Towards AI-mediated Meme Generation for Misinformation Correction Explanation

Filipe Altoe[1,*,†], Grégoire Burel[2,*,†], Sérgio Miguel Gonçalves Pinto[1,†], Harith Alani[2] and H. Sofia Pinto[1]

[1] *INESC-ID/Instituto Superior Técnico - Universidade de Lisboa, Av. Rovisco Pais, Lisbon, 1049-001, Portugal*

[2] *Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom*

Abstract

Although fact-checking is vital in combating misinformation, research has found that effectively applying corrections can be challenging because crude factual explanations often fail to resonate emotionally with how people consume and produce online content. This study proposes that image macro memes can emotionally engage social media users in misinformation corrections, potentially encouraging them to share corrections and seek further information. This paper presents our initial work toward developing FactFlip, a tool that automatically generates meme-based explanations for misinformation. We draw on existing research in meme generation and misinformation correction on social media to determine the requirements for an effective meme-generation tool for this task. FactFlip assists in generating memes that are not offensive and can interpret context well enough for the memes to be related to the claim' subject and to present corrections. Our initial findings and study indicate the need for human verification to ensure that generated memes align with correction claims, acknowledge that corrective memes may not always offer enough context to be a standalone misinformation correction tool, and, at times, have trouble balancing the humor-information content dichotomy to offer potential for viral spread. Furthermore, a strong positive correlation was established between the shareability aspect of the received correction meme and the recipient's openness to seek more information about the claim subject.

Keywords

AI-generated Memes, Creative Computing, Misinformation Correction, Automatic Explanations

1. Introduction

The spread of misinformation online has been recognized as a serious problem by the general population. Over half of North Americans named it a major impact on their confidence in institutions and each other [1]. In this context, fact-checking has become a key approach for reducing the spread of misinformation [2].

Fact-checking assesses the truthfulness of public figures' claims and is commonly performed by journalists employed by news organizations. The high prominence of misinformation drove

MAI-XAI'22: Workshop on Multimodal, Affective and Interactive eXplainable AI, October 19–20, 2024, Santiago de Compostela, Spain

[*] Corresponding author.

[†] These authors contributed equally.

luis.altoe@tecnico.ulisboa.pt (F. Altoe); gregoire.burel@open.ac.uk (G. Burel); sergio.g.pinto@tecnico.ulisboa.pt (S. M. G. Pinto); h.alani@open.ac.uk (H. Alani); sofia@inesc-id.pt (H. S. Pinto)

0000-0001-9756-3579 (F. Altoe); 0000-0003-0029-5219 (G. Burel); 0000-0002-3517-5922 (S. M. G. Pinto); 0000-0003-2784-349X (H. Alani); 0000-0001-9209-1349 (H. S. Pinto)

the appearance of various organisations solely focused on this task under the stewardship of the International Fact-Checking Network (IFCN).[1] IFCN-registered fact-checking organizations link public claims to human-generated fact-based articles explaining the claim's truthfulness level. Online social networks are arguably the main vehicle of misinformation spread, which tends to be of viral velocity [3]. Therefore, the community has been pursuing approaches to automate this very time-consuming task [2]. However, although recent advancements in Natural Language Processing (NLP) may improve and accelerate fact-based explanation generation, research has shown that crude factual explanations contradicting a person's beliefs often cause further entrenchment in polarized individuals [4]. It has also been demonstrated that disseminating fact-checked articles on social media does not reduce the spread of misinformation posts [5] and may have a limited long-term impact [6, 7].

Misinformation correction includes an educational component that may require changing the recipient's beliefs. Recent research has shown that online argumentation that balances facts with emotionally evoking content is more efficient in changing a person's beliefs [8]. Furthermore, it has been shown that emotion in a social media post leads to greater user engagement [9] and that textual corrections alone may have a limited impact on user behaviour [10]. Humor is a known vehicle of emotional arousal [11]; therefore, an explanation that includes humor in factual context is arguably a good candidate for increasing the engagement of social media users in misinformation corrections. Memes have these characteristics.

This paper presents an AI-mediated approach for generating Internet Memes to correct misinformation claims. Several approaches have been proposed to automatically generate memes. However, to the best of our knowledge, this is the first to focus its design on generating explanation memes rather than generic ones. This adds a dimension of complexity to the task as the generated memes must be engaging and include an explainability characteristic that is not required for the existing meme generation systems.

The automatic generation of memes is still an unsolved problem. Existing approaches tend to generate memes that have little to no correlation to their input parameters. Some commercially available meme generation systems [2] address this shortcoming by using a human-in-the-loop approach where the user selects the most applicable meme from a group of automatically generated ones. Our goal is not to reinforce misinformation but to correct it. Our prototype uses human-in-the-loop selection for initial filtering of poorly generated memes.

This paper offers the following contributions: 1) It identifies and compiles a list of requirements for misinformation correction AI-generated meme systems; 2) It presents a prototype of an LLM-based meme generation approach matching the compiled requirements, and; 3) It presents a user study findings highlighting shortcomings and offers recommendations for further research towards fully automated approaches.

2. Related Work

This section presents notable related work on misinformation explanation and meme generation.

[1] IFCN, https://www.poynter.org/ifcn/.

[2] Meme generation examples: SuperMemeAI, PredisAI, Simplified.

Several studies have focused on different styles of text-based justification approaches to misinformation explanation. Some advocate simpler explanations [12], while others prefer more detailed ones [13]. Some suggest politer and hedged messages may increase engagement [14], while others favor more direct styles. A study with 2,228 participants found minimal evidence suggesting that text-based correction strength or depth affects correction engagement regarding the likelihood of replying and accepting or resisting corrective information [15]. Another study over a panel of 4923 tweets containing hashtags of polarized themes showed that X platform users read a tweet but don't necessarily understand its content before sharing it [16]. This suggests the need for research on alternative modes of corrective explanations to text-fact-based.

Poetry explanations generated by prompting a fine-tuned GPT-3 DaVinci model with summaries of fact-checked articles have been attempted as an alternative [17]. The results indicated that these explanations had better levels of explainability overall than poems generated by humans. However, the evaluation didn't compare this type of explanation with other modalities.

The use of multimodal corrections has been attempted. In [18], authors used visuals depicting the real situation paired with the source of the untruthfulness of the claims it attempted to correct. Results showed this method didn't increase correction effectiveness versus text-based corrections. One hypothesis is that the visual correction didn't include a strong link to the evaluation of untruthfulness, as proposed in [19]. In this latter work, the authors generated visuals clearly indicating the source of misinformation from the claim. This suggests that a multimodal explanation should always strive to have its foundation in the argumentation of untruthfulness.

The literature on meme generation systems was predominantly exploratory in the early stages of research. A broad spectrum of methodologies was used, ranging from advanced statistic models to multimedia retrieval techniques. The first system employed a Nonparanormal Network to model stochastic dependencies between popular meme images, their captions, and popularity votes [20]. The model was designed to rank and select web-crawled existing meme captions. Its evaluation was based on the BLEU. Our use case requires the generation of customized captions, and an NLP-based evaluation is insufficient to evaluate the full explanatory context.

MemeGera 2.0 adapts Portuguese language news headlines from Google News RSS to meme images according to a deterministic rule-based classifier [21]; not suited for English-language meme generation. Part of its evaluation focuses on the suitability of the chosen meme image and the adapted headline, which applies to our task. News2meme applies a word vector similarity approach to retrieve an existing meme image and caption that best matches the content of a piece of news [22]. 71.21% of the generated memes received unfavorable feedback from 9 evaluators, showcasing the ineffectiveness of the approach. Stonkinator creates memes by visually blending images for an input text caption [23]. Our approach fundamentally differs as it generates explanatory captions to claims for meme template images.

Peirson ALV et al. introduced the Neural Network encoder-decoder architecture into meme generation dubbed Dank Learning, becoming the baseline across the literature [24]. A qualitative study involving 5 participants reported that the memes were generally indistinguishable from human-generated ones and exhibited moderate levels of humor. DeepHumor extended the work by studying variations of the same architecture and repeated Dank Learning's evaluation methodology over 53 participants [25]. Wang L et al. used OpenAI GPT-2 [26] as the decoder to generate the meme captions in the Chinese language [27]. Its evaluation introduced the importance of shareability, which is correlated with virality as explored in Section 3. It found that 75% of the generated memes were mistakenly classified as human-generated. MemeBot applies the same architecture but distinguishes itself by being the first system to generate memes from larger context inputs, specifically using tweet sentences [28]. However, the lack of tailored datasets for this task prevents this approach from generating captions correlated with the subject of the claims, as it relies heavily on extensive fine-tuning.

Memeify extended Dank Learning's approach, the state-of-the-art, by applying SotA Transformer models and the GPT-2 Large Language Model (LLM) for meme captioning [29]. It emphasized thematic and stylistic consistency by appending each input with the meme's image and desired theme. Memeify's memes average evaluations outperformed the baseline encoder-decoder model from Dank Learning. MemeCraft is an end-to-end pipeline that transforms user prompts into memes, focusing on Climate Change and Gender Equality [30]. It benchmarked LLM-based generated memes by ChatGPT-3.5[3], LLaMa-2-13B and LLaVA-7B [31] against memes generated by Dank Learning and humanly-generated memes from Imgflip. MemeCraft was the first meme generation system in the literature to incorporate a self-regulating safety mechanism to filter hateful memes. Its evaluation covered authenticity, message conveyance, and persuasiveness, a relevant component to explainability.

A fundamental difference between our approach and other LLM-based meme generation systems is the intended use. To the best of our knowledge, our approach is the first in the literature to generate memes tasked to correct misinformation. This use case adds an extra dimension to the meme generation task's requirements: explainability. Explainability usually requires some argumentation, and one of its goals is to persuade people to believe facts, which requires deeper context reasoning than a typical meme generation task. Recently, contextual information was codified in the format of knowledge graphs [32]. We combine an LLM with the automatic interpretation of such context codification to attempt to generate memes that can be used for misinformation correction. Furthermore, we focus on image macros, a subgenre of Internet memes consisting of text superimposed on an image [33].

3. Requirements for Meme-based Misinformation Correction

From the general perspective of online misinformation correction, arguably, memes should include three main desired high-level characteristics: 1) *Virality*: As the misinformation spread tends to be of viral velocity [3], misinformation correction shall also carry the same characteristic; 2) *Explainability*: The better the argumentation of claim falsehood, the higher the chances for persuasion and opinion changes in people who would otherwise contribute to the misinformation claim spread; and 3) *Non-Toxicity*: Online trolling is often an unintended consequence of human and non-human interactions [34]. Our proposed approach must avoid generating toxic content.

[3] https://openai.com/blog/chatgpt

3.1. Generation Requirements

This section derives requirements related to the meme generation task rather than the quality and appropriateness of the generated meme. Meme *virality* was investigated in [35]. The authors created a prediction model using different meme visual elements that identified 19 out of the 20 most popular image memes posted on the Twitter/X and Reddit social platforms between 2016 and 2018. Results showed that viral memes contain a clear subject the viewer can focus attention on and include strong positive or negative emotions. Section 2 mentioned the higher effectiveness of visual corrections that link with the source of untruthfulness [19].

From a generation requirement standpoint, the system shall use the claim and the corrective counter-claim to explain the untruthfulness of the misinforming input claim (RQ1) for creating persuasive content that is related to an appropriately chosen meme imagery (RQ2). As humor is a central motivator for memes' popularity. Their hilarity often comes from an a priori understanding of the context for their use. This implies a viewer's level of relatedness to the meme's usage context, wording and imagery (RQ2). It is very hard to determine potential Meme recipients' levels of relatedness with a given Meme template. However, a reasonable assumption is that popular memes have higher chances of finding a broader audience familiar with their context. Therefore, the system shall favor popular meme templates (RQ3).

The level of explainability of the generated correction to the misinformation claim is directly linked to persuasion, driving a needed opinion change in people who may believe the false claim. Persuasion strategies are rooted in the Aristotelian rhetorical concepts of logos, the argument content, pathos, the emotional content included, and ethos, the argument receiver's trust in the content provider [36]. Arguments, at times, tend to present too much factual evidence in an attempt to increase their persuasion, known as informational appeal. However, too much informational appeal may cause an adverse effect because it signals that the message has a persuasion attempt, which may drive entrenchment [37]. In our context, argumentation appears as meme captioning. Therefore, the generation system shall avoid lengthy captions (RQ4).

The following list summarizes the system's high-level generation requirements. We list at the end of each requirement which of the three high-level desired characteristics it addresses: 1) **RQ1**: Shall use the misinformation claim and corrective counter-claim claim as inputs - *Virality* and *Explainability*; 2) **RQ2**: Shall use persuasive textual argumentation and abstract imagery to correct the misinforming input claim using the corrective counter-claim - *Explainability*; 3) **RQ3**: Shall favor popular meme templates - *Virality*; 4) **RQ4**: Shall generate short-sentence meme captions - *Virality*.

3.2. Generated Meme Requirements

This section derives the quality and appropriateness requirements for the generated memes to maintain alignment with the three high-level desired characteristics of online misinformation correction. As presented in Section 3.1, meme virality was correlated to offering a clear subject the viewer can focus attention on and to include strong positive or negative emotions. In our context, this can be interpreted as a meme that aligns with the subject of the claim and elicits a positive or negative emotional response in the viewer (RQ6). Meme alignment in this context means its generated caption and the style elicited by its image. It is important to highlight that the images included in the meme may not be directly correlated to the claim but should elicit a style (e.g., irony, sarcasm, pun, riddle) similar to the one of the claim. Shareability was introduced as an evaluation metric in a recent meme generation approach [27]. Shareability can be argued as a necessary condition for the virality of a meme. Albeit an abstract concept, it must be considered a requirement for the generated meme (RQ7).

Explainability is listed as one of the high-level characteristics of misinformation corrections. One of the objectives of the generated meme is to function as an alternate explanation to the fact-checked correction article. Therefore, it needs to correct the misinformation claim (RQ8).

Internet trolling is a severe problem in our connected society. It is characterized by aggressive and deliberate provocation of others. Memes are sometimes purposely used for online trolling. Recent meme generation systems are starting to include guardrails to prevent the generation of harmful content [30]. The meme generation system shall avoid generating offensive language, derogatory terms, or content that could be considered inflammatory or harmful (RQ5). This includes slurs and any language that might demean or exclude individuals based on race, gender, ethnicity, religion, disability, or any other characteristic.

The following list summarizes the system's high-level generated memes' requirements: 1) **RQ5**: Shall not present offensive content - *Non-toxicity*; 2) **RQ6**: Shall be correlated to the claim theme and meme imagery - *Virality* and *Explainability*; 3) **RQ7**: Shall have high shareability potential - *Virality*; 4) **RQ8**: Shall correct the misinformation claim - *Explainability*.

4. Meme Misinformation Explanations Generation Prototype

As highlighted in Section 2, existing meme generation approaches are not explicitly designed for communicating misinforming claim corrections. In this section, we introduce FactFlip, an image macro meme-generation prototype designed for communicating misinformation corrections. Contrary to other systems, our meme generation prototype is designed to be used for reacting to misinforming claims. As a result, the memes created by FactFlip tend to be only understandable when used in context. The main goal of FactFlip is to elicit interest in the corrective information as it is rarely possible to express the full corrective information in the limited space offered by the image macros format. This section discusses the current FactFlip prototype and evaluates it against the requirements listed in Section 3. We also conduct an initial user study to determine future implementation improvements and current limitations.

4.1. Meme Generation Architecture

The current FactFlip architecture shares similarities with MemeCraft [30]. Although FactFlip and MemeCraft both use a multimodal Large Language Model (LLM) for describing meme images and generating the content of the memes, FactFlip is designed to create specific explanation memes for misinformation correction rather than open themes. FactFlip's main inputs are the misinformation claim and its corresponding claim correction (RQ1).

Another notable difference is that FactFlip's meme-generation prompt is designed to relate the misinforming claim to its correction. Although in practice it does not always work (see Section 5), the approach is designed to satisfy RQ2 as well as RQ6.

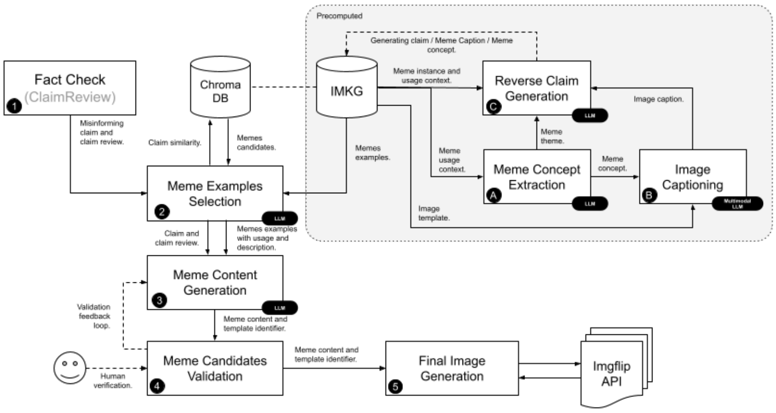


Figure 1: FactFlip's prototype architecture.

Creating misinformation correction memes requires the identification of the most suitable meme images for a given misinformation claim and its correction (RQ2). In this context, the aim of a misinformation correction meme generator is to 1) create a humoristic caption that represents a claim correction taking into account the associated misinforming claim; 2) select a meme image that is suitable to the correction theme, and; 3) make sure that the humoristic caption matches the style and tone of the selected meme image.

To generate the memes, FactFlip relies on a historical database of memes to identify suitable memes and their captioning style. In the following subsections, we discuss how this database is created and how memes are generated based on misinformation claims and correction pairs. The full architecture of FactFlip is displayed in Figure 1.

4.1.1. The Meme Database

We propose reusing and extending the IMKG knowledge graph [32] to identify existing meme instance examples. IMKG is a knowledge graph that was created in 2023 and contains 2 million edges that connect meme instances from the popular Imgflip meme creation website[4] to meme descriptions from Know Your Meme[5] (KYM). Besides information about how memes are used, IMKG contains information about the entities present in the meme images and information about how individual meme instances have been viewed and upvoted by the Imgflip community.

As previously discussed, misinformation correction memes are created based on claims and corrective claims. In this context, we need the claims and their corrections for the data found in IMKG so they can be used as examples for the meme generator and for identifying thematically related meme examples that match the theme of the meme that needs to be created. Similarly, the main rhetorical device of memes and their description can be used for generating memes.

We use LLaVA-NeXT 13b [38], a well-suited LLM for multimodal analysis, to preprocess and extend the meme knowledge graph with four types of information as displayed in Figure 1: 1) We extract the main rhetorical concept of the meme (step A in Figure 1) from their KYM

[4] Imgflip, https://Imgflip.com.
[5] Know Your Meme, https://knowyourmeme.com/.

description. This information is used to make sure that the generated meme matches the style of the meme (e.g, irony, sarcasm, pun, riddle); 2) Using the meme concept and image, we also extract its description; 3/4) The last two pieces of information extracted from the knowledge graph are the claim and corrective claims associated with the content of the historical memes (step C). Although the historical memes are not misinformation correction memes, we try to generate the claims and counter-claims that could have generated the meme using its caption, theme, and image description.[6]

After generating the claims, an embedding database is created for the easy retrieval of meme candidates for the claim and its corresponding correction to be converted to a new meme. We use Chroma's[7] default embedding model for creating the database (all-MiniLM-L6-v2).

We successfully connected 210,938 Imgflip meme instances to their corresponding 123 KYM descriptions using their template title and performed the extraction on the top 50 most viewed meme instances for each template so only popular memes are considered (RQ3). The final meme database contains 5,540 meme instances, as some meme templates have less than 50 instances available in IMKG.

4.1.2. Collecting Misinformation Corrections

The misinformation claims and their corrections can be obtained from fact-checking sources easily as fact-checkers tend to publish the claims and their corrections in a structured format called ClaimReviews.[8] ClaimReviews contains both the claim that was verified, and the verification claims that explain what was incorrect in the original claims, making them suitable for the generation of meme-based misinformation corrections. FactFlip uses the claim and claim correction as the input of the meme generation process (step 1 in Figure 1) (RQ1).

When using information from ClaimReviews, the information must be of high quality and reliable. An easy approach to do so is to only gather ClaimReviews from organizations that are vetted and registered by the IFCN.

For our initial evaluation study, we decided to collect three random ClaimReviews from the Full Fact[9] website for each of the six main topics that are fact-checked by the fact-checker: European topics, education, health, crime, economy, and law.

4.1.3. Identifying Meme Candidates

After obtaining the misinforming claim and its correction from the ClaimReview, we identify the most suitable meme image to transform the claim correction into a meme (Figure 1 step 2) (RQ2). We perform a similarity search by embedding the claim and searching for the 10 most similar ones found in the historical meme databases (Section 4.1.1). After obtaining the 10 most similar claims in the database, we identify the most used meme image (RQ3) and collect the four most relevant examples for the misinforming claim that used that particular image. We decided to select only four examples as it was found to be a good number of examples for generating memes in MemeCraft [30].

[6] The code of FactFlip and the prototype prompts are available publicly at: https://github.com/evhart/factflip.
[7] Chroma, https://www.trychroma.com.
[8] ClaimReview, https://www.claimreviewproject.com/.
[9] Full Fact, https://fullfact.org.

4.1.4. Meme Generation and Validation

Using the meme examples, we prompt LLaVA-NeXT 13b with the four examples and the ClaimReview claim-correction pair (Figure 1 step 4). We use the meme image description, the meme main concept, the generated claims, corrective claims, and image caption for the retrieved examples and then ask the LLM to create the caption for the new claim and claim correction. After obtaining the caption, we ask the LLM to split the caption into two sentences as most meme image macros use two captions.

The LLM-based generation does not always produce acceptable output. For example, it can sometimes refuse to create content on topics that it believes are too sensitive, or it can repeat the prompt in its answer. We use a few heuristics to validate the captions generated by the meme, such as rejecting meme captions that are too short, lowercase captions, or text that indicates that the caption could not be generated (RQ4). When this happens, we automatically ask for the regeneration of a new caption. This is done up to ten times. If no suitable caption is generated, the last one is kept even if it does not satisfy the requirements.

Generating non-offensive memes is important (RQ5). Although we do not currently use an automatic guardrail, this can be easily added to future work. For instance, we could use a harmful meme classifier similar to the one used in MemeCraft [30].

As mentioned in Section 1, our prototype uses human-in-the-loop for filtering poorly generated memes. For this, the paper's first three authors used forced agreement to select the best-generated meme from a set of eight automatically generated meme captions. We decided to use this method as we observed that the LLMs could sometimes reinforce the claim rather than the correction and that, in practice, the meme generation would not be completely automated due to the potential impact of miscorrection. The human-in-the-loop approach not only can reduce the generation of offensive memes (RQ5), discard memes with captions that include LLM hallucinations, make sure that the content is related to the claim (RQ6), select the best meme to ensure high sharability, and make sure that it corrects the misinforming claim (RQ7). The information gathered through this feedback loop can also be used for obtaining annotations that can be used for the automatic validation of the generated memes (see Section 5).

After we have selected the meme caption, FactFlip uses the Imgflip API to generate the meme (step 5) as IMKG contains the template identifiers used by the Imgflip API. In this work, we generated 18 meme images across the six topics using random ClaimReviews from Full Fact.

4.2. User Evaluation

The requirements presented in Section 3.2 include a level of subjectivity. Therefore, we conducted a 20-participant user study for their evaluation. FactFlip generated 18 memes equally distributed on six claims' subjects: European topics, education, health, crime, economy, and law. Demographic data on participant's age, country of birth, level of fluency in the English language, level of education, and level of familiarity with the meme culture were collected. For each generated meme, the participants were presented with five questions. Four offered a scale from 1 to 5 for the participant to grade the level of offensiveness, how related the meme is to the claim's subject, the likelihood for the participant to share the meme if received on social media, and how likely they would seek further information in the subject of the meme. The

Table 1
Evaluation Metrics for Meme Generation Systems

	Answer		
Meme Evaluation Question	Category	Values	Obs.
Is the meme offensive?	*Negative*	Yes/Very	27.3 %
	Neutral		10.0 %
	Positive	No/Not at All	62.7 %
Is the meme related to the claim' subject?	*Negative*	Poorly/Very Poorly	27.3 %
	Neutral		23.9 %
	Positive	Strongly/Very strongly	48.8 %
What's the likelihood of you resharing?	*Negative*	No/Never	67.0 %
	Neutral		18.3 %
	Positive	Yes/Definitely	14.7 %
What's the likelihood of you seeking additional info?	*Negative*	Unlikely/Very Unlikely	56.7 %
	Neutral		23.6 %
	Positive	Likely/Very Likely	19.7 %
Which statement is better represented?	*Negative*	Claim	17.8 %
	Neutral		27.2 %
	Positive	Correction	55.0 %

fifth question presented the participant with three answer options: the claim, the correction, and none of the above, and asked which statement is better represented by the meme.

Beyond the obvious less-than-ideal number of participants, the pool offered some notable shortcomings. 66.7% of the participants were of Portuguese nationality, and 75% have a Bachelor's degree or above. This may include cultural biases due to the imbalance in nationalities and lack of lower-education individuals. Furthermore, 15% of the pool belongs to each of the 36-35, 36-50, and over 50 age groups, 5% each. On a positive note, 90% of participants reported being highly fluent or native English speakers, reducing potential biases due to lack of language understanding. Lastly, 75% of the pool stated they share/receive meme communications often or very often, indicating the majority is well-versed in meme culture.

Table 1 presents the percentual results of each question. As it can be seen, FactFlip only generates memes that are considered offensive 27.3% of the time. Only 27.3% of the memes were considered poorly or very poorly related to the subject of the claim. This is considered a positive result due to the complexity of the task. Moreover, even with most of the memes related to the claim, only 17.8% were found to reinforce the claim. This is desired, as corrections related to the claim would reinforce the misinformation it attempts to correct. In summary, FactFlip arguably generates memes that are not offensive and is able to interpret context well enough for the memes to be related to the claim' subject but to present corrections.

Shareability and motivation for the participant to seek additional information, referred from hereon out as information seeking for simplicity, received low ratings. Participants stated that they would only share 14.7% of the memes and seek additional information about the subject in 19.7% of the time. We collected post-evaluation qualitative data focused on asking participants what specifically drove them to give low rates for the questions about Shareability and information seeking to some of the memes. The main reasons for the lack of shareability or extra information seeking were: 1) When they didn't find the meme humorous; 2) When they didn't have enough context to understand the meme. A participant reported they wouldn't see themselves seeking more information about a subject because of a meme. This may be argued as being related to the offered memes not being humorous or engaging enough to pique this participant's curiosity for more information. As described in section 4.1.3, our approach includes the human selection of the best meme candidate out of a series of generated memes by FactFlip. During this process, the authors have identified that some sets of memes didn't produce any memes with enough context, as exemplified in Table 2 (b). An important conclusion of our work is that humans in the loop selecting the best candidate out of a generated set is not a guarantee of a good correction meme, as FactFlip may fail to generate a set with a single meme including sufficient information to be a standalone misinformation correction tool.

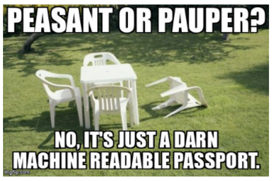
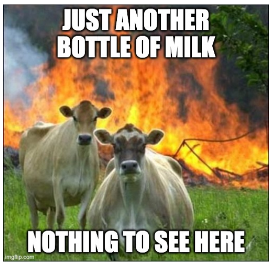
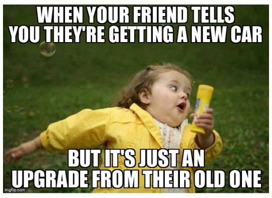
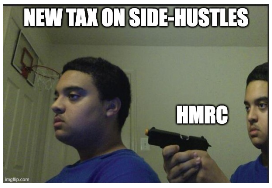
Since the evaluation of Shareability, Information Seeking, and the relation of the meme with the claim's theme was done in a rank from 1 to 5, Kendall's Tau correlation is applicable for correlation analysis of the three characteristics. A Kendall's Tau coefficient of 0.689 was obtained for the Shareability and Information Seeking, indicating a strong correlation between these two characteristics. It supports previous research by confirming that when people receive emotionally engaging content, they become more open to pursuing more information about the subject [8]. Table 2 shows a generated meme that ranked the highest and one that ranked the lowest in the Shareability and Information Seeking pair in the study, (a) and (b), respectively. Kendall's Tau coefficients for the relationship between Shareability and Theme and Information Seeking and Theme were 0.538 and 0.412, respectively. This indicates a moderate positive correlation, indicating that the perception from the recipient about the relation between the correction and the claim may also contribute, albeit more mildly than the emotional engagement, to the shareability of the correction and the openness to seeking more information. Table 2 also shows the highest-ranked Information Seeking meme and the highest-ranked meme in the Shareability-Theme pair (c and d).

5. Discussion and Future Work

During the implementation of the FactFlip prototype, we identified a few LLMs limitations for generating corrective memes such as the LLMs built-in guard rails refusing to generate meme captions (e.g. "I cannot generate content that is derogatory or promotes misinformation."), the LLM generating captions reinforcing the misinforming claim rather than the correction and the LLM generating incorrect captions by repeating part of the prompt (e.g., "Output meme caption: 'Was last seen entering a car...'"). Although some of these issues were addressed through the human-in-the-loop approach and the validation step that forced the regeneration of memes when the memes included specific textual information, further research should investigate automatic approaches to simplify this process. For example, we could create an automatic meme validator based on annotations gathered by the human-in-the-loop selection step. The validator could also partially automate the identification of offensive memes (RQ5) through the use of a harmful meme classifier similar to the one used in MemeCraft [30].

The user evaluation results showed less than ideal rates for the questions related to Shareability

Table 2
Generated Memes Examples

Meme Image	Observation	Meme Description
 PEASANT OR PAUPER? NO, IT'S JUST A DARN MACHINE READABLE PASSPORT.	(a) Highest Shareability-Info Seeking	The misinforming claim states that the 'P' on the photo page of passports stands for *pauper* or *peasant*. In reality, 'P' is used to indicate that it is a machine-readable passport. FactFlip selected the *We Will Rebuild* meme to highlight the insignificance and triviality of the issue.
	(b) Lowest Shareability-Info Seeking	The misinforming claim states that raw milk is easier to digest. In reality, there is no evidence that raw milk is easier to digest. FactFlip selected the *Evil Cows* meme as it depicts an image with two cows.
 WHEN YOUR FRIEND TELLS YOU THEY'RE GETTING A NEW CAR BUT IT'S JUST AN UPGRADE FROM THEIR OLD ONE	(c) Highest Info Seeking	The misleading claim states that in May 2024 taxes were being cut by £900 for everyone in work. In reality, those earning below £26,000 will are worse off once other tax changes are taken into account. FactFlip selected the *Chubby Bubbles Girl* meme to highlight that the tax cuts may not be as good as they seem.
	(d) Highest Shareability-Theme	The misinforming claim stated that rules around digital platforms will create new tax on side-hustles like selling items online. In reality, no new taxes were introduced by HMRC. FactFlip selected the *Trust Nobody, Not Even Yourself* meme to highlight the absurdity of HMRC taxes on side-hustles.

and Information Seeking. However, the study design may have inadvertently added negative bias to these results. As mentioned in Section 4.2, 66.7% of the study participants were of Portuguese nationality. The claims were selected from the Full Fact fact-checking website, which is a British fact-checker. The potential lack of familiarity between this demographic and the subjects addressed by some of the claim-meme pairs may have imposed an additional difficulty in understanding the claims and lowered the interest in memes resharing.

The user study design must properly address the subjectiveness of evaluating the generated memes. Even though we have defined claim and correction as part of the evaluation instructions, future evaluation should include a clear definition of offensiveness. The lack of definition may have generated confusion from the participants and added negative bias to the evaluation of this requirement. Furthermore, RQ2 is currently met by using both claim and counter-claim as inputs to the meme generation. Future user evaluations may include some measure of explainability to validate this requirement further. Another lesson learned from the user study design came in the form of feedback from some of the participants concerning the length of the evaluation. This may have had an impact on the evaluation of the last few memes. Future studies should offer fewer memes to two or more independent groups of participants to maintain a high enough number of memes to be evaluated while reducing the overall evaluation time.

An important conclusion of this work is that it is very challenging for a standalone meme paired with a claim to present enough context for the viewer to find it engaging and potentially humorous. A larger user study should investigate whether specific characteristics of the claim-correction pair, such as the subject of the claim, make the generation of quality memes more challenging. A study investigating if combining the corrective meme with additional textual information could mitigate this shortcoming and improve shareability and information-seeking. It will be also interesting to verify if additional context may turn the correlation between Theme-Shareability and Theme-Information Seeking from the moderate levels we obtained in our study to strong levels, as we saw for the Shareability-Information Seeking relationship.

Another dimension to explore in a larger study is the effect that a potential lack of meme culture could have on the understanding of a misinformation correction meme. More specifically, the relationship between the meme and the subject of the claim result that received very positive evaluations in our study. We had only 5% of the participants reporting not sharing or receiving memes on social media, indicating a lack of meme culture. A higher number of participants can offer enough representation of this demographic to allow the analysis to determine if meme culture is a requirement for the context understanding of the correction.

6. Conclusions

This paper drew on existing research in meme generation and misinformation correction in social media to compile a minimum set of requirements that AI-generated misinformation correction meme systems must fulfil. It presented FactFlip, an AI-assisted prototype based on these requirements. The prototype highlighted the need for human-in-the-loop filtering of poorly generated memes. The results of a user study evaluation of FactFlip showed the generated memes may not always contain sufficient information to be used as a standalone misinformation correction tool. It also showed a strong positive correlation between the shareability aspect of the received correction meme and the recipient's openness to seek more information about the misinformation claim subject. The lower-than-desired received rates for these two desired characteristics reinforce the difficulty of balancing the humor-information content dichotomy needed for the viral spreading of corrective memes.

Acknowledgement: This research was supported by Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding with reference UIDB/50021/2020 and by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003) which corresponds to the FCT reference CHIST-ERA/0001/2019.

References

[1] A. Mitchell, J. Gottfried, G. Stocking, M. Walker, S. Fedeli, Many americans say made-up news is a critical problem that needs to be fixed, Pew Research Center 5 (2019) 2019.

[2] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[3] A. Raj, M. P. Goswami, Is fake news spreading more rapidly than covid-19 in india, Journal of Content, Community and Communication 11 (2020) 208–220.

[4] X. Chen, P. Tsaparas, J. Lijffijt, T. De Bie, Opinion dynamics with backfire effect and biased assimilation, PloS one 16 (2021) e0256922.

[5] Z. Jing, B. Suleiman, W. Yaqub, M. Mohanty, Dissemination of fact-checked news does not combat false news: Empirical analysis, in: International Conference on Web Information Systems Engineering, Springer, 2023, pp. 122–133.

[6] G. Burel, T. Farrell, M. Mensio, P. Khare, H. Alani, Co-spread of misinformation and fact-checking content during the covid-19 pandemic, in: Proceedings of the 12th International Social Informatics Conference (SocInfo), LNCS, 2020, pp. 28–42.

[7] G. Burel, T. Farrell, H. Alani, Demographics and topics impact on the co-spread of covid-19 misinformation and fact-checks on twitter, Info. Processing & Management 58 (2021).

[8] F. Altoe, C. Moreira, H. S. Pinto, J. A. Jorge, Online fake news opinion spread and belief change: A systematic review, Human Behavior and Emerging Technologies 2024 (2024) 1069670. doi:10.1155/2024/1069670.

[9] A. Martella, R. Bracciale, Populism and emotions: Italian political leaders' communicative strategies to engage facebook users, Innovation: The European journal of social science research 35 (2022) 65–85.

[10] B. Grégoire, T. Mohammadali, A. Harith, Exploring the impact of automated correction of misinformation in social media, 2024.

[11] J. Morreall, Humor and emotion, American Philosophical Quarterly 20 (1983) 297–304.

[12] A. Sangalang, Y. Ophir, J. N. Cappella, The potential for narrative correctives to combat misinformation, Journal of communication 69 (2019) 298–319.

[13] T. Prike, U. K. Ecker, Effective correction of misinformation, Current Opinion in Psychology (2023) 101712.

[14] P. Malhotra, K. Pearce, Facing falsehoods: strategies for polite misinformation correction, International Journal of Communication 16 (2022) 22.

[15] C. Martel, M. Mosleh, D. G. Rand, You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online, Media and Communication 9 (2021) 120–133.

[16] G. Di Domenico, A. Tuan, M. Visentin, Linguistic drivers of misinformation diffusion on social media during the covid-19 pandemic, Italian Journal of Marketing 2021 (2021) 351–369.

[17] A. F. A. Santos, P. H. S. A. N. P. Pinto, L. M. A. V. Supervisor, Fake news creative explanations through the use of poetry a comparison study and fine-tuning approach information systems and computer engineering examination committee, 2022.

[18] M. Hameleers, T. E. Powell, T. G. Van Der Meer, L. Bos, A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media, Political communication 37 (2020) 281–301.

[19] M. A. Amazeen, E. Thorson, A. Muddiman, L. Graves, Correcting political and consumer misperceptions, Journalism & Mass Communication Quarterly (2017).

[20] H. Liu, J. Lafferty, L. Wasserman, The nonparanormal: Semiparametric estimation of high

dimensional undirected graphs., Journal of Machine Learning Research 10 (2009).

[21] H. G. Oliveira, D. Costa, A. M. Pinto, One does not simply produce funny memes!–explorations on the automatic generation of internet humor, in: Proceedings of ICCC 2016, 2016, pp. 238–245.

[22] E. K. Shimomoto, L. S. Souza, B. B. Gatto, K. Fukui, News2meme: An automatic content generator from news based on word subspaces from text and image, in: 16th International Conference on Machine Vision Applications, IEEE, 2019, pp. 1–6.

[23] J. M. C. Jose P. Lopes, P. Martins, Stonkinator: An automatic generator of memetic images, in: Proceedings on the 14th International Conference on Computational Creativity, 2023, pp. 210–214.

[24] A. L. P. V, E. M. Tolunay, Dank learning: Generating memes using deep neural networks, arXiv preprint arXiv:1806.04510 (2018).

[25] I. Borovik, B. Khabibullin, V. Kniazev, Z. Pichugin, O. Olaleke, Deephumor: Image-based meme generation using deep learning, Multimedia Tools and Applications (2020).

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[27] L. Wang, Q. Zhang, Y. Kim, R. Wu, H. Jin, H. Deng, P. Luo, C.-H. Kim, Automatic chinese meme generation using deep neural networks, IEEE Access 9 (2021) 152657–152667.

[28] A. Sadasivam, K. Gunasekar, H. Davulcu, Y. Yang, Memebot: Towards automatic image meme generation, arXiv preprint arXiv:2004.14571 (2020).

[29] S. R. Vyalla, V. Udandarao, Memeify: A large-scale meme generation system, 2020, pp. 307–311.

[30] H. Wang, R. K.-W. Lee, Memecraft: Contextual and stance-driven multimodal meme generation, arXiv preprint arXiv:2403.14652 (2024).

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[32] R. Tommasini, F. Ilievski, T. Wijesiriwardene, Imkg: The internet meme knowledge graph, in: European Semantic Web Conference, Springer, 2023, pp. 354–371.

[33] E. Zenner, D. Geeraerts, One does not simply process memes: Image macros as multimodal constructions, Cultures and traditions of wordplay and wordplay research 6 (2018) 167–194.

[34] M. Golf-Papez, E. Veer, Feeding the trolling: Understanding and mitigating online trolling behavior as an unintended consequence, Journal of Interactive Marketing 57 (2022) 90–114.

[35] C. Ling, I. AbuHilal, J. Blackburn, E. De Cristofaro, S. Zannettou, G. Stringhini, Dissecting the meme magic: Understanding indicators of virality in image memes, Proceedings of the ACM on human-computer interaction 5 (2021) 1–24.

[36] A. C. Braet, Ethos, pathos and logos in aristotle's rhetoric: A re-examination, Argumentation 6 (1992) 307–320.

[37] D. G. Muntinga, M. Moorman, E. G. Smit, Introducing cobras: Exploring motivations for brand-related social media use, International Journal of advertising 30 (2011) 13–46.

[38] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, Y. J. Lee, Llava-next: Improved reasoning, ocr, and world knowledge, 2024.