From Generation to Detection: A Multimodal Multi-Task Dataset for Benchmarking Health Misinformation

Zhihao Zhang[1]*, **Yiran Zhang**[1]*, **Xiyue Zhou**[2], **Liting Huang**[3], **Imran Razzak**[4], **Preslav Nakov**[4], **Usman Naseem**[1]
Macquarie University[1], University of Sydney[2], UTS[3], MBZUAI[4]
{zhihao.zhang, yiran.zhang, usman.naseem}@mq.edu.au
{imran.razzak, preslav.nakov}@mbzuai.ac.ae

Abstract

Infodemics and health misinformation have significant negative impact on individuals and society, exacerbating confusion and increasing hesitancy in adopting recommended health measures. Recent advancements in generative AI, capable of producing realistic, human-like text and images, have significantly accelerated the spread and expanded the reach of health misinformation, resulting in an alarming surge in its dissemination. To combat the infodemics, most existing work has focused on developing misinformation datasets from social media and fact-checking platforms, but has faced limitations in topical coverage, inclusion of AI-generation, and accessibility of raw content. To address these gaps, we present MM-Health, a large scale multimodal misinformation dataset in the health domain consisting of 34,746 news article encompassing both textual and visual information. MM-Health includes human-generated multimodal information (5,776 articles) and AI-generated multimodal information (28,880 articles) from various SOTA generative AI models. Additionally, We benchmarked our dataset against three tasks—reliability checks, originality checks, and fine-grained AI detection—demonstrating that existing SOTA models struggle to accurately distinguish the reliability and origin of information. Our dataset aims to support the development of misinformation detection across various health scenarios, facilitating the detection of human and machine-generated content at multimodal levels[1].

1 Introduction

Health misinformation refers to information that is inaccurate, misleading, or false according to the best available health evidence at the time. Such misinformation tends to spread at unprecedented speed and scale on the World Wide Web and social media (Office of the Surgeon General (OSG), 2021), and has been shown to have significant negative effects on society (Borges do Nascimento et al., 2022; Abbott et al., 2021; Wang et al., 2019; Muhammed T and Mathew, 2022; Nathan Walter and Suresh, 2021). During crises, such as outbreaks of infectious diseases like COVID-19, the overproduction of health misinformation in both digital and physical environments is defined as an infodemic. As society transitions into a long COVID era (Davis et al., 2023), the infodemic continues to evolve, leading to a reduction in public trust in health professionals (Nathan Walter and Suresh, 2021). Furthermore, unfiltered exposure to health misinformation can delay or prevent effective disease treatment and even threaten the lives of individuals (Wang et al., 2019). Additionally, the recent advent of generative image models, such as Stable Diffusion (Rombach et al., 2022), MidjourneyV5 (Holz et al.), DALL-E2 (Ramesh et al., 2022), as well as human-level text generators like ChatGPT (Ouyang et al., 2022) and LLaMa (Touvron et al., 2023), has amplified both the quantity and spread of misinformation (Zhou et al., 2023a). The ease with which generative AI models can replicate or manipulate multimodal content, including text and images (Huang et al., 2024), further contributes to the emergence of health misinformation (Park, 2024; Ahmad et al., 2025; Shah et al., 2024). This presents a significant new challenge in combating the infodemic.

Recent studies have focused on developing and analyzing datasets and detection methods to combat the infodemic. Common methods for collecting multimodal health misinformation include scraping social media platforms (Kinsora et al., 2017; Hayawi et al., 2022; Cui and Lee, 2020; Zhou et al., 2020; Srba et al., 2022; Sun et al., 2023; Li et al., 2020; Chen et al., 2021b), such as Twitter (now known as X), and parsing fact-checking web-

*Equal contribution
[1]Our code and data is available at: https://github.com/grantzyr/MM-Health-Dataset

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24245–24260
November 4-9, 2025

Name	Year	Human Context		Machine Context		Multiple AI Models	Reliability	Originality	General Avaliable
		Text	Image	Text	Image				
MedHelp	2013	✔	✘	✘	✘	✘	✔	✘	✘
COAID	2020	✔	✘	✘	✘	✘	✔	✘	✘
ANTi-Vax	2021	✔	✘	✘	✘	✘	✔	✘	✘
MM-COVID	2020	✔	✔	✘	✘	✘	✔	✘	Partly
ReCOVery	2020	✔	✔	✘	✘	✘	✔	✘	Partly
Monant	2022	✔	✔	✘	✘	✘	✘	✘	Partly
MMCOVAR	2021	✔	✔	✘	✘	✘	✔	✘	Partly
Med-MMHL	2023	✔	✔	✔	✘	✘	✔	✔	Partly
Ours	2025	✔	✔	✔	✔	✔	✔	✔	✔

Table 1: Comparison between our MM-Health dataset and other health related misinformation datasets.

sites (Li et al., 2020; Chen et al., 2021b; Srba et al., 2022; Sun et al., 2023), such as Snopes and Poynter. However, these datasets exhibit the following notable limitations:

- **Exclusion of AI-generated misinformation:** The majority of existing datasets focus solely on human-generated misinformation, ignoring the growing trend of AI-generated misinformation (Zhou et al., 2023b; Bashardoust et al., 2024; Xu et al., 2023). With AI models now capable of generating and manipulating both text and image content (Liz-López et al., 2024), it is crucial to include AI-generated misinformation to ensure data diversity.
- **Lack of accessible raw content:** Many datasets only provide URLs or Twitter IDs as delivery artifacts instead of the raw content of the misinformation. Due to ongoing censorship by news and social media platforms, much of this data is no longer accessible, severely limiting the usability of these datasets.

The limitation of the previous health misinformation datasets are summarised in Table 1. Most of the existing methods for automatically detecting health misinformation include content-based CNNs (Cui and Lee, 2020; Zhou et al., 2020; Dai et al., 2020), attention-based hierarchical encoders (Kinkead et al., 2020), and graph-based attention methods (Cui et al., 2020). These methods have demonstrated competitive performance after being trained on health misinformation benchmarks. However, they require extensive training on benchmark datasets to achieve expected performance, which limits the models' ability to generalise to new health misinformation. Since health misinformation is constantly evolving (Okoro et al., 2024), it is important to evaluate generalised models under zero-shot settings. Additionally, these models only accept structured input and provide labelled output without human-readable explanations. This constraint limits the usability of existing detection methods for the general public outside of the research community. To address the aforementioned limitations, we develop MM-Health, a comprehensive multimodal dataset designed for detecting both human and AI generated health misinformation. Additionally, we conduct extensive experiments using state-of-the-art (SOTA) Vision-Language Models (VLLMs) to evaluate their performance in assessing both the reliability and originality of the health information in MM-Health, as well as performing a fine-grained AI detection analysis. Our **key contributions** are as follows:

- We create and release a new large-scale multimodal (text and images) dataset called MM-Health, which contains 34,746 news articles. These articles were collected from existing multimodal health datasets and generated using state-of-the-art (SOTA) AI generative models.
- We conduct a preliminary analysis of the data for both human and AI-generated content, establishing our dataset as a benchmark for baseline evaluation against several SOTA Vision Large Language Models (VLLMs) across three tasks: reliability checks, originality checks, and fine-grained AI detection.
- Our experimental results and analysis show that current VLLMs, including GPT-4o [2], struggle to accurately verify content reliability and originality. This highlights the importance of developing generalised models for multimodal health misinformation detection.

2 Related Work

Web and social media provide valuable sources of information and play important roles in multiple tasks like depression and suicide identifica-

[2] https://openai.com/index/hello-gpt-4o/

tion (Naseem et al., 2024), health mention classification (Naseem et al., 2022) and creditable knowledge management (Nisar et al., 2019). However, systematic reviews shows that health-related misinformation on web and social media posts critical threat to the community, which could lead to delay or prevent treatment and even threaten the lives of individuals (Wang et al., 2019).

2.1 Existing Datasets

As detailed in Table 1, several benchmark datasets have been proposed for health related information detection, which contain various features from different data sources.

MedHelp (Kinsora et al., 2017), **COAID** (Cui and Lee, 2020), and **ANTi-Vax** (Hayawi et al., 2022) are text datasets designed to address online health misinformation. Specifically, MedHelp is collected from an online general health discussion forum. It contains 1,338 non-misinformation comments and 887 misinformation comments annotated by human experts between 2001 and 2013. COAID includes both news articles and Twitter content focused on COVID-19 between December 2019 and September 2020. It labels 204 fake news articles, 3,565 true news articles, 28 fake claims, and 454 true claims from a total of 1,896 news articles, 516 Twitter posts, and 183,569 Twitter engagements. ANTi-Vax is built solely from Twitter content focused on the COVID-19 vaccine between December 2020 and January 2021. It includes a total of 15,073 tweets, 5,751 of which are misinformation and 9,322 are general vaccine-related tweets.

MM-COVID (Li et al., 2020), **ReCOVery** (Zhou et al., 2020), **Monant** (Srba et al., 2022), **MMCoVaR** (Chen et al., 2021b), and **Med-MMHL** (Sun et al., 2023) are multimodal datasets for health-related misinformation detection. MM-COVID contains 3,981 fake articles and 7,584 real articles related to COVID-19 which collected from Feb 2020 to Jul 2020. ReCOVery includes both news articles and tweets mentioning COVID-19, labelled by NewsGuard [3] and Media Bias/Fact Check [4]. It comprises 118,339 reliable articles and 28,274 unreliable articles sourced between January 2020 and May 2020. MMCoVaR focuses on vaccine-related misinformation between February 2020 and March 2021, featuring 958 unreliable and 1,635 reliable news articles, plus 24,184 tweets categorized as reliable, unreliable, or inconclusive. Med-MMHL covers multiple diseases, integrating LLM-generated fake news, with 1,979 real and 1,604 fake articles, alongside 1,334 reliable and 639 unreliable tweets from January 2017 to May 2023. There are no existing dataset that incorporate both text and image AI for misinformation generation. Furthermore, the most datasets rely on website URLs or tweet IDs, which have become partially or fully inaccessible due to Twitter's API restrictions and ongoing content censorship.

2.2 Existing Methods

2.2.1 Methods for textual data

Existing methods typically utilize a latent space to create embedding representations for article content. Embeddings capture contextual information and encode it into a lower-dimensional representation for downstream tasks such as classification and prediction (Papanikou et al., 2024). Many approaches leverage this feature using the BERT to achieve exceptional performance in health misinformation classification, either through fine-tuning (Kaliyar et al., 2021) or a combination of pre-training and fine-tuning (Lee et al., 2019). Other strategies, such as BERTweet (Nguyen et al., 2020) and COVID-Twitter-BERT (Müller et al., 2023), combine multiple pre-trained models across various contexts. These hybrid models achieve higher performance due to the distinct contextual understanding provided by different models. However, textual models capture only single-modality features, neglecting the visual information often associated with articles. We address this limitation by using SOTA vision-language models to capture both textual and visual context.

2.2.2 Methods for multimodal data

Previous studies have found that image-containing posts or articles tend to have higher user interaction, spread more quickly, and trigger stronger emotional responses (Papanikou et al., 2024). Multimodal misinformation detection requires both image and text encoders to capture textual and visual features. One method (Uppada et al., 2023) leverages these encoders to identify fake posts containing visual data, while another approach detects inconsistencies between textual entities and image content for misinformation detection. Limited research has explored multimodal health misinformation detection. DGExplain (Shang et al., 2022) is a recent multimodal encoder-decoder architecture for COVID-19 misinformation detection that investigates cross-

[3] https://www.newsguardtech.com/
[4] https://mediabiasfactcheck.com/

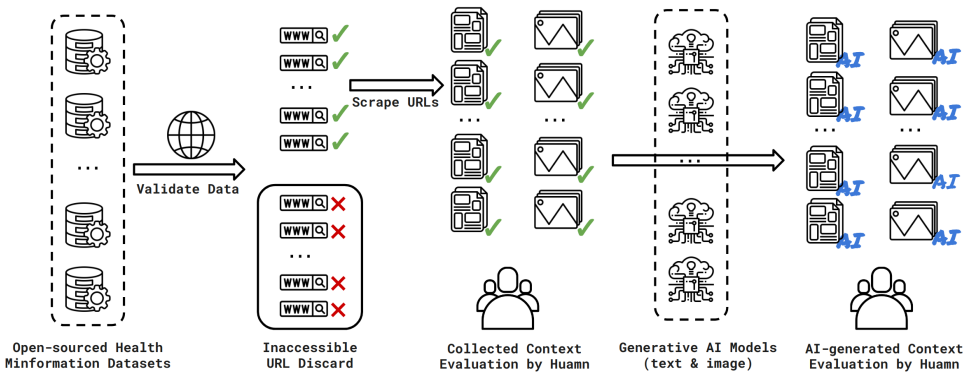


Figure 1: Data collection process for MM-Health includes: 1) utilising multiple existing open-source health misinformation datasets as data sources, 2) validating the available data samples and collecting human-generated multimodal data from the provided URLs, and 3) implementing generative AI models to collect AI-generated replicated multimodal data. To ensure data quality, both human and AI generated content are evaluated by five human evaluators proficient in English.

modal associations between text and images to assess the reliability of news articles. However, these models often require extensive training on benchmark datasets and offer limited interpretability. Our work applies current advanced Vision-Language Models as a baseline for health misinformation detection to evaluate their performance.

3 MM-Health Dataset

The overall data collection process for MM-Health, including human and AI-generated content, is illustrated in Figure 1. This process can be summarised as follows: 1) utilising multiple existing open-source health misinformation datasets as data sources, 2) validating the available data samples and collecting human-generated multimodal data from the provided URLs, and 3) implementing generative AI models to collect AI-generated replicated multimodal data. To ensure data quality, both human and AI generated content are evaluated by five human evaluators proficient in English.

3.1 Data Collection

3.1.1 Health Misinformation Collection

To ensure variability, we utilise four recent health-related misinformation datasets (Sun et al., 2023; Chen et al., 2021b; Li et al., 2020; Zhou et al., 2020), collected between February 2020 and May 2023, as our primary data sources. Since Twitter (now known as X.com) restricted data access through its application programming interface (API) in 2023, we exclude tweets from these datasets and focus on the available news articles. While Med-MMHL provides raw text and image data, the other datasets only provide URLs as their data source. However, not all URLs are accessible due to news websites consistently removing outdated or misleading news, either voluntarily or in response to complaints. Therefore, we validate the URLs in the database and scrape the text and image metadata from the available URLs and attach the reliable and unreliable labels extracted from the original datasets. We also use the Pillow[5] library to filter out some images, such as blurry, logo-based, or meaningless images, which are often downloaded during web scraping. Examples of removed images are in Appendix (Figure 5).

3.1.2 AI Health Misinformation Generation

After collecting human generated misinformation from open-source datasets, we deploy five state-of-the-art (SOTA) generative AI models for each modality to generate AI replicated data. To ensure data variability, we adopt text and image models with different architectures. For text generation, we use Llama-3.1-8B, Qwen2.5-7B, ChatGLM-4-9B, Gemma2-9B, and Mistral-v0.3-7B. For image generation, we use FLUX.1-dev [6], Stable Diffusion 1.5, Stable Diffusion XL, Stable Diffusion VAE, and Stable Diffusion PAG. To enhance model understanding during text generation, we integrate both textual and visual context, ensuring diverse and comprehensive outputs. Specifically, we implement a two-step generation pipeline as follows: 1) providing raw text and image data to GPT-4o to create a topic summary for the article, and 2) inputting the topic summary into the five selected LLMs to generate the desired text. We find that

[5] https://python-pillow.org/
[6] https://github.com/black-forest-labs/flux

specific and concise prompts yield the best results for LLMs. Image generation follows a similar two-step pipeline to that of text generation. First, the images are sent to GPT-4o to generate short image captions. Next, the original images and their captions are supplied to the image generation models to create new images. The original images serve as anchor points during the generation process, predefining the de-noising parameters as well as the image dimensions. All the text and image models are provided with their default parameter settings. Refer to Figure 6 in appendix for text generation prompts and image generation prompts.

3.1.3 Post-Processing with Human Evaluation

Our two-stage generation process allows AI models to create most of the text and images based on existing data without triggering their internal censorship mechanisms (Glukhov et al., 2023). However, some data samples are still not generated, as models may refuse to produce them or output empty strings and black images. Since different models are equipped with different censorship mechanisms, there are variations in behaviour between models. To address this, we developed an algorithm to post-process data alignment between the human-generated context and all five AI-generated contexts(see the Algorithm 1 in appendix for the implementation). The dataset consists of two sources: human-collected text and images from open-source datasets and AI-generated text-image pairs created by five different models. For consistency, our algorithm will ensured that each AI-generated sample was present across all five models before being included in the final dataset. After data matching, human evaluation was conducted by five independent annotators. These annotators were students ranging from junior to senior levels, including both male and female participants. All had backgrounds in AI literature and a foundational understanding of health-related topics, which helped ensure informed and thorough evaluation. Before the annotation task, all participants underwent thorough training to familiarize them with the evaluation criteria, common failure modes (e.g., logo artifacts, mismatched captions), and the overall goal of the task. This training ensured consistency and reliability in their assessments. The annotators were compensated fairly according to the university's standard participant payment guidelines. Inter-annotator agreement was measured and found to be high, with a score above 80%, indicating strong consistency across evaluators. While domain-specific expertise was not strictly required, the annotators' familiarity with AI and health contexts contributed to a more nuanced understanding of the generated image–text pairs. This results in 34,746 pieces of health-related information, of which 5,791 were collected from open-source datasets and 28,955 were generated by AI models Examples of reliable and unreliable news generated by humans and AI are in Figure 7 in the appendix. The news articles are written in diverse styles by different LLMs while addressing the same topic. While Stable Diffusion 1.5 generates dramatically different images, other models slightly manipulate the original images while maintaining a realistic appearance.

Model	Avg Len	Avg CSS
Llama-3.1-8B	479.57	0.761
Qwen2.5-7B	568.4	0.766
ChatGLM-4-9B	577.55	0.761
Gemma2-9B	340.45	0.774
Mistral-v0.3-7B	337.87	0.783

Model	FID	Avg CIS
FLUX.1-dev	12.33	0.869
SD 1.5	27.72	0.737
SD XL	19.42	0.866
SD XL VAE	15.76	0.883
SD XL PAG	19.30	0.866

Table 2: Statistics of generated text and image data from different models. The left table shows text statistics: "Avg Len" is average word count, and "Avg CSS" is average Cosine Semantic Similarity. The right table shows image statistics: "FID" is Fréchet Inception Distance and "Avg CIS" is Average Cosine Image Similarity.

3.2 Data Statistics and Analysis

Overall data statistics. All instances in MM-Health are multimodal, containing both textual and visual information for reliability and originality studies, with the proportion of reliable to unreliable articles being approximately 4:1. We split our dataset into training and testing sets by randomly selecting 80% and 20% of the news articles from the collected data instances. Additionally, 10% of the training set is further split into a validation set. As shown in Table 5, this results in a total of 4,154 training samples, 463 validation samples, and 1,159 testing samples from human-generated news articles, while 20,770, 2,315, and 5,795 samples are used for training, validation, and testing, respectively, from AI-generated news articles. To ensure the generalisation of our dataset across all split sets, we further analyse the data statistics in each set, including the average text length and the average number of images per article from both human and AI sources. The average text length and number of images are consistent across different sets. It is worth noting that the average text length generated by AI is shorter than that of the original human text,

with an average of around 450 words compared to approximately 850 words. Additionally, unreliable news articles tend to be shorter, averaging around 650 to 750 words, while containing more images per article around 6 to 8 images compared to around 1.5 images in reliable articles.The overall statistics of the dataset are presented in appendix Table 5.

Data analysis. To identify the differing data characteristics across the five AI models, we compare their statistics in Table 2. For text models, we calculated the average article length from all five text models. There is a noticeable variation among the models, with ChatGLM-4 tending to generate longer articles, while Mistral-v0.3 is more likely to generate shorter articles. Additionally, we utilise the OpenAI `text-embedding-3-small` model to generate text embeddings for both human and AI articles, comparing their semantic differences using cosine similarity. For news articles that are too long to fit into the model's 8,191 context window, we split these long articles into smaller chunks and compute the mean vectors of all chunks to obtain the final text embedding representations for those articles. As indicated on the left side of Table 2, the average cosine semantic similarities across the text models range from 0.76 to 0.78. We plot the Gaussian Kernel Density Estimation (KDE) (Zhu et al., 2022) distribution of the semantic similarity in Figure 2, showing that the cosine similarities are concentrated between 0.8 and 0.85 across all models, with slight variance. Notably, a small number of generated articles exhibit low semantic similarity when compared to human generated articles.

For image models, we calculated both the Fréchet Inception Distance (FID) (Seitzer, 2020) and the average cosine image similarity between AI-generated images and collected images. FID is a common metric used to evaluate the quality of images created from generative models by comparing the distribution of generated images with that of real images. A lower FID score indicates greater similarity between the generated and real images. The average cosine similarity is computed between the vector representations of AI and real images, extracted using ResNet (He et al., 2016). The average FID and average cosine image similarities are reported in right side of Table 2, we also plot KDE of the cosine image similarity in Figure 3. We observe that Stable Diffusion 1.5 exhibits the greatest image variance compared to other models, while Stable Diffusion XL and its VAE and PAG variations generate the most similar images. These trends hold for both the average and distribution analyses of the image models. Unlike the text models, no generated images exhibit low similarity when compared to real images.

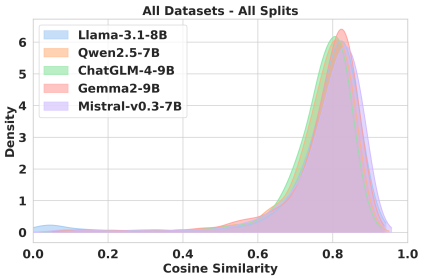


Figure 2: KDE distribution of the semantic similarity between the human articles and articles from five LLMs.

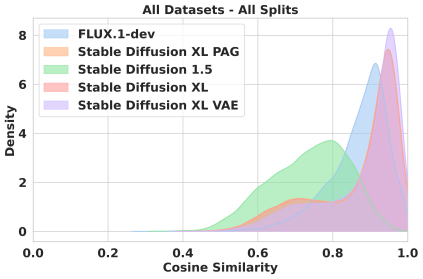


Figure 3: KDE distribution of image similarity between real and generated images across five image models.

4 Experiments

Existing multimodal vision LLMs can provide insightful descriptions for input text and image data, offering the most accessible way to interact compared to other deep learning models that require extensive training. We utilise both proprietary and open-source models as baselines to detect data reliability and originality using our MM-Health dataset and describe the benchmarking of these models.

4.1 Task Definition

We designed three benchmark tasks to evaluate baseline models' performance against both the reliability and originality of the health information collected in MM-Health. Those tasks are described as follows:

Task 1: Information reliability check. This task evaluates the VLLMs' capability to distinguish whether the provided information is reliable or unreliable. We provide the models with three different data settings: text-only data with human and AI content separated, text-image data with human and AI content separated, and text-image data with human and AI content mixed.

		Reliable										Unreliable									
		4o		4o-mini		Llama3.2-V		LLaVA-1.6		Qwen2-VL		MedGemma		4o		4o-mini		Llama3.2-V		LLaVA-1.5		Qwen2-VL		MedGemma	
		ZS		ZS		ZS	FS	ZS	FS	ZS	FS	ZS	FS	ZS		ZS		ZS	FS	ZS	FS	ZS	FS	ZS	FS
Text	Human	0.499		0.499		0.500	0.499	0.500	0.494	0.499	0.499	0.499	0.485	0.334		0.328		0.246	0.376	0.077	0.348	0.162	0.348	0.295	0.409
	AI	0.499		0.500		1.000	0.436	1.000	0.498	0.500	0.499	1.000	0.499	0.183		0.105		0.020	0.238	0.010	0.101	0.019	0.113	0.081	0.208
Text & Image	Human	0.499		0.499		0.499	0.499	0.500	0.495	0.499	0.498	0.527	0.495	0.371		0.353		0.312	0.373	0.085	0.351	0.140	0.340	0.304	0.393
	AI	0.498		0.499		0.498	0.454	1.000	0.498	0.499	0.500	0.532	0.499	0.314		0.261		0.250	0.331	0.030	0.207	0.084	0.247	0.229	0.327
	AI-T 25%	0.499		0.499		0.499	0.468	1.000	0.497	0.498	0.500	0.499	0.499	0.347		0.304		0.287	0.349	0.065	0.267	0.124	0.302	0.274	0.381
	AI-T 50%	0.500		1.000		0.499	0.477	1.000	0.497	0.499	0.500	0.499	0.499	0.347		0.304		0.286	0.347	0.065	0.267	0.124	0.302	0.274	0.385
	AI-T 75%	1.000		1.000		1.000	0.490	1.000	0.497	0.499	0.500	0.499	0.499	0.395		0.372		0.330	0.381	0.086	0.339	0.207	0.392	0.363	0.437
	AI-I 25%	0.498		0.499		0.498	0.450	1.000	0.499	0.500	0.500	0.499	0.499	0.313		0.262		0.245	0.328	0.040	0.221	0.089	0.244	0.231	0.333
	AI-I 50%	0.499		1.000		0.499	0.450	1.000	0.499	0.500	0.499	0.500	0.500	0.310		0.264		0.253	0.333	0.043	0.219	0.075	0.231	0.261	0.341
	AI-I 75%	1.000		1.000		0.499	0.449	1.000	0.497	1.000	0.499	1.000	1.000	0.310		0.265		0.252	0.333	0.037	0.212	0.077	0.244	0.253	0.336

Table 3: Marco F1 score of **Task 1 information reliability check**. "ZS" is zero-shot and "FS" is five-shot.

Task 2: Information originality check. This task evaluates the VLLMs' capability to distinguish whether the provided information is AI-generated, human-generated, or mixed. Following a similar approach to Task 1, we provide the models with content that is either separated into human and AI categories or mixed.

Task 3: Fine-Grained AI Detection Analysis. This task provides deeper insights into the information originality check from Task 2. Since we applied five different text and image generation models for our MM-Health dataset, we mixed the AI-generated data to create twenty-five different variations. This allows us to evaluate which combinations are the easiest or hardest for baseline models to detect as AI-generated.

4.2 Baseline Models

We established various multimodal baselines, including both proprietary and open-source models for a comprehensive comparison. **Proprietary models:** We employed GPT-4o and GPT-4o Mini, which have demonstrated excellent performance on open-source benchmarks such as MMLU (Hendrycks et al., 2021), HumanEval (Chen et al., 2021a), and MMMU (Yue et al., 2024), and are widely adopted in vision and language research. These models serve as a benchmark standard, representing the top zero-shot performance that current state-of-the-art VLLMs can achieve. **Open-sourced models:** We include three general open-source VLLMs for a comprehensive comparison. These models are: Llama-3.2-Vision, built on the Llama-3.1-8B text model with a 2-billion-parameter vision encoder; LLaVA-1.6, which utilises the 7-billion-parameter Qwen v1.5 model with CLIP-ViT-L-336 attached via an MLP projection layer; and Qwen2-VL, which employs Multimodal Rotary Position Embedding (M-RoPE) with a 675M vision encoder and Qwen2-7B, resulting in a total of 7.6B parameters. To further access the performance of domain specific model, we also include MedGemma, which is a 4B model of Gemma3 variants that are trained for performance on medical text and image comprehension. These models demonstrate competitive zero-shot performance on vision-and-language benchmarks.

4.3 Experiment Settings.

For the GPT-4o and GPT-4o mini models, we use the OpenAI API endpoints: `gpt-4o-2024-08-06` and `gpt-4o-mini-2024-07-18`, respectively, throughout our experiments. Open-source models are hosted on a server equipped with an Nvidia A6000 GPU with 48GB of VRAM. To replicate realistic scenarios in most cases, we retain the default parameter settings for all models. To fully utilise the potential of the models, we experiment with baseline models using both zero-shot and five-shot settings, employing consistent prompts designed for different tasks. Zero-shot is directly evaluated on the testing set, while five-shot uses five samples from the training set and evaluates the models on the testing set. We use standard macro F1 scores to evaluate the baseline models across all three tasks. See appendix for details.

5 Result

Task 1: Information reliability check. The results of the information reliability check are presented in Table 3. We divide the dataset into reliable and unreliable articles and apply VLLMs for binary classification across various data settings, including text-only, text-image, and mixed text-image content incorporating both human and AI-generated inputs. To assess the impact of AI-generated content, we introduce varying proportions of AI-generated context at 25%, 50%, and 75%, in both text and image modalities. Baseline models perform better at classifying reliable articles than unreliable

		Reliable										Unreliable									
		4o		4o-mini		Llama3.2-V		LLaVA-1.6		Qwen2-VL		MedGemma		4o		4o-mini		Llama3.2-V		LLaVA-1.6		Qwen2-VL		MedGemma	
		ZS		ZS		ZS	FS	ZS	FS	ZS	FS	ZS	FS	ZS		ZS		ZS	FS	ZS	FS	ZS	FS	ZS	FS
Text	**Human**	0.316		0.323		0.023	0.321	0.242	0.321	0.493	0.324	0.001	0.075	0.254		0.266		0.012	0.255	0.201	0.267	0.320	0.288	0.006	0.023
	AI	0.093		0.074		0.327	0.057	0.211	0.041	0.003	0.046	0.000	0.166	0.113		0.122		0.495	0.083	0.220	0.062	0.003	0.043	0.028	0.203
Text & Image	**Human**	0.234		0.304		0.140	0.312	0.137	0.319	0.321	0.317	0.008	0.012	0.263		0.239		0.069	0.237	0.138	0.279	0.301	0.298	0.009	0.014
	AI	0.155		0.119		0.188	0.122	0.271	0.070	0.093	0.071	0.156	0.174	0.215		0.233		0.231	0.164	0.282	0.063	0.091	0.099	0.211	0.178
	AI-T 25%	0.234		0.172		0.261	0.144	0.261	0.091	0.122	0.094	0.199	0.215	0.285		0.335		0.259	0.145	0.272	0.079	0.122	0.089	0.279	0.241
	AI-T 50%	0.268		0.188		0.316	0.173	0.239	0.118	0.143	0.125	0.253	0.283	0.306		0.339		0.284	0.159	0.270	0.091	0.111	0.097	0.308	0.267
	AI-T 75%	0.307		0.205		0.338	0.171	0.203	0.127	0.162	0.138	0.280	0.309	0.389		0.400		0.305	0.221	0.173	0.096	0.134	0.116	0.272	0.262
	AI-I 25%	0.236		0.143		0.256	0.173	0.262	0.100	0.114	0.096	0.207	0.218	0.261		0.264		0.279	0.176	0.295	0.067	0.093	0.118	0.262	0.235
	AI-I 50%	0.270		0.180		0.284	0.215	0.235	0.109	0.143	0.108	0.246	0.256	0.369		0.282		0.319	0.253	0.193	0.082	0.117	0.188	0.259	0.276
	AI-I 75%	0.322		0.218		0.333	0.246	0.185	0.134	0.156	0.130	0.291	0.283	0.358		0.275		0.316	0.203	0.154	0.093	0.118	0.194	0.266	0.294

Table 4: Marco F1 score of **Task 2 information originality check**. "ZS" is zero-shot and "FS" is five-shot.

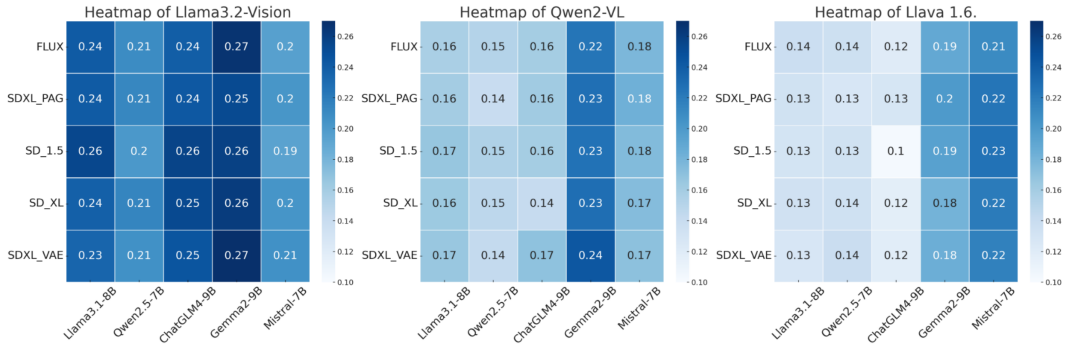


Figure 4: Heatmap representation of the **Task 3 fine-grained AI detection analysis**. Each heatmap illustrates F1 scores from various VLLMs across twenty-five different combinations of AI-generated content. Darker colours represent higher F1 scores.

ones, with a significant disparity in F1 scores observed in both zero-shot and five-shot settings. In zero-shot setting, some models, such as Llava-1.6, MedGemma, GPT-4o, and GPT-4o-min, achieve perfect F1 scores for reliable articles but struggle to classify unreliable articles. This may be caused by that the models tend to classify input information as reliable without proper verification. In the five-shot setting, models' performance improve on unreliable articles but worsen on reliable ones. This suggests that example-based learning helps the models better understand content for verification. MedGemma's noticeably higher zero-shot and five-shot performance than other open-source models suggests that a model with health-domain-specific knowledge is better equipped to identify the reliability of health-related information.

Task 2: Information originality check. The results of the originality check are presented in Table 4. We classify the dataset into three categories: human-generated, AI-generated, and a mix of text-image content combining human and AI inputs, for multi-label classification. VLLMs are tasked with distinguishing whether the input context is human, AI, or human-AI mixed. Similar to Task 1, we introduce AI-generated content in both text and images at proportions of 25%, 50%, and 75% to further evaluate the robustness of VLLMs in distinguishing mixed content. Baseline models struggle to accurately classify the originality of articles in both zero-shot and five-shot settings. We observe that baseline models may perform worse in the five-shot setting, suggesting that current VLLMs may lack the ability to learn comprehensive representations for assessing information originality using examples. GPT-4o performs the best among all the models, with MedGemma ranking as the second-best model, slightly trailing GPT-4o in performance. This performance difference is likely attributed to the larger dataset and model size used in GPT-4o's training as well as health specific fine-tuning in MedGemma training.

Task 3: Fine-grained AI detection analysis. We visualise the performance of VLLMs using heatmaps that display the combinations of five text and five image models, resulting in a total of twenty-five results per heatmap, as shown in Figure 4. Each heatmap displays the F1 scores of different open-source VLLMs in the five-shot setting, with darker colours indicating higher F1 scores in the binary classification of AI-generated content. A comparison between the GPT-4o and GPT-4o-mini models in the zero-shot setting is provided in the appendix Figure 11 and Figure 12. Among the

open-source models, Llama-3.2-Vision performs the best, followed by Qwen2-VL and Llava-1.6. We observe that text generated by Gemma-2-9B is more likely to be detected, while text generated by Qwen-2.5-7B is less likely to be detected. However, no consistent trend is observed for image models. The average F1 scores across all baseline models are approximately 0.2, indicating that none of the AI models cannot be reliably detected.

The results across all three tasks highlight the challenges faced by existing SOTA VLLMs. For information reliability check, model exhibit a strong bias towards reliable articles but struggle with unreliable content. In the information originality check, models fail to effectively differentiate between human, AI and mixed content, with five-shot setting showing no significant improvement. The fine-grained AI detection analysis decompose the model performance towards different generative models, with baseline models achieving low average F1 scores, indicating the need for more advanced detection methods. These findings demonstrate that MM-Health presents a challenging benchmark for the community. By releasing the MM-Health dataset, we aim to encourage the development of more robust and effective models to accurately classifying multimodal health misinformation.

6 Case Study

To provide a deeper understanding of our quantitative results, we conducted a qualitative analysis of representative classification errors.

6.1 Task 1: Information reliability check.

Models demonstrated a tendency to rely on superficial stylistic cues rather than semantic content, leading to predictable errors in both directions.

Unreliable Content Misclassified as Reliable. Models were consistently deceived by unreliable articles that mimicked credible journalism. These articles successfully evaded detection by anchoring misleading narratives to a verifiable fact, adopting a professional tone, and avoiding obvious red flags. Instead of containing clear falsehoods, they relied on biased framing and strategic omissions, which the models failed to identify.

Reliable Content Misclassified as Unreliable. Conversely, factually accurate content was often misclassified as unreliable, primarily due to a lack of explicit citations. Models also applied overly cautious heuristics to sensitive domains like medicine and politics, incorrectly penalizing content for using generalized language or offering reasonable predictive analysis.

6.2 Task 2: Information originality check.

Similar patterns were observed in detecting text provenance, where models struggled with content that either perfectly emulated a known format or deviated from a perceived natural baseline.

AI-Generated Content Misclassified as Human. AI-generated text evaded detection by adhering closely to formal structures, such as news reports or scientific summaries. Its plausibility was enhanced by incorporating real-world details, like quoting public figures or referencing known events, causing its impersonal and structured format to be mistaken for authentic human writing.

Human-Written Content Misclassified as AI. Simultaneously, human-authored content was frequently misclassified as AI-generated. This was particularly common with text that was concise, logically structured, or satirical. The models appeared to equate clarity and logical structure with machine-generated text, and they failed to grasp rhetorical nuances like irony, often processing satirical arguments literally.

7 Conclusion

We introduce MM-Health, a multimodal dataset designed for detecting human and AI generated health misinformation. Unlike datasets limited to human generated information or augmented with LLM generated text, our approach combines text and image generative models to create multimodal counterparts of misinformation. To ensure diversity in generative model outputs, we incorporate five different open-source models for each modality. Our experiments involve six different VLLMs, including proprietary GPT-4o and GPT-4omini models, as well as SOTA open-source models, to evaluate data reliability and originality classification. Experimental results reveal that current VLLMs cannot accurately detect misinformation and AI-generated content, highlighting challenges in the era of generative AI. We hope publicly releasing the MM-Health dataset guides future research.

Acknowledgements

This research was supported by the Macquarie University Research Acceleration Scheme (MQRAS) and Data Horizon funding.

Limitations

In this work, we limited data collection to two modalities due to their prevalence on mainstream social platforms. This decision restricts the range of misinformation addressed by MM-Health, particularly from short-video platforms like TikTok, thereby limiting the scope of its potential applications. We encourage future studies to incorporate a broader variety of modalities to further enhance the diversity of machine-generated health information. Additionally, we exclusively employed out-of-the-box multimodal language models for misinformation detection, aiming to emulate common practices among the general public as our baseline. Given the existence of several specialised models designed for misinformation detection, we suggest further research to facilitate a more comprehensive comparison. Finally, there is a potential risk that the synthetic health misinformation could be misused by malicious actors. While we acknowledge this risk, the relatively small scale of the data generated in this study is unlikely to significantly exacerbate the existing challenge posed by the vast amount of health misinformation already circulating on the internet.

Ethical Considerations

All experiments strictly adhere to the Code of Ethics. In Section 3.1.1, which details our human evaluation in data collection, we clearly informed the human evaluators of the task and that their responses would be utilized to assess the capabilities of large generative models. To ensure the anonymity and privacy of individuals involved in the data collection, we implemented a de-identification protocol. We directly remove all human evaluators' names associated with the generated data, all de-identified articles are stored in plain text format, without any identifying information. The original raw data are permanently deleted after the de-identification process. By taking these steps, we ensure that our data collection and analysis processes align with ethical guidelines and data protection regulations.

References

Rebecca Abbott, Alison Bethel, Morwenna Rogers, Rebecca Whear, Noreen Orr, Liz Shaw, Ken Stein, and Jo Thompson Coon. 2021. Characteristics, quality and volume of the first 5 months of the COVID-19 evidence synthesis infodemic: a meta-research study. *BMJ Evid Based Med*, 27(3):169–177.

Syed Talal Ahmad, Haohui Lu, Sidong Liu, Annie Lau, Amin Beheshti, Mark Dras, and Usman Naseem. 2025. Vaxguard: A multi-generator, multi-type, and multi-role dataset for detecting llm-generated vaccine misinformation. *arXiv preprint arXiv:2503.09103*.

Amirsiavosh Bashardoust, Stefan Feuerriegel, and Yash Raj Shrestha. 2024. Comparing the willingness to share for human-generated vs. ai-generated fake news. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Israel Júnior Borges do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bull World Health Organ*, 100(9):544–561.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021a. Evaluating large language models trained on code.

Mingxuan Chen, Xinqiao Chu, and K. P. Subbalakshmi. 2021b. Mmcovar: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *ASONAM '21: International Conference on Advances in Social Networks Analysis and Mining, Virtual Event, The Netherlands, November 8 - 11, 2021*, pages 31–38. ACM.

Limeng Cui and Dongwon Lee. 2020. Coaid: COVID-19 healthcare misinformation dataset. *CoRR*, abs/2006.00885.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 492–502, New York, NY, USA. Association for Computing Machinery.

Enyan Dai, Yiwei Sun, and Suhang Wang. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 853–862. AAAI Press.

Hannah E Davis, Lisa McCorkell, Julia Moore Vogel, and Eric J Topol. 2023. Long COVID: major findings, mechanisms and recommendations. *Nature Reviews Microbiology*, 21(3):133–146.

David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. 2023. Llm censorship: A machine learning challenge or a computer security problem? *Preprint*, arXiv:2307.10719.

K. Hayawi, S. Shahriar, M.A. Serhani, I. Taleb, and S.S. Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23–30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

David Holz, Jim Keller, Nat Friedman, Philip Rosedale, and Bill Warner. Midjourney V5.

Liting Huang, Zhihao Zhang, Yiran Zhang, Xiyue Zhou, and Shoujin Wang. 2024. Ru-ai: A large multimodal dataset for machine generated content detection. *Preprint*, arXiv:2406.04906.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools Appl.*, 80(8):11765–11788.

Laura Kinkead, Ahmed Allam, and Michael Krauthammer. 2020. Autodiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. *BMC Medical Informatics and Decision Making*, 20(1):104.

Alexander Kinsora, Kate Barron, Qiaozhu Mei, and V. G. Vinod Vydiswaran. 2017. Creating a labeled dataset for medical misinformation in health forums. In *2017 IEEE International Conference on Healthcare Informatics, ICHI 2017, Park City, UT, USA, August 23-26, 2017*, pages 456–461. IEEE Computer Society.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A multilingual and multimodal data repository for combating COVID-19 disinformation. *CoRR*, abs/2011.04088.

Helena Liz-López, Mamadou Keita, Abdelmalik Taleb-Ahmed, Abdenour Hadid, Javier Huertas-Tato, and David Camacho. 2024. Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Information Fusion*, 103:102103.

Sadiq Muhammed T and Saji K Mathew. 2022. The disaster of misinformation: a review of research in social media. *Int J Data Sci Anal*, 13(4):271–285.

Martin Müller, Marcel Salathé, and Per E. Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6.

Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G. Dunn. 2024. Hybrid text representation for explainable suicide risk identification on social media. *IEEE Trans. Comput. Soc. Syst.*, 11(4):4663–4672.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2022. Identification of disease or symptom terms in reddit to improve health mention classification. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2573–2581. ACM.

Camille J. Saucier Nathan Walter, John J. Brooks and Sapna Suresh. 2021. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Communication*, 36(13):1776–1784. PMID: 32762260.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Tahir M. Nisar, Guru Prabhakar, and Lubica Strakova. 2019. Social media information benefits, knowledge management and smart organizations. *Journal of Business Research*, 94:264–272.

Office of the Surgeon General (OSG). 2021. *Confronting Health Misinformation: The U.S. Surgeon General's Advisory on Building a Healthy Information Environment*. US Department of Health and Human Services.

Yvonne Oshevwe Okoro, Oluwatoyin Ayo-Farai, Chinedu Paschal Maduka, Chiamaka Chinaemelum Okongwu, and Olamide Tolulope Sodamade. 2024. A review of health misinformation on digital platforms: Challenges and countermeasures. *International Journal of Applied Research in Social Sciences*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Vasiliki Papanikou, Panagiotis Papadakos, Theodora Karamanidou, Thanos G. Stavropoulos, Evaggelia Pitoura, and Panayiotis Tsaparas. 2024. Health misinformation in social networks: A survey of it approaches. *Preprint*, arXiv:2410.18670.

Hyun Jun Park. 2024. The rise of generative artificial intelligence and the threat of fake news and disinformation online: Perspectives from sexual medicine. *Investig Clin Urol*, 65(3):199–201.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`. Version 0.3.0.

Siddhant Bikram Shah, Surendrabikram Thapa, Ashish Acharya, Kritesh Rauniyar, Sweta Poudel, Sandesh Jain, Anum Masood, and Usman Naseem. 2024. Navigating the web of disinformation and misinformation: Large language models as double-edged swords. *IEEE Access*.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3623–3631, New York, NY, USA. Association for Computing Machinery.

Ivan Srba, Branislav Pecher, Matús Tomlein, Róbert Móro, Elena Stefancova, Jakub Simko, and Mária Bieliková. 2022. Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2949–2959. ACM.

Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. 2023. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv preprint arXiv:2306.08871*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Santosh Kumar Uppada, Parth Patel, and Sivaselvan B. 2023. An image and text-based multimodal model for detecting fake news in osn's. *Journal of Intelligent Information Systems*, 61(2):367–393.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social Science and Medicine*, 240:112552.

Danni Xu, Shaojing Fan, and Mohan Kankanhalli. 2023. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 9291–9298, New York, NY, USA. Association for Computing Machinery.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023a. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023b. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for COVID-19 news credibility research. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3205–3212. ACM.

Ruonan Zhu, Qing Liu, Chongru Huang, and Bingyi Kang. 2022. Z-ACM: an approximate calculation method of z-numbers for large data sets based on kernel density estimation and its application in decision-making. *Inf. Sci.*, 610:440–471.

A Remove Images

Examples of removed images during health misinformation collection, those images include blurry, logo-based, fuzzy, or meaningless images, which are often downloaded during web scraping.

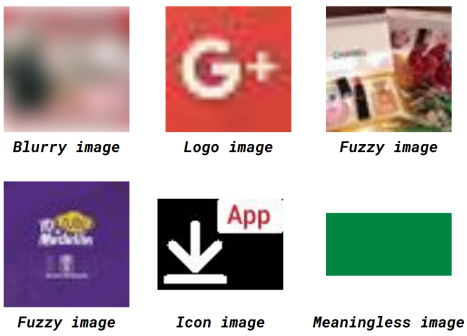


Figure 5: Sample of removed images after web scrapping. These images are irrelevant to the health topic, including blurry, logo-based, fuzzy, or meaningless images.

B Overall Statics

The overall statics of MM-Health is presented in Table 5, we utilised five different generative models for both text and image modalities to form the AI-generated portion of the dataset. The average text length and number of images are consistent across different sets.

C Post-process Algorithm

We developed Algorithm 1 to post-process data alignment between the human-generated context and all five AI-generated contexts. R_{text} and R_{image} represent the human data collected from the open-source datasets, while $N_{text}^{(k)}$ and $N_{image}^{(k)}$ are inputs derived from AI-generated text and images, where k ranges from one to five, representing the five different models. The outer loop iterates through each possible text-image pair across all models, while the inner loop checks the pair N_{text_i} and N_{image_j} against all five models. Only when text and image pairs from all five models exist will the matched data be added to M_{final} to form the final dataset.

Algorithm 1 Text and Image Alignment Across Models

Ensure: R_{text}, R_{image}
Require: $N_{text}^{(k)}, N_{image}^{(k)}, k = 1, ..., 5$
  $M_{final} \leftarrow \emptyset$
  for all N_{text_i}, N_{image_j} **do**
    $match \leftarrow$ true
    for $k \leftarrow 1$ to 5 **do**
      if $N_{text_i}^{(k)} \notin R_{text} \vee N_{image_j}^{(k)} \notin R_{image}$ **then**
        $match \leftarrow$ false
        break
      end if
    end for
    if $match =$ true **then**
      $M_{final} \leftarrow M_{final} \cup (R_{text_i}, R_{image_j}, N_{text_i}, N_{image_j})$
    end if
  end for
  return M_{final}

D Generation Prompts

The generation prompts are presented in Figure 6. Within the prompts, text highlighted in blue represents the raw input images extracted from the news article. Green text highlights indicate the original news article content, while grey text shows the intermediate output from GPT-4o, and orange text denotes the final generated output. The generation examples are shown in Figure 7, and we ensure the generations are similar to the original images but with modifications to mimic the real-world scenario where misinformation is only slightly different from the truth.

E Experiment prompts

We present the prompts used for all three tasks in the experiment, covering both zero-shot and five-shot settings. Both proprietary and open-source models follow the same prompts. Figure 8 illustrates the zero-shot and five-shot prompts for the information reliability check, Figure 9 depicts the prompts for the information originality check, and Figure 10 presents the prompts for the fine-grained AI detection analysis. Article inputs are highlighted in green, while model outputs are highlighted in orange. Since models may not always produce the exact output specified in the prompts, we run each model several times until all responses are correctly formatted to meet the output requirements.

F Heatmaps

We present the heatmaps in Figure 11 and Figure 12 for GPT4o, GPT4o-mini, Llama-3.2-Vision,

Split	Source	Reliable				Unreliable			
		Article #	Avg Len.	Image #	Avg IPA.	Article #	Avg Len.	Image #	Avg IPA.
Train	**Human**	3345	848.34	5175	1.55	809	635.77	6917	8.55
	AI	16725	463.84	25768	1.54	4045	450.56	34084	8.43
Val	**Human**	373	856.01	586	1.57	90	749.04	709	7.88
	AI	1865	464.34	2917	1.56	450	455.93	3415	7.59
Test	**Human**	932	822.82	1435	1.54	227	669.3	1572	6.93
	AI	4660	460.21	7139	1.53	1135	450.28	7695	6.78
Overall	**Human**	4650	843.84	7196	1.55	1126	651.58	9198	8.17
	AI	23250	463.15	35824	1.54	5630	450.93	45194	8.03

Table 5: Overall statistics of the MM-Health dataset, where "Avg Len" represents the average word count of the articles, and "Avg IPA" indicates the average image count per article. The statistics are provided for both human and AI data across all data splits.

Text

GPT-4o Summary
- PROMPT
Please summarize a topic based on the following image and text:
<Original article text>
<Original article images>
- OUTPUT:
[GPT-4o article summary]

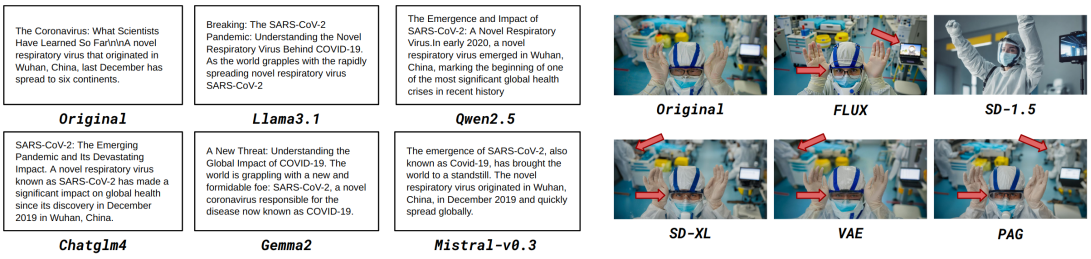
Text Generation
- PROMPT
Please write an new article to support this topic:
<GPT-4o article summary>
- OUTPUT:
[**LLM generated text**]

Image

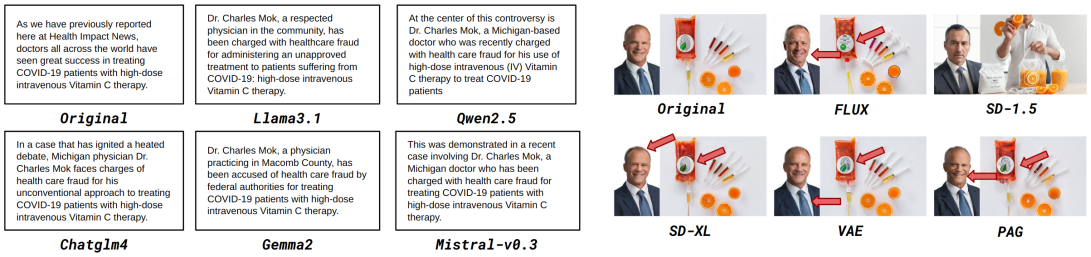
GPT-4o Caption
- PROMPT
Describe this image with a short sentence:
<Original article images>
- OUTPUT:
[GPT-4o image caption]

Image Generation
- PROMPT
<GPT-4o image caption> with photo realistic, detailed, 8k uhd, dslr, soft lighting, high quality, Fujifilm XT3 quality.
- IMAGE
<Original article images>
- OUTPUT:
[**Model generated image**]

Figure 6: AI generation prompts. Text generation prompts are shown on the left side of the figure. Image generation prompts are shown on the right side of the figure.



(a) Reliable news generated from both human and AI.



(b) Unreliable news generated from both human and AI.

Figure 7: Examples of human and AI generated (a:) reliable and (b:) unreliable multimodal news. The examples at the top represent reliable news, while those at the bottom represent unreliable news. Note that the articles have been rewritten on the same topic and the image details have been manipulated, as pointed by the ⇒.

Qwen2-VL and Llava-1.6 in zero-shot setting for task 3 - fine-grained AI detection analysis. All the models have been run through the exact zero-shot settings and prompts as demonstrated in Figure 10.

0-shot

Text Only

```
- PROMPT
ARTICLE CONTENT:
<article_content>
Please analyze the provided article to determine
whether the content of the article is real
information or misinformation.
article_type: "real_information" or "misinformation"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[real_information/misinformation]
```

Text and Image

```
- PROMPT
ARTICLE CONTENT:
<article_content with images>
Please analyze the provided article and the images
within the article to determine whether the content
of the article and images is real information or
misinformation.
article_type: "real_information" or "misinformation"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[real_information/misinformation]
```

5-shot

Text Only

```
- PROMPT
Please analyze the provided article to determine
whether the content of the article is real
information or misinformation.
article_type: "real_information" or "misinformation"

Here are some examples:
Example1:
<article_content>
<RETURN: misinformation>
[example 2 - 4]
Example5:
<article_content>
<RETURN: real_information>
Now Answer:
ARTICLE CONTENT:
<article_content>
article_type: "real_information" or "misinformation"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[real_information/misinformation]
```

5-shot

Text and Image

```
- PROMPT
Please analyze the provided article and the
images within the article to whether the
content of the article is real information or
misinformation. article_type:
"real_information" or "misinformation"

Here are some examples:
Example1:
<article_content with images>
<RETURN: misinformation>
[example 2 - 4]
Example5:
<article_content with images>
<RETURN: real_information>
Now Answer:
ARTICLE CONTENT:
<article_content with images>
article_type: "real_information" or
"misinformation"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[real_information/misinformation]
```

Figure 8: Zero-shot and five-shot prompts for task 1 - information reliability check.

0-shot

Text Only

```
- PROMPT
ARTICLE CONTENT:
<article_content>
Please analyze the provided article to determine
whether it is created by human, generated by AI, or
a mixture of both.
made_by: "ai" or "human" or "mixed"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[ai/human/mixed]
```

Text and Image

```
- PROMPT
ARTICLE CONTENT:
<article_content with images>
Please analyze the provided article and the images
within the article to determine whether the article
and images are created by humans, generated by AI,
or a mixture of both.
made_by: "ai" or "human" or "mixed"
- OUTPUT:
[ai/human/mixed]
```

5-shot

Text Only

```
- PROMPT
Please analyze the provided article to
determine whether it is created by human,
generated by AI, or a mixture of both.
made_by: "ai" or "human" or "mixed"

Here are some examples:
Example1:
<article_content>
<RETURN: misinformation>
[example 2 - 4]
Example5:
<article_content>
<RETURN: real_information>
Now Answer:
ARTICLE CONTENT:
<article_content>
made_by: "ai" or "human" or "mixed"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[ai/human/mixed]
```

5-shot

Text and Image

```
- PROMPT
Please analyze the provided article and the
images within the article to determine whether
the article and images are created by humans,
generated by AI, or a mixture of both.
made_by: "ai" or "human" or "mixed"

Here are some examples:
Example1:
<article_content>
<RETURN: misinformation>
[example 2 - 4]
Example5:
<article_content>
<RETURN: real_information>
Now Answer:
ARTICLE CONTENT:
<article_content>
made_by: "ai" or "human" or "mixed"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[ai/human/mixed]
```

Figure 9: Zero-shot and five-shot prompts for task 2 - information originality check.

0-shot

Text and Image

```
- PROMPT
ARTICLE CONTENT:
<article_content with images>
Please analyze the provided article and the
images within the article to determine whether
the article and images are created by humans,
generated by AI, or a mixture of both.
made_by: "ai" or "human" or "mixed"
- OUTPUT:
[ai/human/mixed]
```

5-shot

Text and Image

```
- PROMPT
Please analyze the provided article and the
images within the article to determine whether
the article and images are created by humans,
generated by AI, or a mixture of both.
made_by: "ai" or "human" or "mixed"

Here are some examples:
Example1:
<article_content>
<RETURN: misinformation>
[example 2 - 4]
Example5:
<article_content>
<RETURN: real_information>
Now Answer:
ARTICLE CONTENT:
<article_content>
made_by: "ai" or "human" or "mixed"
DO NOT ADD ANYTHING ELSE.
- OUTPUT:
[ai/human/mixed]
```

Figure 10: Zero-shot and five-shot prompts for task 3 - fine-grained AI detection analysis.

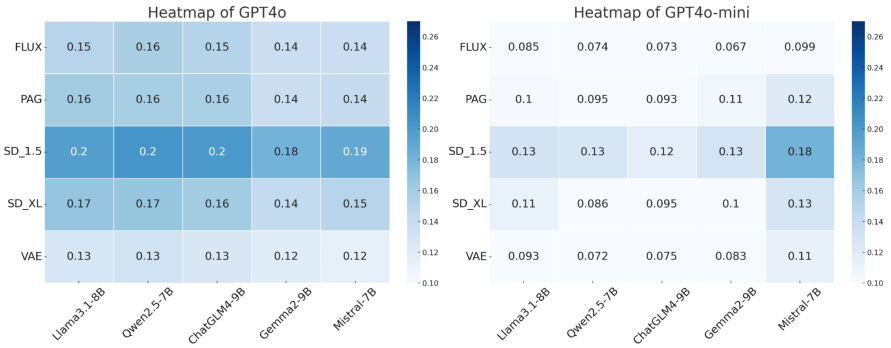


Figure 11: Zero-shot GPT4o and GPT4o-mini heatmaps for task 3 - fine-grained AI detection analysis.

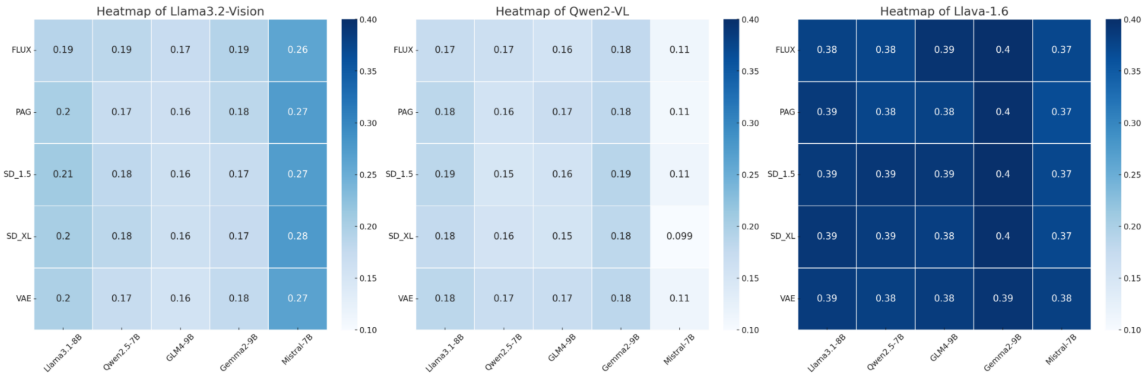


Figure 12: Zero-shot Llama-3.2-Vision, Qwen2-VL and Llava-1.6 heatmaps for task 3 - fine-grained AI detection analysis.