FACTIFY3M: A Benchmark for Multimodal Fact Verification with Explainability through 5W Question-Answering

Megha Chakraborty[1]   Khusbu Pahwa [2]   Anku Rani[1]
Adarsh Mahor[1]   Aditya Pakala[1]   Arghya Sarkar[1]
Harshit Dave[1]   Ishan Paul[1]   Janvita Reddy[1]
Preethi Gurumurthy[1]   Ritvik G[1]   Samahriti Mukherjee[1]
Shreyas Chatterjee[1]   Kinjal Sensharma[1]   Dwip Dalal[1]
Suryavardan S[1]   Shreyash Mishra[1]   Parth Patwa[2]
Aman Chadha†[3,4]   Amit Sheth[1]   Amitava Das[1]
[1]University of South Carolina, USA   [2]UCLA, USA
[3]Amazon AI, USA   [4]Stanford University, USA
meghac@email.sc.edu   amitava@mailbox.sc.edu

Abstract

Combating disinformation is one of the burning societal crises - about 67% of the American population believes that disinformation produces a lot of uncertainty, and 10% of them knowingly propagate disinformation. Evidence shows that disinformation can manipulate democratic processes and public opinion, causing disruption in the share market, panic and anxiety in society, and even death during crises. Therefore, disinformation should be identified promptly and, if possible, mitigated. With approximately 3.2 billion images and 720,000 hours of video shared online daily on social media platforms, scalable detection of multimodal disinformation requires efficient fact verification. Despite progress in automatic text-based fact verification (e.g., FEVER, LIAR), the research community lacks substantial effort in multimodal fact verification. To address this gap, we introduce FACTIFY 3M, a dataset of 3 million samples that pushes the boundaries of the domain of fact verification via a multimodal fake news dataset, in addition to offering explainability through the concept of 5W question-answering. Salient features of the dataset include: *(i) textual claims, (ii) ChatGPT-generated paraphrased claims, (iii) associated images, (iv) stable diffusion-generated additional images (i.e., visual paraphrases), (v) pixel-level image heatmap to foster image-text explainability of the claim, (vi) 5W QA pairs, and (vii) adversarial fake news stories.*

1 FACTIFY 3M - an illustration

We introduce FACTIFY 3M (*3 million*), the largest dataset and benchmark for multimodal fact verification.

Consider the example in fig. 1. A widely distributed image of the sports legend Magic Johnson with an IV line in his arm was accompanied by the claim that he was donating blood, implicitly during the COVID-19 pandemic. If true, this is troubling because Magic Johnson is a well-known victim of AIDS and is prohibited from donating blood. The picture predated the COVID-19 epidemic by a decade and is related

---

†Work does not relate to position at Amazon.

PromptFake3M at a glance									
	Entailment classes	Textual Support	Visual/Image Support	No. of claims	No. of paraphrased claims	No. of images	No. of stable diffusion generated images	5WQA pairs	No. of evidence documents	Adversarial OPT-generated news story
Support	Support_Multimodal	Texts are supporting each other ~similar news	Images are supporting each other	232,000	882,000	232,000	927,000	858,400	232,000	✗
	Support_Text	Texts are supporting each other ~similar news	Images are neither supporting nor refuting	174,000	609,000	169,000	661,000	852,600	174,000	✗
Neutral	Insufficient_Multimodal	Texts are neither supported nor refuted ~may have common words	Images are supporting each other	99,000	366,000	99,000	347,000	375,000	99,000	✗
	Insufficient_Text	Texts are neither supported nor refuted ~may have common words	Images are neither supporting nor refuting	126,000	525,000	123,000	466,000	441,000	126,000	✗
Fake	Refute	Fake claim	Fake image support	316,000	1,193,000	309,000	916,400	1,327,000	316,000	✓135,000
Total				947,000	3,575,000	932,000	3,317,400	3,954,000	947,000	135,000

Table 1: A top-level view of FACTIFY 3M: (i) classes and their respective textual/visual support specifics, (ii) number of claims, paraphrased claims, associated images, generated images, 5W pairs, evidence documents, and adversarial stories.

Figure 1: A tweet referring to the sports personality and known AIDS victim Magic Johnson with a photo that was taken a decade before COVID and, moreover, is not real evidence of blood donation.

to his treatment for AIDS.

Textual claim: The text associated with the claim in fig. 1 purports that the author of this tweet took this photo and assisted Magic Johnson in donating blood. The further implicit declaration is that he is a medical worker and possibly works for a hospital that Magic Johnson visited for blood donation.

ChatGPT-paraphrased claims: To emulate the textual diversity in how news publishing houses report day-to-day happenings, it is important to produce data representative of such variety. Variety in text is embodied through a difference in narrative styles, choice of words, ways of presenting factual information, etc. For e.g., fig. 5a shows a claim and document that address the same topic but differ in their textual prose. To mimic this aspect, we adopted ChatGPT as a paraphraser and generated claims. Fig. 3 shows an example of a claim paraphrased using ChatGPT (OpenAI, 2022).

Associated images: The image included as part of the claim (refer fig. 1 for the image embedded in the tweet and fig. 5a for images included as part of the claim) improves its trustworthiness perception since humans tend to believe visual input much more than mere text prose. Moreover, the text and image compo-

5W QA based Explainability				
Who claims	**What claims**	**When claims**	**Where claims**	**Why claims**
• **Q1**: *Who went to the hospital?* Ans: Magic Johnson • **Q2**: *Who worked with whom?* Ans: the author with Magic Johnson • **Q3**: *Who took the photo?* Ans: the author	• **Q1**: *What did Magic Johnson do at the hospital?* Ans: donated blood • **Q2**: *What process Magic Johnson was part of?* Ans: blood donation	• **Q1**: *When did Magic Johnson visit the hospital?* Ans: last month time of the post = Sept - 1 month from August	• **Q1**: *Where did Magic Johnson pay visit to?* Ans: hospital.	• **Q1**: *Why did Magic Johnson visit hospital?* Ans: to donate blood.
⚠ **caution**	❌ **verified false**	⚠ **caution**	❓ **not verifiable**	❌ **verified false**
Evidence				
• **news1** - url1 - Magic Johnson visits hundreds of kids at SC hospital, on Dec 10, 2019 • No information is available that the author worked with Magic Johnson • No information is available about who took this photo.	• **news1** - url1 - Magic Johnson Shuts Down 'False Story' He Donated Blood • **news 2** - url2 - Magic Johnson **didn't donate blood** to help fight COVID-19 related story - • **news 3** - url3 - Magic Johnson opens up on living with HIV 30 years	• **news1** - url1 - Magic Johnson visits hundreds of kids at SC hospital, on Dec 10, 2019 • **news2** - url2 - Magic Johnson at Howard University Hospital, on Feb 7, 2013	• **news1** - url1 - Magic Johnson visits hundreds of kids at SC hospital, on Dec 10, 2019 • **news2** - url2 - Magic Johnson at Howard University Hospital, on Feb 7, 2013	• **news1** - url1 - Magic Johnson Shuts Down 'False Story' He Donated Blood • **news 2** - url2 - Magic Johnson **didn't donate blood** to help fight COVID-19

Figure 2: An illustration of the proposed 5W QA-based explainable fact verification system. This example illustrates the false claim shown in fig. 1. A typical semantic role labeling (SRL) system processes a sentence and identifies verb-specific semantic roles. Therefore, for the specified example, we have 3 sentences: sentence 1 has two main verbs *work* and *come*, sentence 2 has one verb *meet*, and sentence 3 has one verb *take*. For each verb, a 5W QA pair will be automatically generated (4 × 5 = 20 sets of QA pairs in total for this example). Furthermore, all those 20 5W aspects will be fact-checked. If some aspects end up having *neutral* entailment verdict, possible relevant documents with associated URLs will be listed for the end user to read further and assess. In addition, a reverse image search result will be shown to aid human fact-checkers further.

nents together provide a holistic claim assertion, similar to how news articles convey world updates.

Stable Diffusion-generated additional images aka visual paraphrases: Referring to fig. 5a, the diversity of images associated with the claim and document are apparent. Specifically, the image associated with the claim is that of a frontal face, while that associated with the document, while still the same person, is one of a not-so-visible face but includes multiple other entities. Such diversity is commonly seen in the wild when different news media houses cover the same news. As such, we try to emulate this

Sports star Magic Johnson came to the hospital last month to donate blood to support the COVID-19 crisis.
Prphr 1: Last month, Magic Johnson, a famous athlete, visited a hospital to give blood to help with the COVID-19 pandemic.
Prphr 2: Magic Johnson, a well-known sports figure, went to a hospital last month to donate blood to assist with the COVID-19 crisis.
Prphr 3: Magic Johnson, a sports celebrity, recently visited a hospital and donated blood to contribute to the COVID-19 crisis.
Prphr 4: Recently, Magic Johnson, a well-known sports star, went to a hospital and gave blood to support the COVID-19 crisis.
Prphr 5: To support the COVID-19 crisis, Magic Johnson, a sports celebrity, came to a hospital last month to donate blood.
Prphr 6: In an effort to support the COVID-19 crisis, Magic Johnson, a well-known sports figure, visited a hospital last month to donate blood.

Figure 3: Claims paraphrased using ChatGPT (OpenAI, 2022) to foster their textual diversity.

Sports star Magic Johnson came to the hospital last month to donate blood to support the COVID-19 crisis. He shared his experience on social media, expressing his appreciation to the medical staff at the hospital and encouraging others to donate blood. He said, "I just donated blood at the hospital to help those affected by the coronavirus pandemic. I want to thank all the doctors, nurses, and medical staff for their hard work and dedication. I urge everyone to step up and donate blood if they are able. Every donation makes a difference."

Figure 4: An example of OPT (Zhang et al., 2022) generated fake news that confirms the Magic Johnson blood donation incident.

aspect in our work by harnessing the power of the latest in text-to-image generation. We generate additional images using Stable Diffusion (Rombach et al., 2021). Fig. 5b shows the diversity of generated images in terms of the camera angle, subject, entities, etc., which in turn offers enhanced visual diversity for a multimodal claim.

Pixel-level image heatmap: To clearly delineate and hence explain which aspect of the image is being referred to in the various components of the text caption, we generate pixel-level attribution heatmaps to foster explainability. For e.g., referring to fig. 5c, the heatmap highlights gothic-architecture buildings for the word *capitol* which the building is known for, and highlights the human figure for the term *investigating*. A dataset of this kind would be very helpful in designing explainable multimodal fact verification and possibly visual question-answer-based fact verification.

5WQA: The process of fact verification is inherently intricate, with several questions representing the components within the underlying claim that need answers to reach a verdict on the veracity of the claim. Referring to the example in fig. 1, such questions may include: *(a) who donated blood? (b) when did he donate blood? (c) can Magic Johnson donate blood? (d) what can go wrong if this claim is false?* Manual fact-checking can be labor-intensive, consuming several hours or days (Hassan et al., 2015; Adair et al., 2017).

Contemporary automatic fact-checking systems focus on estimating truthfulness using numerical scores, which are not human-interpretable. Others extract explicit mentions of the candidate's facts in the text as evidence for the candidate's facts, which can be hard to spot directly. Only two recent works (Yang et al., 2022; Kwiatkowski et al., 2019) propose
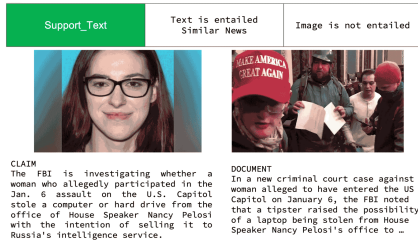
(a) Another example of covering the same news event by two news media houses. Here the same alleged lady is visible in both images, but altogether two images are different, and the text is paraphrased differently.

(b) Stable Diffusion output for the above claim.
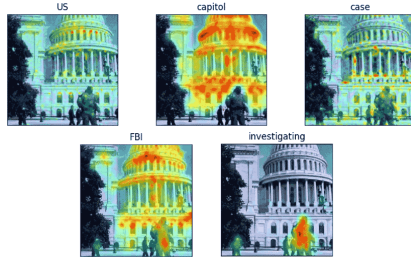
(c) DAAM (Tang et al., 2022) explanation for the above claim.

Figure 5: An example from PromptFake 3M dataset. Multimodal fact verification is a challenging endeavor considering the intricacies of real-life data, where the entailment of images requires understanding the nuances of day-to-day life situations. As such, multimodal entailment is an under-research paradigm with limited progress; current SoTA systems lack the finesse needed to handle the complexity portrayed in these previous examples adeptly.

question answering as a proxy to fact verification explanation, breaking down automated fact-checking into several steps and providing a more detailed analysis of the decision-making processes. Question-answering-based fact explainability is indeed a very promising direction. However, open-ended QA for a fact can be hard to summarize. Therefore, we refine the QA-based explanation using the 5W framework (*who, what, when, where, and why*). Journalists follow an established practice for fact-checking, verifying the so-called 5Ws (Mott, 1942), (Stofer et al., 2009), (Silverman, 2020), (Su et al., 2019), (Smarts, 2017), (Wiki_Article, 2023). This directs verification search and, moreover, identifies missing content in the claim that bears on its validity. One consequence of journalistic practice is that claim rejection is not a matter of degree (as conveyed by popular representations such as a number of Pinocchios or crows, or true, false, half true, half false, pants on fire), but the rather specific, substantive explanation that recipients can themselves evaluate (Dobbs, 2012). Please refer to fig. 2 to look at the possible 5W QA questionnaire for the claim in fig. 1.

Adversarial fake news: Fact verification systems are only as good as the evidence they can reference while verifying a claim's authenticity. Over the past decade, with social media having mushroomed into the masses' numero-uno choice of obtaining world news, fake news articles can be one of the biggest bias-inducers to a person's outlook towards the world. To this end, using the SoTA language model, we generate adversarial news stories to offer a new benchmark that future researchers can utilize to certify the performance of their fact verification systems against adversarial news.

Programmatic detection of AI-generated writing (where an AI is the sole author behind the

article) and its more challenging cousin – AI-assisted writing (where the authorship of the article is split between an AI and a human-in-the-loop) – has been an area of recent focus. While detecting machine-generated text from server-side models (for instance, GPT-3 (Brown et al., 2020), which is primarily utilized through an API, uses techniques like watermarking (Wiggers, 2022b)) is still a topic of investigation, being able to do so for the plethora of open-source LLMs available online is a herculean task. Our adversarial dataset will offer a testbed so that such detection advances can be measured against with the ultimate goal of curbing the proliferation of AI-generated fake news.

2   Related works - data sources and compilation

Automatic fact verification has received significant attention in recent times. Several datasets are available for text-based fact verification, e.g., FEVER (Thorne et al., 2018), Snopes (Vo and Lee, 2020), PolitiFact (Vo and Lee, 2020), FavIQ (Kwiatkowski et al., 2019), HoVer (Jiang et al., 2020), X-Fact (Gupta and Srikumar, 2021), CREAK (Onoe et al., 2021), FEVEROUS (Aly et al., 2021), etc.

Multimodal fact verification has recently started gaining momentum. DEFACTIFY workshop series at AAAI 2022 (Mishra, 2022) and 2023 (Suryavardhan, 2023) has released FACTIFY 1.0 (Mishra et al., 2022) and 2.0 (Mishra et al., 2023) with 50K annotated data each year, which we have embedded as well as part of FACTIFY 3M. Fact verification datasets are mainly classified into three major categories: *(i) support, (ii) neutral, and (iii) refute*. While it is relatively easier to collect data for *support* and *neutral* categories, collecting large-scale *refute* category aka fake news/ claims is relatively

challenging. FEVER (Thorne et al., 2018) - proposed an alternative via manual imaginary claim generation - is complex, strives with scalability issues and this process may generate something unrealistic. Therefore, we decided to merge available datasets, at least for the *refute* category. It is wise to have all these datasets into one, and further, we have generated associated images using stable diffusion. While selecting datasets, we only chose datasets with evidence claim documents as we are interested in 5W QA-based explanation. Other datasets only with fake claims were discarded for this reason. Furthermore, we use OPT (Zhang et al., 2022) to generate adversarial fake news documents of text based on the refute claims as prompts.

We have adopted an automatic method to compile a list of claims for *support* and *neutral* categories. It is often seen that the same event gets reported by two different media houses separately on the same day - therefore, one can be treated as support for the other. With this strategy in mind, we have collected and annotated large-scale data automatically (c.f. appendix A for details). Fig. 6 visualizes how much data from each dataset are compiled.
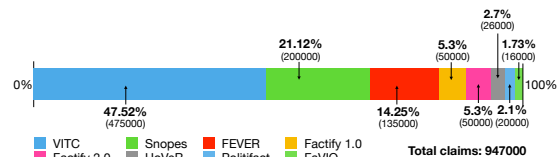

Figure 6: Distribution of our dataset delineating its constituent components.

Table 1 offers a statistical description of the five entailment classes, their definitions in terms of textual/visual support, an approximate count of the claims, paraphrased claims, images, 5W QA pairs, evidence documents, and adversarial stories for the Refute class. Furthermore, to

solidify the idea behind the above categories, fig. 5 offers a walk-through of an example from the FACTIFY 3M dataset.

There are only a handful of other previous works, (Yao et al., 2022), (Abdelnabi et al., 2022), (Roy and Ekbal, 2021), (Nielsen and McConville, 2022), (Jin et al., 2017), (Luo et al., 2021), that have discussed multimodal fact verification. None of them generated large resources like ours and did not discuss QA-based explanation, heatmap-based image explainability, and adversarial assertion.

3   Paraphrasing textual claims

A claim may have multiple diverse manifestations depending on the style and manner in which it was reported. Specifically, the textual component (i.e., prose) may have variations as highlighted in fig. 5a. We seek to echo such heterogeneity to ensure the real-world applicability of our benchmark (c.f. examples in fig. 3). Manual generation of possible paraphrases is undoubtedly ideal but is time-consuming and labor-intensive. On the other hand, automatic paraphrasing has received significant attention in recent times (Sancheti et al., 2022; Xue et al., 2022; Bandel et al., 2022; Garg et al., 2021; Goyal and Durrett, 2020). Our criteria for selecting the most appropriate paraphrasing model was the linguistic correctness of the paraphrased output and the number of paraphrase variations. To achieve this, we propose the following process - let's say we have a claim c, we generate a set of paraphrases of c. Textual paraphrase detection is a well-studied paradigm, there are much state-of-the-art (SoTA) systems (Wang et al., 2021, 2019; Tay et al., 2021). We pick the best model available mostly trained based on the resources available from SNLI (Bowman et al., 2015). Next, we use the entailment model (Wang et al., 2019) to choose the right paraphrase candidate from the generated set, by doing a pairwise entailment check and choosing only examples which exhibit entailment with c. We empirically validated the performance of GPT3 (Brown et al., 2020), Pegasus (Zhang et al., 2020), and ChatGPT (Schulman et al., 2022) models for our use-case and found that GPT3-text-davinci-003 outperformed the rest (c.f. appendix C for details - we have evaluated models on three factors: (i) coverage, (ii) correctness and (iii) diversity).

4   Visual paraphrase - Stable Diffusion based image synthesis

While textual diversity in claims seen in the wild is commonplace, typically the visual components – particularly, images – also show diversity. The concept of AI-based text-to-image generators has been around for the past several years, but their outputs were rudimentary up until recently. In the past year, text prompt-based image generation has emerged in the form of DALL-E (Ramesh et al., 2021), ImageGen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2021). While these new-age systems are significantly more powerful, they are a double-edged sword. They have shown tremendous potential in practical applications but also come with their fair share of unintended use cases. One major caution is the inadvertent misuse of such powerful systems. To further this point, we have utilized Stable Diffusion 2.0 (Rombach et al., 2021) to generate a large amount of fake news data.

Stable Diffusion (Rombach et al., 2021) is a powerful, open-source text-to-image generation model. The model is not only extremely capable of generating high-quality, accurate images to a given prompt, but this process is also far

less computationally expensive than other text-conditional image synthesis approaches such as (Ding et al., 2021; Nichol et al., 2021; Zhou et al., 2021; Gafni et al., 2022). Stable diffusion works on stabilizing the latent diffusion process which has an aspect of randomization, as a result, it generates a different result each time. Moreover, quality control is a challenge. We have generated 5 images for a given claim and then further ranked them, discussed in the next section.

4.1 Re-ranking of generated images

In order to quantitatively assess and rank the images generated by the stable diffusion model, we leverage the CLIP model (Radford et al., 2021a) to obtain the best image conditioned on the prompt. We use CLIP-Score based re-ranking to select the best image corresponding to the prompt. The CLIP-Score denotes the proximity between the final image encodings and the input prompt encoding.

4.2 Pixel-level image heatmap

(Tang et al., 2022) perform a text–image attribution analysis on Stable Diffusion. To produce pixel-level attribution maps, authors propose Diffusion Attentive Attribution Maps (DAAM), a novel interpretability method based on upscaling and aggregating cross-attention activations in the latent denoising subnetwork. We adapt the official code available on Github (Castorini, 2022) to obtain the attribution maps in the generated images for each word in the cleaned prompt (pre-processed prompt after removal of stopwords, links, etc.). See fig. 5c for examples.

4.3 Assessing quality of synthetically generated images

While SD has received great acclaim owing to its stellar performance for a variety of use cases,

to our knowledge, to our knowledge, we are the first to adopt it for fake news generation. As such, to assess the quality of generated images in the context of the fake news generation task, we utilize two evaluation metrics.

We use Fréchet inception distance (FID) (Heusel et al., 2017) which captures both fidelity and diversity and has been the de-facto standard metric for SoTA generative models (Karras et al., 2018, 2019; Ho et al., 2020; Brock et al., 2018). The process we adopted to compute the FID score to quantitatively assess the quality of SD-generated images is detailed in E.1. For a chosen set of 500 claims (100 per category), we obtained an FID score. We obtained an FID score of 8.67 (lower is better) between the set of real images and SD-generated images for the same textual claims.

As our secondary metric, we utilized Mean Opinion Score (MOS) at the claim category level which is a numerical measure of the human-judged perceived quality of artificially generated media (c.f. appendix E.2 for process details). Results of the conducted MOS tests are summarized in fig. 7.
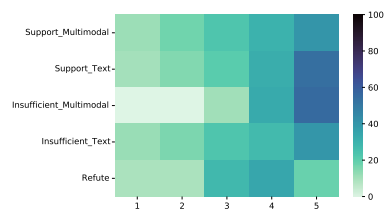

Figure 7: Heatmap of MOS scores with 500 assessed samples for each category.

5 Automatic 5W QA pair generation

A false claim is very likely to have some truth in it, some correct information. In fact, most fake news articles are challenging to detect precisely because they are mostly based on correct infor-

mation, deviating from the facts only in a few aspects. That is, the misinformation in the claim comes from a very specific inaccurate statement. So, given our textual claim and image claim, we generate 5W question-answer pairs by doing semantic role labeling on the given claim. The task is now based on the generated QA pairs, a fact-checking system can extract evidence sentences from existing authentic resources to verify or refute the claim based on each question- *Who, What, When, Where, and Why*.

Our method of using 5W SRL to generate QA pairs and then verify each aspect separately allows us to detect '*exactly where the lie lies*'. This, in turn, provides an explanation of why a particular claim is refutable since we can identify exactly which part of the claim is false.

5.1 5W semantic role labelling

Identification of the functional semantic roles played by various words or phrases in a given sentence is known as semantic role labeling (SRL). SRL is a well-explored area within the NLP community. There are quite a few off-the-shelf tools available: (i) Stanford SRL (Manning et al., 2014), (ii) AllenNLP (AllenNLP, 2020), etc. A typical SRL system first identifies verbs in a given sentence and then marks all the related words/phrases haven relational projection with the verb and assigns appropriate roles. Thematic roles are generally marked by standard roles defined by the Proposition Bank (generally referred to as PropBank) (Palmer et al., 2005), such as: *Arg0, Arg1, Arg2*, and so on. We propose a mapping mechanism to map these Prop-Bank arguments to 5W semantic roles. The conversion table 4 and necessary discussion can be found in appendix F.

5.2 Automatic 5W QA pair generation

We present a system for generating 5W aspect-based questions generation using a language model (LM) that is fed claims as input and uses the SRL outputs as replies to produce 5W questions with respect to the 5W outputs. We experimented with a variety of LMs: BART (Lewis et al., 2019) and ProphetNet (Qi et al., 2020), eventually settling on ProphetNet (see fig. 8) based on empirically figuring out the best fit for our use-case (c.f. appendix G for details).
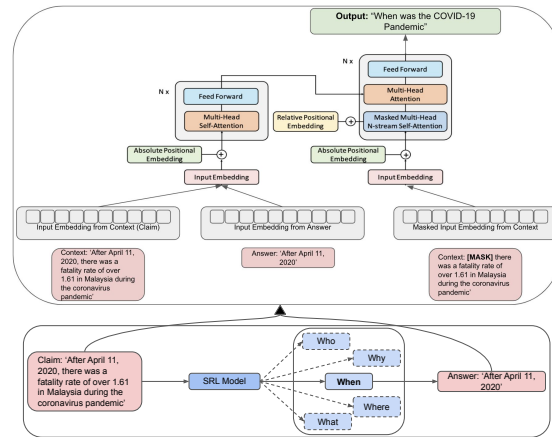
Figure 8: Illustration of 5W QA Generation Pipeline using ProphetNet.

We then create answers using the evidence from the questions generated using ProphetNet by running them through T5 (Yamada et al., 2020) – a SoTA QA model, using the 5W-based generated questions. See section 7.2 for details.

6 Injecting adversarial assertion

The rise of generative AI techniques with capabilities that mimic a human's creative thought process has led to the availability of extraordinary skills at the masses' fingertips. This has led to the proliferation of generated content, which is virtually indistinguishable as real or fake to

the human eye, even for experts in some cases. This poses unparalleled challenges to machines in assessing the veracity of such content.

As one of the novelties of our work, we address this by introducing synthetically generated adversarial fake news documents for all the *refute* claims using OPT (Zhang et al., 2022), a large language model. In doing so, we attempt to confuse the fact verification system by injecting fake examples acting as an adversarial attack. To draw a parallel in a real-world scenario, this could mean the proliferation of such fake news articles online via social media, blog posts, etc. which would eventually lead to a fact-verification system being unable to make a concrete decision on the trustworthiness of the news. Such a scenario would lend itself as a natural manifestation of an adversarial attack by virtue (rather, the "vice") of the fake news articles confusing the fact verification system. We analyze the impact on the performance of our fact verification system in table 2. Our goal in offering these adversarial articles as one of our contributions is to provide future researchers a benchmark using which they can measure (and hence, improve) the performance of their fact verification system.

6.1 Accuracy of text generation

We assess the quality of text generation using the following evaluation metrics:

Fluency: We measure the text fluency using perplexity scores. The perplexity is a measure of the likelihood of the generated sentence on a language model. We use a pre-trained GPT-2 model to evaluate text perplexity. A lower value is preferred. We have used the GPTZero detector to evaluate our perplexity score (Tian, 2023). Checking for paraphrased text generated over a 50 claims (25 original and 25 adversarial), we report an average perplexity score of 129.06 for original claims and 175.80 for adversarial claims (more details in appendix J). We intend to perform more evaluations on standard fluency measures such as ROGUE (Lin, 2004) and BLEURT (Sellam et al., 2020).

7 Experiments - baselines & performance

In this section, we present baselines for: *(i) multimodal entailment, (ii) 5W QA-based validation, and (iii) results of our models after adversarial injections of generated fake news stories.*

7.1 Multimodal entailment: *support or refute?*

In this paper, we model the task of detecting multimodal fake news as multimodal entailment. We assume that each data point contains a reliable source of information, called *document*, and its associated image and another source whose validity must be assessed, called the *claim* which also contains a respective image. The goal is to identify if the claim entails the document. Since we are interested in a multimodal scenario with both image and text, entailment has two verticals, namely textual entailment, and visual entailment, and their respective combinations. This data format is a stepping stone for the fact-checking problem where we have one reliable source of news and want to identify the fake/real claims given a large set of multimodal claims. Therefore the task essentially is: given a textual claim, claim image, text document, and document image, the system has to classify the data sample into one of the five categories: Support_Text, Support_Multimodal, Insufficient_Text, Insufficient_Multimodal, and Refute. Using the Google Cloud Vision API (Google, 2022), we also perform OCR to obtain the text

	Support_Text		Support_Multimodal		Insufficient_Text		Insufficient_Multimodal		Refute		**Average**	
	Text-only	Multimodal	Text-only	Multimodal	Text-only	Multimodal	Text-only	Multimodal	Text-only	Multimodal	Text-only	Multimodal
Pre-adversarial attack (F1)	0.33	0.61	0.15	0.60	0.22	0.58	0.22	0.57	0.31	0.65	0.25	0.60
Post-adversarial attack (F1)	0.15 (55%↓)	0.46 (25%↓)	0.06 (60%↓)	0.43 (28%↓)	0.21 (4%↓)	0.56 (3%↓)	0.11 (64%↓)	0.55 (43%↓)	0.11 (64%↓)	0.37 (43%↓)	0.13 (48%↓)	0.47 (21%↓)

Table 2: Results of the text-only and multimodal baselines pre- and post-adversarial attack.

embedded in images and utilize that as an additional input.

Text-only model: Fig. 9 shows our text-only model, which adopts a siamese architecture focussing only on the textual aspect of the data and ignores the visual information. To this end, we generate sentence embeddings of the claim and document attributes using a pretrained MPNet Sentence BERT model (Reimers and Gurevych, 2019a) (specifically the `all-mpnet-base-v2` variant). Next, we measure the cosine similarity using the generated embeddings. The score, thus generated, is used as the only feature for the dataset, and classification is evaluated based on their F1 scores.
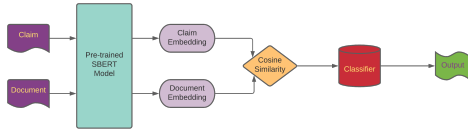

Figure 9: Text-only baseline model which takes only claim text and document text as input.

Table 2 shows the F1 score for the unimodal (i.e., text-only) and multimodal approaches (c.f. pre-adversarial attack row in table 2) trained using their respective feature sets. The multimodal model shows a distinct improvement in performance compared to the text-only model, indicating the value-addition of the visual modality.

7.2   5W QA-based validation

We generate 5W Question-Answer pairs for the claims, thus providing explainability along with evidence in the form of answers generated. To this end, we use the SoTA T5 (Raffel et al.,
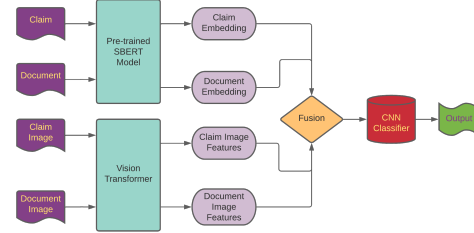
Figure 10: Multimodal baseline model which takes as input: (i) claim text, (ii) claim image, (iii) document text, and (iv) document image.

2020) model for question answering (QA) in this work. It is trained using a modified version of BERT's masked language model, which involves predicting masked words and entities in an entity-annotated corpus from Wikipedia.
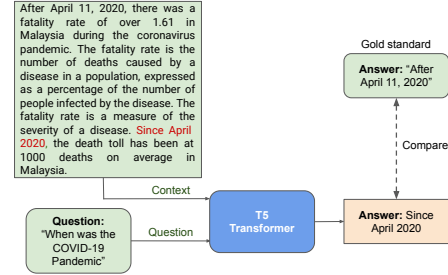
Figure 11: T5-based question answering framework.

7.3   Adversarial attack

With the emergence of prolific ChatGPT, the risk of AI-generated content has reached an alarming apocalypse. ChatGPT has been declared banned by the school system in NYC (Rosenblatt, 2023), Google ads (Grant and Metz, 2022), and Stack Overflow (Makyen and Olson, 1969), while sci-

entific conferences like ACL (Chairs, 2023) and ICML (Foundation, 2023) have released new policies deterring the usage of ChatGPT for scientific writing. Indeed, the detection of AI-generated text has suddenly emerged as a concern that needs imminent attention. While watermarking as a potential solution to the problem is being studied by OpenAI (Wiggers, 2022b), a handful of systems that detect AI-generated text such as GPT-2 output detector (Wiggers, 2022a), GLTR (Strobelt et al., 2022), GPTZero (Tian, 2022), has recently been seen in the wild. Furthermore, these tools typically only produce meaningful output after a minimum (usually, 50+) number of tokens. We tested GPTZero on a randomly selected set of 100 adversarial samples, equally divided into human-generated text and AI-generated text. Our results indicate that these systems are still in their infancy (with a meager 22% accuracy). It is inevitable that AI-generated text detection techniques such as watermarking, perplexity, etc. will emerge as important paradigms in generative AI in the near future, and FACTIFY 3M will serve the community as a benchmark in order to test such techniques for fact verification.

Table 2 shows the F1 score post adversarial attack for the unimodal (i.e., text-only) and multimodal approaches - proving that injecting adversarial news can confuse fact-checking very easily.

8 Conclusion and future avenues

We are introducing FACTIFY 3M, the largest dataset and benchmark for multimodal fact verification. We hope that our dataset facilitates research on multimodal fact verification on several aspects - *(i) visual QA-based explanation of facts, (ii) how to handle adversarial attacks for fact verifications, (iii) whether generated images can be detected, and (iv) 5W QA-based help journalists to fact verify easily for complex facts*. FACTIFY 3M will be made public and open for research purposes.

9 Discussion and limitations

In this section, we mention a few aspects that could be improved and also detail how we plan to improve upon those specific aspects.

9.1 Paraphrasing claims

Manual generation of possible paraphrases is undoubtedly ideal but is time-consuming and labor-intensive. Automatic paraphrasing is a good way to scale quickly, but there could be more complex variations of meaning paraphrases hard to generate automatically. For example - "*It's all about business - a patent infringement case against Pfizer by a rival corporate reveals they knew about COVID in one way!*" and "*Oh my god COVID is not enough now we have to deal with HIV blood in the name of charity!*".

An ideal for this shortcoming would be to manually generate a few thousand paraphrase samples and then fine-tune language models. On the other hand, a new paradigm in-context Learning is gaining momentum (Xun et al., 2017). In-context learning has been magical in adapting a language model to new tasks through just a few demonstration examples without doing gradient descent. There are quite a few recent studies that demonstrate new abilities of language models that learn from a handful of examples in the context (in-context learning - ICL for short). Many studies have shown that LLMs can perform a series of complex tasks with ICL, such as solving mathematical reasoning problems (Wei et al., 2022). These strong abilities have been widely verified as emerging abilities for large language models (Wei et al., 2022). From prompt engineering to chain of thoughts, we are excited to do more experiments with the new paradigm of in-context learning for automatically paraphrasing claims.

9.2 Image synthesis using Stable Diffusion

Although, in general, the quality of the image synthesized by Stable Diffusion is great, it does not perform well in two cases - i) *very long text* (more than 30 words or so, multiple sentence claim, etc.), ii) *text with metaphoric twists* - for example, "*It's all about business - a patent infringement case against Pfizer by a rival corporate reveals they knew about COVID in one way!*" and "*Oh my god COVID is not enough now we have to deal with HIV blood in the name of charity!*". It is worthy seeing how in-domain adaptation could be made for SD image synthesis, inspired from (Ruiz et al., 2023).

9.3 5W SRL

Semantic role labeling is a well-studied sub-discipline, and the mapping mechanism we proposed works well in most cases except in elliptic situations like anaphora and cataphora. In the future, we would like to explore how an anaphora and coreference resolution (Joshi et al., 2019) can aid an improvement.

9.4 5W QA pair generation

5W semantic role-based question generation is one of the major contributions of this paper. While automatic generation aided in scaling up the QA pair generation, it also comes with limitations of generating more complex questions covering multiple Ws and *how* kinds of questions. For example - "*How Moderna is going to get benefited if this Pfizer COVID news turns out to be a rumor?*". For the betterment of FAC-TIFY benchmark, we would like to generate few thousand manually generated abstract QA pairs. Then will proceed towards in-context Learning (Xun et al., 2017).

Abstractive question-answering has received momentum (Zhao et al., 2022), (Pal et al., 2022)

recently. We want to explore how we can generate more abstract QA pairs for the multimodal fact-verification task.

9.5 QA system for the 5W question

Generated performance measures attest the proposed QA model needs a lot more improvement. This is due to the complexity of the problem and we believe that will attract future researchers to try this benchmark and conduct research on multimodal fact verification.

It has been realized by the community that relevant document retrieval is the major bottleneck for fact verification. Recent work introduced a fresh perspective to the problem - named Hypothetical Document Embeddings (HyDE) (Gao et al., 2022) and applied a clever trick even if the wrong answer is more semantically similar to the right answer than the question. This could be an interesting direction to explore and examine how that could aid in retrieving relevant documents and answers.

9.6 Adversarial attack

Precisely, we are the first to formally introduce an adversarial attack for fact verification and introducing large-scale data. While it is a hot topic of discussion how systems can identify AI-generated text, there is no breakthrough so far. We would like to explore more in this direction more, specifically for multimodal fact verification.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14940–14949.

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward " the holy grail " : The continued quest to automate fact-checking.

AllenNLP. 2020. Allennlp semantic role labeling. https://demo.allennlp.org/semantic-role-labeling. [Online; accessed 2023-01-02].

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. *arXiv preprint arXiv:2203.10940*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ali Borji. 2022. Generated faces in the wild: Quantitative comparison of stable diffusion, mid journey and dall-e 2. *arXiv preprint arXiv:2210.00586*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Castorini Castorini. 2022. Castorini/daam: Diffusion attentive attribution maps for interpreting stable diffusion.

Program Chairs. 2023. Acl 2023 policy on ai writing assistance.

Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. 2022. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Michael Dobbs. 2012. The rise of political fact-checking, how reagan inspired a journalistic movement. *New America Foundation*, pages 4–5.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Neural Information Processing Systems Foundation. 2023. Clarification on large language model policy llm.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022.

Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Sonal Garg, Sumanth Prabhu, Hemant Misra, and G Srinivasaraghavan. 2021. Unsupervised contextual paraphrase generation using lexical control and reinforcement learning. *arXiv preprint arXiv:2103.12777*.

Google. 2022. Cloud Vision Api. https://cloud.google.com/vision. [Online; accessed 2023-01-02].

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. *arXiv preprint arXiv:2005.02013*.

Nico Grant and Cade Metz. 2022. A new chat bot is a 'code red' for google's search business.

Ashim Gupta and Vivek Srikumar. 2021. X-FACT: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models.

Asian News International. Asian news international.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. BERT for coreference resolution: Baselines and analysis. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Tero Karras, Samuli Laine, and Timo Aila. 2018. A style-based generator architecture for generative adversarial networks.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. Analyzing and improving the image quality of stylegan.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 2019.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*.

Mod Makyen and Peter Olson. 1969. Temporary policy: Chatgpt is banned.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60. The Association for Computer Linguistics.

Shreyash Mishra. 2022. De-factify 2022. [Online; accessed 2023-01-02].

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2023. Factify 2: A multimodal fake news and satire news dataset. In *proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection*. CEUR.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal

fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*.

Frank Luther Mott. 1942. Trends in newspaper content. *The Annals of the American Academy of Political and Social Science*, 219:60–65.

Cable News Network. Cable news network.

ABC News. Abc news.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2020. Unsupervised paraphrasing with pretrained language models. *arXiv preprint arXiv:2010.12885*.

Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. *OpenReview*.

OpenAI. 2022. Chatgpt. https://openai.com/blog/chatgpt/. [Online; accessed 2023-01-02].

Vaishali Pal, Evangelos Kanoulas, and Maarten Rijke. 2022. Parameter-efficient abstractive question answering over tables or text. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 41–53, Dublin,

Ireland. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2021. On aliased resizing and surprising subtleties in gan evaluation.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pretraining. *arXiv preprint arXiv:2001.04063*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021a. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Nils Reimers. 2022. Sentencetransformers documentation. https://sbert.net/. [Online; accessed 2023-01-02].

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Kalhan Rosenblatt. 2023. Chatgpt banned from new york city public schools' devices and networks.

Arjun Roy and Asif Ekbal. 2021. Mulcob-mulfav: Multimodal content based multilingual fact verification. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Abhilasha Sancheti, Balaji Vasan Srinivasan, and Rachel Rudinger. 2022. Entailment relation aware paraphrase generation. *arXiv preprint arXiv:2203.10483*.

Christoph Schuhmann. 2022. Christophschuhmann/improved-aesthetic-predictor: Clip+mlp aesthetic score predictor.

J Schulman, B Zoph, C Kim, J Hilton, J Menick, J Weng, JFC Uribe, L Fedus, L Metz, M Pokorny, et al. 2022. Chatgpt: Optimizing language models for dialogue.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Craig Silverman. 2020. Verification handbook: Homepage.

Media Smarts. 2017. How to recognize false content online - the new 5 ws.

Kathryn T Stofer, James R Schaffer, and Brian A Rosenthal. 2009. *Sports journalism: An introduction to reporting and writing*. Rowman & Littlefield Publishers.

Northwestern Now Story. 2022. Chatgpt writes convincing fake scientific abstracts that fool reviewers in study. https://news.northwestern.edu/stories/2023/01/chatgpt-writes-convincing-fake-scientific-abstracts-that-fool-reviewers-in-study/?fj=1. [Online; accessed 2023-01-02].

Hendrik Strobelt, Sebastian Gehrmann, and Alexander Rush. 2022. Giant language model test room. http://gltr.io/dist/index.html. [Online; accessed 2023-01-02].

Jing Su, Xiguang Li, and Lianfeng Wang. 2019. The study of a journalism which is almost 99% fake. *Lingue Culture Mediazioni-Languages Cultures Mediation (LCM Journal)*, 5(2):115–137.

S Suryavardhan. 2023. De-factify 2023. [Online; accessed 2023-01-02].

Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Edward Tian. 2022. Gptzero. https://gptzero.me/. [Online; accessed 2023-01-02].

Edward Tian. 2023. Gptzero detec.

Hindustan Times. Hindustan times.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Kyle Wiggers. 2022a. Gpt-2 output detector demo. https://openai-openai-detector.hf.space/. [Online; accessed 2023-01-02].

Kyle Wiggers. 2022b. Openai's attempts to watermark ai text hit limits. https://techcrunch.com/2022/12/10/openais-attempts-to-watermark-ai-text-hit-limits. [Online; accessed 2023-01-02].

Wikipedia Wiki_Article. 2023. Five ws.

Tang Xue, Yuran Zhao, Chaoqi Yang, Gongshen Liu, and Xiaoyong Li. 2022. Sect: A successively conditional transformer for controllable paraphrase generation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Guangxu Xun, Xiaowei Jia, Vishrawas Gopalakrishnan, and Aidong Zhang. 2017. A survey on context learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):38–56.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022. Explainable fact-checking through question answering. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8952–8956.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *arXiv preprint arXiv:2205.12487*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained

transformer language models. *arXiv preprint arXiv:2205.01068.*

Wenting Zhao, Konstantine Arkoudas, Weiqi Sun, and Claire Cardie. 2022. Compositional task-oriented parsing as abstractive question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4418–4427, Seattle, United States. Association for Computational Linguistics.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2021. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792.*

Frequently Asked Questions - FAQs

1. Does Stable Diffusion offer the adeptness to generalize and scale to different real-world scenarios? In other words, Stable Diffusion is great at generating one-off plausible examples but is generalizability to life's combinatorial scenarios a concern?

   Ans. - Fake news is generally written connecting popular topics and personalities, therefore stable diffusion does a decent job. However, there are some limitations, which we discussed in detail in the limitation section 9.2. Moreover, we have generated stable diffusion images for all claims as a visual paraphraser as mentioned in section 4. Furthermore, we have presented a holistic evaluation through objective (FID) and subjective (MOS) metrics. Please refer to the section 4.3, table 7.

2. What are the novel assertions in this paper if the multimodal data is generated automatically using SoTA generative models?

   Ans. - The novelty of this work is three-fold:
   (a) Justification of the classification using 5WQA verification.
   (b) Injecting an adversarial attack in the form of fake news to make the dataset more robust.
   (c) Adding synthetically generated images using Stable Diffusion to enhance multimodal data.

3. 5W SRL is understandable, but how is the quality of the 5W QA pair generation using a language model?

   Ans. - We have evaluated our QA generation against the SoTA model for QA Tasks - T5. Please refer to the appendix section I, table 7 for a detailed description of the process and evaluation. Moreover, please see the discussion in the limitation section 9.4.

4. What is the overarching idea we're trying to highlight by introducing an adversarial attack?

   Ans. - The broader point that the introduction of an adversarial attack indicates is that a fact verification model needs to be more robust in combating synthetically generated fake news, which is easily publishable by wrongdoers on the internet. This is of extreme relevance today as AI-assisted writing has become very popular and miscreants spread fake news taking advantage of LLMs.

5. How does adversarial attack impact the performance?

   Ans. - As reported in table 2, we see that the performance of the model drops across all categories post adversarial attack using fake claims. This is seen in both instances: text-only and multimodal model.

6. Despite the controversies surrounding AI-assisted writing, why have we still chosen to use LLMs as our paraphrasers?

Ans. - The controversy lies mostly in a conversational setting or creative writing. When it comes to paraphrasing news claims, we have empirically found that GPT-3 (specifically the `text-davinci-003` variant) (Brown et al., 2020) performs better in comparison to other models such as Pegasus (Zhang et al., 2020), and ChatGPT (Schulman et al., 2022).

7. What was the chosen metric of evaluation for text generation using LLMs?

   Ans. - For now, we have evaluated adversarial claims using GPTZero text detector. Evaluation on standard metrics such as control and fluency will be made public along with our dataset.

Appendix

This section provides supplementary material in the form of additional examples, implementation details, etc. **To bolster the reader's understanding of the concepts presented in this work.**

A   Data sources and compilation

In this section, we provide additional details on data collection and compilation. As mentioned in section 2 we are only interested in refute category from the available datasets, for support and neutral categories we have collected a significant amount of data from the web. This data collection process is semi-automatic.

For FEVER and VITC, only the claims belonging in the train split were used for making the dataset. FaVIQ (Park et al., 2021) has two sets: Set A and Set R. Set A consists of ambiguous questions and their disambiguations. Set R is made of unambiguous question-answer pairs. We have used claims from set A in our dataset to make the entailment task more challenging. In the case of HoVer (Jiang et al., 2020), we have used all 26171 claims for our dataset.

In the Factify dataset (Mishra et al., 2022), the authors have collected date-wise tweets from Twitter handles of Indian and US news sources: (i) Hindustan Times (Times), ANI (International) for India, and (ii) ABC (News), CNN (Network) for the US, based on accessibility, popularity and posts per day. We drew our motivation from (Mishra et al., 2022). Moreover, these Twitter handles are eminent for their objective and disinterested approach. From each tweet, the tweet text and the tweet image(s) have been extracted. Listing A delineates each attribute in the dataset and its respective description while listing B elaborates on the process we followed for collecting data for Support and Neutral

categories.

Listing A: Attributes

- Claim: Tweet A text
- Claim_image: Tweet A image
- Claim_ocr: Tweet A image OCR
- Document: Tweet B article text
- Document_image: Tweet B image
- Document_ocr: Tweet B image OCR
- Category

Listing B: Procedure for data collection for Support and Neutral categories

- For each tweet of account A, authors got similar tweets from account B. Similarity is measured on the basis of text. Text similarity is measured using Sentence BERT (Reimers and Gurevych, 2019b) first, and then the extent of common words is measured as the second metric.

- Next, the image similarity for the corresponding images of the tweet pair was calculated. Image similarity is measured using histogram similarity and cosine similarity on a pre-trained ResNet50 model.

- According to the scores for each of these measures, the tweet pair is classified into 4 categories: Support_Multimodal, Support_Text, Insufficient_Multimodal, and Insufficient_Text. The various thresholds used for classification are listed in Figure 12.

- From this tweet pair, authors have selected a tweet (say tweet B) and obtained the url for the corresponding article published on the source's website from the tweet text. Then the tweet text was replaced with article contents after scraping it (document in dataset). This is done so as to mimic real world fact checking process, i.e., manually comparing claims with documents or articles.

- The image OCRs were obtained using Google Cloud Vision API (Google, 2022).

B Text and image similarity measures

Table 1 explains the five classes in the dataset. For the appropriate classification of the dataset, two similarity measures were computed.

B.1 Sentence comparison

We adopt two methods to check similarity given a set of two sentences:

- **Sentence BERT:** Sentence BERT (Reimers and Gurevych, 2019b) is a modification of the BERT model that uses a contrastive loss with a siamese network architecture to derive sentence embeddings. These sentence embeddings can be compared with each other to get their corresponding similarity score. Authors use cosine similarity as the textual similarity metric.

- We utilize Sentence BERT (SBERT) (Reimers and Gurevych, 2019b) instead of alternatives such as BERT or RoBERTa, owing to its rich sentence embeddings yielding superior performance while being much more time-efficient (in terms of sentences/sec) (Reimers, 2022). We manually decide on a threshold value *T1* for cosine similarity and classify the text pair accordingly. If the cosine similarity score is greater than *T1*, then it is classified into the Support category. On the other hand, if the cosine similarity score is lower than *T1*, the news may or may not be the same (the evidence at hand is insufficient to judge whether the news is the same or not). Hence it is sent for another check before classifying it into the Insufficient category. **NLTK:** If the cosine similarity of the sentence pair is below *T1*, we use the NLTK library (Bird et al., 2009) to check for common words between the two sentences. If the score of the common word is above a different manually decided threshold *T2*, only then the news pair is classified into the Insufficient category. Not sure what this sentence is trying to say - let's rephrase. Common words are being checked to ensure that the classification task is challenging. To check for common words, both texts in the pair are preprocessed, which included stemming and removing stopwords. The processed texts are then checked for common and similar words, and their corresponding scores are determined. If the common words score is greater than T2, the pair is classified as Insufficient else the pair is dropped.

B.2 Image comparison

We adopt two metrics for assessing image similarity:

- **Histogram Similarity:** The images are converted to normalized histogram format and similarity is measured using the correlation metric cite.

- **Cosine Similarity:** The images are converted to feature vectors using pre-trained ViT (Dosovitskiy et al., 2020) model, and these feature vectors are used to calculate the cosine similarity score. Manually decided thresholds, as described in Figure 12, are used to judge whether the text and image pair is similar or not.

The text pairs are first classified into either Support or Insufficient categories, and then further sub-classified into Support_Text/Support_Multimodal, or Insufficient_Text/Insufficient_Multimodal categories based on the similarity of the image pairs. If the corresponding images for the texts are similar, then they could be used to judge whether news is the same or not. The category where both the images and the texts are similar is

called `Support_Multimodal`. The category where the images are similar but the texts were not is called `Insufficient_Multimodal`. If the corresponding images for the texts were not similar, then they could not be used to judge whether news is the same or not. The category where both the images and the texts are not similar is called `Insufficient_Text`. The category where the texts are similar but the images are not is called `Support_Text`.
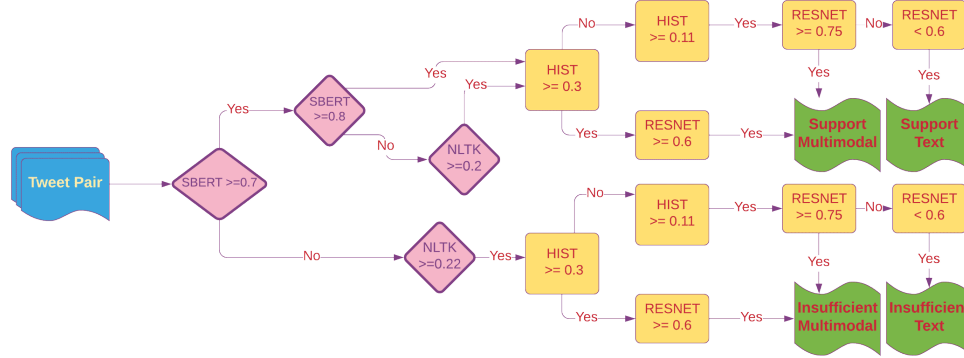

Figure 12: Text and image pair similarity based on classification thresholds on pre-trained models.

For each article published on these websites, we collect the claim (sentence that states the fake news), document (text that proves claim is false), claim images (fake news image, could be screenshot of the fake post), document image (image that is proof of the fake nature of the claim).

C Paraphrasing textual claims

A textual given claim may appear in various different textual forms in real life, owing to variety in the writing styles of different news publishing houses. Incorporating such variations is essential to developing a strong benchmark to ensure a holistic evaluation. This forms our motivation behind paraphrasing textual claims. Manual generation of possible paraphrases is undoubtedly ideal, but that process is time-consuming and labour-intensive. On the other hand, automatic paraphrasing has received significant attention in recent times (Niu et al., 2020) (Zhang et al., 2020) (Nighojkar and Licato, 2021). As mentioned in section 3, for a given claim, we generate multiple paraphrases using various models and perform entailment using (Wang et al., 2019) – a SoTA model trained on the on SNLI task (Bowman et al., 2015) – to detect how many of them are entailed in the actual claim.

In the process of choosing the appropriate model based on a list of available models, the primary question we asked is how to make sure the generated paraphrases are rich in diversity while still being linguistically correct. A top level, we delineate the process followed to achieve this as follows (more details later in this section). Let's say we have a claim c. We generate n paraphrases using a paraphrasing model. This yields a set of paraphrases, denoted by $p_1^c, \ldots, p_n^c$. Next, we make pair-wise comparisons of these paraphrases with c, resulting in $c - p_1^c, \ldots,$ and $c - p_n^c$. At this step, we identify the examples which are entailed, and only those are chosen.

However, there are many other secondary factors, for e.g., a model may only be able to generate a

limited number of paraphrase variations compared to others but others can be more correct and/or consistent. As such, we considered three major dimensions in our evaluation: *(i) coverage, (ii) correctness, and (iii) diversity*. To offer transparency around our experiment process, we detail the aforementioned evaluation dimensions as follows.

1. **Coverage - the number of considerable paraphrase generations that a model generates:** We intend to generate up to 5 paraphrases per given claim. Given all the generated claims, we perform a minimum edit distance (MED) calculation at the word level instead of a character level. If MED is greater than 2 for any given paraphrase candidate (for e.g., $c - p_1^c$ in the above example) with the claim then we further consider that paraphrase, otherwise discarded. We evaluated all four models based on this setup to identify the model of choice which is generating the maximum number of considerable paraphrases.

2. **Correctness - correctness in paraphrase generations:** After the first level of filtration, we performed pairwise entailment and kept only those paraphrase candidates, marked as entailed by the (), SoTA model trained on SNLI.

3. **Diversity - linguistic diversity in paraphrase generations:** We are interested in choosing a model that can produce paraphrases with significant linguistic diversity. This implies that we are interested in checking for dissimilarities between generated paraphrase claims. For e.g., $p_1^c - p_2^c, p_1^c - p_3^c, p_1^c - p_4^c, \ldots, p_1^c - p_n^c$ – this process is repeated for all the other paraphrases and the dissimilarity score is averaged across all paraphrase generations. Since there is no standard metric to measure dissimilarity, we use the inverse of the BLEU score as a proxy metric. This gives us an understanding of the linguistic diversity of a given model.

Based on our experiments centred around the above dimensions, we experimented with three models: (i) GPT3-text-davinci-003 (Brown et al., 2020), (ii) Pegasus (Zhang et al., 2020), and (ii) ChatGPT (Schulman et al., 2022) and found that GPT3-text-davinci-003 was ideal. The results of our experiments are reported in table 3 below.

Model	Coverage	Correctness	Diversity
Pegasus	32.46	94.38%	3.76
T5	30.26	83.84%	3.17
GPT3-text-davinci-003	35.51	88.16%	7.72

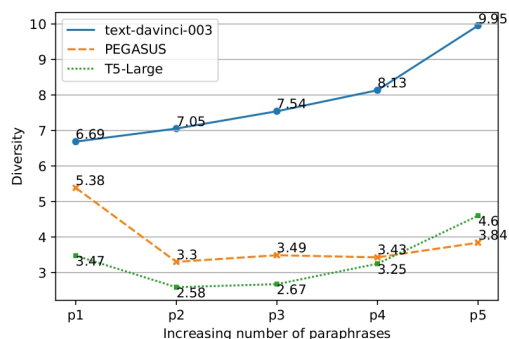Table 3: Evaluation dimensions of textual claim paraphrasers.

Figure 13: Variation of Diversity with Increase in number of paraphrases generated

D   Visual paraphrasing using Stable Diffusion

Building upon section 4, we highlight the process behind visual paraphrasing in this section. Diffusion models are machine learning models that are trained to denoise random gaussian noise step by step to get a sample of interest, such as an image. However one of the major downsides of diffusion models is that the denoising process is both time and memory consumption are very expensive. The main reason for this is that they operate in pixel space which becomes unreasonably expensive, especially when generating high-resolution images. Stable diffusion was introduced to solve this problem as it depends on Latent diffusion. Latent diffusion reduces the memory and computational cost by applying the diffusion process over a lower dimensional latent space instead of on the actual pixel space. It is trained with the objective of "removing successive applications of Gaussian noise to training images", and can be considered as a sequence of denoising autoencoders.

Quality control is a big reason to worry when paraphrasing automatically. There are two aspects we have tested for the available models - (i) variations, and (ii) the number of paraphrases generated.

D.1   Explainability of generated images

E   Assessment of Stable Diffusion generated images

While Stable Diffusion has received great acclaim owing to its stellar performance for a variety of use cases, to our knowledge, we are the first to adopt it for fake news generation. As such, to assess the quality of generated images in the context of the fake news generation task, we utilize two evaluation metrics.

E.1   FID & Relevance Score-based quantitative assessment of Stable Diffusion generated images

While Stable Diffusion has received great acclaim owing to its stellar performance for a variety of use cases, to our knowledge, we are the first to adopt it for fake news generation. As such, to assess the quality of generated images in the context of the fake news generation task, we utilize two evaluation metrics - i) FID (Heusel et al., 2017) and ii) Relevance Score (Hao et al., 2022) - details are discussed in the following paragraphs.

FID Score: In order to compute the FID scores, we first filter out the claims from our dataset that consists of person entities by leveraging the BERT-base-NER model. Following the process adopted in (Borji, 2022), we ran the Mediapipe (Lugaresi et al., 2019) face detector twice: first on the entire image to detect faces, and thereafter on the individual detections to prune false positives, to extract faces from the real and Stable Diffusion generated images corresponding to the filtered set of claims. We then compute the FID between the set of faces extracted from the real and Stable Diffusion generated images using the `clean-fid` package released by (Parmar et al., 2021).

Relevance Score: Considering Figure 14.a is the original image and Figure 14.b is the SD generated image, we compute the relevance scores as the combination of two metrics - i) CLIP score (Radford et al., 2021b) (which measures how relevant the generated image is to the user input prompt) - this is to measure the semantic similarity between text and image modality using pre-trained vision-language model CLIP, and ii) Aesthetic score (Hao et al., 2022), obtained by

(a) Real Image

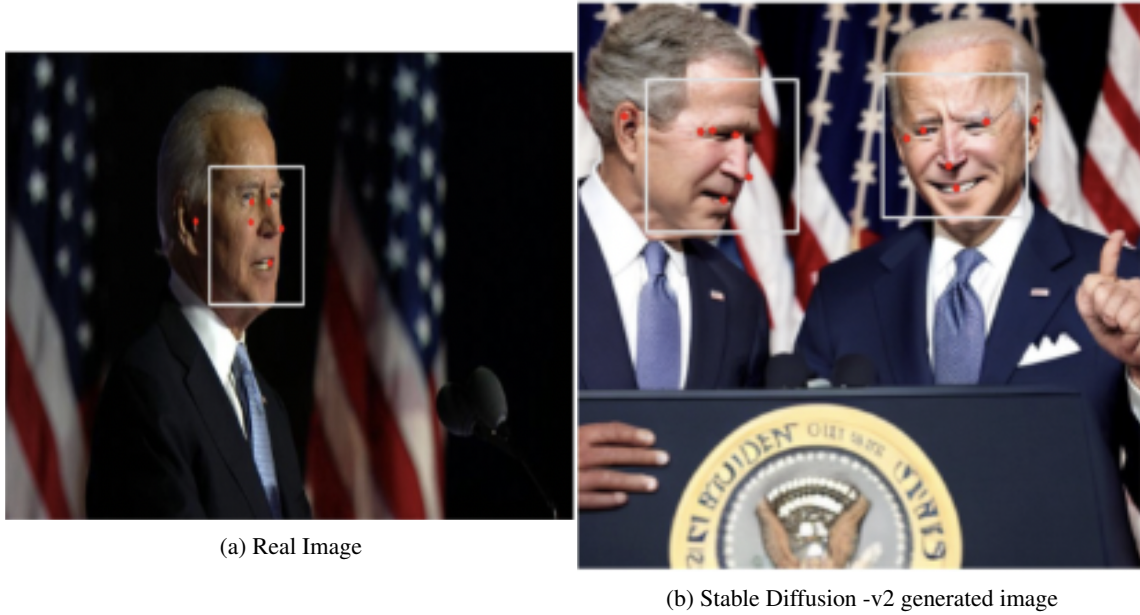(b) Stable Diffusion -v2 generated image

Figure 14: In this example, for the claim : "Former President George W. Bush congratulates President-elect Joe Biden, says election was 'fundamentally fair' and 'its outcome is clear' ", the left image where only Joe Biden is visible is the original claim image, and on the right where George Bush and Joe Biden are visible is the SD generated one. To assess the quality of the generated image we have calculated the pair-wise FID score. First, we extract the faces using Mediapipe (Lugaresi et al., 2019) Face Detector for the real and Stable Diffusion generated image for each claim. We then compute the FID using clean-fid (Parmar et al., 2021) pairwise. Then for a set of 500 randomly selected samples, we average out the pairwise FID scores. It is 8.67, demonstrating a good match overall. The difference between the left image vs. the right one is the number of faces. In such a case, we take the best (lowest) FID score as the FID score for that claim. In this way, we make sure what is the common minimum between an AI-generated image vs. an actual news image.

employing a linear estimator on top of a frozen CLIP model, that is trained by human ratings in the Aesthetic Visual Analysis [MMP12] dataset. This score represents the quality of a generated image based on human evaluation pre-scores, representing what human perceives as aesthetically pleasing.

We use the relevance score as introduced in (Hao et al., 2022) to measure whether the generated images are relevant to the original input prompt. We compute CLIP (Radford et al., 2021b) similarity scores to measure how relevant the generated images and the original input prompts are. The resulting relevance score is defined as:

$$f_{\text{rel}}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \mathbb{E}_{i_{\boldsymbol{y}} \sim \mathcal{G}(\boldsymbol{y})}\left[\min\left(20 * g_{\text{CLIP}}\left(\boldsymbol{x}, i_{\boldsymbol{y}}\right) - 5.6, 0\right)\right] \quad (1)$$

where, $i_{\boldsymbol{y}} \sim \mathcal{G}(\boldsymbol{y})$ means sampling images $i_{\boldsymbol{y}}$ from the text-to-image model $\mathcal{G}$ with $\boldsymbol{y}$ as input prompt, and $g_{\text{CLIP}}\left(\cdot, \cdot\right)$ stands for the CLIP similarity function.

Second, we employ aesthetic predictor as discussed in (Hao et al., 2022) to quantify aesthetic preferences. The aesthetic predictor (Schuhmann, 2022) builds a linear estimator on top of a frozen CLIP model, which is trained by human ratings in the Aesthetic Visual Analysis [MMP12] dataset. The aesthetic score is defined as:

$$f_{\text{aes}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{i_{\boldsymbol{x}} \sim \mathcal{G}(\boldsymbol{x}), i_{\boldsymbol{y}} \sim \mathcal{G}(\boldsymbol{y})} \left[g_{\text{aes}}(i_{\boldsymbol{y}}) - g_{\text{aes}}(i_{\boldsymbol{x}}) \right] \tag{2}$$

where, $g_{\text{aes}}(\cdot)$ denotes the aesthetic predictor, and $i_{\boldsymbol{y}}, i_{\boldsymbol{x}}$ are the images generated by the prompts $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively.

E.2 MOS-based quality assessment of Stable Diffusion generated images

This section delineates the process followed to assess the quality of synthetically generated images, given the prompt used for a generation as context to the human rater. Specifically, we asked 10 raters to assign an integral score from 1 (*bad quality*) to 5 (*excellent quality*) to the generated images in the context of the given prompt. Specifically, similar to (Chambon et al., 2022), the scoring system was verbalized as follows:

5. Life-like generated image with potentially minor error elements, but practically indistinguishable from an original.

4. Good generated image with noticeable errors not influencing the claim's veracity assessment.

3. Moderate errors in the generated image with possible minor negative imacpts to the claim's veracity assessment.

2. Errors leading to hallucinated lesions while still preserving the major theme of the claim but influencing the claim's veracity assessment.

1. Severe errors such as the generated image not following the prompt's major theme resulting in the claim's veracity assessment impossible.

The raters rated the CLIP re-ranked output for each prompt (so 500 images in total), presented in a randomized fashion. As part of a pilot study, we assessed the calibration procedure and the test-retest reliability of 10 raters on a subset of 500 generated images by adding a generated image twice to a larger test set, similar to (Ledig et al., 2017). We observed good reliability and no significant differences between the ratings of the identical images.

F 5W SRL

A typical SRL system first identifies verbs in a given sentence and then marks all the related words/phrases haven relational projection with the verb and assigns appropriate roles. Thematic roles are generally marked by standard roles defined by the Proposition Bank (generally referred to as PropBank) (Palmer et al., 2005), such as: *Arg0, Arg1, Arg2*, and so on. We propose a mapping mechanism to map these PropBank arguments to 5W semantic roles (refer to the conversion table 4).

Not necessarily all the Ws are present in all the sentences. To understand this sparseness, a detailed analysis of the presence of each of the 5W at the sentence level has been done and reported in figure 15.

PropBank Role	Who	What	When	Where	Why
ARG0	**84.48**	0.00	3.33	0.00	0.00
ARG1	10.34	**53.85**	0.00	0.00	0.00
ARG2	0.00	9.89	0.00	0.00	0.00
ARG3	0.00	0.00	0.00	22.86	0.00
ARG4	0.00	3.29	0.00	34.29	0.00
ARGM-TMP	0.00	1.09	**60.00**	0.00	0.00
ARGM-LOC	0.00	1.09	10.00	**25.71**	0.00
ARGM-CAU	0.00	0.00	0.00	0.00	**100.00**
ARGM-ADV	0.00	4.39	20.00	0.00	0.00
ARGM-MNR	0.00	3.85	0.00	8.57	0.00
ARGM-MOD	0.00	4.39	0.00	0.00	0.00
ARGM-DIR	0.00	0.01	0.00	5.71	0.00
ARGM-DIS	0.00	1.65	0.00	0.00	0.00
ARGM-NEG	0.00	1.09	0.00	0.00	0.00

Table 4: A mapping table from PropBank(Palmer et al., 2005) (*Arg0, Arg1, ...*) to 5W (*who, what, when, where, and why*).
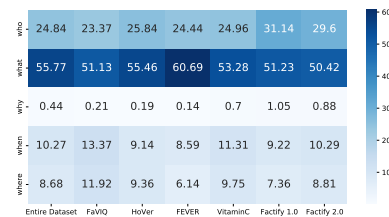

Figure 15: Sentence level co-occurrence of Ws at sentence level across the corpus.

F.1 Human evaluation of 5W SRL

In this study evaluation for the 5W Aspect, based on semantic role labelling is conducted using *mapping accuracy*. This involves accuracy on SRL output mapped with 5Ws.

For the purpose of finding how good the mapping of 5W with semantic roles and generation of semantic roles, human annotation of 3500 data points was conducted, 500 random datapoints from the entire dataset, 500 each from FEVER (Thorne et al., 2018), FavIQ (Park et al., 2021), HoVer (Jiang et al., 2020), ViTC (Schuster et al., 2021), Factify 1.0 (Mishra et al., 2022) and Factify 2.0 (Mishra et al., 2023), see table 5

G 5W QA pairs generation using language model

For the QG task, we shortlisted two pre-trained top-performing models for question generation according to the papers with code leaderboard where the model and code have been released. These models were fine-tuned on various SQuAD datasets (Rajpurkar et al., 2018) by simply appending the answer to the context. A random sampling on 352k data points was done to get of 15% of the datapoint to find the best question-generating model with respect to 5W. For example, given an answer from "who" based on semantic role labeller and context from the claim, it should generate questions containing "who" and not other Ws. By modelling the claims as context and the outputs from the SRL models as answers, the process of generating 5W questions for the task of fact verification was accomplished. The pre-trained models we utilized for QG are as follows:

- **BART:** BART (Lewis et al., 2019) is a denoising autoencoder for pretraining sequence-to-sequence models, trained by (i) corrupting text with an arbitrary noising function, and (ii) learning a model to reconstruct the original text. BART was trained to generate questions

in two ways: casual generation and context-based generation. For this task, we used the `bart-squad-qg-hl` variant focusing on context-based generation. This variant of BART scored 24.15, 25.43, and 52.64 on the BLEU4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L metrics (Lin, 2004), respectively, whereas the current state-of-the-art (SoTA) of the BART model from Textbook 2.0 scores 25.08, 26.73, and 52.55 on the same metrics.

- **ProphetNet:** ProphetNet (Qi et al., 2020) is a generative model that uses multi-lingual pre-training with masked span generation to create shared latent representations across languages. It generates all the masked spans together, given an input sequence, and uses a future n-gram loss to prevent overfitting on strong local correlations. ProphetNet is optimized through an n-step look-ahead prediction, which predicts the next n tokens based on previous context tokens at each time step, encouraging the model to explicitly plan for future tokens. It was evaluated on benchmarks for abstractive summarization and question generation tasks such as CNN/DailyMail, Gigaword, and SQuAD 1.1 (Rajpurkar et al., 2016). ProphetNet has a 12-layer encoder and 12-layer decoder with 1024 embedding/hidden size and 4096 feed-forward filter size. The batch size and training steps were set to 1024 and 500K, respectively, and Adam optimization was used with a learning rate of 3×10^{-4} for pre-training. The input length was set to 512 and masking was done randomly in continuous spans every 64 tokens, with 15% of the total number of tokens masked.

5W QA pair generation is a result of two submodules: (i) 5W SRL, and (ii) 5W-based QA pair generations. We have used pretrained models of context-based question generation models, wrapped in automation infrastructure. Contexts are the actual claim, and the answers are the Semantic Role Labeling outputs. As an example, let's consider a claim, "After April 11, 2020, there was a fatality rate of over 1.61 in Malaysia during the coronavirus pandemic". After applying SRL, we obtain the answer to the "When" of the input sentence, yielding "After April 11, 2020". Next, we feed the answer obtained in the prior step (After April 11, 2020) along with the context ("After April 11, 2020, there was a fatality rate of over 1.61 in Malaysia during the coronavirus pandemic.") as the input to the model. Finally, this yields a question starting with "When", which in this case is "When was the COVID-19 pandemic?".

G.1 Human evaluation of 5W SRL and QA generation

For the evaluation purpose, a random sample of 3500 data points was selected for annotation. The questions generated using the Prophetnet model were utilized for this purpose. The annotators were instructed to evaluate the question-answer pairs in three dimensions: the question is well formed, which means it is syntactically correct, the question is correct which means it is semantically correct with respect to the given claim, and extracted answer from the model is correct. The evaluation results for the datasets are presented in the following analysis, see table 6

	FaVIQ	FEVER	HoVer	VitaminC	Factify1.0	Factify2.0
Who	89%	85%	90%	87%	86%	82%
What	85%	56%	68%	78%	81%	93%
When	86%	90%	95%	98%	83%	75%
Where	93%	100%	90%	97%	93%	86%
Why	0%	-	100%	92%	87%	93%

Table 5: Human evaluation of 5W SRL; It is observed that for most of the datapoints *why* is missing

		FaVIQ	FEVER	HoVer	VitaminC	Factify 1.0	Factify 2.0
Who	Question is well-formed	86%	77%	84%	79%	80%	82%
	Question is correct	90%	82%	86%	83%	87%	89%
	Answer is correct	89%	85%	90%	87%	86%	82%
What	Question is well-formed	71%	53%	68%	79%	77%	72%
	Question is correct	77%	69%	70%	81%	80%	76%
	Answer is correct	85%	56%	68%	78%	81%	93%
When	Question is well-formed	88%	77%	86%	78%	81%	78%
	Question is correct	90%	86%	88%	94%	92%	89%
	Answer is correct	86%	90%	95%	98%	83%	75%
Where	Question is well-formed	90%	95%	68%	87%	91%	88%
	Question is correct	85%	95%	78%	92%	92%	83%
	Answer is correct	93%	97%	90%	97%	93%	86%
Why	Question is well-formed	0%	-	100%	92%	92%	90%
	Question is correct	0%	-	100%	95%	95%	94%
	Answer is correct	0%	-	100%	96%	87%	93%

Table 6: Human evaluation of QA generation

H 5W QA-based validation

To design the 5W QA validation system, we utilized the claims, evidence documents, and 5W questions generated by the question generation system as input. The answer generated by the 5W QG model is treated as the gold standard for comparison between claim and evidence. We experimented with three models, T5-3B (Raffel et al., 2020), T5-Large (Raffel et al., 2020), and Bert-Large (Devlin et al., 2018). The T5 is an encoder-decoder-based language model, that treats this task as text-to-text conversion, with multiple input sequences and produces an output as text. The model is pre-trained using the C4 corpus (Raffel et al., 2020) and fine-tuned on a variety of tasks. T5-Large employs the same encoder-decoder architecture as T5-3B (Raffel et al., 2020), but with a reduced number of parameters. The final model that we experimented with is the Bert-Large (Devlin et al., 2018) model, which utilizes masked language models for pre-training, enabling it to handle various downstream tasks and represent both single and pairs of sentences in a single token sequence. It is trained using MLM and a binarized next-sentence prediction task to understand sentence relationships.

I Selecting the best combination - 5W QAG vs. 5W QA validation

We have utilized off-the-self models both for 5W question-answer generation and 5W question-answer validation. Given that the datasets using for training the models bear an obvious discrepancy in terms of the distribution characteristics compared to our data (world news) which would probably lead to a generalization gap, it was essential to experimentally judge which system offered the best performance for our use-case. Instead of choosing the best system for generation vs. validation, we opted for pair-wise validation to ensure we chose the best combination. Table 7 details our evaluation results – the rows denote the QA models while the columns denote QAG models. From the results in the table, we can see that the best combination in terms of a QAG and QA validation model was identified as T5-3b and ProphetNet respectively.

	ProphetNet								BART							
	Claim				+Paraphrase				Claim				+Paraphrase			
	BLEU	ROUGHEL	Recall	F1	BLEU	ROUGHEL	Recall	F1	BLEU	ROUGHEL	Recall	F1	BLEU	ROUGHEL	Recall	F1
T5-3b	**29.22**	**48.13**	**35.66**	**38.03**	**28.13**	**46.18**	**34.15**	**36.62**	21.78	34.53	28.03	28.07	20.93	33.57	27.65	27.24
T5-Large	28.81	48.02	35.26	37.81	21.46	46.45	27.19	36.76	21.46	34.90	27.41	27.99	20.88	33.69	20.88	27.31
BERT large	28.65	46.25	34.55	36.72	27.27	44.10	32.95	35	20.66	33.19	25.51	26.44	19.74	32.34	25.14	25.71

Table 7: Selecting the best combination - 5W QAG vs. 5W QA validation

J   Injecting adversarial assertion for fake news

The extraordinary capabilities of today's large language models to generate realistic text based on prompts has had an electrifying impact on the scientific community. Per (Story, 2022), "Human reviewers could only detect fake abstracts [of scientific articles] 68% of the time". Given these major advances in language models, it is even easier today to generate and propagate misinformation in the form of fake news that would be extremely difficult, even for human experts, to detect as false without the proper tools to verify its authenticity.

We have thus included some fake news claims synthetically generated by OPT in our dataset to provide a more realistic view of news media in recent times. This adversarial attack would help build more robust fact verification models if they are able to detect these fake claims.
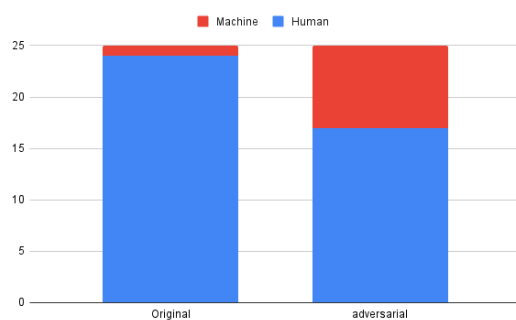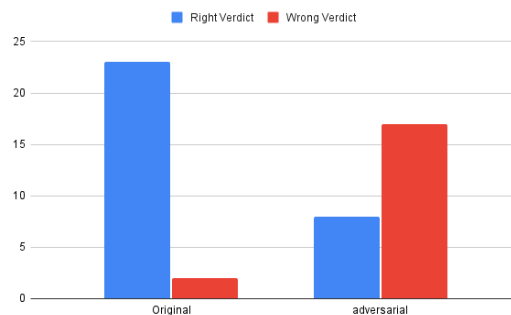

Figure 16: Representation of Human vs Machine

Figure 17: Representation of Right vs Wrong verdicts