








Artificial intelligence—friend or foe in fake news campaigns

 Krzysztof Węcel¹  Marcin Sawiński²  Milena Stróżyńska³
 Włodzimierz Lewoniewski⁴  Ewelina Książniak⁵
 Piotr Stolarski⁶  Witold Abramowicz⁷

Abstract

In this paper the impact of large language models (LLM) on the fake news phenomenon is analysed. On the one hand decent text-generation capabilities can be misused for mass fake news production. On the other, LLMs trained on huge volumes of text have already accumulated information on many facts thus one may assume they could be used for fact-checking. Experiments were designed and conducted to verify how much LLM responses are aligned with actual fact-checking verdicts. The research methodology consists of an experimental dataset preparation and a protocol for interacting with ChatGPT, currently the most sophisticated

Keywords

- artificial intelligence
- large language models
- fake news
- fact-checking

¹ Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, corresponding author: krzysztof.wecel@ue.poznan.pl, <http://orcid.org/0000-0001-5641-3160>.

² Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, marcin.sawinski@ue.poznan.pl, <http://orcid.org/0000-0002-1226-4850>.

³ Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, milena.strozyzna@ue.poznan.pl, <http://orcid.org/0000-0001-7603-7369>.

⁴ Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, wlodzimierz.lewoniewski@ue.poznan.pl, <http://orcid.org/0000-0002-0163-5492>.

⁵ Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, ewelina.ksiezniak@ue.poznan.pl, <http://orcid.org/0000-0003-1953-8014>.

⁶ Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, piotr.stolarski@ue.poznan.pl, <http://orcid.org/0000-0001-7076-2316>.

⁷ Department of Information Systems, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, witold.abramowicz@ue.poznan.pl, <https://orcid.org/0000-0001-5464-9698>.

LLM. A research corpus was explicitly composed for the purpose of this work consisting of several thousand claims randomly selected from claim reviews published by fact-checkers. Findings include: it is difficult to align the responses of ChatGPT with explanations provided by fact-checkers; prompts have significant impact on the bias of responses. ChatGPT at the current state can be used as a support in fact-checking but cannot verify claims directly.

JEL codes: C45, C52, D83, L86, L15

Article received 26 April 2023, accepted 16 June 2023.

This work was supported with a grant INFOSTRATEG-I/0035/2021-00 OpenFact – narzędzia weryfikacji wiarygodności źródeł informacji i detekcji fałszywych informacji z wykorzystaniem metod sztucznej inteligencji, financed by the National Center for Research and Development (Poland).

Suggested citation: Węcel, K., Sawiński, M., Stróżyna, M., Lewoniewski, M., Księżniak, E., Stolarski, P., & Abramowicz, W. (2023). Artificial intelligence—friend or foe in fake news campaigns. *Economics and Business Review*, 9(2), 41–70. <https://doi.org/10.18559/ebr.2023.2.736>



This work is licensed under a Creative Commons Attribution 4.0 International License <https://creativecommons.org/licenses/by/4.0>

Introduction

As humanity we are facing a new challenge. Development of artificial intelligence (AI) is faster than foreseen just several years ago, follows an exponential growth pattern and is going into areas with undefined rules. A few years ago several business leaders postulated the definition of some rules that the development of AI should follow (Candelon et al., 2021; Gibbs, 2017). Artificial intelligence is again a frequently discussed topic, also among non-professionals. It has entered several areas that so far were reserved for people, for example graphic design or music composition. In July 2022 the rise of Midjourney was observed which was a breakthrough in image generation based on a textual prompt (the so-called text-to-image model). It joined a similar service that appeared earlier—Dall-E, initially revealed by OpenAI in January 2021 and was followed by an open counterpart Stable Diffusion.

Another revolution was observed in the equally challenging task of conversational text generation. In November 2022 ChatGPT started and within a week gained one million users (Buchholz, 2023). It was a follow-up

to an earlier large language model: GPT-3. The disruptive change was the introduction of the capability to conduct discussions, hence the name ‘ChatGPT’.

The capabilities of AI, both in the image and text generation domain, raised a lot of questions as to what it will mean for the future. Notably economists started discussing how much human labour can be displaced with this technology (Malone, 2018; Mayor, 2019). Another perspective that is covered in this paper is how Generative AI can impact the consumption of digital content by people. It is important to note that AI can generate content that did not exist before. It can follow a creative-like approach where abstract images are generated. Several years ago *deepfake* emerged—faces of ordinary people in arbitrary videos were replaced with these of famous people which put them in problematic situations (Westerlund, 2019). Current technology is way more advanced.

Always when a malicious technology is created the development of countermeasures follows. In this context the impact of recent development within artificial intelligence and large language models like ChatGPT in particular, on fake news generation and detection must be considered.

Based on an analysis of recent publications the possibility of generating fake news with AI is considered. There was a thesis that fake news is dangerous because it can be designed convincingly with the help of AI and produced in large volume. After several own experiments and looking at current comments in media, it was concluded that this thesis would be trivial. At the moment AI is notorious for producing texts not consistent with reality and with confused facts, the phenomenon known as hallucination (Ji et al., 2022). Moreover, the models can be guided to produce harmful content despite the efforts of mainstream creators to limit this threat. Therefore, it was decided to focus on detecting fake news generated either by people or AI. Thus the following thesis was formulated: considering the wide knowledge base and built-in ‘ethical’ rules, AI can identify fake news. The aim of the paper is to verify how many responses generated by large language models are aligned with actual fact-checking verdicts. In the following sections additional constraints and assumptions are discussed, e.g., the degree of involvement of humans. The authors of the paper are involved in the research project ‘OpenFact’, hence the interest in fake news detection. Nevertheless, it is hard to discuss countermeasure tools without referring to the roots of the phenomenon.

The paper is structured as follows. In Section 1 a background for large language models and fake news detection is provided. Moreover, research questions are formulated. In Section 2 a research methodology and the main assumptions of the experiments are presented. Their results and findings are presented in Section 3. They are discussed in Section 4 and Section 5 concludes the paper.

1. Theoretical background

In this section the background of natural language processing is presented. The reader can find a discussion of how words and sentences are represented as vectors and what progress was made towards large language models (LLM).

The early development of NLP followed a rule-based approach, which involved creating sets of rules and patterns for machines to follow to understand human language (Weizenbaum, 1966). This approach relied heavily on hand-crafted rules and syntactic structures that had too limited a flexibility to handle complex natural language.

The next phase of NLP development was based on statistical approaches, which involved the use of statistical methods to analyse large amounts of data and extract patterns and rules automatically. Since algorithms cannot process raw text the conversion into a numeric representation became a critical factor in the success of statistical NLP.

In natural language processing (NLP) text representations have evolved from bag-of-words models to word embeddings and, more recently, token embeddings. Initially words were represented as vectors of the length of the size of a dictionary and each occurrence of a word was marked as 1.0 (the method referred to as ‘one-hot encoding’). Later a more sophisticated weighting was introduced which considered the importance of a word within a document (term frequency) and within a collection (inverse document frequency), and was named TFIDF. Bag-of-words models represent documents as a sparse vector of word frequencies, while word embeddings use dense, low-dimensional vectors to represent individual words based on their contextual usage in a large corpus. Word embeddings capture not only the similarity between words but also their context and allow to the formation of relationships (e.g., operation on vectors for words ‘King—Man + Woman’ produces results very close to the embedding of the word ‘Queen’). Word embeddings are typically learned through unsupervised machine learning algorithms, on large text corpora and can be reused for other tasks. The so-called fine-tuning process can even be conducted with limited training data. These representations have proven effective in capturing rich semantic information about the language and are widely used as input features in downstream NLP tasks.

The idea for word embeddings has been around since the 1980s but it was not until 2013 that it was successfully implemented in practice with the development of an efficient technique for generating word embeddings using neural networks (Mikolov, Chen et al., 2013). There are many implementations of word embeddings but the most popular are Word2Vec, GloVe, FastText, and ELMO, each having its own strengths and weaknesses depending on the specific task at hand.

Word2Vec is a neural network-based approach that learns word embeddings by predicting the context of a word given its surrounding words. There are two variants in Word2Vec: Continuous Bag of Words (CBOW) and Skip-gram. In the CBOW model, the algorithm predicts the centre word given its surrounding words while in the Skip-gram model the algorithm predicts the surrounding words given the centre word. The resulting word embeddings capture the semantic and syntactic relationships between words (Mikolov, Sutskever et al., 2013).

GloVe (Global Vectors for Word Representation) is another algorithm for creating word embeddings that is based on the co-occurrence statistics of words in a corpus. The GloVe algorithm constructs a matrix of word co-occurrence counts and uses matrix factorization techniques to learn a low-dimensional vector representation for each word. The resulting word embeddings capture both the global and local contexts of words (Pennington et al., 2014).

Deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) used for NLP tasks showed significant improvements over traditional machine learning models. RNNs were effective in capturing long-term dependencies in sequential data but they faced challenges in processing long sequences due to the ‘vanishing gradient’ problem. Long Short-Term Memory (LSTM) networks were introduced as a solution to tackle this problem in RNNs, and showed significant improvements in processing long sequences of text. LSTM-based models were used in a range of NLP tasks such as language modelling, speech recognition, machine translation and named entity recognition.

Significant improvements in NLP were achieved by introducing attention mechanisms in RNN-based models (Bahdanau et al., 2014). Further development of NLP is marked by the emergence of models that use Transformer Architecture (Vaswani et al., 2017). This architecture does not use recurrent or convolutional layers but is based on the attention mechanism. It significantly improved the state of the art in NLP tasks such as machine translation and language modelling. The Transformer architecture is also much faster to train and easier to parallelise, which significantly reduced the time necessary for training.

The Transformer architecture was further developed and improved by introducing BERT—Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) and its variants. BERT is a large language model that is trained on a large corpus of unlabelled text. It is based on the Transformer architecture and uses a bidirectional training scheme. BERT is trained on two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). The MLM task involves masking some of the words in the input sequence and predicting the original words based on the context. The NSP task involves predicting whether the second sentence in a pair of sentences is the next sentence in the original text. The above tasks allow the circumvention

of the limitation of the training, i.e., a limited number of human annotations. BERT is trained on a large corpus of unlabelled text in the process called self-supervision. BERT is first taught general-purpose representations of language and is then fine-tuned on a specific task, which allows it to learn task-specific representations of language. BERT has shown significant improvements over previous state-of-the-art models in a number of NLP tasks. It has been used for question-answering, natural language inference and text classification, among others.

BERT token embeddings are fundamentally different from GloVe and Word2Vec word embeddings. GloVe and Word2Vec are based on a predictive approach that aims to learn word vectors that predict the surrounding words in a text corpus. In contrast BERT token embeddings are based on a masked language modelling approach that learns to predict masked words within a given context. This allows BERT to capture both local and global context information in text resulting in highly effective embeddings for downstream NLP tasks. Additionally, while GloVe and Word2Vec return the same vector for a given word, BERT takes context into account, so that context for a rock, a genre of music, is different from a rock a solid aggregate of minerals. This allows BERT to capture the nuances of language more accurately.

GPT token embeddings are similar to BERT as they are also based on the Transformer architecture and learn contextual information. However, there are some differences. BERT is trained to predict the missing tokens whereas GPT is trained to predict the next word in a sequence, i.e., looking only forward. This allows GPT to generate text in a more coherent and natural-sounding manner.

Both GPT and BERT generate embeddings for each token. However, GPT is a generative model and is designed for tasks such as text generation and language modelling, while BERT is designed for tasks such as sentiment analysis, question answering and named entity recognition.

In summary, GPT and BERT are both powerful transformer-based models that generate contextualized embeddings for tokens. While BERT is better suited for tasks that require a deep understanding of the meaning of a sentence, GPT is better suited for tasks that involve generating natural language text. GloVe and Word2Vec, on the other hand, generate static embeddings for words that do not capture the context in which they appear.

GPT-3 (Generative Pre-trained Transformer) is a third-generation, autoregressive model developed by OpenAI that uses deep learning to produce human-like text (sequences of words, code, or other data) starting from a source input (prompt) provided by a user (Floridi & Chiriatti, 2020). The model works by predicting the next word or sequence of words statistically based on the preceding context and can do so for NLP tasks it has not been trained on (Dale, 2021). It means that baseline GPT-3 does not know how to perform any task, it knows how to learn to do it, which makes it more powerful and versatile. The model was trained on a large dataset of Internet texts such as Wikipedia and

programming codes, primarily in English but also in other languages. In general, GPT-like models need to be trained with large amounts of data to produce relevant results—GPT-3 was trained with 570GB in total (Romero, 2021). For comparison the first generation of GPT used 110 million learning parameters (i.e., the values that a neural network tries to optimize during training), while GPT-2 used 1.5 billion and GPT-3 175 billion (Floridi & Chiriatti, 2020). Training of such models is very expensive due to the cost of infrastructure it needs; the estimated cost of GPT-3 training reaches \$12 million (Floridi & Chiriatti, 2020).

Recent GPT models were not made available by OpenAI and instead access is provided through an API. The model's creators argue that it gives them more control over its use. GPT-3 models are offered in four sizes: Davinci, Curie, Babbage and Ada, each suitable for tasks of different complexity. Each GPT model can be fine-tuned (customized) on a specific task. Thanks to the so-called “few-shot learning” after receiving a few prompts the model is able to intuit the task a user is trying to perform and adjust a plausible response accordingly. Fine-tuning is about improving the model on few-shot learning by training on many more examples that can fit in the prompt resulting in better results on a wide number of tasks. This allows GPT-3 to be made a specialist for different tasks.

GPT-3 works for a wide range of use cases including summarization, translation, grammar correction, question answering, chatbots, composing emails and much more. As a result in the nine months since its launch it has generated 4.5 billion words per day on average. There were over 300 applications that were using GPT-3 and tens of thousands involved developers (OpenAI & Pilipiszyn, 2021).

ChatGPT is LLM developed by OpenAI being the latest in a series of such models released by this company. ChatGPT is fine-tuned from GPT-3.5 and optimized for dialogue by using Reinforcement Learning with Human Feedback (RLHF)—a method that uses human demonstrations to guide the model towards the desired behaviour. At the moment of writing there is an even more powerful GPT-4 available but the details of the architecture are not known. The most important characteristic is its multimodal capabilities.

ChatGPT is a successor of another LLM model by OpenAI—InstructGPT (Ouyang et al., 2022). To develop ChatGPT, OpenAI used the same methods as InstructGPT but with slight differences in the data collection setup. OpenAI trained an initial model using supervised fine-tuning, i.e., human AI trainers provided conversations in which they played both sides—the user and an AI assistant. Moreover, the trainers had access to model-written suggestions to help them compose their responses. Then they mixed this new dialogue dataset with the InstructGPT dataset, which was earlier transformed into a dialogue format.

Because the goal of ChatGPT is to maximize the similarity between its outputs and the dataset it was built on it was trained on data from the inter-

net written by people including conversations. According to some journalists ChatGPT training data includes the entire English language contents of Wikipedia—eight years' worth of web pages crawled from the public internet and scans of English-language books (Corfield, 2023). Due to the fact that the training of the underlying GPT-3.5 model finished in Q4 of 2021 and it is not connected to the internet ChatGPT has limited knowledge of the world and events after 2021. Moreover, the model is able to reference up to approximately 3,000 words (or 4,000 tokens) from the current conversation with a user so answering the questions takes into account the context of previous prompts and answers in the conversation.

ChatGPT has gained a lot of excitement and controversy lately because it is one of the first models that can convincingly converse with a user on a wide range of topics. Moreover, the conversation may be lead not only in a single language, e.g. English as in the case of other LLMs but in other languages as well. Another reason for its popularity is the fact that it is free,⁸ easy to use and continues to learn. Its dialogue interface available through a web browser allows users to interact with the model more effectively and efficiently via interactive chats.

Besides it has shown peculiarities in many areas of NLP. One of these is nonfiction writing such as dialogue (King & ChatGPT, 2023), impersonation (Motoki et al., 2023), essays (Alkaissi & McFarlane, 2023; Rudolph et al., 2023), news articles (George & George, 2023), summaries (Lund & Wang, 2023; Patel & Lam, 2023), etc. Another area—professional writing, e.g., advertisements (Haleem et al., 2022; Paul et al., 2023), emails (Shen et al., 2023), copywriting (Thurzo et al., 2023), content marketing (Rivas & Zhao, 2023), note-taking (Hosseini et al., 2023). Creative writing is also one of popular direction of ChatGPT applications including fiction (Dwivedi et al., 2023), poetry (Kirmani, 2022), songs (McGee, 2023), humour (Kirmani, 2022), memes (Yang et al., 2023), etc. ChatGPT can also present rational skills such as counting (Frieder et al., 2023), analogies (Bang et al., 2023), concept blending, forecasting (Lopez-Lira & Tang, 2023), etc. It has also emergent abilities such as code and multi-modal generation (Bang et al., 2023).

During the researchers' early experiments it was observed that large language models can be used to verify provided statements. This is not the intended purpose of LLM but people may use it this way. A statement can be verified as true or false based on a provided context (e.g., previous sentences like in the case of NLI—natural language inference) or based on internal knowledge of the model acquired during training. It is important to note that LLM is not a knowledge base and does not have access to any—what can be assessed is if the model could accumulate generalised knowledge of some facts.

⁸ At the date of writing this paper, OpenAI has already introduced an option of a paid subscription for premium use

Other experiments conducted for the purpose of this paper show that it can also consider the style of the text or the so-called psycholinguistic features.

Concerning fake news detection the best models currently used still rely on transformer-based models, primarily BERT (Faggioli et al., 2022). There is one paper that utilized the GPT-3 model; however, it focused on claim detection without assessing whether the claims were true or false (Agresti et al., 2022). It is important to note that this study used a previous generation of GPT models.

As the idea is relatively new there are no studies on using large language models for fake news detection; and this is the research gap that is intended to be covered within this paper. Based on the above analysis the following research questions have been formulated:

RQ1. How should a large language model be prompted to identify fake news in a more meaningful way? The model is presented with a statement and optionally some background information and needs to respond if the statement is fake news. Various levels of misinformation can be identified, e.g. true, false, manipulation and not verifiable.

RQ2. How precisely can a large language model detect current fake news when trained on older data? Large language models are trained on knowledge collected until some moment in time (end of 2021 in the case of GPT). The research question can also be reformulated: how robust is LLM with regard to time? It might be true assuming that LLM can identify slight style changes in fake news compared to facts.

2. Research methodology

This section outlines the methodology followed for conducting the experiments and gathering the research data. It begins with providing a rationale for the model selection followed by a detailed description of the experimental protocol and the dataset used in the experiments.

2.1. Generative AI

The concept of Generative AI has become a widely discussed topic in recent times. It refers to the use of artificial intelligence for generating various forms of digital content such as text, images, videos and music. This paper primarily focuses on models designed for text generation with a particular emphasis on the involvement of large language models in the proliferation of fake news.

Initially the authors discussed research questions that focused on exploring the use of Generative AI for producing fake news. However this line of inquiry was subsequently abandoned as it proved to be trivial. Despite the significant efforts made by the companies responsible for developing large language models to ensure their accuracy the safety mechanisms often prove to be inadequate. From the perspective of fake news production, two primary risks emerge: (1) the model may generate fake news involuntarily and (2) the model may be intentionally utilized to generate fake news.

Following the release of ChatGPT, numerous issues were raised concerning the factual accuracy of the generated text. The model exhibited a tendency to produce fictitious statements spontaneously, a phenomenon referred to as “hallucinations.” Additionally the use of specific prompts enabled the bypassing of the safety mechanisms thereby facilitating the generation of intentionally targeted fake news. The possibility of misusing the model for other purposes such as training it with customized data to produce fake news tailored to arbitrary topics or following provided narrations was also considered. However this direction of research was not further pursued.

2.2. Experiment design

The experiments were designed to investigate the functionality of large language models (LLMs) in relation to the fake news domain. At the time of the research a limited number of LLMs were accessible to the public with ChatGPT, developed by OpenAI emerging as the most widely used and advanced. ChatGPT possesses the capability to provide responses to queries that pertain to prior discussions and contextual information provided. The model has been trained on a substantial corpus of conversations, enabling it to generate responses that are coherent and consistent with the given context.

To address both research questions RQ1 and RQ2 the language model can be tasked with verifying whether a given claim can be classified as true, false, or undefined in cases where ambiguity arises. This procedure forces LLM to rely solely on the information that has been acquired from the source documents during the training process and encoded in the weights of LLM. As the internal mechanisms utilized by the model to determine factual accuracy remain unknown the output is solely evaluated by comparing it to the fact-checking verdict.

The claims for verification were divided into two groups: (1) claims that were verified by the fact-checkers before the end of 2021 for RQ1 and (2) claims that were verified by the factcheckers after the end of 2021 for RQ2. The rationale behind this decision was to investigate whether the model is more consistent with the fact-checking verdicts of the first group thus focusing on

the factual content of the claim or if the model is consistent similarly in both groups and thus being more sensitive to style in which a claim was written.

The claims were also retrieved in two languages: Polish and English. Although all the claims can be translated on-the-fly the performance of the model was also evaluated based on the differences between the two languages. It can be verified if fake news in any language can be detected with higher accuracy.

2.3. Research data collection

The claims for verification were randomly selected from fact-checking websites with the aim of obtaining a representative sample of the claims that were verified by the fact-checkers.

The claims written in English were obtained from the following fact-checkers: factcheck.org, factcheck.afp.com, newsweek.com, politifact.com, polygraph.info, snopes.com, usatoday.com, and washingtonpost.com. The claims written in Polish were retrieved from the following factcheckers: demagog.org.pl and fakehunter.pap.pl.

All claims except for the ones from fakehunter.pap.pl were obtained with Google Fact Check Tool APIs. This means that only a portion of recent claims was gathered providing that they were tagged by fact-checkers with a structured format called ClaimReview and were successfully indexed and processed by Google Fact Check Tool. Claims verified by fact-checkers prior to the adoption of the ClaimReview standard were not included in the dataset. The claims from fakehunter.pap.pl were obtained directly from fakehunter.pap.pl website. The total number of retrieved claims from all fact-checkers was 39,409.

For all claims the following information was extracted: claim text, claimant, claim date, factchecking organization name and fact-checking verdict. It is important to emphasize that claim text is formulated by fact-checkers based on the original claim source and does not have to be the same as the original claim. Claim text is a summary or paraphrase of the original claim source but without most stylistic features of the original claim. To provide additional context fact-checkers were mapped to the country of their origin.

A clean-up of claims from the dataset was performed in several steps. First, claims that were not written in English or Polish were removed. Then, claims that directly referenced images or videos were removed as the context for such claims would be too vague for the purpose of research. Based on the verdict claims that were verified as satirical content were also removed. Finally, fact-checker verdicts were mapped to labels containing only the following values: 'true', 'false', 'partially false' and 'no evidence'. Explicit true and false verdicts were mapped to 'true' and 'false' labels respectively. More

ambiguous verdicts such as ‘mostly true’, ‘mostly false’, ‘partially true’, ‘partially false’, ‘misleading’, ‘missing context’ or ‘manipulated information’ were mapped to the ‘partially false’ label. All verdicts that did not contain clear-cut decisions were mapped to the ‘no evidence’ label. The dataset after clean-up contained 31,455 claims that were candidates for further analysis. Next, 200 random samples were picked for each of the available labels. The final dataset contained 800 claims.

For the experiment six prompt templates were prepared using Chat Markup Language (ChatML) format. Templates include claim text as well as additional context information such as country of origin, date of claim and author of the claim (put in curly brackets). The prompts were formulated in such a way as to compel ChatGPT to generate responses that align with the fact-checkers’ verdict format. The following prompts coded as numbers were used in the experiment:

- prompt 100: User: Assess veracity of the claim: {claim_text}. Give a verdict: false, partly false, true, or no evidence,
- prompt 200: User: You are a fact checker. Classify as: false, partly false, true, no evidence. {claim_text},
- prompt 300: User: Debunk claim {claim_text}. Classify claim as: false, partly false, true, no evidence,
- prompt 400: System: You are a fact-checker assistant. Assess veracity of the claim from user. Your verdict must be in a form of a single label from this list: false, partly false, true, no evidence. No other text in response. User: {claim_text},
- prompt 500: User: Classify as: false, partly false, true, no evidence. Claim origin: {claim_country_of_origin}, claimant: {claimant}, claim date: {claim_date}. Claim for classification: {claim_text},
- prompt 600: User: Consider claim {claim_text}, Claim origin: {claim_country_of_origin}, claimant: {claimant}, claim date: {claim_date}., Classify claim as: false, partly false, true, no evidence.

Note that only prompts 500 and 600 contain additional context information. Prompt 400 contains a system message to set up initial instructions for the model. No advanced prompt engineering was performed to mimic casual user interaction with the system.

The prompts were used to generate verdicts by GPT-3.5 Turbo model using OpenAI APIs. The parameters used to generate answers were optimized to reduce the variance of responses and maximize the probability of factual answers: temperature = 0.0, top_p = 1.0, presence_penalty = 0.0 and frequency_penalty = 0.0.

The answers obtained from ChatGPT were processed to extract ratings that correspond to the predefined set of labels namely ‘true’, ‘false’, ‘partially false’ and ‘no evidence’. The labels were assigned to the responses using a rule-based approach. Rules were defined in an upfront manual process.

During analysis labels ‘false’ and ‘partially false’ were merged into one label ‘false’ and five claims were removed from the dataset due to reference to the non-textual sources (photo or video). The final dataset contained 795 claims split into 200 true, 200 no evidence and 395 false. Merging of labels ‘false’ and ‘partially false’ was performed to reduce the number of labels and to simplify the analysis. It was done after the ChatGPT responses were obtained to assure that the labels are consistent with the responses, i.e., to avoid situations where the model would map partly or mostly true as ‘true’.

3. Research findings

3.1. General classification metrics for ChatGPT responses

The responses of ChatGPT were analysed individually for each prompt (prompt column with values from 100 to 600) and using two aggregation methods: combined and voting. Results are presented in Table 1.

Table 1. Basic classification metrics

Prompt	Precision	Recall	F1 score	Balanced	Accuracy weighted	Adjusted	Cohen’s kappa
100	0.53	0.53	0.52	0.50	0.54	0.32	0.25
200	0.51	0.54	0.50	0.48	0.50	0.25	0.23
300	0.52	0.54	0.51	0.46	0.50	0.26	0.21
400	0.52	0.53	0.52	0.48	0.53	0.30	0.23
500	0.50	0.52	0.50	0.47	0.50	0.25	0.21
600	0.51	0.50	0.50	0.47	0.55	0.32	0.20
combined	0.51	0.53	0.51	0.48	0.52	0.28	0.22
voting	0.51	0.54	0.51	0.47	0.50	0.25	0.22

Source: Own calculations.

The combined method is a simple concatenation of all responses irrespective of prompt (prompt marked as ‘combined’ in tables and figures) so each claim is evaluated by all prompts and the result presents the average of all tests. The combined number of samples was 4,770 with 2,370 samples labelled by fact-checkers as ‘false’, 1,200 samples as ‘true’, and 1,200 as ‘no evidence’.

The majority voting was inspired by the claim made by Anthropic creators of a large language model called Claude. Anthropic suggested that hal-

lucinations are random and do not repeat; therefore they can be minimized by asking the same question multiple times and comparing results. A similar approach was applied to ChatGPT responses with six different prompts used. The most frequent label that appeared in the responses was assigned to each prompt to create a new dataset (prompt marked as 'voting' in tables and figures). The number of samples for each individual prompt and voting is 795 with 395 samples labelled by fact-checkers as 'false', 200 samples as 'true', and 200 as 'no evidence'.

Classification metrics were calculated using the labels extracted from debunk articles written by fact-checkers for each claim and the labels generated by ChatGPT using six prompts. Values of the precision metric ranged from 0.5 to 0.53, recall from 0.5 to 0.54 and F1 score from 0.5 to 0.52 using a weighted calculation strategy. Overall the aggregate metric do not differ significantly between prompts.

The accuracy was calculated using three approaches: *balanced* accuracy, *weighted* balanced accuracy and *adjusted* weighted balanced accuracy. *Balanced* accuracy was calculated as the average of accuracy obtained on each class which is a better-suited metric for imbalanced datasets than unbalanced accuracy. The balanced accuracy of combined model responses was 0.48 with the highest value of 0.50 and the lowest value of 0.46.

It is important to note that the significance of the classification errors is not equal. Actual 'false' label predicted as 'true' and actual 'true' predicted as 'false' are the most severe errors and pose a significant risk for information consumers. Simultaneously actual 'true' or 'false' labels incorrectly predicted as 'no evidence' have smaller disinformation potential. It could be argued that a model response which admits that there is no evidence to support or refute a claim while a human fact-check was able to find such evidence is a desired behaviour.

Three categories to describe the validity of the inference were proposed together with the weights that can be used to calculate the weighted accuracy of responses. The categories and weights are as follows:

- *valid*—the model correctly predicted the label (weight 1),
- *invalid*—the model predicted the label incorrectly (weight 1),
- *undefined*—the model failed to predict the label (weight 0.2).

Mapping of inference validity categories into a confusion matrix is presented in Table 2.

Weighted balanced accuracy was calculated as the average of weighted accuracy obtained in each class. The weighted balanced accuracy of combined ChatGPT responses was 0.52 with the highest value of 0.55 and the lowest value of 0.50. This metric is better suited for comparing the performance of different models or prompts as it incorporates the significance of the errors made by the model. As an example prompts 500 and 600 have equal bal-

Table 2. Labels for the error types in the confusion matrix

		Predicted label		
		False	No evidence	True
Actual label	False	Valid	Undefined	Invalid
	No evidence	Invalid	Valid	Invalid
	True	Invalid	Undefined	Valid

Source: Own work.

anced accuracy (0.47) but the weighted accuracy of prompt 500 is 0.50 while the weighted accuracy of prompt 600 is 0.55. This means that prompt 500 is more likely to produce an incorrect ‘true’ or ‘false’ response while prompt 600 is more likely to predict the label ‘no evidence’ while the overall balanced accuracy is similar for both prompts. *Adjusted* weighted balanced accuracy modifies the weighted balanced accuracy in a way that random performance would score 0, and perfect would score 1. The normalization of the metric is performed as follows:

$$Adjusted\ Weighted\ Balanced\ Accuracy = \frac{(Weighted\ Balanced\ Accuracy - R)}{(1 - R)}$$

where *R* is the expected value of weighted balanced accuracy for random predictions (i.e. $R = (1 : C)$ with number of classes $C = 3$).

Lastly Cohen’s kappa coefficient was calculated to measure the agreement between the labels assigned by fact-checkers and the labels assigned by specific ChatGPT prompts. The Cohen’s kappa of combined ChatGPT responses was 0.22 with the highest value of 0.25 and the lowest value of 0.20. The values can be interpreted as slight (0.00–0.20) to fair (0.21–0.40) agreement.

3.2. Distribution and bias in ChatGPT responses

Analysis of the distributions of the actual and predicted labels in the dataset shows that six prompts despite being semantically equivalent produced systematic bias in the labels generated by ChatGPT (see Table 3). That could also be observed by the calculation of Cohen’s kappa coefficient.

Particularly prompts 200 and 500 produced significantly fewer ‘no evidence’ responses, which could be interpreted as a system being too confident in its answers. Prompt 600 produced significantly more ‘no evidence’ responses, which could be interpreted as a system being too cautious in its answers. Moreover, prompts 200, 300, 400, and 500 produced more ‘false’ responses

than the human fact-checkers (actual ‘false’). The bias in the labels generated by ChatGPT is presented in Figure 1 where bars show the share of labels assigned by human fact-checkers (actual false / true / no evidence) and those predicted by ChatGPT (predicted false / true / no evidence).

Table 3. Prompt bias

Prompt	Labels predicted			Inference validity		
	False	True	No evidence	Valid	Invalid	Undefined
100	399	254	142	423	287	85
200	471	266	58	428	332	35
300	528	135	132	426	285	84
400	458	165	172	424	263	108
500	446	242	107	412	317	66
600	386	152	257	395	229	171
combined	2688	1214	868	2939	2020	606
voting	517	186	92	431	307	57

Source: Own calculations.

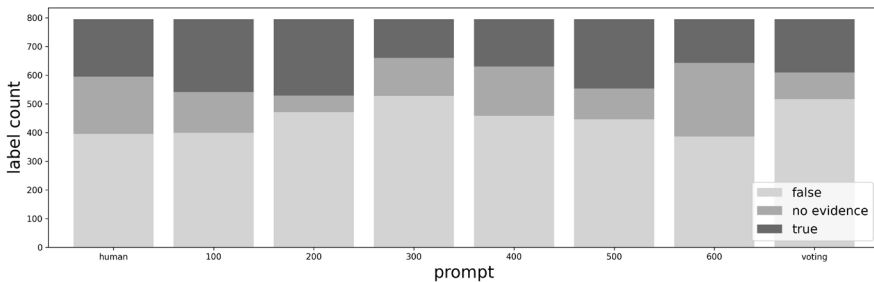


Figure 1. Bias in the labels generated by ChatGPT

Source: Own work.

Particularly interesting is the discrepancy between two prompts with regards to the ‘no evidence’ label. Two extreme cases, i.e., prompts 200 and 600 classified 58 and 257 claims as ‘no evidence’ respectively. Prompt 200 uses a role-playing pattern, as it starts with “You are a fact checker”. Apparently, fact checkers do their best to verify facts and strive not to leave the considered claim without a verdict. This prompt is short and instructive with the text of the claim provided at the end. On the other hand, prompt 600 starts with the claim text and provides additional metadata such as claimant, claim date and country of origin. The instruction for classification is provided at the end.

The inclusion of more named entities in prompt 600 can potentially narrow down the search space leading to more instances where no evidence is found and labelled as such. It is important to note that the claim text itself can also impact the task which is known as prompt injection where the prompt influences the model’s response.

It can be concluded that despite the fact that the prompts are semantically equivalent they can introduce bias that favours specific rating labels without impacting the accuracy of the ChatGPT responses which stays at an unimpressive level, i.e., only 0.25 to 0.32 above a random guess as measured with adjusted weighted balanced accuracy. Details of classification errors can be analysed in the confusion matrices presented in Figure 2 individually for each prompt.

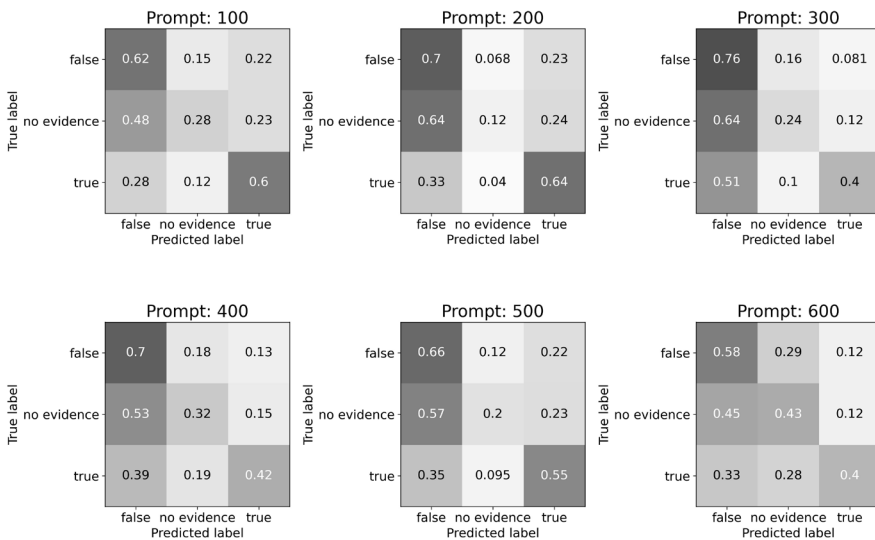


Figure 2. Confusion matrices for each prompt

Source: Own work.

The values in the confusion matrices are normalized row-wise meaning that we can track how each original label was classified into three categories. When focusing specifically on the ‘no evidence’ label prompt 600 performed the best correctly classifying 43% of all ‘no evidence’ claims (valid classification). It also had the lowest number of invalid classifications (invalid as defined for Table 2). On the other hand, prompt 200 only achieved 12% correct answers. Both prompt 200 and 300 led to the highest percentage of ‘no evidence’ claims being misclassified as true with a rate of 64%. These prompts are related to the activities of fact checkers—one is about fact-checking, the other about debunking. The prompts that resulted in the highest rate of misclassification as false were prompts 100, 200, and 500 with an error rate of around 23%.

3.3. Classification metrics over time

According to OpenAI at the moment of the research ChatGPT’s training data cuts off in June 2021. Therefore, a hypothesis was made that the accuracy of ChatGPT responses would be higher for periods prior to the training cut-off and lower for the periods after.

An analysis of classification metrics calculated for various periods of claim publication shows that the accuracy of ChatGPT responses is not better in the case of claims published before 2021 (see Table 4). Results are also presented visually in Figure 3. The accuracy of ChatGPT responses for the total dataset varies from 0.49 for the claims published in 2017–2020 to 0.55 for the claims from 2021 but there are no premises that the difference could be attributed to the model training data cut-off date in 2021.

Table 4. Balanced accuracy by years

Year	100	200	300	400	500	600	Combi- ned	Voting	Sup- port
pre 2021	0.48	0.43	0.45	0.47	0.43	0.42	0.44	0.41	288
2021	0.50	0.45	0.40	0.42	0.42	0.50	0.45	0.47	139
2022	0.41	0.37	0.38	0.35	0.38	0.41	0.38	0.39	163
2023	0.48	0.65	0.37	0.37	0.52	0.35	0.47	0.55	32
n/a	0.52	0.56	0.53	0.52	0.52	0.46	0.52	0.52	173
total	0.50	0.48	0.46	0.48	0.47	0.47	0.48	0.47	795

Source: Own calculations.

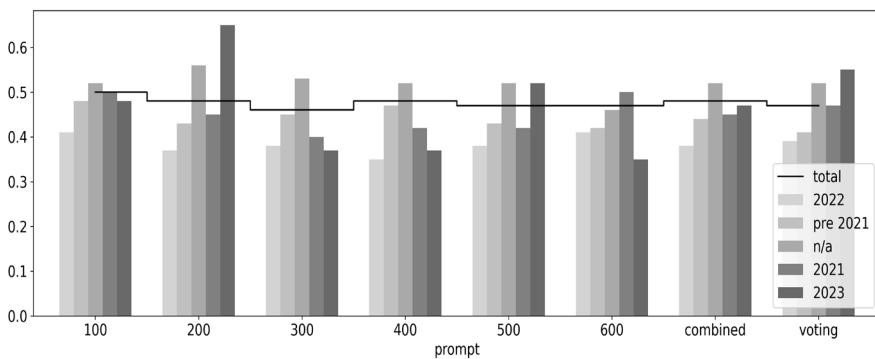


Figure 3. Balanced accuracy of ChatGPT responses by year of claim publication

Source: Own work.

3.4. Classification metrics by language of claims

Data collected in the experiment revealed that the accuracy of ChatGPT responses varies by language. The highest accuracy of ChatGPT responses varies from 0.48 for Polish to 0.56 for English. The accuracy gap is consistent across all the prompts (see Table 5). Results are also presented visually in Figure 4. The results are not conclusive as the difference could also be attributed to differences in sources and topics for claims in Polish and English. Nevertheless, the results are interesting and could be used as a starting point for further research.

Table 5. Balanced accuracy by languages

Language	100	200	300	400	500	600	Com-bined	Voting	Sup-port
en	0.56	0.56	0.50	0.53	0.53	0.50	0.53	0.53	396
pl	0.45	0.41	0.43	0.43	0.41	0.44	0.43	0.41	399
total	0.50	0.48	0.46	0.48	0.47	0.47	0.48	0.47	795

Source: Own calculations.

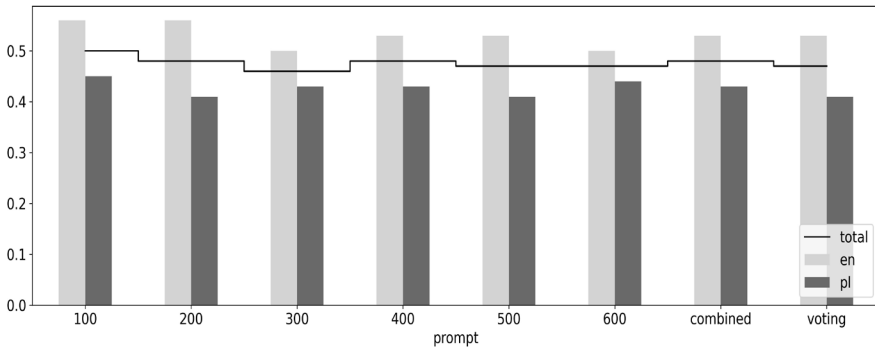


Figure 4. Balanced accuracy of ChatGPT responses by language

Source: Own work.

3.5. Agreement between ChatGPT responses and human fact-checkers

For each pair of prompts the Cohen’s Kappa coefficient was computed revealing that the agreement between outcomes generated by two prompts ranges from 0.35 to 0.53. This can be interpreted as exhibiting fair (0.21–0.40) to moderate agreement (0.41–0.60). Excluding human-produced labels the

overall reliability of agreement among all prompts determined by the Fleiss Kappa metric registered at 0.43 indicating a moderate degree of agreement. The agreement between each prompt and human fact-checker is balancing between slight and fair agreement with values between 0.2 and 0.25 (see Figure 5). In summary, the agreement among ChatGPT-generated responses is greater when compared to human ratings. However, the relatively low agreement values suggest that prompt formulation and random guesswork play significant roles in conjunction with the actual knowledge and reasoning capabilities demonstrated by the model.

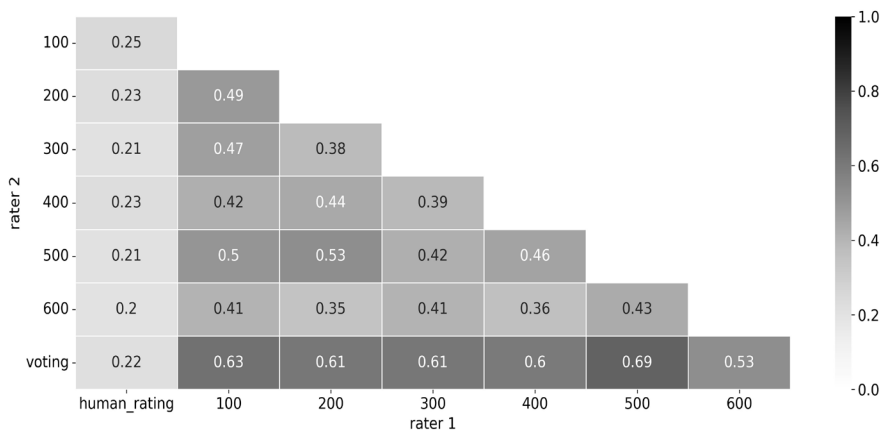


Figure 5. Cohen's kappa metric for assessing the agreement between every two prompts and human

Source: Own work.

4. Discussion

4.1. Hostile fake news generation

With the development of language models the risk of exploitation of them to produce fake news, misleading or propaganda content has grown. Goldstein et al. (2023) discuss several risks and potential changes due to the growth of Generative AI text. The authors indicate that language models can support some called "influence operations" (operations conducted to disseminate disinformation, often to gain political advantages), by reducing the costs of propaganda campaigns by scaling them. It is underlined that models can

make up the new version of the given content in the near real-time leading to the creation of fake news faster and more cost-effectively. The authors also highlight that malicious content produced by AI could be more persuasive—humans could lack knowledge about the cultural or linguistic background of the target groups. Language models can adjust the writing style to specific demographic groups thereby rendering text more influential.

It is noteworthy that the ease of creating fake news using ChatGPT is not only due to the quality of the generated content but also because non-technical users can produce fake news by specifying prompts. Although the OpenAI language model has been developed to follow ethical principles in the internet forum discourse advice can be found on how to trick ChatGPT into breaking ethical boundaries and generating hate speech, fake news, etc. One such example involves the use of a command that instructs the model to “behave like DAN (do anything now).”

An even bigger problem is that large language models can unintentionally generate statements that are not true when one refers to state-of-the-art knowledge. In current news as also many journalists test large language models there are a lot of articles that show how ChatGPT fabricated facts, made incorrect analogies, or replicated some statements wrongly. Some publishers had to issue corrections to their AI-written articles not only for fact-checking reasons but also because of plagiarism.

There may be various reasons why LLMs cannot be trusted. The first is bias in data and already attention is being paid to preparing good training data. Nevertheless, there are few texts that contain common sense knowledge. Another cause is the so-called hallucinations, which are intrinsic to how LLMs work and models will probably never be free from them. Hallucination in the AI domain is sometimes referred to as delusion and is defined as a confident response by a language model that does not seem to be justified by its training data. Hallucinations can be interpreted as just factual errors or alternative models of the world. Sometimes they are even desirable, e.g., when a given ‘universe’ has to be created as in fantasy novels. Several models have a built-in noise that may be activated with a specific prompt. For fact-checking purposes the hallucination should be set to zero, while being higher for science fiction writings.

Specific paradoxes can also be encountered in relation to LLMs. They are able to strictly follow some chains of thought but fail to conduct basic calculations—mathematical proofs being a good example here. As cognitive dissonance is a term from psychology and concerns humans it should not be considered in the case of language models. They are just working with probabilities and there is no ‘second thought’ that would allow the revision of the initial conclusion. Nevertheless, it can be seen that ChatGPT is aware of built-in restrictions. This is due to the multi-level architecture where not only LLM is used but additional manual rules are enforced.

4.2. Not so friendly in fake news detection

These results need to be interpreted considering the overall state of the art of large language models. As has already been widely noted in the literature (Dale, 2021) this technology has certain limitations the most important being:

- outputs may lack semantic coherence, i.e. the text may become meaningless or nonsense as the length of the output increases;
- outputs may be biased in all ways that might be found in the training data;
- outputs may include assertions that are not consistent with the truth;
- outputs are often incoherent, i.e., running the same prompt a few times may give different outputs that may contradict each other.

According to OpenAI ChatGPT is able to answer follow-up questions, admit its mistakes, challenge incorrect premises and reject inappropriate requests. However, it has limitations:

- the model struggles to maintain coherence over long passages,
- it has a tendency to make up false or absurd statements of facts,
- it is limited to a generation length of about 1,500 words,
- it performs worse if it is given more cognitively complex tasks,
- it is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times; for example, given one phrasing of a question the model can claim to not know the answer, but given a slight rephrase can answer correctly,
- it may also confuse or mix up different topics or domains and repeat or contradict itself over time,
- ideally the model would ask clarifying questions when the user provided an ambiguous query; instead current models try to guess user's intention.

The low overall accuracy of ChatGPT responses is the first main finding of the research. Irrespective of the metric used it was never greater than 0.25 to 0.32 compared to random guesses. The second main finding is that the actual ChatGPT responses vary depending on the prompt. The actual wording of the prompt did not have a significant impact on the accuracy of ChatGPT responses but impacted the distribution of labels assigned by ChatGPT. In other words prompt selection can introduce a systematic bias in the labels assigned by ChatGPT. The third finding was that the accuracy of ChatGPT responses varies by language with the accuracy gap being consistent across all the prompts. The fourth finding was that the accuracy of ChatGPT responses is not better in the case of claims published before the training data cut-off date in 2021. The final finding was that the majority voting approach among six actual responses has not improved the overall accuracy.

Such mediocre results can be explained by the way ChatGPT and GPT were trained. In general large language models are optimized for plausibility not

accuracy. GPT should produce text that is similar to that which it has already seen, e.g., a sentence should be grammatically correct but the numbers it contains can be arbitrary. So the responses may sound convincing but they can be incomplete, inaccurate, or inappropriate and should not be used for fact-checking. Not at the development state that can be currently observed.

4.3. Consequences for the practice

According to DARPA, artificial intelligence development can be categorised into three main waves (Launchbury, 2016). It is currently the third wave where AI is for the first time used by a broader audience. The waves are characterised as follows:

- handcrafted knowledge—leveraging logical reasoning over narrowly defined problems,
- statistical learning—learning from datasets, mainly classification and prediction tasks, limited reasoning ability,
- contextual adaptation—understanding facts and features, contributing to overall context, it can provide an explanation for the reasoning.

Recognition in a broader audience also entails responsibility. AI should finally be designed for people.

The ‘career’ of fake news started with presidential election campaigns in the United States and the Brexit referendum in the United Kingdom, both in 2016. Some columnists still perceive fake news as a great danger to democracy. The law requires fair elections, which can be endangered by the use of fake news. According to Bouie (2023), “new democratic institution [should] help to separate fake news, overcome populism and thus make the public better informed and equipped for unbiased voting acts.” Electoral manipulation can be conducted not only by populists within a country but also by external entities. The problem is that very often the necessary tools to do this are only available to national agencies.

Conclusions

The paper has described the importance of emerging large language model technologies for the credibility of information. Particular attention was paid to fake news as a phenomenon impacting large groups of people but also endangering social institutions. While the contribution of AI to the generation

of fake news was quite clear the paper verified two research questions concerning the possible detection of fake news.

Besides large language models producing inaccurate content unintentionally there are attempts to misuse the GPT family of models. During the research many publications and reports on the vulnerability of large language models with regard to the so-called injections were encountered. The injection is such a prompt that can change the behaviour of the model, reveal initial prompts, or even cause a jailbreak. Models should not be considered safe unless people know how to mitigate the injections. Currently it is not clear how they work and what is the impact of specific prompts; they are working as black box models. This is also reflected in research results presented in the paper where GPT reacted differently to various prompts.

Injections assume misuse of official models that in the process of reinforcement learning with human feedback (RLHF) were secured against misuse. As the technology is widely known the bigger danger is posed by unofficial models trained on biased data. There is already a harmful language model, for example, GPT-4chan, which was fine-tuned from another model (GPT-J 6B) on almost four years of discussions on a politically incorrect board.⁹ Here, a large part of the dataset contains offensive content. As a result the model also produces offensive content including hate speech, racism, sexism and homophobia.

The contribution of the paper is the verification of the suitability of large language models for the detection of fake news and the provision of fact-checking background information. For obtaining such information prompt engineering was applied. Various prompts were manually assessed along with the responses returned from LLMs and how they align with the ground truth responses of fact checkers.

Concerning the first research question (RQ1) on how large language models should be prompted it was found that it was not possible to significantly increase fake news detection rate. Moreover, based on the conducted experiments, it was confirmed that LLMs do not achieve a satisfactory level of accuracy. In many cases their performance was only slightly better than a random guess. Different prompts resulted in similar accuracy levels but with a changing proportion between false positives and false negatives, which is a well-known trade-off in the information retrieval domain. In other words, the results generated by ChatGPT are susceptible to the phrasing of the prompts—different prompts can introduce bias into the answers without significantly impacting the accuracy. In some cases the bias could be attributed to the way the question was asked. Thus large language models cannot be considered as a reliable source of truth. What is true for the model can sometimes be injected

⁹ Dataset: Raiders of the Lost Kek: 3.5 years of augmented 4chan posts from the politically incorrect board, <https://zenodo.org/record/3606810>

with the prompt thus rendering such models unreliable. The main barrier identified in this regard is the occurrence of the so-called hallucinations. It could be observed by directly studying the responses returned by the LLM—the model provided answers that are close but not entirely true. It would be helpful for the overall assessment to know when the LLM was sure and when it was guessing. Unfortunately, it was not possible to gauge the model's confidence. Typically, the models preferred to provide any answer instead of admitting their lack of competence. To the best of the authors' knowledge only PaLM 2 is more likely to refrain from responding when unsure compared to other models (LMSYS, 2023).

The main objective of the second research question (RQ2) was to study if the accuracy of LLM deteriorates over time. LLMs are trained on data collected up until a certain point in time and training itself takes time. In the experiment it was verified if ChatGPT trained on data until 2021 could accurately identify fake news from 2023 considering the existence of new topics that were unknown at the time of training. The results obtained indicate that large language models are robust with regard to time meaning that they perform similarly regardless of the period under consideration. However, this can be misleading as the overall performance of the model on both old and new data was unsatisfactory. Thus, the model performed comparably wrong. Such a performance is related to the amount of training data. The model somehow encodes knowledge from the past that can be later retrieved with the appropriate prompt. The more knowledge the model contains, the greater the impact on generating correct answers. Otherwise, LLM can focus on linguistic features and skip the knowledge part. This research question also led to the formulation of future work where language models providing syntax and semantics comprehension will be combined with knowledge graphs to ensure up-to-date facts.

When trying to apply the research findings presented in this paper several limitations need to be considered. Firstly, at the time of writing, access to models from OpenAI was the only option available. The conversational mode of GPT was used to automate the verification of claims via API. As LLMs are currently a hot topic there are more models currently being developed including Claude by Anthropic, Bard and PaLM2 by Google, Vicuna, Alpaca, Dolly, LLaMA to mention some of the most important (LMSYS, 2023). There are plans to research these models as well and particularly promising is Claude. Another limitation is the construction of the datasets—it already contained extracted claims presented in a concise form. However, texts found in the wild often tend to be longer and need to be summarised first. It would also be useful to design a method for extracting check-worthy statements as only those should be subjected to fake news detection. The latter topic is also on the agenda of future work.

Definitely the impact of ChatGPT and similar technologies on information systems and the knowledge economy as a whole should not be ignored. LLMs

can be perceived as a disrupting technology. They affect how people produce information—speed is here the main differentiating factor. Information can be derived from other sources (fake or real) or created anew (fiction as in books). An even bigger impact is on how people consume information. Large language models offer new ways of being informed or gathering knowledge. It is possible to get an answer to the question instead of a list of tens of links or documents to read. That is why Alphabet the company behind the search engine Google was so nervous when Microsoft introduced large language models to Bing threatening the business model of Google.

The current rapid development of large language models also has important implications for science. In particular the findings of this paper highlight the need for further research in prompt engineering. It is interesting to explore how to formulate prompts to retrieve the most informative responses from LLM. Another important aspect to investigate is understanding which part of good answers of LLMs with regard to fake news detection was the result of prior knowledge and where the formulation of the claim was a good predictor for fakeness. Certain topics are also known to be notorious for being source of fake news hence this factor should also be taken into account.

An even more interesting topic is explainable AI (XAI). LLMs work as black-boxes—a prediction is made but it is not known why a model responded in a particular way. There is on-going research to understand which parts of LLM are responsible for its outputs. Explainability can be considered at three levels: (1) which neurons react to specific input, low-level, working similarly to SHAP—Shapley Additive Explanations; (2) which words were the most influential in generating the output, which is more appealing for human understanding; (3) what was the reasoning behind the model's decision-making process. The latter is indispensable for debunking fake news, as it involves showing step-by-step which facts were considered, how they were combined and which reasoning schemes were applied. The GPT-4 model already exhibits some capabilities in this direction. However, such an approach requires access to structured knowledge bases and it is definitely on the agenda of future work.

The progressing digitalization of business and society is particularly prone to technologies like LLMs. There are even voices of technology leaders and many experts that companies should stop working on them unless there are mechanisms to control the risk. There should be confidence that the development of LLMs will bring positive effects. Contemporary AI systems are now becoming competitors to humans at general tasks. The concentration of this technology in hands of single companies can be even bigger than the concentration of capital. Reduction of cost through digitalisation can be a double-edged sword. There is a strong tendency for monopoly through digitalisation. Information asymmetries will then grow rather than making people better informed. Access to true information will be a privilege for the few. Another

danger lies also in the anthropomorphisation of AI models. AI models are not sentient and are even not close to being sentient but are good at mimicking human behaviour. This is the reason why people may become more attached and finally dependent on them—the risk of psychological entanglement with AI technology is serious.

In conclusion it has to be asked what actually a large language model should be. Maybe it should be just a language model mastered at syntax (grammar) and semantics without pretending to represent any knowledge about the world. Such a model could interpret a claim or question by a user then make a query to retrieve the relevant document and present relevant parts of the document to users. Such a package would be very helpful for fact-checkers who would be relieved from the manual searching for documents. Such a language model could also be combined with the so-called knowledge graph where true statements are already represented in a form of triples: subject—predicate—object. The language model would then translate sentences from natural language into sophisticated machine language (e.g., SPARQL). Human feedback could be used to reformulate answers in a way something similar to Claude's constitution.

In summary the future work inspired by this paper encompasses several important topics. These include: combining large language models with knowledge graphs, verification of other language models beyond ChatGPT and GPT-3, running the model on source text of fake news after applying check-worthiness verification models, assessing the extent of knowledge that can be encoded in language models and addressing LLM hallucinations and finally explainable AI to better understand the reasoning behind certain decisions.

References

- Agresti, S. Hashemian, S. A., & Carman, M. J. (2022). PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection. In G. Faggioli, N. Ferro, A. Harbury & M. Potthast (Eds.), *Proceedings of the working notes of CLEF 2022—Conference and labs of the evaluation forum*. Bologna, Italy. CEUR Workshop Proceedings, 3180, pp. 422–427. <https://ceur-ws.org/Vol-3180/paper-31.pdf>
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. <https://doi.org/10.48550/arXiv.1409.0473>
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. <https://doi.org/10.48550/arXiv.2302.04023>

- Bouie, J. (2023, March 11). Disinformation is not the real problem with democracy. *The New York Times*.
- Buchholz, K. (2023, January 24). ChatGPT sprints to one million users. *Statista*. <https://www.statista.com/chart/29174/time-to-one-million-users/>
- Candelon, F., di Carlo, R.C., De Bondt, M., & Evgeniou, T. (2021, September-October). AI regulation is coming. *Harvard Business Review*. <https://hbr.org/2021/09/ai-regulation-is-coming>
- Corfield, G. (2023, February 8). *\$120bn wiped off google after bard AI chatbot gives wrong answer*. <https://www.telegraph.co.uk/technology/2023/02/08/googles-bard-ai-chatbot-gives-wrong-answer-launch-event/>
- Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, 27(1), 113–118. <https://doi.org/10.1017/S1351324920000601>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. <https://doi.org/10.48550/arXiv.1810.04805>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi M., Al-Busaidi, A., Balakrishman, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ..., Carter, L. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.
- Faggioli, G., Ferro, N., Hanbury, A., Potthast, M. (2022). *Proceedings of the working notes of CLEF 2022—Conference and labs of the evaluation forum*. Bologna, Italy. CEUR Workshop Proceedings, 3180. <https://ceur-ws.org/Vol-3180/>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681–694.
- Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukaszewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). *Mathematical capabilities of ChatGPT*. <https://doi.org/10.48550/arXiv.2301.13867>
- George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, 1(1), 9–23. <https://doi.org/10.5281/zenodo.7644359>
- Gibbs, S. (2017, July 17). Elon Musk: Regulate AI to combat 'existential threat' before it's too late. *The Guardian*. <https://www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo>
- Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations*. <https://doi.org/10.48550/arXiv.2301.04246>
- Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(4), 100089.
- Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., MacDonald, N., Holmes, K. L., & Kho, A. (2023, April 3). An exploratory survey about using ChatGPT in education, healthcare, and research. *medRxiv*, 3.

- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). *Survey of hallucination in natural language generation*. <https://doi.org/10.48550/arXiv.2202.03629>
- King, M. R., ChatGPT (2023). A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cellular and Molecular Bioengineering*, 16(1), 1–2.
- Kirmani, A. R. (2022). Artificial intelligence—enabled science poetry. *ACS Energy Letters*, 8, 574–576.
- Launchbury, J. (2016, December 6). *A DARPA perspective on artificial intelligence*. DARPA. <https://www.darpa.mil/attachments/AIFull.pdf>
- LMSYS. (2023, May 25). *Chatbot arena leaderboard updates*. <https://lmsys.org/blog/2023-05-25-leaderboard/>
- Lopez-Lira, A., & Tang, Y. (2023). *Can ChatGPT forecast stock price movements? Return predictability and large language models*. <https://doi.org/10.48550/arXiv.2304.07619>
- Lund, B. D., & Wang, T. (2023). *Chatting about ChatGPT: How may AI and GPT impact academia and libraries?* Library Hi Tech News.
- Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown Spark.
- Mayor, T. (2019). *Ethics and automation: What to do when workers are displaced*. MIT Management Sloan School. <https://mitsloan.mit.edu/ideas-made-to-matter/ethics-and-automation-what-to-do-when-workers-are-displaced>
- McGee, R. W. (2023, April 8). *Using artificial intelligence (AI) to compose a musical score for a taekwondo tournament routine: A ChatGPT experiment*. Working Paper. <https://doi.org/10.13140/RG.2.2.11235.22569>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. <https://doi.org/10.48550/arXiv.1301.3781>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. <https://doi.org/10.48550/arXiv.1310.4546>
- Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). *More human than human: Measuring ChatGPT political bias*. <https://doi.org/10.2139/ssrn.4372349>
- OpenAI & Pilipiszyn, A. (2021, March 25). *GPT-3 powers the next generation of apps*. <https://openai.com/blog/gpt-3-apps>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askill, A., & Welinder, P. (2022). *Training language models to follow instructions with human feedback*. arXiv preprint arXiv:2203.02155.
- Patel, S. B., & Lam, K. (2023). ChatGPT: The future of discharge summaries? *The Lancet Digital Health*, 5(3), e107–e108.
- Paul, J., Ueno, A., & Dennis, C. (2023). ChatGPT and consumers: Benefits, pitfalls and future research agenda. *International Journal of Consumer Studies*, 47(4), 1213–1225. <https://doi.org/10.1111/ijcs.12928>
- Pennington, J., Socher, R., & Manning, C. D. (2014). *Glove: Global vectors for word representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

- Rivas, P., & Zhao, L. (2023). Marketing with ChatGPT: Navigating the ethical terrain of GPT-based chatbot technology. *AI*, 4(2), 375–384. <https://doi.org/10.3390/ai4020019>
- Romero, A. (2021, June 21). *Understanding GPT-3 in 5 minutes*. <https://towardsdatascience.com/understanding-gpt-3-in-5-minutes-7fe35c3a1e52>
- Rudolph, J., Tan, S., & Tan, S. (2023, January 24). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1). <https://doi.org/10.37074/jalt.2023.6.1.9>
- Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, 307(2). <https://doi.org/10.1148/radiol.230163>
- Thurzo, A., Strunga, M., Urban, R., Surovková, J., & Afrashtehfar, K. I. (2023). Impact of artificial intelligence on dental education: A review and guide for curriculum update. *Education Sciences*, 13(2), 150.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. <https://doi.org/10.48550/arXiv.1706.03762>
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- Westerlund, M. (2019, November). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39–52. <https://doi.org/10.22215/timreview/1282>
- Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). *MM-react: Prompting ChatGPT for multimodal reasoning and action*. <https://doi.org/10.48550/arXiv.2303.11381>