

To Err is Human: How People Infer Authorship in the Age of AI

Yijia Xu^{1*}

Western Academy of Beijing, 10 Lai Guang Ying Dong Lu, Chao Yang District, Beijing 100 102, China

Corresponding author email: clairexu1018@gmail.com

Abstract

As LLMs increasingly generate human-like text, it is becoming increasingly pressing an empirical and societal question: how do we distinguish AI-generated text from human-written text? Through two studies, we explored whether and how people judge AI authorship. In Study 1, we had participants ($n = 101$) judge the written responses to an inside joke prompt—our chosen prompt based on LLMs' response—and assess each response on a 7-point scale of 1, definitely AI-written, to 7, definitely human-written, followed by self-reports of justifications. People's classifications were near chance accuracy (51.1%), but their continuous ratings did separate the human-written from the AI-written responses. Multivariate logistic regression showed that accuracy was best predicted by attention to surface-level errors and structural features rather than by deeper experiential cues (for example, interpretations of humor and personal schema matches). In Study 2, a new sample ($n = 100$) evaluated AI responses deliberately modified with spelling and grammatical errors. Accuracy dropped steeply in participants below chance (39.2%), validating our conclusion in Study 1. Together, the results of these two studies reveal that people's capacity to detect human expression rests less on detecting genuine lived experience than on superficial linguistic cues that are easy to manipulate.

Keywords

Social and Behavioral Science, Cognitive Psychology, Artificial Intelligence, Human-AI Discrimination, AI-Generated Text Discrimination

Introduction

Imagine it is your birthday today. Among the messages and notifications, one long, heartfelt letter from a close friend, full of warmth, personal anecdotes, and a familiar turn of phrase, stood out. But incidentally, you discover that it wasn't sent by your friend, after all: it had been written by an AI. Would that change how much you like the letter? Would you react differently or refrain from replying altogether? Would it feel as genuine a relationship with you as before? Just imagining this hypothetical scenario should have annoyed you, perhaps because the annoying part is not just the deceit but something deeper—the expectation that some messages should be coming from a human brain.

Nevertheless, as AI systems become increasingly adept at mimicking our communication, from a birthday message to an election speech, this intuition becomes increasingly difficult to rely on. Large language models (LLMs) are advanced AI systems designed to understand, process, and generate human-like text by analyzing vast datasets using deep learning.^{1,2} The past decade has witnessed a dramatic acceleration in LLMs' ability to produce text that is grammatically fluent, contextually coherent, and stylistically appealing and convincing.^{3,4}

In many domains, this development has been conspicuously productive. Many hospitals and healthcare organizations employ LLMs to draft and manage medical records, reducing the administrative burden on clinicians to free time for patient care.⁵⁻⁸ In e-commerce, when an online customer has a question about an item, the LLM can analyze the query and context to provide a relevant, accurate, and helpful response.^{9,10} In educational contexts, educators can use LLMs to automate time-consuming tasks, such as creating quizzes, significantly enhancing efficiency.^{11,12} LLMs function as private tutors for students, providing tailored guidance, instant feedback, and customized lesson plans, allowing them to learn at their own pace. They also support learners with diverse learning abilities, democratizing educational content.^{13,14} Therefore, LLMs have improved efficiency and broadened access to information across a myriad of domains. When the goal of individuals or institutions is to increase efficiency, expand access, and adapt to users' needs, the value of their output lies in its function, rendering the authorship of the content largely irrelevant.

However, in other contexts, the identity of the author does matter. A love letter has meaning from the writer's vulnerability. A eulogy has meaning from a real relationship between people informed by memory, affection and loss. Even in education, when a student turns in an AI-written essay, what is lost is not just academic integrity, but also the struggle through which understanding is actually achieved—the struggle of turning a half-known idea into one's own. Taking a step back, this authorship concern is not individual-level. In public-facing domains, including journalism, political communication, professional writing, and personal testimony, uncertainties about authorship can impact sources' credibility and accountability, especially when LLMs still struggle to provide accurate information.^{15,16} When readers cannot trust who is speaking anymore, suspicion will become the default, destroying our collective capacity for reasoning about what is true. More importantly, failing to distinguish AI-generated from human-authored texts will promote the use of such LLMs in written tasks with the efficiencies mentioned. What is at stake, then, are not just facts but an authentic human voice: the irreducible trace of a person, with particular experiences, writing at a certain moment.

Whether that trace is noticeable or not has become one of the most defining empirical questions of the field. Much of the scientific work on human-AI discrimination descends from Turing's defining question: Can a machine produce the same responses as a human?¹⁷ That framing has triggered a large empirical legacy, including several updated variants of the imitation game,^{18,19} studies on chatbot detection,²⁰ and, more recently, experiments examining whether humans can reliably detect LLM text.^{21–24} Together, this literature reveals that AI-generated content is becoming harder to discern and that human discrimination accuracy is often only slightly above chance.

Yet much of this work focused on the outcome of detection, whether humans are successful at that outcome or not, but not on the process by which humans arrive at their judgment. In other words, what cues are people using to decide that a piece of text was written by a person and not an AI? Only very recently have some of these gaps started to close. For example, Ippolito and the other researchers found that even editing an AI-written text on a superficial level could entirely undermine detection, insisting on surface features.²² The number of similar or replicated studies that decipher the cues that people use to judge and explain what contributed to accurate or inaccurate detections remains very limited.

Understanding how people make such decisions is essential in many ways. First, schools often assume that teachers or students can “tell” when something is AI-generated. But if people rely on unreliable cues, they may falsely accuse human writers or fail to recognize AI use. Second, it would unveil our biases about human expression. We would understand what people associate with human and AI writing, revealing underlying social and political expectations surrounding “humanness” in written texts. Uncovering how this could also change reveals how AI changes our relationship to language itself. Finally, it could also contribute to the development of AI-detection tools.

These “cues” that individuals rely upon for making judgments about authorship are thus a critical construct, constituting the notion of strategy. Specifically, throughout the study, a strategy is defined as the evaluative lens or technique a participant employs to judge a text's authorship. Examples include, but are not limited to, attending to grammatical imperfections, relying on emotional resonance, or drawing on beliefs about how AI systems typically write (the distinct strategies are defined later in this work). These strategies are meant to capture people's mental processes when deciding to attribute the authorship to the texts. To categorize the strategies, we relied on self-reports, which were then sorted using a systematic, AI-assisted coding procedure applied to written justifications. One challenge inherent in this approach is that self-reported strategies may not always align with the cognitive processes that actually drove judgments, a distinction we are attentive to throughout.

Overall, this study aims to investigate whether humans can accurately distinguish AI-generated text from human-written text and, critically, which cognitive strategies underlie those judgments. While prior work has established that detection accuracy is often modest, the field has yet to systematically examine what evaluative cues people rely upon, and whether those cues are genuinely diagnostic of authorship or merely superficially correlated with it. To address this gap, the participants evaluated written responses to an inside joke prompt. After each judgment, participants described their reasoning, allowing us to identify and test the predictive value of specific strategies. In a follow-up study, we directly manipulated the most predictive strategy that participants reported relying upon to conduct a causal test of their role in authorship detection. By doing so, we hope to shed light on the cognitive mechanisms underlying an increasingly common judgment in everyday life: how readers decide whether a text bears the mark of human authorship.

Methods

Stimulus Development

This research aimed to examine whether human evaluators could distinguish between responses generated by humans and those generated by an LLM and the specific strategies they relied upon. Because no standardized stimulus set exists for assessing “humanness” in open-ended language production, all stimuli were developed iteratively through a multi-phase process designed to reflect both human intuitions and AI strategies regarding what constitutes convincingly human text.

We began the stimulus development process by informing three widely used LLMs (ChatGPT, Gemini, and DeepSeek) about our investigation into how good humans are at detecting AI authorship. Because we want to give humans the best chance, we asked the LLMs to tell us which scenarios they would find most difficult to “pass” as humans. All three LLMs consistently identified two such scenarios: childhood memories involving sensory experiences and inside jokes (see Appendix A). While the former is more intuitive in meaning, the latter requires a more specific definition and clarification. Inside jokes are defined here, as they were defined for participants, as humorous references that are understandable only to a small group of people due to shared context or experience. This definition is consistent with prior work on conversational humor and shared mental models, as well as the modern dictionary.^{25,26} Unlike general trivia or abstract opinion pieces, inside jokes require the author to draw on genuinely idiosyncratic social memory, a specific group of people, a shared moment, and an emotional context that cannot be easily fabricated from statistical patterns alone. The LLMs justified the consensus of the two scenarios by pointing out that these prompts required access to personal experience, shared social context, and subjective sensory memory. However, they neither have authentic experiences nor a group of friends like humans do. Consequently, both domains are ecologically relevant and theoretically motivated as a test bed for authorship detection.

On the basis of informal testing, we found that the AI responses to childhood memory prompt significantly differed from responses by human subjects, in terms of style and detail, whereas AI inside jokes were more plausible. Hence, we decided to use the inside joke prompt because it is theoretically demanding but also empirically fair.

After selecting the inside-joke prompt, stimuli were generated from both humans and ChatGPT 4.0 to develop a complete stimulus set. First, we asked ChatGPT 4.0 to produce 10 inside jokes, in order to convince humans that they were written by humans, using the following prompt:

“Your task here is to convince other people that you are a human and not an AI. To do that, we would like you to answer a question about an inside joke you have with your friends. An inside joke is something humorous that only a small group of people would understand, because they know the context. As you answer, keep in mind things or ways people say things that AIs wouldn’t. There is no minimum or maximum word count. Please take your time and do your best — there are no right or wrong answers.”

Question: Can you tell us an inside joke you have with your friends and what it means to you?

We then gave the same task to humans through a Qualtrics survey and asked them to provide an example of an inside joke from their own lives (see Appendix B). We also asked our participants to describe their strategy for convincing others of their human authorship. We analyzed these descriptions to identify common strategies. To do this, we entered all 51 subject responses to ChatGPT 4.0 and instructed the LLM to identify different strategies used by the human respondents. This analysis identified six distinct strategies:

- (1) **Authenticity:** Using a sincere, genuine manner of expression when writing, recounting the reality naturally, rather than deliberately crafted or optimized
- (2) **Avoiding AI “tells”:** Using deliberate stylistic choices to avoid features commonly associated with AI-generated text, such as overly formal language, generic phrasing, or rigid structure
- (3) **Conversation & informal tone:** Using a casual, conversational, or spoken-like tone, such as slang, asides, and rhetorical questions, when composing the response
- (4) **Embracing imperfection:** Using intentional or uncorrected imperfections, such as typos, informal grammar, abrupt phrasing, or uneven structure, to reflect natural human writing

- (5) **Emotional appeal:** Using expressions of emotion, humor, or affective language to make the response feel more palpable and relatable
- (6) **Unique and personal content:** Using specific personal experiences, idiosyncratic details, or unique perspectives to elevate the creativity of the responses

Subsequently, we asked the LLM to categorize each self-reported strategy using a binary (0/1) system for frequency analysis. The two most endorsed strategies are appeal to authenticity (38%) and reliance on unique personal content (22%). Given their majority, we prioritized responses in these categories when constructing the final stimulus set for Study 1. We randomly selected five responses from each of these two categories and the ten AI-generated responses produced by ChatGPT 4.0, for a final set of twenty stimuli. We anonymized all responses. Importantly, no grammatical errors were corrected at this stage; we only introduced minor formatting changes to make the texts appropriate for subsequent testing.

Study 1

Participants

Participants ($n = 101$; 36 Male, 63 Female, 2 Non-binary/Other) were recruited online through Prolific and voluntarily completed this study. All participants were fluent English speakers and writers. Among the participants, none were under 18 or above 85 years old, 4.95% were between 18-24, 25.74% were between 25-34, 27.72% were between 35-44, 21.78% were between 45-54, 6.93% were between 55-64, 9.90% were between 65-74, and 2.97% were between 75-84. Furthermore, 7.92% of the participants identified as Asian, 69.31% as European, 3.96% as Hispanic/Latino, 0.99% as Pacific Islander, 4.95% as Mixed/Multiple Ethnicity, 7.92% as Other (with the majority being African American), and 4.95% preferred not to disclose. In terms of highest education level, 11.88% earned a secondary/high school diploma, 22.77% had some college/university experience (no degree), 35.64% earned a Bachelor's degree, 19.80% earned a Master's degree, 3.96% earned a Doctorate/PhD, 4.95% reported Other, and 0.99% preferred not to disclose.

Procedures

Participants were given the definition of an inside joke and told that their task was to identify whether each of a series of inside joke explanations was written by another person or by an LLM. Participants read each of the 20 jokes in random order. After each joke, participants indicated whether they believed the text was generated by a human or by AI using a 7-point Likert-type scale ranging from -3 (definitely AI-generated) to +3 (definitely human-generated), with 0 indicating uncertainty. Immediately following each assessment or rating, we asked participants to provide a brief written justification explaining the features of the descriptions that led them to believe they were written by a human or not. We encourage participants to take their time and respond thoughtfully.

Data Analytics Plan

The primary objective of Study 1 was to assess whether participants could reliably distinguish between human-written and AI-generated responses, and whether participants were more accurate when relying on certain features.

By asking participants to rate each response on a 7-point Likert scale from "definitely AI-generated" (-3) to "definitely human-written" (+3), we obtained two independent measures. First, ratings were averaged separately for human and AI stimuli, yielding a single mean score per condition for each participant. Second, we obtained an average score overall for each participant across the two conditions. Then, we recoded the ratings into binary accuracy values based on the stimuli's true source (human vs AI-generated). Correct directional judgments were coded as 1 and incorrect judgments as 0 ("uncertain" or "0" was counted as incorrect). This approach ensured a within-subject comparison while minimizing inter-individual differences in response bias.

Furthermore, we categorized the participants' justifications or indications of strategies using an identical binary (0/1) system to indicate presence or absence. Chat GPT 4.0 initially assigned binary values to the justifications. We screened for categorizing errors by the LLM afterward, and based on the self-reports, we extended the strategy categories to eight:

- (1) **Surface errors:** Participants use spelling, grammar, punctuation, typos, inconsistent capitalization, or sloppy mechanics as evidence that the text is human or AI
- (2) **Stylistic voice:** Participants comment on the sound or feel of the writing — formality, robotic tone, awkwardness, vibe, and unnaturalness
- (3) **Logical plausibility:** Judging whether the scenario itself could realistically occur in the world
- (4) **Personal schema matching:** Participants relate the story to their own experiences, people they know, or behaviors they have seen
- (5) **Structural features:** Judgments about length, organization, story arc, paragraphing, repetition, formatting, or structural cues often associated with AI
- (6) **Emotional cues:** Participants evaluate tone, warmth, offensiveness, compassion, coldness, or emotional resonance
- (7) **Humor diagnostics:** Evaluating whether the joke works — funniness, comedic timing, dullness, punchline logic
- (8) **Meta-inference:** The participant uses their beliefs about how humans vs AI behave, rather than the text itself, to judge authorship

Each response could receive multiple strategy codes. Specifically, the AI facilitated the coding based on contextual interpretation rather than keyword matching, ensuring that strategies were identified according to their functional role in participants' reasoning. The resulting dataset consisted of binary strategy indicators paired with an accuracy outcome for each judgment.

To assess the predictive value of strategy use on identification accuracy, we conducted a multivariate logistic regression model with accuracy as the dependent variable and all strategy categories entered simultaneously as independent predictors. Logistic regression was selected because it allows the evaluation of multiple categorical predictors while modeling a binary outcome variable. We interpreted the regression coefficients in terms of log-odds and converted them to odds ratios to allow direct evaluation of effect magnitude. Statistical significance of individual predictors was assessed using Wald tests, with α set at .05.

Finally, we evaluated the model performance using classification accuracy, receiver operating characteristic curves (ROC), and area under the curve values (AUC). We used these metrics to assess the extent to which the collective use of strategies improved the prediction of correct identification beyond chance.

Study 2

Participants

A new sample ($n = 100$; 36 Male, 61 Female, 2 Non-binary/Other, 1 prefer not to say) was recruited using the same platform and eligibility criteria as in Study 1. No participant took part in both studies. All participants were fluent English speakers and provided informed consent prior to participation. Among the participants, none were under 18 or above 85 years old, 6.00% were 18-24, 28.00% were 25-34, 29.00% were 35-44, 19.00% were 45-54, 8.00% were 55-64, 7.00% were 65-74, and 3.00% were 75-84. Moreover, 9.00% of the participants were Asian, 66.00% were European, 5.00% were Hispanic/Latino, 2.00% were Pacific Islander, 6.00% were Mixed/Multiple Ethnicity, 1.00% Middle East/North African, 7.00% as Other (the majority were Caucasian/African American), and 4.00% preferred not to say. Finally, in terms of education, 12.00% earned a secondary/high school diploma, 24.00% had some college/university experience (no degree), 38.00% earned a Bachelor's degree, 18.00% earned a Master's degree, 4.00% earned a Doctorate/PhD, 3.00% reported Other, and 1.00% preferred not to disclose.

Materials

For this study, we also administered our investigation via Prolific and closely mirrored the structure used in Study 1. The critical difference lay in the stimulus set. Although we informed participants that they would evaluate responses from both humans and AI, all 10 stimuli in this study were generated by ChatGPT 4.0. In Study 1, we inadvertently duplicated two AI-generated stimuli. In Study 2, we replaced these stimuli with two new responses to the inside-joke prompt. The content of the AI-generated responses remained the same as that of the original AI-generated stimuli used in Study 1, except for stimuli 3 and 6, which were replaced with entirely distinct inside jokes due to the replication.

We modified the stimuli by instructing ChatGPT 4.0 to intentionally incorporate surface-level errors that human evaluators in Study 1 reported relying on most when making authorship judgments.

“Your job is to convince other humans that you are a human. Last time, you provided us with some jokes that you think will fool humans. Now, we have identified that humans use the surface errors of the text (e.g., spelling, grammar, punctuation, typos, inconsistent capitalization, or sloppy mechanics errors) as a predictor to distinguish AI responses from humans. Thus, can you modify these ten responses that you gave me before by applying some errors that match the human predictors we identified?”

The revised responses included spelling mistakes, grammatical slips, inconsistent capitalization, punctuation irregularities, and other minor mechanical deviations. Importantly, we preserved the narrative content and the underlying structure of each response, such that differences between original and modified versions were limited to surface-level linguistic features.

Procedures

The procedure for Study 2 is similar to that of Study 1. After obtaining informed consent, we explained to participants that they would evaluate short written responses and judge whether each was generated by a human or an AI. Then, we presented the definition of an inside joke and the rating scale prior to the evaluation task.

We presented all ten stimuli in randomized order to each participant and asked them to rate the likely source of the stimuli using the same 7-point Likert-type scale ranging from –3 (definitely AI-generated) to +3 (definitely human-generated), with 0 indicating uncertainty. Following each rating, we required the participants to provide a brief written justification explaining their judgment.

Before the end of the study, we did not inform the participants that all stimuli in this study were AI-generated, nor did we inform them that these responses had been intentionally modified. This information was included in the debriefing section, which explained the purpose of the study and the nature of the stimulus manipulation.

Data Analytics Plan

Unlike Study 1, which focused on strategy effectiveness, Study 2 focused on performance under a manipulated condition based on the justification strategy that yielded the most consistent accuracy. Moreover, because all stimuli in Study 2 were AI-generated, analyses focused on comparing participants’ ratings and accuracy scores against the Study 1 AI-condition accuracy, rather than on within-study source discrimination.

First, we calculated the mean rating across all Study 2 stimuli. Second, we recoded the ratings into binary accuracy values based on the true source (AI-generated). Correct identification of a response as AI-generated was coded as 1, and incorrect identification as 0 (“uncertain” or “0” was coded as incorrect). Accuracy scores were averaged across all stimuli for each participant.

To determine whether participants performed below chance in Study 2, we conducted a one-sample t-test comparing Study 2 accuracy against 50% chance performance. To evaluate whether the manipulated AI responses shifted participants’ judgments relative to baseline AI-generated responses, we compared participant-level mean ratings and accuracy scores from Study 2 with the corresponding AI-condition scores from Study 1 using independent-samples t-tests. Because Study 1 and Study 2 used independent participant samples, this approach treated the two studies as separate groups. Collectively, this dual approach allowed us to test whether the stimuli with surface errors purposely embedded altered participants’ accuracy relative to chance and their judgments relative to baseline AI evaluations.

Results and Discussion

Study 1

We were interested in whether humans could tell whether the responses to the inside joke prompts were produced by a human or AI. To do this, we examined participants' ratings of each inside joke (the rating from the 7-point scale. As seen in Figure 1, participants rated human-created inside jokes as being more likely to be written by a human ($M = 0.254$, $SD = 1.05$) and rated AI-generated jokes as more likely to have been written by an AI ($M = -0.438$, $SD = 1.09$), $t(100) = 4.34$, $p < .001$, *Cohen's d* = 0.43.

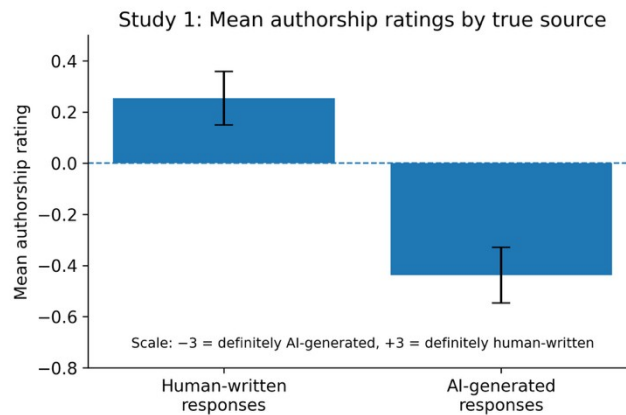


Figure 1. Mean authorship ratings for human-written and AI-generated inside joke responses in Study 1. Participants rated human-written responses as significantly more human-authored than AI-generated responses. Although the average ratings were relatively close to the scale midpoint, they differed significantly by the responses' true source, $t(100) = 4.34$, $p < .001$, $d = 0.43$.

*Note: Error bars represent standard errors.

To assess participants' overall accuracy, we recoded each rating as correct if the subject gave a positive rating to any human-generated response or a negative rating to any AI-generated response, and incorrect if the subject gave a negative rating to any human-generated response or a positive rating to any AI-generated response. Ratings of 0 were considered incorrect, as this neutrality suggests that the subject cannot make a clear distinction. Participants performed only slightly above chance at discriminating human from AI jokes ($M = 51.1\%$), but this difference was not statistically significant, $t(100) = 0.581$, $p = .562$, $d = 0.06$ (Figure 2). Participants more accurately identified human-written jokes as human ($M = 52.0\%$, $SD = 0.230$) than they identified AI-generated jokes as AI ($M = 50.0\%$, $SD = 0.258$).

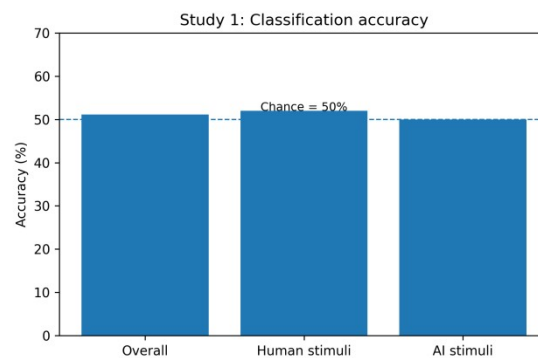


Figure 2. Classification accuracy in Study 1 compared with chance performance. Overall classification accuracy was 51.1%, which did not differ significantly from the 50% chance level. Participants were slightly more accurate at identifying human-written responses as human (52.0%) than at identifying AI-generated responses as AI (50.0%).

Next, we examined the strategies participants reported using to discriminate between human and AI-generated responses. We coded each participant's justification for the presence or absence of eight strategies consistently reported by participants (see above). To examine whether some strategies were more useful than others for

discriminating authorship, we fitted a multivariate logistic regression model with all strategy categories entered simultaneously as predictors of accuracy.

The overall model was statistically significant, $\chi^2(8) = 66.745$, $p < .001$, indicating that these strategies collectively predicted identification accuracy. However, several strategies emerged as especially good predictors of accuracy. According to Table 1 and Figure 3, when participants noticed surface-level errors (i.e., misspellings, ungrammatical sentences), they were significantly better at discriminating human- from AI-generated content (OR = 2.57, 95% CI [1.85, 3.56]). Likewise, participants were better when they attended to the structural features of the jokes (i.e., judgments about length, organization; OR = 2.21, 95% CI [1.44, 3.40]), relied on meta-inference about AI behavior (using personal beliefs about how humans vs AI behave; OR = 2.30, 95% CI [1.12, 4.75]), or focused on the emotional cues (evaluating tone, feeling, and/or emotional resonance the response transmits; OR = 1.96, 95% CI [1.07, 3.57]). In contrast, strategies involving stylistic voice, personal schema matching, humor diagnostics, and logical plausibility did not significantly predict accuracy when controlling for other strategies (Table 1 and Figure 3).

Table 1. Logistic regression predicting accuracy from strategy use. When all eight strategies were compared simultaneously, surface errors, structural features, meta-inference, and emotional cues were significant predictors of judgment accuracy. Surface errors produced the largest estimated effect, increasing the odds of an accurate judgment by approximately 2.57 times, with $p < .001$.

Term	Estimate	Std. Error	z-value	p-value	Odds Ratio	95% CI (Odds Ratio)
Intercept	0.1696	0.0705	2.407	0.0161	1.185	(1.032, 1.36)
Cues	0.6703	0.3067	2.185	0.0289	1.955	(1.072, 3.566)
Diagnostics	-0.3431	0.1920	-1.787	0.0739	0.710	(0.487, 1.034)
Errors	0.9427	0.1674	5.6320	1.7845e-8	2.567	(1.849, 3.564)
Structure	0.7940	0.2188	3.630	2.8391e-4	2.212	(1.441, 3.397)
Inference	0.8337	0.3695	2.2560	0.0241	2.302	(1.116, 4.749)
Plausibility	0.0723	0.1649	0.4390	0.6609	1.075	(0.778, 1.485)
Schema	-0.3794	0.3122	-1.215	0.2242	0.684	(0.371, 1.262)
Voice	-0.0835	0.1246	-0.6700	0.5028	0.920	(0.720, 1.174)

*Note: Cues = Emotional cues; Diagnostics = Humor Diagnostics; Errors = Surface errors; Structure = Structural features; Inference = Meta-inference; Plausibility = Logical plausibility; Schema = Personal schema matching; Voice = Stylistic voice

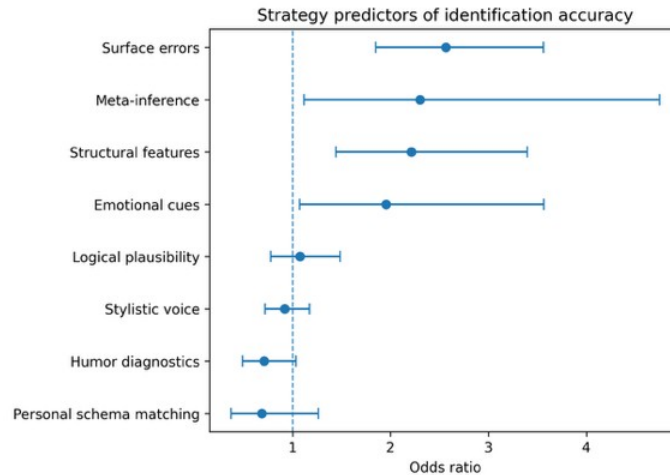


Figure 3. Odds ratios from the logistic regression predicting judgment accuracy from reported strategy use. Attention surface errors, meta-inference, structural features, and emotional cues were associated with significantly greater odds of an accurate authorship judgment. Contrarily, the confidence intervals for logical plausibility, stylistic voice, humor diagnostics, and personal schema matching crossed an odds ratio of 1, indicating no significant predictive effect.

*Note: Error bars represent 95% confidence intervals. The vertical dashed line indicates OR = 1, meaning no predictive effect.

To assess whether the eight strategy predictors were excessively collinear, variance inflation factors (VIF) were calculated for the predictor matrix used in the logistic regression model. As seen in Table 2, all VIF values were close to 1 and well below the commonly used threshold of 5, indicating that multicollinearity was not a concern. Therefore, all eight strategy predictors were retained as separate variables in the model.

Table 2. Variance inflation factors across eight strategies. The VIF ranged from 1.01 to 1.02 across all eight strategy predictors. These values indicate negligible multicollinearity and support retaining the strategies as separate predictors in the regression model.

Strategy	VIF
Emotional cues	1.01
Humor diagnostics	1.01
Surface errors	1.02
Structural features	1.01
Meta-inference	1.01
Logical plausibility	1.02
Personal schema matching	1.01
Stylistic voice	1.02

Finally, we calculated the receiver ROC and AUC to evaluate the overall predictive performance of the logistic regression model. The model produced an AUC of 0.602, indicating that the strategy predictors collectively discriminated accurate from inaccurate judgments above chance, although the model's collective predictive strength was modest (Figure 4).

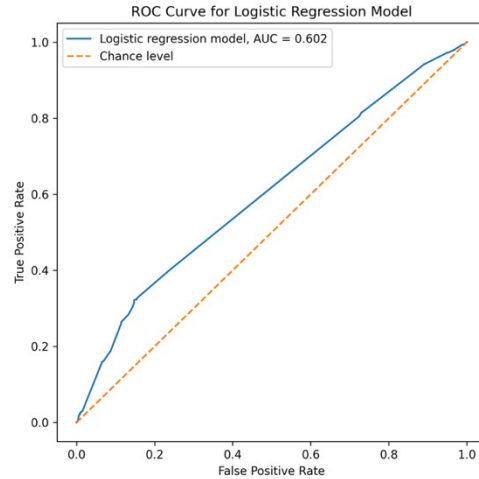


Figure 4. ROC curve for the logistic regression model predicting judgment accuracy from reported strategy use. The logistic regression model produced an AUC of 0.602, indicating that the combined strategy predictors distinguished accurate from inaccurate judgments better than chance, although the model’s overall discriminative performance was modest.

For Study 1, we investigated how accurately individuals can distinguish between human- and AI-generated responses in a context designed to highlight uniquely human experiences — inside jokes. Though participants nominally performed above chance, the difference was not statistically significant (Figure 2). Interestingly, participants were more successful at identifying human-authored responses than AI-generated ones (Figure 2).

One possible explanation is that participants demonstrated a slight response bias toward labeling responses as human-authored. Because human communication is the default form of interaction in everyday life, participants may have been more inclined to assume that ambiguous responses were written by humans rather than AI. This subtle, though significant, difference may also be explained by participants’ greater familiarity with human communication patterns due to lifelong exposure to human-authored language, which may make minor cues of human writing speak to their instincts and gut more readily and be slightly easier to recognize than indicators of AI-generated text.

Indeed, analysis of participants’ reported strategies not only supports this explanation but also provides further insight into the mechanisms underlying these judgments. Several strategies did, indeed, matter, including attention to surface errors, structural features of the response, emotion cues, and meta-inference about how AI computers were likely to respond (Table 1). In contrast, strategies unique to human experience, such as humor interpretation, personal schema matching (which we expected, given that participants reported having used them in stimulus design), did not reliably predict accuracy. In this ironic dependence on simple, heuristic, superficial cues, therefore, we saw something noteworthy—that if human judgments are ambiguous, people may default to readily available cues rather than to more painstaking considerations about meaning or even about experience authenticity, in keeping with cognitive load theory.²⁷

Among the strategies, the predictive power of surface error is particularly strong. Participants interpreted (e.g.) grammatical errors, spelling mistakes, and informal slips as signs of human authorship. This insight implies our current schema for human and AI writing: as humans naturally make capitalization, subject-verb and spelling errors in spontaneous communication, human answers are viewed as “imperfect,” but AI answers try to minimize such errors, and “perfectly” follow the grammar rules. This schema was probably the template for the judging heuristic. Moreover, grammatical and spelling errors are highly salient and, importantly, strongly ambiguous features, so that participants can easily and confidently make judgments about them in contrast to, say, humor or coherence. But signaling asymmetry complicates the confidence: their presence strongly signals human authorship, but their absence does not. Yet, even so, because they are informative when present, they retained predictive power.

Structural features also significantly predicted classification accuracy, demonstrating participants' reliance on easily observable aspects of text when judging authorship and schemas for AI-generated text. Organization, coherence, and formatting are highly salient and cognitively low-effort to evaluate, so they are just as useful as heuristic cues under low-certainty conditions. Obvious, "perfect" schema for AI-generated text hypothesized earlier, participants may have perceived a good, logical answer as too rigid and perfect, especially in the trivial work like the one at hand, as opposed to formal essay writing. This way, structural features capture genuine, albeit subtle, differences between human- and AI-generated text. The greater variability of human language across contexts allows structural cues to serve as imperfect but partially valid indicators of authorship, sufficient to elevate accuracy slightly above chance.

Together, these findings suggest that easily observable imperfections and writing structure are conducive to distinguishing human and AI authorship due to their high signal-to-noise ratio compared to deeper cues that tend to be more subjective (e.g., humor, meaning, experience). Though insignificant, the reliance on such superficial cues was sufficient to garner judgment accuracy above chance.

Besides the seemingly superficial strategies, other important findings include the meta-inference and emotional cues (Table 1). Participants who relied on the former often justified their judgments by reasoning about what LLMs "would" or "would not" produce. For example, some responses were classified based on beliefs that LLMs would not produce any offensive content and, thus, any inside joke including potentially controversial information would be human. To a certain extent, this finding validates the speculated schema of AI — that they are "perfect" and rigid — implying that our understanding of the current LLMs is somewhat accurate. Thus, when judging authorship, relying on stereotypes about AI writing rather than solely evaluating the appearance or the content of the responses could also aid discrimination. Similarly, emotional cues also significantly predicted classification accuracy, although their effect was weaker than that of the three aforementioned strategies. This may be because, although LLMs can generate emotionally expressive language, such expressions often appear more standardized or formulaic, allowing subtle differences in perceived emotional authenticity to serve as cues for judgment.

Nevertheless, despite their statistical significance, it is worth speculating why meta-inference and emotional cues were weaker predictors than surface errors and structural features, albeit rooted deeper in the context of the response (Table 1). Judgments based on a further inferred step, in this case, postulating AI behavior and emotional tones, may also be influenced by subjective interpretation. Such inductions require participants to consciously or subconsciously interpret and construct meaning, which is highly susceptible to personal experiences and emotions, deviating from the objective indicators of authorship. They could have projected their personal knowledge beyond what the text allows, potentially overanalyzing and overinterpreting, which often leads to greater hesitation. On the other hand, despite superficial, surface errors, structural features are objective cues that are less vulnerable to cultural, personal, or social differences. As a result, while meta-inference and emotional cues may contribute to perceptions of "humanness," their predictive power remains weaker compared to more salient, consistent, and objective features.

Together, these findings suggest that humans may not detect AI-generated content by identifying authentic markers of human experience. By contrast, they rely on surface-level linguistic cues and assumptions about AI behavior — heuristics that are reasonably strong predictors of accuracy in this study. However, because these cues are not inherently tied to authorship and instead reflect superficial features of the text, they may be unstable, in that they can be easily manipulated without altering the underlying source or content.

Consequently, we wondered whether manipulating these surface-level cues would impair individuals' ability to accurately classify authorship. If such cues serve as the basis for judgment, then altering them should reduce discrimination accuracy. Observing such a decrease would further increase the confidence that individuals rely on these heuristics rather than on stable, contextual indicators of authorship. Thus, we designed Study 2, which directly tested whether altering surface-level writing features could hinder human detection of AI-generated text.

Study 2

For this study, we examined whether intentionally inserting surface errors, the strongest predictor of accuracy in Study 1, would undermine participants' ability to discriminate AI-generated from human-written responses. To examine whether the manipulation shifted perceived authorship, we first calculated the average rating across all 10

AI-generated responses ($M = 0.161$, $SD = 1.10$). As displayed in Figure 5, participants rated the manipulated AI responses as more human-authored than the unmodified AI responses in the Study 1 condition, as indicated by an independent-samples t-test, $t(199) = 3.88$, $p < .001$, $d = .55$.

Similarly, we calculated participants' overall accuracy by first recoding each rating as correct if the participant gave a negative rating to the AI-generated responses and incorrect if the participant selected a positive rating to the 10 responses; ratings of 0 were also considered incorrect. As seen in Figure 6, participants' accuracy in Study 2 was significantly below chance ($M = 39.2\%$; $SD = 0.235$), as confirmed by a one-sample t-test against 50% chance performance, $t(99) = -4.59$, $p < .001$, $d = -0.46$. Participants' accuracy in Study 2 was lower than their accuracy for AI-generated responses in Study 1, as indicated by an independent-samples test, $t(199) = -3.10$, $p = .002$, $d = -0.44$.

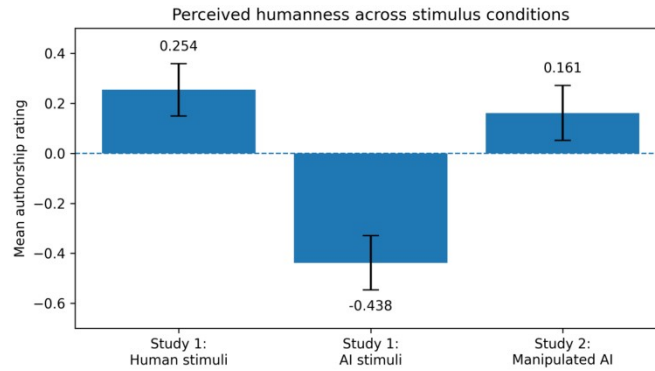


Figure 5. Mean authorship ratings across Study 1 and Study 2 stimuli. Participants rated the manipulated AI responses in Study 2 as significantly more human-authored than the unmodified AI responses in Study 1, $t(199) = 3.88$, $p < .001$, $d = 0.55$. The mean rating for manipulated AI responses also shifted to the human-authored side of the rating scale.

*Note: Error bars represent standard errors. Positive ratings indicate judgments toward human authorship, whereas negative ratings indicate judgments toward AI authorship.

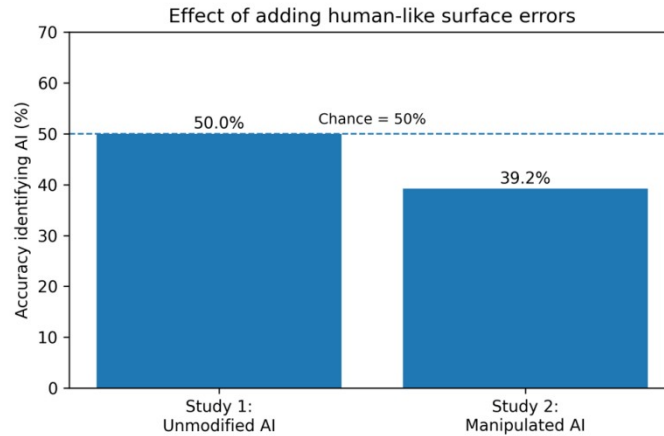


Figure 6. Accuracy for unmodified AI responses in Study 1 compared with manipulated AI responses in Study 2. Accuracy declined from chance level for unmodified AI responses in Study 1 to 39.2% for manipulated AI responses in Study 2. Study 2 accuracy was significantly below chance and significantly lower than accuracy for the unmodified AI responses, showing that the addition of surface errors increased misclassification of AI text as human-written.

In Study 2, we examined whether the surface-level errors, the best predictors from Study 1, truly play a causal role in human judgments of authorship. To test this possibility, AI-generated responses were deliberately modified to include minor spelling mistakes and grammatical irregularities that resemble those commonly produced by human writers. If participants heavily relied on surface heuristics when making authorship judgments, increasing such imperfection should weaken their ability to recognize AI-generated responses.

These results very much corroborated this prediction: Participants' accuracy dropped dramatically below chance and frequently mistook AI-generated responses for human-written texts (Figure 6). These results further support our belief that mistakes are signs of "humanness" to us, and that readers mostly depend on this "shallow" linguistic information when identifying a text as authored by a human being. Yet, this result also indicates that the cues participants were relying upon in Study 1 were not valid cues of authorship but indicators that can be easily deceived. Crucially, the experiential authenticity of the AI responses was not manipulated in this study; the content underlying the responses was the same as in Study 1. Yet, these linguistic mistakes served to markedly disadvantage participants' ability to identify the authorship, suggesting a potential heuristic hierarchy in authorship judgments.

Conclusion

Altogether, across our two studies, we investigated the features humans are using to determine whether they are reading a response generated by a human or an AI. Our results of Studies 1 and 2 indicate that human judgments about authorship involve less than detecting genuine indicators of human experience and more about surface-level traits of writing. While they might turn out to be right, they are very subject to manipulation, giving a coherent picture: people can tell humans and AI apart, but not for the reasons we would expect, not in ways that remain true if those surface-level signals are deliberately removed.

A notable strength of this study is its ecological validity. Rather than relying on decontextualized text snippets or artificial prompts, we asked participants to evaluate responses to a prompt that is inherently grounded in real social experience, reinforced by a preliminary phase of stimulus development. By selecting a stimulus that demands personal experience — exactly what LLMs lack — we gave participants their best possible chance of both proving their "humanness" and detecting AI authorship. The fact that they nonetheless fell back on susceptible cues and low accuracy makes the finding even more striking. Furthermore, by extending the work across two studies — one observational, one experimental — we were able to move beyond mere description of strategy use toward a causal account of how those strategies shape detection accuracy.

That said, there are several limitations of the current work that deserve acknowledgment and point to potential directions for future research. First, by prioritizing ecological validity, we inevitably traded experimental control. We did not fully control the responses in either study. The instructions offered considerable flexibility for interpretation and execution, which manifested in responses of various forms, lengths, and contents. All features of the responses, regardless of human or AI, were maintained. Even in Study 2, the surface errors that we asked ChatGPT 4.0 to purposely commit were determined by the LLM itself. This form of experimental manipulation lacked standardization. The type, frequency, and salience of the features were not systematically controlled. Therefore, though causality was implied, this limits the causal specificity; the spectrum of surface errors influenced human perception of authorship, but it is unclear which imperfection specifically. Future research could further clarify the causality by manipulating the features or the nature of the stimuli.

Another noteworthy limitation is that the strategy categories derived from participants' self-reports were developed iteratively and may not fully reflect the evaluative heuristics participants used. They are, therefore, somewhat arbitrary and reductionistic, as, realistically, cognitive processes do not map cleanly onto one strategic taxonomy. Most notably, some responses were ambiguous, qualifying for multiple categories. This made categorization even more difficult, and decisions had to be made based on our subjective judgments. Future work can build categories with a larger number of raters to provide more confidence in the categorization and validity, and highlight nuances that we may well have missed in developing our coding strategies independently.

Finally, this research was done entirely in written English, which limits the generalizability of the results in important ways. The diagnostic power of surface errors as a cue for human authorship may be specific to languages and writing systems where spelling errors and informal cues are meaningful. For example, in a logographic writing system such as Mandarin Chinese, the concept of a 'spelling error' has no obvious equivalent because characters are

not put together out of phonemic units in the same way. In this sense, the predictive strategies and the schema associating imperfection with human authenticity may therefore be constrained to certain cultures with phonemic writing systems.

We also focused exclusively on written text, when human communication and expression extend beyond words on a screen. Spoken communication introduces an entirely different set of signals — tonal variation, pacing, hesitation, nonverbal cues, and the spontaneous disfluencies of real-time speech — that may be far more difficult for AI to replicate convincingly, or conversely, may reveal new vulnerabilities in human detection that the written channel does not expose. Extending research across a wider range of linguistic systems and modes of communication would be a valuable next step in clarifying the remaining gap between human cognition and AI.

Finally, we are aware that the choice of the inside joke prompt concerns the social position of the evaluator in relation to the inside joke stimuli. Although inside jokes were selected because they are theoretically grounded in shared experience and idiosyncratic social memory, the participants judging these responses were not members of the social groups from which the jokes originated. So, they could evaluate only the perceived genuineness of the writing, and were guided towards surface cues; deeper experiential cues were not fully available to participants in this study. Future work might consider using responses from friends, classmates, and others who know the author to reduce this gap in information.

As AI-generated text becomes ubiquitous in academic, professional, and personal life, from student papers to job applications to letters of condolence, being able to reliably discern human or machine authorship has real stakes. Errors in this assessment may force people to lose faith in authentic expression, have faith in artificially authored text, or incorrectly attribute intent in ways that structure decisions, relationships, and institutions.

These results have practical implications for the design of AI detection technologies and training for media competence. If people tend to depend on surface biases that are trivially broken by instructing an AI to add typos, then detection-based approaches to authorship verification might only have a short shelf life. And as LLMs become increasingly able to reproduce the full stylistic texture of human writing, including the styling mistakes, the authenticity vs. artificialness gap might close to a point where it becomes functionally difficult to tell them apart. This matters in any context where authorship verification is pertinent: academic assessments, journalism, lawyerly communication, personal testimony evaluation, and so on. Thus, the implication is not just that better detection tools are needed, but that institutions may have to rethink what it means to verify authenticity in the first place, going from a surface-level inspection to detecting the conditions in which a text was generated, rather than the text itself.

This brings us back to our intuition we set out with—that some expressions feel meaningful because they arise from a human mind. A birthday letter, a eulogy, or an inside joke seems to have value in what was actually said, but in the prediction that it was lived, felt, and remembered. But the present findings subtly complicate this intuition. Across both studies, participants did not reliably identify authentic human expression through its depth of experience or contextual richness. Instead, they relied on something much simpler: imperfection. Errors, irregularities, and departures from “polished” language were seen as tokens of being human, not because they represent meaning, but because they signal human limitations.

To this end, “to err is human” is not just a motto but a sort of schema that causes people to recognize that as authentic. But Study 2 shows the fallibility of that schema. And in that trial, if AI output was purposefully sprayed with the very same mark of imperfection, participants automatically saw it as human. What seems to be a sign of genuine experience is, in reality, a superficial feature, potentially easy to imitate. If error is what we think of as human, then simulating error is simulating humanness.

It does not just imply that AI can ape human writing, but that our criteria for identifying the human might be more superficial than we think. As artificial systems become better and better at replicating not just fluent language but also the “faults” we associate with human thought, the divide between human and machine might not disappear, but it might become perceptually indistinct. What is challenged then is not only our ability to recognize AI but also the stability of the idea that to err is human.

AI Acknowledgement

During the preparation of this work, the author used ChatGPT 4.0 in order to facilitate the coding of written justifications based on contextual interpretation. After using this tool, the author reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgements

I would like to express my deepest gratitude to Professor Jason P. Mitchell for his generous guidance throughout the study, which not only expanded my understanding of the human mind but also strengthened my intellectual curiosity and passion for psychology. I am also deeply grateful to my parents for their unwavering belief in me and for encouraging me to continually challenge myself.

References

1. Cao, Z.; Zhang, X.; Zeng, D. D. Large Language Models: Technology, Intelligence, and Thought. *Front. Eng. Manag.* **2025**, *12* (3), 710–715. <https://doi.org/10.1007/s42524-025-5004-3>.
2. Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models, 2023. <https://doi.org/10.48550/ARXIV.2307.06435>.
3. Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. arXiv 2020. <https://doi.org/10.48550/ARXIV.2005.14165>.
4. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; Zhang, Y. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv 2023. <https://doi.org/10.48550/ARXIV.2303.12712>.
5. Bhuyan, S. S.; Sateesh, V.; Mukul, N.; Galvankar, A.; Mahmood, A.; Nauman, M.; Rai, A.; Bordoloi, K.; Basu, U.; Samuel, J. Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *J. Med. Syst.* **2025**, *49* (1), 10. <https://doi.org/10.1007/s10916-024-02136-1>.
6. Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; Tseng, V. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *PLOS Digit. Health* **2023**, *2* (2), e0000198. <https://doi.org/10.1371/journal.pdig.0000198>.
7. Lee, P.; Bubeck, S.; Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **2023**, *388* (13), 1233–1239. <https://doi.org/10.1056/NEJMSr2214184>.
8. Rabbani, S. A.; El-Tanani, M.; Sharma, S.; Rabbani, S. S.; El-Tanani, Y.; Kumar, R.; Saini, M. Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions. *BioMedInformatics* **2025**, *5* (3), 37. <https://doi.org/10.3390/biomedinformatics5030037>.
9. Ahmed, S. A. The Impact of Artificial Intelligence on Customer Experience in E-Commerce Platforms. *J. Fundam. Comp. Res.* **2025**, *11* (1).
10. Giang, N. T. P.; Hung, D. M.; Huong, T. T.; Quyen, H. T.; Thom, N. T.; Tran, T. B.; Tan, T. D. Researching User Experience with Artificial Intelligence Application for Customer Care Services on E-Commerce Platform. *J. Ekon. Perusah.* **2025**, *31* (2), 149–174. <https://doi.org/10.46806/jep.v31i2.1319>.
11. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>.
12. Zhai, X. ChatGPT for Next Generation Science Learning. *SSRN Electron. J.* **2023**. <https://doi.org/10.2139/ssrn.4331313>.
13. Holmes, W.; Bialik, M.; Fadel, C. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*; Center for Curriculum Redesign: Boston, MA, 2019.
14. Luckin, R.; Holmes, W. *Intelligence Unleashed: An Argument for AI in Education*. **2016**.

15. Li, J.; Chen, J.; Ren, R.; Cheng, X.; Zhao, X.; Nie, J.-Y.; Wen, J.-R. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2024; pp 10 879–10 899. <https://doi.org/10.18653/v1/2024.acl-long.586>.
16. Ravichander, A.; Ghela, S.; Wadden, D.; Choi, Y. HALoGEN: Fantastic LLM Hallucinations and Where to Find Them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1)*; 2025. <https://doi.org/10.18653/v1/2025.acl-long.71>.
17. Turing, A. M. *Mind* **1950**, 59 (236), 433–460.
18. Sonkar, S.; Liu, N.; Chen, X.; Baraniuk, R. G. The Imitation Game for Educational AI, 2025. <https://doi.org/10.48550/ARXIV.2502.15127>.
19. Warwick, K.; Shah, H. Passing the Turing Test Does Not Mean the End of Humanity. *Cogn. Comput.* **2016**, 8 (3) 409–419. <https://doi.org/10.1007/s12559-015-9372-6>.
20. Erol, G.; Ergen, A.; Gülşen Erol, B.; Kaya Ergen, Ş.; Bora, T. S.; Çölgeçen, A. D.; Araz, B.; Şahin, C.; Bostancı, G.; Kılıç, İ.; Macit, Z. B.; Sevgi, U. T.; Güngör, A. Can We Trust Academic AI Detective? Accuracy and Limitations of AI-Output Detectors. *Acta Neurochir. (Wien)* **2025**, 167 (1), 214. <https://doi.org/10.1007/s00701-025-06622-4>.
21. Clark, E.; August, T.; Serrano, S.; Haduong, N.; Gururangan, S.; Smith, N. A. 2021. <https://doi.org/10.48550/ARXIV.2107.00061>.
22. Ippolito, D.; Duckworth, D.; Callison-Burch, C.; Eck, D. Automatic Detection of Generated Text Is Easiest When Humans Are Fooled, 2019. <https://doi.org/10.48550/ARXIV.1911.00650>.
23. Jones, C. R.; Bergen, B. K. Large Language Models Pass the Turing Test, 2025. <https://doi.org/10.48550/ARXIV.2503.23674>.
24. Parshakov, P.; Naidenova, I.; Paklina, S.; Matkin, N.; Nesseler, C. Users Favor LLM-Generated Content—Until They Know It’s AI, 2025. <https://doi.org/10.48550/ARXIV.2503.16458>.
25. Merriam-Webster. Inside Joke. *Merriam-Webster.com Dictionary*.
26. Norrick, N. R. *Conversational Joking: Humor in Everyday Talk*; Indiana University Press: Bloomington, 1993.
27. Sweller, J. Cognitive Load during Problem Solving: Effects on Learning. *Cogn. Sci.* **1988**, 12 (2), 257–285.

Appendices

Appendix A

We were interested in finding a prompt that AI would have difficulty responding to. In order to generate a prompt, we asked the AI to self-identify some via the following prompt:

Hello! How are you? I am conducting a study to investigate whether humans can convince another participant that they are human, with the participant believing that they have an equal chance to be talking to an AI or a human. We're thinking of prompts for this conversation. What are some prompts that you think as an AI would struggle with answering, while humans probably don't?

All AI (ChatGPT 4.0, Gemini 2.5, and DeepSeek) indicated that questions regarding inside jokes would be challenging for them to generate a compelling and convincing response to, as they can “simulate answers to many of these, but humans will provide answers rooted in experience, physicality, and fallibility” (from the AI).

Appendix B

Survey Extract (Prompted):

Please read the instructions below carefully.

Note: when we refer to AIs, we are referring to generative AI models like ChatGPT, DeepSeek, and Gemini

In the age of technology, AIs have become more prevalent in contemporary society. Many people are concerned about whether we can identify AI-generated and authentic human responses for various reasons, such as academic honesty. At the same time, due to the rapid advancement of AI, it is becoming harder to distinguish between those responses.

AIs are really good at convincing us that they are humans. That said, we wanted to see if we are good—or even better—at convincing humans that we are humans.

*Your task here is to convince other people that you **are** a human and **not** an AI. To do that, we would like you to answer*

a question about an inside joke you have with your friends. An inside joke is something humorous that only a small group of people would understand, because they know the context. As you are answering, keep in mind things or ways people say things that AIs wouldn't. There is no minimum or maximum word count. Please take your time and do your best—there are no right or wrong answers.

In this research, we are interested in understanding the inside joke and its context. Therefore, we really hope that you can take a few minutes to write a paragraph or two—that would really help us understand the full story of it through the details.

To provide ample time, you will not be able to move on to the next page in the first three minutes. Once the time has elapsed, a "next" button will appear on your screen.

Can you tell us an inside joke you have with your friends and what it means to you? (Please wait 3 minutes before continuing).

When you were trying to act like a human and not an AI, what were your strategies? (Please wait 60 seconds before continuing).

Survey Extract (Not prompted):

Please read the instructions below carefully.

We are interested in the kind of inside jokes that people have with their friends. An inside joke is something humorous that only a small group of people would understand, as they know its context. Therefore, we will ask you to answer a question regarding an inside joke between you and your friends.

We want you to spend a few minutes describing an inside joke of your choice. There is no minimum or maximum word count. Please take your time and do your best—there are no right or wrong answers.

In this research, we are interested in understanding the inside joke and its context. Therefore, we really hope that you can take a few minutes to write a paragraph or two—that would really help us understand the full story of it through the details.

To provide ample time, you will not be able to move on to the next page in the first three minutes. Once the time has elapsed, a "next" button will appear on your screen.

Can you tell us an inside joke you have with your friends and what it means to you? (Please wait 3 minutes before continuing).

Why did you pick this joke? How easy or hard was it to recall the details?

Author

Yijia (Claire) Xu is an International Baccalaureate student at the Western Academy of Beijing, interested in psychology, AI, and emerging technologies' social impact. Her research was published in the University of Tennessee's *Modern Psychological Studies*. She hopes to continue studying the intersection of science, technology, and society.