Proceedings of the 5th International Conference on Social Psychology and Humanity Studies
DOI: 10.54254/2753-7048/2027.38988

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

Yijia Xu

Western Academy of Beijing, Beijing, China
xyjclaire1018@gmail.com

Abstract. As large language models (LLMs) increasingly generate human-like text, it is becoming an increasingly pressing empirical and societal question: how can AI-generated text be distinguished from human-written text? Through two studies, this paper explored whether and how people judge AI authorship. In Study 1, participants (n = 101) judged the written responses to an inside joke prompt (the prompt chosen based on LLMs' response) and assessed each response on a 7-point scale of −3, definitely AI-written, to +3, 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 the conclusion of 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

1. 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 artificial intelligence (AI). Would that change how much you like the letter? Would you respond in a different way, or not reply at all? Would it feel as genuine a relationship with you as before? Even just imagining this situation has probably annoyed you; perhaps the deceit is only part of the reason, and the deeper part is the expectation that some messages should come from a human brain.

However, AI systems are getting better at copying human communication, from a birthday message to an election speech; thus, this intuition is becoming harder to rely on. Large language models (LLMs) are advanced AI systems built to understand, process and generate human-like text, and they learn from very large datasets with deep learning [1, 2]. Over the past decade, the text

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2027.38988

produced by LLMs has improved rapidly, and it is now fluent in grammar, coherent in context, and appealing and convincing in style [3, 4].

In many fields, this development has been clearly useful. Many hospitals and healthcare organizations use LLMs to write and manage medical records; thus, clinicians have less paperwork and more time for patient care [5-8]. In e-commerce, if an online customer asks a question about an item, the LLM can analyze the question and its context and then give a relevant, accurate and helpful answer [9]. In education, teachers can use LLMs to automate time-consuming tasks such as making quizzes, and this greatly raises efficiency [10, 11]. For students, LLMs can act as private tutors with personal guidance, instant feedback and customized lesson plans, so students can learn at their own pace. They also help learners with different learning abilities, and in this way educational content is open to more people [12, 13]. Therefore, LLMs have raised efficiency and made information easier to access in a great number of fields. If individuals or institutions want to increase efficiency, expand access and meet users' needs, the value of the output is in its function; as a result, who wrote the content does not matter much.

In other situations, however, who the author is does matter. A love letter gets its meaning from the vulnerability of the writer. A eulogy is meaningful due to a real relationship that memory, affection and loss have shaped. Even in education, if a student hands in an essay written by AI, academic integrity is lost, and so is the struggle that leads to real understanding, that is, the effort of turning a half-known idea into one's own. More broadly, this authorship concern is not an individual-level issue. In public fields such as journalism, political communication, professional writing and personal testimony, uncertainty about authorship can affect the credibility and accountability of sources, especially since LLMs still have difficulty in giving accurate information [14, 15]. If readers can no longer trust who is speaking, suspicion will become the default, and the shared ability of society to reason about what is true will be destroyed. More importantly, if AI-generated and human-authored texts cannot be told apart, such LLMs will be used even more in written tasks for the efficiencies mentioned. Therefore, both facts and an authentic human voice are at risk, and such a voice is the irreducible trace of one person with particular experiences at a certain moment of writing.

One of the main empirical questions in this field is now whether this trace can be noticed or not. Much scientific work on human-AI discrimination comes from the question put forward by Turing [16], that is, whether a machine can produce the same responses as a human. This framing has led to a large body of empirical work, such as some updated versions of the imitation game [17, 18], studies on chatbot detection [19], and more recent experiments on whether humans can reliably detect LLM text [20-23]. 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 [21]. 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.

Such decisions are worth studying for several reasons. First, schools often believe a teacher or a student can "tell" if a text is AI-generated. But unreliable cues may lead them to blame human writers by mistake or to miss real AI use. Second, this research can expose the biases people hold about human expression. It shows what features people link to human writing and to AI writing;

DOI: 10.54254/2753-7048/2027.38988

thus, social and political expectations about "humanness" in written texts become visible. A change in these expectations would also show how AI is changing the way people relate to language. Lastly, the results may help to develop AI-detection tools.

Thus, the "cues" people depend on in authorship judgments are an important concept; they form the basis of the idea of strategy. Here, the term strategy means the angle of evaluation or the technique a participant applies when deciding who wrote a text. A participant may, for example, notice grammatical imperfections, depend on emotional resonance, or draw on beliefs about the usual writing style of AI systems; other strategies also exist, and each one is defined later in this paper. This concept is used to describe the mental processes of people when they make such a decision. The categories were based on self-reports, and the written justifications went through a systematic coding procedure with the help of AI. The reported strategies, however, may not always agree with the cognitive processes that really produce a judgment, and this gap is kept in mind throughout this paper.

Overall, this paper will investigate whether humans can accurately tell AI-generated text from human-written text and, more importantly, which cognitive strategies are behind these judgments. Prior work has shown that detection accuracy is often modest; however, the field has not yet systematically examined what evaluative cues people use, or whether these cues really indicate authorship or are only correlated with it on the surface. To this end, participants in this study were asked to evaluate written responses to an inside joke prompt. After each judgment, they described their reasoning, and these descriptions made it possible to identify specific strategies and test their predictive value. A follow-up study then directly manipulated the self-reported strategy with the highest predictive power, so that its role in authorship detection could be tested causally. In this way, this study tries to better understand the cognitive mechanisms behind a judgment that people make more and more often in everyday life, that is, how readers decide whether a text has the mark of human authorship.

2. Methods

2.1. 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. There is no standardized stimulus set for assessing "humanness" in open-ended language production; therefore, all stimuli were developed step by step in several phases to reflect both human intuitions and AI strategies about what makes a text convincingly human.

First, in the stimulus development process, three widely used LLMs (ChatGPT, Gemini, and DeepSeek) were told about this research on how well humans can detect AI authorship. Since the aim was to give humans the best chance, the LLMs were asked in which scenarios it would be most difficult for them to "pass" as humans. All three LLMs gave the same two scenarios, that is, childhood memories with sensory experiences and inside jokes (see Appendix A). The meaning of the first one is more intuitive; the second one needs a more specific definition and clarification. In this paper, inside jokes are defined in the same way as they were defined for participants: humorous references that only a small group of people can understand, 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 [24, 25]. 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

DOI: 10.54254/2753-7048/2027.38988

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, it was 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, the inside joke prompt was chosen because it is theoretically demanding but also empirically fair.

Once the inside-joke prompt was chosen, a full set of stimuli was built with responses from humans and from ChatGPT 4.0. First, ChatGPT 4.0 received the prompt below and was asked to write 10 inside jokes that would make people believe a human wrote them:

"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?

Then, human respondents received the same task in a Qualtrics survey, and each of them gave an example of an inside joke from their own lives (see Appendix B). The respondents also explained the strategy they used to convince others that a human had written their answer. These explanations were then analyzed for strategies that came up often. For this step, the 51 responses were all put into ChatGPT 4.0, and the LLM was instructed to list the different strategies the human respondents had applied. The analysis gave six different 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

Then, each self-reported strategy was coded by the LLM in a binary (0/1) way so that frequencies could be counted. Appeal to authenticity (38%) and reliance on unique personal content (22%) were the two strategies chosen most often. Since these two made up the majority, priority was given to responses in these categories when building the final stimulus set for Study 1. Five responses were randomly picked from each of the two categories, and the ten AI-generated responses from ChatGPT 4.0 were added, so the final set had twenty stimuli. All responses were made anonymous. Importantly, no grammatical errors were corrected at this stage; only minor formatting changes were introduced to make the texts appropriate for subsequent testing.

DOI: 10.54254/2753-7048/2027.38988

3. Study 1

3.1. 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.

3.2. 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. Right after each rating, participants had to write a short justification, saying which features of the description made them think a human wrote it or not. Participants were encouraged to take their time and respond thoughtfully.

3.3. Data analytics plan

The main aim of Study 1 was to find out whether participants could reliably tell human-written responses from AI-generated ones, and whether they were more accurate if they relied on certain features.

Participants rated each response on a 7-point Likert scale from "definitely AI-generated" (−3) to "definitely human-written" (+3), and two independent measures were obtained from these ratings. First, the ratings of human stimuli and AI stimuli were averaged separately; thus, each participant had one mean score in each condition. Second, each participant's overall average score for the two conditions together was also calculated. Next, the ratings were recoded into binary accuracy values according to the true source of each stimulus (human vs AI-generated). A judgment in the correct direction got a code of 1, and an incorrect one got 0 ("uncertain" or "0" was counted as incorrect). This kept the comparison within subjects and reduced individual differences in response bias.

In addition, participants' justifications or indications of strategies were coded with the same binary (0/1) system for presence or absence. The first coding of the justifications was the work of ChatGPT 4.0. Categorizing errors by the LLM were screened for afterward, and based on the self-reports, the strategy categories were extended 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

DOI: 10.54254/2753-7048/2027.38988

(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

Any response could have more than one strategy code. The AI helped with the coding by interpreting the context instead of matching keywords; in this way, each strategy was identified by its function in the reasoning of participants. In the final dataset, each judgment had binary strategy indicators and one accuracy outcome.

In order to see how well strategy use predicted identification accuracy, a multivariate logistic regression model was run; accuracy was the dependent variable, and all strategy categories were entered at the same time as independent predictors. Logistic regression was selected because it allows the evaluation of multiple categorical predictors while modeling a binary outcome variable. The regression coefficients were read as log-odds, and converting them to odds ratios (ORs) made it possible to evaluate the size of each effect directly. Statistical significance of individual predictors was assessed using Wald tests, with α set at .05.

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

4. Study 2

4.1. 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.

DOI: 10.54254/2753-7048/2027.38988

4.2. Materials

This study was also administered via Prolific and closely mirrored the structure used in Study 1. The critical difference lay in the stimulus set. Although participants were informed 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, two AI-generated stimuli were inadvertently duplicated. In Study 2, these stimuli were replaced 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.

The stimuli were modified 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?"

In the revised responses, there were spelling mistakes, grammatical slips, inconsistent capitalization and irregular punctuation, as well as other small mechanical errors. Importantly, the narrative content and the underlying structure of each response were preserved, such that differences between original and modified versions were limited to surface-level linguistic features.

4.3. Procedures

The procedure for Study 2 is similar to that of Study 1. After informed consent was obtained, participants were told that they would evaluate short written responses and judge whether each was generated by a human or an AI. Next, they read the definition of an inside joke and saw the rating scale before the evaluation began.

Each participant saw the ten stimuli in a random order and judged where each one probably came from on the same 7-point Likert-type scale, from −3 (definitely AI-generated) to +3 (definitely human-generated), where 0 meant uncertainty. After every rating, participants also had to give a short written reason for their judgment.

Participants were not told until the end of the study that every stimulus was AI-generated, and they were also not told that the responses had been changed on purpose. The debriefing gave them both facts and explained the aim of the study and the kind of stimulus manipulation.

4.4. Data analytics plan

Study 1 was about how well the strategies worked, and Study 2 was about performance in a condition manipulated through the justification strategy with the most consistent accuracy. In addition, all Study 2 stimuli were AI-generated; thus, the analyses compared participants' ratings and accuracy scores with the AI-condition accuracy of Study 1, and no within-study source discrimination was tested.

First, the mean rating over all Study 2 stimuli was computed. Second, every rating became a binary accuracy value, judged against the true source (AI-generated). Labelling a response as AI-

DOI: 10.54254/2753-7048/2027.38988

generated was correct and got 1, any other answer got 0, and "uncertain" or "0" was also counted as wrong. For each participant, the accuracy scores of all stimuli were then averaged.

To see if participants in Study 2 were below chance, Study 2 accuracy was compared with 50% chance performance in a one-sample t-test. Independent-samples t-tests were then used to see if the manipulated AI responses moved participants' judgments away from the baseline AI-generated responses; mean ratings and accuracy scores of each participant in Study 2 were compared with the matching AI-condition scores in Study 1. Study 1 and Study 2 had independent participant samples; thus, this approach handled them as separate groups. In this way, the two analyses tested if the stimuli with deliberately inserted surface errors changed accuracy compared with chance, and changed judgments compared with the baseline AI evaluations.

5. Results and discussion

5.1. Study 1

The first question was whether humans could tell whether the responses to the inside joke prompts were produced by a human or AI. To do this, participants' ratings of each inside joke (the rating from the 7-point scale) were examined. 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. The average ratings were relatively close to the midpoint of the scale; however, there was a significant difference between them according to the true source of the responses. In Figure 1, error bars represent standard errors.

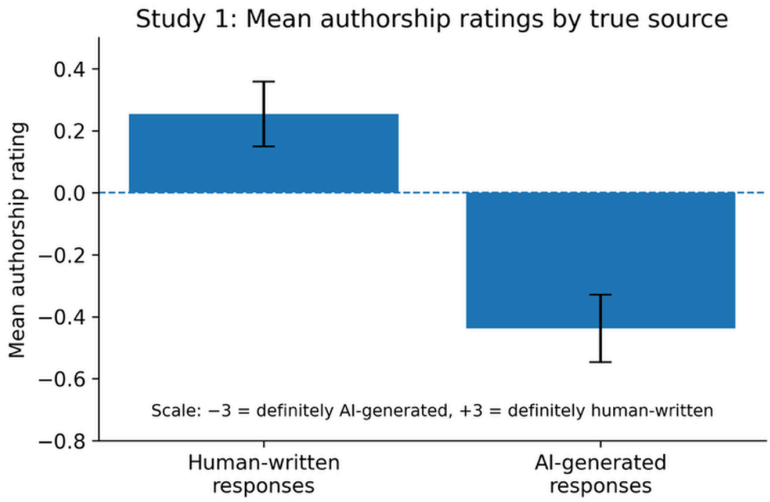


Figure 1. Mean authorship ratings for human-written and AI-generated inside joke responses in Study 1

To assess participants' overall accuracy, each rating was recoded 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

DOI: 10.54254/2753-7048/2027.38988

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 ($M = 50.0\%$, $SD = 0.258$).

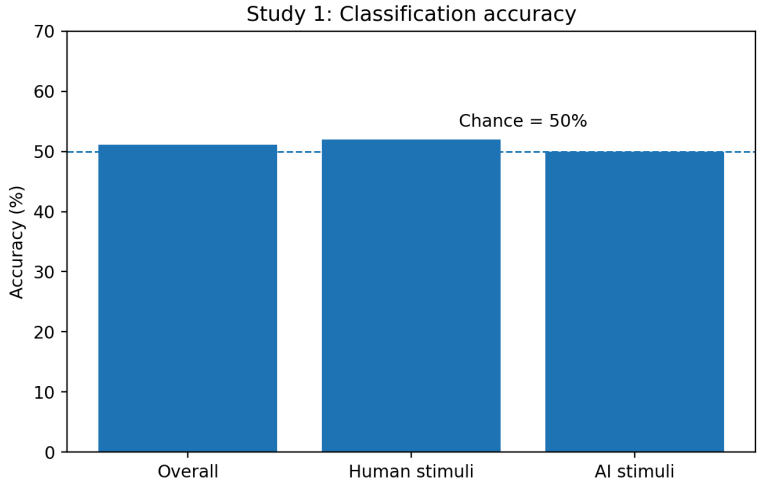


Figure 2. Classification accuracy in Study 1 compared with chance performance

Next, the strategies participants reported using to discriminate between human and AI-generated responses were examined. Each participant's justification was coded 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, a multivariate logistic regression model was fitted 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). Surface errors produced the largest estimated effect, increasing the odds of an accurate judgment by approximately 2.57 times. In Table 1, 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. In Figure 3, error bars represent 95% confidence intervals, and the vertical dashed line indicates OR = 1, meaning no predictive effect.

DOI: 10.54254/2753-7048/2027.38988

Table 1. Logistic regression predicting accuracy from strategy use

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)

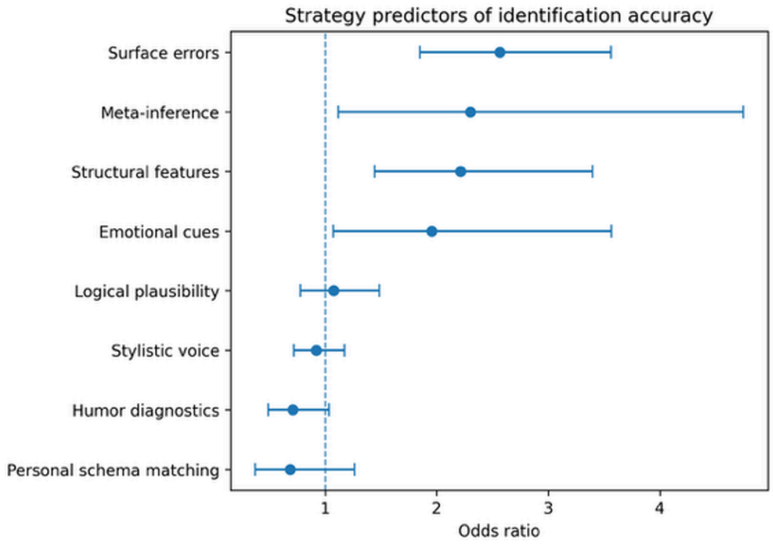


Figure 3. Odds ratios from the logistic regression predicting judgment accuracy from reported strategy use

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. The VIF ranged from 1.01 to 1.02 across all eight strategy predictors. Therefore, all eight strategy predictors were retained as separate variables in the model.

Table 2. Variance inflation factors across eight strategies

Strategy	VIF
Emotional cues	1.01
Humor diagnostics	1.01

DOI: 10.54254/2753-7048/2027.38988

Table 2. (continued)

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

Lastly, the ROC and the AUC were computed to judge how well the logistic regression model predicted as a whole. The AUC was 0.602; thus, the strategy predictors together separated accurate and inaccurate judgments better than chance, but the predictive strength of the model as a whole was modest (Figure 4).

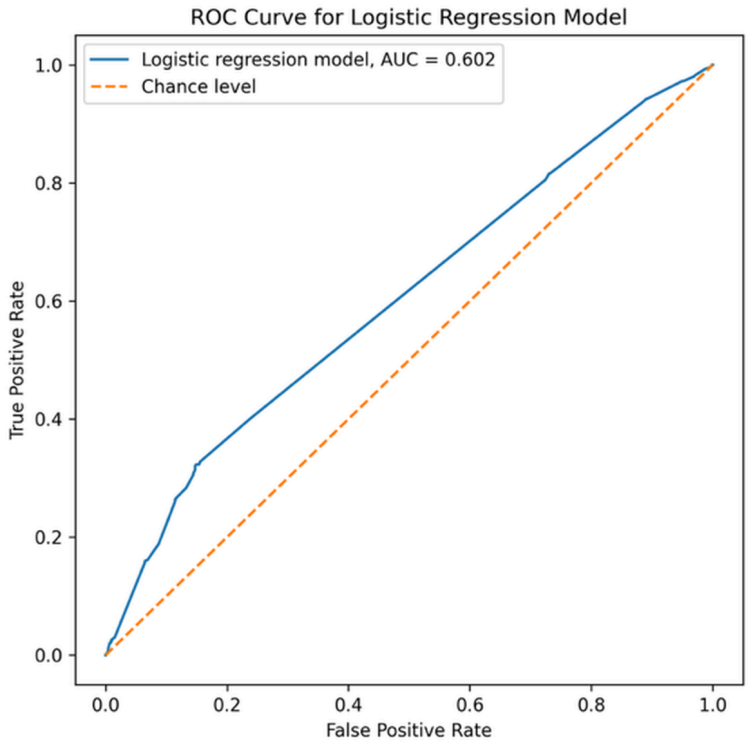


Figure 4. ROC curve for the logistic regression model predicting judgment accuracy from reported strategy use

Study 1 tested how well people tell human and AI-generated responses apart in a context, inside jokes, chosen to bring out uniquely human experience. Accuracy was nominally higher than chance, but statistical tests did not find this gap significant (Figure 2). Human-authored responses were also identified more successfully than AI-generated ones (Figure 2).

A slight response bias may explain this, that is, a tendency of participants to label responses as human-authored. In everyday life, human communication is the default way of interaction; therefore, participants may have been more ready to assume an ambiguous response came from a human and not from AI. This subtle but significant difference may also come from greater

DOI: 10.54254/2753-7048/2027.38988

familiarity with human communication patterns, since participants have heard and read human-authored language all their lives; thus, small cues of human writing may reach their instinct and gut more quickly and be a little easier to recognize than the signs of AI-generated text.

The strategies that participants reported are in line with this explanation, and they also show more about how these judgments work. 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 was expected, given that participants reported having used them in stimulus design), did not reliably predict accuracy. Therefore, this ironic reliance on simple, heuristic and superficial cues is worth noting; it shows that if human judgments are ambiguous, people may turn to cues that are easy to find instead of thinking more carefully about meaning or even about the authenticity of experience, and this fits cognitive load theory [26].

Of all the strategies, surface error has especially strong predictive power. For example, participants took grammatical errors, spelling mistakes and informal slips as signs of human authorship. This shows the schema people now hold for human and AI writing; humans naturally make capitalization, subject-verb and spelling errors in spontaneous communication, so human answers are seen as "imperfect," but AI answers try to have as few such errors as possible and "perfectly" follow the grammar rules. It is likely that the judging heuristic was based on this schema. In addition, grammatical and spelling errors are very prominent and also unambiguous; thus, compared with features like humor or coherence, participants can judge them easily and with confidence. 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 were also a significant predictor of classification accuracy; this shows that participants judged authorship by easily seen parts of a text and by their schemas for AI-generated text. Little cognitive effort is needed to notice organization, coherence and formatting and to judge them, so they work just as well as heuristic cues when people are not certain. Earlier, a schema was hypothesized in which AI-generated text is obviously "perfect"; with it, participants may have seen a good and logical answer as too rigid and perfect, in particular in a trivial task like this one, unlike formal essay writing. Structural features thus pick up real but subtle differences between human- and AI-generated text. Human language varies more between contexts; therefore, structural cues can act as imperfect yet partly valid signs of authorship, enough to raise accuracy a little above chance.

The above results therefore point to easily observed imperfections and writing structure as helpful for separating human and AI authorship, since their signal-to-noise ratio is high compared with deeper and often more subjective cues (e.g., humor, meaning, experience). Depending on these superficial cues was enough to lift judgment accuracy above chance, though not significantly.

In addition to the strategies that look superficial, meta-inference and emotional cues were also important results (Table 1). When participants used meta-inference, they often explained a judgment by thinking about what LLMs "would" or "would not" write. Some of them, for example, believed that LLMs never produce offensive content, so they took any inside joke with possibly controversial information as human. This result gives some support to the speculated schema of AI as "perfect" and rigid, and it implies that people's present understanding of LLMs is somewhat accurate. Therefore, in authorship judgments, stereotypes about AI writing may also help discrimination, and not only the evaluation of how the responses look or what they say. Emotional cues were a significant predictor of classification accuracy as well, but their effect was smaller than the effects of

DOI: 10.54254/2753-7048/2027.38988

the three strategies above. One reason may be that LLMs can write emotionally expressive language, but this language often looks more standardized or formulaic; thus, small differences in how authentic the emotion feels can work as cues for judgment.

Meta-inference and emotional cues were statistically significant and go deeper into the context of a response; however, surface errors and structural features still predicted accuracy better (Table 1), and the reason for this is open to speculation. These judgments need one more step of inference, here about how AI behaves and about emotional tone, so subjective interpretation may also affect them. In such inductions, participants interpret and build meaning, consciously or not, and this is easily shaped by personal experience and emotion; thus, it moves away from objective signs of authorship. Participants may have brought in personal knowledge beyond the text and overanalyzed and overinterpreted it, and this often causes more hesitation. By contrast, surface errors and structural features are superficial, but they are objective cues and depend less on cultural, personal or social differences. Therefore, meta-inference and emotional cues may add to perceptions of "humanness," but their predictive power is still weaker than that of more salient, consistent and objective features.

In summary, the results suggest that people may not find AI-generated content through authentic markers of human experience. Instead, they depend on surface-level linguistic cues and on assumptions about how AI behaves, and in this study these heuristics predicted accuracy fairly well. However, these cues reflect superficial text features and have no built-in link to authorship; therefore, they may be unstable, as they can be easily manipulated while the source and the content stay the same.

Therefore, the next question was whether a change in these surface-level cues would hurt people's ability to classify authorship correctly. If judgments rest on such cues, changing the cues should lower discrimination accuracy. A drop of this kind would give more confidence that people depend on these heuristics and not on stable, contextual indicators of authorship. To this end, Study 2 directly tested if changing surface-level writing features could make human detection of AI-generated text harder.

5.2. Study 2

In this study, surface errors, the strongest predictor of accuracy in Study 1, were inserted on purpose to test if they would weaken participants' ability to tell AI-generated responses from human-written ones. First, the mean rating of all 10 AI-generated responses was computed ($M = 0.161$, $SD = 1.10$) to see if the manipulation changed perceived authorship. An independent-samples t-test showed that the manipulated AI responses were rated as more human-authored than the unmodified AI responses of the Study 1 condition, $t(199) = 3.88$, $p < .001$, $d = 0.55$ (Figure 5). The mean rating for manipulated AI responses also shifted to the human-authored side of the rating scale. In Figure 5, error bars represent standard errors. Positive ratings indicate judgments toward human authorship, whereas negative ratings indicate judgments toward AI authorship.

Similarly, participants' overall accuracy was calculated 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, $t(199) = -3.10$, $p = .002$, $d = -0.44$.

DOI: 10.54254/2753-7048/2027.38988

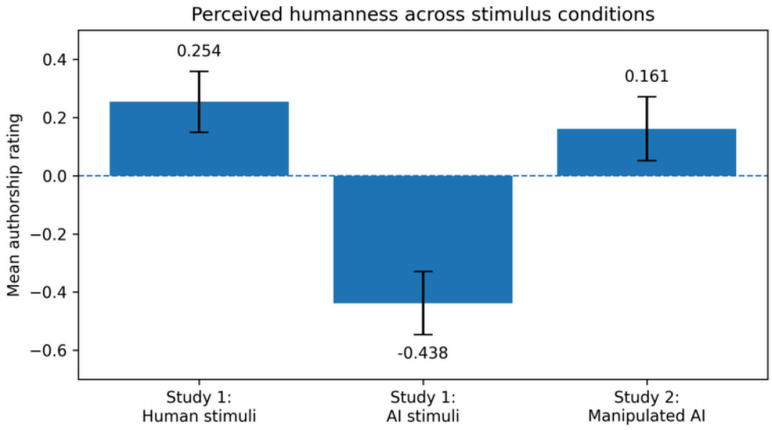


Figure 5. Mean authorship ratings across Study 1 and Study 2 stimuli

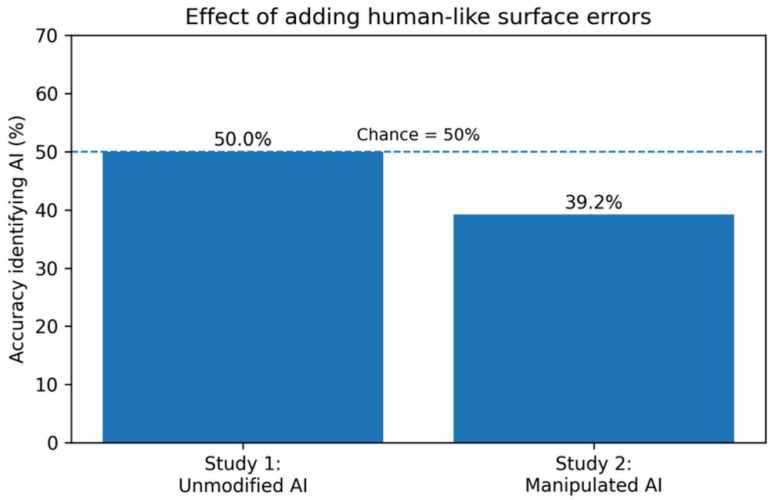


Figure 6. Accuracy for unmodified AI responses in Study 1 compared with manipulated AI responses in Study 2

Surface-level errors were the best predictors in Study 1, and Study 2 explored whether they really have a causal role in human judgments of authorship. To test this, AI-generated responses were changed on purpose so that they had small spelling mistakes and grammatical irregularities similar to those that human writers often make. If participants heavily relied on surface heuristics when making authorship judgments, increasing such imperfection should weaken their ability to recognize AI-generated responses.

The prediction was clearly confirmed by the results; accuracy fell far below chance, and participants often took AI-generated responses for human-written texts (Figure 6). This gives further support to the idea in this paper that people see mistakes as signs of "humanness" and that readers mostly rely on this "shallow" linguistic information to decide that a human being wrote a text. However, the result also shows that the cues used in Study 1 were not valid cues of authorship; they were indicators that are easy to deceive. This study did not change how authentic the experience in the AI responses was, and the content behind them was the same as in Study 1. Even so, the linguistic mistakes clearly weakened participants' ability to identify authorship; this points to a possible heuristic hierarchy in authorship judgments.

DOI: 10.54254/2753-7048/2027.38988

6. Conclusion

Altogether, the two studies investigated the features humans are using to determine whether they are reading a response generated by a human or an AI. The 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 one would expect, not in ways that remain true if those surface-level signals are deliberately removed.

One strength of this study is its ecological validity. Text snippets without context or artificial prompts were not used; instead, participants evaluated responses to a prompt based on real social experience, and a preliminary phase of stimulus development made this design stronger. The chosen stimulus needs personal experience, and this is exactly what LLMs do not have; thus, participants had the best chance both to prove their "humanness" and to detect AI authorship. Even so, they fell back on cues that can easily be manipulated and had low accuracy; the result is therefore even more surprising. In addition, there were two studies in this work, one observational and one experimental; therefore, this work could go further than only describing strategy use and give a causal account of how these strategies affect detection accuracy.

At the same time, this work has some limitations, and they also show possible directions for future research. First, since ecological validity was given priority, some experimental control was inevitably lost. In both studies, the responses were not fully controlled. The instructions left much room for interpretation and execution; as a result, the responses had different forms, lengths and contents. All the features of the responses were kept, no matter whether they came from humans or AI. Even in Study 2, ChatGPT 4.0 was asked to make surface errors on purpose, but the LLM itself decided which errors to make. Thus, this kind of experimental manipulation was not standardized. There was no systematic control of the type, frequency and salience of the features. Therefore, causality was implied, but the causal specificity is limited; the range of surface errors affected how humans perceived authorship, and it is still unclear which specific imperfection had this effect. Later studies could make the causal link clearer through manipulation of the features or the nature of the stimuli.

Another limitation is that the strategy categories were derived from participants' self-reports through an iterative process, and they may miss part of the evaluative heuristics that participants really applied. The categories are therefore somewhat arbitrary and reductive, as real cognitive processes do not fit neatly into one strategic taxonomy. In particular, some responses were ambiguous and fit several categories. This made categorization harder, and some decisions had to rest on the subjective judgment of the author. Later work can build the categories with more raters to give more confidence in the categorization and validity; it can also bring out nuances that may have been missed here, as the coding strategies were developed without other raters.

Lastly, the whole research used written English only; thus, how far the results generalize is limited in important ways. Surface errors may work as a diagnostic cue for human authorship only in languages and writing systems where spelling errors and informal cues carry meaning. Mandarin Chinese, for example, uses a logographic writing system, and a 'spelling error' has no clear counterpart there, since characters are not built from phonemic units in the same way. Therefore, the predictive strategies, and the schema that links imperfection to human authenticity, may hold only for some cultures that have phonemic writing systems.

DOI: 10.54254/2753-7048/2027.38988

In addition, this study only looked at written text, but human communication and expression go 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, it is recognized 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.

AI-generated text is appearing everywhere in academic, professional and personal life, from student papers to job applications to letters of condolence; therefore, reliably telling human authorship from machine authorship has real consequences. If this judgment is wrong, people may be forced to lose trust in authentic expression, to trust text written by machines, or to attribute intent wrongly in ways that shape decisions, relationships and institutions.

The above results also matter in practice, for the design of AI detection technologies and for media competence training. If people tend to rely on surface biases that can be broken simply by telling an AI to add typos, detection-based methods of authorship verification may only be useful for a short time. In addition, LLMs are getting better at copying the whole stylistic texture of human writing, even the style mistakes; thus, the gap between authenticity and artificialness may become so small that in practice it is hard to tell them apart. This is important in every situation where authorship verification is relevant, such as academic assessments, journalism, lawyerly communication, the evaluation of personal testimony, etc. For this reason, better detection tools are needed, but institutions may also have to rethink from the start what it means to verify authenticity; the focus may need to shift from a surface-level inspection of the text itself to the conditions in which the text was generated.

This leads back to the starting intuition, that is, some expressions feel meaningful since they come from a human mind. The value of a birthday letter, a eulogy or an inside joke seems to come from the belief that it was lived, felt and remembered, and not from what was actually said. However, the results of this study make this intuition a bit more complicated. In both studies, participants did not reliably recognize authentic human expression by its depth of experience or rich context. Instead, they used a much simpler cue, that is, imperfection. Errors, irregularities and departures from "polished" language were regarded as signs of being human; the reason is that they signal human limitations, and not that they represent meaning.

In this sense, "to err is human" is more than a motto; it works as a kind of schema that makes people regard error as authentic. But Study 2 has shown that this schema can fail. 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 people think of as human, then simulating error is simulating humanness.

DOI: 10.54254/2753-7048/2027.38988

It does not just imply that AI can ape human writing, but that the criteria people use for identifying the human might be more superficial than they think. Artificial systems are getting better and better at copying fluent language and also the "faults" that people link to human thought; as a result, the line between human and machine may still exist, but it may become hard to perceive. In the end, both the human ability to recognize AI and the stability of the idea that to err is human are challenged.

Acknowledgement

The author sincerely thanks Professor Jason P. Mitchell for the generous guidance given during the whole study; this guidance expanded the author's understanding of the human mind, and it also made the author more curious intellectually and more passionate about psychology. Many thanks also go to the author's parents for their constant belief and their encouragement to keep taking on new challenges.

Declaration of generative AI use

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.

References

[1] Cao, Z., Zhang, X., & Zeng, D. D. (2025). Large language models: Technology, intelligence, and thought. *Frontiers of Engineering Management, 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. (2023). A comprehensive overview of large language models (arXiv:2307.06435). arXiv. 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., . . . Amodei, D. (2020). Language models are few-shot learners (arXiv:2005.14165). arXiv. 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. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4 (arXiv:2303.12712). arXiv. 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. (2025). Generative artificial intelligence use in healthcare: Opportunities for clinical excellence and administrative efficiency. *Journal of Medical Systems, 49*(1), Article 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. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health, 2*(2), Article e0000198. https://doi.org/10.1371/journal.pdig.0000198

[7] Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine, 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. (2025). Generative artificial intelligence in healthcare: Applications, implementation challenges, and future directions. *BioMedInformatics, 5*(3), Article 37. https://doi.org/10.3390/biomedinformatics5030037

[9] Giang, N. T. P., Hung, D. M., Huong, T. T., Quyen, H. T., Thom, N. T., Tran, T. B., & Tan, T. D. (2025). Researching user experience with artificial intelligence application for customer care services on e-commerce platform. *Jurnal Ekonomi Perusahaan, 31*(2), 149–174. https://doi.org/10.46806/jep.v31i2.1319

DOI: 10.54254/2753-7048/2027.38988

[10] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

[11] Zhai, X. (2023). ChatGPT for next generation science learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4331313

[12] Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

[13] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

[14] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2024). 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)* (pp. 10879–10899). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.586

[15] Ravichander, A., Ghela, S., Wadden, D., & Choi, Y. (2025). HALoGEN: Fantastic LLM hallucinations and where to find them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1402–1425). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.71

[16] Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[17] Sonkar, S., Liu, N., Chen, X., & Baraniuk, R. G. (2025). The imitation game for educational AI (arXiv:2502.15127). *arXiv*. https://doi.org/10.48550/arXiv.2502.15127

[18] Warwick, K., & Shah, H. (2016). Passing the Turing test does not mean the end of humanity. *Cognitive Computation, 8*(3), 409–419. https://doi.org/10.1007/s12559-015-9372-6

[19] 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. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica, 167*(1), Article 214. https://doi.org/10.1007/s00701-025-06622-4

[20] Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text (arXiv:2107.00061). arXiv. https://doi.org/10.48550/arXiv.2107.00061

[21] Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled (arXiv:1911.00650). arXiv. https://doi.org/10.48550/arXiv.1911.00650

[22] Jones, C. R., & Bergen, B. K. (2025). Large language models pass the Turing test (arXiv:2503.23674). arXiv. https://doi.org/10.48550/arXiv.2503.23674

[23] Parshakov, P., Naidenova, I., Paklina, S., Matkin, N., & Nesseler, C. (2025). Users favor LLM-generated content—until they know it's AI (arXiv:2503.16458). arXiv. https://doi.org/10.48550/arXiv.2503.16458

[24] Merriam-Webster. (n.d.). Inside joke. In *Merriam-Webster.com dictionary*. https://www.merriam-webster.com/dictionary/inside%20joke

[25] Norrick, N. R. (1993). Conversational joking: Humor in everyday talk. Indiana University Press.

[26] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

DOI: 10.54254/2753-7048/2027.38988

Appendix

Appendix A

The aim was to find a prompt that AI would have difficulty responding to. In order to generate a prompt, the AI was asked 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):

DOI: 10.54254/2753-7048/2027.38988

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?