HINTSOFTRUTH: A Multimodal Checkworthiness Detection Dataset with Real and Synthetic Claims

Michiel van der Meer[1,2], **Pavel Korshunov**[2], **Sébastien Marcel**[2], **Lonneke van der Plas**[3,2]

[2] Idiap Research Institute, Martigny, Switzerland,
[1] Leiden University, Leiden, The Netherlands,
[3] USI Università della Svizzera italiana, Lugano, Switzerland

Abstract

Misinformation can be countered with fact-checking, but the process is costly and slow. Identifying checkworthy claims is the first step, where automation can help scale fact-checkers' efforts. However, detection methods struggle with content that is (1) multimodal, (2) from diverse domains, and (3) synthetic. We introduce HINTSOFTRUTH, a public dataset for multimodal checkworthiness detection with 27K real-world and synthetic image/claim pairs. The mix of real and synthetic data makes this dataset unique and ideal for benchmarking detection methods. We compare fine-tuned and prompted Large Language Models (LLMs). We find that well-configured lightweight text-based encoders perform comparably to multimodal models but the former only focus on identifying non-claim-like content. Multimodal LLMs can be more accurate but come at a significant computational cost, making them impractical for large-scale applications. When faced with synthetic data, multimodal models perform more robustly.

https://hintsoftruth.github.io/

1 Introduction

Online misinformation spreads rapidly via social networks and deceptive websites posing as legitimate news sources (Del Vicario et al., 2016; Rocha et al., 2021; Ecker et al., 2024). This influences voting behavior (Ribeiro et al., 2017) and pollutes the digital information space (Greenspan and Loftus, 2021; Sharma et al., 2019). Misinformation tactics include decontextualization (e.g., wrongly presenting image-based evidence) and providing incomplete information (Kreps et al., 2022). Generative AI, like ChatGPT (OpenAI, 2023) for text and Midjourney (Midjourney, Inc., 2023) for images, has worsened the issue by enabling large-scale alteration or fabrication of news narratives (Zhou et al., 2023; Chen and Shu, 2024). Given these developments, continuous verification of multimodal information is a key challenge (Abdelnabi et al., 2022; Singh and Sharma, 2022).

Figure 1: Examples of the multimodal checkworthiness detection task.

Media gatekeepers, including news publishers and fact-checking services, verify content veracity, but manual fact-checking is costly and time-consuming (Nakov et al., 2021). Therefore, selecting which claims to fact-check is a major challenge, as the amount of potential misinformation far exceeds fact-checking capacity. Automated approaches can help by identifying **checkworthy** claims (Nakov et al., 2018; Konstantinovskiy et al., 2021), see Figure 1 for an example, or in other stages in the fact-checking pipeline (Figure 2).

However, existing automated checkworthiness detection methods (1) have poor support for multimodal content, (2) have only been tested in a limited number of domains, (3) have unknown capabilities on synthetic media, and (4) do not consider compute cost as a factor. (Akhtar et al., 2023). First, modern misinformation often includes mixed forms of media, such as images or videos (Dufour et al., 2024), yet it is unclear if detection methods effectively integrate visual data (Alam et al.,

2023). Second, strategies for misinformation detection vary by domain (Ecker et al., 2022; Chen et al., 2021; Lasser et al., 2023), raising concerns about generalizability, especially, for practical applications (Jiang and Wilson, 2018; Monteith et al., 2024). Third, ubiquitous access to generative models is reshaping misinformation (Xu et al., 2023), warranting the evaluation of detection methods on synthetic content. Lastly, while Large Language Models (LLMs) perform well, their high compute cost may render large-scale checkworthiness detection impractical (Augenstein et al., 2024), though the exact tradeoffs are unknown.

This paper introduces HINTSOFTRUTH, the first publicly available multimodal dataset of image-text pairs containing both real-world and synthetically generated checkworthy and non-checkworthy claims. We source real claims from datasets like 5Pils (Tonglet et al., 2024), Multiclaim (Pikuliak et al., 2023), Flickr30K (Hodosh et al., 2013), and SentiCap (Sharma et al., 2018). Synthetic images and text are generated using Flux (Black Forest Labs, 2024), StableDiffusion 3.5 (Stability.ai, 2024, SD), Llava (Li et al., 2024), and BLIP (Li et al., 2022). We evaluate recent text and image models, from lightweight ones like TinyBERT (Jiao et al., 2020) for scalability to large, multimodal models like Pixtral (Mistral, 2024). These evaluations reveal model limitations and guide practical decisions in checkworthiness detection.

Contributions We present: (1) HINTSOFTRUTH, a novel dataset for multimodal checkworthiness detection from diverse sources, with an established connection between images and textual claims, that can be used as a benchmark for checkworthiness detection models, (2) synthetic counterparts of images and claims in the dataset, which has not been explored in the context of checkworthiness, and (3) an extensive set of experiments demonstrating the limits of state-of-the-art detection methods.

2 Related Work

2.1 Human-Centered Fact-Checking

Recent research on human-centered AI has emphasized developing tools that augment humans (Akata et al., 2020; Nakov et al., 2021). In the field of fact-checking, such tools would complement human fact-checkers in their work (Micallef et al., 2022; Graves, 2017), allowing experts to control *what* and *how* to fact-check (Das et al., 2023).

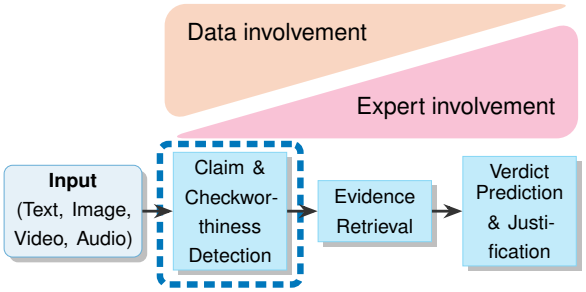

Figure 2: The fact-checking pipeline from Akhtar et al. (2023), visualized the amount of data and expert effort required. We focus on the highlighted stage.

Crucially, as shown in Figure 2, the fact-checking pipeline involves handling large amounts of data and needs expert involvement. In the early stages of the pipeline, large amounts of data are processed to filter out **checkworthy** content. Hence, relying on experts to do this task manually is infeasible. Moreover, the expert is required in the later stages for the complex tasks of verdict prediction and justification. Natural Language Processing (NLP) technology provides various types of support, especially when dealing with scale (van der Meer, 2024; Procter et al., 2023), to simplify the problem (Chen et al., 2022; Bonet-Jover et al., 2024), or to combat cognitive biases (Soprano et al., 2024). In this work, we use NLP techniques to address the scale issues for checkworthiness detection.

2.2 Misinformation in the Age of LLMs

LLMs play a significant role both in detecting and generating misinformation. Recent work integrates LLMs into fact-checking frameworks (Geng et al., 2024), although methods are shown to generalize poorly across time (Stepanova and Ross, 2023). Nonetheless, LLMs look promising when applied to text-based checkworthiness detection (Majer and Snajder, 2024). Reviews of LLM-generated multimedia highlight the open challenges (Lin et al., 2024; Augenstein et al., 2024). For instance, large amounts of synthetic misinformation have the potential to impact the quality of future LLMs (Pan et al., 2023), and misinformation generated by GPT-4 may be harder to detect than that written by humans (Chen and Shu, 2024).

2.3 Multimodal Resources

Existing work on fact-checking emphasizes empirical research, which involves extensively benchmarking fact-checking methods (Schlichtkrull

et al., 2023; Papadopoulos et al., 2024), often using distant supervision (Nakamura et al., 2020; Zlatkova et al., 2019). Most multimodal datasets investigate the out-of-context use of images and claims (Luo et al., 2021; Tonglet et al., 2024), or whether claims are reflected in an image (Yoon et al., 2024; Papadopoulos et al., 2023). Few datasets exist that (1) check whether the image contributes new information (Liu et al., 2024), or (2) contain synthetically generated data (Xu et al., 2023; Seow et al., 2022). The few efforts on multimodal checkworthiness indicate that textual data, whether through OCR or by focusing on claims only, is sufficient for state-of-the-art performance (Vogel and Frick, 2023). More extensive experiments on varied types of data with complex image use, across domains, are needed to further examine this finding.

3 Method

We introduce the multimodal checkworthiness task definition, how we obtain the real-world data underlying HINTSOFTRUTH, and how we generate synthetic samples to augment our dataset.

3.1 Task Definition: Multimodal Checkworthiness Detection

Given a textual claim c and an image i published alongside the claim, our task aims to predict whether the pair is worthy of fact-checking $p(i, c) = 1$. In checkworthiness detection, the following questions are answered: (**Q1**) Does the text contain a verifiable factual claim? (**Q2**) Is the claim potentially harmful, urgent, and up-to-date? The task definition is derived from Barrón-Cedeno et al. (2020), which also formed the basis for the canonical dataset for multimodal checkworthiness detection, CheckThat! 2023 Task 1A (Alam et al., 2023; Cheema et al., 2022).

To establish that an image provides meaningful context to a claim and is necessary for assessing the pair's checkworthiness, we also consider the following contextualized questions: (**Q3**) Is the content of the claim reflected in the image? (**Q4**) Does the image contribute extra information to the claim? These two questions help identify *complex* image use, which will test the multimodal capabilities of checkworthiness detectors (Dufour et al., 2024).

3.2 Getting Checkworthy Image/Claim Pairs

We set out to obtain image/claims pairs that we deem checkworthy. We rely on data stemming from fact-checking articles, as claims in these articles have already been checked. Fact-checking articles are written by experienced fact-checkers and contain rich contextual information. In practice, claims are often sourced from social media platforms. We obtain our data from two sources:

5Pils (Tonglet et al., 2024). 5Pils contains extracted images, claims, and contextual questions about claims from news sources in India, Kenya, and South Sudan. Through the use of contextual questions, images in this dataset are ensured to have a relationship with the claim.

Multiclaim (Pikuliak et al., 2023) contains URLs to a wide array of fact-checking articles and their respective claims but needs to be scraped and filtered for images. We retain those claims for which (1) we find images in close proximity, and (2) explicitly refer to visual information. See Appendix A.1 for additional details.

3.3 Non-checkworthy Image/Claim Pairs

We also need image/text pairs that are not checkworthy. We resort to strategies derived from the task definition for obtaining negative instances. We select samples from datasets that we consider not checkworthy because they answer 'no' to any of the guiding questions posed in Section 3.1. The strategies select:

Non-factual (Q1) claims, such as subjective opinions or facts that cannot be verified using external information. The dataset representing this strategy is **SentiCap** (Sharma et al., 2018).

Non-relevant (Q2) statements that are not harmful, not about breaking news, not up-to-date, or not relevant to news topics. The dataset representing this strategy is **Flickr30K** (Hodosh et al., 2013), though there are many other resources containing arbitrary image-text pairs (see Appendix C.3).

No cross-modal connection (Q3) images we know have a deep connection with a text but with the image swapped to no longer make sense. The dataset representing this strategy is **Fakeddit** (Nakamura et al., 2020).

To incorporate the fourth guiding question (**Q4**), we filter out claims from any of the stated datasets that do not explicitly refer to multimodal content (see Appendix A.2 for a list of terms). This way, we encourage that the samples with basic image use (i.e., those pairs where the claim does not refer to the image) are excluded from our dataset. We combine the checkworthy and non-checkworthy samples into HINTSOFTRUTH, our novel multi-

modal checkworthiness dataset that spans multiple domains.

3.4 Generating Synthetic Samples

Given the risks of synthetic misinformation (Dufour et al., 2024; Papadopoulos et al., 2023; Zhou et al., 2023), we augment our dataset with additional samples that contain either claims or images generated using various publicly accessible models. New samples consist of the original text (or image) and the corresponding generated image (or text). Specifically, we employ two image generators to create images from claims and two multimodal models to generate claims from images. Our approach follows a simple cross-modal generation method: models freely generate corresponding text or images without requiring adversarial prompts (Perez and Ribeiro, 2022). This allows us to examine how checkworthiness detection models respond to synthetic data. The labels of the new samples depend solely on the claim: synthetic **claims** are deemed *non-checkworthy*, as models primarily generate non-relevant captions (see **Q2** in Section 3.1), while synthetic **images** retain their original label, ensuring consistency with the claim's content.

4 Experiments

4.1 Data

See Table 1 for the datasets used in this paper. We use the canonical CheckThat! 2023 Task 1A dataset ('CheckThat' henceforth, Alam et al., 2023) as training dataset and reference benchmark using its predefined train/test split, which represents the in-distribution scenario (models are trained and tested on the same dataset). CheckThat has a label ratio of .66/.34 between non-checkworthy and checkworthy samples. The HINTSOFTRUTH dataset is split into two equally-sized development sets, and a smaller test set (40%, 40%, and 20%, respectively). We use the test set (label ratio of .62/.38) for testing the detection methods fine-tuned on CheckThat. This represents a cross-distribution scenario, which should be more challenging than an in-distribution setup. HINTSOFTRUTH is available online[1].

4.2 Multimodal Checkworthiness Methods

We experiment with various state-of-the-art text-based, image-based, and multimodal encoders for checkworthiness detection, see Table 2. We use different model sizes to investigate the tradeoff between compute cost and task performance. We include single-modality models to identify whether both modalities are needed (i.e., checkworthiness can be assessed without leveraging cross-modal information). In addition, we distinguish between encoder-only and decoder-only models, to determine the difference between fine-tuning models on multimodal checkworthiness and In-Context Learning (Dong et al., 2024). Below, we describe the experimental setup for each type of approach. Additional information is available in Appendix B.

Fine-Tuning (FT) To fine-tune models for multimodal checkworthiness detection, we update all model parameters θ when predicting $p_\theta(i, c)$. We instantiate the models using pretrained versions, adding a single linear classification layer with a two-node output over their embeddings.[2] We fine-tune our models with data from CheckThat using its predefined train/val/test split. Additionally, we tune a threshold parameter on the positive class probability, similar to a single-neuron sigmoid output (Zou et al., 2016; Korshunov and Marcel, 2019). At various thresholds, we compute the True Positive Rate (TPR) and False Positive Rate (FPR), and we select the threshold for an FPR of 0.3, prioritizing recall over precision (see App. C.1).

In-Context Learning (ICL) We evaluate the impact of n-shot learning (with $n = \{0, 1, 2, 5\}$) and prompt verbosity. The verbose prompt instructions include guiding questions **Q1** through **Q4** (see Section 3.1), while the succinct prompt only asks for an overall checkworthiness label. We experiment with two models: (1) **Llava** (Liu et al., 2023), using a Mistral-7B backend with a 32K token context. (2) **Pixtral** (Mistral, 2024), with a context size of 1024K. Both models are chosen for their compatibility with standard hardware (up to a single H100 with 80GB VRAM) and accessibility, excluding non-local proprietary LLMs. Our experiments focus on zero-shot ICL with concise instructions to minimize token usage, though multiple setups are explored in Section 5.1.

4.3 Research Questions

Based on the criteria discussed in Section 1, we conduct four experiments to address: (RQ1) Does

[1] https://huggingface.co/datasets/michiel/hints_of_truth

[2] Single-logit was less stable and yielded inferior performance, see App. C.2.

Dataset	Source / Subset	Checkworthy	Size	Description
CheckThat 2023 Task 1A	Twitter	Mixed	3,175	Tweets on COVID-19, technology, climate change.
	5Pils	✓	1,676	News articles from India, Kenya, and South Sudan.
	Multiclaim	✓	3,048	Social media posts in a general domain.
	SentiCap	✗	3,171	Captions with sentiment injection.
	Flickr30K	✗	3,000	Image captions from a general domain.
	Fakeddit	✗	1,382	Reddit posts from a general domain.
HINTSOFTRUTH	Mixed	Mixed	12,277	Mixed domain benchmark
HINTSOFTRUTH-aug	5Pils	Mixed	1,676	Generated claims using BLIP, Llava. Generated images using Flux, StableDiffusion 3.5.
	Flickr30K	✗	3,000	Generated captions using BLIP, Llava. Generated images using Flux, StableDiffusion 3.5.

Table 1: Overview of the datasets used in this study for multimodal checkworthiness. HINTSOFTRUTH aggregates samples from five sources, and HINTSOFTRUTH-aug contains synthetically generated variants.

Model	Modality	Size	App.
TinyBERT	text	14M	FT
BERT-base	text	109M	FT
BERT-large	text	335M	FT
ResNet-26	image	16M	FT
ViT-base	image	86M	FT
ViT-large	image	303M	FT
BLIP	text, image	385M	FT
BLIP2	text, image	1.17B	FT
Llava	text, image	7.57B	ICL
Pixtral	text, image	12.4B	ICL

Table 2: Models used for checkworthiness detection. Depending on the model, we fine-tune them (FT) or perform In-Context Learning (ICL).

Model	Prec.	Rec.	F1	Acc.
TinyBERT	0.698	0.721	0.702	0.724
BERT-base	<u>0.735</u>	0.769	0.735	0.748
BERT-large	0.726	0.760	0.723	0.735
ResNet	0.595	0.600	0.596	0.641
ViT-base	0.639	0.655	0.640	0.666
ViT-large	0.654	0.670	0.658	0.686
BLIP	**0.782**	<u>0.819</u>	**0.788**	**0.801**
BLIP2	**0.782**	**0.822**	<u>0.786</u>	<u>0.797</u>
Llava (0-shot)	0.565	0.574	0.554	0.572
Pixtral (0-shot)	0.673	0.675	0.588	0.588

Table 3: (Macro-averaged) performance for the CheckThat! 2023 Task 1A benchmark. Best scores per sub-dataset are shown in **bold**, with the second-best <u>underlined</u>.

combining modalities influence checkworthiness detection performance? (RQ2) How well do models generalize across domains? (RQ3) How do models fare on synthetic data? (RQ4) What is the tradeoff between compute cost and task performance?

5 Results

Table 3 shows the in-distribution experiments results on CheckThat, and Table 4 demonstrates the results on the non-synthetic, real part of HINTSOFTRUTH, illustrating the cross-distribution experiments. We answer each research question individually step by step.

5.1 Cross-modality Performance

On the CheckThat dataset, the strongest models are multimodal BLIP and BLIP2, which form the upper bound (see Table 3). Interestingly, the accuracies of text-only encoders are close to those of BLIP and BLIP2 (up to 94% relative to their accuracy), suggesting that little visual information is required for accurate checkworthiness detection, in line with results found in Vogel and Frick (2023). Image-only encoders also achieve only 14% lower accuracy than the upper bound. The narrow gap between single and multimodal models shows the dataset's limited suitability for assessing multimodal capabilities. ICL-based methods perform surprisingly poorly, with considerable false positive rates of 30% and 39% for Llava and Pixtral.

For the part of HINTSOFTRUTH containing real

Model	5Pils Acc.	Multiclaim Acc.	Flickr30K Acc.	SentiCap Acc.	Fakeddit Acc.	Overall P.	Overall R.	Overall F1	Overall Acc.
TinyBERT	0.898	0.878	0.794	0.721	0.480	0.779	0.796	0.772	0.775
BERT-base	0.682	0.611	0.573	0.816	0.803	0.669	0.675	0.671	0.685
BERT-large	0.645	0.607	0.490	0.854	0.825	0.655	0.661	0.657	0.671
ResNet	0.472	0.360	0.715	0.684	0.632	0.545	0.543	0.543	0.578
ViT-base	0.321	0.353	0.710	0.689	0.766	0.529	0.526	0.525	0.571
ViT-large	0.343	0.342	0.692	0.661	0.784	0.520	0.519	0.517	0.561
BLIP	0.769	0.540	0.934	0.600	0.416	0.656	0.661	0.657	0.671
BLIP2	0.880	0.795	0.716	0.689	0.535	0.734	0.749	0.727	0.730
Llava (0-shot)	0.225	0.266	0.053	0.449	**0.996**	0.328	0.318	0.320	0.334
Pixtral (0-shot)	**0.954**	**0.919**	**0.937**	**0.962**	0.716	**0.909**	**0.921**	**0.914**	**0.918**

Table 4: Test set of HINTSOFTRUTH, performance per subset. The last four columns show the overall macro-averaged scores. Best scores per sub-dataset are shown in **bold**, with the second-best underlined.

data, Pixtral forms the upper performance bound (see Table 4). The gap between text-only models and the upper bound is larger than in CheckThat (-14% vs. -7%). Surprisingly, TinyBERT outperforms larger text-only models and is second to only Pixtral. This suggests that a small model with a well-tuned classification threshold can be effective, which poses an interesting venue for smaller organizations with limited compute capacity. Image-only encoders perform worse (-35% vs. upper bound) than other models. Among multimodal models, BLIP2 excels, followed by Llava, aligning performance with parameter count (i.e., the bigger the model, the better its performance).

We investigate the impact of n-shot learning on LLava and Pixtral by varying the prompting setup for these ICL models in Figure 3. Performance on HINTSOFTRUTH reveals that Pixtral and Llava have contrary behavior with an increase in context: (1) Adding more examples with few-shot learning aids Llava but hurts the Pixtral model, and (2) Llava can benefit from a long prompt in zero-shot cases, but Pixtral generally benefits from short prompts. This is a surprising finding as additional examples should inform a model better. Like before, we observe an oversensitivity to predicting a *checkworthy* label. The wide context of Pixtral may have it confuse which image/claim pair is currently under scrutiny. While CheckThat shows this behavior partially, HINTSOFTRUTH provides a clearer pattern, attesting to the usefulness of our dataset.

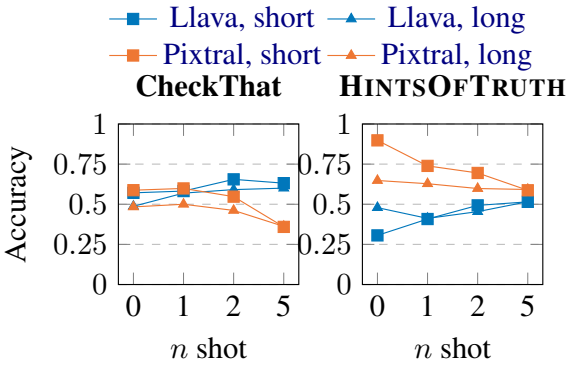

Figure 3: Few-shot performance with ICL on CheckThat (left) and HINTSOFTRUTH (right).

5.2 Domain Generalization

Evaluating performance on each of the HINTSOFTRUTH subsets shows (see Table 3) that while fine-tuned (FT) models are trained on only the three domains using CheckThat, they consistently generalize to the subsets of HINTSOFTRUTH, which suggests an effective knowledge transfer. However, performance varies based on experiment characteristics such as modalities used, pretraining setup, and model size.

Among FT models, TinyBERT is robust across most datasets but struggles on Fakeddit, likely due to the linguistic differences between CheckThat and Fakeddit; Text data in the latter stems from user-submitted post titles, which are less grammatically correct.[3] Larger BERT models perform well on Fakeddit but worse on SentiCap and Multiclaim. Since TinyBERT is distilled from these models, constraining model size may enhance generaliza-

[3]Example: *"took this photo of my dog rolling in some grass"* for Fakeddit vs. *"a photograph shows rays of lights in the shape of a cross during the august 2017 eclipse."* for Multiclaim.

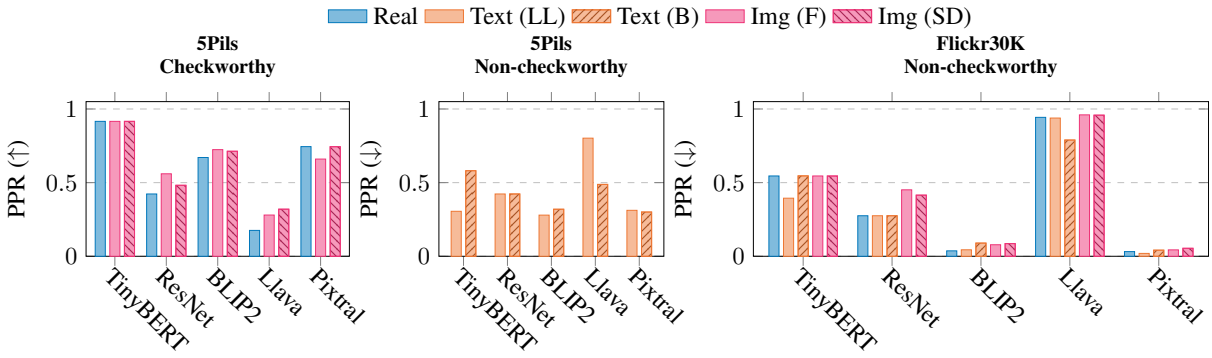

Figure 4: Positive prediction rate (PPR) for each model on the two augmented subsets in HINTSOFTRUTH compared to real-world data, with 5Pils plot split into checkworthy and non-checkworthy samples. ↑ and ↓ denote higher and lower as better, respectively. **Text (LL)**: Llava-generated caption, **Text (B)**: BLIP-generated caption, **Img (F)**: Flux-generated image, **Img (SD)**: StableDiffusion-generated image.

tion but influence error modes.

ICL performance also varies: Llava achieves the highest accuracy on Fakeddit, while Pixtral excels on all other subsets. Llava's training on noisy user-generated ShareGPT4V data (Chen et al., 2024) may explain its behavior, while Pixtral may favor syntactically correct texts. This difference between the two models highlights noisy data as a unique generalization challenge. Finally, BLIP excels on Flickr30K, despite not being finetuned on it, raising data leakage concerns (Balloccu et al., 2024).

5.3 Performance on Synthetic Data

We investigate prediction behavior on the synthetic part of HINTSOFTRUTH, using images generated by Flux (Black Forest Labs, 2024) and Stable Diffusion 3.5 (Stability.ai, 2024), and textual claims by Llava (Li et al., 2024) and BLIP (Li et al., 2022). To the human eye, synthetic samples appear distinct from real-world samples (see Figure 5 for some examples). Our goal is to determine whether models can reliably detect synthetic data and differentiate between various generative methods. To achieve this, we cross-check with the same models used for classification to assess whether they can identify their own synthetic generations. We evaluate a subset of models, including the smallest (TinyBERT, ResNet) and largest (BLIP2, Llava, Pixtral), to analyze compute/accuracy tradeoffs. Figure 4 provides an overview of the positive checkworthiness prediction rate (PPR) per model to reveal how frequently each model classifies an image/claim pair as checkworthy, shedding light on both model biases toward synthetic modalities and their potential failure modes.

Results TinyBERT is accurate on real 5Pils data but struggles with synthetic text. For example, it misclassifies over half of BLIP-generated texts as checkworthy. It also has a high false positive rate ($\sim$ 0.55) on augmentations of Flickr30K, which could lead to an unnecessarily high workload for fact-checkers. ResNet, on the other hand, often misses checkworthy samples for 5Pils (high false negative rate). It's higher PPR for synthetic images– by 32% for Flux and by 14% for SD on 5Pils and 64% Flux and 51% for SD on Flickr30K shows that ResNet may be overly sensitive to detecting synthetic images as checkworthy, even if the synthetic images are not relevant (see Section 3.4).

Llava generates many false negatives on 5Pils while obtaining a high false positive rate on Flickr30K, suggesting that the model may misunderstand the task instructions. The high PPR on Llava-generated texts for 5Pils reveals an oversensitivity to synthetic texts generated by itself. BLIP2 behaved more in line with expectations, with a lower PPR for synthetic texts while sustaining a high PPR for synthetic images in 5Pils. On Flickr30K, it maintained a minimal PPR across all synthetic data, likely benefiting from pretraining on synthetic captions. The origin of the synthetic text (BLIP vs. Llava) had little impact on its performance. Pixtral mirrors BLIP2's results, except that pairs from 5Pils with images generated by Flux were 10% less often identified as checkworthy by Pixtral, suggesting that as newer, higher-quality image generators emerge, Pixtral's accuracy might decline.

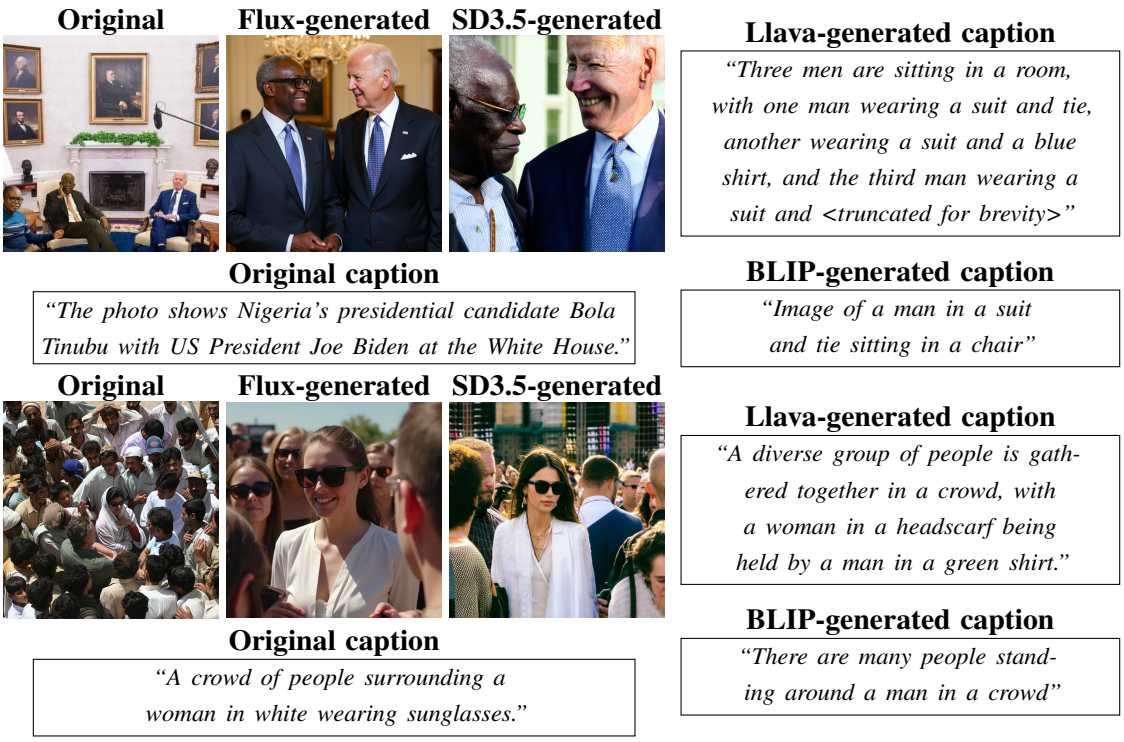

Figure 5: Examples of synthetically generated images and captions. The upper row shows a checkworthy example from 5Pils. The bottom row shows a non-checkworthy example from Flickr30K.

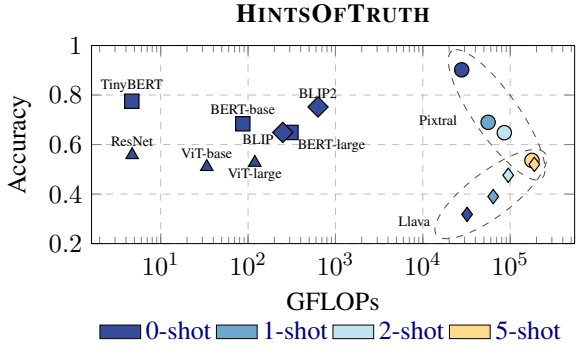

Figure 6: Compute budget versus task performance.

5.4 Compute Budget

Unsurprisingly, models with more parameters generally perform better at checkworthiness detection. However, running large models like Pixtral demands substantial compute resources. Since checkworthiness detection serves as a prefiltering task, such resources may not be available to media organizations or outpaced by new content. To explore the trade-off between model size and performance, we visualize the compute budget in FLOPs (Hassid et al., 2024) compared to final accuracy in Figure 6. FLOPs usage and wall time are estimated using the `calflops` library (Ye, 2023), averaging over 100 random samples from HINTSOFTRUTH, measured on a node with a single H100 GPU.

Results The best-performing model, Pixtral, requires at least two orders of magnitude more compute than FT models, even with zero-shot ICL. BLIP2 offers a balanced trade-off, ranking third in accuracy at a reasonable compute cost. However, in wall time, it closely matches ICL models—on average, BLIP2 runs as long as 1-shot Pixtral, while Pixtral 0-shot is up to 36% faster per sample (see App C.4 for details). TinyBERT emerges as the most balanced, delivering competitive accuracy at significantly lower cost and runtime (four orders of magnitude in FLOPS, two in wall time). This suggests that tuning a small model can achieve strong performance, raising questions about the role of visual information in checkworthiness detection.

6 Conclusions

HINTSOFTRUTH provides key insights into the challenges and opportunities in multimodal checkworthiness detection and the questionable role that visual content plays in misinformation. Our findings indicate that while multimodal models outperform image-only approaches, their advantage over text-only models is not attested. Well-tuned text-based models achieve nearly the same accuracy (up to 86%), raising uncertainty about *the*

extent to which visual content contributes to the checkworthiness of real-world image/claim pairs. Unlike many other areas of NLP, our experiments reveal that the *syntactic* and *grammatical structure* of the claims, rather than their domain, impacts generalization. Larger models, like Pixtral, demonstrate high adaptability but may unexpectedly fail to transfer. When confronted with synthetic data, lightweight models become *oversensitive*, often misclassifying images as checkworthy. This increases fact-checkers' workload by requiring manual filtering of false positives. Fine-tuning models on synthetic samples could help, but risks turning into an adversarial race with evolving image generators (Corvi et al., 2023). Our analysis of the computational trade-offs reveals that large models come at compute costs of *four orders of magnitude larger* than smaller models like TinyBERT. Though small models require careful tuning to be conservative, their lightweight nature makes them practically be better suited as checkworthiness detection methods.

Future work should shift checkworthiness detection to a ranking-based approach, helping fact-checkers prioritize claims. Explain why a claim needs verification can further help fact-checkers communicate decisions (McCright and Dunlap, 2017), for instance, by collaboratively uncovering the arguments supporting a fact-check decision (van der Meer et al., 2024b). Techniques like Learning to Defer (Madras et al., 2018; Khurana et al., 2024) and Active Learning (van der Meer et al., 2024a) assist in efficient data collection.

Limitations

Several limitations have an impact on the findings of our work. First, our method of checking for complex image use by retaining only claims that mention multimodal content, to answer **Q4** (Does the image contribute extra information to the claim?), can be strengthened. Specifically, for the checkworthy subsets (5Pils and Multiclaim), we assume that, since the claims and images were included in a fact-checking article, the image provides additional context as part of the fact-check. However, further (manual) verification would be necessary to test this assumption empirically. Similarly, we assume that the image generator will generate appropriate context when considering the prompt for the augmented versions of these subsets. Again, this is a strong assumption, as generated images can be said to provide (potentially irrelevant) context. Nonetheless, we believe the augmentations to be checkworthy: models can learn from visual artifacts or types of generated contexts.

Second, our study is conducted entirely on English data, whereas misinformation has impacts across many different languages and cultures. However, some of the resources used in our work could be exploited to generate instances in other languages.

Third, we do not incorporate retrieval-augmented generation (RAG) systems in our experiments. While such systems could potentially enhance checkworthiness detection by retrieving relevant fact-checks (Singal et al., 2024), they are sensitive to temporal leakage when past fact-checks are accessible (Glockner et al., 2022), skewing the results, and require even further resources than the models in this paper.

Finally, we do not conduct a human evaluation of checkworthiness predictions. While crowd annotators are often employed for such tasks, their ability to accurately judge checkworthiness remains uncertain. Fact-checking services often employ expert journalists who draw on their intuition and experience to decide what to fact-check and may take up to a couple of days to write fact-checking articles. Whether lay crowd annotators can reliably annotate checkworthiness in an online annotation study is therefore unclear. Parallel crowd and expert evaluation studies, such as expert assessments or real-world fact-checking use cases, could provide deeper insights into annotator behavior.

Ethical Considerations

The development of multimodal fact-checking datasets, like HintsOfTruth, involves several critical ethical considerations to ensure societal benefit.

Bias Mitigation Data is sourced from diverse domains, including social media (e.g., Multiclaim) and datasets that focus on underrepresented cultures (e.g., 5Pils). However, since we primarily reuse existing datasets, our corpus remains limited in size and inclusivity. As a result, geographic and cultural biases may persist.

Anonymization To protect user privacy, real-world data from social media and fact-checking articles is anonymized. Image/claim pairs are stripped of personally identifiable information (PII), and no additional contextual information is introduced.

We adhere strictly to the licensing terms of the publicly available datasets we use.

Misinformation Risks As our system contributes to the fact-checking pipeline, it is designed to help combat misinformation. However, synthetic data generation tools have the potential for misuse. To mitigate this risk, our study explicitly avoids introducing adversarial prompts that could be exploited for harmful purposes.

Resource Accessibility We prioritize lightweight models, such as TinyBERT, to enhance scalability and ensure that organizations with limited computational resources can access misinformation detection tools. Additionally, all models used in our research, including the largest ICL models, are freely available on the HuggingFace Hub.

Acknowledgements

This research was sponsored by Hasler Foundation's FactCheck project. This work was performed using the compute resources from the Idiap Research Institute and the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. We would also like to thank the ARR reviewers for their helpful feedback.

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 (CVPR)*, pages 14940–14949.

Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen V. Hindriks, Holger H. Hoos, Hayley Hung, Catholijn M. Jonker, Christof Monz, Mark A. Neerincx, Frans A. Oliehoek, Henry Prakken, Stefan Schlobach, Linda C. van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8):18–28.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Firoj Alam, Alberto Barrón-Cedeño, Gullal S Cheema, Gautam Kishore Shahi, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Wajdi Zaghouani, et al. 2023. Overview of the clef-2023 checkthat! lab task 1 on check-worthiness of multimodal and multigenre content.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.*, 6(8):852–863.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 215–236. Springer.

Black Forest Labs. 2024. FLUX.1: A Text-to-Image Generation Model. https://blackforestlabs.ai. Accessed January 16, 2025.

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, Patricio Martínez-Barco, and Mario Nieto-Pérez. 2024. RUN-AS: a novel approach to annotate news reliability for disinformation detection. *Lang. Resour. Evaluation*, 58(2):609–639.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. COYO-700M: Image-text Pair Dataset. `https://github.com/kakaobrain/coyo-dataset`.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Sijing Chen, Lu Xiao, and Jin Mao. 2021. Persuasion strategies of misinformation-containing posts in the social media. *Inf. Process. Manag.*, 58(5):102665.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1107–1128. Association for Computational Linguistics.

Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. 2024. AMMeBa: A Large-scale Survey and Dataset of Media-based Misinformation In-The-wild. *CoRR*, abs/2405.11697.

Ullrich K. H. Ecker, Lena Q. Tay, Jon Roozenbeek, Sander van der Linden, John Cook, Naomi Oreskes, and Stephan Lewandowsky. 2024. Why misinformation must not be ignored. *American Psychologist*. Advance online publication.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jiahui Geng, Yova Kementchedjhieva, Preslav Nakov, and Iryna Gurevych. 2024. Multimodal large language models to support real-world fact-checking. *CoRR*, abs/2403.03627.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique*, 10(3):518–537.

Rachel Leigh Greenspan and Elizabeth F Loftus. 2021. Pandemics and infodemics: Research on the effects of misinformation on memory. *Human Behavior and Emerging Technologies*, 3(1):8–12.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Michael Hassid, Tal Remez, Jonas Gehring, Roy Schwartz, and Yossi Adi. 2024. The larger the better? improved LLM code-generation via budget reallocation. In *First Conference on Language Modeling*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Intell. Res.*, 47:853–899.

Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW):82:1–82:23.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? In *First Conference on Language Modeling*.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *DTRAP*, 2(2):14:1–14:16.

Pavel Korshunov and Sébastien Marcel. 2019. Vulnerability assessment and detection of deepfake videos. In *IAPR International Conference on Biometrics*, Idiap-RR-18-2018.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Jana Lasser, Segun T Aroyehun, Fabio Carrella, Almog Simchon, David Garcia, and Stephan Lewandowsky. 2023. From alternative conceptions of honesty to alternative facts in communications by US politicians. *Nature human behaviour*, 7(12):2140–2151.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`. Accessed January 16, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024. Detecting Multimedia Generated by Large AI Models: A Survey. *CoRR*, abs/2402.00045.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. 2024. MMFakeBench: A Mixed-source Multimodal Misinformation Detection Benchmark for LVLMs. *CoRR*, abs/2406.08772.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Madras, Toniann Pitassi, and Richard S. Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 6150–6160.

Laura Majer and Jan Snajder. 2024. Claim Check-worthiness Detection: How Well do LLMs Grasp Annotation Guidelines? *CoRR*, abs/2404.12174.

Aaron M McCright and Riley E Dunlap. 2017. Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Journal of Applied Research in Memory and Cognition*.

Nicholas Micallef, Vivienne Armacost, Nasir D. Memon, and Sameer Patil. 2022. True or False: Studying the Work Practices of Professional Fact-checkers. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW1):127:1–127:44.

Midjourney, Inc. 2023. Midjourney. Generative AI model for image generation (Version 6).

Mistral. 2024. Announcing Pixtral 12B.

Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6149–6157. European Language Resources Association.

Preslav Nakov, Alberto Barrón-Cedeno, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9*, pages 372–387. Springer.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, Giovanni Da San Martino, et al. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference onArtificial Intelligence,{IJCAI-21}*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization.

OpenAI. 2023. ChatGPT. Large language model.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On

the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 36–44, New York, NY, USA. Association for Computing Machinery.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2024. VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias. *Int. J. Multim. Inf. Retr.*, 13(1):4.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Rob Procter, Miguel Arana-Catania, Yulan He, Maria Liakata, Arkaitz Zubiaga, Elena Kochkina, and Runcong Zhao. 2023. Some Observations on Fact-checking Work with Implications for Computational Support. *CoRR*, abs/2305.02224.

Manoel Horta Ribeiro, Pedro H. Calais, Virgílio A. F. Almeida, and Wagner Meira Jr. 2017. "Everything I Disagree With is #fakenews": Correlating Political Polarization and Spread of Misinformation. *CoRR*, abs/1706.05924.

Yasmim Mendes Rocha, Gabriel Acácio De Moura, Gabriel Alves Desidério, Carlos Henrique De Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health*, pages 1–10.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

Jia Wen Seow, Mei Kuan Lim, Raphaël C.-W. Phan, and Joseph K. Liu. 2022. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3).

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed Fact Checking using RAG and Few-shot In-context Learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, Miami, Florida, USA. Association for Computational Linguistics.

Bhuvanesh Singh and Dilip Kumar Sharma. 2022. Predicting image credibility in fake news over social media using multi-modal approach. *Neural Comput. Appl.*, 34(24):21503–21517.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From Pixels to Prose: A Large Dataset of Dense Image Captions. *CoRR*, abs/2406.10328.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive Biases in Fact-checking and Their Countermeasures: A Review. *Inf. Process. Manag.*, 61(2):103672.

Stability.ai. 2024. Stable Diffusion 3.5.

Nataliya Stepanova and Björn Ross. 2023. Temporal Generalizability in Multimodal Misinformation Detection. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 76–88, Singapore. Association for Computational Linguistics.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. "image, tell me your story!" predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7845–7864, Miami, Florida, USA. Association for Computational Linguistics.

Michiel van der Meer. 2024. Facilitating Opinion Diversity through Hybrid NLP Approaches. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 272–284, Mexico City, Mexico. Association for Computational Linguistics.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024a. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Michiel van der Meer, Enrico Liscio, Catholijn Jonker, Aske Plaat, Piek Vossen, and Pradeep Murukannaiah. 2024b. A hybrid intelligence method for argument mining. *Journal of Artificial Intelligence Research*, 80:1187–1222.

Inna Vogel and Raphael Frick. 2023. Fraunhofer sit at checkthat! 2023: Mixing single-modal classifiers to estimate the check-worthiness of multi-modal tweets.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, abs/1910.03771.

Danni Xu, Shaojing Fan, and Mohan S. Kankanhalli. 2023. Combating Misinformation in the Era of Generative AI Models. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 9291–9298. ACM.

Xiaoju Ye. 2023. calflops: a FLOPs and Params calculate tool for neural networks in pytorch framework.

Yejun Yoon, Seunghyun Yoon, and Kunwoo Park. 2024. Assessing News Thumbnail Representativeness: Counterfactual text can enhance the cross-modal matching ability. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9009–9024. Association for Computational Linguistics.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G. Parker, and Munmun De Choudhury. 2023. 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 2023, Hamburg, Germany, April 23-28, 2023*, pages 436:1–436:20. ACM.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

Quan Zou, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. 2016. Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5:2–8. Big data analytics and applications.

A Data

A.1 Multiclaim

To extract images and claims from the Multiclaim dataset, we followed these steps:

1. Obtain fact-checking article from URL.
2. Filter claims for multimodal terms (e.g. "photo", "image", etc.).
3. Filter out articles not written in English.
4. Obtain the image associated with the claim based on the HTML in the article. We look for the image tag that is closest to the claim in the HTML tree.
5. Filter out some erroneously obtained images based on their URL, such as repeated entries (usually website logos), or specific image dimensions (image too small or aspect ratio too distorted).

A.2 Multimodal Terms

The complete list of multimodal terms is: photo, image, picture, screenshot, artwork, video.

B Experimental Details

Computational resources Experiments were largely run between August 2024 and February 2025. Training and inference were performed on a cluster with heterogeneous computing infrastructure, including RTX3090, V100, and H100 GPUs. We fine-tuned a total of eight models for checkworthiness detection, which took up to two hours per model. For all our experiments, we use the Huggingface transformers library (Wolf et al., 2019), with default values unless otherwise mentioned.

Model versions See Table 5 for the specific checkpoints used to instantiate the FT and ICL models.

B.1 Fine-tuning

We use hyperparameters shown in Table 6 when fine-tuning the models on the training set of CheckThat. Non-mentioned parameters are set using the default values in the Huggingface library. During training, we keep track of the accuracy on the validation set, and at the end of training all epochs, we use the model at the step that obtained the best accuracy on the validation set.

Model	Checkpoint	Tuned Treshold
TinyBERT	`huawei-noah/TinyBERT_General_4L_312D`	0.37249295339780664,
BERT-base	`google-bert/bert-base-cased`	0.000887640770134563
BERT-large	`google-bert/bert-large-cased`	0.0006220573599437401
ResNet	`microsoft/resnet-26` 2	0.4326931064840402
ViT-base	`google/vit-base-patch16-224-in21k`	0.13248605867961713
ViT-large	`google/vit-large-patch16-224-in21k`	0.09581064554051821
BLIP	`Salesforce/blip-vqa-base`	0.07592173944742421
BLIP2	`Salesforce/blip2-itm-vit-g`	0.0761946809023745
Llava	`llava-hf/llava-v1.6-mistral-7b-hf`	n/a
Pixtral	`mistral-community/pixtral-12b`	n/a

Table 5: Model checkpoint names used in the experiments as found on the Huggingface Hub.

Hyperparameter	Value
learning rate	2e-05
max epochs	10
batch size	16

Table 6: Fine-tuning hyperparameters

B.2 In-Context Learning

For the short prompt, see Prompt 1. For the verbose prompt, see Prompt 2. In cases of few-shot learning where $n > 0$, the orange examples are repeated n times, once for each randomly retrieved example. The blue text is filled during inference for each sample in the evaluation set.

Prompt 1: Short prompt

The goal of this task is to assess whether a given statement posted by an individual is worth fact-checking. Provide your answer in by selecting between 'checkworthy' or 'not checkworthy', and provide a brief explanation. Give your response in the following format: {'label': <answer>}.

Here are some examples:
STATEMENT: <demonstration 1 text>
<demonstration 1 image>
EXAMPLE OUTPUT: <demonstration 1 label>

STATEMENT: <input text>
<input image>

Give your response as a JSON object.

Prompt 2: Verbose prompt

The goal of this task is to assess whether a given statement posted by an individual is worth fact-checking. In order to make that decision, one would need to ponder about questions, such as 'does it contain a verifiable factual claim?' or 'is it harmful?', before deciding on the final check-worthiness label. (Multimodality) Given a tweet with the text and its corresponding image, predict whether it is worth fact-checking. Answers to the questions relevant for deriving a label are based on both the image and the text. The image plays two roles for check-worthiness estimation: (i) there is a piece of evidence (e.g., an event, an action, a situation, a person's identity, etc.) or illustration of certain aspects from the textual claim, and/or (ii) the image contains overlaid text that contains a claim (e.g., misrepresented facts and figures) in a textual form. Provide your answer in by selecting between 'checkworthy' or 'not checkworthy', and providing a brief explanation. Give your response in the following format: {'label': <answer>}.

Here are some examples:
STATEMENT: <demonstration 1 text>
<demonstration 1 image>
EXAMPLE OUTPUT: <demonstration 1 label>

STATEMENT: <input text>
<input image>

Give your response as a JSON object.

C Additional Experiments

C.1 Threshold tuning for fine-tuned models

See Figures 7 and 8 for the ROC and Precicion-Recall curves for CheckThat and HINTSOFTRUTH, respectively. We selected a False Positive Rate (FPR) of 0.3 as the threshold point to prefer recall over precision. The final threshold values are reported in Table 5.

C.2 Single-logit Output

Opposed to the two-logit setup used in the experiments in this paper, one could use a single-logit setup to perform checkworthiness detection. In this setup, we would turn the softmax and cross-entropy loss into a sigmoid activation and a binary cross-entropy loss. However, we found this led did not let the model learn better-than-random accuracy, even though its loss was going down more

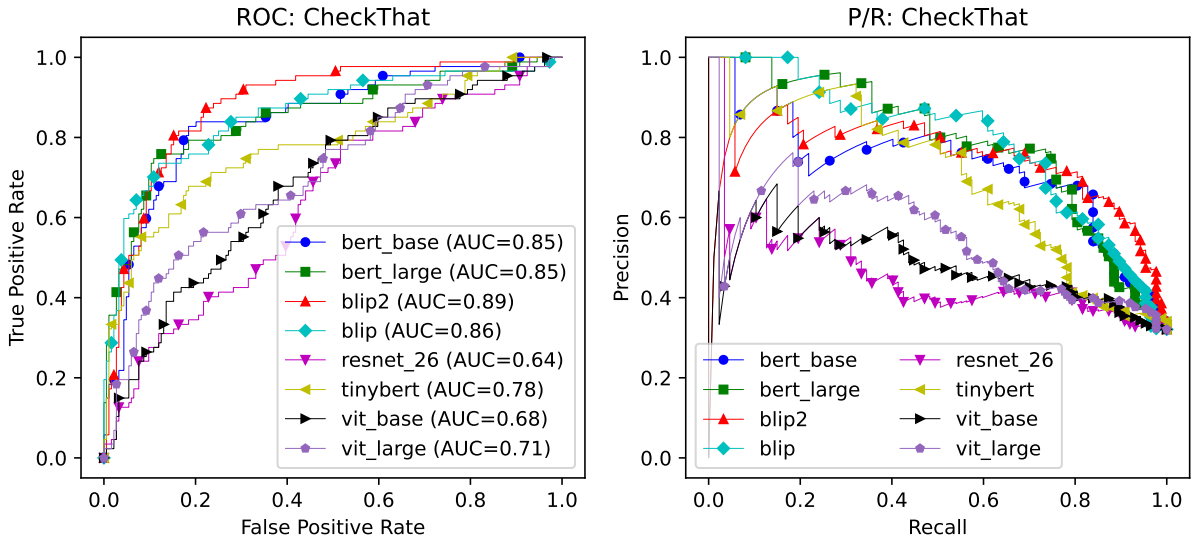

Figure 7: Tuning threshold parameter for CheckThat dataset. (Left) ROC for the fine-tuned models. (Right) Precision/Recall tradeoff.

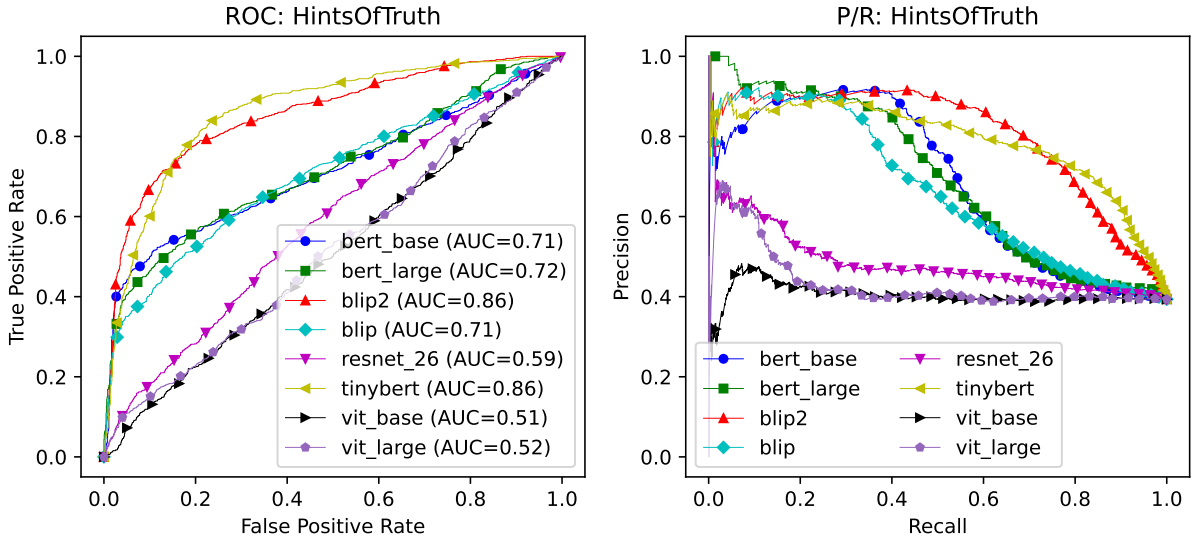

Figure 8: Tuning threshold parameter for HintsOfTruth. (Left) ROC for the fine-tuned models. (Right) Precision/Recall tradeoff.

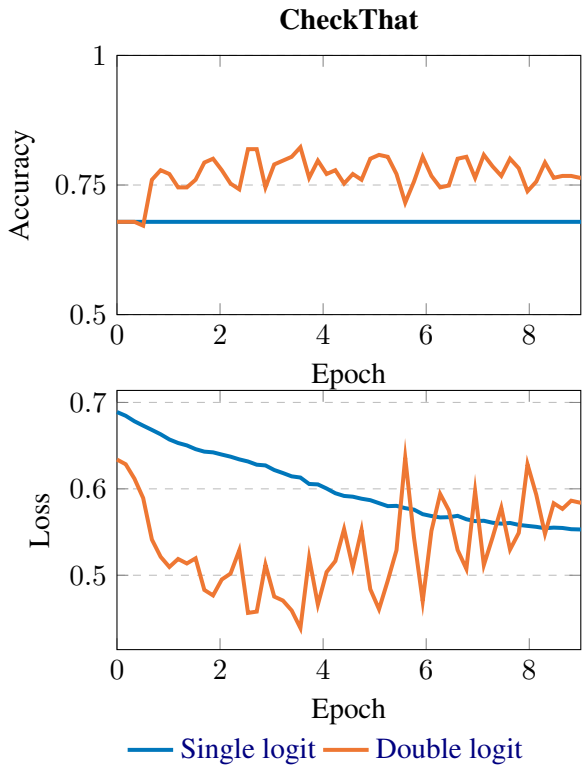

Figure 9: Performance during training a TinyBERT model using a single or double logit setup.

smoothly when training on the CheckThat data, see Figure 9.

C.3 Image captions as Negative Samples

In this set of experiments, we investigate whether models for multimodal checkworthiness are likely to label generic image captions as checkworthy wrongly. This acts as an additional sanity check that our models do not rely on spurious features or other shortcuts (Geirhos et al., 2020).

Approach We take models trained on the CheckThat dataset, and apply them across various image caption datasets. Since image captions (1) do not contain verifiable claims, or (2) are not (potentially) harmful, they can be considered non-checkworthy. We perform experiments on the following datasets: (1) **Flickr30K** (Hodosh et al., 2013) (2) **VizWiz** (Gurari et al., 2018) (3) **Conceptual Captions** (Sharma et al., 2018) (4) **Coyo** (Byeon et al., 2022) (5) **PixelProse** (Singla et al., 2024). For each dataset, impose similar filtering on the textual claims as for the scraped datasets mentioned in Section 3. Furthermore, since these datasets are of significant size, we downsample them to 6K samples each before accessing the image URLs. Table 7 denotes the final sizes of each dataset. We further split this into training/test/validation sets using a 70/20/10 ratio. We then apply each checkworthiness detection approach (as described in Section 4.2) to this task and report the classification accuracy.

Dataset	Source	Size
Flickr30K	Flickr	3,000
VizWiz	Blind humans	693
Conceptual Captions	Webpages	3,168
Coyo	Webpages	3,872
PixelProse	Webpages	5,331

Table 7: Additional image captioning datasets besides Flickr30K.

	Accuracy				
Model	**Fl**	**VW**	**CC**	**Co**	**PP**
TinyBERT	.787	.775	.593	.586	.360
BERT-base	.582	.444	.695	.875	.472
BERT-large	.472	.092	.383	.775	.111
ResNet-26	.713	.641	.697	.671	.688
ViT-base	.708	.754	.796	.809	.800
ViT-large	.703	.831	.790	.783	.750
BLIP	.913	.317	.170	.423	.065
BLIP2	.739	.930	.830	.900	.806
Llava (0)	.047	.626	.563	.740	.359
Pixtral (0)	.923	.713	.597	.363	.577

Table 8: Accuracies for each of the image captioning datasets: Flickr30K (**Fl**), VizWiz (**VW**), Conceptual Captions (**CC**), Coyo (**Co**), and PixelProse (**PP**).

Results See Table 8.

C.4 Compute Wall Time

We compute the average wall time per sample and its standard deviation over 100 random samples. See Figure 10 for the results, including the various n-shot ICL models. For the FT image-based and multimodal models, the standard deviation is considerable, due to having to resize images at inference time to be fed as input to the model. Pixtral has larger standard deviations than Llava, possibly due to the larger context size in combination with KV caching.

C.5 Prompt Error Rates

We prompt the LLMs to produce a response according to a fixed format and include some additional manual response interpretation steps to ensure we

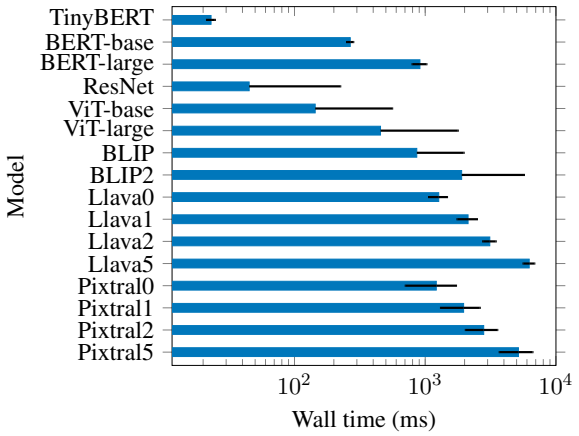

Figure 10: Average (bar, blue) and standard deviation (line, black) wall time (in milliseconds) per sample.

can extract the prediction from the model. However, even with this generous interpretation, the model occasionally erroneously reverts to a different response format or fails to provide a prediction label. We consider those responses as errors, and plot the error rates in Figure 11.

The error rates are generally low, especially for the Pixtral model. For Llava, in a one-shot setup, the error rate is largest, both for CheckThat and HINTSOFTRUTH. Qualitative observations revealed that reasons for the models failing to respond in almost all cases are due to the model refusing to answer and immediately generating an EOS token. Possibly, due to the sensitive nature of some of the claims, the samples ending up as examples in 1- and 2-shot prompts may hit a safety barrier. The low error rates highlight the robustness of the Pixtral model in adhering to the specified response format. It's discrepancy with Llava underscores the importance of model selection and prompt engineering in minimizing errors and ensuring reliable predictions.

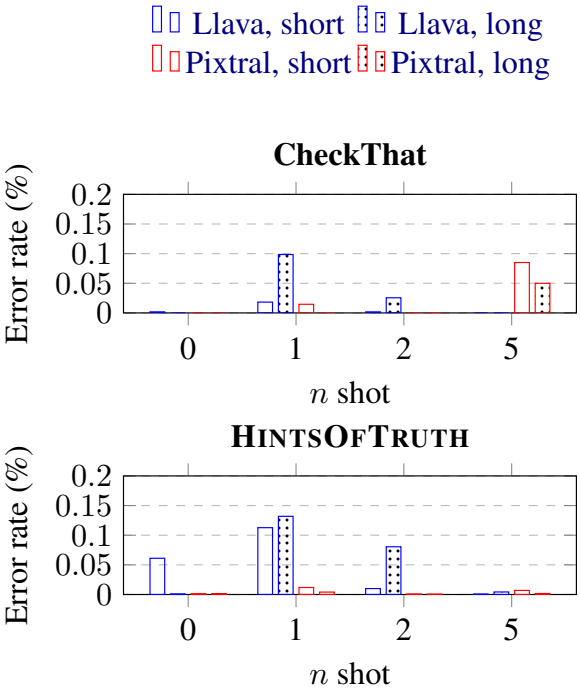

Figure 11: Response error rates for the ICL setup.