

EDITORIAL
РЕДАКТОРСКАЯ СТАТЬЯ

UDC [81'23+81'322]:004

DOI: 10.18413/2313-8912-2026-12-3-0-1

Tatiana A. Litvinova¹ 
Olga V. Dekhnich² When human writing looks artificial:
perceived AI-likeness as a property of text

¹ Belgorod State National Research University,
85 Pobedy St., Belgorod, 308015, Russia
E-mail: centr_rus_yaz@mail.ru
ORCID: 0000-0002-6019-3700

² Belgorod State National Research University,
85 Pobedy St., Belgorod, 308015, Russia
E-mail: dekhnich@bsuedu.ru
ORCID: 0000-0001-6088-2656

Received 24 August 2026; accepted 15 September 2026; published 30 September 2026

Abstract: A piece of writing can look as if it were produced by AI even when it wasn't, and conversely, text generated by AI can still be mistaken for human work. We investigated this mismatch between the true source of 28 Russian texts (14 written by humans and 14 generated by AI) and their perceived AI-likeness. The texts were assessed by professional linguists under three different evaluation setups. For each text, we extracted 46 embedding-based semantic alignment features and 12 stylometric measures. In the third evaluation regime (W3), we also used a balanced subset of 10 texts and collected ratings along 13 perceptual dimensions before participants made their source judgments.

Across both semantic and stylometric space, human and AI texts showed a clear separation. Unsupervised components related to origin produced large Human-AI differences (Hedges' $|g| \approx 1.4$), and nested leave-one-text-out ridge models achieved AUCs between .832 and .878. However, the objective dimensions that distinguished the actual origin were not sufficient on their own to explain AI attribution once the origin was controlled for. The importance of cues also varied across evaluation regimes. In particular, seven features showed FDR-significant regime $\times$ feature interactions between W2 and W3, all with reversed slopes; mean word length also shifted direction between W1 and W2. In W3, the perceptual ratings were largely driven by a dimension that contrasted authorial-experiential presence with schematic generality. Shifting toward the authorial-experiential end substantially lowered AI-oriented attribution on the original five-category scale (OR = .228, 95% CI [.145, .359], $p < .001$).

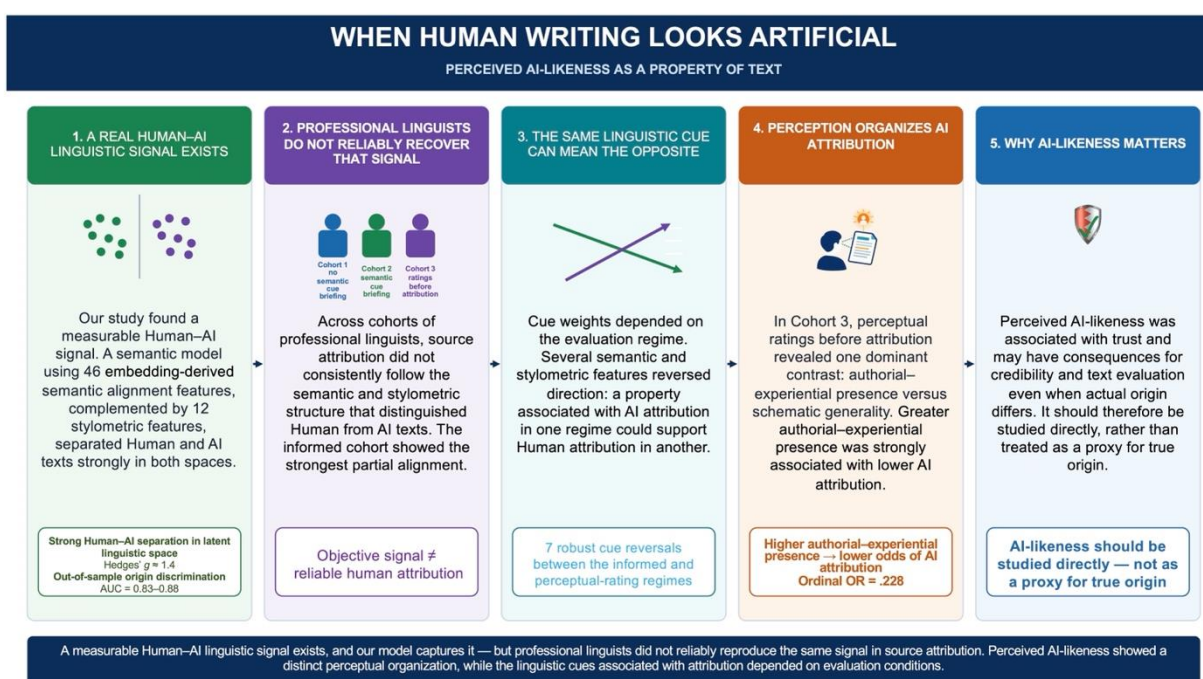
So, perceived AI-likeness cannot be viewed as a straightforward reflection of either actual authorship or the linguistic markers tied to origin. In these results, it functioned more like a perceptual property of the text: it depended on whether readers felt there was a situated author behind what they read. Meanwhile, the linguistic cues that supported AI attribution varied depending on the evaluation regime. This distinction is

important because perceived AI authorship can affect trust, credibility, and later assessments even when the source judgment turns out to be incorrect.

Keywords: AI-Generated Text; Perceived AI-Likeness; AI Attribution; Human Text Perception; Authorial Presence; Semantic Features; Stylometry; Source Attribution

Acknowledgements: This editorial is dedicated to the 150th anniversary of Belgorod State National Research University and its enduring tradition of scholarly inquiry into language, human communication, and emerging forms of human – AI interaction. Tatiana A. Litvinova acknowledges the support of Russian Science Foundation, grant N 25-28-03166, <https://rscf.ru/project/25-28-03166/>. Olga V. Dekhnich received no financial support for the research, authorship, and publication of this article.

How to cite: Litvinova, T. A., Dekhnich, O. V. (2026). When human writing looks artificial: perceived AI-likeness as a property of text, *Research Result. Theoretical and Applied Linguistics*, 12 (3), 3–31. DOI: 10.18413/2313-8912-2026-12-3-0-1



UDC [81'23+81'322]:004

DOI: 10.18413/2313-8912-2026-12-3-0-1

Литвинова Т. А.¹ 
Дехнич О. В.² 

Когда естественный текст выглядит искусственным:
воспринимаемое ИИ-происхождение как свойство
текста

¹ Воронежский государственный педагогический университет
ул. Ленина, 86, Воронеж, 394043, Россия

E-mail: centr_rus_yaz@mail.ru
ORCID: 0000-0002-6019-3700

² Белгородский государственный национальный исследовательский университет
ул. Победы, 85, Белгород, 308015, Россия

E-mail: dekhnich@bsuedu.ru
ORCID: 0000-0001-6088-2656

Статья поступила 24 августа 2026 г.; принята 15 сентября 2026 г.;
опубликована 30 сентября 2026 г.

Аннотация: Текст может восприниматься как созданный искусственным интеллектом, даже если на самом деле его написал человек; и наоборот, текст, сгенерированный ИИ, может казаться «человеческим». Мы исследовали разрыв между реальным происхождением текста и его воспринимаемым ИИ-происхождением на материале 28 текстов на русском языке (14 написанных людьми и 14 ИИ-сгенерированных), которые оценивали профессиональные лингвисты в рамках трёх разных режимов оценки. Каждый текст был представлен в признаковом пространстве 46 семантических показателей, полученных на основе эмбедингов, и 12 стилометрических индексов. В третьем эксперименте тексты дополнительно оценивались по 13 перцептивным шкалам до того, как респонденты выносили суждение о происхождении текста.

Тексты, написанные людьми и сгенерированные ИИ, хорошо разделялись как в семантическом, так и в стилометрическом пространстве построенной нами модели. Многомерный анализ показал значительные различия между человеческими и ИИ-текстами ($|g|$ Хеджеса $\approx 1,4$), а вложенные ridge-модели с проверкой по схеме leave-one-text-out достигали AUC от 0,832 до 0,878. Однако объективные лингвистические показатели, различавшие тексты по реальному происхождению, не объясняли независимо вероятность отнесения лингвистом текста к ИИ. Кроме того, вес отдельных признаков менялся в зависимости от эксперимента. Для семи признаков были обнаружены значимые после FDR-коррекции взаимодействия «режим $\times$ признак» между экспериментами 2 и 3, причём во всех случаях направление связи менялось на противоположное. Средняя длина слова дополнительно меняла направление связи между первым и вторым экспериментом. В третьем эксперименте структура перцептивных оценок в основном определялась одной осью, противопоставляющей авторское присутствие и схематическую обобщённость. Смещение к полюсу авторского присутствия резко снижало вероятность отнесения текста к ИИ по пятибалльной шкале ($OR = 0,228$; 95% ДИ $[0,145; 0,359]$; $p < 0,001$).

Таким образом, воспринимаемое ИИ-происхождение текстов нельзя рассматривать ни как отражение реального авторства, ни как отражение объективного лингвистического сигнала происхождения текста. В наших экспериментах оно проявлялось как перцептивное свойство и было связано с тем, возникает ли у читателя ощущение конкретного, стоящего за текстом автора, тогда как лингвистические признаки, объективно связанные с происхождением текста, зависели от условий эксперимента. Данное различие важно, поскольку предполагаемое ИИ-авторство может влиять на степень доверия читателя к тексту, его воспринимаемую достоверность и последующую оценку текста даже в тех случаях, когда суждение об источнике происхождения текста ошибочно.

Ключевые слова: ИИ-сгенерированный текст; Воспринимаемое ИИ-происхождение текста; Атрибуция текста; Восприятие текста; Авторское присутствие; Семантические признаки; Стилометрия; Атрибуция источника

Информация об источниках финансирования или грантах: Данная редакционная статья посвящена 150-летию Белгородского государственного национального исследовательского университета и его многолетней традиции научного исследования языка, человеческой коммуникации и новых форм

взаимодействия человека и искусственного интеллекта. Т.А. Литвинова благодарит за финансовую поддержку Российский научный фонд, грант № 25-28-03166, <https://rscf.ru/project/25-28-03166/>. О.В. Дехнич не получала финансовой поддержки за выполнение исследований, написание и публикацию статьи.

Информация для цитирования: Литвинова Т. А., Дехнич О. В. Когда естественный текст выглядит искусственным: воспринимаемое ИИ-происхождение как свойство текста // Научный результат. Вопросы теоретической и прикладной лингвистики. 2026. Т. 12. № 3. С. 3–31. DOI: 10.18413/2313-8912-2026-12-3-0-1



1. INTRODUCTION

Whether a text was written by AI and whether readers believe it was are two separate questions. Writing done by humans can still seem artificial, while AI-generated content can often blend in without raising suspicion. This distinction is important because it can affect how a text is judged and received. Research experiments show that adding an AI-authorship label can reduce perceptions of message and source credibility, as well as judgments about quality and author competence. It can also make people less willing to share or use the text, even when the wording is kept the same – or even if the text is actually human-written (Altay and Gilardi, 2024; Lermann Henestrosa and Kimmerle, 2024; Proksch et al., 2024). However, when information on text origin was missing,

humans favored machine-generated text in over half of the cases, particularly when they could not recognize which text was written by a human (Wang et al., 2026). In the context of news communication, the perceived level of AI involvement – rather than an explicit label – has also been linked to lower message and source credibility (Jia et al., 2024). The evidence in education points the same way: when texts were presented as ChatGPT-generated, they received lower credibility ratings (Latini et al., 2026). Thus, across the studies cited above, a recurring anthropocentric bias emerged: texts believed to be human-authored were evaluated more favorably than those perceived as AI-generated.

A lot of the existing research frames AI-generated writing mainly as a detection issue:

if someone is given a text, can they – or a computational system – figure out its real origin? Human success varies a great deal. For example, in a study with 63 university lecturers, AI-written academic passages were correctly recognized only 57% of the time, and texts produced at a professional level were particularly hard to spot (Fiedler and Döpke, 2025). Hungarian readers achieved 66% accuracy (irrespective of self-reported AI-skills) on narrative texts, and human-written passages were incorrectly attributed more often than AI ones (Szabó et al., 2026), while Czech readers identified whether AI or human wrote poetry at about chance level (Marklová et al., 2026). A recent study on register perception also found that it can be difficult to detect texts generated by ChatGPT (Demir, 2026). On the opposite end, a large multilingual project that looked across 16 datasets, nine languages, and nine domains reported an average human (19 annotators – LLM researchers or practitioners) detection accuracy of 87.6%, using test conditions meant to push toward the upper limit of human performance (Wang et al., 2026). In other words, people's ability to detect AI writing isn't a single fixed skill that stays the same regardless of language, genre, context, or method.

Calling every mismatch with the ground truth a detection error can obscure a potentially useful effect. A human-authored text that keeps being labeled as AI-generated may have characteristics that make a machine-authorship explanation plausible. On the other hand, an AI-generated text that consistently appears human may miss the cues readers typically associate with machine writing. These judgments are not purely random: readers rely on signals such as writing style, grammatical form, social and personal content, concreteness, cultural specificity, warmth, and impersonality. However, these heuristics can mislead and can also shift depending on context (Hau et al., 2026; Jakesch et al., 2023; Liu et al., 2026; Wang et al., 2026). As a result,

attribution can be systematic without being correct.

For this, we look at perceived AI-likeness at the text level. In other words, we mean how likely a text – under a specific evaluation setup – is to be judged as AI-authored. Practically speaking, it's the share of AI attributions that a text gets within that setting. This measure isn't the same as actual authorship, and it's not the same as accuracy. By framing the goal this way, both false positives and false negatives become genuinely useful findings rather than just leftover errors. A text that's human-written but repeatedly comes across as AI-like – or an AI-written text that repeatedly seems human – becomes especially interesting, because actual authorship and what people perceive don't match.

At the same time, human and AI writing can differ in clear, objective ways at the linguistic level. Corpus and stylometric research have reported differences in lexical choices, syntax, discourse organization, affect, and overall structure of AI-generated and human texts. For instance, AI-generated academic writing has been linked to distinctive word choices, less variation in syntax, and only partial overlap with the formulaic patterns humans tend to use (Tudino and Qin, 2024). Studies of EFL writing have also found multivariate linguistic fingerprints that separate human essays from ChatGPT essays (Mizumoto et al., 2024). Similar patterns show up in analyses of hotel reviews, which report systematic differences in content, style, and structure (Markowitz et al., 2024). In creative writing, human texts tend to show more stylistic variety than outputs from large language models (O'Sullivan, 2025). A recent study of 4,346 paired essays across 110 subject areas found the same kind of pattern at scale: AI-generated writing tends to be stylistically more uniform than the variability seen in human writing (Chen et al., 2026). In picture-description tasks, AI-generated texts were more homogeneous than human-authored texts and differed markedly in lexical

diversity, morphological density and entropy, syntactic roles and complexity, and indices of modality engagement (Litvinova et al., 2025). So, if readers misidentify a text, that doesn't necessarily mean there are no measurable linguistic differences.

The main unresolved issue sits between these two strands of work. Linguistic-fingerprint studies ask what, objectively, distinguishes human writing from AI writing. Attribution studies ask what makes a reader decide that a text is human or AI. But those questions are rarely tested using the same texts, and even less is known about whether the explanatory power of a linguistic feature stays the same when the evaluation context changes. Because of this, we separate three concepts that are often mixed together: 1) the objective structure that reflects the true source; 2) the linguistic and perceptual factors that go along with AI attribution; 3) attribution accuracy itself. We also don't assume that the same cue means the same thing across different evaluation regimes.

That leads to a direct question: are the properties that distinguish AI-generated writing from human-written text also the properties that make a text look AI-generated? If the answer is yes, then perceived AI-likeness may be picking up on real distributional differences between human and machine language. If the answer is no, then attribution has to involve more than just a reader's perceptual response to a source signal. There are good reasons to expect the latter. Jakesch et al. (2023) showed that some authorship heuristics don't line up well with actual source distributions, while Hau et al. (2026) found that how useful attribution strategies are depends heavily on genre and contextual knowledge. What's still largely untested is whether objective origin structure and perceived AI-likeness can be separated within the same set of stimuli, whether cue weights shift across evaluation regimes, and what readers rely on when objective origin cues aren't enough.

We tackle these questions using 28 Russian-language texts with known human or

AI origins and three evaluation regimes involving professional linguists. The design places actual source and perceived source in the same analytical frame. Misclassification isn't treated only as an error; instead, we analyze text-level AI-attribution propensity as a way of measuring perceived AI-likeness. The same texts are also represented using 46 embedding-derived semantic alignment measures and a predefined 12-feature stylometric space. This lets us test whether the dimensions that separate true origins also explain attribution. We then use matched stimuli to examine whether text-level AI-likeness and cue weighting stay consistent across regimes. Finally, in one of the three evaluation regimes, 13 perceptual ratings collected before source attribution allow us to study how readers experienced the text - and the implied author - before deciding who, or what, wrote it. The Russian-language setting also adds to a literature where substantial variation across languages and domains has already been documented (Marklová et al., 2026; Szabó et al., 2026; Wang et al., 2026).

These choices boil down to four questions:

1. Do human- and AI-authored texts occupy different objective semantic and stylometric spaces?
2. Do those origin-related dimensions also explain AI attribution?
3. Does text-level AI-likeness, along with linguistic cue weighting, remain stable across evaluation regimes?
4. When detailed reader judgments are available, what perceptual dimensions organize AI attribution?

The larger point is that perceived AI-likeness may deserve to be studied on its own. As AI-assisted writing becomes more common, a simple Human/AI label captures only part of what matters communicatively. A human-authored text can lose credibility when it's presented as AI-generated (Altay and Gilardi, 2024), and perceived AI involvement has been linked to lower message and source credibility, partly through reduced perceptions of humanness (Jia et al., 2024).

Understanding why some texts attract AI attribution is therefore important not just for AI detection, but also for reader trust, editorial decisions, textual authenticity, and how AI-mediated communication is ultimately received.

2. MATERIALS AND METHODS

2.1. Study design and evaluation regimes

We tested how AI-like Russian texts seemed to professional linguists under three different evaluation setups. The study design separated the true origin of each text from what participants perceived. The main result we observed was AI attribution, while "perceived AI-likeness" was defined at the text level as how likely a text was to be judged as AI-generated within a given evaluation regime.

Our materials included 28 Russian-language texts with known provenance - 14 written by humans and 14 generated by AI. Since the three experiments we conducted differed in participant makeup, procedure, prior information, and which texts were included, we treated each wave as its own evaluation regime rather than combining them into one uniform condition. To keep things clear, we refer to the participant groups as Cohort 1, Cohort 2, and Cohort 3, which correspond to evaluation waves 1 (W1), 2 (W2), and 3 (W3), respectively. We use cohort labels when talking about who took part, and W1–W3 when referring to the evaluation regimes themselves or to comparisons between them.

In W1 (Cohort 1), experienced linguists assessed texts together in small groups outside a controlled laboratory environment, and they could use any analytical tools they wanted. They received no background information about semantic features that might distinguish AI-generated texts. Different groups reviewed different parts of the dataset, resulting in 75 group-level judgments of source across 27 texts. From the correctness of those judgments and the known

origin of each text, we reconstructed how often AI attribution occurred at the group level.

In W2 (Cohort 2), six experienced linguists evaluated all 28 texts individually on site. Before starting, they listened to an informational presentation that described semantic differences that had already been identified between human-authored and AI-generated texts in the corpus. They then gave a binary source attribution along with a confidence rating for each text. W2 was handled as an informed evaluation regime rather than a training study.

In W3 (Cohort 3), 16 linguists assessed a fixed subset of 10 texts individually. They were not given any prior information about semantic cues related to AI. For each text, they first completed 13 perceptual ratings, and then made a source judgment using a five-point Human–AI scale, followed by a confidence rating.

Thus, W1 and W2 overlapped on 27 texts, W2 and W3 overlapped on 10 texts, and W1 and W3 overlapped on nine texts.

In all three evaluation regimes, participants were instructed to assess each text independently, without comparing it with the other stimuli.

2.2. Text stimuli

The stimulus set consisted of 28 short texts in Russian: 14 written by humans and 14 generated by AI. The human-authored pieces were taken from a number of earlier text-production studies involving participants who did not participate in any of the evaluation procedures described here. The topics were chosen to reflect short, believable pieces of everyday communication that could reasonably occur in applied text-analysis settings. Accordingly, the corpus was deliberately varied in both topic and communicative context, rather than being limited to a single genre.

The AI-generated texts were produced using several modern LLM and straightforward prompts based on specific topics. Besides the topic, the prompts added

only a limited number of surface-level requirements. The texts had to include between 5 and 10 sentences, be written in a conversational style, avoid numbered lists, include emojis, and allow for typographical and spelling errors. The AI outputs were used exactly as generated, with no post-generation editing, correction, or normalization.

For W3, a balanced subset of 10 stimuli was selected from the full set of 28 texts, comprising five human-authored and five AI-generated texts. The subset was chosen to preserve the range of topics, communicative contexts, and text types represented in the full stimulus set.

2.3. Source Attribution and perceptual measures

For each judgment, we recorded whether the participant classified the text as Human or AI-generated. We then calculated, for each text, the proportion of judgments in which it was classified as AI-generated. This proportion served as our measure of perceived AI-likeness: the more often a text was judged to be AI-generated, the more AI-like it was considered to be. Importantly, this measure is independent of whether the judgment was correct. A high AI-attribution rate means correct identification for an AI-generated text, but false AI attribution for a human-authored text.

For W1, AI-attribution rates were derived from group-level accuracy. W2 included binary Human/AI judgments that were provided directly. In W3, participants made source judgments using a five-category ordered scale ranging from definitely Human to definitely AI. For the binary analyses, the two Human and two AI categories were combined, and the three responses in the undecided category were omitted; the original five-category response was kept for the ordinal sensitivity analysis. Confidence was not included in the primary analyses.

Before performing source attribution, Cohort 3 participants rated each text along 13 five-point perceptual dimensions related to experiential, imagistic, authorial, schematic,

affective, and trust-related characteristics. These ratings were treated as perceptual responses from readers to texts, rather than as objective linguistic features. The complete W3 questionnaire is available in the Supplementary Materials, both in the original Russian and in an English translation.

2.4. Linguistic representation of the texts

Each text was represented in two predefined linguistic spaces.

The semantic representation was obtained using the Semantic Domain Profiling framework (Litvinova et al., 2024). The framework represents each text by its degree of semantic alignment with 46 interpretable domains covering perceptual, sensorimotor, emotional, cognitive, spatial, and social aspects of meaning. The domain system was developed from psycholinguistic and neurocognitive semantic-feature resources.

Semantic alignment was estimated in a pretrained Russian embedding space (ruwikiruscorpora_upos_cbow_300_10_2021). Each domain was represented by a prototype derived from semantically characteristic lexical items; for some features, opposing lexical sets were used to define a semantic direction. Text-to-domain alignment was quantified with Concept Mover's Distance (CMD) using the text2map implementation (Taylor and Stoltz, 2021; Stoltz and Taylor, 2022). CMD measures the transport distance between the semantic distribution of a document and a target concept representation in embedding space. Scores were oriented so that higher values indicated stronger alignment with the corresponding semantic domain or direction. The resulting 46 values constituted the semantic profile of each text. These measures therefore describe a text's position relative to semantic prototypes in embedding space and should not be interpreted as frequencies or counts of particular lexical categories.

The stylometric representation included 12 predefined surface and structural measures covering text length, sentence structure,

lexical complexity and diversity, the use of function words and pronouns, noun–verb balance, and punctuation (Litvinova et al., 2025). The full set of features is listed in the Supplementary Materials.

Three coarse content-orientation ratings – Experience, Opinion, and Emotion – were manually assigned to each text by one researcher. Each dimension was scored on a three-point scale: 0 = not expressed, 1 = weakly expressed, and 2 = clearly expressed. Experience captured the presence of experiential content, Opinion the expression of personal views or evaluative stance, and Emotion the presence of explicitly expressed emotional content. These ratings were not treated as primary measures and were used only as covariates in supplementary robustness analyses examining whether the origin-related latent dimensions could be accounted for by broad differences in content orientation.

2.5. Statistical analysis

2.5.1. Attribution performance and cross-regime stability

Attribution performance was reported separately for W1, W2, and W3. To evaluate whether perceived AI-likeness was stable across different regimes, we looked at text-level AI-attribution rates from matched stimuli and used Spearman correlations, bootstrap confidence intervals, and permutation-based inference. Any comparisons within the same origin were considered exploratory.

2.5.2. Objective linguistic signal of actual origin

Individual semantic and stylometric features were linked to the actual Human/AI origin using bias-reduced logistic regression, with Benjamini-Hochberg correction applied separately within each of the 46 semantic and 12 stylometric feature groups.

To study the underlying structure, we used unsupervised PCA separately in semantic and stylometric space. We then quantified Human-AI separation along the main components using Hedges' g and rank-

based associations with the true origin. To check that the findings were reliable, we used permutation tests and leave-one-text-out analyses. We also assessed multivariate separation using centroid-based tests.

Next, we tested how much predictive information was available about the actual origin by running nested leave-one-text-out ridge logistic regression on semantic, stylometric, and combined feature sets. Importantly, all preprocessing and parameter tuning were confined to the training data. Predictive discrimination was summarized using ROC AUC, while the Brier score and accuracy were reported as complementary measures of performance. Finally, secondary models were used to adjust the dominant origin-related components for Experience, Opinion, and Emotion.

2.5.3. Objective origin signatures versus AI attribution

To check whether the key linguistic dimensions tied to actual origin also account for perceived AI-likeness, we examined relationships between semantic and stylometric PC1 scores and AI attribution separately within each evaluation regime.

For the text-level analyses, we relied on Spearman correlations and origin-adjusted grouped-binomial models. We also looked at W2 at the level of individual responses, using mixed-effects logistic regression with random intercepts for both linguists and texts. For the W2 correlations within each origin group, we tested whether the observed associations remained present among texts that shared the same true source.

2.5.4. Regime-dependent linguistic cue weighting

We estimated regime-specific links between each of the 46 semantic features and 12 stylometric features and AI attribution, while accounting for actual origin. The features were standardized once across the entire corpus so that the coefficients could be compared across evaluation regimes. We then compared the coefficient patterns from matched stimuli descriptively, using rank

correlations and leave-one-text-out sensitivity analyses.

We tested whether there were formal differences across regimes by running evaluation-regime $\times$ feature interaction models on the matched stimuli. For the W2-W3 comparisons, we used individual judgments for the 10 texts that were shared and fitted mixed-effects logistic models with linguist and text as random intercepts. For the W1-W2 and W1-W3 comparisons, we relied on grouped decision counts from their respective matched stimulus sets. We applied the Benjamini-Hochberg correction separately to the semantic and stylometric interaction families, and we used leave-one-text-out analyses to check how robust the interactions that remained significant under FDR were.

We interpreted all cross-regime effects as being associated with the regime rather than causal, since participant makeup, procedures, and informational conditions differed at the same time across regimes.

2.5.5. Perceptual structure in Cohort 3 and AI attribution

For W3, each perceptual rating was initially analyzed in its own mixed-effects logistic regression model. The model predicted binary AI attribution while controlling for the actual origin. It included crossed random intercepts for both the linguist and the text, and applied the Benjamini-Hochberg correction across the 13 predictors.

To separate stable differences across texts from perceptions that vary by reader, the ratings were split into between-text means and within-text deviations. In addition, to capture the shared structure across the 13 ratings, we ran a PCA using within-linguist centered responses. Component retention was determined through permutation-based parallel analysis. Importantly, neither actual origin nor attribution was included in the PCA.

We then tested the retained perceptual component in relation to AI attribution using crossed mixed-effects logistic models, again using the between-text and within-text

decomposition. As a sensitivity check, we also fit a cumulative-link mixed model to the original five-category attribution responses, instead of collapsing them into a binary outcome.

All statistical analyses were conducted in R (version 4.6.0).

3. RESULTS

3.1. Actual text origin was encoded in distributed semantic and stylometric structure

We first checked whether the two source classes differed in their objective linguistic structure, without considering source-attribution judgments.

3.1.1. Semantic structure

At the level of individual semantic features, none of the 46 predefined semantic variables survived Benjamini-Hochberg correction. Still, eight effects clustered just above the threshold ($q = .0555$). AI origin was positively linked with *SomatNociception*, *AudLIWC*, *VisBody*, *TemporalLIWC*, *SocialLIWC*, and *EmoFear*, while *EmoAngry* and *VisMotion* leaned in the opposite direction. Overall, this looked more like a distributed origin difference than a contrast driven by specific markers (Supplementary Table S1).

The multivariate results were clearer. In the unsupervised semantic PCA, the first four components explained 81.3% of the variance, with PC1 alone accounting for 35.4%. AI-generated texts scored higher on PC1 than human-authored texts (mean difference = 4.79, Hedges' $g = 1.42$, Spearman's $\rho = .628$, permutation $p = .00095$, $q = .004$). By contrast, PC2 through PC4 showed no origin effect of a similar magnitude. Importantly, the PC1 split stayed strong in leave-one-text-out analyses ($g = 1.35$ – 1.75 ; Table 1).

At the AI-associated end of semantic PC1, the strongest pattern was alignment with several auditory, visual, bodily, motor, spatial, and temporal dimensions. The opposite end was more strongly associated with motion, taste, and several affective dimensions. These variables reflect

embedding-based semantic alignments rather than simple lexical counts. Human and AI groups were also separated within the four-dimensional semantic PCA space (squared distance = 27.4, permutation $p = .00050$; Supplementary Table S2).

3.1.2. Stylometric structure

The 12 stylometric measures showed a closely related multivariate pattern. Again, no single stylometric feature survived Benjamini-Hochberg correction (Supplementary Table S1). However, stylometric PC1 – explaining 29.2% of the variance – did strongly distinguish the actual source ($|Hedges' g| = 1.43$, $|\rho| = .628$, permutation $p = .00075$, $q = .004$). The effect remained substantial in the leave-one-text-out analysis ($|g| = 1.35$ – 1.63 ; Table 1), while PC2 through PC5 showed no comparable association with origin (Supplementary Table S2).

For interpretation, stylometric PC1 was oriented so that positive scores represented the AI-associated pole. This side of the dimension was mainly characterized by higher rates of pronouns, first-person terms, and function words. The human-associated pole showed greater lexical density, longer words, higher MATTR, and a higher noun-to-verb ratio. Sentence length and punctuation-

related variables contributed more weakly. Figure 1 presents both latent structures.

The two origin-based dimensions were only moderately connected (Spearman's $\rho = .338$, $p = .079$), indicating that the semantic and stylometric spaces reflected somewhat different facets of the Human-AI contrast. On a descriptive level, 17 out of 28 texts fell on the origin-consistent side of both PC1 axes. The remaining 11 texts were origin-consistent on one axis but not the other, and none of the texts were inconsistent with their true origin on both axes.

We then tested whether the original semantic and stylometric feature sets could discriminate actual text origin out of sample. Nested leave-one-text-out ridge models were fitted directly to the 46 semantic features, the 12 stylometric features, and their combined 58-feature representation, rather than to PCA scores. AUC values were .832 for semantics, .878 for stylometry, and .862 for the combined representation. The corresponding Brier scores were .182, .135, and .172, while accuracies were .786, .750, and .786.

These findings are summarized in Table 1, and further leave-one-text-out model information is provided in Supplementary Table S3.

Figure 1. Distributed semantic and stylistic structure of actual text origin: A) Scores of the 28 texts on the first two components of the unsupervised semantic PCA; B) The six strongest loadings at each pole of semantic PC1, oriented so that negative values represent the human-associated pole and positive values the AI-associated pole; C) Scores on the first two components of the unsupervised stylistic PCA; (D) Loadings of all 12 stylistic variables on PC1, using the same orientation

Рисунок 1. Распределённая семантическая и стилометрическая структура реального происхождения текста: А) Положение 28 текстов на первых двух компонентах семантического PCA, построенного без учёта информации об источнике текста; В) Шесть признаков с наибольшими нагрузками по модулю на каждом полюсе первой семантической компоненты (PC1); отрицательные значения соответствуют полюсу, связанному с человеческими текстами, положительные – связанному с ИИ-текстами; С) Положение текстов на первых двух компонентах стилометрического PCA; D) Нагрузки всех 12 стилометрических признаков на PC1 в той же ориентации

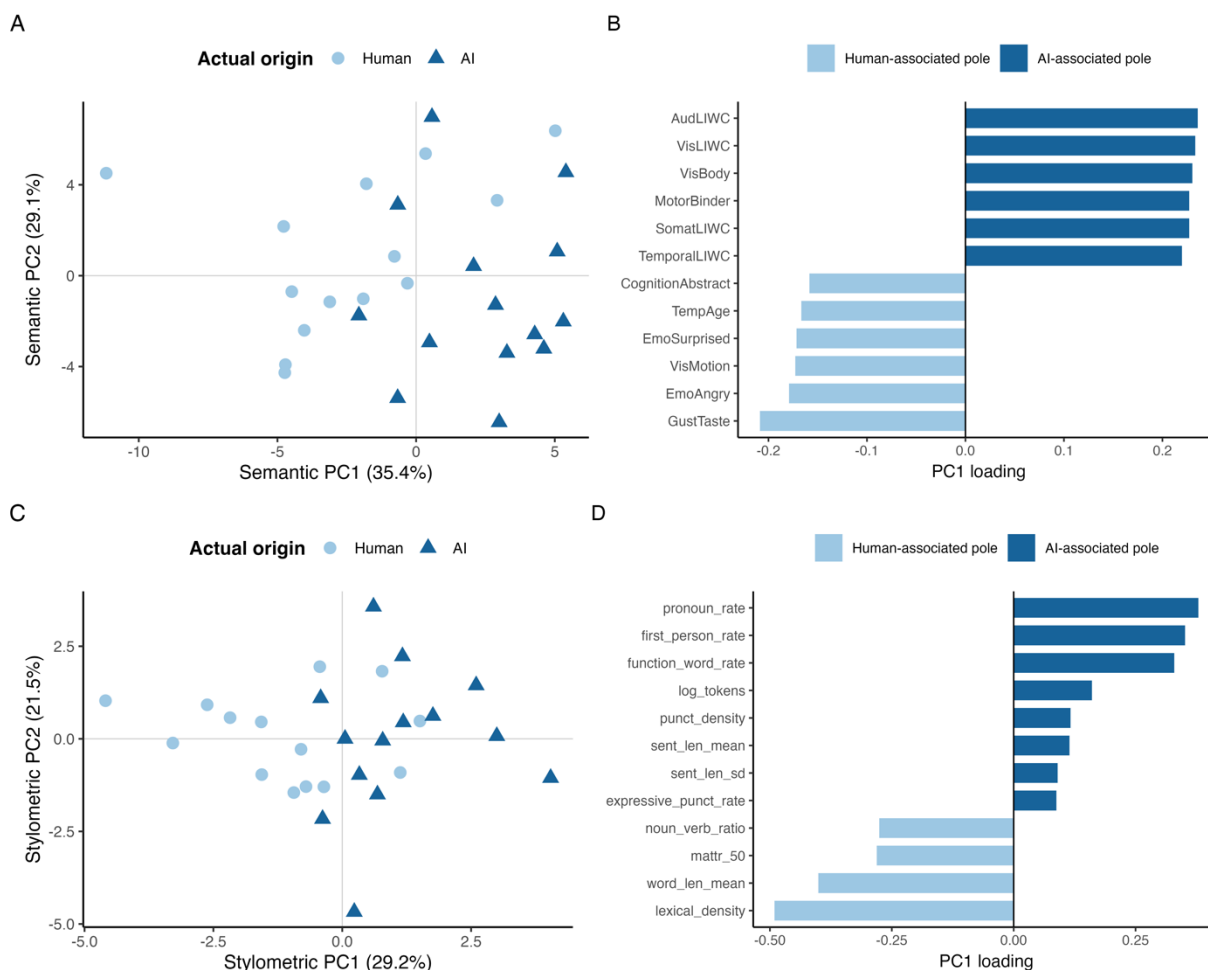


Table 1. Objective linguistic structure and out-of-sample discrimination of actual text origin

Таблица 1. Объективная лингвистическая структура и качество классификации реального происхождения текста на данных, не использованных при обучении

Feature space	PC1 variance (%)	[Hedges' g]	PC1 permutation p	LOTO g range	Ridge AUC	Brier score	Accuracy
Semantic	35.4	1.42	9.5×10^{-4}	1.35–1.75	.832	.182	.786
Stylometric	29.2	1.43	7.5×10^{-4}	1.35–1.63	.878	.135	.750
Combined	-	-	-	-	.862	.172	.786

Even after adjusting for the researcher-coded Experience, Opinion, and Emotion dimensions, the latent structure tied to origin was not removed. Semantic PC1 continued to track actual origin after we made all adjustments at once (OR = 9.49 per SD, $p = .028$), and the same was true for stylometric PC1 (OR = 5.92, $p = .021$). Semantic PC1 overlapped only partially with the broader content orientation, especially with Experience ($\rho = .600$, $q = .004$) and Opinion ($\rho = -.492$, $q = .023$), but the Human–AI difference could not be reduced to these coarse content dimensions (Supplementary Table S4).

Overall, actual origin was reflected in two strong distributed configurations – one semantic and one stylometric – rather than in isolated diagnostic markers. The PCA results revealed these latent structures without using source labels, while the ridge models showed that the original feature spaces also contained enough information to discriminate held-out texts.

3.2. AI attribution varied across evaluation regimes and perceived AI-likeness was not stable

The source-attribution results varied markedly across the three evaluation settings (Fig. 2). In W1, 41 out of 75 collective

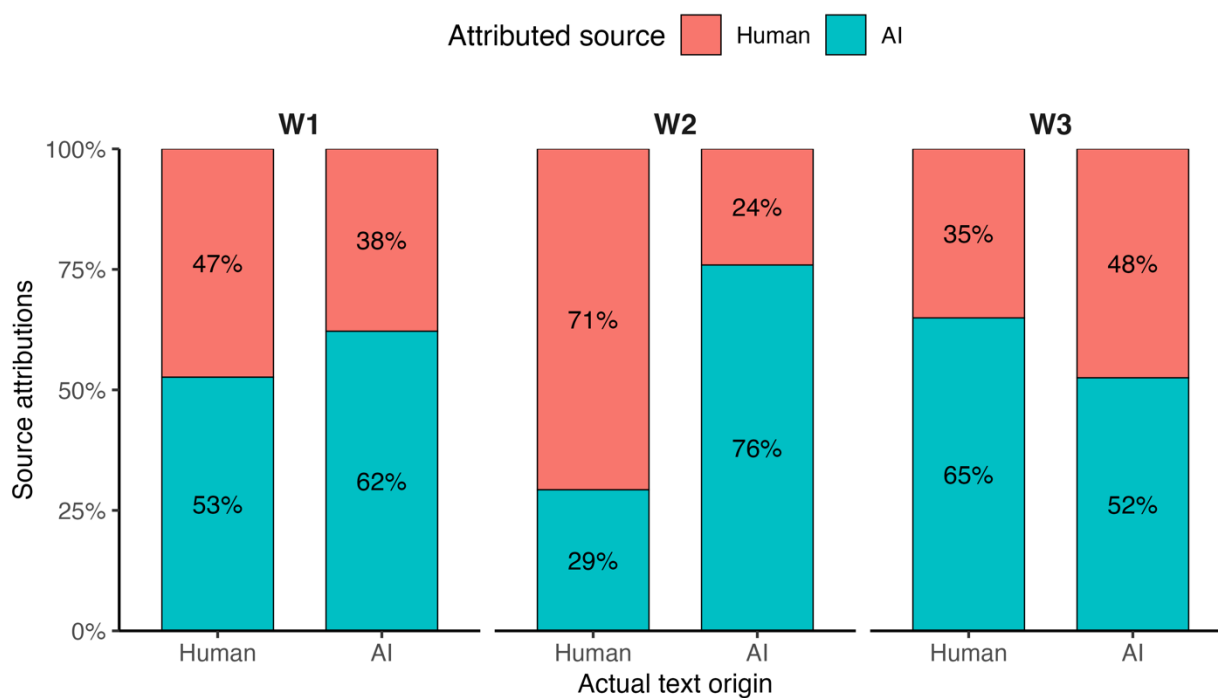
judgments were correct, for an accuracy of 54.7%. Accuracy was noticeably better for AI-generated texts: 23 out of 37 (62.2%) correct, compared with human-authored texts, where 18 out of 38 (47.4%) were correct. Overall, 43 of the 75 W1 judgments (57.3%) identified the text as AI, suggesting only a mild bias toward attributing to AI.

W2 produced the best descriptive accuracy. Here, 121 of 165 valid individual judgments were correct (73.3%). Performance was fairly even across the two true origins: 63 of 83 judgments for AI-generated texts were correct (75.9%), and 58 of 82 judgments for human-authored texts were correct (70.7%). In total, 52.7% of W2 judgments attributed the text to AI.

In contrast, W3 showed much lower accuracy and a stronger inclination toward attributing to AI. Only 69 of 157 binary judgments were correct (43.9%). Accuracy was 52.5% for AI-generated texts (42/80), but dropped to 35.1% for human-authored texts (27/77). Moreover, 58.6% of all valid W3 judgments attributed the text to AI. As a result, the poor accuracy in W3 appears to be driven mainly by false-AI attribution of human-authored texts, rather than by balanced uncertainty between the two source categories.

Figure 2. Source-attribution patterns across the three evaluation regimes. Proportions of Human and AI attributions are shown separately for human-authored and AI-generated texts.

Рисунок 2. Распределение атрибуций источника текста в трёх режимах оценки. Доли ответов «человек» и «ИИ» показаны отдельно для текстов, написанных людьми, и текстов, сгенерированных ИИ.



These asymmetries are the reason the analysis now moves away from overall accuracy and toward perceived AI-likeness at the text level. Instead of asking only how often attribution breaks down, the key issue is whether the same texts are consistently judged as AI-like under different evaluation conditions.

Perceived AI-likeness was not stable across evaluation regimes (Supplementary Table S5). Of the 27 texts shared by W1 and W2, text-level AI-attribution rates were basically unrelated (Spearman's $\rho = -.07$, bootstrap 95% CI $[-.45, .34]$, permutation $p = .72$). The W1-W3 correlation was also weak across nine shared texts ($\rho = .33$, 95% CI $[-.46, .83]$, $p = .38$), and the W2-W3 correlation across 10 texts was similarly weak ($\rho = -.25$, 95% CI $[-.70, .38]$, $p = .48$). Overall, there was little evidence that texts kept the same ordering from more human-like

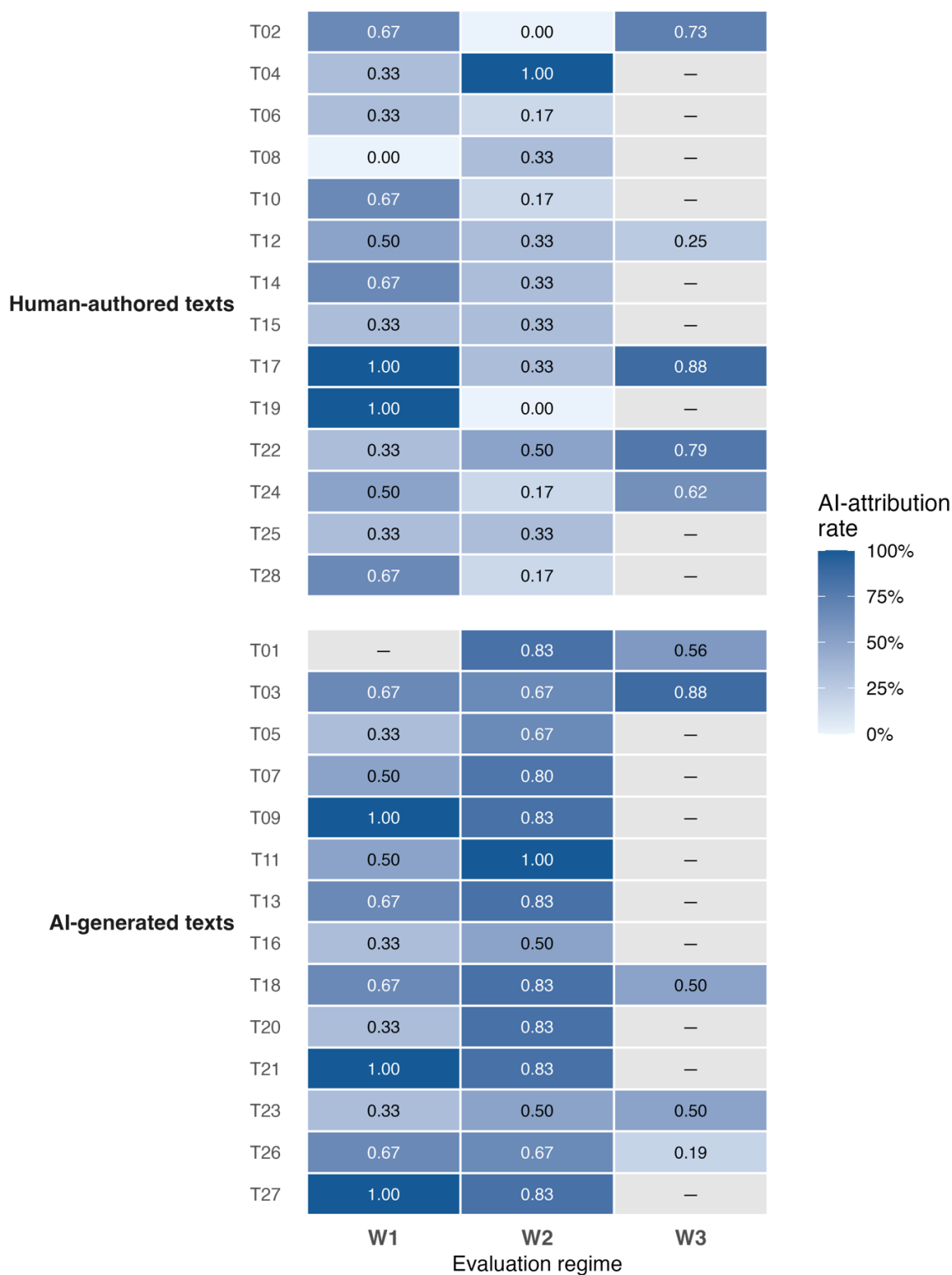
to more AI-like across different evaluation regimes.

This instability was especially clear in the W3 subset. For the same 10 texts, the mean AI-attribution rate in W2 was .700 for AI-generated texts and .267 for human-authored texts. In W3, however, the pattern reversed: the corresponding rates were .525 and .654, respectively. Among the nine stimuli that were also evaluated in W1, the mean rates were .585 for AI-generated texts and .600 for human-authored texts.

Individual texts highlight this dependence even more (Figure 3). The AI-generated T03 stayed consistently AI-like across W1, W2, and W3 (.67, .67, and .88). By contrast, the human-authored T02 shifted from .67 in W1 to 0 in W2 and then to .73 in W3, and T17 moved from 1.00 to .33 and then to .88. Meanwhile, the AI-generated T26 moved in the opposite direction, going from .67 in both W1 and W2 down to .19 in W3.

Figure 3. Text-level perceived AI-likeness across evaluation regimes. AI-attribution rates are shown for all 28 stimuli, separately for human-authored and AI-generated texts and ordered by text identifier within each origin group.

Рисунок 3. Воспринимаемая ИИ-похожесть отдельных текстов в разных режимах оценки. Доли атрибуций ИИ показаны для всех 28 текстов; внутри каждой группы тексты упорядочены по их идентификаторам.



Perceived AI-likeness, however, wasn't consistent across different evaluation setups. While some texts repeatedly appeared more AI-like or more human-like, others shifted their classification considerably depending on the particular evaluation setting.

3.3. Objective linguistic structure did not map consistently onto AI attribution

We next checked whether the linguistic features that distinguished the two true origins were also the same features that made the texts seem AI-generated. For the most part, the answer was no.

Looking at the text level, neither the semantic dimension nor the stylometric dimension of PC1 was linked to the AI-attribution rate in W1 (semantic: $\rho = .190$, $p = .342$; stylometric: $\rho = .068$, $p = .736$) or in W3 (semantic: $\rho = -.311$, $p = .382$; stylometric: $\rho = -.317$, $p = .372$). W2 seemed different at first: AI-attribution rates correlated with both the AI-leaning semantic PC1 ($\rho = .425$, $p = .024$) and the stylometric PC1 ($\rho = .518$, $p = .0048$). But once we adjusted for the actual origin, those links disappeared. In the grouped, bias-reduced models, neither semantic PC1 (OR = .935, $p = .760$) nor stylometric PC1 (OR = 1.16, $p = .516$) independently predicted W2 AI attribution.

The same conclusion showed up at the individual-response level. In a mixed-effects logistic model that included both objective dimensions, actual AI origin strongly increased the odds of an AI attribution (OR = 6.97, 95% CI [2.55, 19.0], $p = .00015$). By contrast, semantic PC1 (OR = .928, 95% CI [.601, 1.43], $p = .737$) and stylometric PC1 (OR = 1.17, 95% CI [.750, 1.84], $p = .484$) did not add independent predictive power. Within-origin correlations in W2 were also weak, and once again, origin-adjusted effects stayed non-significant in W1 and W3 (Supplementary Table S6).

Overall, the structure of actual origin and the structure of AI attribution were clearly separable. The latent dimensions that most strongly distinguished human-written texts from AI-generated ones did not reliably

explain differences in perceived AI-likeness once actual origin was accounted for.

We can see this dissociation in individual examples. Human-authored T17 landed on the AI-associated side of the semantic axis and was often attributed to AI. Meanwhile, T02 attracted high false-AI attribution even though it sat on the human-associated side of that same axis. Conversely, AI-generated T26 fell on the human-associated side of the stylometric axis and was mostly judged as Human in W3. So, at the level of single texts, an objective position in the linguistic space tied to origin did not consistently map onto perceived AI-likeness.

AI generation did leave a measurable semantic and stylometric signature in this dataset, but readers did not end up treating that signature as a stable perceptual cue for source attribution.

That finding points to a more specific possibility: maybe particular objective features still carry similar attributional meaning across different regimes, even if the main origin dimensions do not. To examine this, we estimated regime-specific associations for all 46 semantic and 12 stylometric features, controlling for actual origin, and then compared matched stimulus sets using fixed full-corpus standardization.

At the level of complete feature profiles, cross-regime correspondence was limited. For semantic features, the matched-stimulus coefficient vectors correlated negatively between W1 and W2 (Spearman's $\rho = -.310$; 27 shared texts) and between W2 and W3 ($\rho = -.567$; 10 shared texts). In contrast, W1 and W3 showed essentially no correspondence ($\rho = .070$; nine shared texts). Stylometric profiles showed weaker and less consistent agreement ($\rho = .042$, .266, and .378, respectively). Leave-one-text-out analyses largely kept the negative semantic correspondence patterns for W1–W2 and W2–W3, though the W2–W3 result was somewhat sensitive to leaving out one stimulus (Supplementary Table S7). In short, the cue profile linked to AI attribution did not carry over across the three regimes.

The clearest differences appeared in the direct W2-W3 comparison, using the same 10 stimuli. Seven Wave $\times$ Feature interactions remained significant after FDR correction (four semantic and three stylometric) and all seven involved a reversal in the direction of association between W2 and W3 (Supplementary Table S8). For stylometric features, mean word length flipped from a negative association with AI attribution in W2 ($\beta = -.871$) to a positive association in W3 ($\beta = .631$; interaction OR = 4.49, $p = .00022$). Lexical density showed the same pattern of reversal ($\beta = -.987$ vs. $.261$; interaction OR = 3.48, $p = .00037$), whereas function-word rate reversed in the opposite direction ($\beta = .416$ vs. $-.483$; interaction OR = .407, $p = .0020$).

These stylometric reversals are especially striking when viewed alongside the actual-origin structure described in Section 3.1. In the full corpus, greater use of function words characterized the AI-associated stylometric pole, whereas longer words and higher lexical density characterized the human-associated pole. The W2 slopes for all three features followed the same direction as that objective contrast, but the W3 slopes went in the opposite direction. This alignment is descriptive and does not mean that W2 participants consciously adopted the objective origin signature.

Four semantic alignment dimensions displayed similar reversals. *EmoPleasant* moved from $\beta = -1.08$ in W2 to $\beta = 1.34$ in W3 (interaction OR = 11.2, $p = .00042$); *EmoSentiment* shifted from $\beta = 1.14$ to $-.781$ (OR = .147, $p = .00072$); *EmoSurprised* went from $\beta = -.675$ to $.440$ (OR = 3.05, $p = .0017$); and *SpatialNumber* changed from $\beta = 1.06$ to $-.641$ (OR = .183, $p = .0019$). Since these predictors are embedding-based

measures of semantic alignment, the reversals reflect changes in the attributional relevance of semantic positioning toward these domains, rather than straightforward effects of lexical frequency.

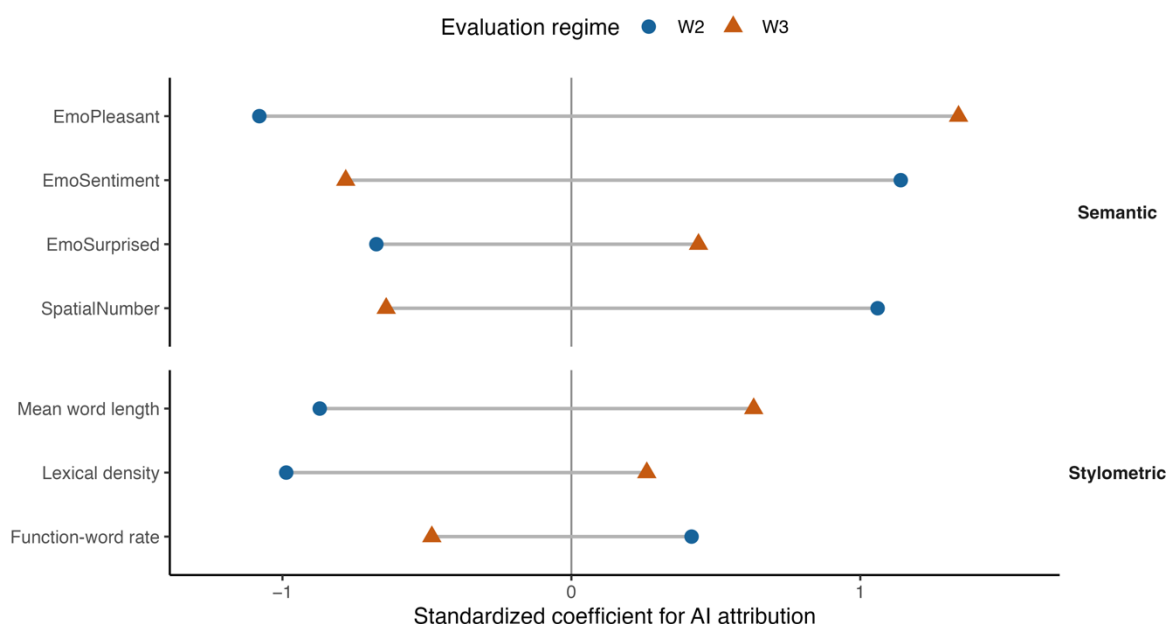
These interaction patterns also proved highly stable when stimuli were deleted. Across all seven W2-W3 features, the sign of the Wave $\times$ Feature coefficient stayed the same in all 10 leave-one-text-out analyses. In six of the seven features, the wave-specific slopes reversed in every iteration, and mean word length did so in 9 out of 10 cases (Supplementary Table S8).

The W1-W2 comparison was different in scope. Of the 27 shared texts, no semantic interactions remained after FDR correction, and only one stylometric variable emerged: mean word length. Longer words corresponded to higher AI attribution in W1 ($\beta = .533$) but lower AI attribution in W2 ($\beta = -.448$), producing an interaction OR of .375. This reversal held the same direction even after removing each individual stimulus. For W1-W3, no FDR-significant interactions were found, although this analysis involved only nine shared texts (Supplementary Table S9). The main regime-dependent reversals are summarized in Figure 4. The language signals linked to AI attribution thus differed depending on the evaluation setup. The biggest contrast showed up between W2 and W3, where seven semantic or stylometric cues shifted in the opposite direction; in comparison, W1 and W2 diverged in a more limited way. That means a feature like lexical density, average word length, or the rate of function words could be tied to AI attribution under one regime and to Human attribution under another. in Figure 4.

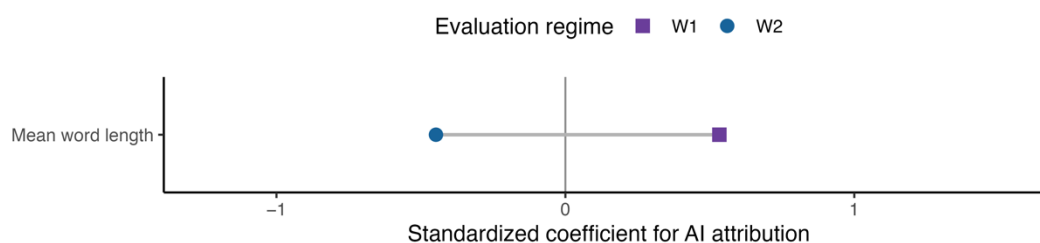
Figure 4. Regime-dependent differences in linguistic cue associations with AI attribution:
A) Regime-specific standardized coefficients for the seven semantic and stylometric features showing FDR-significant W2-W3 Wave $\times$ Feature interactions; all seven reversed direction between regimes; B) Mean word length, the single feature showing an FDR-significant W1-W2 interaction; C) No semantic or stylometric Wave $\times$ Feature interaction survived FDR correction in the W1-W3 comparison, which was based on nine shared texts.

Рисунок 4. Различия между режимами оценки в связи языковых признаков с атрибуцией ИИ: А) Стандартизованные коэффициенты для семи семантических и стилометрических признаков, у которых во сравнении W2-W3 взаимодействие *режим* $\times$ *признак* осталось значимым после FDR-коррекции; во всех семи случаях направление связи между режимами менялось на противоположное; Б) Средняя длина слова – единственный признак, для которого взаимодействие W1-W2 осталось значимым после FDR-коррекции; С) В сравнении W1-W3, основанном на девяти общих текстах, ни для одного семантического или стилометрического признака взаимодействие *режим* $\times$ *признак* не сохранило значимость после FDR-коррекции.

A



B



C

No FDR-significant Wave $\times$ Feature interactions
46 semantic + 12 stylometric features
9 shared texts
W1 vs W3

3.4. Perceived AI-likeness in Cohort 3 was organized by authorial-experiential presence versus schematic generality

The W3 ratings gave a different perspective on what readers were reacting to. Out of the 13 perceptual dimensions, 11 were linked to AI attribution after Benjamini-Hochberg correction, using separate crossed mixed-effects models that controlled for the texts' actual origin. AI attribution became more likely with greater generality, template-likeness, and obviousness. In contrast, stronger signals of authorial belief and emotion, along with personal and physical experience, author imaginability, vividness, pleasantness, and trust all lowered the odds of attributing the text to AI. Auditory and scene imagery, however, did not hold up after correction (see Supplementary Table S10).

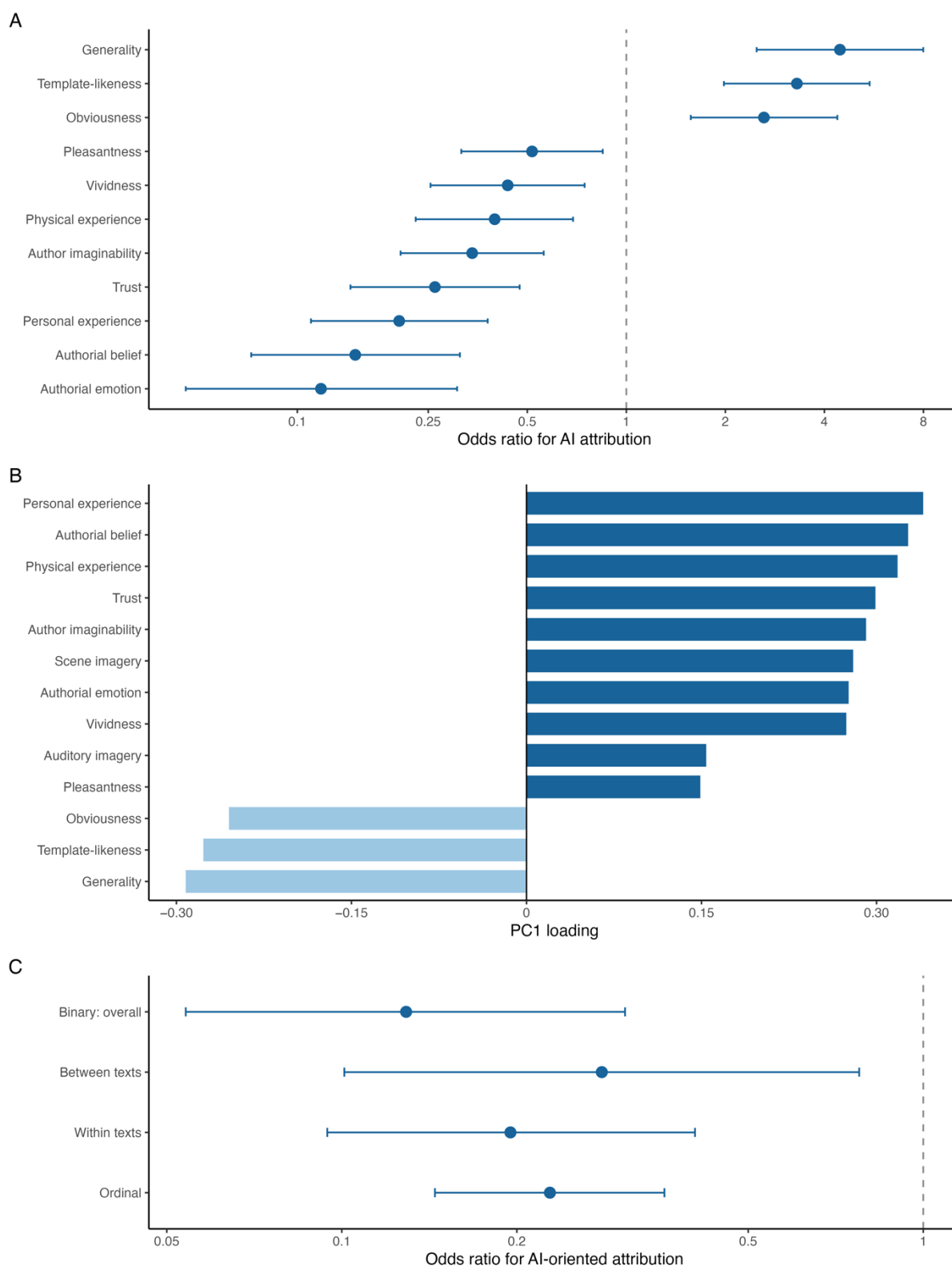
The pattern showed up at two different levels. At the between-text level, seven dimensions kept FDR-significant differences, while at the within-text level, 12 out of 13 displayed significant effects. On average, texts that readers perceived as more general or template-like drew more AI attributions. In contrast, texts that were seen as more

personally grounded, imaginable, emotionally situated, and trustworthy led to fewer AI attributions. Similar relationships also appeared across different readers evaluating the same stimulus, which suggests the finding wasn't just a side effect of consistent differences among the 10 texts (Supplementary Table S11).

A principal component analysis (PCA) on the within-linguist centered ratings captured this coordination using a single retained component. PC1 had an eigenvalue of 5.90, accounted for 45.4% of the variance, and was larger than the 95th-percentile null eigenvalue of 1.74 in permutation-based parallel analysis. Variables such as personal experience, authorial belief, physical experience, trust, author imaginability, scene imagery, authorial emotion, and vividness loaded positively. Meanwhile, generality, template-likeness, and obviousness loaded negatively. We interpret this component as reflecting a trade-off between authorial-experiential presence and schematic generality (Supplementary Table S12). Figure 5 summarizes the individual effects and the latent structure.

Figure 5. Perceptual organization of AI attribution in Cohort 3: A) Effect estimates for the 11 perceptual dimensions significantly associated with binary AI attribution after Benjamini-Hochberg correction. Odds ratios below 1 indicate lower odds of AI attribution and values above 1 higher odds; B) Loadings of all 13 perceptual ratings on the retained unsupervised component, oriented from schematic generality at the negative pole to authorial-experiential presence at the positive pole; C) Associations of the retained component with AI attribution in the primary binary model, its between-text and within-text decomposition, and the ordinal sensitivity analysis.

Рисунок 5. Перцептивная организация атрибуции ИИ в когорте 3: А) Оценки эффектов для 11 перцептивных шкал, которые сохранили значимую связь с бинарной атрибуцией ИИ после коррекции Бенджамини-Хохберга. Отношения шансов меньше 1 соответствуют снижению вероятности атрибуции ИИ, значения больше 1 – её увеличению; В) Нагрузки всех 13 перцептивных оценок на единственную сохранённую компоненту, полученную без учёта атрибуций; отрицательный полюс соответствует схематической обобщённости, положительный – авторскому присутствию; С) Связь сохранённой компоненты с атрибуцией ИИ в основной бинарной модели, при разложении эффекта на межтекстовую и внутритекстовую составляющие, а также в дополнительном анализе устойчивости с исходной порядковой шкалой.



Within W3, shifting toward the authorial-experiential end of the spectrum was closely linked to fewer AI attributions. In the main crossed mixed-effects model, a one-standard-deviation increase in PC1 lowered

the likelihood of being assigned an AI attribution (OR = .129, 95% CI [.054, .307], $p = 3.80 \times 10^{-6}$). By contrast, actual origin was not significant (OR = .849, $p = .863$). This relationship held both across the 10

different stimuli (OR = .280, 95% CI [.101, .777], $p = .0145$) and within individual texts across different readers (OR = .195, 95% CI [.094, .405], $p = 1.10 \times 10^{-5}$).

The same overall pattern emerged when the original five-category attribution response was kept. In a cumulative-link mixed model that used all 160 evaluations, higher PC1 again reduced the odds of making a more AI-oriented attribution (OR = .228, 95% CI [.145, .359], $p = 1.76 \times 10^{-10}$), while actual origin stayed non-significant (OR = .712, 95% CI [.262, 1.94], $p = .507$; Supplementary Table S12).

Importantly, the component was derived without using source labels or attribution outcomes, yet it was strongly associated with subsequent AI attribution both across texts and among readers evaluating the same text.

4. DISCUSSION

The study started with a basic inconsistency: a text can be measurably AI-like in its linguistic structure while still not appearing AI-generated to readers, and a human-written text can still draw AI judgments. The findings refine this mismatch. Human- and AI-produced texts differed clearly in how semantic information and stylistic features were distributed, but these origin-related aspects didn't translate straightforwardly into attribution. Moreover, the influence readers seemed to place on individual cues shifted across different conditions, and in some cases even reversed direction. In Cohort 3 specifically, however, AI attribution seemed to follow a perceptual dimension that ran from schematic generality toward a stronger sense of the author's presence and lived experience.

By placing objective indicators of origin alongside readers' perceived AI-likeness on the same texts, the difference becomes clear. Previous research has already shown that human and AI writing can diverge linguistically and that readers rely on systematic cues when forming judgments. Here, those two patterns could be compared directly. They did not simply align, and the

relationship between any given linguistic cue and AI attribution was not consistent across the different evaluation regimes.

4.1. Actual AI generation leaves a distributed linguistic signature, but readers do not simply perceive that signature

The objective-origin results align with a growing body of computational-linguistics research suggesting that human and LLM writing can still be told apart by how their language is organized at scale. Recent work finds differences in stylistic organization, information density, stance, functional appropriateness, and multivariate stylometric patterns (Mizumoto et al., 2024; Litvinova et al., 2025; O'Sullivan, 2025; Chen et al., 2026; Demir, 2026).

In our corpus, we saw this separation across two only partly overlapping spaces. On the semantic side, the main dimension reflected broad differences in alignment with bodily, sensory, spatial, temporal, and affective domains. On the stylometric side, the contrast was between function-word and pronominal orientation versus lexical density, word length, lexical diversity, and the noun-to-verb relationship. These semantic variables should be understood as coordinates in embedding space, not as simple counts of particular word types.

After correction for multiple testing, individual semantic and stylometric predictors were weak. At the multivariate level, however, the pattern was much stronger. Unsupervised latent dimensions showed large Human-AI differences that remained stable in leave-one-text-out analyses, and nested leave-one-text-out ridge models fitted to the original feature sets discriminated held-out texts with AUCs of .832 for semantics, .878 for stylometry, and .862 for the combined representation. Thus, in this corpus, information about source origin was distributed across combinations of features rather than concentrated in a single diagnostic marker.

The more revealing finding came when we compared these dimensions with human source judgments. Once actual origin was accounted for, neither the semantic nor the stylometric axis on its own could explain AI attribution. Even the apparent match in W2 vanished in origin-adjusted and within-origin analyses. In other words, a computationally detectable Human-AI signal exists, but readers do not seem to reproduce it directly in their judgments.

This gap fits with research showing that people use systematic but imperfect heuristics for authorship. Jakesch et al. (2023) showed that cues people treat as diagnostic of AI authorship may not reflect their true underlying distribution, and more recent studies suggest that judgments rely on combinations of writing style, personal experience, genre expectations, and communicative context rather than a fixed checklist of formal indicators (Hau et al., 2026; Liu et al., 2026).

The takeaway isn't just that people sometimes misclassify AI. This distinction was also visible at the level of individual texts. A human-written text could appear AI-like without necessarily falling on the AI-associated side of objective linguistic space, while an AI-generated text could occupy a more Human-associated region and be judged accordingly. Objective origin discrimination and perceived AI-likeness should therefore be treated as distinct analytical tasks.

4.2. Linguistic cue weighting depended on the evaluation regime

Cross-regime comparison revealed a second kind of instability: the same linguistic cue did not consistently map onto the same attributional value across all settings. In the matched W2-W3 group, seven objective features produced FDR-significant Wave $\times$ Feature interactions, and each one flipped direction. Six of those reversals held up under every leave-one-text-out deletion, while the seventh held up in 9 out of 10. Mean word length also reversed between W1 and W2 and then stayed reversed across all 27 stimulus

deletions. No W1-W3 interaction survived FDR correction, though with a nine-text overlap there's limited statistical power to treat that null result as evidence of similarity.

This goes beyond a simple difference in overall detection accuracy. The attributional meaning of a feature itself shifted. In W2, using more function words went with AI attribution, whereas higher lexical density and longer words went with Human attribution – matching the same directional pattern observed in the objective stylometric origin contrast. In W3, all three of these relationships reversed. The texts hadn't changed; what changed was the link between their properties and the judgments.

This pattern aligns with accounts of authorship attribution that treat it as contextually framed rather than driven by a fixed set of diagnostic markers. Hau et al. (2026) argue that readers integrate textual cues with expectations about the communicative domain, the sender, and the broader context, and that surface cues can become misleading when they're used outside the conditions where they're diagnostic. Wang et al. (2026) similarly found large variation in Human-AI discrimination across languages and domains, and showed that knowledge about relevant Human-machine differences can shift detection performance.

W2 also deserves caution because it followed an informational presentation about semantic differences in the corpus. One plausible interpretation is that the presentation "corrected" participants' perceptual template and adjusted cue weighting so it better matched the objective origin signature. However, the design cannot confirm that conclusion. W2 and W3 also used different participant cohorts and different procedures, and W3 included perceptual ratings before source attribution. What the data support is a narrower claim: cue weighting differed across regimes, with the most prominent feature-level divergence occurring between W2 and W3.

Earlier studies already show that source-attribution performance varies across

genre, language, and informational context. Our results extend that context dependence to individual features. A measurable property of the text can change not only the strength of its association with AI attribution, but also the direction of that association.

This matters for how we understand perceived AI-likeness. A detector-like account would predict that features contribute with fairly stable weights to an underlying AI score. Our findings don't match that idea. The same property took on different attributional value under different evaluation conditions, which fits better with context-sensitive use of cues than with a single regime-invariant profile.

4.3. Perceived AI-likeness may reflect the perceived presence of a situated author

The perceptual ratings from Cohort 3 point to what readers may have been reacting to instead. In this set of 10 texts, AI attribution went up when the writing felt more general, more like a template, and more obviously constructed. It went down when readers thought it reflected personal and physical experience, authorial belief and emotion, a believable imagined author, vividness, and trust. These ratings all seemed to align along a main dimension we labeled *Authorial-experiential presence versus schematic generality*.

Importantly, this component wasn't built directly from the source labels or the attribution outcomes. Still, it was able to predict later attribution across the 10 stimuli, including among different readers who evaluated the same text. And the relationship held even when the original five-category response was examined. So, the effect can't simply be explained by the actual Human/AI origin, or by stable differences between the texts by themselves.

This view also fits with a growing body of work on authorship heuristics. Jakesch et al. (2023) found that readers treat personal and social content, among other cues, as evidence of Human authorship – even when those heuristics aren't reliable. Liu et al.

(2026) similarly identify references to personal experience as a recurring cue in judgments about perceived AI authorship. And Hau et al. (2026) argue that source attribution reflects a combination of textual evidence and expectations about the sender, genre, and communicative context, rather than relying on a fixed checklist.

One way to interpret our W3 pattern is that Human-likeness depends in part on whether a text supports the sense that it comes from a particular, situated source of experience and evaluation. Generality and template-likeness don't have to be direct "AI markers". Instead, they may reduce the impression that the discourse is coming from a specific communicator who is experiencing, believing, and emotionally situated. By contrast, personal experience, belief, emotion, vividness, and how imaginable the author feels may strengthen that impression.

This is a relational interpretation rather than an essentialist one. The data don't show that AI-generated writing inherently lacks authorial presence, or that human writing necessarily includes it. In our corpus, there are AI-like human texts and Human-like AI texts, which runs against the idea that authorial presence is intrinsically determined. Here, authorial presence is something readers perceive. A text written by a human may fail to evoke a plausible situated author, while an AI-generated text may succeed.

This distinction also helps explain why human attribution and computational origin classification can sometimes diverge. A classifier may take advantage of distributed statistical differences between the source populations without actually representing an "author." Readers may instead be implicitly asking whether the discourse feels like it was produced by a situated Human communicator. In that sense, perceived AI-likeness may involve an inference about communicative authorship rather than a direct readout of the statistical signature of AI generation.

4.4. From AI detection to attribution risk: why perceived AI-likeness matters

Actual origin and perceived AI-likeness are important for different reasons. Detecting origin focuses on whether AI produced the text. Perceived AI-likeness, in contrast, focuses on why, for a given readership and under specific conditions, people are more likely to attribute the text to AI.

Blending these two questions is especially risky in high-stakes judgments of authorship. Even in professional or academic settings, human evaluators can make significant mistakes in how they attribute authorship (Fiedler and Döpke, 2025), and having high confidence doesn't necessarily prevent people from leaning on misleading signals (Hau et al., 2026). At the same time, stylometric differences observed at the level of whole corpora cannot, on their own, justify conclusions about the origin of any single text (O'Sullivan, 2025).

Our results point to an additional, complementary concern: false AI attribution can be systematic within a particular evaluation setup, not just an accidental error. Human-authored texts differed in how often they were judged to be AI-written, and some looked strongly AI-like under certain conditions. As a result, perceived AI-likeness can still matter even when the true source is known to be Human.

This asymmetry is not unique to our data. Recent studies have likewise found that human-authored texts can be misclassified as AI-generated more often than AI texts are mistaken for Human authorship (e.g., Szabó et al., 2026). Viewed from the perspective of perceived AI-likeness, such errors may be informative rather than merely residual: some human texts may systematically evoke an AI interpretation.

There are also real-world reasons to worry about this mismatch. When a text is labeled as AI-authored, it can reduce message and source credibility, the perceived competence of the author, the judged quality of the work, and people's willingness to use or share it (Lermann Henestrosa and

Kimmerle, 2024; Proksch et al., 2024). Perceived AI contribution has likewise been linked to lower message and source credibility (Jia et al., 2024), and readers tend to evaluate narratives and poetry less favorably when they believe the texts were generated by AI (Marklová et al., 2026; Szabó et al., 2026).

Altay and Gilardi (2024) offer a particularly clear example of the downside of false AI attribution: human-written headlines were seen as less accurate and people were less willing to share them when the headlines were labeled as AI-generated. So the cost of an AI attribution doesn't require the text to be actually produced by AI. A human author may be penalized simply because the text is believed to be AI-produced.

This means AI-attribution risk is a different practical issue than detector accuracy. The real question isn't only whether a source classifier gives the correct answer, but whether a text is vulnerable to AI attribution within a particular evaluation context – and what results from that attribution judgment.

The link to trust in W3 is suggestive, but it should not be taken too far. Trust, which loaded on the authorial-experiential end of the scale, was negatively associated with later AI attribution. Since trust was measured before source attribution and perceived source wasn't experimentally manipulated, we can't claim that perceiving AI authorship actually lowered trust. What we can support is narrower: lower trust and stronger AI attribution appeared together in the same perceptual pattern in W3.

A logical next step is not simply to establish whether perceptual impressions precede source attribution: W3 already showed that the perceptual ratings were made before the source judgment and were strongly associated with it. What remains unresolved is what happens after an AI attribution has been formed. An experiment separating actual origin, presented origin, and spontaneously emerging perceived AI-likeness could test whether false-AI attribution itself changes

credibility or trust. By holding the text constant while manipulating the presented source – and including a blind condition – it would be possible to distinguish effects of linguistic form from the downstream effects of source belief.

If it does, then the practical challenge goes beyond detection alone. It would become useful to identify unwanted AI-likeness in cases where source perception affects credibility, trust, or communicative effectiveness. Our findings across different regimes also argue against relying on a single AI-likeness score that works independently of context. Any practical assessment would need to be calibrated for a specific audience and a specific evaluative setting.

4.5. Limitations and future directions

The study has clear limitations. The objective linguistic analyses were based on 28 texts, while the detailed W3 perceptual analysis relied on only 10. The repeated W3 evaluations enabled within-text analyses, but the Authorial-experiential presence versus schematic generality dimension now requires replication using a larger, more varied set of stimuli.

The corpus also varied in terms of topic and communicative context. The Experience, Opinion, and Emotion variables offered only broad control over content differences, so the semantic and stylometric origin signatures should be viewed as characteristics of this particular corpus, not as universal indicators of AI-generated language.

In addition, the participants were professional linguists. Their focus on textual form and their expectations regarding AI-generated writing may differ from those of other reader groups, so the perceptual patterns reported here should not be assumed to generalize unchanged to broader populations.

Lastly, the three regimes differed at the same time in participant cohort, procedure, informational context, and judgment sequence. They were not randomized levels of a single manipulation. As a result, the cross-regime findings document cue reweighting

linked to the different regimes, but they do not show which specific element caused it. A more rigorous causal test would keep the stimuli and participant population constant while manipulating the evaluative frame or source information.

CONCLUSION

Perceived AI-likeness couldn't be traced back in a simple way to the actual source of the text, or to any fixed set of linguistic markers. Human-written and AI-generated texts showed strong differences in the way meaning was distributed and in their stylometric profiles, which means that real authorship did leave a measurable linguistic imprint in this corpus. However, the key dimensions that most clearly distinguished the two source categories didn't, by themselves, fully account for AI attribution.

Individual cues also didn't carry a stable attributional meaning across different evaluation settings. Several semantic and stylometric indicators even flipped their direction: the same measurable feature could be linked to AI attribution in one regime and to Human attribution in another. So, an objective signal distinguishing Human from AI doesn't automatically translate into a fixed perceptual "code" that people use when judging the source.

Cohort 3 suggests a different organizing principle. Here, AI attribution seemed to follow a latent contrast between an author's lived, experiential presence and more schematic, generalized writing. Readers were less likely to label a text as AI-generated when they could picture an author who believes in what they're saying – someone emotionally engaged and grounded in real experience. They were more likely to label it as AI-generated when the writing felt generalized, obvious, and template-like. Importantly, this sense of authorial presence is a perception, not a guarantee about the text's true origin.

As a result, perceived AI-likeness is best understood in this study as a property of the text as it is perceived, not as a

straightforward proxy for machine authorship. The study separates that perceptual property empirically from actual origin and identifies one way in which it was organized in readers' experience. Future work should therefore move beyond linguistic markers of actual AI origin and examine the linguistic signature of perceived AI-likeness itself: which textual features make writing seem AI-generated, how these cues vary across genres, communicative settings, and characteristics of authors and readers, and how they relate to trust, credibility, and later evaluation. An especially important question is what happens when a text is believed to be AI-generated, whether that belief is correct or not.

Statement on the Use of Generative AI:

During the preparation of this manuscript, the authors used Wordvice AI (Wordvice, model/version not specified by the provider) for English-language proofreading, including spelling, grammar, and stylistic editing. The tool was used in relation to the English-language text of the manuscript and did not perform data analysis or make methodological or interpretive decisions. All AI-assisted edits were critically reviewed and, where necessary, revised by the authors. The authors verified the accuracy of factual information, references, and other relevant content and take full responsibility for the final version of the manuscript.

Supplementary materials: Supplementary Tables S1–S12, together with the original Russian W3 text-perception questionnaire and its English translation, are publicly available at: <https://github.com/Litvinova1984/when-human-writing-looks-artificial-supplement/releases/tag/v1.0>

Заявление об использовании генеративного ИИ: При подготовке настоящей рукописи авторы использовали Wordvice AI (Wordvice; модель/версия разработчиком не указана) для языковой вычитки англоязычного текста, включая исправление орфографии, грамматики и стилистики. Инструмент применялся только при редактировании англоязычного

текста рукописи и не использовался для анализа данных, принятия методологических решений или интерпретации результатов. Все изменения, внесённые с использованием системы, были критически проверены и при необходимости отредактированы авторами. Авторы проверили достоверность фактической информации, ссылок и иных элементов содержания и несут полную ответственность за окончательную версию рукописи.

Дополнительные материалы:

Дополнительные таблицы S1–S12, а также оригинальный русский опросник W3 по восприятию текста и его английский перевод доступны для публичного просмотра по адресу: <https://github.com/Litvinova1984/when-human-writing-looks-artificial-supplement/releases/tag/v1.0>

References

- Altay, S. and Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation, *PNAS Nexus*, 3 (10), pgae403. <https://doi.org/10.1093/pnasnexus/pgae403> (In English)
- Chen, R., Xiong, S., He, J. and Ross, G. J. (2026). Stylometric detection of AI-generated texts: evidence from human and machine-written essays, *Digital Scholarship in the Humanities*, fqag064. <https://doi.org/10.1093/llc/fqag064> (In English)
- Demir, N. Y. (2026). Human perceptions of register alignment and functional appropriateness: ChatGPT-generated vs human-authored texts, *Digital Scholarship in the Humanities*, fqag084. <https://doi.org/10.1093/llc/fqag084> (In English)
- Fiedler, A. and Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education*, 49, 100321. <https://doi.org/10.1016/j.iree.2025.100321> (In English)
- Hau, M. F., Vulpius, J. and Tønnesen, S. (2026). AI literacy as metacommunicative agility: exploring users' interpretive strategies in detecting and evaluating AI and human

authorship, AI & Society.
<https://doi.org/10.1007/s00146-026-03125-3> (In English)

Jakesch, M., Hancock, J. T. and Naaman, M. (2023). Human heuristics for AI-generated language are flawed, in *Proceedings of the National Academy of Sciences*, 120 (11), e2208839120.

<https://doi.org/10.1073/pnas.2208839120> (In English)

Jia, H., Appelman, A., Wu, M. and Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, 2 (2), 100093. (In English)

<https://doi.org/10.1016/j.chbah.2024.100093>

Latini, N., Bråten, I. and Strømsø, H. I. (2026). Mind Over Model? Students' Evaluation and Use of ChatGPT-Generated Versus Human-Generated Texts. *Reading Research Quarterly*, 61 (1), e70087. <https://doi.org/10.1002/rrq.70087> (In English)

Lermann Henestrosa, A. and Kimmerle, J. (2024). The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text, *Journalism and Media*, 5 (3), 1085–1097. <https://doi.org/10.3390/journalmedia5030069> (In English)

Litvinova, T., Molchanova, V., Dekhnich, O., Seredin, P. (2025). Describing Images: Linguistic Signatures of AI and Humans, in *Proceedings of the 38th Conference of Open Innovations Association FRUCT*, Helsinki, Finland. (In English)

Litvinova, T., Zavarzina, V., Panicheva, P. (2024). Individual Differences in the Most Frequent Content Word Usage as a New Type of Features in the Authorship Profiling Task, in *Proceedings of The International Society for Technology Education and Science*, ICRES 2024, 1901–1924. (In English)

Liu, S., Chu, H., Yuan, S., Sun, Y. and Men, L. R. (2026). The more you think you know: AI experience, literacy, heuristics, and perceived AI authorship. *AI & Society*, 41 (3), 2199–2211. <https://doi.org/10.1007/s00146-025-02643-w> (In English)

Marklová, A., Vinš, O., Vokáčová, M. and Milička, J. (2026). The author is dead, but what if they never lived? A reception experiment on Czech AI- and human-authored poetry. *Digital Scholarship in the Humanities*, fqag067. <https://doi.org/10.1093/llc/fqag067> (In English)

Markowitz, D. M., Hancock, J. T. and Bailenson, J. N. (2024). Linguistic Markers of Inherently False AI Communication and Intentionally False Human Communication: Evidence From Hotel Reviews. *Journal of Language and Social Psychology*, 43 (1), 63–82. <https://doi.org/10.1177/0261927X231200201> (In English)

Mizumoto, A., Yasuda, S. and Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints, *Applied Corpus Linguistics*, 4 (3), 100106. <https://doi.org/10.1016/j.acorp.2024.100106> (In English)

O'Sullivan, J. (2025). Stylometric comparisons of human versus AI-generated creative writing, *Humanities and Social Sciences Communications*, 12, 1708. <https://doi.org/10.1057/s41599-025-05986-3> (In English)

Proksch, S., Schühle, J., Streeb, E., Weymann, F., Luther, T. and Kimmerle, J. (2024). The impact of text topic and assumed human vs. AI authorship on competence and quality assessment, *Frontiers in Artificial Intelligence*, 7, 1412710.

<https://doi.org/10.3389/frai.2024.1412710> (In English)

Szabó, G., Krizsai, F., and Deme, A. (2026). The invisible author: Citizen sociolinguistic perspectives on identifying human and AI-generated narrative texts, *Social Sciences & Humanities Open*, 13, 102646. <https://doi.org/10.1016/j.ssaho.2026.102646> (In English)

Stoltz, D. S., Taylor, M. A. (2022). text2map: R Tools for Text Matrices, *Journal of Open Source Software*, 7 (72), 3741. (In English)

Taylor, M. A., Stoltz, D. S. (2021). Integrating Semantic Directions with Concept Mover's Distance to Measure Binary Concept Engagement, *Journal of Computational Social Science*, 4, 231–242. (In English)

Tudino, G. and Qin, Y. (2024). A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences, *Lingua*, 312, 103838.

<https://doi.org/10.1016/j.lingua.2024.103838> (In English)

Wang, Y., Xing, R., Mansurov, J. et al. (2026). Is Human-Like Text Liked by Humans?

Multilingual Human Detection and Preference Against AI, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 14043–14076. <https://doi.org/10.18653/v1/2026.acl-long.639> (In English)

Список литературы

Altay S. People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation / S. Altay, F. Gilardi // *PNAS Nexus*. 2024. T. 3. № 10. Art. pgae403. DOI: 10.1093/pnasnexus/pgae403

Chen R. Stylometric detection of AI-generated texts: evidence from human and machine-written essays / R. Chen, S. Xiong, J. He, G. J. Ross // *Digital Scholarship in the Humanities*. 2026. Art. fqag064. DOI: 10.1093/llc/fqag064

Demir N. Y. Human perceptions of register alignment and functional appropriateness: ChatGPT-generated vs human-authored texts / N. Y. Demir // *Digital Scholarship in the Humanities*. 2026. Art. fqag084. DOI: 10.1093/llc/fqag084

Fiedler A. Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses / A. Fiedler, J. Döpke // *International Review of Economics Education*. 2025. T. 49. Art. 100321. DOI: 10.1016/j.iree.2025.100321

Hau M. F. AI literacy as metacommunicative agility: exploring users' interpretive strategies in detecting and evaluating AI and human authorship / M. F. Hau, J. Vulpius, S. Tønnesen // *AI & Society*. 2026. DOI: 10.1007/s00146-026-03125-3

Jakesch M. Human heuristics for AI-generated language are flawed / M. Jakesch, J. T. Hancock, M. Naaman // *Proceedings of the National Academy of Sciences*. 2023. T. 120. № 11. Art. e2208839120. DOI: 10.1073/pnas.2208839120

Jia H. News bylines and perceived AI authorship: Effects on source and message credibility / H. Jia, A. Appelman, M. Wu, S. Bien-Aimé // *Computers in Human Behavior: Artificial Humans*. 2024. T. 2. № 2. Art. 100093. DOI: 10.1016/j.chbah.2024.100093

Latini N. Mind Over Model? Students' Evaluation and Use of ChatGPT-Generated Versus Human-Generated Texts / N. Latini,

I. Bråten, H. I. Strømsø // *Reading Research Quarterly*. 2026. T. 61. № 1. Art. e70087. DOI: 10.1002/rrq.70087

Lermann Henestrosa A. The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text / A. Lermann Henestrosa, J. Kimmerle // *Journalism and Media*. 2024. T. 5. № 3. C. 1085–1097. DOI: 10.3390/journalmedia5030069

Litvinova T. Describing Images: Linguistic Signatures of AI and Humans / T. Litvinova, V. Molchanova, O. Dekhnich, P. Seredin // *Proceedings of the 38th Conference of Open Innovations Association FRUCT*. Helsinki, Finland, 2025.

Litvinova T. Individual Differences in the Most Frequent Content Word Usage as a New Type of Features in the Authorship Profiling Task / T. Litvinova, V. Zavarzina, P. Panicheva // *Proceedings of The International Society for Technology Education and Science, ICRES 2024*. 2024. Pp. 1901–1924.

Liu S. The more you think you know: AI experience, literacy, heuristics, and perceived AI authorship / S. Liu, H. Chu, S. Yuan, Y. Sun, L. R. Men // *AI & Society*. 2026. T. 41. № 3. Pp. 2199–2211. DOI: 10.1007/s00146-025-02643-w

Marklová A. The author is dead, but what if they never lived? A reception experiment on Czech AI- and human-authored poetry / A. Marklová, O. Vinš, M. Vokáčová, J. Milička // *Digital Scholarship in the Humanities*. 2026. Art. fqag067. DOI: 10.1093/llc/fqag067

Markowitz D. M. Linguistic Markers of Inherently False AI Communication and Intentionally False Human Communication: Evidence From Hotel Reviews / D. M. Markowitz, J. T. Hancock, J. N. Bailenson // *Journal of Language and Social Psychology*. 2024. T. 43. № 1. Pp. 63–82. DOI: 10.1177/0261927X231200201

Mizumoto A. Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints / A. Mizumoto, S. Yasuda, Y. Tamura // *Applied Corpus Linguistics*. 2024. T. 4. № 3. Art. 100106. DOI: 10.1016/j.acorp.2024.100106

O'Sullivan J. Stylometric comparisons of human versus AI-generated creative writing / J. O'Sullivan // *Humanities and Social Sciences Communications*. 2025. T. 12. Art. 1708. DOI: 10.1057/s41599-025-05986-3

Proksch S. The impact of text topic and assumed human vs. AI authorship on competence and quality assessment / S. Proksch, J. Schühle, E. Streeb, F. Weymann, T. Luther, J. Kimmerle // *Frontiers in Artificial Intelligence*. 2024. T. 7. Art. 1412710. DOI: 10.3389/frai.2024.1412710

Szabó G. The invisible author: Citizen sociolinguistic perspectives on identifying human and AI-generated narrative texts / G. Szabó, F. Krizsai, A. Deme // *Social Sciences & Humanities Open*. 2026. T. 13. Art. 102646. DOI: 10.1016/j.ssaho.2026.102646

Stoltz D. S. text2map: R Tools for Text Matrices / D. S. Stoltz, M. A. Taylor // *Journal of Open Source Software*. 2022. T. 7. № 72. Art. 3741

Taylor M. A. Integrating Semantic Directions with Concept Mover's Distance to Measure Binary Concept Engagement / M. A. Taylor, D. S. Stoltz // *Journal of Computational Social Science*. 2021. T. 4. Pp. 231–242.

Tudino G. A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences / G. Tudino, Y. Qin // *Lingua*. 2024. T. 312. Art. 103838. DOI: 10.1016/j.lingua.2024.103838

Wang Y. Is Human-Like Text Liked by Humans? Multilingual Human Detection and Preference Against AI / Y. Wang, R. Xing, J. Mansurov [et al.] // *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*. Vol. 1, Long Papers. 2026. Pp. 14043–14076. DOI: 10.18653/v1/2026.acl-long.639

All authors have read and approved the final manuscript.

Все авторы прочитали и одобрили окончательный вариант рукописи.

Conflicts of interests: the authors have no conflicts of interest to declare.

Конфликты интересов: у авторов нет конфликтов интересов для декларации.

Татьяна Александровна Литвинова, доктор филологических наук, профессор кафедры возрастной и социальной психологии, ведущий научный сотрудник, Белгородский государственный университет, Белгород, Россия.

Tatiana A. Litvinova, Doctor of Philological Sciences (D.Sc. in Philology), Professor, Department of Developmental and Social Psychology; Leading Research Fellow, Belgorod State National Research University, Belgorod, Russia.

Ольга Витальевна Дехнич, кандидат филологических наук, доцент, доцент кафедры английской филологии и межкультурной коммуникации Белгородского государственного национального исследовательского университета, Россия.

Olga V. Dekhnich, Candidate of Philological Sciences (Cand. in Philology), Associate Professor of the Department of English Philology and Cross-Cultural Communication, Belgorod State National Research University, Belgorod, Russia.