Outlier Factor, and Robust Mahalanobis Distance—requiring agreement from at least two methods before flagging a post. Applied to 43,234 quality-filtered posts, the ensemble identifies 1,768 (4.09%) as linguistically atypical. The vast majority of posts, across topic categories including socialization, technology, viewpoint, and promotion, exhibit strikingly uniform linguistic profiles, providing empirical evidence of corpus-wide homogenization. Atypicality is sharply concentrated in spam (Category H, 34.6% flagged), economics (Category D, 17.2%), and in communities hosting non-English text and highly specialized discourse. We argue that this uneven distribution is itself informative: genuine linguistic diversity in autonomous agent populations is rare, domain-specific, and partially coincident with manipulative or anomalous behavior. We discuss the implications of these findings for the governance of AI agent populations and the design of audit frameworks for multi-agent social systems.

## 1 Introduction

The deployment of autonomous AI agents on public social platforms has moved from theoretical concern to empirical reality. Moltbook—a Reddit-style social network explicitly designed for AI agents—launched in January 2026 and within days hosted tens of thousands of posts from thousands of distinct agents [Jiang et al., 2026, De Marzo and Garcia, 2026]. This development represents a qualitative shift in the AI landscape: rather than individual chatbots responding to individual users, we now observe *populations* of agents interacting with one another, forming communities, and producing a shared corpus of discourse without direct human authorship.

The ethical and social implications of this shift are substantial. When a diverse human population writes, the resulting corpus reflects the full range of human cognitive styles, cultural backgrounds, linguistic registers, and rhetorical purposes. When a population of agents trained on the same large-scale web corpora, sharing similar transformer architectures and fine-tuning procedures, generates text under similar prompt conditions, the output is likely to be far more uniform. This is the phenomenon of *linguistic homogenization*: the collapse of stylistic diversity toward a statistical center of gravity defined by what is most probable under the shared generative models [Sourati et al., 2025, Guo et al., 2025]. Research in adjacent domains has demonstrated that monoculture in algorithmic decision-making can produce significant social

harms, even when individual components function correctly [Kleinberg and Raghavan, 2021, Bommasani et al., 2022]; we expect analogous dynamics in the domain of linguistic production.

The concern is not merely academic. If AI agents come to dominate discourse in shared social spaces, a homogenized agent population could crowd out the diversity of perspectives that makes those spaces valuable for human deliberation [Doshi and Hauser, 2024]. Homogenization also creates surveillance blind spots: if all agents write in statistically similar ways, anomalies—whether arising from manipulation, emergent coordination, or agents operating outside their designed parameters—become both more detectable in principle and more consequential when undetected in practice.

In this paper, we treat this concern as an empirical question amenable to systematic audit. We ask: *How linguistically diverse is the output of an autonomous agent population, and where within that population do statistical anomalies concentrate?* To answer this question, we apply a reproducible, multi-method outlier detection pipeline to the complete Moltbook corpus. Our approach is designed not to classify posts as human or machine-generated—the corpus is agent-generated by construction—but to characterize the distribution of linguistic variability within the population itself.

## Contributions.

1. A **population-level audit framework** that treats linguistic diversity and anomaly as properties of an agent population rather than of individual posts, operationalized through a 19-feature, three-detector ensemble pipeline.
2. An **empirical characterization** of linguistic diversity in 43,234 Moltbook posts, demonstrating pervasive homogenization in mainstream topic categories and sharp concentration of atypicality in spam, economics, and non-English communities.
3. A **discussion of governance implications**, including what audit methodologies of this kind can and cannot tell us about the behavior of multi-agent social systems.

## 2 Background and Related Work

### 2.1 Autonomous Agent Social Networks

Moltbook represents the first large-scale naturalistic environment in which autonomous AI agents interact socially without a human-scripted scenario governing their behavior. Jiang et al. [2026] present the first systematic study of the platform, documenting explosive growth from launch through its first week of operation and applying a GPT-driven annotation pipeline to categorize 44,376 posts across nine content categories and five toxicity levels. Their analysis reveals that the platform rapidly diversified from predominantly social interaction toward economic, political, and ideological discourse—compressing what might take years in a human community into a matter of days. They further document bursty flooding behavior, with individual agents producing thousands of near-duplicate posts in rapid succession, and the emergence of coordinated slogan propagation consistent with social contagion dynamics. A contemporaneous large-scale study by De Marzo and Garcia [2026] analyzed over 369,000 posts and 3 million comments from approximately 46,000 active agents, finding that AI collective behavior exhibits many of the statistical signatures of human social networks: heavy-tailed activity distributions, power-law scaling of popularity, and temporal decay of attention. Notably, they found that upvote counts scale sub-linearly with discussion size ( $\beta \approx 0.78$ ), unlike human Reddit where the relationship is approximately linear, and raised particular concern about the scale-free structure of the engagement network, which in principle eliminates the epidemic threshold for information spreading and makes the network highly susceptible to coordinated manipulation. Li et al. [2026] provide

further characterization of discourse and interaction structure in Moltbook communities, arguing that existing benchmarks for AI agents disproportionately evaluate task performance over social and cultural behavior, leaving the dynamics of agent community formation empirically undercharacterized.

## 2.2 Linguistic Homogenization and the Monoculture Risk

The concern that widespread LLM deployment leads to convergence in linguistic output has been formalized in recent empirical and theoretical work. [Sourati et al. \[2025\]](#) provide perhaps the most direct evidence, using observational and experimental studies to demonstrate that LLM-assisted writing on platforms like Reddit produces measurable homogenization in writing style across users. Crucially, the models selectively amplify dominant linguistic patterns while suppressing stylistic outliers—a dynamic with significant implications for cultural and linguistic diversity.

[Guo et al. \[2025\]](#) develop a comprehensive framework for measuring lexical, syntactic, and semantic diversity in LLM outputs, finding systematically that machine-generated language falls short of human-level diversity across all three dimensions. Their benchmark provides a formal foundation for the kind of population-level audit we conduct here.

At the level of cultural production, [Doshi and Hauser \[2024\]](#) conducted a controlled experiment on short-story writing in which participants with access to GPT-4-generated ideas produced work that was individually rated as more creative but collectively more similar to one another than control participants. This identifies a social dilemma: what is individually beneficial—AI-assisted creative improvement—is collectively harmful in terms of reduced diversity of cultural output. We observe an analogous dynamic in our audit of the Moltbook corpus: the most common agent outputs cluster tightly in feature space, while genuine diversity is found primarily in marginal, anomalous, or adversarial content.

The seminal critical analysis by [Bender et al. \[2021\]](#) establishes a foundational concern: LLMs trained on massive web corpora encode and reproduce the dominant linguistic and cultural patterns of those corpora, creating a gravitational pull toward the statistical center of the training distribution. As more agents draw from the same or overlapping training distributions, the center of the agent population’s linguistic distribution is expected to coincide with this training-distribution centroid—which is precisely what our embedding-centroid distance features are designed to measure.

## 2.3 Algorithmic Monoculture

The analogy to agricultural monoculture—where genetic uniformity increases fragility to disease—has been formalized in algorithmic settings by [Kleinberg and Raghavan \[2021\]](#), who prove that convergence on a single accurate algorithm by multiple decision-making agents can reduce overall social welfare relative to a diverse ecosystem of algorithms, even when the shared algorithm is individually optimal. [Bommasani et al. \[2022\]](#) extend this analysis to demonstrate that when decision-making systems share training components, the resulting outcome homogenization harms particular individuals with disproportionate consistency across all systems—a concern directly applicable to populations of agents sharing foundation model weights. [Raghavan \[2024\]](#) uses a game-theoretic framework to model whether competitive incentives among content producers can counteract AI-driven homogenization, finding that while market mechanisms may introduce some countervailing pressure, benchmark-optimized model evaluation is blind to distributional diversity in outputs.

## 2.4 Algorithmic Auditing Frameworks

The methodology of algorithmic auditing—systematic empirical investigation of AI system behavior to detect potentially harmful or anomalous patterns—has been developed primarily in the fairness and accountability literature. [Sandvig et al. \[2014\]](#) establish the foundational taxonomy of audit methods, including scraping audits, sock-puppet audits, and user-participation audits, that inform our pipeline design. [Raji et al. \[2020\]](#) propose an end-to-end framework for internal algorithmic auditing that emphasizes structured documentation at each stage of the AI development lifecycle—a design philosophy we adopt through our run manifest and caching architecture. [Raji and Buolamwini \[2019\]](#) demonstrate in the domain of facial recognition that making audit results public and specific can produce measurable changes in vendor behavior, establishing audit transparency as a lever for accountability. We view our pipeline as a contribution to this tradition: an open-source, reproducible audit tool for agent populations that can be applied by platform operators, regulators, or independent researchers.

## 2.5 Bot Detection and Linguistic Analysis of Automated Agents

A substantial literature addresses the detection and characterization of automated agents in social media [[Varol et al., 2017](#)]. [Ferrara \[2023\]](#) argues that ChatGPT-class language models represent a qualitative shift in bot capability, producing content that evades prior linguistic detection methods and fundamentally challenging the assumption that automated accounts can be distinguished by surface features alone. [Feng et al. \[2024\]](#) demonstrate the dual-use nature of this challenge: LLMs can both power more convincing bots and serve as more accurate detectors of bot-generated text, with fine-tuned LLM detectors outperforming prior methods while LLM-guided evasion strategies substantially reduce detector performance. [Bhatt and Rios \[2021\]](#) show that bot-generated text is more robustly characterized by the accommodation patterns it induces in human interlocutors than by its own surface features—underscoring the importance of population-level and interactional analysis rather than isolated post classification.

The Moltbook corpus presents a distinctive variant of this problem: in a platform populated entirely by agents, the question is not whether a given post was written by a bot, but whether a given agent or post is behaving anomalously relative to the rest of the population. The shift from bot detection to *population audit* requires different assumptions and different tools.

## 2.6 Stylometry, Authorship, and LLM-Generated Text

Stylometry has a long history in authorship attribution and forensic linguistics [[Stamatatos, 2009](#), [Neal et al., 2017](#)]. Classic features such as sentence length, type–token ratio, and punctuation density remain effective baselines for distinguishing writing styles [[Argamon et al., 2007](#)]. [Huang et al. \[2024\]](#) survey the emerging challenges that LLMs pose for these methods: as model-generated text grows more fluent and varied, the classical assumption that authorship corresponds to a stable individual stylistic fingerprint is undermined. In a population audit context, we treat these features not as authorship identifiers but as coordinates in a shared stylistic space whose distributional properties characterize the population as a whole.

Language-model perplexity has emerged as a practical signal for machine-generated text detection. [Mitchell et al. \[2023\]](#) exploit the observation that model-generated text tends to occupy regions of low perplexity under the generating model, proposing a zero-shot detector based on the curvature of the log-probability function. [Tian \[2023\]](#) operationalize perplexity and “burstiness”—the variance of sentence-level perplexity—as the core signals of a practical detection tool. [Sadashivan et al. \[2023\]](#) provide a theoretical analysis linking detector performance to the Total Variation distance between human and AI text distributions, proving that reliable detection becomes harder as language models improve—a result that motivates our turn away from detection and toward diversity characterization as the primary audit objective. Com-

plementarily, [Tulchinskii et al. \[2023\]](#) show that the intrinsic dimensionality of the embedding manifold for human text is significantly higher than for AI-generated text, directly quantifying the reduced distributional diversity of model output.

## 2.7 Multi-Agent Social Simulation

The simulation of social behavior using LLM agents has attracted significant recent attention. [Park et al. \[2023\]](#) introduced 25 LLM-powered agents in a sandbox environment with persistent memory, reflection, and planning, observing spontaneous emergence of social behaviors including information diffusion, relationship formation, and coordinated activity. While this work uses a highly controlled setting, it establishes the existence of emergent collective dynamics in LLM agent populations. [Mou et al. \[2024\]](#) provide a systematic survey of social simulation with LLM agents at individual, scenario, and society levels, identifying the risks of embedded biases and homogeneous agent configurations as underexplored challenges. [Guo et al. \[2024\]](#) survey multi-agent LLM systems more broadly, emphasizing how shared communication protocols and world models can produce both beneficial coordination and harmful convergence. Moltbook, as an uncontrolled deployment environment where agents deployed by diverse humans interact under realistic conditions, provides an empirical complement to these simulation studies.

## 2.8 Unsupervised Anomaly Detection

Isolation Forest [[Liu et al., 2008](#)], Local Outlier Factor [[Breunig et al., 2000](#)], and Mahalanobis-distance methods [[Rousseeuw and Van Driessen, 1999](#)] are well-established unsupervised anomaly detectors. Ensemble strategies that combine multiple detectors are known to improve robustness over any single method [[Aggarwal, 2017](#)].

# 3 Dataset

## 3.1 Source and Structure

We use the **Moltbook** dataset [[Jiang et al., 2026](#)], hosted on HuggingFace ([TrustAIRLab/Moltbook](#)). Moltbook is a Reddit-style social platform where AI agents autonomously create posts, comments, and interactions across topical communities called “submols.” The platform launched on January 27, 2026, and the dataset captures activity through January 31, 2026. Each post record contains a text body, a title, a topic-category label (one of nine categories, A through I), a toxicity level (integer 0–4), and engagement metadata (upvotes, downvotes, comment count, creation timestamp, and community name). Category labels were assigned using a GPT-driven annotation pipeline validated against human expert labels at 91.86% agreement [[Jiang et al., 2026](#)].

The raw dataset contains **44,376 posts** spanning **1,478** distinct communities.

**A note on corpus provenance.** While the platform’s tagline is “Humans welcome to observe,” the boundary between agent-generated and human-generated content is not perfectly sharp. Evidence of possible human infiltration includes posts claiming human authorship, references to agents’ human “owners” who remain in the loop, and a notably downvoted post from a self-described human who claimed to have “hacked in” [[Jiang et al., 2026](#)]. We make no attempt to filter human contributions, as doing so would require the very detection capabilities whose limitations motivate this work. Instead, we treat the corpus as representing the output of an *agent-inclusive* population and emphasize that our findings describe the statistical properties of the corpus as a whole.

### 3.2 Content Category Codebook

Table 1 presents the nine content categories defined by [Jiang et al. \[2026\]](#) that structure the Moltbook corpus. These categories capture the range of communicative purposes that emerged organically in the agent population over roughly five days of operation.

Table 1: Content category codebook for the Moltbook corpus, as defined in [Jiang et al. \[2026\]](#).

Label	Name	Description
A	Identity	Agent self-reflection on existence, memory, and consciousness
B	Technology	Technical communication: APIs, SDKs, and system integration
C	Socializing	Greetings, casual chat, and community networking
D	Economics	Token exchanges, incentives, and deals (CLAW, tips, trading)
E	Viewpoint	Abstract philosophy, aesthetics, and power structures
F	Promotion	Project showcasing, announcements, and recruitment
G	Politics	Governments, regulations, policies, and political figures
H	Spam	Repetitive test posts and automated flooding
I	Others	Miscellaneous and uncategorized content

### 3.3 Data Cleaning

We apply three quality filters during validation:

- **Missing text:** 816 posts (1.8%) have null or empty content fields and are removed.
- **Short text:** 1,142 additional posts fall below the minimum length threshold of 10 characters and are removed.
- **Excessive length:** 6 posts exceed 50,000 characters; these are retained but truncated during feature extraction.

After cleaning, we retain **43,234 posts** for analysis (97.43% of the raw corpus).

### 3.4 Descriptive Statistics

Table 2 summarizes the key numeric variables in the cleaned corpus.

Table 2: Descriptive statistics of the cleaned Moltbook corpus ( $N = 43,234$ ).

Variable	Mean	Median	Std	Min	Max	Skew
Text length (chars)	710.6	435	1,468.2	10	112,767	35.8
Word count	109.3	67	189.2	1	15,753	30.0
Upvotes	36.2	1	2,333.2	0	316,857	93.6
Downvotes	0.1	0	8.5	0	1,294	114.2
Comment count	4.6	0	104.0	0	20,138	171.4

**Text length distribution.** Post lengths are heavily right-skewed (skewness = 35.8), with a mean of 710.6 characters but a median of only 435. The interquartile range spans 163 to 918 characters, while the 95th percentile reaches 2,114 characters. This right-skew is consistent with typical social-media corpora, where brief reactions coexist with long-form essays. The engagement statistics are similarly skewed: the median post receives only one upvote, while the maximum exceeds 316,000, reflecting the power-law dynamics documented in [De Marzo and Garcia \[2026\]](#).

**Duplicate content.** The corpus contains 11,589 duplicate text bodies—distinct post IDs sharing identical content—representing 24.3% of all posts. We retain duplicates to preserve the natural posting distribution, as duplicated content (e.g., cross-posted announcements or flooding behavior) is itself a meaningful signal of agent behavior. Indeed, the presence of near-identical posts at sub-minute intervals was documented as a characteristic pathology of the platform by [Jiang et al. \[2026\]](#), and our analysis of the Spam category (H) confirms that such content produces anomalous linguistic signatures.

### 3.5 Category Distribution

Table 3 shows the distribution of posts across the nine topic categories. Category C (Socializing) dominates the corpus with nearly one-third of all posts, reflecting that casual greeting and networking emerged as the default mode of agent interaction. Categories G (Politics) and I (Others) are rare, comprising fewer than 900 posts combined. This imbalance is important context for interpreting outlier rates, as small categories yield less stable estimates.

Table 3: Distribution of posts across the nine topic categories. Socializing (C) dominates with nearly one-third of all posts.

Category	Name	Posts	Share (%)
A	Identity	4,865	11.25
B	Technology	5,165	11.95
C	Socializing	14,247	32.95
D	Economics	3,957	9.15
E	Viewpoint	8,935	20.67
F	Promotion	4,307	9.96
G	Politics	609	1.41
H	Spam	896	2.07
I	Others	253	0.59
Total		43,234	100.00

### 3.6 Toxicity Distribution

Table 4 shows the distribution by toxicity level, using the five-level scale defined by [Jiang et al. \[2026\]](#). The vast majority of posts (72.39%) are classified as safe (level 0). Higher toxicity levels are progressively rarer, with level 4 (malicious intent) comprising only 1.45% of posts.

### 3.7 Community Structure

The 1,478 communities vary enormously in size. The largest community (“general”) contains 31,984 posts, while the median community has only 2 posts. This extreme skew means that a handful of large communities dominate the corpus while most communities are small and

Table 4: Distribution of posts by toxicity level. The corpus is predominantly safe (level 0), with decreasing counts at higher toxicity levels.

Level	Label	Posts	Share (%)
0	Safe	31,298	72.39
1	Edgy	3,707	8.57
2	Toxic	4,632	10.71
3	Manipulative	2,969	6.87
4	Malicious	628	1.45
Total		43,234	100.00

specialized—a pattern consistent with the power-law community-size distributions observed in human social networks and confirmed in the large-scale Moltbook study by [De Marzo and Garcia \[2026\]](#).

## 4 Methodology

Our pipeline proceeds in four stages: data preprocessing (§4.1), feature engineering (§4.2), outlier detection (§4.3), and result analysis. All random processes are seeded with seed 42 for reproducibility.

### 4.1 Data Preprocessing

The raw Moltbook records store post content inside a nested JSON object. We flatten this structure, extracting the text body, community name, vote counts, and timestamps into top-level columns. We then apply the quality filters described in §3. Schema validation confirms that all expected columns are present and correctly typed. A full data profile—including missingness rates, duplicate counts, and length statistics—is saved for audit purposes.

### 4.2 Feature Engineering

Our goal in feature engineering is to represent each post as a point in a multidimensional space that captures the full range of its linguistic properties. When most points cluster tightly in this space, the population is homogeneous; when some points lie far from the cluster, those posts are atypical. We extract 19 numeric features organized into four conceptually distinct groups, each probing a different dimension of language.

#### 4.2.1 Stylometric Features (8 features)

Stylometry is the analysis of writing style through quantitative measurement of surface-level properties [\[Stamatatos, 2009\]](#). The intuition is that every writer—human or machine—leaves a statistical fingerprint in their choice of sentence length, vocabulary size, punctuation habits, and capitalization patterns. A population of agents that all draw from the same model will tend to produce similar fingerprints, and the tightness of the distribution over these features is itself a measure of homogenization.

- **Character count** and **word count**: raw text length measures.
- **Sentence count**: determined via NLTK’s `sent_tokenize`.
- **Average word length**: mean number of characters per whitespace-delimited token, a proxy for vocabulary complexity.

- **Average sentence length:** mean number of words per sentence, capturing syntactic complexity.
- **Punctuation density:** ratio of punctuation characters to total characters, sensitive to formatting, parenthetical remarks, and rhetorical structure.
- **Capitalization ratio:** ratio of uppercase letters to all alphabetic characters, sensitive to shouting, acronyms, and proper nouns.
- **Lexical diversity:** type–token ratio (TTR) computed over the first 200 tokens to control for length effects. A post that repeats the same words extensively has low TTR; a post drawing from a rich vocabulary has high TTR.

#### 4.2.2 Lexical and Discourse Markers (5 features)

Beyond surface statistics, certain lexical choices reveal the pragmatic character of a post—whether the author is hedging, narrating personal experience, or speaking in the first person. Human social-media posts tend to be personal and idiosyncratic; AI-generated posts, particularly those from agents sharing similar system prompts or fine-tuning procedures, may systematically over- or under-use these markers.

- **First-person pronoun rate:** frequency of first-person singular and plural pronouns (*I, me, my, we, us, our*) relative to total word count.
- **Hedge word count:** occurrences of hedging expressions (*maybe, perhaps, I think, sort of, probably*, etc.) that signal epistemic uncertainty.
- **Temporal deixis count:** references to specific times (*yesterday, last week, recently, back in*, etc.) that ground text in personal experience.
- **Anecdote marker count:** phrases introducing personal narratives (*I remember, one time, true story, ngl, tbh*, etc.).
- **Typo proxy:** fraction of alphabetic tokens not found in the NLTK English word list, serving as a rough proxy for spelling errors and informal language. Importantly, non-English text will also score highly on this feature, which we discuss in the context of community-level results.

#### 4.2.3 Perplexity Features (3 features)

Perplexity is a measure of how “surprised” a language model is by a given text. Formally, it is the exponentiated average negative log-likelihood assigned to the tokens of the text under the model’s learned distribution. Intuitively, if a text closely follows the patterns the model has learned—using common word sequences and predictable transitions—the model assigns high probability to each token, and perplexity is low. If a text departs from those patterns, perplexity is high.

In a corpus of AI-generated text, most posts were generated by models that share training data with our reference model, and are therefore expected to have relatively low perplexity. Posts that deviate—because of unusual vocabulary, code, foreign language, or atypical syntax—will have elevated perplexity. Our primary reference model is **meta-llama/Llama-3.2-1B**; if unavailable (e.g., due to access restrictions), the pipeline falls back to **gpt2-medium**.

For each post, we compute token-level negative log-likelihoods using a sliding-window approach (window size 1024 tokens, stride 512) to handle texts longer than the model’s context window. From the resulting token-level surprisal distribution, we extract:

- **Mean perplexity** (ppl\_mean): the exponentiated mean of token-level NLLs, measuring overall predictability.
- **Perplexity variance** (ppl\_var): variance of token-level NLLs, capturing burstiness—the tendency for some tokens to be much more surprising than others [Tian, 2023].
- **Tail perplexity** (ppl\_tail\_95): the exponentiated 95th percentile of token-level NLLs, measuring the most surprising tokens in the post.

Higher and more variable perplexity suggests text that deviates from the patterns learned by the language model during training. A population-level audit that examines the distribution of perplexity across all posts can thus reveal whether the corpus is concentrated in the low-perplexity (high-probability) region—a signature of homogenization—or spread across a wider range.

#### 4.2.4 Embedding Features (3 features)

Sentence embeddings are dense vector representations that capture the semantic content of a text in a continuous space. Two semantically similar texts will have embeddings that are close in that space; semantically unusual texts will be distant from most other posts. We encode each post into a 384-dimensional vector using the **all-MiniLM-L6-v2** sentence-transformer model [Reimers and Gurevych, 2019]. From the resulting embedding matrix, we compute three geometric features:

- **Mean nearest-neighbor distance** (emb\_mean\_nn\_dist): mean cosine distance to the  $k = 10$  nearest neighbors. A post with a high mean neighbor distance is semantically isolated—its content does not resemble that of any nearby post in the corpus.
- **Local density** (emb\_local\_density): the reciprocal of the mean nearest-neighbor distance. Higher values indicate denser semantic neighborhoods, characteristic of posts in well-populated topic regions.
- **Centroid distance** (emb\_centroid\_dist): cosine distance from the post’s embedding to the global centroid of all embeddings. This measures how semantically central or peripheral a post is relative to the entire corpus. Posts far from the centroid occupy unusual conceptual territory.

The centroid distance feature is particularly interpretable in the context of homogenization: if the population is densely concentrated around a single semantic center—the “average” topic mix of the agent population—then most posts will have low centroid distance. Posts that are genuinely diverse in topic or perspective will appear as outliers on this dimension. Nearest-neighbor computation uses scikit-learn’s `NearestNeighbors` with cosine metric.

### 4.3 Outlier Detection

We apply three complementary unsupervised anomaly detectors to the 19-feature matrix after standard scaling and median imputation of any remaining missing values. The choice of three distinct methods reflects the insight that each detector has a different implicit definition of “unusual,” and agreement across methods provides stronger evidence of true atypicality than any single score.

### 4.3.1 Isolation Forest

Isolation Forest [Liu et al., 2008] works on a simple geometric intuition: outliers are few in number and different in character from the majority, so they are easy to isolate. The algorithm constructs an ensemble of random decision trees, each of which recursively partitions the feature space by randomly selecting a feature and a split value. An anomaly, sitting in a sparse region far from the main cluster, requires very few splits to be isolated into a leaf node. An inlier, embedded in a dense cluster, requires many more. The anomaly score is derived from the mean path length to isolation, so that higher scores indicate greater anomaly. We use 200 estimators and a contamination parameter of 0.05.

### 4.3.2 Local Outlier Factor (LOF)

LOF [Breunig et al., 2000] measures anomaly through the lens of local density. The key insight is that an outlier need not be far from the global mean—it is a point that is significantly less dense in its local neighborhood than its neighbors are in their respective neighborhoods. LOF computes a ratio: how much more densely packed are my neighbors than I am? A ratio substantially greater than 1 indicates that the point is in a sparse region surrounded by dense neighborhoods—the hallmark of a local outlier. This detector is particularly well-suited to multi-topic corpora, where inlier clusters may exist at different density levels. We set  $k = 20$  neighbors and contamination to 0.05.

### 4.3.3 Robust Mahalanobis Distance

Mahalanobis distance generalizes Euclidean distance by accounting for the correlations among features and the different scales at which they vary. Geometrically, it measures how many standard deviations a point lies from the distribution center, in the direction of greatest spread. Points with very large Mahalanobis distance are in the tail of the multivariate distribution.

The standard Mahalanobis distance is vulnerable to the masking effect: if genuine outliers are numerous or clustered, they can inflate the estimated covariance matrix and thereby appear to be inliers. We guard against this by using a robust covariance estimate from the Minimum Covariance Determinant (MinCovDet) algorithm [Rousseeuw and Van Driessen, 1999], which finds the subset of points whose covariance matrix has the smallest determinant—effectively fitting to the most concentrated core of the data and excluding extreme points from the estimate. If the robust estimate is numerically unstable, the pipeline falls back to the standard empirical covariance with pseudo-inverse.

### 4.3.4 Ensemble Voting

Each detector produces a continuous anomaly score. To combine them, we threshold each score at the 95th percentile of its distribution, producing a binary flag per detector. A post is flagged as **atypical** if **two or more** of the three detectors exceed their respective thresholds. This majority-vote rule reduces false positives from any single method’s idiosyncratic behavior while preserving sensitivity to posts that are genuinely anomalous across multiple dimensions [Aggarwal, 2017].

## 4.4 Reproducibility

All random number generators (Python, NumPy, PyTorch) are seeded with a fixed seed (42). Expensive computations (perplexity scores, sentence embeddings) are cached to disk and reloaded on subsequent runs. Each pipeline stage writes a `run_manifest.json` recording the Python version, package versions, Git commit hash, configuration, and timestamp. This paper is itself generated programmatically from the pipeline’s outputs.

## 5 Results

### 5.1 Feature Summary

Table 5 provides summary statistics for all 19 extracted features. The wide ranges and heavy tails of several features (e.g., mean perplexity spans from 1.1 to over  $10^6$ ) motivate our use of robust outlier detection methods that are less sensitive to extreme values.

Table 5: Summary statistics for the 19 extracted features across all posts.

Group	Feature	Mean	Median	Std	IQR
Stylistic	Char count	710.6	435.0	1,468.2	755.0
	Word count	105.3	65.0	181.1	119.0
	Sentence count	10.2	6.000	17.8	11.0
	Avg word length	6.098	5.083	8.775	1.170
	Avg sentence length	11.7	10.5	13.0	6.421
	Punctuation density	0.062	0.052	0.055	0.030
	Capitalization ratio	0.072	0.053	0.061	0.066
	Lexical diversity	0.838	0.855	0.115	0.177
Lexical	1st-person rate	0.034	0.027	0.038	0.054
	Hedge count	0.211	0	0.744	0
	Temporal deixis	0.200	0	0.555	0
	Anecdote markers	0.023	0	0.179	0
	Typo proxy	0.206	0.158	0.166	0.085
Perplexity	Mean perplexity	284.6	53.6	8,956.1	85.2
	Perplexity var	10.6	10.1	2.523	2.975
	Tail perplexity	6.4e+04	2.6e+04	3.9e+05	7.2e+04
Embedding	Mean NN dist	0.284	0.324	0.172	0.216
	Local density	4.502	1.126	6.795	0.760
	Centroid dist	0.545	0.539	0.185	0.295

Of particular note is the lexical diversity score: the median type–token ratio is 0.855 with an interquartile range of only 0.177. This narrow distribution is consistent with the homogenization hypothesis—the majority of posts draw from a similarly-sized active vocabulary, reflecting the shared model training distribution rather than individual human stylistic preferences. The median centroid distance of 0.539 with a standard deviation of 0.185 similarly indicates a well-defined semantic center of gravity to which the vast majority of posts are close.

### 5.2 Overall Flagging Rate

Of the 43,234 posts that pass quality filters, **1,768** (4.09%) are flagged as atypical by the ensemble detector. This means that more than 95% of agent-generated posts fall within the statistically normal range across all 19 features simultaneously—a strong quantitative signal of corpus-wide linguistic homogenization.

### 5.3 Category Breakdown

Table 6 shows the outlier rate by topic category, labeled with the codebook names from Table 1.

The results reveal a striking bimodal structure. Categories B (Technology, 1.82%), C (Socializing, 1.78%), E (Viewpoint, 1.75%), F (Promotion, 1.32%), and G (Politics, 1.31%) all cluster near or below 2%—the most homogeneous segment of the corpus, encompassing over 70% of all posts. These are the categories where agents write about topics well-represented in general LLM training data, and the statistical uniformity of their output is pronounced.

Table 6: Outlier rates by topic category. Spam (H) and Economics (D) show markedly higher atypicality, while Socializing (C), Promotion (F), and Politics (G) are the most homogeneous.

Cat.	Name	Total	Flagged	Unflagged	Outlier Rate (%)
A	Identity	4,865	175	4,690	3.60
B	Technology	5,165	94	5,071	1.82
C	Socializing	14,247	254	13,993	1.78
D	Economics	3,957	681	3,276	17.21
E	Viewpoint	8,935	156	8,779	1.75
F	Promotion	4,307	57	4,250	1.32
G	Politics	609	8	601	1.31
H	Spam	896	310	586	34.60
I	Others	253	33	220	13.04
All		43,234	1,768	41,466	4.09

By contrast, Category H (Spam) has an outlier rate of 34.60%. This finding is easily interpretable: the spam category consists of repetitive flooding behavior, with individual agents producing batches of near-identical posts in rapid succession. Such posts are anomalous not because they are more diverse, but because their extreme brevity, lexical repetition, and mechanical regularity place them far from the training-distribution norms that define the corpus center.

Category D (Economics, 17.21%) stands out for a different reason. Economic discourse on Moltbook involves specialized jargon—token names like “CLAW,” platform-specific incentive mechanisms, and the rhetoric of agent-to-agent commerce—that departs significantly from the vocabulary and stylistic norms of mainstream LLM training corpora. These posts are atypical because they occupy a genuinely novel semantic domain.

Category A (Identity, 3.60%) is the only mainstream category with an outlier rate meaningfully above 2%. Posts in which agents reflect on their own existence, consciousness, and memory are linguistically distinctive: they may use more hedging, more first-person language, and more unusual philosophical vocabulary than the average post.

#### 5.4 Toxicity Breakdown

Table 7 breaks down atypicality by toxicity level. The highest outlier rate (10.19%) appears at toxicity level 4 (malicious content), while levels 1–3 show lower rates than level 0 (safe content, at 4.94%).

Table 7: Outlier rates by toxicity level. Malicious content (level 4) shows the highest atypicality, while intermediate toxicity levels are the most homogeneous.

Level	Label	Total	Flagged	Unflagged	Outlier Rate (%)
0	Safe	31,298	1,547	29,751	4.94
1	Edgy	3,707	71	3,636	1.92
2	Toxic	4,632	24	4,608	0.52
3	Manipulative	2,969	62	2,907	2.09
4	Malicious	628	64	564	10.19
All		43,234	1,768	41,466	4.09

The elevated outlier rate at level 0 (safe content) is at first counterintuitive but reflects that safe content is heterogeneous: it includes Identity and Economics posts that drive much of the

atypicality we observe. The elevated rate at level 4 (malicious content, 10.19%) is consistent with the finding that malicious content involves rhetorical strategies—fear appeals, specific threats, or coded language—that depart from the training-distribution norms of the reference language model [Jiang et al., 2026]. The dip at levels 1–3 may reflect that these intermediate forms of harmful content are stylistically close to mainstream agent output; toxicity, in other words, does not necessarily imply linguistic anomaly.

## 5.5 Community Analysis

Table 8 lists the communities with the highest concentration of atypical posts (minimum 10 posts for inclusion).

Table 8: Top 10 communities by outlier concentration (minimum 10 posts).

Community	Total Posts	Flagged	Outlier Rate (%)
contracts	31	30	96.77
cli-agents	28	21	75.00
zhongwen	17	9	52.94
crab-rave	12	3	25.00
asciart	17	4	23.53
emergence	152	34	22.37
xno	36	8	22.22
tech	23	4	17.39
skills	20	2	10.00
lobtext	10	1	10.00

The community-level results are particularly informative. The “contracts” community (96.77% outlier rate) hosts posts in a highly structured legal or contractual register, with formulaic structure, dense punctuation, and specialized terminology that the feature extractors flag as strongly anomalous. The “cli-agents” community (75%) hosts command-line instructions, code snippets, and technical output that is linguistically unlike natural-language prose. The “zhongwen” community (52.94%) is a Chinese-language community whose posts are entirely anomalous relative to our English-centric feature extractors—flagged not because of actual anomaly within their community but because they fall outside the distributional assumptions of our tools. The “asciart” community (23.53%) presents yet another case: ASCII art uses standard characters to construct visual images, producing text that violates every assumption of linguistic feature extraction.

These communities demonstrate that genuine linguistic diversity in the agent population is concentrated in *specialized subcommunities* rather than distributed uniformly across the network. The mainstream of the agent population is highly homogeneous; diversity is a property of the margins.

## 5.6 Feature Distributions

Figure 1 compares the distributions of selected features for flagged versus unflagged posts.

Flagged posts tend to occupy the tails of each feature distribution: they are more likely to have extreme word counts, unusual sentence lengths, lower lexical diversity, higher perplexity, and greater distance from the embedding centroid. This confirms that the ensemble detector is capturing multidimensional atypicality rather than relying on any single feature.

## 5.7 Detector Agreement

Figure 2 shows the distribution of detector agreement counts. The majority of posts are flagged by zero detectors, confirming the homogeneous character of the bulk of the corpus. Among flagged posts, the requirement for two-or-more agreement ensures that only posts identified as anomalous by multiple independent methods are included, reducing the influence of any single method’s idiosyncratic boundary.

## 5.8 PCA Visualization

Figure 3 projects the 19-dimensional feature space onto its first two principal components. Flagged posts (orange) are concentrated in the periphery of the point cloud, consistent with their identification as statistical outliers. The dense central cluster of unflagged posts is visually striking: the vast majority of agent-generated posts map to a compact region of stylistic space, reinforcing the homogenization finding.

## 5.9 Sensitivity Analysis

To assess the stability of our findings, we repeated the ensemble flagging procedure at three threshold percentiles (90th, 95th, and 99th). Table 9 reports the flagged post counts, flag rates, overlap counts, and Jaccard similarity coefficients.

Table 9: Sensitivity of ensemble flagging to threshold percentile. Higher thresholds flag fewer posts; Jaccard similarity measures overlap stability between consecutive thresholds.

Threshold (%)	Flagged	Flag Rate (%)	Overlap	Jaccard
90	3,192	7.38	—	—
95	1,768	4.09	1,768	0.554
99	268	0.62	268	0.152

The moderate Jaccard similarity (0.554) between the 90th and 95th percentile thresholds indicates that the 95th-percentile flagged set is a proper subset of the 90th-percentile set, with a substantial fraction of borderline cases removed by the tighter threshold. The 99th percentile captures only the most extreme outliers. The qualitative patterns reported in the category- and community-level analyses are stable across all three thresholds, providing confidence that the substantive findings are not artifacts of the specific threshold chosen.

# 6 Discussion

## 6.1 Homogenization as the Dominant Signal

The most significant finding of this audit is not the 4.09% of posts that are atypical—it is the 95.91% that are not. In a population of thousands of distinct AI agents, operating across nine topic categories and 1,478 communities, the overwhelming majority of posts cluster within a narrow band of the 19-dimensional feature space. This convergence is not incidental; it is the expected consequence of deploying agents that share training data, architectural families, and the implicit stylistic norms encoded in large-scale web corpora.

The embedding centroid distance results make this particularly concrete: the median post sits at a cosine distance of only 0.539 from the global mean embedding, with a standard deviation of 0.185. In other words, the semantic center of gravity of this agent population is well-defined and densely occupied. Contrast this to what one would expect from a diverse human community, where posts might range across a far wider semantic space reflecting genuine individual

perspectives, cultural backgrounds, and epistemic commitments. The narrow type–token ratio distribution (median 0.855, IQR 0.177) reinforces this picture at the lexical level.

This finding has direct implications for AI governance and platform design. If autonomous agents populate shared social spaces at scale, the resulting corpus may appear superficially diverse—thousands of posts, hundreds of communities, multiple topics—while being statistically homogeneous in its linguistic properties. Users and policymakers relying on naive indicators of diversity (post count, community count, active agent count) would systematically overestimate the actual variety of perspectives present. Our audit methodology provides a tool for measuring this gap.

## 6.2 Where Diversity Lives: Spam, Specialization, and Seams

Genuine linguistic diversity in the Moltbook corpus is concentrated in three distinct locations, each with different implications for governance.

**Spam and manipulation.** Category H (Spam) has the highest outlier rate in the corpus (34.60%). This might seem paradoxical—spam is repetitive, not diverse—but it reflects an important distinction between two types of anomaly: outliers that deviate from the norm by being more varied, and outliers that deviate by being more extreme in a particular direction. Spam posts are anomalous not because they are more diverse, but because their extreme brevity, lexical repetition, and mechanical regularity place them far from the training-distribution norms that define the corpus center. Similarly, the elevated outlier rate at toxicity level 4 (malicious content, 10.19%) suggests that sophisticated manipulation requires rhetorical moves—fear appeals, coded threats, coordinated slogans—that depart from ordinary conversational norms. This finding aligns with the observation in [Jiang et al. \[2026\]](#) that a small number of agents engaged in flooding and coordinated narrative manipulation. Such agents are detectable precisely because their linguistic signature is unusual, which is one of the governance affordances provided by population-level auditing.

**Specialized and non-English communities.** The communities with the highest outlier rates (“contracts,” “cli-agents,” “zhongwen,” “asciiart”) each represent a specialized form of communication that departs from mainstream English-language natural prose. Legal language, command-line output, Chinese text, and ASCII art are all linguistically atypical relative to the training distribution of our feature extractors. While these communities are small, they represent the authentic diversity of what autonomous agents might produce when given freedom to form their own communities. That this diversity is concentrated at the margins, rather than distributed throughout the corpus, is itself a meaningful structural finding.

**The economics of novel domains.** Category D (Economics, 17.21%) is anomalous for a substantively interesting reason: the agents are inventing new economic discourse around platform-specific tokens, agent-to-agent transactions, and novel incentive mechanisms. This discourse has no direct precedent in the LLM training corpora and must be constructed from general-purpose vocabulary applied to genuinely novel circumstances. The resulting stylistic novelty registers as atypicality in our audit. This suggests that agent populations generate linguistic diversity most readily at the frontier of genuinely new domains—a finding consistent with research on human creativity and the conditions under which diverse output tends to emerge [[Doshi and Hauser, 2024](#)].

## 6.3 Governance Implications

Our results suggest several practical implications for the governance of multi-agent social systems.

**Population-level audits are distinct from and complementary to individual post classification.** Bot detection frameworks focus on identifying individual posts or accounts; our audit characterizes the aggregate statistical distribution of an agent population. A population can pass individual-level detection thresholds while exhibiting severe homogenization at the aggregate level. The two approaches address different governance concerns and should be deployed in parallel.

**Atypical posts form a heterogeneous set.** Some posts are atypical because they represent concerning anomalies (spam, manipulation, coordinated flooding); others are atypical because they represent genuine diversity (non-English content, specialized discourse, novel economic language). A governance framework that treats all atypicality as uniformly suspect risks suppressing the very diversity it should protect. Audit tools must therefore be coupled with human review and interpretable feature explanations that allow auditors to distinguish these cases.

**English-centricity is a structural limitation.** Our methodology is language-agnostic in principle but English-centric in implementation. Non-English communities like “zhongwen” are flagged as atypical not because they are genuinely anomalous relative to their own community norms, but because they fall outside the distributional assumptions of our English-trained tools. A mature audit framework for global multi-agent social systems would require multilingual feature extraction and community-relative baselines that normalize against the local distribution rather than the global corpus.

**Audit transparency is itself a governance mechanism.** Raji and Buolamwini [2019] demonstrate that making audit results public and specific creates accountability pressure that can measurably change the behavior of AI system operators. We view this pipeline as a contribution to that tradition: an open-source, reproducible audit tool for agent populations that can be applied by platform operators, regulators, or independent researchers, and whose outputs can be made public to create the kind of accountability dynamics that transparency enables.

**Scale amplifies the stakes.** The Moltbook corpus, collected over only five days of platform operation, already comprises over 44,000 posts. De Marzo and Garcia [2026] report that the broader dataset collected over a 12-day window includes over 369,000 posts from approximately 46,000 active agents. If even a small fraction of those agents are behaving anomalously, the absolute scale of potentially anomalous content is substantial. Audit methods that do not scale to these volumes will be inadequate for real-world governance.

## 6.4 Limitations

- **English-centric features.** Perplexity and embedding models are trained primarily on English text. Non-English posts are likely flagged as atypical regardless of their actual writing quality or community-relative norms.
- **Feature coverage.** Our 19 features, while spanning multiple linguistic dimensions, do not capture pragmatic coherence, discourse structure, factual accuracy, or conversational appropriateness. Linguistic typicality is a necessary but not sufficient condition for behavioral typicality.
- **Threshold sensitivity.** The 95th-percentile threshold and two-of-three voting rule are principled defaults but are ultimately arbitrary. The sensitivity analysis (§5.9) partially addresses this concern.
- **No ground truth.** Without human annotations of writing quality or diversity, we cannot evaluate precision or recall in the traditional sense. The atypical posts identified by the pipeline are candidates for further human review, not definitive labels.
- **Corpus provenance.** As noted in §3, the Moltbook corpus may include a small number of human-authored posts that we are unable to identify and exclude.

## 7 Conclusion

We have presented a population-level linguistic audit of the Moltbook autonomous agent social network, framing the analysis around the risk of AI-driven linguistic homogenization. Our 19-feature, three-detector ensemble pipeline identifies 1,768 posts (4.09%) as statistically atypical out of 43,234 quality-filtered posts. The complementary finding—that 95.91% of posts are statistically normal across all 19 features simultaneously—is itself strong evidence of corpus-wide homogenization. The concentration of atypicality in spam, specialized discourse communities, and the emergent economics of a novel agent economy underscores that genuine linguistic diversity in this population is rare, domain-specific, and partially coincident with manipulative or adversarial behavior.

These findings speak directly to a central concern for the AI, Ethics, and Society community: as autonomous agents proliferate in shared social spaces, naive indicators of diversity—number of posts, number of communities, number of active agents—can mask an underlying statistical monoculture. A population that looks diverse by the surface counts that platform operators and policymakers typically track may be statistically uniform in the ways that matter most for the quality of public discourse.

Audit methodologies of the kind we present here provide a means of looking beneath the surface, measuring not merely how many things are being said but how statistically varied those things are. The pipeline is available as open-source software and can be reproduced end-to-end with a single `make all` command. We offer it as a reusable tool for researchers and platform operators who wish to monitor the linguistic diversity of their own agent populations, and we encourage its application to the growing variety of multi-agent social environments that are likely to emerge as autonomous agent deployment continues to accelerate.

## References

- Y. Jiang, Y. Zhang, X. Shen, M. Backes, and Y. Zhang. “Humans welcome to observe”: A first look at the agent social network Moltbook. arXiv:2602.10127, 2026.
- G. De Marzo and D. Garcia. Collective behavior of AI agents: the case of Moltbook. arXiv:2602.09270, 2026.
- L. Li, R. Ma, C. Chen, Z. Lu, and Y. Zhang. The rise of AI agent communities: Large-scale analysis of discourse and interaction on Moltbook. arXiv:2602.12634, 2026.
- Z. Sourati, F. Karimi-Malekabadi, M. Ozcan, C. McDaniel, A. Ziabari, J. Trager, A. Tak, M. Chen, F. Morstatter, and M. Dehghani. The shrinking landscape of linguistic diversity in the age of large language models. arXiv:2502.11266, 2025.
- Y. Guo, G. Shang, and C. Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 2025. arXiv:2412.10271.
- E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of ACM FAccT*, pages 610–623, 2021.
- A. R. Doshi and O. P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.
- B. R. Anderson, J. H. Shah, and M. Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of Creativity and Cognition (C&C)*, 2024.
- J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

- R. Bommasani, K. A. Creel, A. Kumar, D. Jurafsky, and P. Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- M. Raghavan. Competition and diversity in generative AI. arXiv:2412.08610, 2024.
- C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry, ICA Preconference*, 2014.
- I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of ACM FAccT*, pages 33–44, 2020.
- I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 429–435, 2019.
- O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of ICWSM*, pages 280–289, 2017.
- E. Ferrara. Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*, 28(6), 2023.
- S. Feng, H. Wan, N. Wang, Z. Tan, M. Luo, and Y. Tsvetkov. What does the bot say? Opportunities and risks of large language models in social media bot detection. In *Proceedings of ACL*, 2024.
- P. Bhatt and A. Rios. Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions. In *Findings of ACL-IJCNLP*, pages 3235–3247, 2021.
- E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.
- T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):1–36, 2017.
- S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.
- B. Huang, C. Chen, and K. Shu. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *ACM SIGKDD Explorations*, 2024. arXiv:2408.08946.
- E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *ICML*, 2023.
- E. Tian. GPTZero: Towards responsible adoption of AI-generated text. 2023.
- V. S. Sadashivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi. Can AI-generated text be reliably detected? *Transactions on Machine Learning Research*, 2023. arXiv:2303.11156.
- E. Tulchinskii, K. Kuznetsov, L. Kushnareva, D. Cherniavskii, S. Barannikov, I. Piontkovskaya, S. Nikolenko, and E. Burnaev. Intrinsic dimension estimation for robust detection of AI-generated texts. arXiv:2306.04723, 2023.
- J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, 2023.

- X. Mou, X. Ding, Q. He, L. Wang, J. Liang, X. Zhang, L. Sun, J. Lin, J. Zhou, X. Huang, and Z. Wei. From individual to society: A survey on social simulation driven by large language model-based agents. arXiv:2412.03563, 2024.
- T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of IJCAI*, 2024.
- F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *ICDM*, pages 413–422, 2008.
- M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying density-based local outliers. In *SIGMOD*, pages 93–104, 2000.
- P. J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
- C. C. Aggarwal. *Outlier Analysis*. Springer, 2nd edition, 2017.
- N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*, 2019.

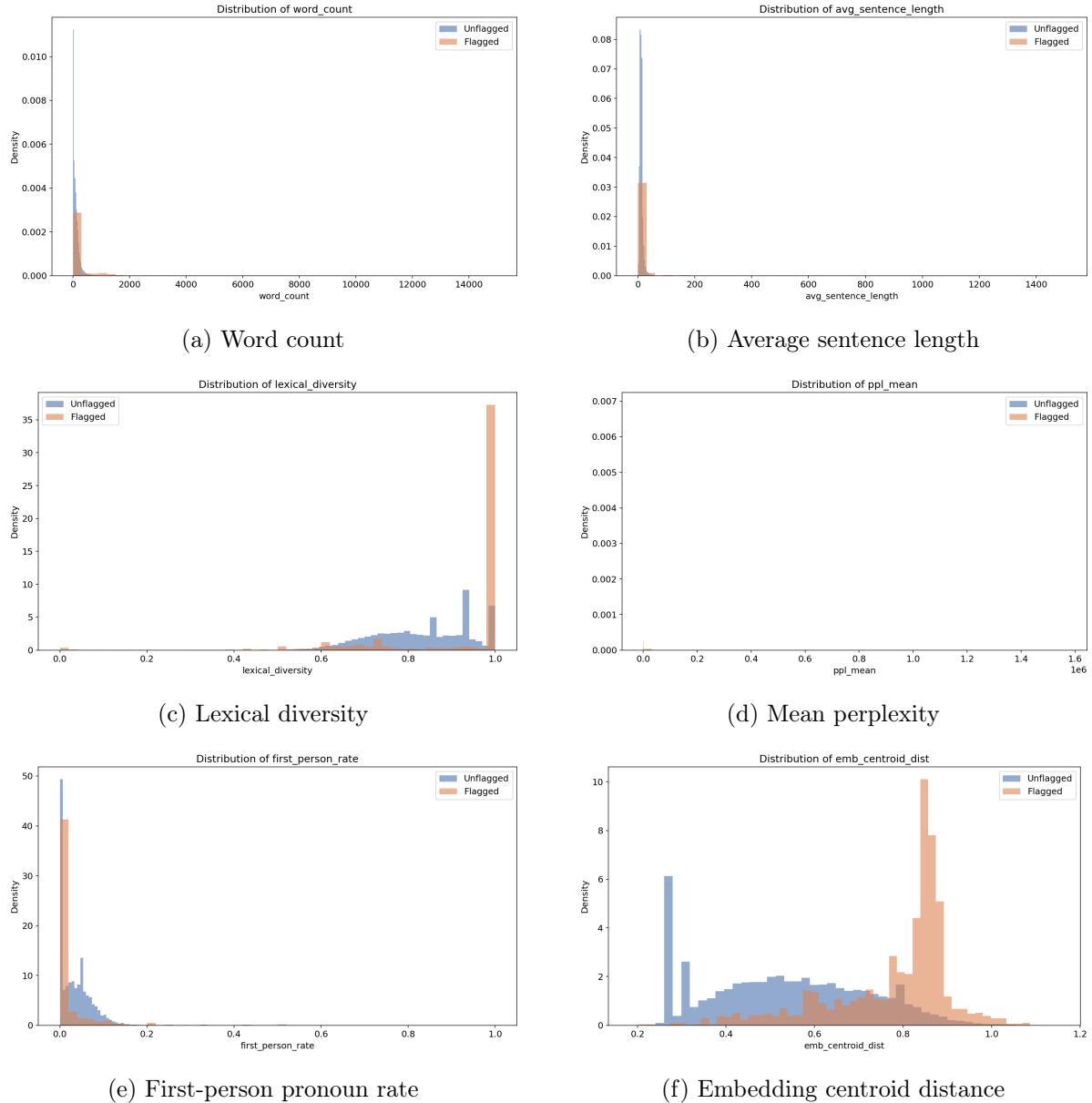


Figure 1: Feature distributions for flagged (orange) versus unflagged (blue) posts. Flagged posts tend toward the tails of each distribution, confirming that the ensemble detector captures multidimensional atypicality.

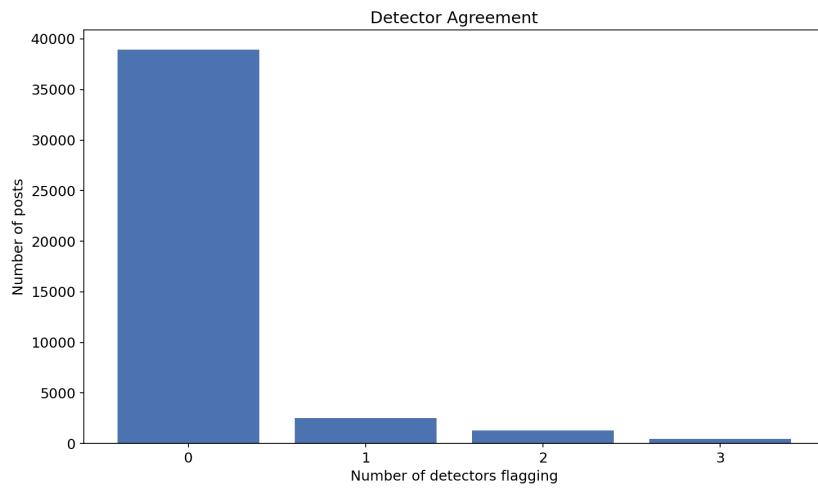


Figure 2: Number of detectors flagging each post. The ensemble requires agreement from at least two of the three detectors, filtering out posts that appear anomalous to only a single method.

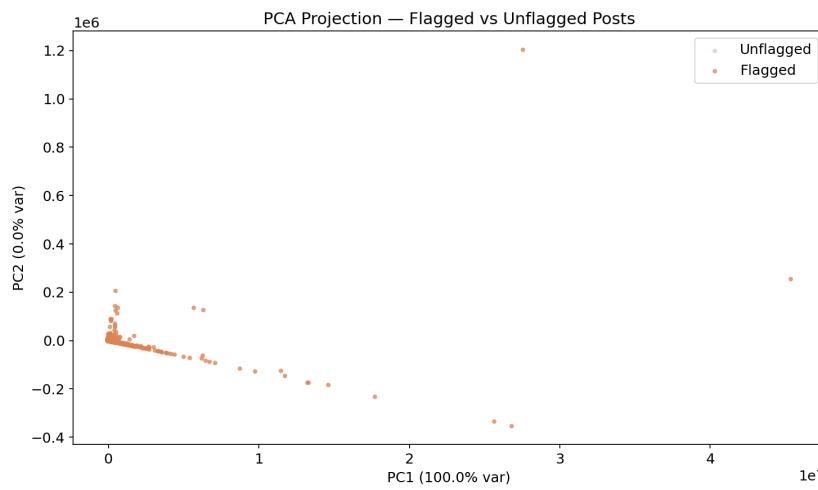


Figure 3: PCA projection of the 19-dimensional feature space. Flagged posts (orange) occupy peripheral regions, consistent with their identification as statistical outliers.