MemeFact: Fact-Checking News with Memes

PIC2 - Master in Computer Science and Engineering
Instituto Superior Técnico, Universidade de Lisboa

Sérgio Miguel Gonçalves Pinto — 93614*
sergio.g.pinto@tecnico.ulisboa.pt

Advisor: Helena Sofia Pinto
Co-advisor: Daniel Gonçalves

Abstract Social Networks allowed for the rapid and widespread proliferation of misinformation online, making fact-checking services essential to such type of content. Yet, research in fact-checking shows that current methods only identify a small fraction of fake news and flag content well after it has reached peak dissemination. Additionally, traditional textual factual explanations fail to engage a broad audience, as the general public seldom reads them. In response to findings that emotion-based explanations more effectively alter beliefs, this thesis introduces a novel fact-checking approach using memes, leveraging their concise, humorous nature and viral potential. A comprehensive literature review highlights a notable gap. This review supports integrating state-of-the-art (SOTA) large language models (LLMs) enhanced by multimodal pre-training and contextual data augmentation to improve relevance and coherence into the generated content. From this review, we derive critical requirements for MemeFact, an AI-mediated system that automatically generates memes to explain the verdict of online claims based on their textual rationale. This study proposes to study a Multimodal Debate Mechanism with a fine-tuned judge model or a single fine-tuned LLM enhanced with Reinforcement Learning from Human Feedback (RLHF) for meme captioning. It plans to select the LLMs and respective learning techniques by evaluating the generated memes through a user survey and further assess the practicality of meme-based fact-checking with an additional user study. This research aims to make a significant societal impact by addressing the issue of misinformation and encouraging further research on creative approaches to fact-checking.

Keywords — Memes, Fact-Checking, Fake News, Misinformation, Creative Explanations

*I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa (https://nape.tecnico.ulisboa.pt/en/apoio-ao-estudante/documentos-importantes/regulamentos-da-universidade-de-lisboa/).

Contents

1 Introduction

Twenty centuries ago, news would take an estimated 20 days to travel between Olisipo and Rome [1]; nowadays, a mobile phone allows your horse to travel near the speed of light. While the fast pace of news propagation dramatically increased our access to knowledge, it has not granted us omniscience. This limitation is partly due to deception tactics that have been used throughout human history. Three thousand three hundred years ago, Pharaoh Ramesses 2nd returned home a hero to its people for allegedly beating the Hittites single-handedly after being ambushed and abandoned by his army in the Battle of Kadesh—in reality, it was a humiliating draw that would have cost his position [2]. In the 18th century, the Catholic Church's misleading attribution of supernatural causes to the 1755 Lisbon earthquake backfired, sparking the Enlightenment era [3]. A century later, sensationalist journalism pushed the war between US and Spain over Cuba [4]. Ironically, the Byzantine Generals Problem, a classic in Computer Science, illustrates this timeless dilemma: ensuring the desired consensus among users through unreliable communication channels [5].

Fast forward to the 21st century, technological advances such as the internet and social networks disrupted the news ecosystem by reducing the constraints on news dissemination, compromising the journalistic norms of objectivity, ethics, and balance—obtained to end 20th-century propaganda. These conditions were optimal for the rapid and widespread proliferation of misinformation—"false or misleading information"—and disinformation—"false information that is purposely spread to deceive people"—which now manifest with a modern rebranding: Fake News— "fabricated information that mimics news media content in form but not in organizational process or intent" [6].

1.1 Why is it interesting and important?

The significance of this issue resides in its worldwide transversal ongoing impact. At the heart of Democracy is the act of voting, which is meaningful if public deliberation is grounded in equitable distribution of information free from interference or manipulation [7]. Starting with the 2016 US elections: 25% of the news contained by tweets five months preceding the election were fake or extremely biased [8]; 1 in 4 Americans visited a fake news website from October 7th to November 14th, with Trump supporters being the majority of the visitors [9]; on Facebook, fabricated news—such as "Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement" [1]—favoring Trump outnumbered those favoring Clinton by 22 million shares. The same happened for the 2017 French Presidential Election [10]: on X (formerly Twitter), bot accounts used to support Trump resurged from inactivity to join the MacronLeaks disinformation campaign from a total of 17 million tweets published from April 27th to the election day.

Concerning the COVID-19 pandemic, the misinformation around it has had clear public health implications: individuals with higher susceptibility to false information were more likely to be anti-vaccination and/or reluctant to take the vaccine and follow suggested health guidelines [11, 12]. In more detail, there was a decline in intent to definitely take the vaccine by 6.2 percentage points in the UK and 6.4 pp. in the US [13]. Furthermore, it was found that psychological disorders, panic, fear, depression, and fatigue are not only consequences of the disinformation spread [14] but also factors that contribute to its proliferation [15]. Additionally, the type of media through which people consume news affect their susceptibility to misinformation and conspiracy beliefs. Specifically, traditional media such as the newspapers, TV, and radio tend to mitigate the prevalence of such beliefs while digital media and political figures amplify them. However, despite their role in informing the public, traditional media outlets have experienced a significant decline in trust. In 2022, a survey found that, within the US, 38% of respondents expressed having *None at all*, 28% *Not very much*, and 34% a *Great deal of* trust in them [16]. Thus, it is important to understand the factors contributing to its proliferation to counter it effectively.

1.2 Why is it hard?

The complexity of this issue is largely due to the inherent human biases such as selective exposure and confirmation bias. Additionally, how social media algorithms are built, and the overwhelming volume of content that

[1] https://shorturl.at/mtRT0

flows in social media platforms aggravated by the armies of automated bots, sophisticated with AI generative capabilities, that are used by counterintelligence organizations seeking to destabilize the international scene when countries are at war. About human biases, people struggle to distinguish between truth from falsehood when they do not apply critical reasoning or when their prior beliefs are inaccurate. Headlines that evoke strong emotions are more likely to be believed regardless of their truthfulness and the likelihood of people sharing misinformation is higher than their ability to judge its truthfulness [17].

With more than half of Americans saying they get their news from social media *Often* or *Sometimes*, platforms like Facebook, X, and Google play an important role [18]. Their business models use complex statistics to maximize engagement, which they monetize through advertising [19]. Inadvertently, they promote selective exposure—a human bias that makes people prefer information that confirms their preexistent beliefs or knowledge. Bakshy et al. confirmed this by finding that algorithmic filtering on Facebook contributed to users finding about 15% less ideologically diverse content and engaging with 70% less of such content when it was presented [20].

To conclude, it is important to address the pivotal role of social bots—accounts programmed to mimic human behavior on social media platforms. Currently, accounts spreading misinformation are more likely to be social bots than genuine users [21]. They aim to disseminate fake news during incubation, adopting tactics such as tagging influential users. This strategy has been effective, as humans interact with content from bots almost as likely as humans. Until now, we have provided a broader context of the problem to contextualize the reader. We will now turn our attention to the existing solutions.

1.3 What is wrong with previous proposed solutions? How does ours differ?

There are two categories of intervention for the fake news problem: empowering individuals to discern fake news and implementing structural changes to limit exposure. Our focus lies on the former, primarily through fact-checking, defined "as the assignment of a truth value to a claim made in a particular context" [6, 22]. We refine this definition to the *"process of evaluating the accuracy of a claim made in a particular context, accompanied by an explanation detailing the rationale behind its verdict"*, that we refer to as Textual Fact-Checking.

Fact-checking—given the volume of fake news—faces scalability and slowness issues—most false claims are not flagged, and when they are, it is late to prevent their peak viral spreading [17]. Additionally, it suffers from the implied truth effect—headlines without warning tags are perceived as more credible or accurate, simply because they have not been explicitly marked as false [23]. Its effectiveness in improving accuracy is short-term [24], and repetition of false headlines increases perceived accuracy even on disputed claims [25].

The other solution, textual fact-checking, differs from the previous by including textual explanations for the verdict. Similarly, it faces a lot of challenges that limit its effectiveness. First of all, it does not reach the intended audience, only 2.7% of respondents search for fact-checking articles verifying fake news from visited untrustworthy websites [9]. It improves factual knowledge of voters but does not change beliefs and the less factual the priors from voters the more false claims deviate them from the truth [26]. Furthermore, the effectiveness weakens when verdicts are presented on a spectrum of truthfulness, and most importantly, with complex explanations full of jargon. Also, its effectiveness is lower among less politically knowledgeable people, and it is dependent on people's ideology [27]. At last, people do not read long articles and textual fact-checking has more impact than to traditional fact-checking due to its higher sharing intentions [28]. Despite these challenges, earlier concerns in the literature about fact-checking causing the backfire effect have been debunked, with recent research indicating the prevalence of this phenomenon to be rare. The backfire effect is the phenomenon where exposure to a false claim followed by a correction makes a respondent less accurate than when solely exposed to a false claim [29].

There is much research on these two solutions that will not be mentioned for brevity reasons—we mentioned mainly their weaknesses as opportunities and motivation for our subsequent work. Our meta-analysis shows an increasing acceptance of fact-checking as conditionally effective (its success varies according to several variables)—there is no consensus that they are overall effective. Current approaches using textual fact-checking are professional fact-checking (e.g., FactCheck.org[2], PolitiFact[3], Polígrafo[4]) and crowdsourcing (e.g., Community Notes from X which has recently been shown to be ineffective in reducing engagement with misinforma-

[2] https://www.factcheck.org
[3] https://www.politifact.com
[4] https://www.poligrafo.sapo.pt

Figure 1: Drake Hotline Bling Meme generated with the ImgFlipAI website (c.f. Appendix A for the prompt).

tion [30]). Nonetheless, the literature focuses disproportionately on the US, and the analysis is mainly from an ideological and partisan standpoint.

Given this context, our approach to the task of explaining fake news diverges by using a different modality with a different genre of explanation: respectively, images with humor, i.e., memes (visual argumentation) (c.f. Figure 1 for an example).

1.4 What are memes?

The term meme, originally introduced by Richard Dawkins to describe small units of culture that spread from person to person [31], has now evolved into Internet Memes—digital content that is rapidly circulated and imitated online [32]. We focus on one specific subdomain of Internet Memes: solely image-based medium memes with natural language captions (also known as image macros)—from now on, memes will refer to this. They are a medium of innovative cultural transmission that adeptly select, combine, and synthesize pre-existing ideas. They align with a sociocultural creativity approach as they emerge and circulate within social networks. Their applicability is diverse: from political propaganda [33], and hate speech [34], to coping mechanisms [35], and educational strategies [36], but primarily serving as entertainment. In 2019, memes emerged as a tool in Nigeria to spread COVID-19 awareness to the general public [37]. We plan to use memes similarly as a strategic method to mitigate misinformation and educate the general public.

There is some important terminology surrounding memes. The captions are the text embedded into the placeholders of the meme template; the meme template is the set of placeholders—invisible boxes that delimiter where the captions can fit into the image; a meme's label is a short description of the meme; a meme's background knowledge is the context portrayed by the meme which is not necessarily related with the image content: it includes the original media frame (the original scene that inspires the meme), the lore of its usage, and its respective meaning (c.f. Figure 2 for a clarification of the concepts) [38].

We aim to integrate explanatory text into the meme template related to the content (image and captions) that fits cohesively into the form (template) spaces and is coherent with its context and stance—tone, attitude, and standpoint that guide how the audience will receive and interpret the message. Therefore, we will create memes via meme captioning and will not change their semantics—this includes the image and the template. Lastly, memes have three characteristics that make their creation particularly challenging: (1) the multimodal nature that requires techniques to extract and integrate textual and visual information; (2) they succinctly convey a complex, unexpected interplay of ideas, which may limit systems to understand them without adequate background knowledge and therefore generate less coherent captions/memes; (3) and their fluidity, i.e., how easily they are subject to variations and alterations which makes datasets chaotic given that a single meme will have different instances requiring careful data processing [38].

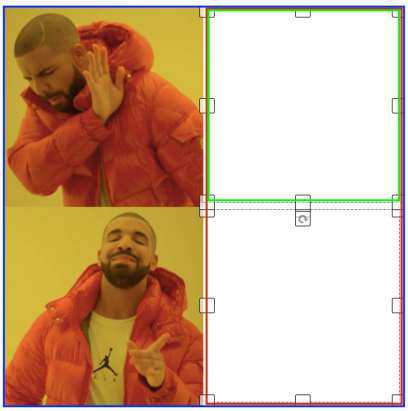

Figure 2: In Figure 1, we illustrate a meme with the label "Drake Hotline Bling" whose original media frame is the Hotline Bling music video. It has two captions: "Immigrants taking union jobs" and "Immigrants taking day laborer positions" embedded in placeholders that correspond to the green square in this Figure. The blue square denotes the meme image, and the red square is the template. Conceptually, this meme symbolizes a visual strict preference x $\succ$ y, where x represents the preferred scenario in the lower placeholder. Individuals commonly use it to express contrasting opinions or choices on social, political, and cultural preferences.

1.5 Why use memes?

It would be naïve to believe memes could solve all the issues previously mentioned while keeping the objectivity and correctness of textual fact-checking. Our approach, grounded in Aristotle's rhetorical appeals will trade some appeal based on logic and reason for the appeal to evoke persuasive emotions [39]. Recent literature has shown that emotionally evoking content is more efficient in changing people's beliefs [40]. Memes have intrinsic desired characteristics to tackle the problems of audience reach, scalability, and slowness: they are brief and straight to the point, therefore requiring a small attention span, are humorous, and have the potential to go viral [41]. An often-overlooked advantage of fact-checking is that it holds the entities that spread the false claims accountable [42]; satire, one of the humor types used in memes [43], is very effective in making politicians accountable [44]. A prior workshop on meme-based fact-checking emphasized the importance of memes being accompanied by contextual information and that they should be used with audiences that acknowledge their contextual digital culture [28]. Additionally, research suggests these explanations should strive for a balanced truthful presentation, avoiding the distortion of facts and ensuring that the critical information is conveyed. This involves using the right balance of facts and creative content that ultimately should play a role in educating the audience [45].

1.6 Work Objectives

The objective of this thesis is to address the novel task *Fact-Checking Meme Captioning–automatic generation of captions for memes that serve as the tool for fact-checking*–by proposing an AI-mediated system that automatically generates memes to explain the verdict of claims given their rationale. To achieve this, we plan to execute the following research tasks:

- **Phased System Development:** Build a system architecture in compliance with the System Requirements in Section 4.2, where (an) LLM(s) generate(s) meme captions given a claim, its verdict, the rationale, and an optional meme image.
  - Concatenate the captions and images with the ImgFlip API and filter hateful content through a hateful memes detection model.
  - Incorporate Retrieval Augmented Generation (RAG) through a vector embeddings database with data from a Knowledge Graph and a Knowledge Base.

  - Incorporate LLMs Multimodal Debate for the captions generation or Reinforcement Learning from Human Feedback (RLHF) to refine outputs based on user interactions on memes.

- **Development of a Curated Dataset:** Construct a curated dataset from a crowdsourcing platform featuring the textual input, their corresponding generated memes, and detailed evaluation metrics.
- **LLM Selection Study:** Test the system's performance with different state-of-the-art (SotA) large language models (LLMs) and their respective most effective learning techniques.
- **Conduct an Ablation Study:** Evaluate enhancement in performance provided by the different methodologies against the core architecture with a fine-tuned LLM and RAG.
- **Conduct two User Surveys:**
  - **System's LLM Selection:** Evaluate the compliance with the Memes Requirements for generated memes from the LLM Selection Study and Ablation Study.
  - **Meme-based Fact-Checking Study:** Evaluate the compliance with the Creative Explanations Requirements and compare with other Creative Explanations.

1.7 CIMPLE

This proposal and the subsequent dissertation are being developed under CIMPLE (Countering Creative Information Manipulation with Explainable Artificial Intelligence)[6]; a CHIST-ERA project approved under the 2019 call "Explainable Machine Learning-based Artificial Intelligence". The project focuses on advancing AI explainability by enhancing explanation personalization to maximize understandability. It emphasizes structuring AI decisions using Knowledge Graphs to provide context-rich explanations that improve user trust, particularly in complex areas like misinformation detection and manipulation. We address Task 5.3 of Work Package 5, which oversees "the design and development of computational creativity methods for information manipulation."

Building on the foundational goals of CIMPLE and of our specific Task 5.3, we aim to develop a system that generates visual explanations for fact-checking articles. Thus, our mission is grounded in a dual-reason argument. The first reason is that it seeks to contribute to research fields such as Natural Language Processing, Reinforcement Learning, Representation Learning, Information Retrieval and ultimately to make progress in Meme Generation Systems. Secondly, it aspires to make a significant societal impact by developing a system designed to tackle the real-world problem of misinformation. Our vision is that this work will be a pioneering effort at fact-checking and promoting further research on other creative means, enhancing the capability of fact-checking tools. To our knowledge, this is the first initiative that tries to explain fake news via an image-based medium and the first approach to develop a meme generation system for the task of fact-checking.

1.8 Outline

We introduce informal definitions for key concepts, how they evolved through time, and link them to the present study in Section 2. Section 3 delves into the relevant literature to our task and provides summarization tables. The system requirements and the architecture are proposed in Section 4. Section 5 outlines hands-on work done to this proposal where Related Work systems were tested for our downstream task. Section 6 discusses the evaluation methodologies to measure our objectives. Section 7 presents a Gantt chart with a roadmap of the subsequent tasks and their timelines. The proposal concludes with Section 8.

2 Background

This section explains the research areas in which this project is inserted. We address the role of creativity in Sections 2.1 and 2.2. Section 2.3 addresses the role of explainability and we intersect the latter two in Section 2.4. Lastly, we discuss the roles of virality and humor to memes in Section 2.5.

[6] https://www.chistera.eu/projects/cimple

2.1 Creativity

Creativity research started with J.P. Guilford's 1950 presidential address to the American Psychological Association—considered the founding figure in this field—and the subsequent publication of his journal article [46]. Koestler proposes that creativity manifests through humor, discovery, or art [47]; it arises from the interaction—called bisociation—of two independent and unrelated frames of reference or matrices of thought. The former distinction resides in the emotional charge that accompanies the bisociation. Boden defines creativity as "the ability to generate novel, and valuable ideas" [48]; the author argues there are two types of novelty: psychological and historical. The former consists of ideas "new to the person who generated it", whereas the latter is a psychological idea that "has never occurred in history before". In earlier work, Boden discusses how novel ideas can arise through combination, exploration, and transformation [49]. The first one "produces unfamiliar combinations of familiar ideas"; the second explores new ideas "on some culturally accepted style of thinking, or conceptual space"; on the third, the "space or style itself is transformed", allowing for ideas that "simply could not have been generated before the change" [48]. As the field matures, different definitions of creativity emerge. The individualist school of thought simplifies it as creating new externalized mental combinations. On the other hand, the sociocultural tribe defines it as the generation of novel and valuable products for a given social group [50].

2.2 Computational Creativity

Since Boden controversially introduced creativity into Artificial Intelligence (AI) discussions [51], it took some years for Computational Creativity (CC) to be considered a subdomain within AI—now regarded by some as its final frontier [52]. Wiggins defines CC as the "study and support, through computational means and methods, of behavior, exhibited by natural and artificial systems, which would be deemed creative if exhibited by humans" [53]. This interdisciplinary field adopts an explicitly algorithmic perspective on creativity, answering questions such as if creativity can be attributed to computers—a question viewed by some as not scientific but philosophical [48,54]. Wiggins later refined its definition to the "philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative" [52]. Perspectives on the role of these systems vary among authors, with some regarding them solely as Creative Support Systems, enhancing human creativity by collaboration, and allowing for the democratization of the Creative Act [55, 56]; while others also envision them as autonomous independent creative entities [54]. Modern creative systems are capable of producing a wider range of artifacts, including text-based artifacts–"any artifact that can be represented simply as a string of characters"—and graphics-based artifacts–"any type of artifact that can be visually understood" [57]. Memes are essentially a fusion of these two, a graphics-based artifact with one or multiple text-based artifacts within, manifesting as arguments or jokes.

2.3 Explainable Artificial Intelligence

Deep Learning models are often perceived as black box predictors—the "internals are either unknown to the observer or known but uninterpretable by humans". This lack of transparency violates the General Data Protection Regulation's right of explanation for all individuals to obtain meaningful explanations of the logic involved when automated decision-making takes place [58]. It is also critical for fact-checking because it raises trust issues if there are no justifications for the claim's verdict [27]. Explainable Artificial Intelligence (XAI) thus is "the field that tries to produce AI techniques capable of justifying themselves". Its shortcomings are failing to provide explanations with a careful balance of appeal to facts and emotions that laypeople can understand [57].

2.4 Creative Explanation

Within CIMPLE, the concept of Creative Explanation (CE) is defined as an "artifact that justifies why a piece of online information is false in a creative and targeted way" making use of techniques from CC to tackle the shortcomings of XAI [57]. Recent work introduced Personalization into Fake News explanations because these personalized explanations maximize explanation acceptance [59]. Underlying is the belief in explanatory pluralism, which consists of a given event having multiple plausible explanations. The authors mention that the

challenge remains to determine if a given individual changed a pre-conceived belief after receiving the explanation (self-reflective inducement state).

2.5 Virality and Humor in Memes

Social networks transformed modern communication from broadcasting (one-to-many) to social dialogues (many-to-many). They originated a new variant for the definition of virality—"the rapid spread (...) of Dawkins' memes across individuals and communities growing exponentially with each cycle"—an intrinsic characteristic of memes. A meme's virality depends on its likeability—the degree to which the meme is engaging—and shareability—the degree to which the receiver believes others will share the likeability. In more detail, close-up scale memes, the presence of characters, and facial emotions are indicatives of virality, while long text and content lacking clear focal subjects for the viewer are the opposite. Humor is defined as "the ability to be amused by something seen, heard or thought about, sometimes causing you to smile or laugh (...)".[7] Virality and humor are related, and we can explain the relationship via a conceptual framework that uses the source-message-channel-receiver (SMCR) communication model [43]. A sender can have four humor styles: affiliative, self-enhancing, self-defeating, and aggressive. In print media, a message can be of seven different types of humor: comparison, personification, exaggeration, pun, sarcasm, silliness, and surprise (c.f. Appendix C for the definitions). To some extent, a meme's success lies in the appropriateness of the relationship between the humor and the message. A study analyzed a thousand memes from a Facebook page and found that affiliative and aggressive are the most prevalent styles, while sarcasm and silliness are the most common types of humor [43]. They also found that the self-defeating style is the most engaging, and the exaggeration type generates the highest number of likes. However, there were no significant differences in virality between types of humor. A recent study proclaimed there are eight types of humor: fun (affiliative), benevolent humor (self-enhancing), nonsense, wit, irony, satire, sarcasm (aggressive), and cynicism [60, 61].

3 Related Work

This section offers a comprehensive review around Meme Generation Systems (MGS) for Fact-Checking. We start by studying Meme Detection and Classification (MDC) and Meme Semantic Understanding (MSU) since most AI-related research efforts, around meme-related, tasks focus on these topics. We delve into these subfields in Sections 3.1 and 3.2, summarizing their findings and relating them to our desired downstream task. Only more recently research focused on MGSs, which we discuss in Section 3.3. We investigate academic MGSs and look into commercial MGSs in Section 3.3.1 to find out what is being used and if we could leverage any for our work. Lastly, we analyze the literature on the second major domain of this work, Fact-Checking, by reviewing Misinformation Mitigation Systems in Section 3.4. We conclude by examining Creative Explanation Systems in Section 3.4.1, a subtype of the latter systems of which our task is a specific instance, and by discussing how the findings relate to our solution requirements in Section 3.5.

3.1 Memes Detection & Classification

We turn our attention to meme detection systems that revolve around detecting harmful content, since the majority of the literature addresses this topic due to moderation mechanisms initiatives from social media platforms.

Facebook, organized the Hateful Memes Challenge (HMC) and published the benchmark dataset for this task featuring examples that only received high confidence ratings from trained annotators with a moderate agreement on Cohen's Kappa [62]. The novelty in this dataset from previous ones in the literature was including benign cofounders—copies of hateful memes modified with either benign text or images–to prevent models from exploiting unimodal priors. It was benchmarked on several models: unimodal, multimodal, pre-trained on unimodal data, and multimodal data. Multimodal models outperformed marginally, suggesting that the sub-field of multimodal pre-training is still in its infancy, which is a limitation for meme generation. The gap between human evaluations and the best-performing model, Visual Bidirectional Encoder Representations from Transformers (BERT) [63] trained on the Common Objects in Context dataset (COCO) [64], is very substantial with

[7] https://dictionary.cambridge.org/dictionary/english/humor

84.7% versus 69.47% accuracy, respectively. ISSUES addresses the HMC from Facebook through the use of a Contrastive Language-Image Pre-training model (CLIP) [65], enhanced by the textual inversion technique, 2-stage training, and a multimodal feature fusion function (Combiner Network) [34]. This approach achieved 77.70% accuracy on the HMC dataset. Another approach also used CLIP, enhanced with prompt engineering techniques [66]. The system takes as input a meme and outputs either the label hateful or non-hateful depending on the class with the highest cosine similarity. Without prompt engineering, CLIP achieved an accuracy of 57.8% through a zero-shot classification approach. At the same time, after accurately finding the threshold of the cosine similarity, the authors claim the system achieved an accuracy of 87.42% with no false positives, surpassing ISSUES by almost 10%. Such a significant improvement from a model using solely prompt engineering leaves questions about the scientific rigor of the methodologies employed. Unfortunately, the results cannot be reproduced since the code is unavailable.

UNITOR extends the previous systems' task by classifying memes for their referring topics [67]. The system employs UmBERTo[8], a BERT-based model pre-trained on Wikipedia, to encode textual content, and integrates the ResNET152 [68] CNN to leverage visual information. UmBERTo was fine-tuned on HaSpeedDe, another Hate Speech Detection dataset from Facebook previous to the HMC dataset. While the task primarily relies on text embeddings, the visual embeddings were found to have a positive impact, although marginally, confirming the findings from Facebook's initial paper. The Multimedia Automatic Misogyny Identification challenge also identifies and classifies memes but specifically for misogyny [69]. The dataset created for the challenge was obtained through manual extraction and web scraping from social media platforms, as well as websites like KnowYourMeme.[9] This resulted in 15k memes, labeled by three human annotators with a fair to moderate Fleiss-k agreement on their categories of misogyny. The top-performing teams used ensemble models of machine learning techniques, ensembles of different pre-trained models, and CLIP paired with either logistic regression or Long Short-Term Memory (LSTM). The models were biased towards misclassification of not misogynous memes as misogynous.

Other attempts augmented models with knowledge graphs. KERMIT tackles the task of hateful meme detection by building a knowledge-enriched information network with memory-augmented neural networks whose information is fetched from ConceptNet[10] and is composed of the meme entities and their relationships [70]. Afterward, through the attention mechanism, KERMIT has a dynamic learning mechanism that identifies the most informative segment of the information surrounding the meme for the classification. MemeGraphs not only uses knowledge graphs to represent the visual content of the memes but also leverages scene graphs—express memes in terms of detected visual objects and relations between them [71]. A human-annotated dataset is provided to assess their quality. Although one of the annotators classified 22% of the detected objects and 32% of the relations as incorrect, the fully automatic augmentations achieved the highest scores. MemeGraphs outperforms ImgBERT on all settings and TextBERT on some, all variants of BERT [72]. The best score was obtained using TextBERT with only scene graph input.

ExplainHM is the most advanced system and is the first to tackle the explainability for harmful meme detection systems [73]. It prompts two LLaVAs [74] with the meme and a harmfulness label (hateful/not hateful), directing them to generate a specific rationale. The latter explains how the interplay between the meme text and the meme image could contribute to the given label. These conflicting perspectives will enter a debate to promote divergent complex reasoning pathways so that another LLM, acting as a co-judge, will use them as prompt parameters to infer the final classification. Due to the huge number of parameters of current LLMs, the authors fine-tuned a smaller LLM, the final judge, that accepts as input the concatenation of the meme text and the two rationales in order with the co-judge preference. The authors compared the system on three publicly available datasets to the SOTA models for meme detection, such as VisualBERT, obtaining advancements in the Macro F1-Score metric of around 3,14%. Regarding explainability, the authors evaluated GPT-4[11] and humans. The former scored explanations from ChatGPT and LLaVa higher than human evaluation did and vice versa. For the latter, human explanations only performed better on Conciseness, whereas for the rest (Informativeness, Readability, Soundness, and Persuasiveness), they obtained worse results than LLaVa and ChatGPT. Ten professional linguists annotators evaluated this from 50 randomly chosen memes. On the other hand, the number of

[8] https://github.com/musixmatchresearch/umberto
[9] https://knowyourmeme.com/
[10] https://conceptnet.io/
[11] https://openai.com/index/gpt-4-research/

individuals who participated and their demographic factors are unknown. The paper concludes with a figure that contrasts the performance of each model by considering their number of parameters, aligning with GreenAI principles. The authors conclude the system is effective and not overly dependent on the parameter size of the Small LLM judge to enhance performance.

Overall, multimodal pre-training on multimodal models and augmenting models with contextual information improves the performance. CLIP and VisualBert are the models generally used by the best-performing systems, though both are outperformed by SOTA LLMs such as GPT-4 and LLaVa. We provide Table 1 with a summary of the reviewed systems and their relevance to our Proposed Solution.

Table 1: Memes Detection & Classification reviewed work.

Name	Task	Architecture	Relevant Cues
Facebook's HMC [62]	HMD	-*	Multimodal models multimodally pre-trained are the best-performing.
ISSUES [65]	HMD	CLIP	Desintagling image and textual embeddings. Using the Textual Inversion Technique or a Mutimodal Fusion Function.
Meme Hate Speech Detection [66]	HMD	CLIP	Prompt Engineering improves detection.
KERMIT [70]	HMD	ViT	Augmented embeddings improve detection.
MemeGraphs [71]	HMD	Scene graphs.	Augmented embeddings improve detection.
ExplainHM [73]	HMD	ChatGPT-4, LLaVa	Explainability metrics.
MAMI [69]	HMC	ERNIE-Vil, Uniter, VisualBERT, CLIP	Ensembles of pre-trained models are the best-performing.
UNITOR [67]	HMC	RestNet152, UmBERTo	Integration with visual embeddings enhances performance.

Note: In the Relevant Cues column, we include insights relevant to our Proposed Solution in Section 4. HMD stands for Hateful Memes Detection, and HMC for Classification. *This paper did not propose an architecture, although the best-performing model was the Visual BERT.

3.2 Memes Semantic Understanding

We move from detecting and classifying offensiveness on memes to analyzing and interpreting their semantics. This approach examines memes' contextual background information, visual metaphors, and the sentiments they convey.

Early systems started by approaching the task of meme retrieval [75]. SimMeme is a meme-dedicated search engine that uses a graph-based data model enriched with the usual textual and visual features, plus additional meme-related tags with ontology edges that describe the semantic relations between them. The information surrounding memes was very limited, and therefore, a new approach applying Named Entity Recognition (NER) and Relation Extraction (RE) followed [76]. The authors introduced the MERMAID dataset composed of memes from ImgFlip and the entities and relationships between them in their captions. Three annotators identified the latter with an inter-annotator agreement core of 0.602, measured by Krippendorff's alpha. They further created the MERF framework, trained it on the MERMAID dataset, and benchmarked its performance for both subtasks. For NER, the performance was exceptional and similar between BERT and ROBERTA. As stated by the authors, most captions predominantly describe the entities themselves. However, RE proved to be more challenging, reflecting the task's inherent complexity. Among the tested architectures, BERT-CLIP emerged as the best performer, reaching a Micro-F1 score of .701 and an accuracy rate of 65.1%. Eventually, the challenge became to identify the information from the vast memes' context sufficient to explain their meaning [77]. MemeX takes a meme and a related context and determines if the context explains the meme, producing a corresponding label. This system includes three innovative ideas: firstly, it augments the multimodal meme representation with the encodings of the captions' meaning and semantics; secondly, it encodes the meme contextual document using a pre-trained BERT encoder and integrates it with the enriched meme representation on the Transformer

Encoder; finally, the output of the encoder is refined by a custom LSTM before the feed-forward layer that gives the classification. A dataset, curated by two annotators, was introduced, which had 3400 memes and their related contextual documents with evidence sentences. The former were web-scraped from Google Images[12] and Reddit[13] while the latter was predominantly retrieved from Wikipedia.[14] Empirical results demonstrate that their architecture outperforms MMBT [78], the best-performing baseline, by 5.34% in accuracy when benchmarked against their dataset.

The literature later moved from extracting information from the memes to generating natural language explanations. LUMEN, a multimodal, multi-task learning framework, generates explanations for visual semantic role labeling in memes [79]. This framework uses three techniques to understand the meme: Optical Character Recognition to extract the meme's textual content; the Vision Transformer [80] to encode the meme into a visual embedding; and the Omnipotent Feature Aggregator image captioning model that generates descriptive text for the meme. LUMEN outperformed baseline models across multiple evaluation metrics. The results underscore the effectiveness of combining textual and visual analysis to generate meme explanations over unimodal solutions, as stated in Section 3.1. A different approach generates explanations for the meaning of memes by focusing on their visual metaphors [81]. The authors introduced MemeCap, a dataset with 6384 memes sourced through the Reddit's API.[15] Each entry comprises the title, a literal image caption, and a human-written caption that explains the meme's meaning. The latter was obtained by manual annotation conducted on Amazon Mechanical Turk.[16] It was found that 81% of memes had metaphors, of which 44% were complementary. Additionally, 53% of memes have a person or a character as the metaphor vehicle type. The human evaluation setup introduced new metrics such as Appropriate Length and Faithfulness: the former evaluates if the generated captions are too verbose. In contrast, the latter evaluates if the generated captions contain hallucinations. The results indicate that all tested models perform significantly worse than humans on all metrics except length appropriateness.

Similarly to the previous section, modern LLMs were employed to perform the literature downstream tasks [73]. PromptMTopic followed this trend by addressing the task of topic modeling for memes with ChatGPT [82]. The system has two phases: a topic generation phase where ChatGPT learns and generates topics related to the meme via in-context learning and a second phase where redundant and irrelevant topics are filtered via Prompt-Based Matching. The authors found two problems around using ChatGPT: it is prone to hallucinations when it does not understand the context well enough, and its strict content moderation layer prevents it from generating text. ChatGPT gave better qualitative results, and the model outperformed all the topic modeling baseline competitors. Nonetheless, at least with the meme examples provided, the topics generated by the system failed to capture the meme's intended message.

Thus, enriching meme embeddings with information such as the literal meme image descriptions or their entities and relationships, among others, consistently proved to provide better results. We provide Table 2 with a summary of the reviewed systems and their relevance to our Proposed Solution.

[12] https://images.google.com/
[13] https://www.reddit.com/
[14] https://www.wikipedia.org/
[15] https://www.reddit.com/dev/api/
[16] https://www.mturk.com/

Table 2: Memes Semantic Understanding reviewed work.

Name	Architecture	Relevant Cues
SimMeme [75]	-	Semantic similarity measures to compare concepts.
MERMAID [76]	BERT, BERT-CLIP	Augment embeddings with entity recognition and entity relation.
MemeX [77]	BERT, GCN, MMBT, LSTM	Enriching memes embeddings with their semantics. Gated multimodal fusion to augment embbedings. Context-aware Transformer and LSTM.
LUMEN [79]	ViT, Omnipotent Feature Aggregator	Textual and visual analysis enhances performance for the generation of explanations for visual semantic role labeling.
MemeCap [81]	-*	Introduction of metrics to evaluate the meme captions.
PromptMTopic [73]	ChatGPT**	ChatGPT model limitations: hallucinations and strict safety guardrails.

Note:*The paper only proposed a dataset but benchmarked it on Flamingo [83], MiniGPT4 [84] and LLaMa [85]. ** The authors do not disclose which ChatGPT model they used.

3.2.1 Memes Knowledge Representation

Understanding the importance of contextualizing models with the meme's background information, researchers codified this information in Knowledge Graphs (KG) and Knowledge Bases.

The Internet Memes Knowledge Graph (IMKG) is the first aggregator to capture memes stratified semantics explicitly in text and vision and implicitly through references from background knowledge [38]. The information was extracted from a meme encyclopedia (KnowYourMeme), a meme generator website (ImgFlip), and an existing open KG (Wikidata[17]). The visual enrichment of the data is performed with Google Vision API, which extracts entities from the meme images. After all the data integration, the knowledge graph has, on average, 2593 memes per meme image. A more recent KB is the Know Your Meme Knowledge Base (KYMKB), comprised of more than 54000 memes from 5220 meme images and detailed information about them [86]. The authors also propose a non-parametric majority-based classifier that assigns meme images to random memes by the distance between their vector representations. Encoding the About section, the template itself, its instances, or a combination of some can be used to classify memes. The results showed this to be a competitive method to address several classification tasks regarding memes while being more efficient computationally.

We plan to leverage both these systems for our task of meme generation. We showed in Section 3.1 that training models on multimodal data yielded better results. Additionally, research in Section 3.2 demonstrated that providing models in the context of memes enhances their performance. Accordingly, the most advanced MGS [87], which we analyze in the following Section, used IMKG for generating memes. We explain how we will use these systems in Section 4.

3.3 Meme Generation Systems

At last, after an extensive analysis of tasks parallel to ours, we focus our attention on the systems that generate memes. The literature on MGSs 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. Afterward, the literature converged into the encoder-decoder architecture until SOTA Transformer models and LLMs surpassed these and became the status quo, mirroring what we observed in the domain of Meme Detection and Classification in Section 3.1.

The first approach employed a Nonparanormal Network, an extension of the normal Gaussian distribution that leverages Copula Theory. It modeled non-linear stochastic dependencies between popular meme images, their captions, and popularity votes [88]. The model was designed to rank meme captions fetched online and selecting the one with the highest posterior. Its evaluation was based on the BLEU score [89], a metric used

[17] https://www.wikidata.org/wiki/Wikidata:Main_Page

to evaluate machine-generated text comparing it to human-generated text. 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 creates memes by adapting Portuguese news headlines from the Google News RSS feed to meme images according to a deterministic rule-based classifier [90]. Although the authors claim the system is language-independent, the linguistic tools are biased towards Portuguese, making it unsuitable for our needs for English-language meme generation. The evaluation involved 52 participants who assessed the memes on coherence, suitability, surprise, and humor. The findings were that MemeGera could not beat humans on any metric. However, human-generated memes were also evaluated, and they did not get the maximum scores, demonstrating the inherent difficulty of meme creation. Lastly, we argue that the selection criteria for the generated memes should be detailed to rule out any cherry-picking bias. Stonkinator is a simple system that takes as input a text caption and extracts the relevant keywords to use them as a search query for retrieving images online [91]. Upon fetching the images, two are randomly selected and blended with three different techniques to create a Stonks Meme.[18] Our approach is fundamentally different since we plan to generate explanatory captions for fact-checking articles for meme images. In contrast, News2meme is akin to our approach since it applies word vector similarity to retrieve an existing meme image and caption that best matches the content of a piece of news [92]. It was found that 71.21% of the generated memes received unfavorable feedback from 9 evaluators, showcasing the ineffectiveness of the approach.

Peirson et al. introduced the Neural Network encoder-decoder architecture into meme generation dubbed Dank Learning, becoming the baseline across the literature [93]. It comprises a Convolutional Neural Network (CNN) encoder that takes as input a meme image and its label and an LSTM decoder that generates the humorous caption for the meme. The encoder uses a pre-trained Inception v3 network on over one million images. A qualitative study with 5 participants assessed the differentiability of the generated memes with human-generated ones and their hilarity. The findings report that the memes were generally indistinguishable from human-generated ones, and that the hilarity always exceeded 5.5 on a scale from 0 to 10, with a model variation reaching 6.9, showcasing moderate levels of humor. The demographically undefined sample of 5 individuals is not statistically representative. DeepHumor extended the work by studying variations of the same architecture with a more representative evaluation of 53 participants [94]. They repeated Dank Learning's evaluation and complemented it with the metric coherence observed in MemeGera [90]. The results indicate that the generated memes were generally recognized as artificial. Wang et al. diverged by using a Transformer, specifically OpenAI GPT-2 [85], as the decoder to generate the meme captions in the Chinese language [95]. They continued the status quo by evaluating memes on hilarity and coherence, with the average score of both metrics being higher for the generated memes than the ones created by humans. However, selecting only 20 random memes raises concerns about the robustness against poorly constructed outliers. In a second user survey on differentiability, the results reported that 75% of generated memes were mistakenly classified as real, outperforming Dank Learning [93]. MemeBot applies the same architecture but distinguishes itself by being the first system to generate memes from larger context inputs, specifically using tweet sentences [96]. It resembles our work since it will also incorporate large contextual inputs. Their human evaluation assessed the memes' coherence and relevance: "Can you understand the message conveyed through the meme?" and "Is the meme contextually relevant to the text (tweet sentence)?". Through human voters acquired via Amazon Mechanical Turk, 66.5% of the generated memes were considered coherent, and 66.25% were considered relevant. However, the number of participants remained undisclosed after inquiries to the authors. The novelty was introducing the study of shareability, an important characteristic of memes discussed in Section 2.5. Memeify, similarly to Wang et al., leverages GPT-2 model for caption generation, emphasizing thematic and stylistic consistency by appending each input with the meme's class (meme image) and a desired theme [97]. These themes resulted from a clustering algorithm on caption encodings, which identified five distinct clusters. Each meme was classified into six themes: Depressing, Frustrated, Savage, Unexpected, and Wholesome, and a category for outliers. For evaluation, Memeify was benchmarked against Dank Learning and original memes, with 20 social media experts participants assessing memes on the caption content, humor, and originality (1 to 5). Across all themes, Memeify's memes outperformed the baseline model yet fell short of the originals. In another evaluation, they were also misclassified as original 66.67% of the time. Our task's lack of tailored datasets prevents this approach from generating captions correlated with our input requirements, as it relies heavily on extensive fine-tuning. Thus, the adoption of the

[18] https://knowyourmeme.com/memes/stonks

encoder-decoder architecture is dismissed.

MemeCraft is an end-to-end pipeline that transforms user prompts into memes advocating specific social movements such as Climate Change or Gender Equality [87]. It starts by sampling random meme images from ImgFlip and using them as input for VLMs to generate their descriptions. Then, with instruction tuning and few-shot learning, they generate the captions for the meme through LLMs and overlay them into the meme templates through the ImgFflip API. In the end, they incorporate a Multimodal BiTransformers model pre-trained on the Hateful Meme Challenge mentioned in Section 3.1. This is the first meme generation system in the literature to incorporate a self-regulating safety mechanism to filter hateful memes. The authors generated memes using ChatGPT-3.5[19], LLaMa-2-13B [85] and LLaVa-7B [74] as the LLMs on their architecture and benchmarked them against memes from Dank Learning and humanly created memes on ImgFlip. A group of 12 undergraduate students evaluated them on Authenticity, Hilarity, Message Conveyance, Hatefulness and Persuasiveness, a relevant component of explainability. The GPT variant provided the best results across all metrics, yet the LLaMa model was very competitive regarding hilarity and persuasiveness. Another interesting finding is that 75% of the human-created memes have a moderate to high level of humor. At the same time, for ChatGPT, it is around 58%, confirming that SOTA LLMs are still far from humans in humor generation. Finally, regarding hatefulness, the authors presented two scores: hateful memes identified by the safety mechanism and hateful memes identified by the evaluators from the set of post-filtered. For the gender equality social cause and the GPT variant, the safety mechanism identified 55% of the memes as hateful. After filtering, only 2% were identified as hateful by the annotators. Future work could provide the percentage of memes identified by the annotators before filtering to understand the prevalence of false positives. The fundamental difference between our approach and MemeCraft is the intended use. Our task of generating memes to correct misinformation 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. In conclusion, Table 3 and Table 4 summarize the academic MGSs. For a detailed visualization of the chronological progression of the Related Work studied, observe Figure 3.

Table 3: Memes Generation Systems reviewed.

Name	Architecture	Input	Best Performing Model
I Can Haz Chezburger [88]	Nonparanormal Network	Image	All Text + Vision features(Named Entities, Semantics, etc...)
MemeGera 2.0 [90]	Rule Based Classifier	News headline	-
Stonkinator [91]	Image Blending	Input sentence	-
News2meme [92]	Mutual Subspace Method	News articles	-
Dank Learning [93]	Encoder-Decoder	Meme Image and its label	Inception-v3-attention-based LSTM
Memeify [97]	Encoder-Decoder	Image or Theme and Meme Image	VGG16 and GPT-2
DeepHumor [94]	Encoder-Decoder	Image, Image and label or Image and spatial features	Image-only(ResNet) LSTM Decoder with character-level tokenization
Automatic Chinese [85]	Encoder-Decoder	Meme images	ResNet-50 and GPT-2
MemeBot [95]	Encoder-Decoder	Sentence from a tweet	Noun phrases and verbs based input with BERT and a base Transformer custom architecture
MemeCraft [87]	LLMs and VLMs	User prompts	ChatGPT-3.5

[19] https://openai.com/index/chatgpt/

Table 4: Overview of the Memes Generation Systems features.

Name	Languages	Code/Dataset	Human Eval.	Safety Mech.
I Can Haz Chezburger [88]	English	-	✗	✗
MemeGera 2.0 [90]	Portuguese*	-	✓	✗
Stonkinator [91]	English	-	✓	✗
News2meme [92]	English	C, D	✓	✗
Dank Learning [93]	English	C, D	✓	✗
Memeify [97]	English	C, D	✓	✗
DeepHumor [94]	English	C, D	✓	✗
Automatic Chinese [85]	Chinese	-	✓	✗
MemeBot [95]	English	-	✓	✗
MemeCraft [87]	English	C, D	✓	✓

Note: In the Code/Dataset column, "C, D" means the paper made the code and the dataset available, while "-" means none.

The Languages column indicates the languages in which the system was trained or the languages intended to generate the captions. * The authors argue that the system is language-independent.

3.3.1 Commercial Tools

While plenty of websites allow for manual meme generation, only recently some have incorporated AI-powered systems capable of automatic production. Since we aim to create memes, we should analyze available meme-generation websites to understand if we can leverage any. We divided the systems into two groups: those with textual input and those with image input.

Onto the former, the most famous is the ImgFlip website, which generates a meme for a given textual input, allowing the user to choose the meme image between two distinct AI models. The oldest is built on a deep convolutional neural network (DCNN) custom-developed by the authors. From a vast dataset of 100M meme captions, after selecting the 48 most famous meme templates, 45M training examples were gathered for a character-level generation approach. While the dataset is not publicly available, the code is.[20] On the other hand, the most recent model leverages the ChatGPT API according to the information provided on the website. However, the API only allows the use of the DCNN model. Although the latter does not accept textual prompts that enable us to manually test its reasoning capabilities, from around every ten generated memes, 2 include profanity, which is expected since the authors state that no filtering was done on the training data. On top of that, the generated captions are often semantically incoherent, making this an unviable option for our system requirements. SuperMemeAI[21] makes an introduction to human-assisted meme generation systems where the user chooses the preferred one from a list of 10 generated memes, either as images or gifs. This approach has several benefits: a meme of good quality is very situational for the meme image, and it is a hard problem to guess the most fitting—even for humans, as observed in [90]; it leaves the human to decide what he/she considers offensive and humorous, both intrinsically related with the individual characteristics of the user [98, 99]. It is mentioned that the model "generates original memes powered by OpenAI", but no API is available. PredisAI[22] is similar to the previous one but generates fewer memes and has an API. Lastly, there is Simplified[23] that differentiates its product by allowing the user to choose the language and the tone of the generated captions while also providing an API.

Moving onto image-based systems, they work around any image the user uploads instead of solely providing meme images. AIMeme[24] personalizes the service by providing options for the user to choose his personality type, the art style of the generation process meme, and the language. On the other hand, MemeCam[25] takes as context a word tag to direct the generation process and the desired language. While the former states that

[20] https://shorturl.at/ciBI3
[21] https://www.supermeme.ai/
[22] https://predis.ai/
[23] https://simplified.com/ai-meme-generator
[24] https://ai-meme.com/
[25] https://www.memecam.io/

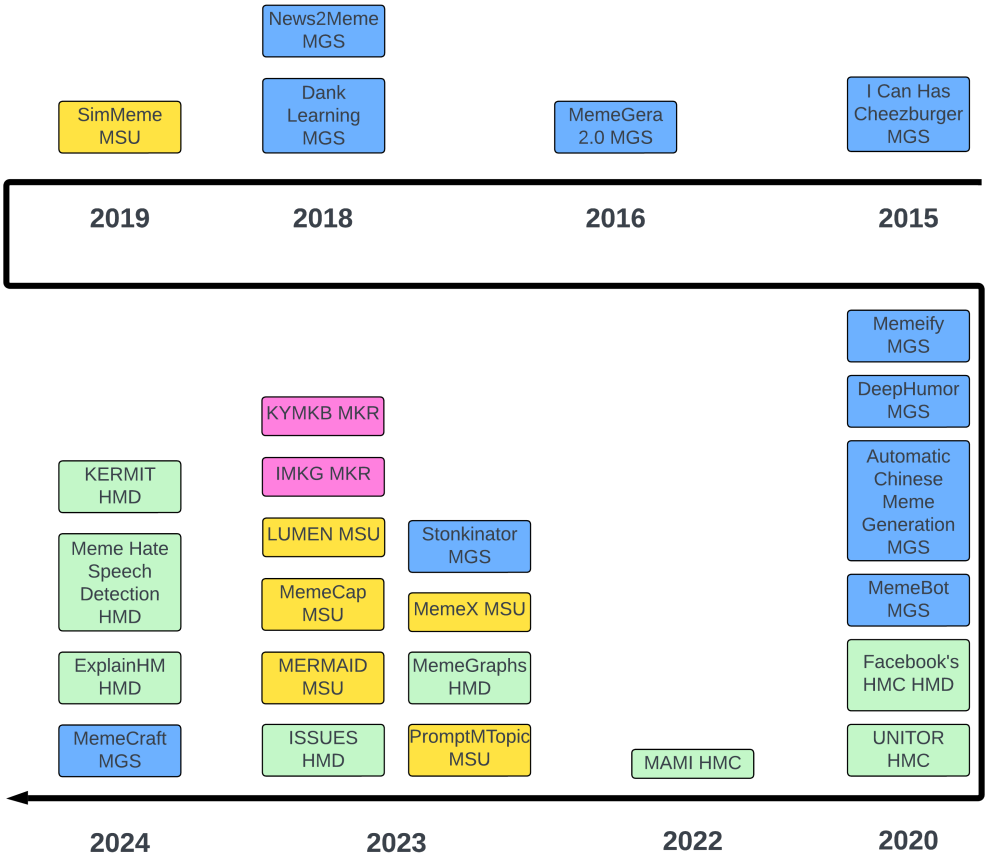

Figure 3: Chronological progression of the Meme-Related reviewed systems, distinguished by colors: Meme Generation Systems (MGS) in blue, Meme Knowledge Representation (MKR) in pink, Meme Semantic Understanding (MSU) in yellow, and Hateful Meme Classification (HMC) and Detection (HMD) in green.

it generates memes using the GPT-4 Vision Model, the latter states that it "combines BLIP image recognition and GPT-3.5 AI-driven caption generation". Both have no API. The last system, AIMemeGenerator[26], is not a commercial system since it is a free application hosted on HuggingFace developed on top of the IDEFICS model. However, given that it allows for generating memes, we will include it here. It combines all the input types from the systems above, allowing for text or text and uploaded image/provided meme images. The generator offers an API.

We contacted all the websites inquiring about their implementation details, obtaining an answer from the creator of MemeCam, who kindly confirmed that his system had no fine-tuning nor any other learning technique to improve the process. Likewise to Section 3.3, we condensed our overview in Table 5. Nonetheless, none of the available APIs allowed for augmenting the generation process, which is one of the main goals of this thesis, and thus, using a commercial system was ruled out, so we proceeded to create our own from scratch.

3.4 Misinformation Mitigation Systems

This section analyses some Misinformation Mitigation Systems. Most of the emphasis is on detection systems, with more recent approaches pursuing explainability.

Onto detection, FACT-GPT is a framework for Misinformation Detection designed to automate the claim-matching phase of fact-checking using LLMs [100]. Claim Matching is matching previously fact-checked claims

[26] https://huggingface.co/spaces/HuggingFaceM4/AI_Meme_Generator

Table 5: Commercial meme AI-generators.

Name	Architecture	Input	API endpoint
ImgFlipAI[27]	ChatGPT API or DCNN	Text or text and meme image.	✗
SuperMemeAI[28]	An OpenAI model	Text.	✗
PredisAI[29]	-	Text and context parameters.	✓
Simplified[30]	-	Text and context parameters.	✓
AIMeme [31]	GPT-4 Vision model	Image and context parameters.	✗
MemeCam[32]	BLIP and GPT-3.5	Image and context parameters.	✗
AIMemeGenerator[33]	IDEFICS	Text or text and image/meme image.	✓

with new instances from various sources. It is built upon the ideas of augmented intelligence, which aims to augment human decision-making capabilities without replacing them with AI-integrated tools. This study focuses on the text entailment task by prompting LLMs to classify pairwise relationships of debunked claims and synthetically generated social media posts with GPT-4 on Entailment, Neutral, or Contradiction. For example, it assesses if Claim C is false, then the social media post P also has to be false (entailment). LLMs reliably matched claims, offering performance comparable to human ratings. Thus, it contributes to all stakeholders in the fact-checking pipeline by reducing redundant verification, aiding automatic content moderation, and improving misinformation analysis on a large corpus.

Moving to recent systems, MisinfoCorrect is a counter-misinformation system built on a reinforcement learning framework on top of GPT-2 that learns to generate responses designed to refute misinformation posts from Twitter [101]. These responses are crafted to explicitly dispute the misinformation, thereby correcting the spreader and mitigating the impact of false claims. The system ensures that each response not only refutes the misinformation but also incorporates supporting evidence to enhance its credibility, maintains a polite tone to improve receptivity, and exhibits fluency and relevance to ensure the communication is both engaging and directly addresses the misinformation post. Rather than augmenting professional fact-checkers, it aims to empower non-expert ordinary users by generating effective responses to online arguments. Overall, the proposed model generated high-quality quantitatively and qualitatively better responses than baseline models and human responses, respectively.

While MisinfoCorrect focuses on textual explanations, a recent direction in misinformation correction is exploring the visual modality. In one of the studies, researchers used visuals that represented real-life scenarios in conjunction with sources that highlighted the falsehoods in the claims that were targeted to be corrected. Nonetheless, the findings demonstrated that these visual methods did not improve the efficacy of corrections when compared to textual approaches [102]. Another study emphasizes that for multimodal explanations to be successful, they must firmly integrate their visual components within a clear and logical framework that argues against the misinformation, ensuring every element effectively contributes to a compelling counter-argument [103]. This may have been the potential reason for the findings in the former study.

3.4.1 Creative Explanations Systems

Early work already developed under the CIMPLE project started with a system whose Creative Explanations for Fake News were through the use of poetry [45]. The author participated in several meetings with important fact-checking agencies from different parts of the world that shared the main argument for using summarized explanations of Fake News for this system: the general public does not read long explanations. Mainly journalists read them. The summarized explanations for the Fake News chosen were the If Your Time Is Short Section on PolitiFact Articles.[34] At the time, no models were benchmarked on the generation of poetry. Therefore, a comparison study was elaborated from where GPT-3 DaVinci was selected based on three criteria and manual evaluation with quantitative and qualitative metrics. These three criteria were that the model needed to receive as input a long text sequence and generate poems from it; the generated poems had to be in English; and lastly, the model needed to be available to be used and tested within a reasonable time and computational resources.

[34] https://shorturl.at/nPZ47

So that GPT-3 DaVinci would learn the desired downstream task, the author created three custom datasets and fine-tuned three variants under them. These variants were qualitatively evaluated on metrics such as Coherence, Readability, Explainability, and if the output was considered Humorous. The human evaluation questionnaire got 103 answers, and the following conclusions were obtained from the data: the human-generated poems quality was compromised since only 60% of the individuals classified real humanly-written poems as human-written, which could have skewed results for machine-generated poems; machine-generated poems were more accepted as an explanation and were also said to explain better than their counterparts; human written poems were preferred in terms of liking, humor, and creativity. Finally, a good degree of similarity with the human-written poems was achieved since there was an even split for machine-generated poems, with 38% of the respondents classifying as Machine and 37% as Human. Additionally, some preventive measures in the questionnaire were considered to reduce biases in the results: did not mention Creative Explanations for Fake News or poems anywhere and only addressed them as explanations; did not mention there was Machine generated explanations until the very end; and balanced the distribution of types of poems across the four versions of the questionnaire.

3.5 Discussion

This section explains how we derived the System Requirements for our proposed solution in Section 4 from the literature review findings and the CIMPLE project guidelines in Appendix B. For brevity, we will only discuss requirements whose arguments have not been previously covered in this proposal.

Multimodal pre-training on multimodal models and augmenting models with contextual information improves the performance, fulfilling Requirement 2 (**R2**). One of the arguments for using memes discussed in 1.5 was because they are humorous. Some degree of their hilarity depends on prior understanding of their context [28]. We assume popular meme images have a higher chance of resonating with a broader audience; therefore, they should be favored (**R3**). We mentioned current problems of SOTA LLMs discussed in the PromptM-Topic system, such as hallucinations and answer refusal [82]. The latter will likely be prevalent, given the predominance of political content in fact-checking articles. Thus, our system should always generate explanations (**R4**). Additionally, likewise, to the commercial MGSs, our system should generate several explanations (**R4**) and let the user choose its preference since humor styles are also inherently correlated with specific personality traits [99]. At last, any content containing "offensive language or derogatory terms that might demean or exclude individuals based on race, gender, ethnicity, religion, disability, or any other characteristic" should be filtered (**R5**).

To prevent the proliferation of misinformation, our system should filter out hallucinations (**R7**). Creative Explanations aim to enhance explanation acceptance by personalizing and emotionally charging the explanations. Thus, we should evaluate our acceptance rate against text-based factual explanations (**R9**). Additionally, our system should play a role in educating the audience, as Guideline 4 suggests (**GL4**). We aim to do this by encouraging users to seek out related fact-checking articles (**R10**) by providing their references with the generated memes (**R8**). Lastly, while **R7** assesses that the generated creative explanations should not be factually incorrect, they also should promote the user's beliefs closer to the truth (**R11**).

The generated meme captions should be concise and fit cohesively into the meme template to ensure readability (**R12**). When people receive emotionally engaging content, they become more open to pursuing more information about the subject [40]. On top of that, it has been shown that emotion in a social media post leads to greater user engagement (**R14**) [104]. Regarding explainability, the better the argumentation, the higher the chances for persuasion and opinion changes in people. Accordingly, the generated memes should be related to the subject of the claim (**R13**) and should be able to explain the claim's verdict (**R15**).

Finally, we derived the requirements for the evaluation process. **GL2** emphasizes the importance of not presenting raw corrective information before the context is given (**R17**). According to Santos et al. [45], our evaluation process should not mention Creative Explanations for Fake News but rather explanations (**R18**), and these should not be proclaimed as machine-generated (**R19**).

4 Proposed Solution

In this section, we start by formalizing the problem which our task solves in Section 4.1. In Section 4.2, we enumerate all the system requirements. Lastly, Section 4.3 presents and explains the system architecture.

4.1 Problem Formulation

Let $\mathcal{M}$ denote a meme composed of an image (meme image) $\mathcal{I}_{\mathcal{M}}$ and an embedded text (meme captions) $\mathcal{T}_{\mathcal{M}}$. As mentioned in Section 1.4, memes have a background knowledge $\mathcal{K}_{\mathcal{M}}$. Now let $\mathcal{F}$ be a Fact-Checking article. It is constituted of a claim $\mathcal{C}_{\mathcal{F}}$, the claim verdict $\mathcal{V}_{\mathcal{C}_{\mathcal{F}}}$ and the rationale $\mathcal{R}_{\mathcal{F}}$ that contains the full textual explanation for the classification of $\mathcal{C}_{\mathcal{F}}$ with $\mathcal{V}_{\mathcal{C}_{\mathcal{F}}}$. A claim is defined as a statement made by a public person that was analyzed by a reputable fact-checking entity. The rationale is the factual textual-explanation for the verdict given to the claim. Lastly, the verdict is the accuracy of the claim that could be true, false, or other (c.f. Appendix D for example verdict definitions).[35] Additionally, there is the If Your Time is Short (IYTIS) section $\mathcal{S}_{\mathcal{F}}$ (S stands for a summary), which is a summarized version of $\mathcal{R}_{\mathcal{F}}$. As a result, we frame the task of fact-checking meme captioning as a function $\mathcal{F}$, which has two distinct use cases depending on the input scenario. Specifically, we define two sub-functions, $\mathcal{F}_1$ and $\mathcal{F}_2$, as follows:

$$\mathcal{F}_1 : (\mathcal{C}_{\mathcal{F}}, \mathcal{V}_{\mathcal{C}_{\mathcal{F}}}, \mathcal{S}_{\mathcal{F}}, \mathcal{I}_{\mathcal{M}}) \rightarrow \{\mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^k\} \quad \text{if} \quad \mathcal{I}_{\mathcal{M}} \text{ is provided}$$

$$\mathcal{F}_2 : (\mathcal{C}_{\mathcal{F}}, \mathcal{V}_{\mathcal{C}_{\mathcal{F}}}, \mathcal{S}_{\mathcal{F}}) \rightarrow \{\mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^k\} \quad \text{if} \quad \mathcal{I}_{\mathcal{M}} \text{ is not provided}$$

$\mathcal{F}_1$ operates when the user provides a specific meme image $\mathcal{I}_{\mathcal{M}}$. It will generate k memes $\mathcal{M}^i$ for the specific provided meme image $\mathcal{I}_{\mathcal{M}}$. Each meme $\mathcal{M}^i$ consists of the image $\mathcal{I}_{\mathcal{M}}$ paired with a generated caption $\mathcal{T}_{\mathcal{M}}{}^i$. $\mathcal{F}_2$ operates when the user provides no specific meme image. It will generate k memes $\mathcal{M}^i$, each composed of a system-selected meme image $\mathcal{I}_{\mathcal{M}}{}^i$ and a corresponding generated caption $\mathcal{T}_{\mathcal{M}}{}^i$.

From an alternative perspective, the task of fact-checking meme captioning can be conceptualized as a modality transformation where the explanation $\mathcal{S}_{\mathcal{F}}$ for the claim's verdict $\mathcal{V}_{\mathcal{C}_{\mathcal{F}}}$, initially presented in textual form, is converted into a visual form. In particular, this visual form $\mathcal{I}_{\mathcal{M}}$ will incorporate embedded text $\mathcal{T}_{\mathcal{M}}$ such that the concatenation of both will represent the visual argumentative explanation $\mathcal{M}$.

4.2 System Requirements

In this section, we exhibit the requirements the system **should** comply with. The system should:

1. **R1**: Accept as input a claim, its verdict, the rationale and an optional meme image.
2. **R2**: Incorporate multimodal pre-training and augment data with contextual embeddings.
3. **R3**: Favor popular meme images that are suitable for Fact-Checking.
4. **R4**: Always generate several explanations.
5. **R5**: Not produce hateful content.

4.2.1 Creative Explanations Requirements

The generated creative explanations should:

1. **R6**: Use a mix of factual data, humor, and artistic content to ensure they are balanced and engaging.
2. **R7**: Not distort the facts being presented.
3. **R8**: Be accompanied by contextual information.
4. **R9**: Have a higher acceptance rate than text-based factual explanations.

[35] Example of a Fact-Checking article: https://shorturl.at/nPZ47

5. **R10**: Encourage users to seek out related fact-checking articles.
6. **R11**: Promote the user's beliefs closer to the truth.

4.2.2 Memes Requirements

The generated memes should:

1. **R12**: Have legible and concise captions that fit cohesively to the meme template.
2. **R13**: Be related to the subject of the claim.
3. **R14**: Be emotionally evoking.
4. **R15**: Explain the claim's verdict.
5. **R16**: Have high virality potential.

4.2.3 Evaluation Process Requirements

The evaluation process should:

1. **R17**: Present the claim clearly to users and only afterward the meme explanation.
2. **R18**: Not mention Creative Explanations for Fake News but rather explanations.
3. **R19**: Not mention that the explanations are machine-generated.

4.3 Architecture

This section discusses the system's architecture illustrated in Figure 4. The system automatically generates a meme (**R6**) given textual input data regarding a Fact-Checking article and possibly a meme image (**R1**). Module 1 represents the system's input already discussed in section 4.1. We will use PolitiFact articles that were web-scraped and collected in a GitHub repository.[36] We will later assess if we should update with more recent articles. We will also study if the IYTIS section has enough factual context for the LLM(s) to generate the captions or if they benefit more by using the full textual explanation. Regarding the input meme image, we will determine which availables images from the IMKG [38] to use as part of the proposal's scope. Additionally, we will study the optimal number of output-generated memes (**R4**). For each, we will prompt the LLM(s) in Module 3 with the information obtained from Module 2 and 2.1 to generate English captions. The latter modules are discussed in Section 4.3.2. Module 4 will concatenate each image to their meme caption via the ImgFlipAPI. At last, in Module 5, we incorporate a Hateful Memes Detection model to filter hateful content (**R5**). We will incorporate the model used in the MemeCraft system [87]. If any meme is flagged, we will prompt the LLMs in Module 3 again until we have enough generated memes (**R4**).

We have two main **ideas** for our system:

1. Multimodal LLMs Debate with Fine-Tuned Judge
2. Fine-Tuned LLM with Reinforcement Learning from Human Feedback (RLHF)

We will study the feasibility to integrate both methodologies for our time constraints. Regarding the first, we would deploy two LLMs tasked with generating captions and a third acting as a judge. The idea came from the system ExplainHM [73]. We would ask one generator, LLM, to generate a more factual caption while the other would generate a more humorous caption. The judge LLM, would be trained on a curated dataset from a crowdsourcing platform and select the optimal caption based on criteria such as humor, factual accuracy, and user engagement potential. This debate would foster creativity, circumvent answer refusals and captions repeating

[36] Fact-checking articles from https://github.com/hugo-ribeiro-36/Tese, originally sourced from https://github.com/crazyplayy/Fake-News-Creative-Explanations.

the prompts, and minimize bias and error using different models for the generator LLM. An ablation study against a basic architecture with only one LLM would confirm if this multimodal debate enhances performance. On the other hand, with the second idea, we would only have one generator LLM also fine-tuned on a curated dataset that would dynamically adjust its output from user interactions on the generated memes such as likes, evaluations, and shares (Module 6). This would align the meme generation more closely with human behavior. Like the first idea, we also do an ablation study against a basic architecture with a single fine-tuned LLM.

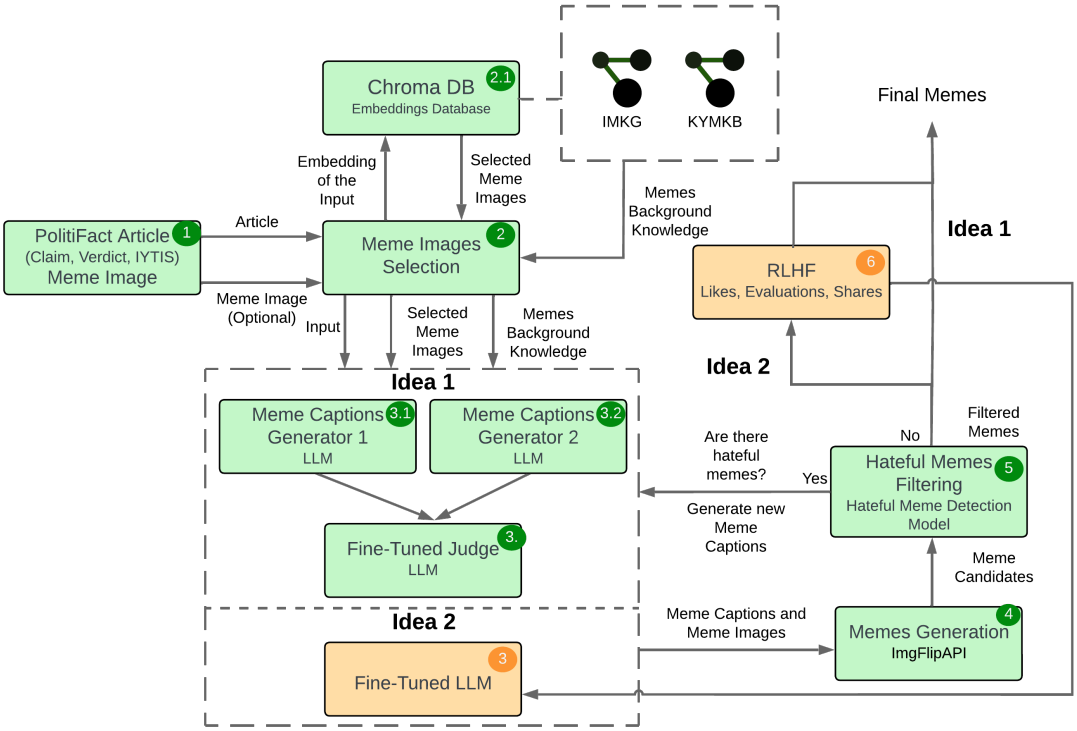

Figure 4: MemeFact's architecture.

4.3.1 LLM Selection

This section briefly discusses our strategy for selecting the most suitable LLMs for our architecture. One of the insights we obtained from our review of the related work suggested that training models with multimodal inputs enhanced performance across all meme-related tasks. Therefore, one of our initial pre-selection criteria for our LLMs will be the model's capability to process multimodal data (**R2**). Additionally, we will filter LLMs that have not been rigorously benchmarked on English language tasks. We will prioritize open-source models that are designed to have ease of integration. Furthermore, in line with the principles of GreenAI, we will prioritize models with fewer parameters when the results are similar within a specific threshold.

The models discussed in the literature review can be looked up in Table 1, 2, 3 and 5 in Section 3. These vary from text-based models such as GPT-2 [85] and 3.5 [105], and the LLaMa series [85]; to image-based models including Inception [106], ResNet [68], and VGG [107]; and well as multimodal models such as OpenAI's CLIP [65], BLIP [108] and GPT-4[37], MiniGPT4 [84], DeepMinds' Flamingo [83], IDEFICS, LLaVa [74], Visual BERT [63], ERNIE-ViL [109] and UNITER [110]. We will only be interested in the latter.

Nonetheless, we will discuss which LLMs will be tested. Post-selection criteria can be observed in Table 7 Section 6.1. These will focus on the content generated by the LLMs and not on their design characteristics.

[37] https://openai.com/index/gpt-4-research/

4.3.2 Retrieval Augmented Generation

This section explains our approach to implementing Retrieval Augmented Generation (RAG) to enhance the capabilities of our architecture by integrating relevant information from Internet Memes Knowledge Graph [38] and Know Your Meme Knowledge Base [86] discussed in Section 3.2.1 (**R2**). RAG addresses two problems with current LLMs: their training data is often out-of-date with the real world, and they make false but plausible-sounding statements (hallucinations) when they lack knowledge regarding the prompt. Both are risks for our architecture: memes are inherently dynamic, with their meanings evolving rapidly with cultural trends, and occurrences of hallucinations would further worsen disinformation instead of mitigating it (**R7**).

To counter these risks, we will augment the LLMs that generate captions with information such as the specific memes' image captioning style and their style of humor (Appendix C), their description, meaning, how they are used, and their popularity (**R3**). Some of this information is not readily available in the IMKG and KYMKB; therefore, we will enrich them with LLMs by artificially generating them. We also plan to add the triplets of the claim, verdict, and rationale that could have generated the existing meme examples.

To facilitate the retrieval, our architecture integrates a vector embedding database[38] (Module 2.1 in Figure 4), whose embeddings of the textual input are used to select suitable meme images for the textual content. These meme images will be obtained from the corresponding meme examples of the generated claims on the enriched KG whose embeddings are similar to one embedding of the textual input. For the use case, when the meme image is given, we will return only meme examples for the specific meme image.

Thus, we aim to prompt the LLMs responsible for generating the captions with our systems input and the meme examples retrieved from the embeddings database with their corresponding information. This includes the generated claims, the meme's style of humor, meaning, description, and the number of placeholders in the meme template. One limitation of our system is that neither IMKG nor KYMKB are updated in real-time with the latest meme trends. Without a robust scientific basis, we hypothesize that more recent meme trends would lead to better engaging memes. This will not be explored but is an important direction for future work.

5 First Steps

In this Section, we present hands-on work done with this thesis proposal. The academic MGSs reviewed in Section 3 were included in the Related Work of a paper submited to the Multimodal, Affective and Interactive eXplainable AI Workshop[39] that is part of the 27th European Conference on Artificial Intelligence. The paper in question, Towards AI-generated Memes for Misinformation Correction, co-authored by us, differs from ours regarding objectives by generating memes for false claims, therefore only creating corrective explanations. We further continued the study on these generation systems by investigating the quality of generated memes for our downstream task. We tested these systems, when possible, under the same meme image and PolitiFact article used for the meme in Figure 1. The prompt used is in Appendix A.

Table 4 shows that five academic systems published their code. Starting with News2Meme [92] we can observe its generated meme in Figure 6. We gave as input to the system the prompt only since it only allows textual input, and it generated a completely incoherent meme. We faced operational difficulties with several other systems: Dank Learning [93], Memeify [97], and DeepHumor [94] could not be tested due to issues such as deprecated code, recently deleted code, and a no longer available dataset, respectively. MemeCraft's [87] is compromised, since its code attempts to access models on HuggingFace that are no longer available. We find an overall lack of commitment toward ensuring reproducibility, maintaining systems, and providing adequate usage documentation.

Table 5 shows that three commercial systems provided API endpoints. PredisAI does not support immediate meme generation; it processes requests on its servers and requires users to wait for a webhook notification upon completion. Additionally, upon further inspection, we noticed Simplified has discontinued its API for meme generation, and AIMemeGenerator is currently unresponsive to requests.

[38] https://www.trychroma.com.
[39] https://sites.google.com/view/mai-xai24/home

6 Evaluation

This Section discusses how we plan to evaluate the memes generated with MemeFact. The literature surrounding MGSs mainly uses the BLEU score as a quantitative meme evaluation. It is relevant to assess coherence and if the generated caption is similar to the ones produced by humans for that specific meme image. However, it tells us nothing about explainability or hilarity. Recent literature has automatically evaluated these by leveraging SOTA LLMs [73]. To the best of our knowledge, there are no known automatic evaluation metrics for evaluating the quality of memes. Thus, due to the subjectivity of the task, we will solely use a qualitative evaluation, particularly through user surveys.

6.1 Qualitative Evaluation

We plan to do two user surveys:

- MemeFact's LLM Selection and Ablation Study
- Meme-based Fact-Checking Study

The first one will evaluate the generated memes that result from the LLM Selection study in Section 4.3.1, with the metrics in Table 7; the second evaluates the compliance with the creative explanations requirements from Section 4.2.1. Both surveys will consider the evaluation process requirements from Section 4.2.3. Table 6 summarizes the related work surrounding metrics for generated explanations and memes. We excluded metrics that were only used by one system from it. ExplainHM [73] uses Informativeness, which evaluates if the explanation provides new information irrelevant to our work. MemeCap's [81] Visual Completeness assesses if the caption describes all the important elements in the input image, which is also irrelevant since our target message is conveyed through text. On the other hand, Textual Completeness is similar to the latter, but for textual input, which is already covered with the Conveyance metric (CON). Faithfulness evaluates if the answer has made-up elements. We also consider it to be included within CON. MemeGera 2.0 [90] evaluated if the memes were surprising and Memeify [97] if they were original–both not relevant. Alternatively, MemeCraft [87] evaluates if the memes are offensive (Toxicity), which is important to measure compliance with **R5**. At last, the only system in the literature to address virality and engagement is MemeBot [95]. They asked users if they liked the generated memes (likeability). We mentioned in Section 2.5 that a meme's virality depends on the likeability and shareability potential users perceive. We will assess both these components to evaluate the compliance with **R16**, as we can observe in Table 7. However, one might question how accurately assessing memes for likeability and shareability through a user survey would mirror actual sharing behavior on a social network. We considered proposing to deploy a bot account on X to test the engagement of the generated memes. Given the constraints of the proposed timeline, we decided to leave this as a bonus objective, acknowledging the additional credibility it would lend to our findings. The second user survey will evaluate compliance with the Creative Explanation Requirements outlined in Section 4.2.1. As part of subsequent work, we will detail the methodology for explicitly measuring the latter, as well as the structure of the questions for both surveys.

7 Work Schedule

This Section outlines the work schedule for the thesis, detailing the critical stages and timelines involved. The project is structured into several key phases: the Architecture Setup, the Development of the Dataset, the LLM Selection Study, the Ablation Study, the User Surveys, and the writing. A Gantt chart is included in Figure 5 to visually depict these phases and their expected completion times, providing a clear roadmap for the project's progression.

8 Conclusion

In this thesis proposal, we have highlighted the problems with current fact-checking approaches and further explained how our innovative meme-based fact-checking approach could be useful in tackling these. By con-

Table 6: Evaluation Metrics from the Related Work.

System	COH	CON	HUM	DIFF	PERS	READ	CONC	SUIT
PoetryCE [45]	✓	✓	✓			✓		
ExplainHM [73]	✓	✓			✓	✓	✓	
MemeCap [81]		✓					✓	
Automatic Chinese [85]	✓		✓	✓				
MemeCraft [87]		✓	✓	✓	✓			
MemeGera 2.0 [90]	✓		✓					✓
Dank Learning [93]			✓	✓				
DeepHumor [94]	✓		✓	✓				
MemeBot [96]		✓						✓
Memeify [97]	✓		✓	✓				

Legend:

COH	**Coherence** - The generated meme's caption is syntactically and semantically coherent.
CON	**Conveyance** - The generated meme communicates the intended message.
HUM	**Humor** - The generated meme is hilarious.
DIFF	**Differentiability** - The generated meme resembles a human-generated one.
PERS	**Persuasiveness** - The generated meme is persuasive.
READ	**Readability** - The generated meme's caption is easy to read.
CONC	**Conciseness** - The generated meme's caption is succinct.
SUIT	**Suitability** - The generated meme's image is appropriate for the intended context.
LIKE	**Likeability** - The generated meme is appealing and enjoyable.
SHAR	**Shareability** - The generated meme is likely to be shared.

Table 7: Metrics to evaluate Meme's Requirements.

Reqs.	COH	CON	LIKE	SHAR	HUM	PERS	READ	CONC	TOX
R5									✓
R12							✓	✓	✓
R13	✓	✓							
R14					✓				
R15		✓				✓			
R16			✓	✓					

ducting an extensive cross-domain literature review, we identified the critical requirements that support the architecture of the proposed MemeFact system. We seek to transform how fact-checking content is perceived and consumed, making it more accessible, understandable, and engaging for the general public. This proposal represents more than just another AI system. It manifests our commitment to AI for Social Good.

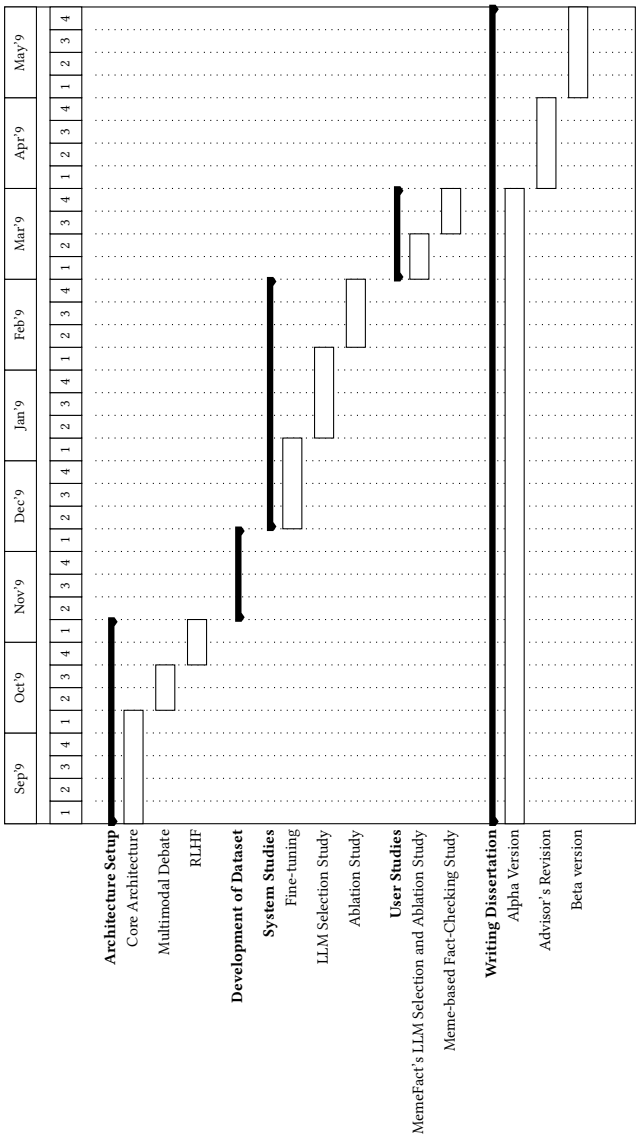

Figure 5: Thesis Schedule Proposal.

Bibliography

[1] W. Scheidel, "Orbis: The stanford geospatial network model of the roman world," 2015.

[2] W. Weir, *History's Greatest Lies: The Startling Truths Behind World Events Our History Books Got Wrong*. Quarto Publishing Group USA, 2009.

[3] R. R. Dynes, "The lisbon earthquake in 1755: Contested meanings in the first modern disaster," 1997.

[4] J. W. Cortada and W. Aspray, *Fake news nation: the long history of lies and misinterpretations in America*. Rowman Littlefield, 2019.

[5] L. Lamport, R. Shostak, and M. Pease, *The Byzantine generals problem*, 2019, pp. 203–226.

[6] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news: Addressing fake news requires a multidisciplinary effort," *Science*, vol. 359, pp. 1094–1096, 3 2018.

[7] S. Issacharoff, "Democracy and collective decision making," *International Journal of Constitutional Law*, vol. 6, pp. 231–266, 4 2008. [Online]. Available: https://doi.org/10.1093/icon/mon003

[8] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nature communications*, vol. 10, p. 7, 2019.

[9] A. M. Guess, B. Nyhan, and J. Reifler, "Exposure to untrustworthy websites in the 2016 us election," *Nature human behaviour*, vol. 4, pp. 472–480, 2020.

[10] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *arXiv preprint arXiv:1707.00086*, 2017.

[11] I. Montagni, K. Ouazzani-Touhami, A. Mebarki, N. Texier, S. Schück, C. Tzourio, and the CONFINS group, "Acceptance of a covid-19 vaccine is associated with ability to detect fake news and health literacy," *Journal of Public Health*, vol. 43, pp. 695–702, 12 2021. [Online]. Available: https://doi.org/10.1093/pubmed/fdab028

[12] A. Bridgman, E. Merkley, P. J. Loewen, T. Owen, D. Ruths, L. Teichmann, and O. Zhilin, "The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media," *Harvard Kennedy School Misinformation Review*, vol. 1, 2020.

[13] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, pp. 337–348, 2021.

[14] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, "The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review," *Journal of Public Health*, pp. 1–10, 2021.

[15] D. D. Coninck, T. Frissen, K. Matthijs, L. d'Haenens, G. Lits, O. Champagne-Poirier, M.-E. Carignan, M. D. David, N. Pignard-Cheynel, and S. Salerno, "Beliefs in conspiracy theories and misinformation about covid-19: Comparative perspectives on the role of anxiety, depression and exposure to and trust in information sources," *Frontiers in psychology*, vol. 12, p. 646394, 2021.

[16] A. Swift, "Americans' trust in mass media sinks to new low," *Gallup News*, vol. 14, 2016.

[17] G. Pennycook and D. G. Rand, "The psychology of fake news," pp. 388–402, 5 2021.

[18] E. Shearer and A. Mitchell, "News use across social media platforms in 2020," 2021.

[19] A. Enders, H. Hungenberg, H.-P. Denker, and S. Mauch, "The long tail of social networking.: Revenue models of social networking sites," *European Management Journal*, vol. 26, pp. 199–211, 2008.

[20] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, vol. 348, pp. 1130–1132, 2015.

[21] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, "The spread of fake news by social bots," *arXiv preprint arXiv:1707.07592*, vol. 96, p. 104, 2017.

[22] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," 2014, pp. 18–22.

[23] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management science*, vol. 66, pp. 4944–4957, 2020.

[24] R. H. Grady, P. H. Ditto, and E. F. Loftus, "Nevertheless, partisanship persisted: Fake news warnings help briefly, but bias returns with time," *Cognitive research: principles and implications*, vol. 6, pp. 1–16, 2021.

[25] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general*, vol. 147, p. 1865, 2018.

[26] O. Barrera, S. Guriev, E. Henry, and E. Zhuravskaya, "Facts, alternative facts, and fact checking in times of post-truth politics," *Journal of Public Economics*, vol. 182, p. 104123, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047272719301859

[27] N. Walter, J. Cohen, R. L. Holbert, and Y. Morag, "Fact-checking: A meta-analysis of what works and for whom," *Political Communication*, vol. 37, pp. 350–375, 2020.

[28] "Deliverable 1.2-user-centred requirements project title countering creative information manipulation with explainable ai." [Online]. Available: http://cimple.eu/

[29] T. Wood and E. Porter, "The elusive backfire effect: Mass attitudes' steadfast factual adherence," *Political Behavior*, vol. 41, pp. 135–163, 2019.

[30] Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, "The roll-out of community notes did not reduce engagement with misinformation on twitter," 7 2023. [Online]. Available: http://arxiv.org/abs/2307.07960

[31] R. Dawkins, *The selfish gene*. Oxford university press, 2016.

[32] L. Shifman, *Memes in digital culture*. MIT press, 2013.

[33] J. T. Nieubuurt, "Internet memes: leaflet propaganda of the digital age," *Frontiers in Communication*, p. 116, 2021.

[34] G. Burbi, A. Baldrati, L. Agnolucci, M. Bertini, and A. D. Bimbo, "Mapping memes to words for multimodal hateful meme classification," 2023, pp. 2832–2836.

[35] U. Akram, K. Irvine, S. F. Allen, J. C. Stevenson, J. G. Ellis, and J. Drabble, "Internet memes related to the covid-19 pandemic as a potential coping mechanism for anxiety," *Scientific reports*, vol. 11, p. 22305, 2021.

[36] R. Vacca, K. DesPortes, M. Tes, M. Silander, A. Amato, C. Matuk, and P. J. Woods, "What do you meme? students communicating their experiences, intuitions, and biases surrounding data through memes," 2022, pp. 212–224.

[37] A. E. Msugheter, "Internet meme as a campaign tool to the fight against covid-19 in nigeria," *Global Journal of Human-Social Science: A Arts Humanities–Psychology*, vol. 20, pp. 27–39, 2020.

[38] R. Tommasini, F. Ilievski, and T. Wijesiriwardene, "Imkg: The internet meme knowledge graph." Springer, 2023, pp. 354–371.

[39] Aristotle and W. R. Roberts, *Rhetoric*. Modern Library, 1954.

[40] F. Altoe, C. Moreira, H. S. Pinto, and J. A. Jorge, "Online fake news opinion spread and belief change: A systematic review," *Human Behavior and Emerging Technologies*, vol. 2024, p. 1069670, 2024. [Online]. Available: https://doi.org/10.1155/2024/1069670

[41] C. Ling, I. AbuHilal, J. Blackburn, E. D. Cristofaro, S. Zannettou, and G. Stringhini, "Dissecting the meme magic: Understanding indicators of virality in image memes," *Proceedings of the ACM on human-computer interaction*, vol. 5, pp. 1–24, 2021.

[42] F. Fact, "Fact checking doesn't work (the way you think it does)," *Full Fact*, 2019.

[43] V. Taecharungroj and P. Nueangjamnong, "Humour 2.0: Styles and types of humour and virality of memes on facebook," *Journal of Creative Communications*, vol. 10, pp. 288–302, 2015.

[44] J. C. Richmond and D. V. Porpora, "Entertainment politics as a modernist project in a baudrillard world," *Communication Theory*, vol. 29, pp. 421–440, 2019.

[45] A. F. A. Santos, "Fake news creative explanations through the use of poetry a comparison study and fine-tuning approach," 2022, thesis for Mestrado em Engenharia Informática e de Computadores. [Online]. Available: https://fenix.tecnico.ulisboa.pt/cursos/meic-a/dissertacao/565303595503110

[46] J. P. Guilford, "Fundamental statistics in psychology and education," 1950.

[47] A. Koestler, "The act of creation," 1964.

[48] M. A. Boden, "Computer models of creativity," *AI Magazine*, vol. 30, pp. 23–34, 2009.

[49] M. ABoden, "The creative mind: Myths and mechanisms, second edition."

[50] M. D. Mumford, W. A. Baughman, and C. E. Sager, *Picking the right material: Cognitive processing skills and their role in creative thought.* Hampton Press, 2003, pp. 19–68.

[51] M. Boden, "Artificial intelligence and natural man," *Synthese*, vol. 43, 1980.

[52] S. Colton and G. A. Wiggins, "Computational creativity: The final frontier?" vol. 12. Montpelier, 2012, pp. 21–26.

[53] G. A. Wiggins, "A preliminary framework for description, analysis and comparison of creative systems," *Knowledge-Based Systems*, vol. 19, pp. 449–458, 11 2006.

[54] T. Veale and F. A. Cardoso, *Computational creativity: The philosophy and engineering of autonomously creative systems.* Springer, 2019.

[55] A. Gabriel, D. Monticolo, M. Camargo, and M. Bourgault, "Creativity support systems: A systematic mapping study," *Thinking Skills and Creativity*, vol. 21, pp. 109–122, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1871187116300293

[56] K. Wang and J. V. Nickerson, "A literature review on individual creativity support systems," *Computers in Human Behavior*, vol. 74, pp. 139–151, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563217302777

[57] P. Quintas and H. S. Pinto, "Report on the state of the art on creative xai," 2022.

[58] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, pp. 1–42, 2018.

[59] F. Altoe and H. S. Pinto, "Towards a personalized online fake news taxonomy," 2023. [Online]. Available: https://doi.org/10.1145/3565472.3592963

[60] W. Ruch, S. Heintz, T. Platt, L. Wagner, and R. T. Proyer, "Broadening humor: Comic styles differentially tap into temperament, character, and ability," *Frontiers in Psychology*, vol. 9, p. 6, 2018.

[61] S. Heintz and W. Ruch, "From four to nine styles: An update on individual differences in humor," *Personality and Individual Differences*, vol. 141, pp. 7–12, 2019.

[62] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[63] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[64] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context." Springer, 2014, pp. 740–755.

[65] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark, "Learning transferable visual models from natural language supervision." PMLR, 2021, pp. 8748–8763.

[66] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, and T. M. Ghazal, "Multimodal hate speech detection in memes using contrastive language-image pre-training," *IEEE Access*, 2024.

[67] C. Breazzano, E. Rubino, D. Croce, and R. Basili, "Unitor@ dankmemes: Combining convolutional models and transformer-based architectures for accurate meme management," vol. 2765. CEUR-WS, 2020.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, pp. 770–778.

[69] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen, "Semeval-2022 task 5: Multimedia automatic misogyny identification," 2022, pp. 533–549.

[70] B. Grasso, V. L. Gatta, V. Moscato, and G. Sperlì, "Kermit: Knowledge-empowered model in harmful meme detection," *Information Fusion*, vol. 106, p. 102269, 2024.

[71] V. Kougia, S. Fetzel, T. Kirchmair, E. Çano, S. M. Baharlou, S. Sharifzadeh, and B. Roth, "Memegraphs: Linking memes to knowledge graphs." Springer, 2023, pp. 534–551.

[72] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[73] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang, "Towards explainable harmful meme detection through multimodal debate between large language models," *arXiv preprint arXiv:2401.13298*, 2024.

[74] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[75] T. Milo, A. Somech, and B. Youngmann, "Simmeme: A search engine for internet memes." IEEE, 2019, pp. 974–985.

[76] S. Toh, A. Kuek, W.-H. Chong, and R. K.-W. Lee, "Mermaid: A dataset and framework for multimodal meme semantic understanding." IEEE, 2023, pp. 433–442.

[77] S. Sharma, U. Arora, M. S. Akhtar, and T. Chakraborty, "Memex: Detecting explanatory evidence for memes via knowledge-enriched contextualization," *arXiv preprint arXiv:2305.15913*, 2023.

[78] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[79] S. Sharma, S. Agarwal, T. Suresh, P. Nakov, M. S. Akhtar, and T. Chakraborty, "What do you meme? generating explanations for visual semantic role labelling in memes," vol. 37, 2023, pp. 9763–9771.

[80] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[81] E. Hwang and V. Shwartz, "Memecap: A dataset for captioning and interpreting memes," *arXiv preprint arXiv:2305.13703*, 2023.

[82] N. Prakash, H. Wang, N. K. Hoang, M. S. Hee, and R. K.-W. Lee, "Promptmtopic: Unsupervised multimodal topic modeling of memes using large language models," 2023, pp. 621–631.

[83] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, and M. Reynolds, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[84] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[85] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, p. 9, 2019.

[86] L. Bates, P. E. Christensen, P. Nakov, and I. Gurevych, "A template is all you meme," *arXiv preprint arXiv:2311.06649*, 2023.

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

[88] W. Y. Wang and M. Wen, "I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions," 2015, pp. 355–365.

[89] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," 2002, pp. 311–318.

[90] H. G. Oliveira, D. Costa, and A. M. Pinto, "One does not simply produce funny memes!–explorations on the automatic generation of internet humor," 2016.

[91] J. P. Lopes, J. M. Cunha, and P. Martins, "Stonkinator: An automatic generator of memetic images."

[92] E. K. Shimomoto, L. S. Souza, B. B. Gatto, and K. Fukui, "News2meme: An automatic content generator from news based on word subspaces from text and image." IEEE, 2019, pp. 1–6.

[93] A. L. P. V and E. M. Tolunay, "Dank learning: Generating memes using deep neural networks," *arXiv preprint arXiv:1806.04510*, 2018.

[94] I. Borovik, B. Khabibullin, V. Kniazev, Z. Pichugin, and O. Olaleke, "Deephumor: Image-based meme generation using deep learning."

[95] L. Wang, Q. Zhang, Y. Kim, R. Wu, H. Jin, H. Deng, P. Luo, and C.-H. Kim, "Automatic chinese meme generation using deep neural networks," *IEEE Access*, vol. 9, pp. 152 657–152 667, 2021.

[96] A. Sadasivam, K. Gunasekar, H. Davulcu, and Y. Yang, "Memebot: Towards automatic image meme generation," *arXiv preprint arXiv:2004.14571*, 2020.

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

[98] L. L. Jacobi, "Perceptions of profanity: How race, gender, and expletive choice affect perceived offensiveness." *North American Journal of Psychology*, vol. 16, 2014.

[99] A. Mendiburo-Seguel, D. Páez, and F. Martínez-Sánchez, "Humor styles and personality: A meta-analysis of the relation between humor styles and the big five personality traits," *Scandinavian journal of psychology*, vol. 56, pp. 335–340, 2015.

[100] E. C. Choi and E. Ferrara, "Fact-gpt: Fact-checking augmentation via claim matching with llms," *arXiv preprint arXiv:2402.05904*, 2024.

[101] B. He, M. Ahamad, and S. Kumar, "Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation," 2023, pp. 2698–2709.

[102] M. Hameleers, T. E. Powell, T. G. L. A. V. D. Meer, and L. Bos, "A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media," *Political communication*, vol. 37, pp. 281–301, 2020.

[103] M. A. Amazeen, E. Thorson, A. Muddiman, and L. Graves, "Correcting political and consumer misperceptions," *Journalism Mass Communication Quarterly*, 2017.

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

[105] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[106] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2015, pp. 1–9.

[107] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[108] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." PMLR, 2022, pp. 12 888–12 900.

[109] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs," vol. 35, 2021, pp. 3208–3216.

[110] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning." Springer, 2020, pp. 104–120.

A Drake Hotline Bling Meme Prompt

The prompt used was:

> *This claim is mostly false: "The biggest threat to your unions is millions of people coming across the border, because you're not gonna have your jobs anymore." The if-your-time-is-short explanation by Politifact is the following: "Economy and labor experts told PolitiFact immigrants who recently crossed the U.S. border likely aren't taking Michigan's union jobs. Instead, newly arrived migrants are likelier to work in jobs Americans don't want to do, such as day laborer positions. These aren't union jobs. There is a correlation between an increase in immigration and a drop in unionization. However, experts said that's not evidence that immigrants are taking union jobs." Now generate a meme reproducing this explanation.*

B CIMPLE Guidelines

- **GL1**: Creative Explanations should present a balanced version of the truth, and a serious attempt should be made to not distort the facts being presented (...) (**R7**).
- **GL2**: Creative Explanations should not present raw corrective information before context is given (**R17**). (...) In a broader sense, this guideline draws attention to the fact that the order in which information is presented within a Creative Explanation is of great importance.
- **GL3**: Creative Explanations should use a combination of factual information, information about its inner workings, humor, and artistic content in order to be balanced and interesting to interact with. No single Creative Explanation should rely entirely on a single strategy or argument to transmit the information to the user (...) (**R6**).
- **GL4**: Creative Explanations should play a role in educating their audience. Many experts from the field of journalism believe that the key to fighting the current "misinformation pandemic" is through proper education of the masses regarding online information sanity (...) (**R10**).

C Humor Definitions

D Politifact Verdict Labels

Table 9: Politifact classification labels definitions.

Label	Definition
True	The statement is accurate and there is nothing significant missing.
Mostly True	The statement is accurate but needs clarification or additional information.
Half True	The statement is partially accurate but leaves out important details or takes things out of context.
Mostly False	The statement contains an element of truth but ignores critical facts that would give a different impression.
False	The statement is not accurate.
Pants on Fire	The statement is not accurate and makes a ridiculous claim.

Table 8: Definitions of the Styles of Humor (†) and Comic Styles (‡).

Concept	Definition
Affiliative †	Creators of affiliative humor memes say funny things, jokes, and witty banters to amuse others and facilitate relationships.
Aggressive †	Creators of aggressive humor memes express humor without regard for its impact on others by saying funny things that are likely to hurt or alienate others.
Self-defeating †	Creators of self-defeating humor memes amuse others by doing or saying humorous and disparaging things at one's own expense.
Self-enhancing †	Creators of self-enhancing humor memes have a humorous outlook on life, are amused by incongruities and maintain a humorous perspective in adversity.
Comparison ‡	Putting two or more elements together to produce a humorous situation.
Exaggeration ‡	Overstating and magnifying something out of proportion.
Personification ‡	Attributes human characteristics to animals, plants, and objects.
Pun ‡	Using elements of language to create new meanings, which result in humor.
Sarcasm ‡	Blatant ironic responses or situations.
Silliness ‡	Making funny faces to ludicrous situations.
Surprise ‡	Humor arises from unexpected situations.

E Generated Memes for the Related Work Systems

Figure 6: News2Meme System generated meme.