TURING: Evaluating Human Abilities to Identify AI-Generated Texts

Natalia Kalashnikova, Nicolas De Bufala, Sophie Fayad, Laurent Cervoni
Centre de Recherche et d'Innovation de Talan
14 rue Pergolèse, Paris, France
{natalia.kalashnikova, nicolas.de-bufala, sophie.fayad, laurent.cervoni}@talan.com

Abstract

This study analyzes humans' ability to identify AI-generated texts across 10 genres. We collected 9164 annotations from 214 participants on 500 texts (half human, half LLM-produced), and analyzed 7943 after quality screening. Our main findings are that the humans accuracy was above chance but far from perfect (around 59%), with a slight tendency to label texts as "Human-generated". Their performance is influenced by the text genre (structural/factual formats easier to identify *vs.* complex genres) and by generating LLM. Annotators optionally selected three-level descriptors to justify decisions. While they had very limited effects on accuracy, their usage showed some association between text features (monotony, lack of cohesion or coherence) and "AI-generated" labeling. However, the linguistic features of the texts appear to have no robust impact after correction on human judgment. A small learning effect emerged but was practically negligible (0.1-0.2%), and personal characteristics of annotators had an impact on their accuracy, except age, which showed no effect. Finally, two automated detection tools were tested, reaching 88% accuracy on our distribution, clearly above humans, highlighting the value of human-tool combinations.

Keywords: AI-Generated text detection, NLP, hybrid dataset, Natural Language Features, AI-Generated texts descriptors, Comparative study

1. Introduction

The ability to determine whether content has been generated by a machine or not seems increasingly crucial in light of the many potential risks posed by artificial content, such as the quality and reliability of information, reputation damage, privacy issues, etc. (Park et al., 2023). As the technology behind text generation continues to advance, the value of studying humans' ability to detect whether a text was generated by AI appears all the more relevant. Recent studies estimate that more than a third of posts on some platforms (Quora, Medium) are AI-generated (Sun et al., 2025).

Most research on this topic (Alshareef et al., 2025; Kadiyala et al., 2025; Su et al., 2025) focuses more on the ability of an algorithm/program/AI to correctly identify the nature of content, rather than on human capability in performing the same task. However, in an era where artificial content is only becoming more frequent in almost all spheres of our lives, it seems interesting to evaluate this ability across different dimensions, such as text genres, textual features, personal characteristics, etc. We believe that the ability to recognize that a text is not "human" and may potentially contain false information or ethical issues is an important skill in the professional world.

Recent research has highlighted the challenges that are faced by humans and automatic methods in distinguishing between AI-generated and human-authored texts. Several studies converge on the observation that accuracy often remains close to chance. For instance, Fleckenstein et al. (2024) showed that novice teachers were essentially guessing, while even experienced teachers only improved when evaluating high-quality texts. Liu et al. (2024) similarly reported recognition rates barely above random performance. Multiple studies have shown that despite this limited accuracy, participants frequently felt certain about their classifications while systematically misjudging the true origin of texts (Chein et al., 2024; Fleckenstein et al., 2024; Milička et al., 2025).

Recent studies suggest that training and feedback can partially mitigate these difficulties. Thus, Milička et al. (2025) demonstrated that participants who received feedback progressively improved their performance in recognizing the origin of text, whereas those without feedback relied on clues such as cohesion, readability, and thematic structure. Russell et al. (2025) extended this perspective by showing that prior experience with LLM influences detection strategies. Non-experts (participants who had not used LLM for content creation) tended to rely on superficial features like grammatical correctness, low-frequency vocabulary, or neutral tone, while experts (participants who had used LLM for content creation) were more sensitive to subtle stylistic and lexical characteristics of AI, and were able to detect such patterns even across different models.

Beyond expertise, strong inter-individual variability has also been observed. Chein et al. (2024) found that fluid intelligence (that is defined as the capacity for analytical thinking) was the most reliable predictor of participants' ability to detect the origin of text. In contrast, lifestyle factors such as

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4527–4535
11–16 May 2026.

frequent use of smartphones and social media were associated with systematic errors, particularly the misattribution of AI-generated content to human authors.

These findings indicate that this ability is not evenly distributed, but rather shaped by both cognitive and experiential differences.

Another field of research concerns the linguistic and stylistic markers perceived as clues of AI authorship. Annotators frequently pointed to several characteristics, such as "AI vocabulary" (e.g., "vibrant", "crucial", "significantly"), overly optimistic conclusions, and reduced originality (Russell et al., 2025). Najjar et al. (2025), focusing on cybersecurity discourse, further showed that human texts tend to rely on more practical and concrete terms such as "use" or "allow", while ChatGPT-produced texts employed more abstract and formal expressions like "realm" or "employ". The authors also noted that short texts are harder to classify reliably than longer ones, since diagnostic cues are less evident.

A final line of research has compared human performance with automatic detection systems. While Russell et al. (2025) reported that human annotators maintained stable accuracy even with paraphrased or "humanized" AI texts, detection tools produced inconsistent results, with only one closed-source system achieving parity with humans. Moreover, when LLMs were tasked with classification under the same instructions as human annotators, their performance was inferior. Improvements were only observed when models were prompted with structured reasoning approaches, such as chain-of-thought, and when they incorporated the same explanatory cues identified by human participants.

Few studies have yet examined the influence of personal characteristics, such as age, field of expertise, or education level, on the ability to detect AI-generated text. Moreover, most existing research has focused on English and on a limited range of text types. To our knowledge, none have explored the capacity to identify texts across multiple genres in French. In addition, while humans appear to rely on stylistic cues, no unified framework has been proposed to quantify the impact of different linguistic features. To fill this gap, the present study investigates the influence of participants' personal characteristics and linguistic features of texts across multiple genres on humans' ability to recognize AI-generated text, and compares their efficiency with recent tools.

The article is organized as follows. Section 2 introduces the dataset, the experimental protocol, and the linguistic and personal features. Subsection 2.6 details data-quality screening and the statistical analysis framework. Section 3.1 reports preprocessing outcomes; Subsection 3.2 presents overall performance via the confusion matrix; Subsections 3.3 and 3.4 analyze textual factors and annotator-provided descriptors; Subsection 3.5 examines personal characteristics. Section 4 compares human judgments with automatic detectors, and the final section concludes our study, while discussing limitations and future work.

2. Methodology

For clarity and conciseness, we adopt the following conventions throughout this article: "AI texts" refers to AI-generated texts, and "Human texts" refers to human-written texts.

2.1. Main Research Questions

In this study, we aim to address the following research questions:

1. How accurately can French speakers detect AI text?
2. To what extent do personal characteristics influence a person's ability to perform this task?
3. Does performance vary depending on linguistic features (genre of the text, length, complexity, etc.)?
4. Are there linguistic differences (e.g., typing errors, overuse of adjectives, etc.) between AI texts and Human texts? If so, do they impact the efficiency of humans on this detection task?
5. Is there a learning effect when participants receive feedback with the correct answer?
6. How efficient are humans on this task when compared to publicly available tools?

2.2. Dataset

Our dataset consists of 500 French texts of various genres. Around half of the texts are human-written, collected from different sources across multiple domains, including recipes, emails, jokes, poems, novels, political speeches, literary criticism, newspaper articles, and discussions. In parallel, we used several LLMs[1] , such as GPT, Mistral, Claude, etc., to generate texts in the same genres. Table 1 summarizes key characteristics of the resulting dataset.

[1]As the text generation step of this experiment was conducted in mid 2024, public available models were GPT-3.5, Claude 3, Gemini 1.5, Mistral Small ("LeChat"), and Copilot which was based on GPT 4-Turbo.

Genre of text	Origin of text	Number of texts	Average number of sentences	Average number of unique words	Average words per sentence	Average total words
Newspaper article	AI	30	19.3	206.1	21.8	416.1
	Human	30	16.2	203	24.8	377.3
Joke	AI	20	10.5	65.8	13.5	129.7
	Human	20	8.7	62.3	11.9	99.9
Literary criticism	AI	30	5.8	74.8	18.8	107.4
	Human	30	9.8	142.2	26.2	230
Political speech	AI	21	18.8	197.7	20.4	374.7
	Human	21	18.9	217.2	24.5	426.3
Discussion	AI	22	12.6	66.1	11	112.9
	Human	16	8.4	76.2	29.1	152.5
Email	AI	50	3.5	32.6	11.6	41.8
	Human	20	4.6	33	7.8	36.1
Theater play	AI	15	14.8	103.2	12.1	166.9
	Human	15	18.7	84.9	8	134
Poem	AI	30	7.8	93.3	21.4	150.7
	Human	30	5.4	76.4	36.2	117.3
Recipe	AI	20	11.25	65.35	11.5	113.4
	Human	20	12.5	97.8	15	178.7
Novel	AI	30	27.5	223.9	16.3	412.1
	Human	30	24.9	238	21.2	462.8

Table 1: Description of dataset

2.3. Experimental Design

We developed a web application to conduct the experiment within our company. On this platform, users log in using their professional email address and are informed about the context of the experiment, their task, and the potential collection of personal data, such as their position within the company or age, at a later stage. The login message is as follows:

"You will be presented with ten texts of different types (theater play, novel, poem, political speech, etc.). These texts were either written by humans or by artificial intelligences such as ChatGPT. Your task is to determine whether the text you are reading was written by a human or an AI. You can explain your reasoning using the different options displayed on the page or by writing a comment in the free-text field. Then confirm your choice by clicking one of the buttons at the bottom of the page, which will load the next text. You may stop at any time by simply closing the page. Thank you for your participation. As part of this experiment, some personal data will be collected, such as your age, level of education, current position, and professional rank. Your data will then be anonymized before being used by our research center. All data is hosted on secure servers in France, in compliance with GDPR regulations, and will only be used for statistical analysis across all participants." The experiment is presented as a series of texts, displayed one by one.

The user is asked to read the content and then answer the main question of this evaluation: *"In your opinion, what is the origin of this text?"*, with three possible answers: *AI, Human,* and *I don't know*.

The user can optionally select among 10 predefined descriptors to characterize the text, and explain their choice about the text's origin.

- Incoherent content / Lacks common sense *vs.* The content appears coherent and logical
- Lack of cohesion or continuity in the text *vs.* The text flows well and maintains continuity
- Strange or inappropriate vocabulary *vs.* Vocabulary is coherent and appropriate for the text
- The tone is inappropriate (too formal or too informal) *vs.* The tone is normal and appropriate
- The text lacks variety or is too monotonous *vs.* The text is varied and sounds natural
- Too many repetitions or paraphrasing *vs.* The text gets straight to the point
- Presence of spelling, grammar, or typing errors *vs.* No mistakes found in the text
- Presence of anglicisms or abbreviations *vs.* No anglicisms or abbreviations

- Overuse of superlatives or adjectives *vs.* Normal amount of adjectives and superlatives
- Too many details and specifications *vs.* Few or an appropriate amount of details

The goal of these descriptors is to better understand what motivated the user's answer and to understand the user's intuition.

After submission, a pop-up message appears indicating the correct answer. This process is then repeated, and every 10 texts, the user gets a recap of their performance on this test. The platform was designed to automatically log the review time (RT) for each annotation, stored with a 1-second resolution.

A total of 214 participants took part in the experiment, yielding 9164 responses in total.

2.4. Linguistic Features

We were inspired by the research of Fredrick and Craven (2025); Chen and Hong (2025) to calculate the following linguistic features for each text in the dataset:

- Length - total number of tokens per text.
- Number of sentences - number of sentences per text.
- Average number of tokens per sentence - length/number of sentences.
- TTR (BERT) - ratio between the number of unique words and the total number of words per text. Tokenization was performed using BERT (Devlin et al., 2019).
- TTR (spaCy) - ratio between the number of unique words and the total number of words per text. Tokenization was performed using spaCy (Honnibal et al., 2020).
- Maximum syntactic depth - maximum depth of the syntactic trees in the text.
- Average syntactic depth - average depth of the syntactic trees in the text.
- Position number - number indicating the text order for each participant.
- Position block - number indicating each participant's block order.
- Review time (RT) - elapsed time between text display and answer submission, in seconds (s).
- Speed (words/s) - time of analyzing the text (length / RT).
- Time per word (s/word) (RT / length)

2.5. Personal Characteristics

We transmitted the list of participants to our company's HR department to collect personal characteristics, such as age, educational degree, field of study, and current job family. We then anonymized the participants' responses by replacing their email addresses with identification numbers. After that, we added personal characteristics to the responses and deleted the list containing the emails and participant numbers, ensuring complete anonymity. All procedures between the HR department and the authors of this study were secured by a non-disclosure agreement (NDA).

2.6. Statistical Analysis

2.6.1. Data quality and preprocessing

Annotations indicating "I don't know", included to avoid forcing a binary choice and thereby reduce guessing bias, were cleaned from all analyses along with annotations with null review time. Uninformed/random responses were screened using RT and speed features. We first examined accuracy as a function of RT. For each 1-second bin, we estimated accuracy with a Wilson 95% confidence interval (CI) and tested against chance using a one-sided binomial test. Multiplicity was controlled across bins with Benjamini–Hochberg False Discovery Rate (FDR). To remove residual ultra-fast, likely uninformed responses, we then excluded the fastest 10% of annotations, thus retaining 90% of the clean dataset. We also tested sequence effects by performing logistic regression of prediction correctness on the within-annotator text rank of each item in the annotation sequence. Population-level effects were estimated using binomial generalized estimating equation (GEE, logit link) clustered by annotator with an exchangeable working correlation and standard errors (SE). As descriptive checks, we also inspected per-annotator slopes.

2.6.2. Confusion Matrix and Performance Metrics

After preprocessing, a 2x2 confusion matrix (counts and normalized rates) was computed between the true origin (AI/Human) and the annotator's decision (AI/Human). We report the following metrics: overall accuracy (with Wilson 95% CI), balanced accuracy (mean recall), macro/weighted precision/recall/F1-score, per-class metrics, and Matthews correlation coefficient (MCC). In single-label classification, micro-averaged precision, recall and F1 coincide with accuracy. We therefore report accuracy and omit micro-averaged metrics.

2.6.3. Textual Factors

Variations in performance according to category of text (accuracy) and LLM of origin (AI recall) were described with Wilson 95% CI and tested with χ^2 homogeneity tests followed by Holm-corrected pairwise comparisons. To quantify effects on performance, we fit separate one-predictor GEEs (with text category as covariate). For continuous linguistic features (length, syntactic depth, lexical diversity proxies, RT, speed, time per word, text rank) predictors were z-scored and coefficients are interpreted as odds ratio per 1 standard deviation (OR / 1 SD). Spearman correlations are also provided descriptively (direction and magnitude). Three-level descriptors selected by annotators to justify their decision (-1 problematic / 0 neutral / +1 adequate) were encoded with 0 as baseline. We estimated three contrasts: (i) positive: +1 vs 0, (ii) negative: -1 vs 0, and (iii) polarity: +1 vs -1. For families of related tests (such as the full set of descriptor contrasts or the set of continuous predictors), we applied Holm correction (two-sided, α = 0.05).

2.6.4. Personal Characteristics

For categorical characteristics (degree, specialty, job family), we presented Wilson barplots with a χ^2 homogeneity test and Holm-corrected pairwise post-hoc comparisons. For continuous characteristics (age and diploma), we first reported Spearman descriptives, then fit a GEE of correctness on z-scored age and diploma age (with category as a covariate), reporting OR per 1 SD with 95% CI and Holm-adjusted p-values. This isolates the trait effect from genre composition while accounting for within-annotator dependence.

2.6.5. Implementation Details

All analyses were conducted in Python: statsmodels (Seabold and Perktold, 2010) for GEE/GLM (robust covariance; exchangeable correlation), SciPy for statistical tests, and patsy for design matrices. Uncertainty is reported as Wilson 95% CI for proportions and as robust SE-based 95% CI for model coefficients/ORs. We applied listwise deletion within each analysis (no outcome imputation). All GEEs were clustered by annotator. Two-sided tests were used unless otherwise specified (RT vs chance is one-sided).

3. Results

3.1. Data Preprocessing

Annotations indicating "I don't know" represented 3.55% of the raw dataset (n = 325 / 9164). They were excluded, along with 13 annotations showing null review time (RT). All subsequent analyses were therefore performed on n = 8 826 annotations, which we refer to as "clean dataset". Prior to our analysis, we screened uninformed/random responses, using first review time (RT) as a quality screen. To this end, for each one-second bin of RT, we estimated the accuracy with a Wilson 95% CI and performed a one-sided binomial test vs. 50%, controlling across bins by Benjamini-Hochberg FDR (Figure 1). We defined the minimal reliable RT as the first second that initiates a run of two consecutive bins whose lower 95% CI bound exceeded 50% accuracy and whose FDR-adjusted p-value was < 0.05. This approach yielded a RT_lower = 7s threshold below which we excluded 593 annotations, thus retaining 93.3% of the clean dataset (n = 8233 / 8826). Post-filter, the speed distribution showed a thin tail corresponding to ultra-fast annotations. We imposed a proportional cap chosen so that the final dataset represents 90% of the 8826 clean volume. This corresponded to the removal of the 290 fastest responses after the RT filter and a speed cap of 23.25 words/s. Mean accuracy remained constant at 0.593 at all 3 stages of pre-processing (raw, review time filter, speed trimming), indicating that these steps cleaned the dataset without shifting the global signal. After RT filter and speed capping, **the final dataset represented n = 7943 annotations**. Finally, to check for sequence effects along annotations, we fitted a user-clustered logistic GEE model. Using text rank in the annotation sequence as a predictor of performance, we observed a small positive slope (β=0.0018, OR=1.0018 per item; p<0.001). These magnitudes are statistically detectable but practically negligible, so we treat observations as effectively exchangeable with respect to sequence

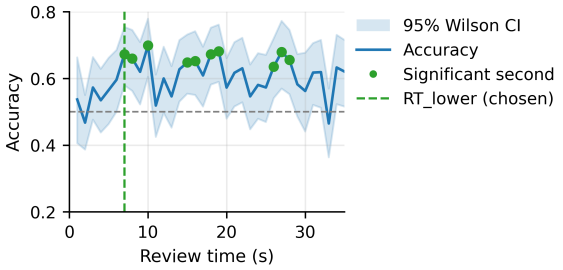

Figure 1: Review time threshold analysis. Accuracy ± Wilson 95% CI (estimated as the proportion of correct answers), according to review time (s). Green dots mark seconds where accuracy is significantly above chance (one-sided binomial test vs 0.5, Benjamini–Hochberg FDR). The vertical dashed line indicates the chosen lower time threshold (RT_lower = 7s), defined as the first second that starts a run of $\geq$ 2 significant bins. Seconds all contained at least 20 observations.

position.

3.2. Confusion Matrix

After data preprocessing, we analyzed the final dataset (n = 7943, AI texts = 4253, Human texts = 3690). Overall, annotators performed above chance (0.593, 95% CI = [0.582, 0.604], MCC = 0.19, cf. Table 2). Interestingly, errors were asymmetric, with 43.6% of AI texts mislabeled as Human *vs* 37.3% of Human texts mislabeled as AI (Figure 2). This suggests a "humanization" bias: the AI label was more precise but under-called (precision = 0.635, recall = 0.564), whereas the Human label showed higher recall but lower precision (recall = 0.626, precision = 0.555). This asymmetry held across categories: AI recall ranged from 0.44 (novels) to 0.70 (articles), whereas Human recall ranged from 0.54 (recipes) to 0.70 (jokes). In 5/9 categories (jokes, literary criticism, speeches, poetry, novels), AI recall was lower than Human recall. Consistently, AI prediction instead of Human spanned 0.30-0.56 (lowest in articles, highest in novels) versus 0.30-0.46 for Human instead of AI.

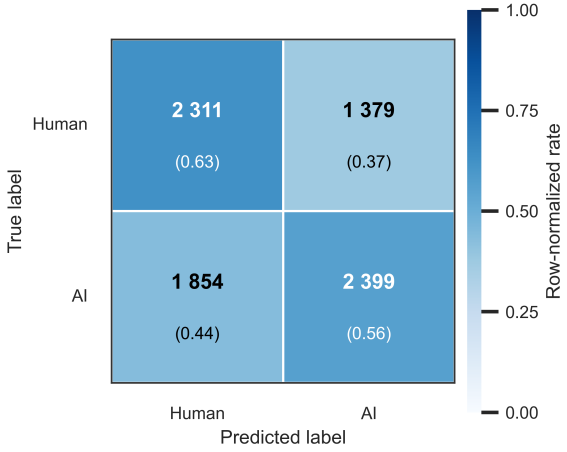

Figure 2: Overall confusion matrix. Rows are ground truth (Human, AI); columns are annotator predictions. Each cell reports the count and within-row rate (normalized) which is mapped to by color scale.

3.3. Contribution of Textual Factors to Performance

When analyzing accuracy according to the category of text, a global test shows strong differences (χ^2: p < 1e-16, Figure 3 A). Pairwise comparisons show the highest accuracies for articles, jokes, and poems, compared to the lowest accuracies for speeches, theater excerpts, and recipes. These results confirm that text category explains a significant part of data variance with easier identification

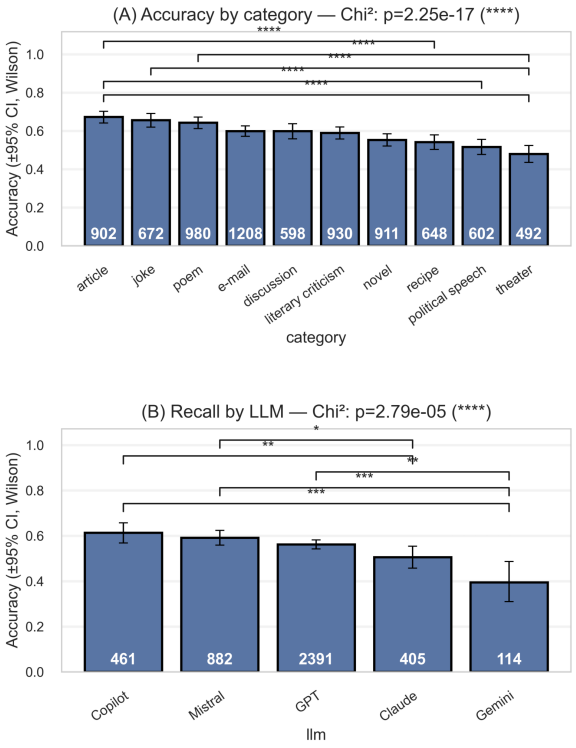

Figure 3: Accuracy by text category and LLM. (A) Bar plot of mean accuracy (proportion correct answers) for each text category with Wilson 95% CIs. (B) Bar plot of AI recall (true positive rate for the AI class) according to the LLM used for generation. Sample sizes are indicated in white on each bar. Significance is encoded as p<0.05 (*), 0.01 (**), 0.001 (***), and 10e-4 (****).

for structured/factual formats compared to speech or theaters, which seem harder to detect. Across LLMs used for generation, accuracy also varied (χ^2: 2.79e-5, Figure 3 B). Copilot (0.614 [0.569, 0.657], n = 461) and Mistral (0.592 [0.559, 0.624], n = 882) outperformed Gemini (0.395 [0.310, 0.486], n=114, p-adjusted $\leq$ 0.001, but with small size) and Claude (0.506 [0.458, 0.555], n = 405, p-adjusted = 0.024). Differences between GPT (0.562 [0.542, 0.582], n = 2391) and Mistral/Copilot were not significant after multiplicity correction. In order to understand such differences, several text features were analyzed. One predictor GEEs clustered by annotator and adjusted on category found no robust effect for any of these parameters (length, number of sentences, average number of words per sentence, syntactic depth, lexical diversity proxies = ttr, review time, and time per word. However, text rank within annotation sequences showed a small but consistent association with accuracy (odds ratio / 1 SD = 1.12, p-adjusted < 0.001), which confirms the small learning effect previously observed (cf. 3.1).

Metrics	*Overall*	*Weighted*	*Per-class (true Human)*	*Per-class (true AI)*
N	**7943**	–	**3690**	**4253**
Accuracy	**0.593**	–	–	–
Balanced accuracy (mean recall)	**0.595**	–	–	–
Precision	–	0.5978	0.5549	**0.635**
Recall	–	0.5930	**0.6263**	0.5641
F1 score	–	0.5932	0.5884	0.5974

Table 2: Evaluation metrics. Micro-averaged metrics equal accuracy in single-label setting. Highest values are in bold and cited in the text.

3.4. Contribution of Textual Descriptors to Performance

Standardized descriptors were proposed to annotators to optionally justify their answers, presented as three-level factors (-1: problematic, 0: neutral/not used, +1: adequate). We thus estimated three contrasts: positive (+1 vs 0), negative (-1 vs 0) and polarity (+1 vs -1) in GEE clustered by user and adjusted on category in order to compute the associated odds ratios. After Holm correction, only coherent/logical content remained significant on the polarity contrast, increasing the probability of correct judgment by 40 to 50% (odds ratio = 1.4-1.5, p-adjusted $<$ 0.01). Other descriptors were direction-consistent (positive cues help, negative cues penalize) but did not pass multiplicity correction. This points to perceived coherence or logic as the most stable cue beyond category.

We observed an interesting pattern in how annotators used descriptors in this experiment: certain descriptor values were strongly associated with a specific judgment. As shown in Figure 4, there was a bias toward "Human" labels when descriptors were used, with high contrasts between some descriptor values. For instance, when annotators noted that the text was "Too monotonous/Lacked diversity", 80% labeled the text as "AI", while only 30% did when they noted the opposite descriptor value. The bias towards positive descriptor values reflects the tendency of some annotators to use descriptor positive values as a checklist to analyze a text.

3.5. Contribution of Personal Characteristics to Performance

We further examined whether personal characteristics were associated with different detection accuracy. Global tests indicated heterogeneity across degrees (p=0.00637), specialties (p=0.000335), and job families (p=2.4e-8). Extremes were driven by small-n groups (MBA high; Associate low). Mainstream groups (High-school, Bachelor, Master, PhD) clustered around 0.55–0.62 with overlapping CIs. Several pairwise differences remained significant after correction in subfamily analyses,

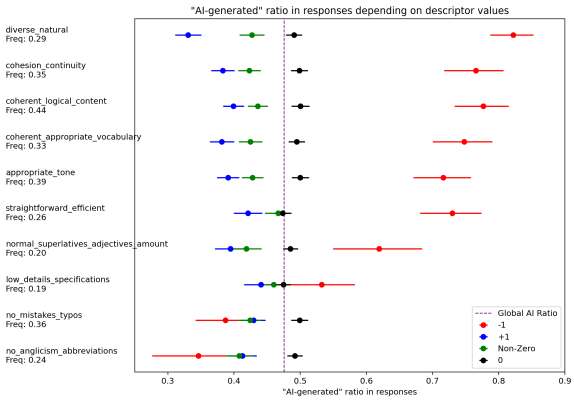

Figure 4: AI Ratio in responses by annotators, depending on their descriptors value choices, with 95% Wilson CI. Non-Zero corresponds to annotators that selected either +1 or -1. The vertical dashed line indicates the AI Ratio in responses in all annotations.

suggesting real but localized gaps. For age, an unadjusted Spearman correlation was small but significant (ρ = -0.042, p = 2.5e-4). In a GEE clustered by annotator and adjusted on category, the independent age effect was tiny and non-robust (OR / 1 SD $\approx$ 0.93, 95% CI [0.86, 1.01], p-adjusted = 0.18), corresponding to at most 1–2 percent points around the mean. The Spearman vs. GEE discrepancy is consistent with category confounding and within-user dependence: once controlled, age contributes little. Overall, personal characteristics explain a smaller share of variance than text-side determinants. Differences exist and are statistically detectable in places, but effects are modest, and age is near-null after proper adjustment.

4. Comparison with Automatic Tools

In previous studies (Fraser et al., 2025; Russell et al., 2025; Fariello et al., 2025), researchers noted the high performance of two open-source automatic tools: Binoculars (Hans et al., 2024) and FastDetectGPT (Bao et al., 2024) for detecting AI texts.

Fast-DetectGPT is a one-shot detector that in-

troduces the concept of *conditional probability curvature* to account for divergences in word choice between LLMs and humans within a given context. The hypothesis of Bao et al. (2024) is that LLMs tend to favor tokens with higher statistical probabilities due to pre-training on large-scale corpora, whereas humans construct sentences based on underlying meanings and contextual cues.

The Binoculars method introduces a zero-shot detector based on the combination of two statistical measures analyzed by two pre-trained language models. The first model estimates the perplexity of the text, while the second model generates token predictions that are then evaluated by the first model to compute the cross-perplexity, which reflects how "surprising" the predictions are for the model. Cross-perplexity serves as a normalization mechanism: if a text closely resembles what the second model would have produced, cross-perplexity decreases and the tool suggests that the text may be AI (Hans et al., 2024).

Both tools performed better than humans, achieving 0.88 of accuracy *vs.* 0.59 for humans' accuracy. This finding indicates that humans are not effective at this task on our distribution. Moreover, given the continuous improvement of AI systems, at present, automated detectors outperform unaided humans on our distribution. Future work should study human-tool combinations and cost-sensitive settings.

5. Conclusion

This study examined humans' ability to accurately detect AI texts, and our findings reveal that this accuracy is approximately 59%, better than chance, but substantially lower than the almost 90% accuracy of automatic detectors (open source in our comparison) reached 0.88 on our distribution, clearly above humans (0.59). Annotators showed a small bias towards labeling texts as Human (52.5%).

Detection accuracy is affected by some of the text characteristics, mainly the text category, with annotators performing better on structural/factual formats (articles, jokes, and poems) than on more complex genres (political speeches, theater excerpts, and recipes). The source LLM behind AI texts also influenced performance: Copilot-generated texts were correctly identified 62% of the time, compared to 56% for GPT and 50% for Claude. Other text characteristics—including length, complexity, and lexical diversity—showed no significant effect. A small learning effect was found for this task, with around 0.1% accuracy gain per text evaluated.

Descriptors were available to users to explain their decision/intuition on the text origin, but close to no positive effects on accuracy were found behind these descriptors when compared to annotations without these descriptors. Certain descriptors' values were associated with classifying a text as "AI" (lack of diversity, monotonous text) or as "Human" (for its opposite).

Annotator characteristics, including education level, job category, and professional specialty, influenced detection ability, even if age showed no significant effect.

To improve and continue this study, the following points could be explored :

- Evaluating detection accuracy with texts generated by newer LLMs as rapid improvement of this technology may yield texts that more closely mimics human writing.
- Standardization of text-generation protocols to reduce quality variability across samples.
- A follow-up experiment could be done by systematically removing obvious detection cues to determine where these cues are inherent to models or if they originate from suboptimal prompting.

In the end, the ability to correctly identify if a text is AI seems to be limited in general, but the use of tools for this task, either to automatically detect or to give insight to the user, may become necessary in the future.

6. Bibliographical References

Abdulrhman M. Alshareef, Aisha Alsobhi, Alaa O. Khadidos, Khaled H. Alyoubi, Adil O. Khadidos, and Mahmoud Ragab. 2025. Automated detection of chatgpt-generated text vs. human text using gannet-optimized deep learning. *Alexandria Engineering Journal*, 124:495–512.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature.

Jason Chein, Steven Martinez, and Alexander Barone. 2024. Can human intelligence safeguard against artificial intelligence? exploring individual differences in the discernment of human from ai texts.

Jialei Chen and Qiufeng Hong. 2025. Lexical diversity and syntactic complexity in ai-translated legislative texts. *Theory and Practice in Language Studies*, 15(9).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Serena Fariello, Giuseppe Fenza, Flavia Forte, Mariacristina Gallo, and Martina Marotta. 2025. Distinguishing human from machine: A review of advances and challenges in ai-generated text detection. *International Journal of Interactive Multimedia and Artificial Intelligence*, 9:6–18.

Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, Stefan D. Keller, Olaf Köller, and Jens Möller. 2024. Do teachers spot ai? evaluating the detectability of ai-generated texts among student essays. *Computers and Education: Artificial Intelligence*, 6:100209.

Kathleen C. Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2025. Detecting ai-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, 82:2233–2278.

Daniel R. Fredrick and Laurence Craven. 2025. Lexical diversity, syntactic complexity, and readability: a corpus-based analysis of chatgpt and l2 student essays. *Frontiers in Education*, Volume 10 - 2025.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. Software library, version as used.

Ram Mohan Rao Kadiyala, Siddartha Pullakhandam, Kanwal Mehreen, Drishti Sharma, Siddhant Gupta, Jebish Purbey, Ashay Srivastava, Subhasya TippaReddy, Arvind Reddy Bobbili, Suraj Telugara Chandrashekhar, Modabbir Adeeb, Srinadh Vura, Suman Debnath, and Hamza Farooq. 2025. Robust and fine-grained detection of ai generated texts.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024. On the detectability of chatgpt content: Benchmarking, methodology, and evaluation through the lens of academic writing. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, page 2236–2250, New York, NY, USA. Association for Computing Machinery.

Jiří Milička, Anna Marklová, Ondřej Drobil, and Eva Pospíšilová. 2025. Humans can learn to detect ai-generated texts, or at least learn when they can't.

Ayat A. Najjar, Huthaifa I. Ashqar, Omar A. Darwish, and Eman Hammad. 2025. Detecting ai-generated text in educational content: Leveraging machine learning and explainable ai for academic integrity.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023. Ai deception: A survey of examples, risks, and potential solutions.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Zhixiong Su, Yichen Wang, Herun Wan, Zhaohan Zhang, and Minnan Luo. 2025. Haco-det: A study towards fine-grained machine-generated text detection under human-ai coauthoring.

Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are we in the ai-generated text world already? quantifying and monitoring aigt on social media.