# Human or AI? An Interpretable Method for Auditing Free-Text Data in Online Psychological Research

Joanna Kuc[1*], Anthony Hills[2], Greg Cooper[3], Mahmud Elahi Akhter[2],

Talia Tseriotou[2], Maria Liakata[2,4], Daniel R Lametti[1,5], Jeremy I Skipper[1]

[1] Department of Experimental Psychology, University College London.

[2]Queen Mary University of London, London, United Kingdom

[3]Department of Clinical, Educational & Health Psychology, University College London

[4]The Alan Turing Institute, London, United Kingdom

[5]Department of Psychology, Acadia University, Wolfville, Canada

*Corresponding author email address: joanna.kuc.22@ucl.ac.uk

Online psychological studies increasingly rely on free-text responses, but the availability of generative AI creates new challenges for evaluating the authenticity and integrity of data pertaining to human introspection. We present an interpretable text-analysis workflow for identifying linguistic cues associated with perceived AI authorship in digital journal entries. Across a two-week journalling study, 251 participants submitted 2,808 text entries, which were independently labelled by human annotators as human-written, AI-generated or invalid, and compared with classifications from a commercial AI-detection tool. We extracted 127 transparent features capturing length, semantic coherence, stylistic consistency, readability, natural-language-inference-based consistency, lexical overlap and Linguistic Inquiry Word Count 2022 psycholinguistic categories. On a high-confidence subset defined by convergence between human judgements and automated detection, interpretable classifiers strongly distinguished entries perceived as human from those perceived as AI-generated. A decision tree achieved 95.3% accuracy using a small set of auditable rules based on verb rate, syllable count, article use and sentence count, while a linear SVM identified 18.6% of entries as ambiguous cases suitable for human review. Entries perceived as AI-generated were longer, more coherent, more stylistically uniform and more formally structured, whereas entries perceived as human showed greater present focus, self-reference and stylistic variability. These findings provide a transparent framework for auditing possible AI mediation in remote psychological text data, while showing how interpretable models can support human review rather than replace it.

**Generative AI;** Online psychological research; Free-text data; Digital journalling; Fraudulent participation; Text analysis; Research integrity

# Introduction

Online psychological studies, especially in the domain of digital mental health, increasingly integrate free-text responses collected via smartphone or web apps (Torous et al., 2025). The growing popularity of modern natural language processing (NLP) approaches, including large language models (LLMs) and, more broadly, generative artificial intelligence (GenAI), offers significant opportunities in capturing and interpreting such large-scale data, moving beyond the typically used structured questionnaires towards free-text responses (Demszky et al., 2023). This transition is where the value for mental health research emerges: participants can describe their experiences in their own words and in whatever form feels most natural to them, capturing emotional nuance and meaning-making that a fixed set of questionnaire items may flatten or miss entirely (McCombie et al., 2024). Daily diaries, expressive writing tasks, open-ended survey responses, blogs or social media reflections and other forms of introspective text provide rich information about participants' thoughts, feelings and experiences (Pennebaker & King, 1999). Such text has been used extensively in computational approaches to mental health, including the detection of depression, anxiety and suicidal ideation from online content (Chancellor & De Choudhury, 2020; Coppersmith et al., 2014; De Choudhury et al., 2021), as well as the longitudinal tracking of individuals' psychological states over time (Hills et al., 2024; Tsakalidis et al., 2022; Tseriotou et al., 2025). In remote-monitoring contexts, these self-generated accounts may provide a complementary source of psychological information to passively collected behavioural data, because they capture how participants interpret and describe their own experiences (Jacobson et al., 2019; Kappen et al., 2023). Nevertheless, the growing accessibility of GenAI also introduces new methodological challenges for researchers collecting such data remotely, particularly in paid studies, where participants may be motivated by incentives beyond a genuine desire to contribute to research. When LLMs are available to compose, edit, or replace responses to study questions, researchers can no longer assume that introspective text straightforwardly reflects unmediated human self-expression.

This challenge is particularly important for journalling and other forms of personal reflection. First, the therapeutic and self-regulatory benefits of expressive writing (Sohal et al., 2022) are thought to arise from genuine self-disclosure and sense-making, rather than from producing polished prose (Pennebaker, 1997). Second, in research contexts, diary-style responses are often treated as behavioural or psychological data: they may be used to infer emotional states, cognitive processes, coping strategies, or changes over time (Jung et al., 2024; Kelly et al., 2021; Nepal et al., 2024). If some entries are generated or heavily mediated by AI, this may compromise the validity of downstream analyses. At the same time, simply excluding responses flagged by commercial AI detectors can be problematic. Existing detectors (e.g., GPTZero, Turnitin, DetectGPT) were developed primarily for formal or academic genres (Weber-Wulff et al., 2023), and their performance on personal, introspective writing remains unclear. Furthermore, the proprietary nature and lack of transparency not only incurs unnecessary research costs but also makes it more difficult for researchers to understand why a given text has been classified as AI-generated (Dik et al., 2025; Liang et al., 2023).

Most work on AI-generated text has framed the problem in terms of detection accuracy: can a model, tool, or human judge correctly distinguish AI-written from human-written text? This framing is useful but incomplete. In many psychological research settings, the relevant methodological question is not only whether a text can be classified as AI-generated, but which observable features make it appear more or less likely to reflect authentic human expression. A transparent account of these features would allow researchers to audit free-text datasets, identify ambiguous cases for human review, and report inclusion decisions more clearly. It would also reduce dependence on opaque commercial systems whose training data, decision rules, and error patterns are not fully available for inspection and may change with updates to software.

The present study addresses this methodological gap by developing an interpretable text-analysis workflow for identifying linguistic cues associated with perceived AI authorship in a digital journaling dataset. As part of a separate project on language and mental health (Kuc et al., 2023), we ran a two-week journalling study in which participants submitted daily entries via a custom

Telegram chatbot. During data collection, a subset of entries appeared to involve generative AI use, including some that retained ChatGPT-style headers or prompts, and several participants acknowledged using AI when contacted. This observation motivated a systematic evaluation of participation integrity and led to the present analysis. A group of university students were asked to independently label chunks of the anonymised dataset as human-written, AI-generated, or invalid. We then compared these labels with classifications from a commercial AI-detection tool and constructed a high-confidence subset in which human judgements and automated probabilities converged. No ground-truth authorship labels exist for this dataset, as entries were not experimentally generated by AI under known conditions, so all analyses necessarily model perceived rather than verified AI authorship. We partially address this by requiring convergence between two independent signals, human judgement and automated detection, rather than relying on either alone; this subset therefore provides a more defensible basis for modelling linguistic features associated with strong perceived authorship judgements than either signal in isolation, while a full-dataset robustness analysis allows us to assess how performance changes when more ambiguous cases are included. For each entry, we extracted a broad set of interpretable features capturing length, semantic coherence, stylistic consistency, lexical overlap, readability, natural-language-inference-based consistency, and psycholinguistic categories from Linguistic Inquiry Word Count, 2022 version (LIWC-22). We then trained different meta-classifiers on this unified feature set to aid with the interpretable prediction.

In summary, the present paper makes three contributions. First, it characterises human and automated judgements of AI authorship in a large corpus of digital journal entries. Second, it identifies interpretable linguistic features associated with perceived human versus AI authorship in introspective text. Third, it evaluates a transparent classification and triage workflow that can support practical auditing of free-text data in psychological research. By focusing on interpretability, ambiguity and human review, the proposed method offers a more cautious alternative to opaque AI-detection tools and a practical framework for researchers working with sensitive, personal, and remotely collected text data.

# Methods

## Ethical statement

Ethics approval for the current study was provided by the UCL Ethics Committee under the reference number EP_2023_010. All participants provided informed consent before taking part. Participation was voluntary, and participants were informed that they could withdraw from the study in accordance with the approved study procedures. The study was conducted in line with the ethical standards of UCL and with appropriate safeguards for participant privacy and confidentiality. All data were handled securely and analysed only for the purposes described in the approved research protocol.

## Study context, participants and procedure

The present dataset was collected as part of a separate project on language and mental health (Kuc et al., 2023). We ran a two-week online journalling study in which participants submitted daily entries via a custom Telegram chatbot. Recruitment took place on social media platforms, including Twitter, Facebook and Instagram, relevant Reddit communities, including r/PaidStudies, r/Journaling and r/DigitalJournaling, and the Sona Systems, which is a participant pool management platform research, which enables researchers to recruit students registered within their institution's subject pool to participate in studies, typically in exchange for course credit. Eligible participants were adults aged 18 years or older who used a smartphone and were proficient in English. Participants who met these criteria received a unique user ID, registered with the chatbot, and were asked to provide one journal entry per day for 14 days. Participants who completed all surveys and submitted more than 60% of the required entries were offered a choice of a £25 shopping voucher or university course credits. The Telegram chatbot, “Journalling Study Bot”, was built on the OneReach.ai Generative Studio X platform[1] and provided a conversational interface for data collection. Generative Studio X is a no-code/low-code AI orchestration platform

[1] https://onereach.ai/enterprise/

designed for building multi-channel conversational agents, supporting integrations with large language models and messaging platforms such as Telegram. Access to the platform was provided through an academic fellowship awarded by OneReach.ai, which supported the development of this research tool. On enrolment, users selected a preferred reminder time for the 14-day period. Participants could respond with text or voice; voice responses, however, were excluded from the present analysis for two reasons. First, voice notes required transcription before linguistic features could be extracted, introducing additional processing decisions around punctuation and sentence segmentation. They were also considered less likely to reflect AI-composed submissions, meaning their inclusion would have introduced a different authorship and modality profile. The analysis therefore focused exclusively on typed entries. Each daily prompt invited an unstructured reflective response:

*"It is time for your daily journal entry! [USERNAME], take a moment to reflect on your day before sharing your thoughts, feelings, and experiences."*

In total, 264 participants enrolled in the study, of whom 263 contributed 3,069 entries after excluding duplicate and empty submissions. Out of those 261 were voice notes, which were excluded. The final dataset comprised 2,808 text entries from 251 participants. On average, each participant submitted 11.7 entries, with a mean of 116.0 words, 7.7 sentences and 638.2 characters.

## Text anonymisation

We anonymised the journal texts by replacing detected named entities with bracketed tags using a multilingual bert-base named-entity-recognition model[2]. A multilingual model was chosen as a conservative choice, as participants were not required to be native English speakers or reside in English-speaking countries, and named entities in diary text may reflect a variety of linguistic and cultural backgrounds. Each identified span was substituted with its entity-group label, for example [PER], [LOC] or [ORG], while preserving punctuation and word order. Replacement was

[2] https://huggingface.co/Davlan/bert-base-multilingual-cased-ner-hrl

case-insensitive. The resulting corpus retained linguistic form while reducing the presence of direct identifiers.

## Human annotation protocol

We recruited 20 UCL students to complete a human annotation task, alongside the researcher JK, for a total of 21 annotators.. The 2,808 entries were partitioned into three approximately equal blocks (~900, ~940, and ~970 entries), and each student was assigned to annotate one block, with 6-7 per block; the author JK additionally annotated a subset of entries across multiple blocks. The students were instructed to read each entry and select one of three categories: human, AI-generated, or invalid. The invalid category was used for responses that did not constitute journal entries, such as questions to the bot mistakenly saved as entries, or minimal messages such as “ok” or “thank you”.

## Label aggregation

We aggregated the human ratings for each entry using simple majority rules. An entry was labelled human if more than half (>50%) of its raters selected human, and labelled AI-generated if more than half selected AI-generated. For example, this required at least 4 of 6 raters, 4 of 7 raters, or 5 of 8 raters to agree on the same label. Entries without a strict majority for either human or AI-generated, as well as entries with a majority invalid label, were treated as invalid and excluded from subsequent analyses. This aggregation procedure was chosen to provide a transparent and reproducible labelling rule.

## Automated AI-detection comparison

To compare human ratings with automated detection, we submitted each journal entry to GPTZero[3], a commercial AI-detection service, via an API that returns a probability that the text is AI-generated together with a three-way discrete classification (human, AI, or mixed). The mixed

[3] https://gptzero.me/

classification indicates that the document is judged to contain both human-written and AI-generated text. We collapsed this three-way output to a binary prediction by mapping mixed classifications to the AI-generated class, on the rationale that any detected AI involvement constitutes AI use. As a robustness check, we repeated the agreement analysis with mixed classifications mapped to the human-written class instead. We then computed agreement metrics between GPTZero's predictions and majority-vote human labels, including accuracy, precision, recall and specificity, treating the AI-generated label as the positive class. The purpose of this comparison was not to treat GPTZero as ground truth; rather, the automated detector served as an independent comparison signal that allowed us to examine convergence and disagreement between human judgements and a widely used commercial AI-detection tool.

Although GPTZero is a commercial product whose trained model and decision thresholds are proprietary, it is not a complete black box: its developers publicly describe a detection approach grounded in two text-level statistics, perplexity and burstiness (*GPTZero*, n.d.) . Perplexity quantifies how predictable a text is to a reference language model, computed as the exponential of the mean negative log-likelihood of its tokens; machine-generated text typically attains lower perplexity, because language models tend to sample high-probability continuations, whereas spontaneous human writing is comparatively less predictable (Gehrmann et al., 2019; Mitchell et al., 2023). Burstiness captures how much this predictability varies within a document, for instance the sentence-to-sentence variance in perplexity, with human writing tending to be more bursty as it alternates between simple and complex constructions while generated text remains more uniform. GPTZero has since augmented this foundation with additional proprietary deep-learning components, so the precise features driving any individual classification cannot be fully recovered.

Although perplexity and burstiness are conceptually interpretable, we did not include analogous measures in our own feature set. Both are detector-proximal measures whose values depend on the choice of reference language model, tokenisation procedure and operational definition of burstiness. More importantly, GPTZero contributed to the definition of the high-confidence subset; including features designed to approximate its publicly described detection logic could therefore

cause the classifiers to partially reproduce the same automated signal used in sample selection. We instead prioritised linguistically interpretable features that characterised the entries across multiple levels, including readability, lexical-category use, semantic coherence, logical consistency and stylistic variation. Perplexity and burstiness could usefully be evaluated as supplementary detector-specific features in future work, ideally using independently ground-truthed authorship data.

## High-confidence subset

We constructed a high-confidence subset in which both human raters and automated detection strongly agreed. In the absence of ground-truth data (i.e., no entries explicitly generated with AI, only perceived AI authorship), this subset was intended to provide a more defensible basis for modelling linguistic features associated with clear authorship judgements.

An entry was included in the high-confidence subset if it satisfied two criteria: first, after excluding ratings marked 'invalid', at least 80% (a priori set threshold) of the remaining ratings agreed on the same label (either human or AI-generated); second, GPTZero's AI-generation probability was consistent with that label, above 0.80 for AI-labelled entries and below 0.20 for human-labelled entries. Because the 80% agreement criterion can be trivially satisfied when the number of substantive (non-invalid) ratings per entry is very small, we additionally verified that each entry retained a sufficient number of such ratings before applying this threshold (rater-coverage statistics are reported in *Results*).

This dual-threshold approach was used to identify cases where independent signals converged, rather than relying on either human perception or automated detection alone. The high-confidence subset was not intended to provide an unbiased estimate of real-world AI-detection performance; instead, it was used to identify linguistic features associated with strong and convergent perceived authorship judgements. To assess the robustness of the resulting models, we also repeated the main classification analyses on the full valid dataset, including more ambiguous entries.

## Feature extraction

We computed journal-level features on the anonymised texts to capture length, semantic coherence, stylistic consistency, readability, internal logical structure and psycholinguistic properties. The goal was to construct a transparent feature set that could be inspected, reproduced and adapted by other researchers working with online free-text data.

As part of preprocessing, entries containing only one sentence were excluded from all subsequent analyses. This decision was made because several sentence-pair based features, including semantic coherence, style similarity and NLI-derived consistency measures, require at least two sentences to compute. Excluding one-sentence entries avoided introducing structurally missing values or arbitrary imputations that could confound linguistic feature estimates with entry length. This exclusion was applied consistently across all analyses, including the LIWC-only model, the BLEU analyses and the permutation controls, so that every model was estimated on the same complete-case sample without imputation.

### Length features

We derived three length features from each anonymised entry: word count, sentence count and character count. Word count was computed using whitespace-delimited tokens. Sentence count was obtained using rule-based sentence splitting on terminal punctuation. Character count was computed as the total number of characters in the entry.

### Semantic coherence

We estimated local semantic coherence as the mean cosine similarity between adjacent sentence embeddings. Each sentence was embedded with all-mpnet-base-v2 via SentenceTransformers[4]. Higher scores indicate that adjacent sentences discuss semantically related content.

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

### Style similarity

We estimated content-independent stylistic consistency using the same adjacent-sentence mean cosine similarity procedure, but with StyleDistance[5] embeddings. This model is fine-tuned to capture surface-level stylistic features, such as sentence structure, punctuation patterns and formality, independently of topic. High style similarity indicates a uniform register, whereas lower values indicate shifts in tone or formality.

### Lexical overlap

We quantified within-participant lexical overlap using self-BLEU scores computed with sacreBLEU[6]. At the group level, we computed sentence-level BLEU between all pairs of entries within each label group (perceived AI, perceived human) and averaged them; higher scores indicate more formulaic, mutually similar language within a group. At the participant level, we computed the mean pairwise BLEU across each participant's own entries as an index of within-writer lexical continuity. BLEU scores were computed only on entries with at least two sentences, consistent with the complete-case policy. The participant-level score was excluded from the classifier feature set.

### Internal consistency using natural language inference

We assessed internal logical consistency using natural language inference (NLI). Each entry was split into sentences, filtering out segments of 10 characters or fewer. For all sentence pairs within each entry, we classified the relationship as entailment, contradiction or neutral using BART-large-MNLI[7] via zero-shot classification. We recorded three entry-level rates: entailment, contradiction and neutral, each expressed as a proportion of total sentence pairs. We also defined a consistency score as:

*consistency = 1 − contradiction rate*

[5] https://huggingface.co/StyleDistance/styledistance
[6] https://github.com/mjpost/sacrebleu
[7] https://huggingface.co/facebook/bart-large-mnli

All sentence pairs were evaluated, not only consecutive pairs. This metric therefore captures global logical coherence across the entire entry.

### Readability

We computed seven readability metrics per entry using the textstat[8] Python library: Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index, Automated Readability Index, Coleman-Liau Index and total syllable count. These metrics quantify text complexity from complementary angles, including word length, sentence length, syllable density and polysyllabic word frequency. Higher grade-level scores indicate more complex text, whereas higher Flesch Reading Ease scores indicate simpler text.

### LIWC analysis

We also analysed all journal entries with LIWC-22 to obtain category scores spanning function words, affect, cognition, perception, social processes and topical lexicons. We excluded seven LIWC summary variables from the meta-feature set: four proprietary composite scores (Analytic, Clout, Authentic and Tone), whose internal algorithms are not publicly documented; two word-count proxies (WC and WPS) that duplicate information already captured by our dedicated length features; and one dictionary-coverage variable (Dic), which reports the percentage of tokens matched by the LIWC dictionary rather than a substantive word category. LIWC-22 produces proportional scores: each category score represents the percentage of words in the entry belonging to that category's dictionary. We treated these scores as features for downstream analyses and standardised them before modelling.

## Unified meta-feature set

To support interpretable classification beyond individual feature groups, we merged all entry-level features into a single dataset. This unified representation included semantic coherence and style similarity scores, length counts, readability metrics, NLI consistency measures, participant-level

[8] https://github.com/textstat/textstat

BLEU scores and LIWC-22 category scores, excluding the seven LIWC summary variables described above. The unified meta-feature set comprised 127 features per entry. This representation was used for the main meta-classifier analyses trained on the high-confidence subset, and for the full-dataset robustness analysis.

## Interpretable meta-classifiers

To move beyond black-box detection towards transparent and auditable classification, we trained five models on the unified meta-feature set using the high-confidence subset. The models were selected to balance interpretability and benchmark performance.

All models were evaluated using nested cross-validation to prevent hyperparameter leakage. An outer five-fold stratified loop provided unbiased performance estimates, while an inner three-fold loop tuned hyperparameters within each training fold only. Two redundant NLI metadata features, sentence count and pair count, which correlated $r > .88$ with the dedicated sentence count feature, and the participant-level BLEU score were excluded from the model feature set.

To prevent participant-level data leakage, both the outer and inner cross-validation loops used group-aware stratified folds. Specifically, we used StratifiedGroupKFold, ensuring that all entries from a given participant appeared in the same fold at every stage of the pipeline. This was important because most participants submitted entries with a single majority label. Because entries with fewer than two sentences were excluded from all analyses, the feature matrix contained no missing values and no imputation was performed at any stage; features were standardised where required.

### LASSO logistic regression

We trained an L1-penalised logistic regression model with balanced class weights. The regularisation strength C was selected from a log-spaced grid of 10 values ranging from 0.001 to 100. L1 regularisation simultaneously performs feature selection and provides coefficient-based

interpretability: each non-zero coefficient indicates the direction and magnitude of a feature's association with perceived AI authorship.

### Decision tree

We trained a classification decision tree with balanced class weights, optimising maximum depth and minimum samples per leaf within the same nested, group-aware cross-validation framework as the LASSO (outer 5-fold and inner 3-fold StratifiedGroupKFold on ParticipantID). Decision trees produce explicit, human-readable rules of the form "if feature exceeds threshold, then perceived AI", and therefore provide the most directly auditable classification logic in the workflow. For visualisation we refit the tree on the full high-confidence subset and post-pruned it by collapsing every internal node whose two child leaves predicted the same class. Such splits arise when the Gini criterion, combined with balanced class weights, rewards splits that improve probability ranking even when they do not change the discrete prediction. Collapsing them does not change the model's predictions but yields a more directly readable rule set.

### Linear support vector machine

We trained a support vector machine with a linear kernel and balanced class weights, tuning the regularisation parameter C via grid search. Beyond classification, the linear SVM provides a signed distance, $d$, from each entry to the decision boundary. Entries farther from the boundary are classified with greater confidence, whereas entries within the margin ($|d| \leq 1$) are flagged as ambiguous candidates for human review. The hyperplanes at $d = \pm 1$ define the SVM margin. During training, the hinge loss penalises points that fall within this region, meaning that $|d| \leq 1$ identifies entries that the model itself treats as insufficiently separated from the opposing class. This margin-based procedure provides a practical triage workflow in which automated classification handles clearer cases and uncertain entries are escalated.

### Random forest and XGBoost

We included a random forest and XGBoost as ensemble benchmarks. The random forest used 200 trees and balanced class weights. XGBoost used 200 estimators and positive-class weighting. For both models, depth and learning rate parameters were tuned where applicable. These models were included to contextualise the performance of the more interpretable classifiers.

### Model interpretation

To enable cross-model comparison of feature importance, we computed SHAP values (Lundberg & Lee, 2017) for all five models. For tree-based models, we used TreeExplainer from the SHAP library[9]; for LASSO and the linear Support Vector Machine (SVM), we used LinearExplainer. SHAP values quantify each feature's contribution to individual predictions, and mean absolute SHAP values provide a model-agnostic measure of global feature importance. We report model performance alongside interpretable outputs: LASSO coefficients, decision-tree rules, SVM margin-based ambiguity flags and SHAP feature rankings. This allows the workflow to be inspected and adapted by other researchers collecting remote free-text data.

### Evaluation metrics

For each classifier, we report ROC-AUC, accuracy, F1 score, precision and recall. AI-generated entries were treated as the positive class. Performance was first evaluated on the high-confidence subset, where human and automated judgements converged, and then repeated on the full valid dataset as a robustness check. This second analysis provides a more conservative estimate of performance when ambiguous entries are retained.

Because majority-vote labels were expected to be unevenly distributed across classes, a classifier could in principle attain high accuracy by defaulting to the majority class. Several design choices guard against this. All classifiers were trained with balanced class weights, so errors on the minority class were weighted equally to errors on the majority class, and cross-validation folds

[9] https://github.com/shap/shap

were stratified by label. We report AUC, which is insensitive to class prevalence (a majority-default classifier attains an AUC of 0.5), alongside per-class metrics (precision, recall and specificity), for which a majority-default strategy would yield zero recall on the positive class. The participant-level permutation control (described below) additionally preserves the label imbalance while breaking the text-label correspondence, providing a direct test of whether the imbalance alone confers any classification advantage.

### Permutation-test sanity check

The L1-penalised logistic regression was re-run on both the high-confidence subset and the full dataset under two label-permutation controls. Each control used 50 permutations and the same group-aware five-fold cross-validation as the main analysis. The first control shuffled labels uniformly at random across all entries. The second control reassigned each participant's majority label to a randomly chosen participant, preserving the participant-level label distribution while breaking the text-to-label correspondence. The participant-level shuffle is a stricter test because it leaves any participant-level label imbalance intact and only severs the link between writing and label.

### Full-dataset robustness check

To contextualise the high-confidence subset results, the same nested cross-validation pipeline used for Table 5 was re-run on all valid entries containing at least two sentences, with the same 127 features.

### Leave-one-rate-out sensitivity analysis

This analysis tested whether the classifier's performance and feature profile were structurally robust to variation in individual rater stringency, or whether they were disproportionately driven by one or a few annotators. Individual raters varied substantially in how stringently they applied the AI label, with AI labelling rates ranging from 14.4% to 72.1% of substantive votes (see Results). To assess whether this variation influenced the main classification results, we conducted a

leave-one-rater-out (LORO) sensitivity analysis. We focused this analysis on the LASSO because it was one of the strongest-performing interpretable models and produces a sparse, signed coefficient profile, allowing both feature selection and the direction of feature associations to be compared directly across LORO iterations. The purpose was therefore to test the stability of the aggregated labels and the resulting interpretable feature profile, rather than to repeat the full model-comparison analysis for all five classifiers.

For each of the 21 raters in turn, we removed that rater's labels from the dataset, recomputed majority-vote labels from the remaining 20 raters, rebuilt the high-confidence subset using the same dual-threshold criteria (at least 80% rater agreement and GPTZero convergence), and retrained the LASSO logistic regression with the same group-aware nested cross-validation pipeline described above. We then quantified the stability of the resulting majority-vote labels, model performance and feature importance across all 21 folds. To assess whether the same features were consistently selected regardless of which rater was removed, we computed Kendall's coefficient of concordance (W) over the top 20 features ranked by median absolute LASSO coefficient across folds.

### Detector decomposition

The main analyses used a convergent label defined by agreement between human raters and GPTZero. To examine whether the two detection methods relied on similar or distinct linguistic features, we trained three separate LASSO logistic regressions on a reduced dataset (n = 1,901) with different binary targets. We excluded 565 entries containing fewer than two qualifying sentences, for which coherence, style similarity, and NLI metrics could not be computed; all excluded entries were human-labelled, and imputing systematically missing values risked introducing class-indicator artefacts. Model A predicted the human majority-vote label (AI versus human). Model B predicted GPTZero's binary classification label, with mixed classifications treated as AI. Model C predicted disagreement between the two methods, that is, whether the human majority-vote label and the GPTZero label differed for a given entry. All three models used the same pipeline as the main analysis: L1-penalised logistic regression with balanced class weights,

nested group-aware cross-validation (outer five-fold and inner three-fold StratifiedGroupKFold on ParticipantID), regularisation strength C tuned over the same log-spaced grid, and standardisation; the participant-level BLEU score was excluded from the feature set. We additionally trained a LASSO regression (rather than classification) model predicting GPTZero's continuous AI-generation probability, using the same nested cross-validation framework with GroupKFold and mean squared error as the scoring metric. To compare feature profiles across models, we computed Spearman rank correlations between the full 127-feature coefficient vectors of Models A, B and C. Finally, we computed mean feature values for the top discriminative features within each of the four cells of the human-by-GPTZero agreement matrix (convergent human, convergent AI, human-labelled but GPTZero-flagged, and AI-labelled but GPTZero-cleared) to characterise the feature profiles of entries on which the two methods agreed or diverged.

# Results

## Demographics

### Journaling study participant demographics

Of the 251 participants who contributed text entries retained for analysis, the reported mean age was 28.6 years (SD = 7.0, range = 18 to 56). The sample was 68.5% male and 28.7% female, with 2.8% missing gender data. Most participants resided in the United Kingdom (86.1%), with a smaller proportion residing in the United States (6.8%). 40.2% identified as White/Caucasian, 39.8% as Black/African/Caribbean/Black British, 3.2% as South Asian, 3.2% as Asian British, and smaller proportions as East Asian, Southeast Asian, Mixed, or other groups. Educational attainment was predominantly at degree level: 39.8% held a bachelor's degree, 39.0% a master's degree, 3.2% a PhD, and 11.2% had completed high school. English was the first language for 86.2% of participants. Among the non-native speakers (n = 31), most rated their English proficiency highly, with 83.9% selecting one of the two highest categories ("excellent" or "very good") on a 6-point scale.

### Human labeler demographics

Of the 20 student annotators (90% female), the mean age was 19.2 years (SD = 0.8, range = 18 to 21). Most were in Undergraduate Year 1 (85.0%), mainly from Psychology courses. Further demographic information, including study background, prior journaling experience prior AI-exposure profile is summarised in Table 1.

| Variable | n | % |
|---|---|---|
| **Age (years)** | M = 19.2 (SD = 0.8), range 18-21 | |
| **Gender** | | |
| Female | 18 | 90.0 |
| Male | 2 | 10.0 |
| **Year of study** | | |
| Undergraduate Year 1 | 17 | 85.0 |
| Undergraduate Year 2 | 3 | 15.0 |
| **Programme** | | |
| BSc Psychology | 12 | 60.0 |
| MSci Psychology | 3 | 15.0 |
| BA Linguistics | 4 | 20.0 |
| Not reported | 1 | 5.0 |
| **Native language** | | |
| English | 12 | 60.0 |
| Other † | 8 | 40.0 |
| **English proficiency** | | |
| Native speaker | 15 | 75.0 |
| Near-native / bilingual | 4 | 20.0 |
| Advanced (C1-C2) | 1 | 5.0 |
| **AI familiarity** | | |
| Slightly familiar | 8 | 40.0 |
| Moderately familiar | 4 | 20.0 |
| Very familiar | 7 | 35.0 |
| Extremely familiar | 1 | 5.0 |
| **Journalling experience** | | |
| Never journalled | 4 | 20.0 |
| Tried journalling but not regularly | 12 | 60.0 |
| Journalled occasionally | 2 | 10.0 |
| Journalled regularly | 2 | 10.0 |
| **Labelling confidence** | | |
| Slightly confident | 3 | 15.0 |
| Moderately confident | 10 | 50.0 |
| Very confident | 4 | 20.0 |
| Extremely confident | 3 | 15.0 |
| **Prior AI-detection experience (multi-select)** | | |
| Compared AI vs human text out of curiosity | 5 | 25.0 |
| Checked submissions as TA/tutor/marker | 2 | 10.0 |
| Used AI detection tools | 15 | 75.0 |
| Read articles/research on AI-text detection | 4 | 20.0 |
| Discussed AI writing socially or in class | 7 | 35.0 |
| Participated in other detection tasks | 1 | 5.0 |
| None of the above | 3 | 15.0 |

**Table 1. Demographics, study background and prior AI exposure of the 20 UCL student annotators.** † Other native languages were Bengali, Cantonese, Hindi, Mandarin Chinese, Polish, and Uzbek.

## Classification outcomes

Across the 2,808 entries, each received a median of 7 independent ratings (IQR 7–8; range 5–8; mean 7.26, SD 0.68); this variation reflects the block-based student allocation (6-7 students per block) combined with JK's additional cross-block coverage of 1,723 entries (61.4%). Of the 2,808 labelled entries, 1,727 were judged as human-written by majority vote, representing 61.5%. A further 739 entries were judged as AI-generated, representing 26.3%. The remaining 342 entries, or 12.2%, were classified as invalid. This comprised 270 entries with invalid-majority labels and 72 entries with no clear majority. All 342 invalid entries were excluded, leaving 2,466 entries for analysis. Of these retained entries, 1,727 were perceived as human-written, representing 70.0%, and 739 were perceived as AI-generated, representing 30.0%.

## Inter-rater reliability

Inter-rater reliability was assessed on all entries with at least two independent labels (n=2,808). Agreement across all three categories (human, AI, invalid) was substantial: Krippendorff's alpha (nominal) = 0.638, and pairwise Cohen's kappa (with the invalid entries dropped) averaged 0.642 (weighted mean kappa = 0.635) across 77 rater pairs with overlapping entries. Both statistics range from 0 to 1, where higher values indicate greater agreement.

## Automated detection performance

We compared classifications from a commercial AI detection tool, GPTZero, against human majority-vote labels. GPTZero agreed with human majority-vote labels on 94.3% of entries (Table 2), concurring on 712 of 739 AI-labelled entries and 1,614 of 1,727 human-labelled entries. The confusion matrix (Figure 1) shows that the majority of disagreements were entries labelled human by raters but flagged as AI by GPTZero: 113 entries, or 6.5% of human-labelled entries. The reverse was relatively rare: 27 entries, or 3.7% of AI-labelled entries.

| Metric | Value |
|---|---|
| Accuracy | 94.3% |
| Precision | 0.863 |
| Recall | 0.963 |
| Specificity | 0.935 |
| F1 score | 0.910 |
| *GPTZero AI probability* | |
| AI-labelled entries: mean (SD) | 0.930 (0.199) |
| human-labelled entries: mean (SD) | 0.063 (0.224) |

**Table 2. GPTZero detection performance against human majority-vote labels. AI-generated is the positive class.**

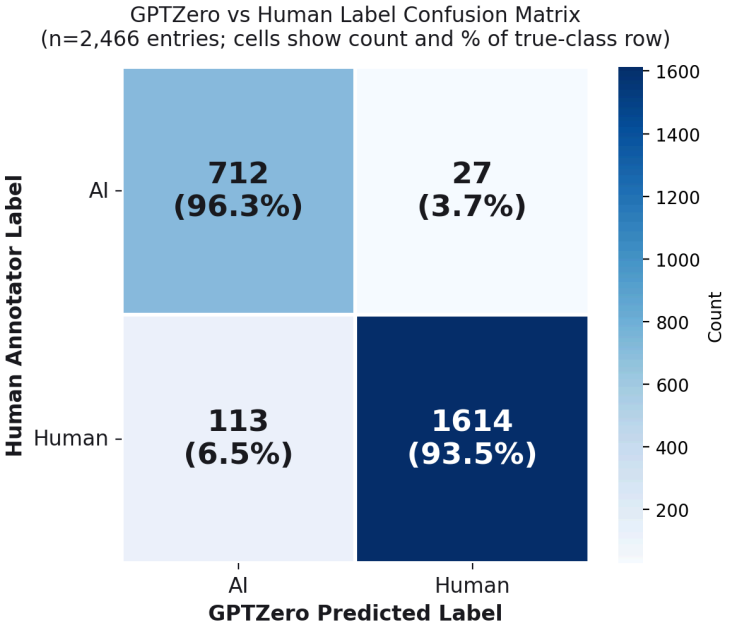


**Figure 1. Confusion matrix showing GPTZero classification outcomes against human majority-vote labels (N=2,466 entries).**

## High-confidence subset

Applying the dual-threshold criteria described in the methods yielded 1,967 entries, representing 79.8% of the 2,466 valid entries. Of these, 485 were labelled as AI-generated and 1,482 were labelled as human-written, giving a 3.1:1 ratio. Within this subset, GPTZero's confidence was very high. AI-labelled entries had a mean AI probability of 0.991, while human-labelled entries had a mean AI probability of 0.002. In addition, 73.4% of entries had at least 90% human rater agreement.

## Demographic comparison by AI usage

At the participant level, 159 participants submitted only human-labelled entries, 62 submitted a mix of human- and AI-labelled entries, and 30 submitted exclusively AI-labelled entries. Participants who submitted any AI-labelled entries differed from the human-only group in several respects. They reported being more predominantly male, at 85.9% compared with 58.4%; more likely to be native English speakers, at 95.2% compared with 80.7%; more likely to identify as White/Caucasian, at 62.0% compared with 27.7%; and submitted more entries on average, 11.7 compared with 8.8. Age and educational attainment were comparable between groups.

| | Human only | Mixed | AI only |
|---|---|---|---|
| | (*n* = 159) | (*n* = 62) | (*n* = 30) |
| **Entries per participant** | | | |
| Total (sum) | 1393 | 738 | 335 |
| *M* | 8.8 | 11.9 | 11.2 |
| Median | 8.0 | 12.0 | 13.0 |
| Mean % AI-labelled | 0.0% | 53.7% | 100.0% |
| **Age (years)** | | | |
| *M* (*SD*) | 28.4 (7.2) | 30.3 (6.7) | 26.5 (6.2) |
| Range | 18-56 | 19-49 | 18-38 |
| Not reported | 5.0% | 1.6% | 0.0% |
| **Gender** | | | |
| Male | 58.5% | 87.1% | 83.3% |
| Female | 37.1% | 12.9% | 16.7% |
| Not reported | 4.4% | 0.0% | 0.0% |
| **Native language** | | | |
| English | 71.1% | 80.6% | 100.0% |
| Other | 17.0% | 6.5% | 0.0% |
| Not reported | 11.9% | 12.9% | 0.0% |
| **Ethnicity** | | | |
| White/Caucasian | 27.7% | 58.1% | 70.0% |
| Black/African/Caribbean | 46.5% | 29.0% | 26.7% |
| Asian (incl. British) | 14.5% | 8.1% | 3.3% |
| Mixed/Other | 6.9% | 4.8% | 0.0% |
| Not reported | 4.4% | 0.0% | 0.0% |
| **Education** | | | |
| High School | 11.9% | 4.8% | 20.0% |
| Bachelor’s degree | 35.8% | 43.5% | 53.3% |
| Master’s degree | 40.9% | 41.9% | 23.3% |
| PhD | 1.3% | 9.7% | 0.0% |
| Other / postgrad. diploma | 5.7% | 0.0% | 3.3% |
| Not reported | 4.4% | 0.0% | 0.0% |
| **Country of residence** | | | |
| United Kingdom | 83.0% | 90.3% | 93.3% |
| United States | 7.5% | 8.1% | 0.0% |
| Other | 5.0% | 1.6% | 6.7% |
| Not reported | 4.4% | 0.0% | 0.0% |

**Table 3. Demographic comparison of participants by AI usage.** “Human only” refers to participants whose entries were all labelled as human-written. “Has AI” refers to participants with at least one AI-generated entry.

## Semantic and structural characteristics

After excluding one-sentence entries, 1,424 of the 1,967 high-confidence entries (72.4%) were retained for the feature-based and meta-classifier analyses (485 AI-labelled, 939 human-labelled); the corresponding full-dataset sample comprised 1,897 of 2,466 valid entries (739 AI, 1,158 human). All excluded entries were human-labelled, as every AI-generated entry already contained at least two sentences.

| Feature | Labelled as Human (n = 939) | Labelled as AI (n = 485) | Difference |
|---|---|---|---|
| Word count | 104.718 | 194.216 | 89.499 |
| Sentence count | 6.142 | 10.056 | +3.914 |
| Character count | 538.015 | 1,199.753 | +661.738 |
| Coherence | 0.336 | 0.478 | +0.142 |
| Style similarity | 0.815 | 0.865 | +0.050 |
| BLEU (entry lvl) | 0.012 | 0.018 | +0.006 |
| BLEU (within ppt) | 0.047 | 0.031 | -0.016 |
| NLI consistency | 0.435 | 0.807 | +0.373 |
| Entailment rate | 0.397 | 0.805 | +0.408 |
| Neutral rate | 0.038 | 0.003 | -0.035 |
| FK grade level | 7.270 | 12.031 | +4.761 |
| Reading ease | 75.112 | 43.978 | -31.134 |

**Table 4. Semantic and structural features of the high-confidence subset: labelled as human-written vs labelled as AI-generated entries.**

BLEU analysis indicated lower lexical diversity among AI-labelled entries. AI-labelled entries had a higher self-BLEU score of 0.018, compared with 0.012 for human-labelled entries, corresponding to approximately 46% more n-gram overlap. At the participant level, however, the direction reversed: human-only participants showed numerically higher within-participant self-repetition than AI-using participants (mean within-participant BLEU 0.047 vs. 0.031).

NLI-based consistency analysis showed that entries perceived as AI-generated had higher internal consistency than entries judged as human-written, with mean scores of 0.807 and 0.435, respectively. Sentence-pair patterns also differed substantially: 80.5% of sentence pairs in AI-labelled entries showed entailment relationships, compared with 39.7% in human-labelled entries. In contrast, human-labelled entries contained more neutral, or tangential, sentence pairs, at 3.8% compared with 0.3%.

Entries judged as AI-generated also had a higher Flesch-Kincaid Grade Level, with a mean of 12.0 compared with 7.3 for human-labelled entries. They also had lower Flesch Reading Ease scores, 44.0 compared with 75.1, indicating that AI-labelled entries were less easy to read on average.

## Psycholinguistic features

The LASSO logistic regression trained on LIWC-22 features (excluding composite scores *Analytic*, *Clout*, *Authentic* and *Tone*, as well as word count, words per sentence, and *Dic* words belonging to LIWC corpus) achieved strong cross-validated performance. The model had an AUC of 0.983 (SD ±0.009), accuracy of 0.937 (SD ±0.019), and F1 score of 0.922 (SD ±0.023). L1 regularisation selected 48 of 111 LIWC features.

As represented by Figure 2 (left), the top 10 features most predictive of perceived AI authorship included BigWords, commas, articles, determiners, prepositions, perception words, memory words, auditory words, curiosity-related words, and second-person pronouns. The top 10 features most associated with perceived human authorship included adverbs, present-focused language, cognition words, quantity terms, numbers, verbs, leisure-related words, allure-related words, first-person singular pronouns, and netspeak. Figure 2 (right) shows the highest-frequency words within six representative LIWC categories. AI-associated categories (Perception, curiosity, memory) were dominated by sensory, exploratory, and reflective vocabulary, while human-associated categories (allure, focus-present, personal pronouns) were dominated by everyday-register and self-referential words.

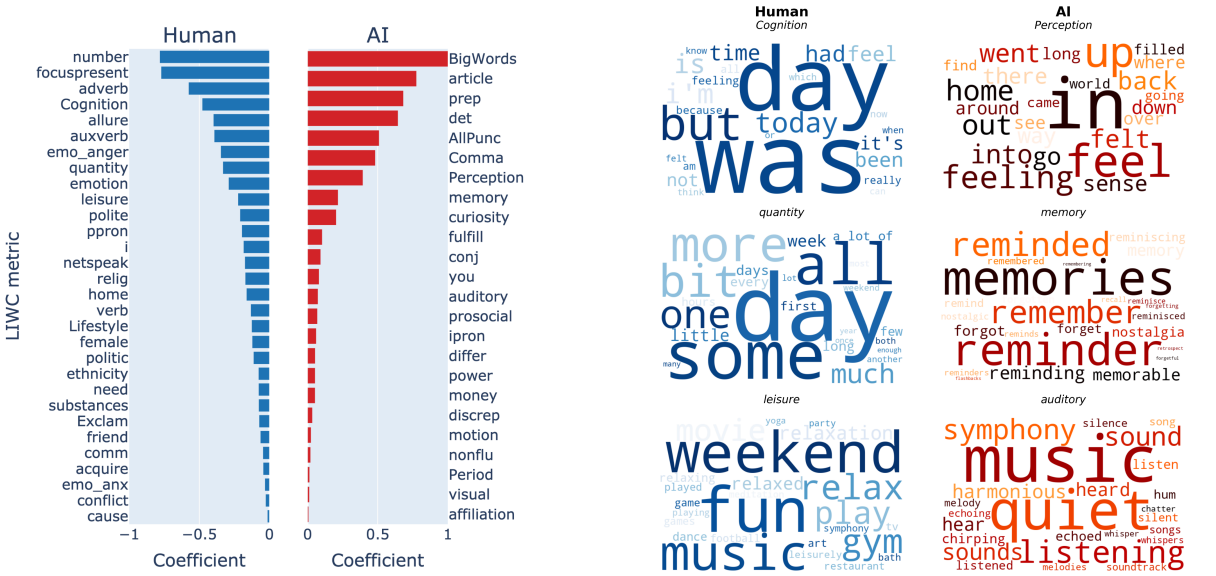


**Figure 2. Linguistic markers distinguishing perceived human and AI authorship.**

*Left panel:* LASSO coefficients for word-based LIWC-22 features predicting perceived AI versus human authorship. Features associated with entries judged as human-written are represented by negative coefficients, while features associated with entries judged as AI-generated are represented by positive coefficients. Only features with non-zero coefficients after L1 regularisation are shown. *Right panel:* Wordclouds for the top three content LIWC categories per direction. The left column shows categories associated with entries judged as human-written: cognition, quantity, and leisure. The right column shows categories associated with entries judged as AI-generated: perception, memory, and auditory perception. Word size reflects frequency within each category.

## Interpretable meta-classifier results

All five meta-classifiers were trained on the high-confidence subset, which included 1,424 entries and 127 features. The features comprised 111 LIWC-22 word categories, 5 length and structure features (word count, sentence count, character count, semantic coherence, and style similarity), 7 readability metrics (Flesch Reading Ease, Flesch-Kincaid Grade, Gunning Fog, SMOG, Automated Readability Index, Coleman-Liau Index, and syllable count), and 4 NLI consistency metrics (consistency score, contradiction rate, entailment rate, and neutral rate). All five models achieved strong cross-validated performance (Table 5), with ROC-AUC at or above 0.98, accuracy at or above 0.95, and F1 at or above 0.93.

| Model | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|
| LASSO | 0.9996 ± 0.0003 | 0.990 ± 0.007 | 0.986 ± 0.010 | 0.980 ± 0.014 | 0.992 ± 0.008 |
| Decision tree | 0.9853 ± 0.0075 | 0.953 ± 0.016 | 0.932 ± 0.024 | 0.915 ± 0.026 | 0.951 ± 0.033 |
| Random forest | 0.9992 ± 0.0004 | 0.982 ± 0.009 | 0.974 ± 0.013 | 0.994 ± 0.008 | 0.955 ± 0.025 |
| XGBoost | 0.9993 ± 0.0005 | 0.986 ± 0.005 | 0.979 ± 0.008 | 0.987 ± 0.008 | 0.971 ± 0.012 |
| Linear SVM | 0.9995 ± 0.0004 | 0.991 ± 0.004 | 0.987 ± 0.005 | 0.986 ± 0.008 | 0.988 ± 0.008 |

**Table 5. Fully group-aware nested cross-validated performance on the high-confidence subset (n=1,424).** Values are reported as mean ± standard deviation.

### LASSO feature selection

The extended LASSO selected 43 of 127 features, spanning multiple feature groups. The strongest positive coefficients, associated with perceived AI authorship, were the Coleman-Liau readability index (+1.53), LIWC BigWords (+1.27), LIWC article (+1.23), LIWC Comma (+1.04), sentence count (+0.81), style similarity (+0.68) and semantic coherence (+0.57). The strongest negative coefficients, associated with perceived human authorship, were LIWC Cognition (−1.12), LIWC adverb (−0.80), LIWC first-person singular i (−0.51) and LIWC number (−0.49). The extended model incorporating all feature types outperformed the LIWC-only LASSO across all metrics, with AUC of 0.9996, accuracy of 0.990, and F1 of 0.986, compared with 0.983, 0.937, and 0.922, respectively, for the LIWC-only model.

### Decision tree

The decision tree achieved 95.3% accuracy at a best depth of 6 and minimum leaf size of 10 under nested cross-validation. Post-pruning left predictions, accuracy, recall and F1 unchanged, and reduced in-sample AUC only marginally, from 0.999 to 0.994. The pruned tree consisted of nine leaves driven by eight features: LIWC verb rate, syllable count, LIWC allure, LIWC BigWords, LIWC article rate, the Coleman-Liau index, LIWC comma rate and sentence count. The primary split was on LIWC verb rate. Entries with 13.4% or fewer verbs and more than 128 syllables were assigned to the perceived-AI class when big-word usage exceeded 16.9%. Entries with higher verb rates were assigned to the perceived-AI class only if article rate exceeded 8.5%, comma rate

exceeded 4.3%, and sentence count exceeded 5.5, or, when article rate was low, if the Coleman-Liau index exceeded 9.8.

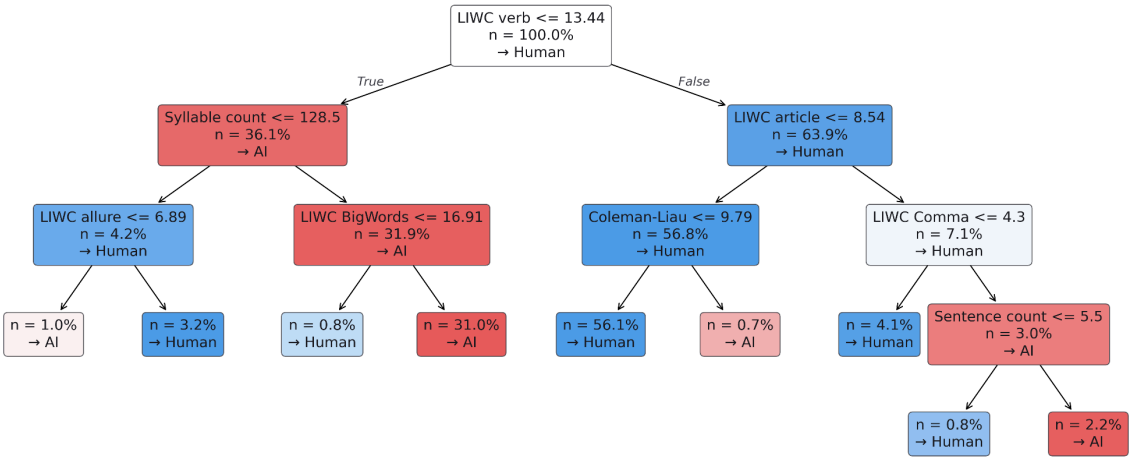


**Figure 3. Post-pruned decision tree for classifying entries as perceived human versus AI.** The tree was originally fit at depth 6. Nodes whose two child leaves predicted the same class were then collapsed. Predictions, accuracy, recall and F1 were unchanged, while in-sample AUC decreased from 0.999 to 0.994. Each node shows the splitting feature and threshold, the percentage of entries reaching it, and the predicted class.

## Random Forest

The Random Forest achieved AUC of 0.9992, accuracy of 0.982, F1 of 0.974, and recall of 0.955, with the highest precision (0.994) among the models. The fitted forest used a cross-validated maximum depth of 5, so all 200 trees reached depth 5, with an average of 19 leaves per tree. The features with the highest Gini importance were LIWC BigWords (0.082), LIWC verb (0.081), LIWC article (0.079), LIWC auxverb (0.078) and Flesch Reading Ease (0.061). Four of the Random Forest's top 10 features overlapped with the LASSO's top 10: LIWC BigWords, LIWC article, the Coleman-Liau index and LIWC Cognition.

### XGBoost

XGBoost achieved AUC of 0.9993, accuracy of 0.986, and F1 of 0.979 at a best maximum depth of 3 and learning rate of 0.1 selected by inner cross-validation. Its feature importance remained concentrated, with LIWC verb alone accounting for 45.0% of the total gain across all 200 boosting rounds. The next strongest contributors were LIWC BigWords (0.160), character count (0.100), LIWC article (0.045) and the Coleman-Liau index (0.029). Five of XGBoost's top 10 features overlapped with the LASSO's top 10: LIWC article, LIWC BigWords, the Coleman-Liau index, LIWC Comma and sentence count. The primary split of the decision tree (Figure 3) was on the same feature (LIWC verb).

### SVM margin-based flagging

The linear SVM achieved AUC of 0.9995, accuracy of 0.991, and F1 of 0.987, with best C selected by inner cross-validation. Its strongest positive coefficients, associated with perceived AI authorship, were LIWC article (+0.11), LIWC Comma (+0.11), LIWC det (+0.08) and LIWC BigWords (+0.08). The strongest negative coefficients, associated with perceived human authorship, were LIWC adverb (−0.09), LIWC Linguistic (−0.07), LIWC verb (−0.07) and LIWC focuspresent (−0.07).

Of the 1,424 entries, 265 (18.6%) fell within the SVM decision margin, defined as a signed distance to the boundary of magnitude 1 or less. The margin contained 175 human-labelled and 90 AI-labelled entries. Predictions agreed with majority-vote labels in all 1,159 entries outside the margin, and in 254 of the 265 entries inside the margin (95.8%).

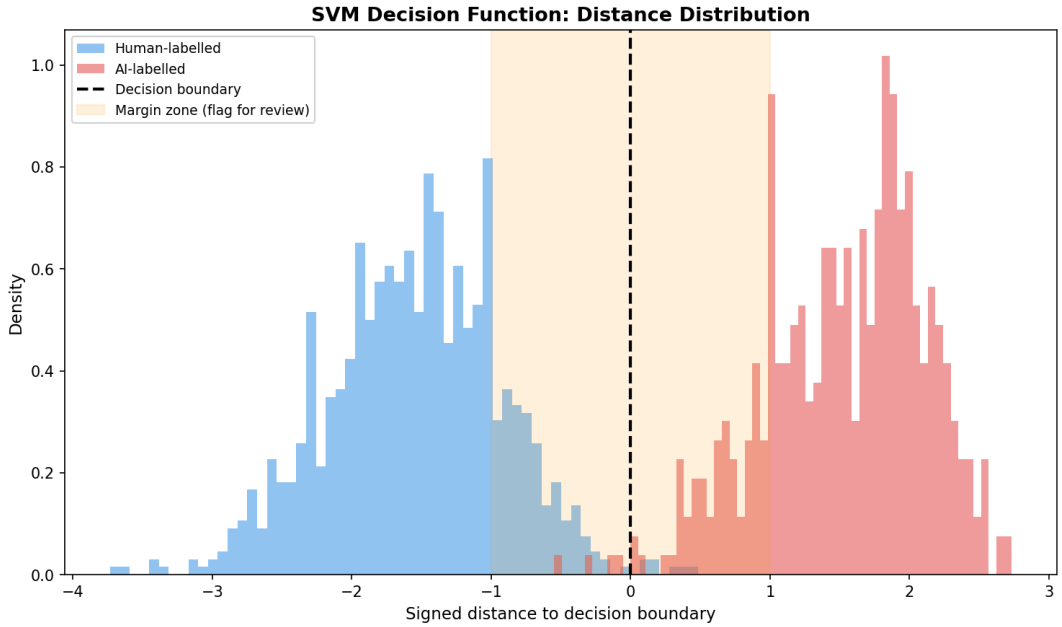


**Figure 4. Distribution of signed distances to the SVM decision boundary.** Entries within the margin zone, defined as an absolute distance of 1 or less, are flagged as ambiguous candidates for human review. Negative distances correspond to entries classified as human, and positive distances correspond to entries classified as AI.

## Feature importance across models

Figure 5A shows the top 15 features by mean rank across the five meta-classifiers. LIWC Articles had the lowest mean rank, 2.2, and was the only feature in the top 3 of every model, ranking first in XGBoost and the linear SVM and third in the LASSO, Decision Tree and Random Forest. The next four features by mean rank were LIWC BigWords (3.0), the Coleman-Liau readability index (4.4), LIWC Commas (5.4) and LIWC Adverbs (8.8).

Figure 5B reports a complementary, per-instance view from an XGBoost model trained on the broader feature set of LIWC word categories with retained word count, five readability metrics, and four NLI consistency metrics. The eight features with the largest mean absolute SHAP values were LIWC Articles, the Coleman-Liau index, LIWC Verbs, LIWC Big Words, word count, LIWC Commas, LIWC Allure, and LIWC Cognition. Higher values of LIWC Articles, the Coleman-Liau index, LIWC Big Words, word count and LIWC Commas pushed predictions toward the AI class, whereas higher values of LIWC Verbs, LIWC Allure and LIWC Cognition pushed predictions toward the human class.

Figure 5C shows the distributions of the same 15 features in the high-confidence subset (n = 1,424), grouped by perceived label. Entries judged as AI-generated had higher values than entries judged as human-written for eight features: LIWC Articles, LIWC Big Words, the Coleman-Liau readability index, LIWC Commas, sentence count, semantic coherence, syllable count and LIWC Determiners. Entries judged as human-written had higher values for the remaining seven: LIWC Adverbs, LIWC Cognition, LIWC Verbs, LIWC Linguistic, LIWC Allure, LIWC Function words and LIWC Pronouns. Separation was visually clearest for LIWC Articles, LIWC Big Words, the Coleman-Liau index and LIWC Commas, while the count-based features (sentence count, character count and syllable count) showed long upper tails and greater overlap despite well-separated medians.

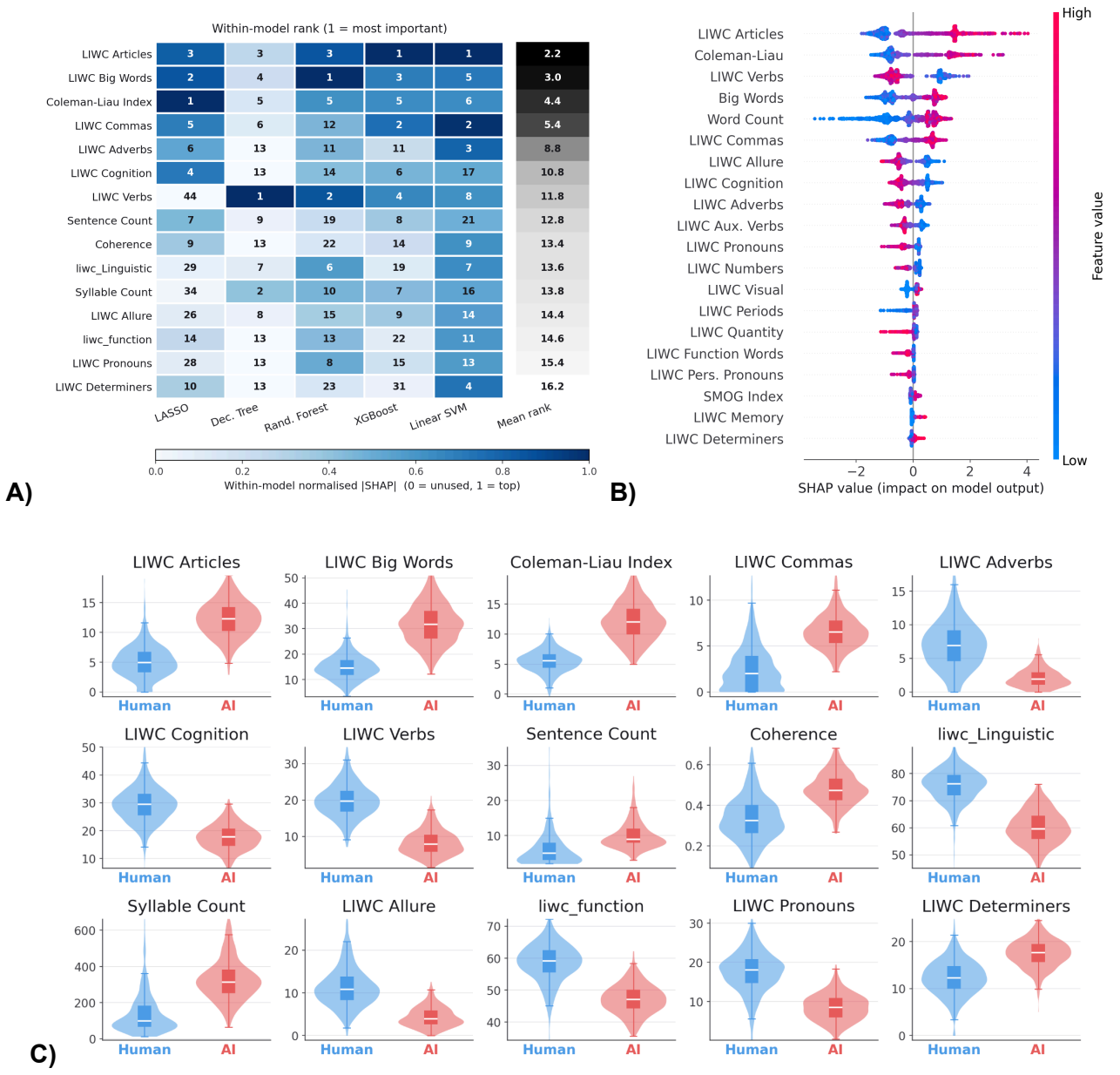


**Figure 5. Feature importance and class-conditional distributions for the top discriminative features in the high-confidence subset (n = 1,424).** *A) Cross-model comparison of feature importance for the top 15 features, ordered by mean rank across the five meta-classifiers*. Cell shading shows within-model normalised absolute SHAP values, where 0 indicates unused and 1 indicates the top feature in that model. Cell labels give the integer rank in that model, with 1 indicating the most important feature. The right-hand strip reports the mean rank across the five models. *B) SHAP beeswarm summary for an XGBoost classifier trained on LIWC word categories with retained word count, readability, and NLI consistency features.* Each dot is one entry. Horizontal position is the feature's SHAP contribution to the model's logit, with rightward movement pushing predictions toward AI and leftward movement pushing predictions toward human. Colour encodes the raw feature value, with blue indicating low values and red indicating high values. Features are sorted from top to bottom by mean absolute SHAP value. *C) Violin plots of key meta-features, grouped by perceived label.* Blue indicates entries judged as human-written, and red indicates entries judged as AI-generated.

## Permutation-test sanity check

Real-label LASSO AUC was 0.9996 in the high-confidence subset and 0.9819 in the full dataset (see Table 6). Under the random-shuffle control, mean AUC collapsed to chance: 0.500 ± 0.025 in the high-confidence subset, with a 95% range from 0.461 to 0.542, and 0.501 ± 0.018 in the full dataset, with a 95% range from 0.463 to 0.531. The participant-level label shuffle yielded the same null result: 0.516 ± 0.035 in the high-confidence subset, with a 95% range from 0.465 to 0.588, and 0.497 ± 0.040 in the full dataset, with a 95% range from 0.416 to 0.557. Across the 200 permutation runs spanning the two controls and the two subsets, the highest observed AUC was 0.603, from a single participant-shuffle run in the high-confidence subset; all remaining runs stayed below 0.58.

| Subset | Real AUC | Random shuffle (mean ± SD, 95% range) | Participant shuffle (mean ± SD, 95% range) |
|---|---|---|---|
| High-confidence | 0.9996 | 0.500 ± 0.025, [0.461, 0.542] | 0.516 ± 0.035, [0.465, 0.588] |
| Full dataset | 0.9819 | 0.501 ± 0.018, [0.463, 0.531] | 0.497 ± 0.040, [0.416, 0.557] |

**Table 6. Permutation-test results across two label-shuffling schemes and two subsets, 50 permutations per cell.**

## Full-dataset robustness check

Performance dropped on the full dataset relative to the high-confidence subset for every model (see Table 7). XGBoost AUC fell, albeit modestly, from 0.9993 in the high-confidence subset to 0.9897 in the full dataset, accuracy from 0.986 to 0.947, and F1 from 0.979 to 0.934. The same direction of change held for the other models. LASSO AUC fell from 0.9996 to 0.9868. Random Forest AUC fell from 0.9992 to 0.9892. Linear SVM AUC fell from 0.9995 to 0.9878. The Decision Tree showed the largest absolute drop in AUC, from 0.9853 to 0.9619, while the Linear SVM and LASSO showed the largest drops in F1, from 0.987 to 0.921 and from 0.986 to 0.922, respectively. The LASSO and Linear SVM were the strongest models in the high-confidence subset, whereas XGBoost was the top performer on the full dataset; the Decision Tree was the weakest model in both subsets.

| Model | Subset | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| LASSO | HC (n=1,424) | 0.9996 ± 0.0003 | 0.990 ± 0.007 | 0.986 ± 0.010 | 0.980 ± 0.014 | 0.992 ± 0.008 |
| LASSO | Full (n=1,897) | 0.9868 ± 0.0091 | 0.937 ± 0.017 | 0.922 ± 0.021 | 0.893 ± 0.038 | 0.954 ± 0.031 |
| Decision Tree | HC | 0.9853 ± 0.0075 | 0.953 ± 0.016 | 0.932 ± 0.024 | 0.915 ± 0.026 | 0.951 ± 0.033 |
| Decision Tree | Full | 0.9619 ± 0.0194 | 0.922 ± 0.017 | 0.900 ± 0.023 | 0.893 ± 0.033 | 0.910 ± 0.052 |
| Random Forest | HC | 0.9992 ± 0.0004 | 0.982 ± 0.009 | 0.974 ± 0.013 | 0.994 ± 0.008 | 0.955 ± 0.025 |
| Random Forest | Full | 0.9892 ± 0.0052 | 0.942 ± 0.016 | 0.924 ± 0.022 | 0.930 ± 0.021 | 0.921 ± 0.044 |
| XGBoost | HC | 0.9993 ± 0.0005 | 0.986 ± 0.005 | 0.979 ± 0.008 | 0.987 ± 0.008 | 0.971 ± 0.012 |
| XGBoost | Full | 0.9897 ± 0.0071 | 0.947 ± 0.011 | 0.934 ± 0.013 | 0.918 ± 0.029 | 0.952 ± 0.020 |
| Linear SVM | HC | 0.9995 ± 0.0004 | 0.991 ± 0.004 | 0.987 ± 0.005 | 0.986 ± 0.008 | 0.988 ± 0.008 |
| Linear SVM | Full | 0.9878 ± 0.0086 | 0.936 ± 0.017 | 0.921 ± 0.019 | 0.887 ± 0.043 | 0.960 ± 0.029 |

**Table 7. Group-aware nested cross-validated performance on the high-confidence subset (n = 1,424) and the full dataset (n = 1,897).** Values are reported as mean ± standard deviation across the five outer folds.

## Leave-one-rater-out sensitivity analysis

Individual raters varied substantially in how often they labelled entries as AI-generated, with AI labelling rates ranging from 14.4% to 72.1% of substantive votes (Figure 6, Panel A). Despite this variation, removing any single rater produced no changes to the majority-vote labels across the full dataset. The high-confidence subset size varied only marginally across the 21 folds, ranging from 1,425 to 1,482 entries compared with 1,424 in the analysis after short-entry exclusion. LASSO classification performance was effectively invariant: mean AUC across all 21 leave-one-rater-out folds was 0.9993 (SD = 0.0004) and mean accuracy was 0.985 (SD = 0.002). Feature importance rankings were also stable, with Kendall's W = 0.791 across the top 20 features, indicating strong concordance, and no sign flips among the top-ranked features in any fold (Figure 6, Panels B and C). Dropping the researcher (JK), who rated 59.6% of the reduced dataset's entries produced the lowest fold-level AUC (0.9992), which was still within 0.0002 of the mean. These results indicate that neither the majority-vote labels nor the classifier's feature profile depended on any individual rater's judgements.

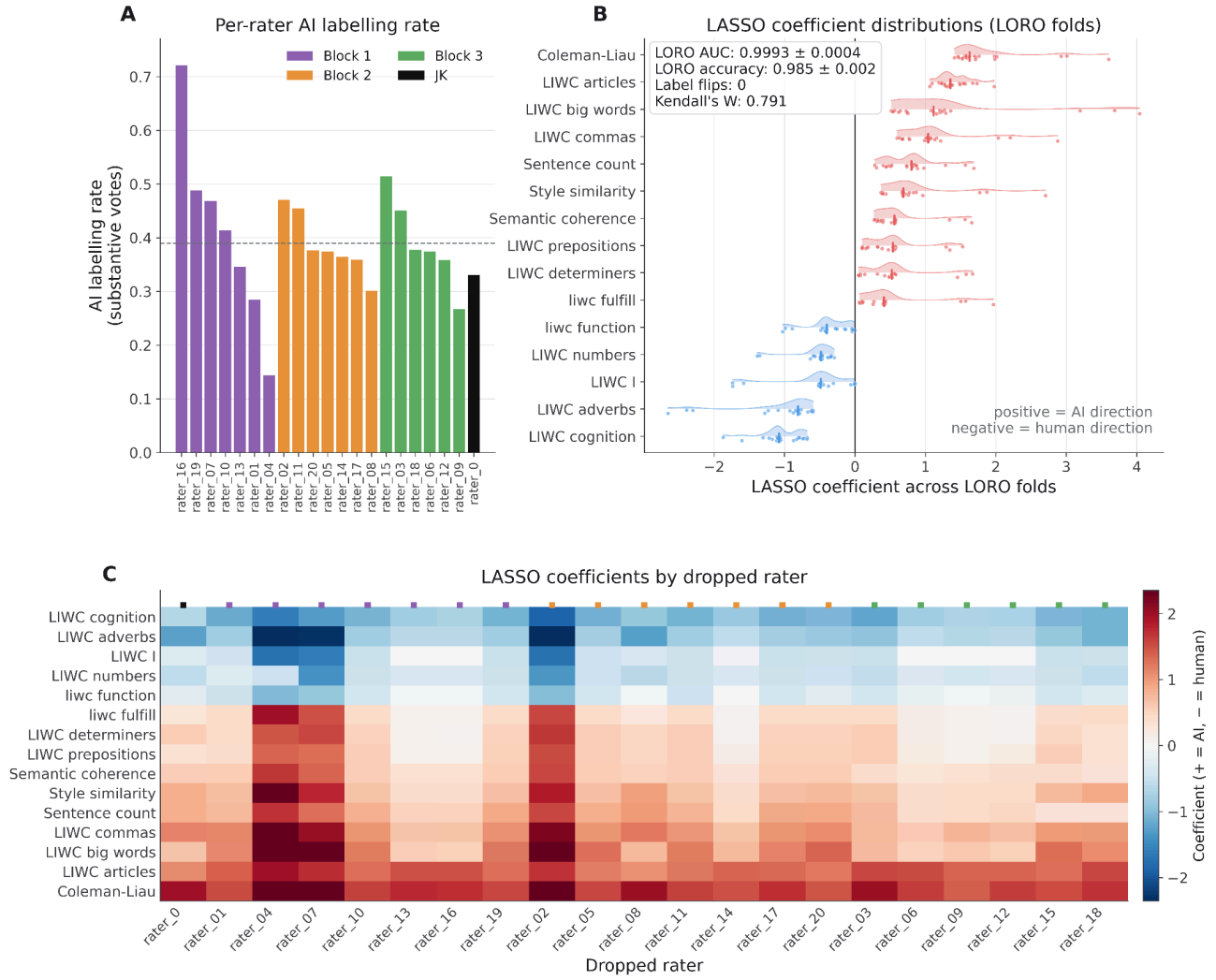


**Figure 6. Leave-one-rater-out sensitivity analysis.** (A) Per-rater AI labelling rates across the 21 annotators, grouped by rating block. The dashed line indicates the overall majority-vote AI rate (38.9%) after short-entry exclusion. Raters varied from 14.4% to 72.1% in the proportion of substantive votes labelled as AI-generated. (B) Distributions of LASSO coefficients across the 21 leave-one-rater-out folds for the top 15 features ranked by median absolute coefficient. Each point represents one fold. Shaded areas show kernel density estimates. Vertical lines indicate the median coefficient. Features above the zero line (positive coefficients) are associated with perceived AI authorship; features below (negative coefficients) are associated with perceived human authorship. The inset reports summary statistics: mean cross-validated AUC and accuracy across folds, total label flips, and Kendall's W for feature-ranking concordance. (C) Heatmap of LASSO coefficients for the same 15 features across all 21 leave-one-rater-out folds, with columns ordered by rater block (block-colour indicators along the top). Positive values (red) indicate features associated with perceived AI authorship; negative values (blue) indicate features associated with perceived human authorship

## Detector decomposition

To disentangle the contributions of human and automated detection to the convergent label, we trained separate classifiers on the reduced dataset (n = 1,897; short entries excluded). Models A and B, predicting human majority-vote labels and GPTZero binary labels respectively, both achieved strong cross-validated performance on the full dataset. Model A achieved AUC of 0.987 and accuracy of 0.949. Model B achieved AUC of 0.987 and accuracy of 0.939 (Figure 7, Panel A). The LASSO regression predicting GPTZero's continuous AI-generation probability achieved a cross-validated R-squared of 0.751 (SD = 0.031) and mean squared error of 0.055 (SD = 0.007). Model C, predicting disagreement between the two methods, achieved above-chance but substantially lower performance, with AUC of 0.778 and accuracy of 0.772, reflecting the extreme class imbalance (137 disagreement entries versus 1760 agreement entries, or 7.2%).

The Spearman rank correlation between the full coefficient vectors of Models A and B was $\rho = 0.44$ ($p < 10^{-6}$), indicating that human raters and GPTZero responded to partially overlapping but non-redundant feature sets (Figure 7, Panel B).

Several features were highly ranked in both models, including LIWC Articles, LIWC Commas, and LIWC Adverbs. However, each model also weighted distinctive features: Model A placed greater emphasis on readability metrics such as the Coleman-Liau index and syllable count, whereas Model B placed greater emphasis on punctuation, function-word, and semantic features such as LIWC Apostrophes, LIWC Auxiliary Verbs, style similarity, and semantic coherence. The correlation between Models A and C was insignificant (rho = 0.04, p = 0.69), indicating that the features predicting authorship labels were largely orthogonal to the features predicting where the two methods disagreed (Figure 7, Panel C).

The agreement-cell feature profiles (Figure 7, Panel D) showed that the two convergent cells (both methods agreeing on human or both agreeing on AI) were cleanly separated on all features, as expected. The two divergent cells occupied intermediate positions. The human-labelled, GPTZero-flagged cell showed the highest style similarity of any cell (0.88, exceeding even

convergent-AI entries at 0.86), whereas the rater-AI, GPTZero-cleared cell had style similarity (0.82) closer to convergent-human entries (0.82). Both divergent cells showed moderate readability and word complexity scores between the two convergent cells. The largest difference between the two divergent cells was in character count: entries labelled as human by raters but flagged as AI by GPTZero were approximately 210 characters shorter than entries labelled as AI by raters but cleared by GPTZero. LIWC Verbs and LIWC Cognition were higher in the human-labelled, GPTZero-flagged cell, consistent with the pattern that informal language markers satisfied human raters but not the automated detector.

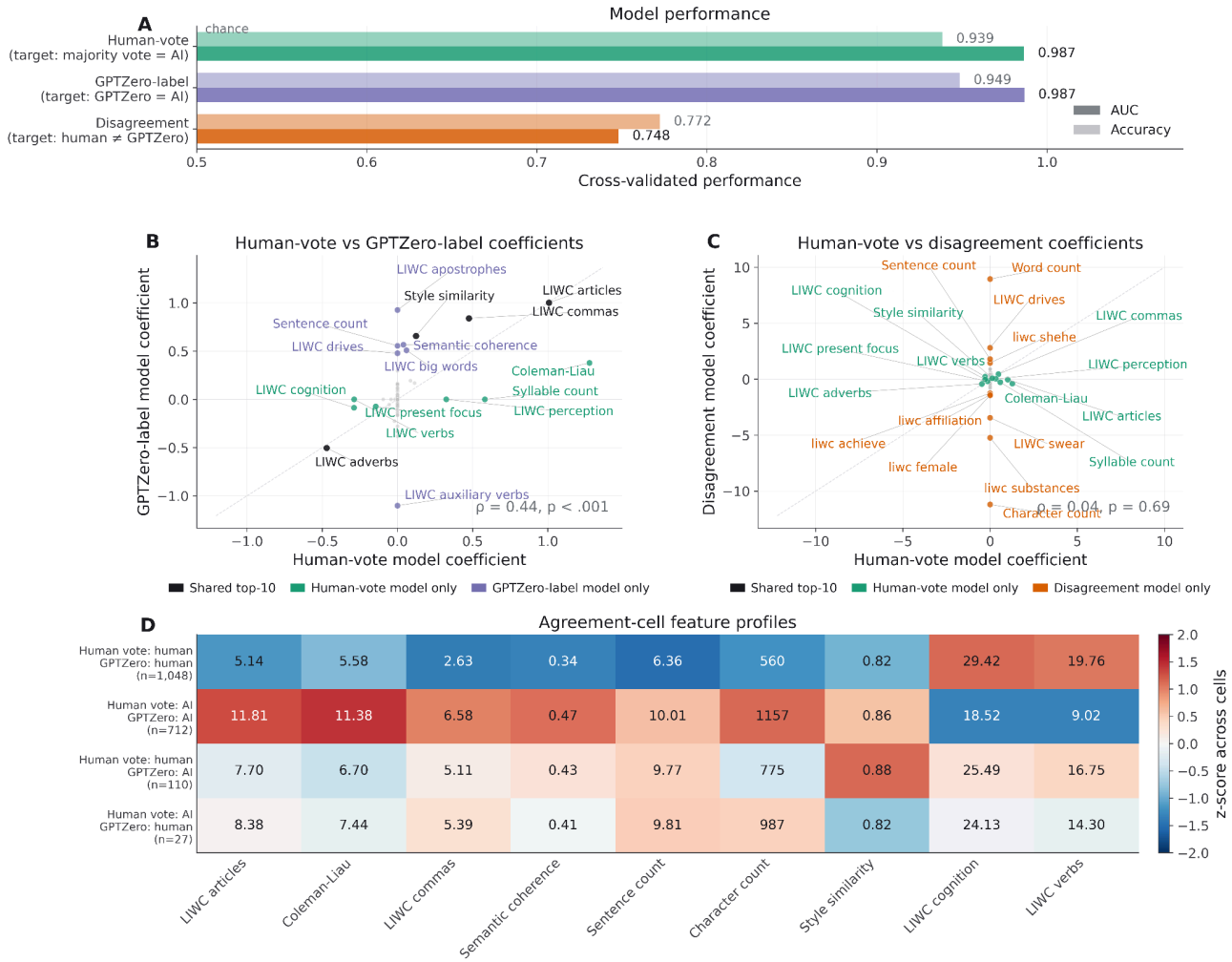


**Figure 7. Detector decomposition: comparing human and automated detection signals.** (A) Cross-validated performance (AUC and accuracy) for three LASSO models trained on the reduced dataset after short-entry exclusion (n = 1,897). Model A predicts the human majority-vote label, Model B predicts the GPTZero binary label, and Model C predicts disagreement between the two methods. The dotted line indicates chance performance. (B) Scatter plot comparing LASSO coefficients between the human-vote model (Model A, x-axis) and the GPTZero-label model (Model

B, y-axis). Each point represents one of 127 features. Labelled points indicate features ranked in the top 10 by absolute coefficient for at least one model, colour-coded by whether they are shared (dark) or unique to one model. The Spearman rank correlation is reported in the lower right. (C) Scatter plot comparing LASSO coefficients between the human-vote model (Model A, x-axis) and the disagreement model (Model C, y-axis), using the same conventions as Panel B. (D) Mean feature values for the top discriminative features across the four cells of the human-by-GPTZero agreement matrix. Values are z-scored across cells for the colour scale, with raw means annotated in each cell. Sample sizes are shown on the y-axis.

# Discussion

The current study developed an interpretable text-analysis workflow for identifying linguistic cues associated with perceived AI authorship in introspective writing, and evaluated this workflow as a transparent alternative to opaque commercial detectors in psychological research. Overall, the findings suggest that perceived AI and human authorship were distinguishable not by a single marker, but by a compact and interpretable constellation of linguistic features. Human raters and GPTZero showed strong convergence, agreeing on 94.3 percent of entries (Figure 1), indicating that both sources of judgement were responding to a broadly shared signal in the dataset. Across feature-selection methods, the same small set of variables consistently emerged as most informative, particularly LIWC Articles, LIWC BigWords, the Coleman-Liau readability index and LIWC Commas (Figure 5A). Finally, the linear SVM provided a practical basis for triage: most entries fell far from the decision boundary and were classified in agreement with the majority-vote labels, whereas a smaller group fell within the margin and would be appropriate for human review (Figure 4). Together, these findings support the value of interpretable linguistic modelling as a way to identify high-confidence and low-confidence cases without relying solely on black-box AI-detection scores.

## Linguistic features of perceived authorship

Two contrasting registers emerged from these results. First, entries judged as AI-generated were marked by greater surface complexity. They were longer, more difficult to read according to readability-grade metrics, and characterised by a denser, more elevated lexical register, including

longer words, heavier function-word scaffolding, and more elaborated punctuation. This register was carried most strongly by high rates of articles and determiners, consistent with more noun-phrase-heavy, nominal prose characteristic of formal edited writing, and by correspondingly low verb rates. Verb rate was the single most influential feature in the tree-based models, consistent with a nominal rather than verbal or narrative style. Human-labelled entries showed the opposite, verb-dense pattern typical of spontaneous first-person narration. A similar pattern has been reported in academic abstracts, where AI-generated texts showed lower readability, higher lexical density, and lower lexical diversity than human-written abstracts, suggesting a register that is formally dense but less lexically varied and more repetitive (Zou et al., 2026). In the context of introspective writing, this appears as an additional level of polish and abstraction atypical of spontaneous self-reflection. The within-entry style-similarity measure pointed in the same direction at a finer-grained level: AI-labelled entries maintained a more uniform register from sentence to sentence, whereas human-labelled entries showed greater stylistic variation as attention moved across topics.

Beyond surface complexity, AI-labelled entries were also more coherent, which was evident through two complementary measures: semantic coherence, indicating how closely adjacent sentences were related in meaning, and NLI-based consistency, reflecting logical continuity across sentence pairs. Both were substantially higher in AI-labelled entries, with around four fifths of sentence pairs showing entailment relationships, compared with around two fifths in human-labelled entries, while human-labelled entries contained substantially more neutral or tangential transitions. This pattern is especially relevant in the context of journalling. Coherence in finished prose emerges from planning and revision (Amjad & Tanveer, 2026), processes that are less central to spontaneous reflection. Mind wandering moves freely between topics without linear constraint (Christoff et al., 2016), and evidence from expressive writing research suggests that authentic self-disclosure involves increasing cognitive elaboration over time rather than polished organisation from the outset (Pennebaker et al., 1997), a pattern that may extend naturally to open reflective journalling. High coherence, in this genre, is therefore not a marker of quality but a potential signal of AI mediation.

Entries judged as human showed the opposite profile. Linguistic distancing, defined as the use of language that increases psychological distance from experience by reducing first-person singular pronouns and present-tense verbs, provides a useful framework for interpreting this contrast (Nook et al., 2017). Human-labelled entries showed consistently lower distancing, with higher rates of present-focused language and first-person self-reference, consistent with LIWC-based operationalisations of verbal immediacy (Bazarova et al., 2013; Pennebaker & King, 1999), as well as more auxiliary verbs, adverbs, affective language including anger and frustration, numerals and quantity terms consistent with the concrete, situated detail of everyday self-report, desire- and motivation-related vocabulary captured by LIWC Allure, and references to concrete everyday topics. Together, these features suggest that entries perceived as human remained closer to the writer's immediate perspective, as would be expected in journal entries collected daily, where participants are likely to offer relatively raw and situated reflections on everyday experience. AI-labelled entries, by contrast, more often presented experience as already organised, coherent and reflected upon. These contrasting registers are visible both in the LIWC coefficient profiles and representative category vocabularies (Figure 2) and in the broader class-conditional feature distributions (Figure 5C)

The BLEU analyses extended this pattern from individual features to broader stylistic continuity. AI-labelled entries showed greater BLEU-based n-gram overlap than human-labelled entries, indicating cross-writer convergence in phrasing, that is, a tendency for AI-labelled texts to reuse similar words and phrase patterns across authors. This parallels the idea of LLM-induced homogenisation, where LLM-assisted writing makes texts by different authors more similar while reducing diversity in both wording and content (Padmakumar & He, 2023). At the same time, writers who used AI showed lower within-writer BLEU across their own entries than writers who did not, indicating weaker lexical continuity over time. Because authorship attribution relies on recurrent linguistic patterns that distinguish one writer's style from another (Stamatatos, 2017), this suggests that AI assistance may weaken the consistency of an individual's voice over time. Together, these findings suggest a dual erosion of individuality: greater shared phrasing across writers, but weaker continuity within the same writer across the individual entries.

## Methodological validation

The classifier results provide methodological validation for this interpretation. Rather than depending on a single model or opaque detector score, the workflow produced converging evidence across models with different assumptions. LASSO, Decision Tree, Random Forest, XGBoost and linear SVM all identified a similar compact feature set, with LIWC Articles, LIWC BigWords, the Coleman-Liau index and LIWC Commas consistently among the most informative variables (Figure 5A). This convergence suggests that the distinction between perceived human and AI authorship was not driven by idiosyncrasies of a single modelling approach, but by a stable linguistic signal distributed across surface complexity, coherence and function-word use.

The detector decomposition analysis provided further evidence that this convergence rested on partially independent signals. Human raters and GPTZero responded to overlapping but non-redundant feature sets (Spearman rho = 0.44 between their respective LASSO coefficient vectors), with human raters placing greater weight on readability metrics and GPTZero placing greater weight on punctuation and function-word features (Figure 7B). This moderate correlation suggests that each method contributed distinct information to the convergent label, strengthening the rationale for requiring agreement between the two rather than relying on either alone. This is notable because none of the 127 features in our own classifier encode token-level model probability, the likelihood-based signal underlying GPTZero's perplexity and burstiness statistics (GPTZero, n.d.); the convergence between human raters and GPTZero therefore holds across at least partially independent feature spaces, rather than reflecting two instruments reading the same underlying cue.

The disagreement model further showed that the 137 entries on which the two methods diverged were not random noise but a characterisable subpopulation with intermediate feature profiles, predicted by a largely orthogonal set of features (rho = 0.04, p = .69 between authorship and disagreement models; Figure 7C-D). This suggests that the entries most likely to be misclassified by one method alone occupy a distinctive region of feature space that the dual-convergence criterion is specifically designed to handle.

The permutation controls strengthened this conclusion: both random label shuffling and participant-level label shuffling returned chance-level AUCs, making it unlikely that the high classification performance reflected leakage or writer-specific fingerprints. The leave-one-rater-out analysis provided a further robustness check on the annotation side: despite individual raters varying from 14.4% to 72.1% in their AI labelling rates, removing any single rater changed zero majority-vote labels and left classifier performance and feature rankings effectively unchanged (Kendall's W = 0.791; Figure 6). Finally, although performance declined on the full dataset relative to the high-confidence subset, as expected, it remained strong, suggesting that the signal was not confined to the clearest cases.

## A transparent workflow for auditing free-text data

For researchers collecting free-text responses in psychological studies, this combination of features and models offers a more cautious alternative to opaque commercial detectors. This aligns with recent feature-based approaches to AI-generated text detection, which suggest that linguistically and statistically grounded classifiers can perform competitively while providing more transparent decision rules (Sen & Hasan, 2025). Such transparency is especially important because commercial detectors have already been shown to be unreliable in the academic-integrity contexts for which they are most often used, with performance varying across tools and declining further after paraphrasing or machine translation (Weber-Wulff et al., 2023). Recent work further suggests that detector scores vary not only across detection tools, but also across source language models and even across files generated under the same conditions, reinforcing the need to treat these scores as provisional rather than definitive (Malik & Amjad, 2025). The behaviour of these tools in introspective writing is therefore even less certain: to our knowledge, they have not been systematically evaluated on diary-style responses, a genre that differs from academic prose in its immediacy, self-reference, affective language and topical discontinuity.

The present workflow addresses this problem by making both classification and uncertainty visible. Feature contributions can be inspected through coefficients, decision-tree rules and SHAP values, allowing researchers to identify which linguistic properties pushed an entry toward a perceived-AI

classification (Figure 3). The SVM margin adds a further layer of caution by separating high-confidence cases from entries close to the decision boundary, which can be flagged for human review rather than automatically excluded (Figure 4). This is important because the 113 entries on which GPTZero disagreed with the human majority vote show that even high overall agreement leaves a non-trivial set of residual cases requiring interpretation. The detector decomposition results reinforce this point: because the two methods weight different aspects of the text, entries that fall in the margin of one detector are not necessarily marginal for the other. The divergent cells in the agreement matrix showed feature profiles intermediate between the two convergent cells (Figure 7D), suggesting that ambiguous cases are not uniformly difficult but rather occupy a specific region of the feature space where formality signals are present but informal-language markers have not been fully suppressed. By grounding classification in standard psycholinguistic and structural features, the workflow therefore allows inclusion decisions to be reported transparently in downstream studies, rather than being delegated to a black-box probability from a commercial service.

This same transparency, however, carries a dual-use cost. Because the workflow reports explicit thresholds (e.g, verb rate, syllable count and article rate) rather than an opaque probability score, it is also, by construction, a roadmap for evasion: the same information that allows researchers to audit and reproduce the method could in principle be used to fine-tune or prompt a model to avoid these specific markers. This tension is not unique to the present workflow; any interpretable detection method faces a trade-off between transparency and robustness to circumvention, whereas opaque commercial detectors are harder to game precisely because their decision rules are hidden. We do not see this as a reason to withhold interpretability, but it does mean that any published feature set and threshold should be treated as provisional rather than durable, and that auditing workflows will likely require periodic recalibration as generative models, and the strategies used to disguise their output, continue to change.

## Implications for digital mental health and psychological research

Beyond the methodological question of how to flag AI-generated entries, these findings illuminate what readers implicitly treat as authentically human in introspective writing. The preference for present-focused language, self-reference and stylistic variability suggests that perceived authenticity rests on recognisable markers of cognitive and emotional presence. If the features that make writing feel genuinely human include apparent flaws, such as lower coherence, hedged affect and topical discontinuity, then optimising free-text responses for conventional writing-quality metrics may paradoxically reduce their perceived genuineness in personal contexts.

This has direct implications for how diary-style responses are collected and analysed in digital mental health and remote-monitoring studies, where free-text data are increasingly used to infer emotional states and change over time (Kelly et al., 2021; Nepal et al., 2024). When diary entries are treated as behavioural or psychological data, their value depends not only on excluding AI-mediated responses but also on preserving the linguistic irregularities through which lived experience is expressed. What might otherwise appear to be noise is therefore part of the signal that makes diary-style data psychologically meaningful.

## Limitations

Several features of the dataset limit how broadly these findings can be interpreted. First, demographic differences between participants with and without AI-labelled entries introduce potential confounds. Participants with AI-labelled entries were more predominantly male, more likely to be native English speakers, and more likely to identify as White or Caucasian. Some linguistic patterns captured by the meta-features, including readability level and function-word usage, may therefore partly reflect writer demographics rather than perceived AI authorship alone. This comparison should also be interpreted cautiously because demographic information was self-reported, and participants willing to submit AI-mediated text may also have been more willing to misreport other study information. Future work should therefore model writer demographics explicitly, ideally using demographic information collected independently of the writing task.

A further limitation concerns the annotators themselves. Because the labels reflect perceived rather than verified authorship, the cues identified here may partly reflect the expectations of a relatively narrow rater group. More broadly, because no entries in this dataset have confirmed ground-truth authorship, the "high-confidence" subset reflects convergence between human judgement and GPTZero rather than validated accuracy. Although the detector decomposition analysis suggests these two signals are only moderately correlated (Spearman rho = 0.44) and rely on partially distinct features, it remains possible that both are responding to a shared surface cue, such as formality or length, that is associated with but not equivalent to actual AI mediation. Convergence between two imperfect and partially correlated signals therefore increases confidence relative to either signal alone, but cannot substitute for verified ground truth. The composition of the rater group is a further source of uncertainty in this convergence: the annotators were mostly young university students, predominantly female, and many had prior exposure to AI-detection tools, while few reported regular journalling experience. Perceived authenticity in this dataset may therefore reflect not only properties of the entries, but also the assumptions this particular group brought to diary-style writing. Future work should test whether the same perceptual cues emerge with more diverse annotators, including experienced diary or mental-health researchers and raters from different cultural and linguistic backgrounds.

A related practical constraint is that part of the feature set relies on LIWC-22, a proprietary psycholinguistic dictionary. Although some open LIWC-like resources are available (e.g., Quatenda Dictionaries[10]), their comparability with LIWC-22 is difficult to assess because the underlying dictionary is not publicly accessible. This may limit exact reproducibility and reduce uptake among researchers without institutional access. Future work should therefore evaluate licence-free approaches, including open lexical resources, syntactic features, readability and repetition metrics, and transformer-derived concept axes. In particular, embedding-based methods could be made interpretable by projecting entries onto labelled semantic or stylistic directions, rather than treating dense embeddings as opaque predictors. This would allow future auditing workflows to preserve the interpretability of the present approach while reducing dependence on licensed dictionaries.

[10] https://github.com/kbenoit/quanteda.dictionaries

A further limitation is that the present analyses do not establish robustness to deliberately evasive AI-generated writing. The AI-labelled entries in this dataset appear to reflect naturally occurring or opportunistic AI mediation, rather than texts generated under prompts designed to avoid detection. It therefore remains unclear whether the same feature profile would hold for out-of-sample AI-generated journal entries explicitly prompted to sound more human, for example by using shorter sentences, more present-tense first-person language, less formal structure, lower coherence, more tangential topic shifts, or more idiosyncratic affective expression. Future validation should therefore include adversarial or prompt-engineered AI journal entries, ideally generated across multiple language models and prompt styles, to test whether the proposed workflow detects AI mediation beyond the specific linguistic regularities observed here.

A related limitation concerns the specific generative models that plausibly produced the AI-labelled entries. Data collection took place from September 2023 until August 2024, a period in which widely accessible models included GPT-3.5 and GPT-4 (via ChatGPT), Claude 2, and Google's Bard/Gemini; participants were not asked which tool they used, so the exact models are unknown. As language models are updated frequently and newer generations may produce more informal, present-tense, first-person prose that better approximates natural journalling style, the linguistic markers identified here may not generalise to text from newer or different models. The present workflow should therefore be understood as characterising perceived AI authorship for models available at the time of data collection, with the same feature profile and thresholds requiring re-validation against subsequent model generations.

Finally, the study was conducted in English, with participants recruited through a specific online study context and asked to respond to a particular daily journalling prompt. The generalisability of these perceptual heuristics across languages, cultures, platforms and forms of introspective writing therefore remains an open question. This is especially important because language models are changing rapidly and may become increasingly capable of producing informal first-person prose. Features that marked perceived AI authorship in this dataset should therefore be treated as historically situated and genre-specific cues rather than stable signatures of AI-generated text.

## Conclusion

This study shows that perceived AI authorship in diary-style writing is not marked by a single linguistic cue, but by a compact and interpretable constellation of features. Entries perceived as human were characterised by greater present focus, self-reference and idiosyncratic stylistic variability, whereas entries perceived as AI showed more formal function-word scaffolding, higher readability grade levels, greater coherence and a more polished uniform register. These findings suggest that, in introspective writing, the markers of perceived authenticity are not the conventional markers of writing quality. Instead, readers appear to treat immediacy, affective presence and stylistic irregularity as signals of human authorship.

Methodologically, the study provides a transparent workflow for auditing possible AI mediation in free-text psychological data. The workflow combined interpretable linguistic features with an explicit triage mechanism, identifying entries that could be treated as higher-confidence cases and those that required human review. This offers a reportable alternative to opaque commercial detector scores. More broadly, the findings show how psychological researchers can respond to AI contamination without treating detection as a purely automated verdict: by making the linguistic basis of classification visible, preserving ambiguous cases for review and recognising that the very irregularities often cleaned from text may be central to its psychological meaning.


# Acknowledgements

We wish to thank the participants who contributed daily journal entries over the two-week study period, and the UCL student annotators who completed the human labelling task.

# Funding statement

J.K. is supported by UCL's Computing Career Development Studentship. The development of experience sampling chatbot for the study was supported by an academic fellowship received by J.K. from the technology company OneReach.ai, which provided the access to Generative Studio

X allowing for building and deploying the chatbot, associated training, and technical support. OneReach.ai had no role in the study's design, data collection, analysis, or interpretation. J.I.S. is supported by the Wellcome Leap. A.H. was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant EP/V030302/1), Keystone grant funding from Responsible AI UK (grant EP/Y009800/1), the Alan Turing Institute (grant EP/N510129/1), and an MRC grant (MR/X030725/1).

## Data availability

The dataset generated and analysed during the current study includes sensitive, personally identifiable content (i.e., journal transcripts). Fully anonymised transcripts, with all identifying information removed, are available from the corresponding author upon reasonable request, subject to appropriate ethical approval and data use agreements.

## Competing interests statement

No authors declare conflicts of interests.

## Authors’ contributions

J.K.: Conceptualisation, Methodology, Software, Formal analysis, Data curation, Visualisation, Writing - original draft, Funding acquisition, Project administration.

A.H.: Methodology, Formal analysis, Writing - review & editing.

G.C.: Methodology (rater sensitivity analysis, detector decomposition), Formal analysis, Visualisation, Writing - original draft (supplementary analyses), Writing - review & editing.

M.E.A.: Methodology.

T.T.: Methodology.

M.L.: Methodology.

D.L.: Software, Writing - review & editing.

J.I.S.: Supervision, Conceptualisation, Project administration, Writing - review & editing.

# References


Amjad, M., & Tanveer, H. (2026). A comparative discourse analysis of AI generated and human-written texts: Coherence, framing, and persuasion. *Journal of Arts and Linguistics Studies*, *4*(1), 387–402.

Bazarova, N. N., Taft, J. G., Choi, Y. H., & Cosley, D. (2013). Managing impressions and relationships on Facebook: Self-presentational and relational concerns revealed through the analysis of language style. *Journal of Language and Social Psychology*, *32*(2), 121–141.

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, *3*. https://doi.org/10.1038/s41746-020-0233-7

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews. Neuroscience*, *17*(11), 718–731.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying Mental Health Signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021). Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688–701.

Dik, S., Erdem, O., & Dik, M. (2025). Assessing GPTZero's accuracy in identifying AI vs. Human-written essays. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.2506.23517

Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. *Proceedings of the 57th Annual Meeting of the Association for Computational*

*Linguistics: System Demonstrations*, 111–116.

*GPTZero*. (n.d.). GPTZero. Retrieved June 10, 2026, from https://gptzero.me/

Hills, A., Tseriotou, T., Miscouridou, X., Tsakalidis, A., & Liakata, M. (2024). Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. *Findings of the Association for Computational Linguistics ACL 2024*, 12526–12537.

Jacobson, N. C., Weingarden, H., & Wilhelm, S. (2019). Digital biomarkers of mood disorders and symptom change. *Npj Digital Medicine*, *2*(1), 3.

Jung, G., Choi, S., Kang, Y., & Kim, J. (2024). MyListener: An AI-mediated journaling mobile application for alleviating depression and loneliness using contextual data. *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 137–141.

Kappen, M., Vanderhasselt, M.-A., & Slavich, G. M. (2023). Speech as a promising biosignal in precision psychiatry. *Neuroscience and Biobehavioral Reviews*, *148*(105121), 105121.

Kelly, D. L., Spaderna, M., Hodzic, V., Coppersmith, G., Chen, S., & Resnik, P. (2021). Can language use in social media help in the treatment of severe mental illness? *Current Research in Psychiatry*, *1*(1), 1–4.

Kuc, J., Lametti, D. R., Camburu, O.-M., & Skipper, J. I. (2023). *The Psychological Benefits of Journaling and Feedback through a Large Language Model-Driven Conversational Bot*. https://osf.io/3utsm/resources

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2304.02819

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.1705.07874

Malik, M. A., & Amjad, A. I. (2025). AI vs AI: How effective are Turnitin, ZeroGPT, GPTZero, and Writer AI in detecting text generated by ChatGPT, Perplexity, and Gemini? *Journal of Applied Learning & Teaching*, *8*(1). https://doi.org/10.37074/jalt.2025.8.1.9

McCombie, C., Esponda, G. M., Ouazzane, H., Knowles, G., Gayer-Anderson, C., Schmidt, U., & Lawrence, V. (2024). Qualitative digital diary methods: participant-led values for ethical and

insightful mental health research. *International Journal of Qualitative Methods*, *23*, 16094069241296189.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2301.11305

Nepal, S., Pillai, A., Campbell, W., Massachi, T., Heinz, M. V., Kunwar, A., Choi, E. S., Xu, X., Kuc, J., Huckins, J. F., Holden, J., Preum, S. M., Depp, C., Jacobson, N., Czerwinski, M. P., Granholm, E., & Campbell, A. T. (2024). MindScape study: Integrating LLM and behavioral sensing for personalized AI-driven journaling experiences. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(4), 1–44.

Nook, E. C., Schleider, J. L., & Somerville, L. H. (2017). A linguistic signature of psychological distancing in emotion regulation. *Journal of Experimental Psychology. General*, *146*(3), 337–346.

Padmakumar, V., & He, H. (2023). Does writing with language models reduce content diversity? In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2309.05196

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, *8*(3), 162–166.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, *72*(4), 863–871.

Sen, V., & Hasan, R. (2025). A feature-based linguistic and statistical framework for AI-generated text detection. *2025 International Conference on Machine Learning and Applications (ICMLA)*, 184–191.

Sohal, M., Singh, P., Dhillon, B. S., & Gill, H. S. (2022). Efficacy of journaling in the management of mental illness: a systematic review and meta-analysis. *Family Medicine and Community Health*, *10*(1). https://doi.org/10.1136/fmch-2021-001154

Stamatatos, E. (2017). Authorship Attribution Using Text Distortion. *Proceedings of the 15th*

*Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1138–1149.

Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A., & Firth, J. (2025). The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *24*(2), 156–174.

Tsakalidis, A., Nanni, F., Hills, A., Chim, J., Song, J., & Liakata, M. (2022). Identifying moments of change from longitudinal user text. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4647–4660.

Tseriotou, T., Chim, J., Klein, A., Shamir, A., Dvir, G., Ali, I., Kennedy, C., Singh Kohli, G., Hills, A., Zirikly, A., Atzil-Slonim, D., & Liakata, M. (2025). Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, 193–217.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 26.

Zou, Y., Kuek, F., Ng, K. H., & Cheng, X. (2026). Comparative analysis of text readability and writing styles in AI-generated vs. Human-written academic abstracts. *PloS One*, *21*(4), e0343163.