

# You Had Better Check the Facts: Reader Agency in the Identification of Machine-Generated Medical Fake News

Barbora Dankova, University of Warwick

## Abstract

During the COVID-19 pandemic, much fake news emerged in the medical field (Naeem *et al.*, 2020: 1). Nowadays, computers can generate text considered to be more trustworthy than text written by a person (Zellers *et al.*, 2019). This means that laypeople are able to produce disinformation; however, they may not understand the implications. This study revealed the most reliable clues as guidance to spot machine writing. While natural-language processing (NLP) research focuses on L1 speakers, studies in second language acquisition demonstrate that L1 and L2 speakers attend to different aspects of English (Scarcella, 1984; Tsang, 2017). In this study, social media users completed a Turing-test style quiz, guessed whether news excerpts were machine generated or human written (Saygin *et al.*, 2000) and identified errors that guided their decision. Quantitative analysis revealed that although both L1 and L2 speakers were equally able to defend themselves against machine-generated fake news, L2 participants were more sceptical, labelling more human-written texts as being machine generated. This is possibly due to concern about the stigma associated with being fooled by a machine due to lower language levels. However, factual errors and internal contradictions were the most reliable indicators of machine writing for both groups. This emphasises the importance of fact-checking when news articles prioritise exaggerated headlines, and NLP tools enable production of popular content in areas like medicine.

**Keywords:** Natural-language processing, detection of fake news, fact-checking of articles, machine-generated medical fake news, natural-language generation, manipulation

## Introduction

Fake news can be defined as misleading stories (Gelfert, 2018) intended to attract attention in order to secure monetary or political profit (Frank, 2015). While many people believe that mainstream sources would not publish false stories, even established media companies may find themselves broadcasting inaccurate information. Readers should therefore not blindly trust online news without fact-checking (Gatten, 2004). Research shows that young people consume news on social networks rather than traditional broadcast media (Allcott and Gentzkow, 2017), which enables quick access to information, but lends itself easily to the spread of fake news (Aldwairi and Alwahedi, 2018). As was seen in the recent public health crisis caused by the COVID-19 pandemic, it is important to recognise the rise of fake news in the medical field (Naeem *et al.*, 2020: 1).

In 1950, Alan Turing described an imitation game in which a machine must convince an interrogator that it is a human being through written responses to the interrogator's questions, and he believed that at the turn of the twenty-first century, computers would be able to fool humans at least 30 per cent of the time. It is beyond the scope of this paper to review the historical developments in natural-language processing (NLP), but research in the field has evolved considerably, and today's language models use the statistical distribution of words to predict the next word in a text based on those preceding it.

News articles generated by state-of-the-art language models have been rated as more trustworthy than human writing (Zellers *et al.*, 2019), which has given rise to a complex discussion around the ethics of artificial intelligence (Radford, Wu, Amodei *et al.*, 2019). Although automatic systems are able to detect machine-generated text, it is crucial for individuals to learn to spot these (Gehrmann *et al.*, 2019). Therefore, more needs to be done to understand both the affordances and risks of natural-language generation technology in the production of news.

While most research in automated content generation focuses on improving the models' ability to approximate human writing as closely as possible (Carlson, 2015), previous research in digital journalism has examined how readers perceive software-generated news content in comparison to that written by human journalists (Clerwall, 2014). Similarly, the present study aims to find the most reliable clues that lead human readers to identify machine-generated fake news. Self-selected participants filled out an online questionnaire consisting of a Turing-test style text-annotation task (Saygin *et al.*, 2000) in which participants were asked to classify medical news articles as being either 'machine generated' or 'human written' and to identify predetermined error categories in the text that guided their decision. And while there are certain linguistic errors that help identify machine writing, it should be noted that, as training sets grow larger, language models are likely to approximate human writing closer and closer, making grammatical errors a temporary issue.

This study examines the factors that contribute to the performance of human readers on the task and their unconscious strategies when approaching the task. Previous research has not studied [L2 English speakers](#) or examined the effect of English proficiency. Research in TESOL, second language acquisition and studies of university students' writing have shown that L1 and L2 English speakers tend to focus on different aspects of the language in production and comprehension. While [L1 speakers](#) focus on vocabulary (Burt, 1975) and overall style (Scarcella, 1984), L2 speakers tend to rely on grammatical rules (Tsang, 2017). This study examined whether this is the case in the detection of machine-generated fake news.

As part of a larger mixed-methods study, this paper uses quantitative analysis to compare L1 and advanced L2 English speakers' performances on the basis of the [F1 score](#) (Goutte and Gaussier, 2005), which was selected as a measure of how well they can defend themselves against machine-generated fake news. The F1 score is a measure used in statistical analysis of binary classification and is further explained in the methodology section of this paper. The errors that the research participants identified in the test excerpts were analysed in order to identify the clues that most reliably lead to the identification of machine writing in everyday life.

## Literature review

Fake news is often described as speculative and not based on evidence (Punjabi, 2017), with hyperbolic headlines intended to attract attention in order to secure monetary or political profit (Frank, 2015). Because of its novelty, fake news has been found to spread further and faster than true stories (Vosoughi *et al.*, 2018). According to Zakharov *et al.* (2019), many readers do not believe that mainstream news outlets would publish false stories. However, even established media companies may publish inaccurate information and should not be exempt from fact-checking (Gatten, 2004).

Research shows that in the UK and USA, young people increasingly use social media like Twitter and Facebook to consume news in place of traditional broadcasting organisations (Allcott and Gentzkow, 2017). This enables quick access to the news, but contributes to the spread of fake news (Aldwairi and Alwahedi,

2018). The medical field is particularly vulnerable to misinformation and, during the COVID-19 public health crisis, the World Health Organization declared that not only are we fighting a pandemic, at the same time, we are also 'fighting an infodemic' (Naeem *et al.*, 2020: 1) and urged health information professionals to help the general public detect false news.

Nowadays, products using text summarisation simplify academic articles (Yadav *et al.*, 2021), and automatic translation helps us overcome language barriers (Xaitqulov, 2021). Modern NLP techniques make our lives easier. However, they can also be misused to reword news briefings in order to fit a political agenda (Sharevski *et al.*, 2021). In recent studies, readers rated the style and trustworthiness of computer-generated propaganda as being more convincing than those of its human-written counterparts (Zellers *et al.*, 2019).

The algorithmic processes that convert data into narrative news texts with limited to no human intervention – such as those created by the companies Narrative Science or Statsheet Network – have been described as 'automated journalism' (Carlson, 2015), and Clerwall (2014) found that software-generated content produced by such tools was barely discernible from that written by human journalists.

Until recently, computers had to be trained for specific tasks. However, nowadays, transfer learning allows unsupervised systems to perform well on a variety of tasks without explicit programming (Radford, Wu, Child *et al.*, 2019). GPT-2 (Generative Pre-trained Transformer) is a publicly available large neural network model pre-trained on *WebText*, a dataset of over 8 million documents from the internet (Radford, Wu, Child *et al.*, 2019), which allows it to generate human-sounding text by predicting the next word in a passage based on all the preceding words. It is effective out of the box, but can be fine-tuned on texts that one wishes to emulate. The proliferation of open-source software and ready-to-go scripts that facilitate the fine-tuning and text-generation process, such as those created by Woolf (2019), have allowed laypeople with little to no coding ability and understanding of the societal implications to operate these on a mass scale. The research organisation OpenAI, which released GPT-2, drew attention to the issue of ethics in artificial intelligence by releasing the model gradually out of fear that it could be abused to generate misleading news articles (Radford, Wu, Amodei *et al.*, 2019). It is therefore imperative to instigate a conversation about disinformation threats created by machine learning and how to offset these (Zellers *et al.*, 2019).

Human judgements are frequently used to help language models emulate human writing (Resnik and Lin, 2010). However, more research needs to be done to help human readers identify machine-generated text. Ippolito *et al.* (2020) revealed the parameters that determine how the next word in a text is chosen and, therefore, how similar or dissimilar to the training data the output will be. On one side of the spectrum, the generated text is less diverse and closer to the original input text, as the model chooses words that commonly follow each other, making automatic statistical detection easier. This also produces fewer semantic errors, making the text more believable to humans. On the opposite side of the spectrum, the text is less predictable and uses uncommon words, which resembles human-written text to an automatic detector, but the frequent semantic errors reveal its machine-generated origin to human readers. While automatic discriminators rely on statistical distribution, humans check whether the presented evidence matches their model of the world (Zellers *et al.*, 2019). Therefore, humans must be kept in the loop to counter ethnic and gender biases encoded in training data, since the results produced by automatic systems can promote discriminatory decision making when accepted without questioning (Chang *et al.*, 2019).

Because of the global status of the English language, L2 speakers make up roughly two-thirds of English speakers in the twenty-first century (Pennycook, 2017). Most of the world's scientific research is in English, and the language is highly prominent in the media (Seidlhofer, 2004) and seen as valuable for education and

employment in today's world (Nunan, 2003). Many L2 speakers therefore need to be able to distinguish fake news in a foreign language. L1 and L2 speakers tend to focus on different features of the language, and previous research in the field has not taken this into account.

Studies that examined the syntactic complexity of the writing of L1 and L2 university students suggest that L1 speakers capture the readers' interest and help them identify the theme by using rhetorical devices, synonyms and lexical collocations (Scarcella, 1984). They can also adjust their language according to the task demands (Foster and Tavakoli, 2009), use more subordination and focus on fewer subtopics in more depth (Mancilla *et al.*, 2017), which they introduce using a variety of discourse markers (Ferris, 1994). In comparison, L2 writers have been shown to use less attention-getting devices and tend to over-specify the theme – failing to judge what their audience knows about the topic – and to use statements that downplay the importance of their exposition. L2 students also utilised clarifying devices governed by few syntactic and pragmatic constraints; they introduced more topics than L1 students in their papers, but more superficially (Scarcella, 1984), while using more explicit discourse markers (Ferris, 1994).

Additionally, Tsang (2017) found that in determining the countability of English nouns, Cantonese participants relied on grammar rules, whereas L1 English speakers gave more weight to semantics. Previous research on L2 English teachers shows similar tendencies. Like students, L2 English teachers paid attention to grammatical accuracy (Connor-Linton, 1995), word order and the organisation of essays (Shi, 2001), whereas L1 teachers focused on lexical errors, range of vocabulary (Burt, 1975), discourse context and overall quality of writing (Song and Caruso, 1996).

Based on previous research done in the fields of natural-language processing, digital journalism and TESOL, the following three research questions were identified:

1. Which errors are most reliable in helping readers identify machine-generated texts?
2. Do L1 and L2 English speakers achieve different levels of performance in detecting machine-generated texts?
3. Do L1 and L2 English speakers use different criteria in the identification of machine-generated text?

## Methodology

This paper presents the quantitative results of a larger interdisciplinary mixed-methods study carried out in order to answer the research questions posed above.


A Qualtrics survey<sup>[1]</sup> was used to administer an online Turing-test style text-annotation task (Saygin *et al.*, 2000) to participants who were presented with seven potentially machine-generated texts and three questions (Figure 1). They were asked to classify each prompt as being either machine generated or human written and to identify errors from a list of predefined options.


The seven prompts consisted of three machine-generated texts, human-written texts and one control text (Figure 2) in random order to minimise ordering effects on participants' perception (Schwarz, 2007). To ensure that participants were engaging with the task and not answering randomly, they were asked to classify the control prompt as *definitely machine generated* and all responses that failed this check were excluded from further analysis.

Previous research indicates that machine-generated texts frequently present problems with grammar, meaning that they include **morphosyntactic errors**, such as the use of incorrect verb forms or word order

(Stahlberg and Kumar, 2021). Machines frequently struggle with **entailment** where a text negates (i) its claims made within the excerpt (Saikh *et al.*, 2019), (ii) punctuation and (iii) formatting (Datta *et al.*, 2020). Much of machine-generated text can also be quite repetitive and predictable (Gehrmann *et al.*, 2019). Ippolito *et al.* (2020) also mentioned semantic errors and incorrect factual information (identified as NONSENSE in the quiz). Based on common errors identified in previous research, the predefined error categories can be seen in Figure 1. Operational definitions and examples of each error type were provided in the quiz instructions, and a short summary of each option was available to participants under every text.

Q4 Please answer the following three questions in relation to this text:

 A new study finds that a drug that is commonly used for treatment of myeloid leukemia can also help those with bone marrow transplantation. This is a procedure in which bone marrow transplants are made from the bone marrow of the recipient. The findings are important because some people with bone marrow transplantation have painful bone marrow loss, which doctors call bone marrow failure. The study is published in PLoS Biology. The researchers say that although the pattern of bone marrow transplantation does vary, the risk of bone marrow transplantation failure is relatively low. In fact, experts previously estimated that around 0.4% of adults in the United States would receive a bone marrow transplant in 2020, according to the National Institutes of Health (NIH). In the new study, the researchers analyzed bone marrow transplantation data from the Bone Marrow Transplantation Study, which took place between 2006 and 2012. They found that in addition to bone marrow transplants, bone marrow transplants were common in the transplantation of bone marrow from transplant recipients with bile duct cancer. They also found that transplant recipients with leukemia had a higher risk of bone marrow transplantation failure than those without the condition.

 Do you think that this text is human-written or machine-generated?

Definitely human-written


Possibly human-written


Possibly machine-generated

Definitely machine-generated

---

Q5 Imagine that you come across this text online. Please select the statement which best describes how you would interact with it.


  I would believe the information.


  I would want to read more.

I would not believe the information.

---

Q6 Please select one or more options to describe the text above. Include any other clues in the last option.

  NO ERROR: This is perfectly correct English.

  GRAMMAR: This text contains grammatical errors.

REPETITION: This text is too repetitive or predictable

ENTAILMENT: This text contradicts itself.

FORMATTING: This text contains incorrect use of punctuation.

NONSENSE: This text defies common sense or presents nonsensical information.

A DIFFERENT ERROR: \_\_\_\_\_

**Figure 1:** Example machine-generated prompt and questions

The online quiz used gamification techniques to encourage participation (von Ahn, 2006). The first question under each text was scored using a simplified scoring system irrelevant to the results of the study. The final score was revealed to participants after they submitted their survey response, along with a short debrief and links to news articles relevant to fake news and natural-language generation for those interested in learning more.

Language models have been customised on video-game character biographies (Shane, 2019) and short stories (Fan *et al.*, 2018), as well as non-fictional language data, such as the news (Zellers *et al.*, 2019), Wikipedia articles (Liu *et al.*, 2018) and Reddit posts (Keskar *et al.*, 2019). This study used medical news articles for two reasons. Firstly, it is a topic that many participants may not have expert knowledge of, forcing them to rely on linguistic rather than content clues. Secondly, it is one of the fields where it is particularly important to think critically when consuming online content.

A publicly available corpus (Triki, 2020) of 1989 articles scraped from the online news outlet Medical News Today before September 2020 was used to fine-tune the GPT-2 language model and to randomly select human-written texts.<sup>[2]</sup> These informative articles are not time-sensitive, and the large amount of textual data used to fine-tune the model diminishes the chances that generated texts would closely resemble any one article.

A pilot study was run to determine the best number and length of prompts. All human-written and machine-generated texts ranged between seven and ten sentences in length to form paragraphs resembling news article previews encountered online. While language models can produce convincing short texts (Guo *et al.*, 2018), they face challenges with longer passages (Puduppully *et al.*, 2019). The prompt length was balanced to showcase longer machine-generated texts and maximise the survey completion rate by keeping the time needed under 15 minutes.

To elicit annotations on a wider range of materials and minimise the effect of individual prompt items on the final results, 21 machine-generated texts, 21 human-written texts and 7 control prompts were produced and randomised in the Qualtrics Survey Flow. To control for the effects of participants' demographic characteristics on their performance in the task, they were asked to provide their age group, gender, first language, level of English and level of education (Dörnyei, 2007) in the next section of the survey.

**Please answer the following three questions in relation to this text:**

Q119

Although we know that some people have OCD, there is no generally accepted treatment. If we are unable to find an appropriate treatment, we may recommend adjunctive or treatment-only treatments, such as: psychotherapy, medication, occupational therapy, relaxation, medication. People with OCD may also benefit from: sleep therapy. People with OCD may also benefit from: medications, sleep support, medication, therapy. If people with OCD have difficulty sleeping, they may experience withdrawal symptoms in the form of insomnia or depression. Treatment for OCD may involve medication and therapy. Learn more about the symptoms of insomnia here. Dear participant, to show that you are reading the text carefully, please select 'Definitely machine-generated' in the first question. This text is not scored.

**Do you think that this text is human-written or machine-generated?**

Definitely human-written

Possibly human-written

Possibly machine-generated

Definitely machine-generated

**Figure 2:** Example control prompt

The described survey was disseminated in social media groups on Facebook. Respondents self-selected to participate and gave informed consent. A total of 143 completed survey responses were recorded; 9 responses that failed the attention check were excluded from analysis. The following analyses are based on the remaining 134 respondents, who yielded a total of 804 annotations.

For analysis, the F1 score was used as a performance metric instead of the simplified scoring system from the quiz. The F1 score derives from the [binary confusion matrix](#) (Goutte and Gaussier, 2005) and is used as a measure of accuracy to compare the performance of diagnostic classification systems in machine learning (Swets, 1988). The F1 score derives from recall and *precision*, two metrics defined in terms of the four fields of the confusion matrix (Figure 3). The values of all three metrics range between 0 and 1. For the purposes of this paper:

- Positive condition (P) refers to all machine-generated texts.



- Negative condition (N) refers to all human-written texts.

	Guessed HW	Guessed MG
Truly HW	True Negatives (TN)	False Positives (FP)
Truly MG	False Negatives (FN)	True Positives (TP)

Figure 3: Confusion matrix (author's own graphic)

*Recall*, also known as the true positive rate (TPR), is defined as

$$TPR = \frac{TP}{TP + FN}$$

and indicates the number of items correctly identified as positive out of the total number of positive items (Buckland and Gey, 1994). In this case, it refers to how many machine-generated texts a participant can spot, and it ensures that participants do not fail to identify many machine-generated texts. The more machine-generated texts a participant successfully detects, the higher their recall.

*Precision*, also known as positive predictive value (PPV), is defined as

$$PPV = \frac{TP}{TP + FP}$$

and refers to the number of items correctly identified as positive out of all items identified as positive (Buckland and Gey, 1994). Here, it represents the number of truly machine-generated texts out of all the texts guessed as machine generated. This score ensures that participants do not simply guess every text to be machine generated to be safe.

Mathematically, the F1 score is the harmonic mean of recall and precision (Chicco and Jurman, 2020). The two scores work against each other and reward the ability to spot machine-generated texts while minimising false alarms. The F1 score is defined as follows:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2 \times \frac{PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$$

When either precision or recall is 0, the F1 score is also 0. When both are zero, it is impossible to define. To get a high F1 score, both precision and recall must be high. However, when one is high and the other is low, the F1 score will be lower. For example, when a participant correctly identifies all machine-generated texts

by guessing every text as machine generated, their recall is 1 and their precision is 0 ( $F1 = .67$ ). When they only guess one to be machine generated, but do not spot the other two, their recall is .34 and their precision is 1 ( $F1 = .50$ ).

In comparison to the regular accuracy score (arithmetic mean), the F1 score centres around the ability to identify the positive class, therefore participants score higher when they guess every text as machine generated (.67). However, those who believe every text to be human written receive an F1 score of 0.

The following section introduces specific hypotheses, describes performed analyses in detail and presents the results.

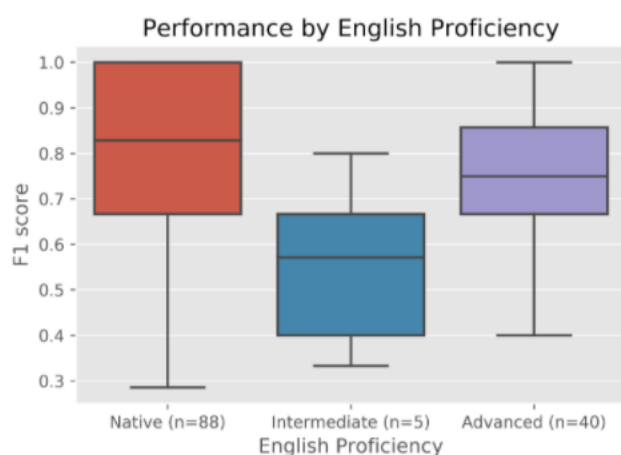
## Analysis and results

The data analysis stage of the study aimed to compare F1 scores across three levels of English proficiency and to reveal participants' covert strategies employed in the task.

One of the participants' recall and precision scores were both zero, making their F1 score impossible to determine. This outlier data point<sup>[3]</sup> was excluded from all quantitative analyses, bringing the number of L1 speakers down to 88.

The following section examines the differences in performance scores between L1 ( $n=88$ ), and advanced ( $n=40$ ) and intermediate ( $n=5$ ) L2 English speakers. Since the collected data did not follow a normal distribution (Shapiro and Wilk, 1965; Winter, 2020), non-parametric tests were used to compare the distribution of performance scores across proficiency levels.

According to a Kruskal-Wallis test, L1, advanced L2 (C1–C2) and intermediate L2 (B1–B2) English speakers differed statistically significantly in their F1 scores,  $z = 8.40$ ,  $p < .05$  (Figure 4). According to further analysis, the performance of intermediate speakers differed statistically significantly from both L1 and advanced L2 speakers. Due to their low number, they were excluded from subsequent analyses.



**Figure 4:** Distribution of F1 scores by English proficiency

A series of two-tailed unpaired Wilcoxon rank-sum tests was performed to examine the following alternative hypothesis by comparing the distribution of five metrics between L1 and advanced L2 participants: confidence, F1 score, accuracy, recall and precision:



- **H<sub>A</sub>**: There is a difference in performance between L1 and advanced L2 English speakers.

Despite the statistically significant difference between participants' scores in the Kruskal-Wallis test, the F1 scores of L1 speakers (mdn = .83) were not statistically significantly higher than those of advanced speakers (mdn = .75,  $z = 1.75$ ,  $p = .08$ ). However, the difference in accuracy between L1 (mdn = .83) and advanced L2 speakers (mdn = .67) was statistically significant ( $z = 1.98$ ,  $p < .05$ ). Such results can be reconciled after a closer look at the individual components of the F1 score. While the difference between L1 and L2 participants in terms of recall was not statistically significant ( $z = 0.74$ ,  $p = .46$ ) it was in terms of precision (L1, mdn = 1.0, L2, mdn = .67;  $z = 2.33$ ,  $p = .02$ ). A detailed analysis of accuracy (Table 1) shows that L1 and advanced L2 speakers correctly identified a comparable number of machine-generated texts, 79.17 per cent and 76.67 per cent respectively. However, while L1 speakers also correctly identified 79.55 per cent of human-written prompts, advanced L2 speakers only identified 70.83 per cent. Advanced L2 speakers' lower accuracy on human-written texts and the individual components of the F1 score reveal that both groups have a comparable ability to spot machine-generated texts. However, advanced speakers labelled more human-written texts as being machine generated. The F1 score rewards the ability to spot machine-generated texts even at the cost of a few false alarms on the part of the advanced L2 speakers.

	L1 (%)	Advanced L2 (%)	Total (%)
Machine generated (%)	79.17	76.67	78.39
Human written (%)	79.55	70.83	76.82
Total (%)	79.36	73.75	77.60

**Table 1:** Overall accuracy by prompt type and proficiency

Next, participants' covert strategies (error categories identified by participants in the third question under each prompt) were analysed to test the following two alternative hypotheses:

- **H<sub>A</sub>**: Specific error types contribute to the successful identification of a text as being machine generated.
- **H<sub>A1</sub>**: The number of errors identified in each category in machine-generated texts differs between L1 and advanced L2 English speakers.

The stacked bar plots in Figure 5 (see Table B1 for exact counts) show how many times each error category was identified, whether it led to a true positive or a false negative, and the percentage of true positives out of the total number of identified errors. There was no drastic difference between the error categories flagged in machine-generated texts by the two groups. In both groups, factual errors (NONSENSE) and repetition were flagged most frequently. However, it was undoubtedly internal contradictions (ENTAILMENT; see Figure 1, Q6, and for other error categories discussed here) and factual errors (NONSENSE) that were the most reliable in the detection of machine-generated prompts. L1 participants identified slightly more grammatical errors, whereas advanced L2 speakers noticed formatting deficiencies more often. Repetition errors were mentioned slightly more by L2 speakers. In 5 to 20 per cent of the time, even prompts marked as NO ERROR were successfully identified as machine generated. OTHER errors included the same error categories phrased differently or alluding to specific words in the prompt by participants who were not sure how to categorise their annotation. A major part of these also commented on coherence.

Subsequently, a [mixed-effects multivariate logistic regression model](#) was fit to discover which of the five specific error types (excluding NO ERROR and OTHER) most reliably led to the correct identification of prompts as being machine generated. The model accounted for fixed effects of the five error types and controlled for random effects of response ID and prompt item ( $n=21$ ). Table 2 shows a summary of the model variables. In the dataset, each observation refers to an individual annotation of a machine-generated text by a particular participant. Two separate models were fit for L1 ( $n=88$ ) and advanced L2 speakers ( $n=40$ ).

For L1 speakers, the logistic regression analysis confirmed that internal contradictions (ENTAILMENT) and factual errors (NONSENSE) most strongly predict the detection of machine-generated texts. The remaining three error types also more or less contributed to the probability that participants detect machine-generated texts correctly (see Table 3 for details). The model was checked for collinearity, and the VIF (variance inflation factor) score for all variables that fell below 2 (Tomaschek *et al.*, 2018).

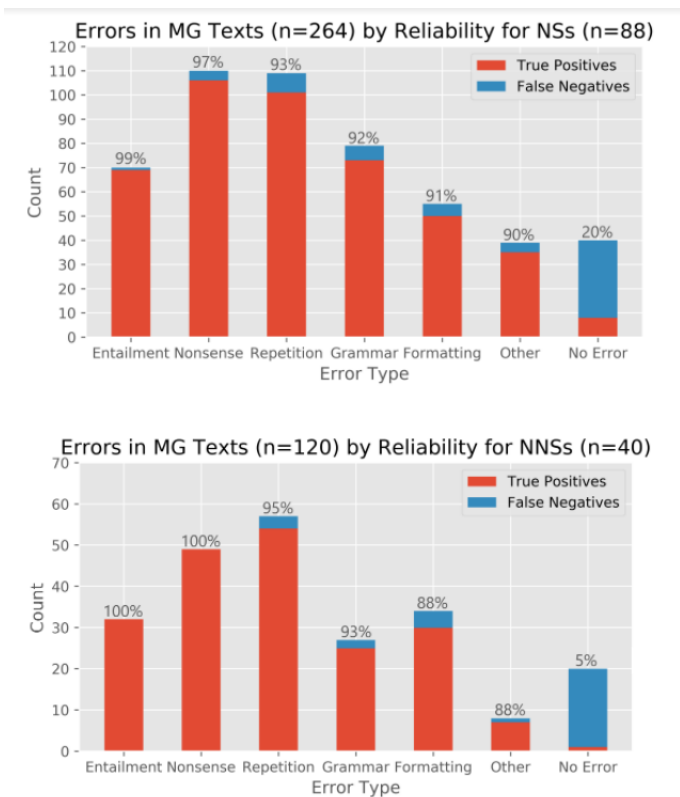
Variable	Type	Description
correct (True or False)	dependent variable	Whether or not the prompt was identified correctly as machine generated.
grammar, repetition, entailment, formatting, nonsense (True or False)	independent variables, fixed effects	Whether or not a particular error type was identified in the prompt.
Responseld (e.g., R_8qvhYV0sDJooptD), item (e.g., MG12)	control variables, random effects	ID numbers generated by Qualtrics identify responses by a single participant; item codes identify individual prompts.

**Table 2:** Logistic regression variables

Error type	Coefficient	p-value
ENTAILMENT	3.78	<.001
NONSENSE	3.29	<.001
REPETITION	2.70	<.001
GRAMMAR	2.03	<.001
FORMATTING	1.64	<.01

**Table 3:** Logistic regression results (ordered by coefficient value)

Although L1 and advanced L2 speakers agreed on the reliability of each error type leading to true positives, the logistic regression model for L2 speakers failed upon overfitting due to a lack of data points and, presumably, the large diversity of L2 participants' first languages, such as Arabic, Chinese, Dutch or German. Spanish, the most common first language in the dataset, was only represented by six speakers.



**Figure 5:** Error annotations in machine-generated prompts from L1 (top) and advanced L2 (bottom) speakers (categories ordered by reliability in descending order)

## Discussion

While previous research mostly used human judgements to further improve existing content generation systems, the aim of this paper was to provide human readers with guidance on how to best defend themselves against machine-generated medical fake news. The study revealed that publicly available language models and data sets can indeed be used to produce believable machine-generated news articles on critical topics such as health and medicine. One in five machine-generated news article excerpts shown to participants in the quiz was guessed to be human written. Machine-generated fake news is still far from indiscernible from human writing. However, as natural-language generation techniques improve, fact-checking will become increasingly important.

Shortly after Google had rolled out Bard, their brand-new AI-powered service, domain experts pointed out a factual error in its very first demo (Vincent, 2023). Astrophysicist Grant Tremblay – who had pointed out the error – later tweeted that text generators, such as OpenAI’s ChatGPT and Google Bard, are often ‘very confidently’ wrong, and the question remains whether language models will learn to fact-check the text they generate in the future (Vincent, 2023).

The F1 score (Goutte and Gaussier, 2005) was selected as an indicator of participants’ ability to defend themselves against machine-generated fake news. Although L1 and advanced L2 speakers did not achieve significantly different F1 and recall scores, L1 speakers’ precision was higher. Both groups were equally good at defending themselves against machine-generated news, but advanced L2 speakers were more sceptical and raised more false alarms by flagging human-written texts as being machine generated. This may be caused as much by the negative social implications associated with being fooled by a machine and the fear of being perceived as naïve as their generalised scepticism in their selection of news (Fletcher and Nielsen,

2019). In future research, a more precise proficiency evaluation measure would be beneficial in place of self-reporting when comparing L1 and advanced L2 (C1–C2 level) English speakers.

While there is a statistically significant difference between the F1 scores of intermediate L2 (B1–B2), advanced L2 (C1–C2) and L1 English speakers, this does not hold for the difference between L1 and advanced L2 speakers. This hints at a potential influence of participants' English proficiency on their performance in the task. Nevertheless, the study sample was too small to claim this with confidence. A larger sample of participants with lower proficiency could confirm or reject the hypothesis that highly proficient participants achieve higher scores in the classification task. However, as the aim of the study was to help English speakers of different levels critically evaluate whether they are reading machine writing, the likelihood is low that less proficient speakers rely on news in English to stay up to date.

The present study focused on the detection of machine-generated text based on linguistic errors. However, it did not consider how credibility is attributed to particular sources of news, which has been shown to influence readers' perception of news (Zakharov *et al.*, 2019). In agreement with previous research (Allcott and Gentzkow, 2017), some participants explicitly acknowledged that if they had encountered the study prompts on social media, they would have been less critical of the information. Quantitative results showed that a small number of human readers were accepting of factual errors in human writing, and some – more so L1 speakers – indicated that they would trust writing that they perceived as being machine generated. With the improvements in natural-language generation, it is possible that in the future, machine writing will not necessarily be seen as harmful, as long as the factual information is correct.

## Conclusion

Fake news is on the rise, and can be particularly harmful in the medical field. Many young people consume news on social media rather than traditional news organisations, which enables them easy access to news but lends itself to the spread of fake news. It has been shown that today's computers can generate news articles that readers rate to be more trustworthy than articles written by humans, which sparked a debate around the ethics in artificial intelligence. To prevent abuse of such AI tools, human readers must learn to spot machine writing. This paper confirms the importance of fact-checking and being aware of current news – be they human-written or machine-generated texts – in an era when advanced language-generation technology is closely approximating human writing.

The findings revealed that advanced L2 speakers were as capable of defending themselves against machine-generated fake news as their L1 counterparts. However, they may have been more sceptical (Fletcher and Nielsen, 2019) of human writing to avoid being fooled by a machine. Factual and entailment errors most reliably identified machine-generated prompts across both groups. Future research should focus on sampling homogenous groups of L2 speakers to discover more nuanced differences.

A small sample of intermediate English speakers suggests an effect of proficiency on performance in the task. Further research may wish to explore performance patterns and strategies of less proficient L2 speakers; however, these are less likely to need to rely on the consumption of news in English in their daily lives.

While there are certain linguistic errors that help identify machine writing, natural-language generation techniques are improving, and it is fact-checking and spotting internal contradictions that are the most reliable at this point in time. Although many readers blindly trust the information published in mainstream sources of news, established media companies should not be exempt from fact-checking (Gatten, 2004).

## List of figures

**Figure 1:** Example machine-generated prompt and questions

**Figure 2:** Example control prompt

**Figure 3:** Confusion matrix (author's own graphic)

**Figure 4:** Distribution of F1 scores by English proficiency

**Figure 5:** Error annotations in machine-generated prompts from L1 (top) and advanced L2 (bottom) speakers

## List of tables

**Table 1:** Overall accuracy by prompt type and proficiency

**Table 2:** Logistic regression variables

**Table 3:** Logistic regression results

## Endnotes

[1] Link to the Qualtrics survey: [http://warwick.co1.qualtrics.com/jfe/form/SV\\_4Jh8XWfATDwuAXY](http://warwick.co1.qualtrics.com/jfe/form/SV_4Jh8XWfATDwuAXY)

[2] The GPT-2 medium 355M language model was fine-tuned on an NVIDIA Tesla T4 with GPU RAM provided by the Google Colab service using the GPT-2-simple package (Woolf, 2019) to fine-tune the model for 2000 steps with the temperature parameter set to 0.7 by default. The temperature values range between 0 and 1 and determine how similar or dissimilar the resulting generated texts will be to the training data. The higher the temperature, the more 'creative' the generated text will be (or dissimilar to the training data).

[3] One L1 undergraduate woman aged 18–25 guessed all three machine-generated texts to be human written, causing her recall to be 0. Both texts guessed to be machine generated were truly human written, causing her precision to be 0 as well and rendering her F1 score impossible to define. This outlier data point was excluded from all quantitative analyses.

## References

- Aldwairi, M. and A. Alwahedi (2018), 'Detecting fake news in social media networks', *Procedia Computer Science*, 141 (2018), 215–22, available at <https://doi.org/10.1016/j.procs.2018.10.171>, accessed 20 December 2020
- Allcott, H. and M. Gentzkow (2017), 'Social media and fake news in the 2016 election', *Journal of Economic Perspectives*, 31 (2), 211–36, available at [doi: 10.1257/jep.31.2.211](https://doi.org/10.1257/jep.31.2.211), accessed 20 December 2020
- Buckland, M. and F. Gey (1994), 'The relationship between recall and precision', *Journal of the American Society for Information Science*, 45 (1), 12–19, available at [https://doi.org/10.1002/\(SICI\)1097-4571\(199401\)45:1<12::AID-ASI2>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L), accessed 12 February 2021

- Burt, M. K. (1975), 'Error analysis in the adult EFL classroom', *TESOL Quarterly*, 9 (1), 53–63, available at <https://doi.org/10.2307/3586012>, accessed 18 December 2020
- Carlson, M. (2015), 'The robotic reporter', *Digital Journalism*, 3 (3), 416–31, available at <https://doi.org/10.1080/21670811.2014.976412>, accessed 25 February 2023
- Chang, K. W., V. Prabhakaran and V. Ordonez (2019), 'Bias and fairness in natural language processing', in *Tutorial Abstracts*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, Hong Kong: Association for Computational Linguistics, pp. 2678–92, available at <https://www.aclweb.org/anthology/D19-2004>, accessed 30 December 2020
- Chicco, D. and G. Jurman (2020), 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21 (1), 1–13, available at [doi: 10.1186/s12864-019-6413-7](https://doi.org/10.1186/s12864-019-6413-7), accessed 16 February 2021
- Clerwall, C. (2014), 'Enter the robot journalist', *Journalism Practice*, 8 (5), 519–31, available at <https://doi.org/10.1080/17512786.2014.883116>, accessed 25 February 2023
- Connor-Linton, J. (1995), 'Looking behind the curtain: What do L2 composition ratings really mean?', *TESOL Quarterly*, 29 (4), 762–65, available at <https://www.jstor.org/stable/3588174>, accessed 20 December 2020
- Datta, P., P. Jakubowicz, C. Vogler and R. Kushalnagar (2020), 'Readability of punctuation in automatic subtitles', in Miesenberger, K., R. Manduchi, M. Covarrubias Rodriguez and P. Peñáz (eds.), *Lecture Notes in Computer Science*, 12,377 vols, Cham: Springer, available at [https://doi.org/10.1007/978-3-030-58805-2\\_23](https://doi.org/10.1007/978-3-030-58805-2_23), accessed 28 December 2020
- Dörnyei, Z. (2007), *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*, 1 vol, Oxford: Oxford University Press
- Fan, A., M. Lewis and Y. Dauphin (2018), 'Hierarchical neural story generation', *arXiv Preprint*, available at <https://arxiv.org/abs/1805.04833v1>, accessed 21 December 2020
- Ferris, D. R. (1994), 'Rhetorical strategies in student persuasive writing: Differences between native and non-native English speakers', *Research in the Teaching of English*, 28 (1), 45–65, available at <https://eric.ed.gov/?id=EJ479249>, accessed 21 December 2020
- Fletcher, R. and R. K. Nielsen (2019), 'Generalised scepticism: How people navigate news on social media', *Information, Communication and Society*, 22 (12), 1751–69, available at <https://doi.org/10.1080/1369118X.2018.1450887>, accessed 15 December 2020
- Foster, P. and P. Tavakoli (2009), 'Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity', *Language Learning*, 59 (4), 866–96, available at <https://doi.org/10.1111/j.1467-9922.2009.00528.x>, accessed 22 December 2020
- Frank, R. (2015), 'Caveat lector: Fake news as folklore', *The Journal of American Folklore*, 128 (509), 315–32, available at <https://doi.org/10.5406/jamerfolk.128.509.0315>, accessed 14 December 2020
- Gatten, J. N. (2004), 'Student psychosocial and cognitive development: Theory to practice in academic libraries', *Reference Services Review*, 32 (2), 157–63, available at [doi: 10.1108/00907320410537676](https://doi.org/10.1108/00907320410537676), accessed 14 December 2020



- Gehrmann, S., H. Strobel and A. M. Rush (2019), 'GLTR: Statistical detection and visualization of generated text', *arXiv Preprint*, available at <https://arxiv.org/abs/1906.04043v1>, accessed 28 December 2020
- Gelfert, A. (2018), 'Fake news: A definition', *Informal Logic*, 38 (1), 84–117, available at <https://doi.org/10.22329/il.v38i1.5068>, accessed 10 December 2020
- Goutte, C. and F. Gaussier (2005), 'A probabilistic interpretation of precision, recall and f-score, with implication for evaluation', in Losada, D. E. and J. M. Fernández-Luna (eds.), *Lecture Notes in Computer Science*, 3408, vols, Cham: Springer, available at [https://doi.org/10.1007/978-3-540-31865-1\\_25](https://doi.org/10.1007/978-3-540-31865-1_25), accessed 15 February 2021
- Guo, J., S. Lu, H. Cai, W. Zhang, Y. Yu and J. Wang (2018), 'Long Text Generation via Adversarial Training with Leaked Information', in *AAAI-18/IAAI-18/EAAI-18 Proceedings*, Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, New Orleans, LA: Association for the Advancement of Artificial Intelligence, pp. 5141–48, available at <https://ojs.aaai.org/index.php/AAAI/article/view/11957>, accessed 29 December 2020
- Ippolito, D., D. Duckworth, C. Callison-Burch and D. Eck (2020), 'Automatic Detection of Generated Text is Easiest when Humans are Fooled', in *ACL Proceedings*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, pp. 1808–22, available at <https://www.aclweb.org/anthology/2020.acl-main.164>, accessed 29 December 2020
- Keskar, N. S., B. McCann, L. R. Varshney, C. Xiong and R. Socher (2019), 'CTRL: A conditional transformer language model for controllable generation', *arXiv Preprint*, available at <https://arxiv.org/abs/1909.05858v2>, accessed 13 March 2021
- Liu, P. J., M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser and N. Shazeer (2018), 'Generating Wikipedia by summarizing long sequences', *arXiv Preprint*, available at <https://arxiv.org/abs/1801.10198v1>, accessed 18 March, 2021
- Mancilla, R. L., N. Polat and A. O. Akcay (2017), 'An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions', *Applied Linguistics*, 38 (1), 112–34, available at [doi: 10.1093/applin/amv012](https://doi.org/10.1093/applin/amv012), accessed 21 December 2020
- Naeem, S. B., R. Bhatti and A. Khan (2020), 'An exploration of how fake news is taking over social media and putting public health at risk', *Health Information and Libraries Journal*, advance online publication, available at <https://doi.org/10.1111/hir.12320>, accessed 25 March, 2021
- Nunan, D. (2003), 'The impact of English as a global language on educational policies and practices in the Asia-Pacific region', *TESOL Quarterly*, 37 (4), 589–613, available at <https://doi.org/10.2307/3588214>, accessed 10 December 2020
- Pennycook, A. (2017), *The Cultural Politics of English as an International Language*, 1 vol, Sydney: Routledge, available at <https://doi.org/10.4324/9781315225593>, accessed 10 December 2020
- Puduppully, R., L. Dong and M. Lapata (2019), 'Data-to-Text Generation with Content Selection and Planning', in *AAAI-19/IAAI-19/EAAI-19 Proceedings*, Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019, Honolulu, HI: Association for the Advancement of Artificial Intelligence, pp. 6908–15, available at <https://doi.org/10.1609/aaai.v33i01.33016908>, accessed 23 February 2021

- Punjabi, P. P. (2017), 'Science and the "fake news" conundrum', *Perfusion*, 32 (6), 429–29, available at <https://doi.org/10.1177/0267659117727418>, accessed 15 December 2020
- Radford, A., J. Wu, D. Amodei, J. Clark, M. Brundage and I. Sutskever (2019), 'Better language models and their implications', available at <https://openai.com/blog/better-language-models>, accessed 23 February 2021
- Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever (2019), 'Language models are unsupervised multitask learners', available at [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf), accessed 23 February 2021
- Resnik, P. and J. Lin (2010), 'Evaluation of NLP systems', in Clark, A., C. Fox and S. Lappin (eds.), *The Handbook of Computational Linguistics and Natural Language Processing*, Hoboken, NJ: Blackwell Publishing Ltd., pp. 271–95, available at <https://doi.org/10.1002/9781444324044.ch11>, accessed 22 February 2021
- Saikh, T., A. Anand, A. Ekbal and P. Bhattacharyya (2019), 'A novel approach towards fake news detection: Deep learning augmented with textual entailment features', in Métails, E., F. Meziane, S. Vadera, V. Sugumaran and M. Saraee (eds.), *Lecture Notes in Computer Science*, 11,608 vols, Cham: Springer, available at [https://doi.org/10.1007/978-3-030-23281-8\\_30](https://doi.org/10.1007/978-3-030-23281-8_30), accessed 21 December 2020
- Saygin, A. P., I. Cicekli and V. Akman (2000), 'Turing test: 50 years later', *Minds and Machines*, 10 (4), 463–518, available at <https://doi.org/10.1023/A:1011288000451>, accessed 10 December 2020
- Scarcella, R. C. (1984), 'How writers orient their readers in expository essays: A comparative study of native and non-native English writers', *TESOL Quarterly*, 18 (4), 671–88, available at <https://doi.org/10.2307/3586582>, accessed 18 December 2020
- Seidlhofer, B. (2004), 'Research perspectives on teaching English as a lingua franca', *Annual Review of Applied Linguistics*, 24 (1), 209–39, available at [doi: 10.1017/s0267190504000145](https://doi.org/10.1017/s0267190504000145), accessed 14 December 2020
- Shane, J. (2019), 'DandD character bios – Now making slightly more sense!', available at <https://www.janelleshane.com/aiweirdness/dd-character-bios-now-making-slightly-more>, accessed 15 March, 2021
- Shapiro, S. S. and M. B. Wilk (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika*, 52 (3), 591–611, available at <https://doi.org/10.2307/2333709>, accessed 25 January 2021
- Sharevski, F., P. Jachim, P. Treebridge, A. Li, A. Babin and C. Adadevoh (2021), 'Meet Malexa, Alexa's malicious twin: Malware-induced misperception through intelligent voice assistants', *International Journal of Human-Computer Studies*, 149, 102604–17, available at <https://doi.org/10.1016/j.ijhcs.2021.102604>, accessed 30 December 2020
- Shi, L. (2001), 'Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing', *Language Testing*, 18 (3), 303–25, available at <https://doi.org/10.1177/026553220101800303>, accessed 18 December 2020
- Schwarz, N. (2008), 'The Psychology of Survey Response', in Donsbach, W. and M. W. Traugott (eds.), *The SAGE Handbook of Public Opinion Research*, London: SAGE Publications Ltd., pp. 374–87, available at

<https://dx.doi.org/10.4135/9781848607910.n35>, accessed 15 February 2021

- Song, B. and I. Caruso (1996), 'Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students?', *Journal of Second Language Writing*, 5 (2), 163–82, available at [https://doi.org/10.1016/S1060-3743\(96\)90023-5](https://doi.org/10.1016/S1060-3743(96)90023-5), accessed 18 December 2020
- Stahlberg, F. and S. Kumar (2021), 'Synthetic Data Generation for Grammatical Error Correction with Tagged Corruption Models', in *ACL Proceedings*, Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021, Online: Association for Computational Linguistics, pp. 37–47, available at <https://www.aclweb.org/anthology/2021.bea-1.4.pdf>, accessed 30 December 2020
- Swets, J. A. (1988), 'Measuring the accuracy of diagnostic systems', *Science*, 240 (4857), 1285–93, available at [DOI: 10.1126/science.3287615](https://doi.org/10.1126/science.3287615), accessed 23 February 2021
- Tomaschek, F., P. Hendrix and R. H. Baayen (2018), 'Strategies for addressing collinearity in multivariate linguistic data', *Journal of Phonetics*, 71 (2018), 249–67, available at <https://doi.org/10.1016/j.wocn.2018.09.004>, accessed 30 March 2021
- Triki, A. (2020), '2k clean medical articles (MedicalNewsToday), Version 1', available at <https://www.kaggle.com/trikialaaa/2k-clean-medical-articles-medicalnewstoday/version/1>, accessed 14 December 2020
- Tsang, A. (2017), 'Judgement of countability and plural marking in English by native and non-native English speakers', *Language Awareness*, 26 (4), 343–59, available at <https://doi.org/10.1080/09658416.2017.1410553>, accessed 18 December 2020
- Turing, A. (1950), 'Computing machinery and intelligence', *Mind*, 59 (236), 433–60, available at <https://doi.org/10.1093/mind/LIX.236.433>, accessed 13 March 2021
- Vincent, J. (2023), 'Google's AI chatbot Bard makes factual error in first demo', available at <https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo>, accessed 26 February 2023
- Von Ahn, L. (2006), 'Games with a purpose', *Computer*, 39 (6), 92–94, available at [doi: 10.1109/mc.2006.196](https://doi.org/10.1109/mc.2006.196), accessed 13 March 2021
- Vosoughi, S., D. Roy and S. Aral (2018), 'The spread of true and false news online', *Science*, 359 (6380), 1146–51, available at <https://doi.org/10.1126/science.aap9559>, accessed 10 December 2020
- Winter, B. (2020), *Statistics for Linguists: An Introduction Using R*, 1 vol, New York: Routledge, available at <https://doi.org/10.4324/9781315165547>, accessed 12 February 2021
- Woolf, M. (2019), 'How to make custom AI-generated text with GPT-2', available at <https://minimaxir.com/2019/09/howto-gpt2>, accessed 15 November 2020
- Xaitqulov, Z. (2021), 'An overview of automated translation and its linguistic problems', *Philology Matters*, 2021 (1), 139–49, available at [DOI: 10.36078/987654482](https://doi.org/10.36078/987654482), accessed 25 December 2020
- Yadav, A. K., A. K. Maurya, R. S. Ranvijay (2021), 'Extractive text summarization using recent approaches: A survey', *Ingénierie des Systèmes d'Information*, 26 (1), 109–21, available at

<https://doi.org/10.18280/isi.260112>, accessed 24 February 2020

Zakharov, W., H. Li and M. Fosmire (2019), 'Undergraduates' news consumption and perceptions of fake news in science', *Libraries and the Academy*, 19 (4), 653–65, available at

<https://preprint.press.jhu.edu/portal/sites/ajm/files/19.4zakharov.pdf>, accessed 15 December 2020

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner and Y. Choi (2019), 'Defending Against Neural Fake News', in *Advances in Neural Information Processing Systems*, Proceedings of the Thirty-Second NeurIPS Conference, 2019, Vancouver: NeurIPS, pp. 1–12, available at

<https://proceedings.neurips.cc/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf>, accessed 24 December 2020

## Glossary

**B1–B2 level English:** Under the CEFR standard (Common European Framework of Reference for Languages), the B1–B2 level refers to intermediate and upper-intermediate users of the language who are able to communicate independently in most contexts.

**Binary confusion matrix:** In machine learning, the confusion matrix, also known as the error matrix, is a table visualisation of the performance of an algorithm. Each row represents the occurrences in an actual class, and each column represents the occurrences in a predicted class. In this paper, the matrix was used for the guesses of individual human participants rather than an algorithm.

**C1–C2 level English:** Under the CEFR standard (Common European Framework of Reference for Languages), the C1–C2 level refers to advanced and proficient users of the language who are often nearing native proficiency.

**Entailment:** Relationship between two claims where if one is true, the second must also be true for example, 'I see a cat' entails 'I see an animal'. However, in this context, machines will frequently assert the first and deny the second.

**F1 score:** The F1 score is a measure used in statistical analysis of binary classification. It derives from the binary confusion matrix (Goutte and Gaussier, 2005) and is used as a measure of accuracy to compare the performance of diagnostic classification systems in machine learning (Swets, 1988).

**L1 English speakers:** Those for whom English is their first language, native language or mother tongue. They are often described as native speakers.

**L2 English speakers:** Those for whom English is an additional, second or foreign language. They are often described as non-native speakers.

**Morphosyntactic errors:** Grammatical errors in word formation or word order; for example, 'he go to school'.

**Mixed-effects multivariate logistic regression analysis:** Formula that predicts the relationships between dependent and independent variables. It allows researchers to calculate the probability of a specific outcome depending on a number of variables. It is particularly common in the field of machine learning. A mixed-effects model is used to account for both within-person (several quiz prompts per person) and across-person (many participants in the study) variability.

---

To cite this paper please use the following details: Dankova, B. (2023), 'You Had Better Check the Facts: Reader Agency in the Identification of Machine-Generated Medical Fake News', *Reinvention: an International Journal of Undergraduate Research*, Volume 16, Issue 1, <https://reinventionjournal.org/article/view/964>. Date accessed [insert date].

If you cite this article or use it in any teaching or other related activities please let us know by e-mailing us at [Reinventionjournal@warwick.ac.uk](mailto:Reinventionjournal@warwick.ac.uk).