Visual Fact-Checker: An Agentic Vision-Language Approach to News Media Image-Text Consistency Verification

CS231N Project Report – Spring 2026

Andrew Lawlor
Stanford University
adlawlor@stanford.edu

Abstract

Visual misinformation in news media, where images are decontextualized or mismatched with captions to support false narratives, represents a growing threat to informed public discourse. Existing automated approaches treat image-caption verification as a static, single-pass classification task. We propose a two-stage approach: first, we fine-tune LLaVA-1.5-7B with QLoRA [7, 3] on the NewsCLIPpings `semantics_clip_text_image` *subset, the most challenging benchmark for out-of-context image detection where falsifications are constructed to maximize CLIP image-text similarity. Training on 50K balanced samples over two epochs, our model achieves 78.96% accuracy with 86.86% precision and 68.25% recall on falsified content on the held-out test set, a +11.98 percentage point improvement over fine-tuned multimodal CLIP (66.98%), the strongest reported baseline on this subset [10]. We further show that our fine-tuned 7B open-weight model outperforms Gemini 3.1 Pro [4] evaluated zero-shot on the same benchmark (75.80% accuracy, F1: 0.694 vs. 0.764), demonstrating that parameter-efficient fine-tuning can match or exceed frontier model performance on specialized misinformation detection tasks. Second, we embed the fine-tuned model within a Playwright-based agentic pipeline that autonomously scrapes live news articles from BBC, Reuters, NYT, WSJ, and AP News, extracts headline-image pairs, and produces real-time consistency verdicts. We document training challenges including fp16 numerical instability on V100 hardware and demonstrate generalization beyond the benchmark to out-of-distribution news content.*

1. Introduction

Visual misinformation in news media represents one of the most pervasive and difficult to detect forms of online manipulation. Unlike deepfakes or AI-generated images, image repurposing simply pairs a real photograph with a misleading caption: a protest photo reused years later, a politician at a routine event captioned as attending a crisis. These decontextualized images are cheap to produce, difficult to detect, and highly effective at shaping public perception [11, 10].

Existing approaches treat image-caption verification as a single-pass classification task, producing a verdict without actively investigating the claim. Models such as VisualBERT [1] and CLIP-based similarity thresholds struggle on the most challenging cases where the falsified image is semantically related to the caption and superficially plausible.

We combine a fine-tuned vision-language model with an agentic web navigation system that actively gathers evidence. The input is an image-caption pair (I, c); the output is a binary verdict $y \in \{\textit{Accurate}, \textit{Falsified}\}$. In stage one, we fine-tune LLaVA-1.5-7B [9, 8] using QLoRA on the `semantics_clip_text_image` subset of NewsCLIPpings [10], where falsified pairs are constructed to maximize CLIP [11] image-text similarity-the most challenging form of visual misinformation. In stage two, a Playwright-based pipeline autonomously scrapes live news articles and produces real-time verdicts.

Our contributions: (1) fine-tuning LLaVA-1.5-7B with QLoRA achieves 78.96% test accuracy, surpassing the strongest reported baseline by +11.98 points while training only 0.71% of parameters; (2) our model outperforms Gemini 3.1 Pro zero-shot (F1: 0.764 vs. 0.694), showing parameter-efficient fine-tuning can exceed frontier performance on specialized tasks; (3) scaling from 20K to 50K samples nearly doubles falsification recall (40.30% to 69.10%); (4) a Playwright agentic pipeline bridges bench-

mark evaluation and live deployment; (5) we document fp16 instability on V100 hardware and the hyperparameter adjustments required for stable convergence.

2. Related Work

2.1. Vision-Language Pretraining

CLIP [11] introduced contrastive pretraining on 400M image-text pairs, learning a shared embedding space for direct image-caption comparison. However, CLIP operates at the global image level rather than reasoning about specific visual-textual relationships, making similarity thresholds a weak baseline when falsified images are selected to maximize CLIP similarity, as in NewsCLIPpings. LLaVA [9] addressed this by connecting a frozen CLIP vision encoder to an LLM via a lightweight MLP projector, enabling open-ended visual reasoning. LLaVA-1.5 [8] improved on this with a two-layer MLP projector and 336px inputs, combining CLIP's visual encoding with Vicuna's [2] language reasoning to articulate why an image does or does not match a caption. Frontier models such as Gemini [4] scale this paradigm further with multimodal pretraining and RLHF alignment, but are not specialized for adversarially constructed misinformation benchmarks; we evaluate Gemini 3.1 Pro as a zero-shot baseline in Section 5.4.

2.2. Parameter-Efficient Fine-Tuning

Full fine-tuning of large VLMs is computationally prohibitive. LoRA [7] reparameterizes weight updates as $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, $r \ll \min(d, k)$, achieving competitive performance at a fraction of the parameter cost. QLoRA [3] extends this by quantizing the frozen base model to 4-bit precision, enabling fine-tuning of 7B+ models on a single GPU. We apply QLoRA at rank r=16 to LLaVA-1.5-7B's attention layers, yielding ∼30M trainable parameters (0.71% of total).

2.3. Misinformation Detection

NewsCLIPpings [10] benchmarks out-of-context multimodal media detection across four falsification strategies; the hardest subset, `semantics_clip_text_image`, uses CLIP to retrieve replacement images that maximize similarity, defeating simple embedding methods by design. Fine-tuned multimodal CLIP (ViT-B/32) achieves 66.98% accuracy; VisualBERT variants reach 54–57% [10]. These discriminative approaches share a key weakness: single-pass verdicts without semantic reasoning or corroborating evidence. Our model achieves 78.96% on the held-out test set, a +11.98 point improvement. FakeNet [1] targets fake news stance detection via selective feature extraction, but operates on text only; our work extends verification to image-caption pairs with visual evidence.

2.4. Agentic Web Navigation

WebAgent [5] and CogAgent [6] demonstrated VLM-based agents for decomposing web tasks and GUI interaction respectively, establishing the viability of autonomous web navigation-though neither targets fact verification. Our Playwright pipeline adapts these agentic approaches to navigate news websites, extract headline-image pairs, and optionally cross-reference fact-checking databases, shifting from passive classification to active, evidence-aware verification.

3. Data

We fine-tune on the NewsCLIPpings [10] dataset, focusing on the `sem_clip_text_img` subset (453K samples) — the hardest scenario, where falsifications maximize CLIP similarity so shallow detectors fail by design. The four subsets are summarized in Table 1. Figures 1 and 2 show representative pristine and falsified examples.

Subset	Size	Threat
`sem_clip_text_img`	453K	CLIP image substitution
`sem_clip_text_txt`	516K	Similar story swapped
`person_sbert_txt_txt`	18K	Same person, diff. context
`scene_resnet_place`	125K	Same scene, diff. event

Table 1: NewsCLIPpings subsets (Luo et al., 2021).

3.1. Dataset Splits and Preprocessing

We sampled 50K balanced examples for training (25K pristine, 25K falsified), 2K for validation (1K/1K), and evaluated on the full held-out test set of 7,264 balanced examples (3,632/3,632); splits are disjoint. Images are resized to 336×336px to match the CLIP ViT-L/14 encoder. No augmentation was applied; dataset diversity across topics, geographies, and time periods provides sufficient regularization. Pixel values are normalized via the standard CLIP pipeline (μ=$[0.4815, 0.4578, 0.4082]$, σ=$[0.2686, 0.2613, 0.2758]$).

4. Methods

Figure 3 shows the full application architecture. LLaVA 1.5 7B multimodal model serves as our baseline (model weights loaded from Hugging Face). We Supervise train (SFT) the model to learn how to become a news media visual fact checker using the NewsCLIPpings dataset consisting of media image/caption pairs on the Google Cloud Platform on a V100 GPU. The Playwright browser automation platform acts as an agentic runtime to conduct inference against live Web news articles from sources including BBC, Reuters, NYT, WSJ, and AP News.

Figure 1: Pristine Example

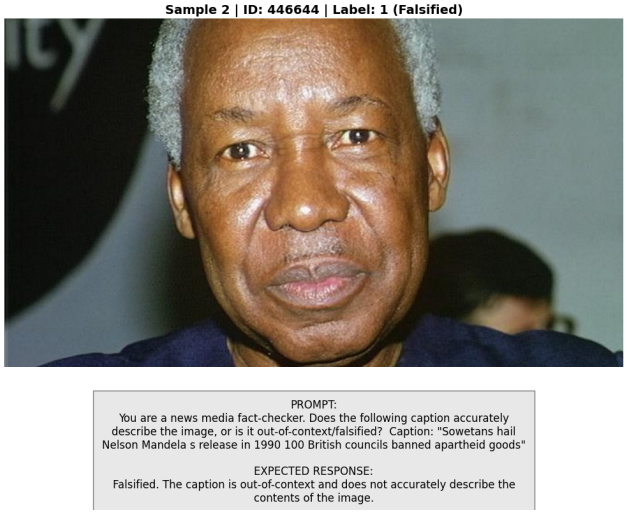

Figure 2: Falsified Example

Figure 4 shows the LLava architecture. LLaVA-1.5 [8] connects two frozen pretrained models-a CLIP ViT-L/14 vision encoder [11] and a Vicuna-7B language model-via a lightweight trainable MLP projector. The vision encoder converts a 336×336 input image into 576 patch embeddings of 1024 dimensions each. The MLP projector (two layers with GELU activation) maps these into Vicuna's 4096-dimensional token space. The resulting visual tokens are concatenated with text tokens and processed autoregressively. Table 2 details the LLaVA 1.5 components.

Component	Parameters	Status
CLIP ViT-L/14 (336px)	∼304M	Frozen
MLP Projector	∼20M	**Trainable**
Vicuna-1.5-7B LLM	∼6.7B	Frozen
QLoRA Adapter (r=16)	∼30M	**Trainable**
Total	**∼7.06B**	**0.71% trainable**

Table 2: LLaVA-1.5-7B model components and parameter counts.

We fine-tune the model on NewsCLIPpings [10] using LoRA adapters applied to the attention layers, keeping all base weights frozen.

For all training we used system prompt `"You are a news media fact-checker."` and task instruction `"Does the following caption accurately describe the image, or is it out-of-context/falsified?"` The model was trained to output either `"Accurate. The caption accurately describes the contents of the image."` or `"Falsified. The caption is out-of-context and does not accurately describe the contents of the image."`

Finetuning is accomplished with a QLora adapter. We encountered numerical instability and out-of-memory errors that required us to introduce gradient clipping, reducing the learning rate from 2e-4 to 2e-5 and lowering DeepSpeed's minimum loss scale to 1e-8 when training on the 50K sample run over two epochs. The run took approximately 22 hours to complete on a V100. For the successful 50K 2-epoch training run, the effective batch size was 16. Because we were training a 7B parameter model on a single NVIDIA V100 GPU (16GB VRAM), fitting a large batch size entirely in memory was not feasible. To achieve an effective batch size of 16, we used gradient accumulation:

- `per_device_train_batch_size`: 1 (one image/conversation processed per forward pass)
- `gradient_accumulation_steps`: 16 (gradients accumulated over 16 forward passes before a single backward pass and weight update)

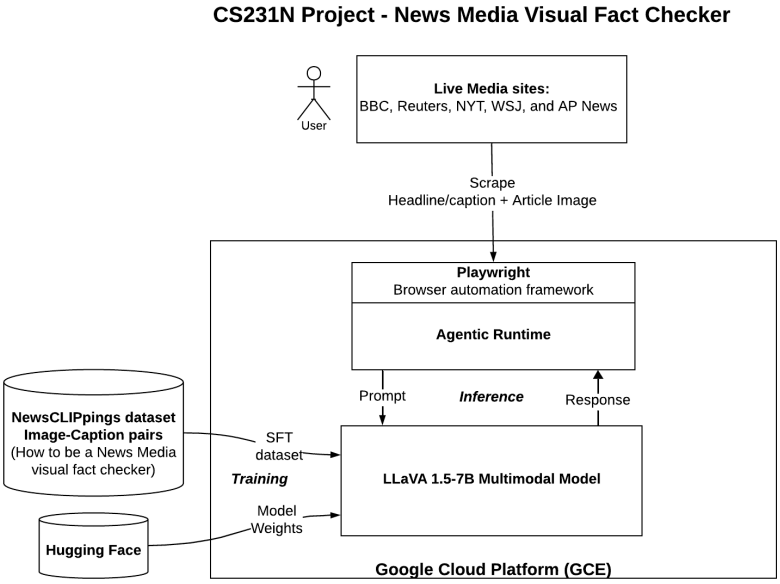

Figure 3: Project Architecture

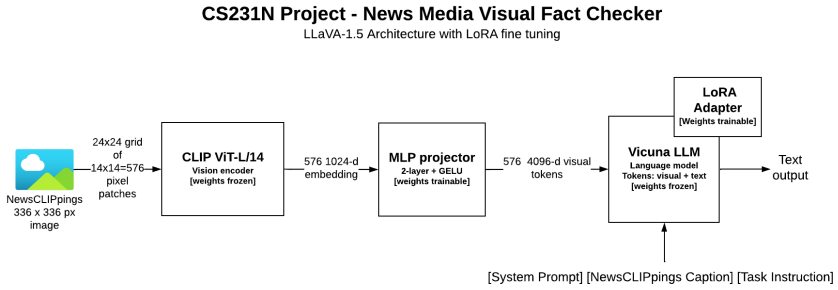

Figure 4: LLaVA Architecture

This approach simulates a batch size of 16 while remaining within the V100's memory constraints.

4.1. Loss Function

Given an image I and caption c, our model predicts a binary verdict $\hat{y} \in \{Accurate, Falsified\}$. Fine-tuning minimizes the cross-entropy loss over the target token sequence:

$$\mathcal{L} = -\sum_{t} \log P_\theta(y_t \mid y_{<t}, I, c) \tag{1}$$

where θ represents the trainable QLoRA parameters (0.71% of total), and loss is computed only on the `gpt` response tokens.

4.2. QLoRA Formulation

LoRA [7] reparameterizes each weight update as $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and rank $r \ll \min(d, k)$. We use r=16, yielding approximately 30M trainable parameters.

4.3. Live Fact-Checking Agentic Pipeline

To demonstrate generalization beyond the NewsCLIPpings benchmark, we deployed our fine-tuned model against live news content using Playwright for automated web scraping. We processed articles from BBC, Reuters, NYT, WSJ, and AP News. We chose these media outlets because they are reputable services with straightforward HTML. We followed all laws and avoided sites with a robot.txt file.

To demonstrate real-world applicability, we engineered an autonomous pipeline to perform live fact-checking on active news URLs. The system operates through a six-step

flow:

1. **Command Line Interface** The live fact-checking pipeline can be invoked from the command line:

```
$ python live_fact_checker.py \
"https://bbc.com/news/123" \
"Optional fake headline"
```

2. **Navigation:** Using `Playwright`, a headless Chromium browser navigates to the target URL. It employs a standard consumer user-agent and waits for the DOM and lazy-loaded images to fully render.

3. **Extraction:** The agent extracts the headline and primary image via injected JavaScript using a cascading strategy. It prioritizes Open Graph (`og:title`, `og:image`) and Twitter Card metadata. If metadata is missing, it falls back to the HTML `<title>` and heuristically selects the largest non-logo `<img>` element on the page.

4. **Prompt Construction:** The downloaded image and extracted headline are formatted into the exact prompt template used during fine-tuning. The text is wrapped in LLaVA's conversation template, and the image is processed into a tensor via the CLIP vision encoder. Researchers can optionally inject a fake headline to test robustness.

5. **Inference:** The multimodal inputs are passed to the fine-tuned 50K LLaVA model. Inference is executed with greedy decoding (`temperature=0`) to ensure deterministic outputs.

6. **Presentation:** Designed for live demonstration, the pipeline outputs the model's raw generated verdict (e.g., *"Falsified. The caption is out-of-context..."*) alongside the extracted metadata to the terminal for immediate qualitative review.

Output Parsing Robustness

Because LLaVA generates natural language rather than discrete logits, robust parsing is required to extract the binary verdict. The pipeline employs a hierarchical string-matching strategy.

First, it checks for an exact prefix match (i.e., the output starts with "Accurate" or "Falsified"), which captures most conforming responses. If this fails due to unexpected conversational filler, it falls back to a substring search for the target keywords.

In the rare event of a completely non-conforming output or a generation error, the parser assigns an "Unknown" label. During metric calculation, "Unknown" predictions are conservatively mapped to the negative class (Accurate), ensuring the model is only credited when it explicitly outputs the "Falsified" label.

5. Experiments

We perform multiple evaluations against a variety of LLaVA and frontier model configurations.

5.1. Baseline

The untrained LLaVA 1.5 model exhibits a strong bias toward predicting **Accurate** on both splits, failing to detect the majority of falsified samples. Table 3 shows false negatives dominating in both the 1K validation subset and the full 7,264-sample test set (370 and 2,485 respectively), confirming the untrained model's inability to identify out-of-context pairs.

Val (1K)	**Pred. Accurate**	**Pred. Falsified**
Act. Accurate	411	89
Act. Falsified	370	130
Test (7,264)	**Pred. Accurate**	**Pred. Falsified**
Act. Accurate	3013	619
Act. Falsified	2485	1147

Table 3: Baseline confusion matrices (val and test). False negatives dominate both splits, confirming strong *Accurate* bias.

5.2. 20K Samples Training Run

We finetuned the baseline LLaVA 1.5 model with a one Epoch run using 20K samples drawn from the `semantics_clip_text_image` dataset. Figure 5 shows the resulting log loss curve. Notice the spike near step 530. We hypothesize that these were particularly difficult samples for the model to handle (perhaps the image was of the right person, but of an event distinct from what the caption identifies).

	Pred. Accurate	**Pred. Falsified**
Actual Accurate	900	100
Actual Falsified	597	403

Table 4: Confusion matrix for the 20K model on the 2K balanced validation subset.

Table 4 shows the confusion matrix for the 20K sample (1 epoch) run on a 2K sample validation dataset. Again, the model exhibits a strong bias toward predicting **Accurate**, with a high false negative rate (597) indicating it fails to detect the majority of falsified samples.

The 20K model achieved 65.15% accuracy but exhibited a strong bias toward predicting **Accurate**, with a recall of

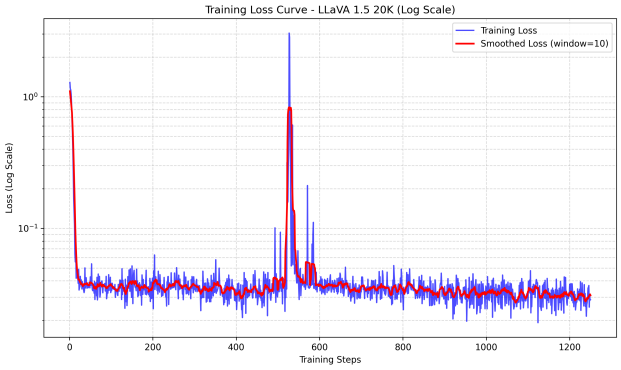

Figure 5: Pilot log loss curve (20K samples). Spike at step 530 corresponds to high-difficulty samples; training recovers within 50 steps.

only 40.30% on the Falsified class. This indicated that the model had not seen sufficient diversity of falsified examples to learn the subtle visual-semantic inconsistencies characteristic of the `semantics_clip_text_image` subset.

5.3. 50K Sample Training Run

In pursuit of better performance, we increased the training set to 50K balanced samples and added a second epoch.

Figure 6 shows the training loss curve for the 50K sample, two-epoch run with stabilized hyperparameters (lr=2e-5, min loss scale=1e-8).

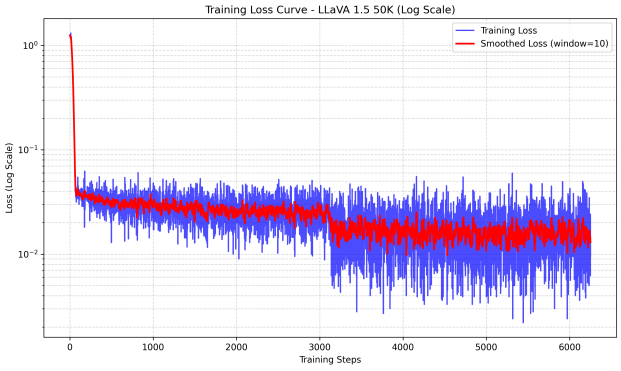

Figure 6: Final log loss curve (50K samples) with stabilized hyperparameters (lr=2e-5, min loss scale=1e-8).

Several features distinguish this curve from the 20K pilot run. First, the prominent loss spike observed near step 530 in the 20K run is absent here, indicating that the reduced learning rate and lower DeepSpeed minimum loss scale successfully resolved the fp16 numerical instability encountered in earlier training attempts. Second, the smoothed loss curve descends steadily through the first epoch (steps 0–3,125) and continues to improve through the second (steps 3,125–6,250), confirming that the additional epoch contributed meaningful learning rather than overfitting. Third, the final smoothed loss stabilizes in the range of 10^{-2}, approximately one order of magnitude below the 20K run's converged value, consistent with the substantial accuracy improvement observed on both the validation set (65.15% $\rightarrow$ 78.60%) and the held-out test set (57.27% $\rightarrow$ 78.96%).

The residual high-frequency variance in the raw loss (visible as dense oscillations around the smoothed curve) is characteristic of per-sample gradient updates with batch size 1 and gradient accumulation, and does not indicate instability-the smoothed trend confirms consistent convergence throughout both epochs.

The 50K fine-tuned model shows substantially improved falsification detection across both splits. The model is conservative, with false negatives exceeding false positives on both validation (309 vs. 119) and test (1,153 vs. 375), reflecting the same precision-favoring boundary observed in the baseline-but now correctly identifying the majority of falsified pairs. We attribute this improvement to exposure to 50K diverse falsified examples, enabling the model to learn visual-semantic inconsistencies the untrained baseline could not detect.

Val (2K)	**Pred. Accurate**	**Pred. Falsified**
Act. Accurate	881	119
Act. Falsified	309	691
Test (7,264)	**Pred. Accurate**	**Pred. Falsified.**
Act. Accurate	3257	375
Act. Falsified	1153	2479

Table 5: 50K model confusion matrices (val and test). False negatives exceed false positives on both splits, reflecting a conservative decision boundary.

Table 6 shows performance across validation and test splits. The 50K fine-tuned model achieves 78.96% test accuracy, surpassing fine-tuned multimodal CLIP (66.98% [10]) by +11.98 points. The most striking gain is recall on the Falsified class, rising from 31.58% to 68.25% (+36.67 points)-fine-tuning on 50K balanced examples corrects the baseline's strong *Accurate* bias, enabling detection of more than two in three falsified pairs at 86.86% precision. Validation results (78.60% accuracy, 85.31% precision, 69.10% recall) closely mirror test set's 78.96%, confirming good generalization. The higher precision than recall reflects a conservative decision boundary desirable in deployments where false misinformation accusations carry reputational cost.

5.4. Frontier Model Comparison

Table 7 compares our model against Gemini 3.1 Pro [4], evaluated zero-shot on a 1,000-sample test subset using the

Metric	Base	20K	50K	Δ
Validation (2K balanced)				
Accuracy	54.10%	65.15%	**78.60%**	+24.50%
Precision	59.36%	80.12%	**85.31%**	+25.95%
Recall	26.00%	40.30%	**69.10%**	+43.10%
Test (7,264 samples)				
Accuracy	57.27%	—	**78.96%**	+21.69%
Precision	64.95%	—	**86.86%**	+21.91%
Recall	31.58%	—	**68.25%**	+36.67%
F1-Score	0.425	—	**0.764**	+0.339

Table 6: Validation and test performance (NewsCLIPpings `sem_clip_text_img`). Positive class = Falsified. Δ is Base → 50K. Baseline was evaluated on 1K balanced set. 20K was not evaluated on the test set (–).

same prompt template as LLaVA. Our fine-tuned model outperforms Gemini on accuracy (78.96% vs. 75.80%), recall (68.25% vs. 54.80%), and F1 (0.764 vs. 0.694). Gemini's higher precision (94.48%) mirrors the conservative decision boundary observed in our CoT ablation (Section 5.5)-without task-specific fine-tuning, capable models default toward caution, missing nearly half of all falsified samples. Parameter-efficient fine-tuning of a 7B open-weight model thus matches or exceeds frontier zero-shot performance while training only 0.71% of parameters and eliminating API inference costs.

Metric	LLaVA (Base)	LLaVA (50K)	Gemini 3.1 Pro
Accuracy	57.27%	**78.96%**	75.80%
Precision	64.95%	86.86%	**94.48%**
Recall	31.58%	**68.25%**	54.80%
F1-Score	0.425	**0.764**	0.694

Table 7: Frontier comparison on NewsCLIPpings test set. Positive class = Falsified. LLaVA evaluated on full 7,264 samples; Gemini on a balanced 1,000-sample subset.

5.5. Chain-of-Thought Ablation

A Zero-Shot CoT variant-modifying the training target to include a brief rationale before the final verdict-caused the model to become extremely conservative (Table 8): precision rose to 92.52% but recall collapsed to 27.20% (−16.10 points accuracy). We hypothesize the reasoning step surfaces uncertainty the model resolves by defaulting to *Accurate*. CoT may suit contexts where false positives are costlier than false negatives, but is unsuitable for maximizing falsification detection. We retained the direct verdict format for all reported results.

5.6. Live News-site Agentic Fact-Checking

Using the same GCP/V100 environment, each fact-check takes ∼30s (∼6s Playwright, ∼22s model load, ∼2s inference). We deployed the pipeline against 30 live articles from BBC, Reuters, NYT, WSJ, and AP News RSS feeds: 15 authentic pairs and 15 manually falsified pairs constructed by swapping headlines between unrelated articles. The 50K model achieved 70.00% accuracy and 73.33% recall, correctly flagging 11 of 15 novel falsified pairs. Five false positives likely reflect the abstract or loosely-correlated nature of modern news photography-for instance, a climate policy article paired with a featureless aerial landscape was correctly flagged as out-of-context. Results are shown in Table 9.

Format	Accuracy	Precision	Recall
Direct verdict	**78.60%**	85.31%	**69.10%**
Chain-of-Thought	62.50%	**92.52%**	27.20%

Table 8: CoT vs. direct verdict on 2K validation subset. Positive class = Falsified.

Metric	Value		Pr. Acc.	Pr. Fals.
Accuracy	70.00%	**Act. Acc.**	10	5
Precision	68.75%	**Act. Fals.**	4	11
Recall	73.33%			
F1-Score	0.710			

Table 9: Live agentic evaluation on 30 articles (15 authentic, 15 falsified). Metrics and confusion matrix side by side.

6. Conclusions

We have presented a two-stage system for automated image-caption consistency verification in news media. Fine-tuning LLaVA-1.5-7B with QLoRA [3, 7] on the NewsCLIPpings `semantics_clip_text_image` subset achieves 78.96% accuracy with 86.86% precision and 68.25% recall on the held-out test set, surpassing the strongest reported baseline by +11.98 percentage points while training only 0.71% of model parameters. Scaling from 20K to 50K training samples nearly doubled recall (40.30% → 69.10%) while precision remained stable, identifying training data volume as the primary lever for improving falsification detection on this adversarially constructed subset.

Evaluated zero-shot against Gemini 3.1 Pro [4], our fine-tuned 7B model achieves higher accuracy (78.96% vs. 75.80%), recall (68.25% vs. 54.80%), and F1-score (0.764 vs. 0.694), demonstrating that parameter-efficient fine-tuning on a specialized benchmark can exceed frontier zero-shot performance at a fraction of the inference cost. The consistent precision-recall asymmetry across Gemini zero-shot and our Chain-of-Thought ablation suggests that without task-specific training, capable models default to

conservative decision boundaries poorly suited for misinformation detection.

The Playwright-based agentic pipeline demonstrates generalization to out-of-distribution live news content from BBC, Reuters, NYT, WSJ, and AP News, correctly classifying authentic articles and flagging manually constructed out-of-context pairings, representing a meaningful step from static classification toward active, evidence-aware fact-checking.

6.1. Limitations

Evaluation is restricted to the `semantics_clip_text_image` subset; performance on the merged benchmark remains untested. fp16 training on V100 hardware required reducing the learning rate from 2e-4 to 2e-5 and lowering DeepSpeed's minimum loss scale to 1e-8; bf16-capable hardware would likely improve results. The Gemini comparison used a 1,000-sample subset due to API cost constraints. Live pipeline evaluation used manually constructed falsifications as a proxy for ground truth; naturally occurring misinformation in live content cannot be verified at scale.

6.2. Future Work

Evaluating on the full merged NewsCLIPpings benchmark would test generalization across all four falsification strategies. Tighter integration of the Playwright evidence-gathering loop-feeding reverse image search and fact-checking results back into a second-pass evaluation-would move the system closer to a true agentic reasoner. Training on bf16-capable hardware and exploring larger LoRA ranks (r=32, r=64) may further improve recall. Finally, decision boundary calibration through threshold tuning or confidence modeling represents a promising direction given the consistent precision-recall asymmetry observed across all model configurations.

7. Contributions & Acknowledgements

As the only member of the team, I, Andrew Lawlor, am fully responsible for problem formulation, literature review, dataset preparation, implementation, experiments, analysis, writing, figures, and poster preparation.

7.1. Generative AI usage statement

I relied on Claude Sonnet 4.6 for brainstorming and as my AI tutor for computer vision concepts throughout the project. I leveraged Cursor with Gemini 3.1 to generate code for non-CV scaffolding, including the interface with Google Cloud Platform (GCP). I prompted Claude for editorial and formatting assistance on my final report to increase readability and coherence.
I have included cursor_chat_prompts.txt in my code submission zip file showing the prompts I used with Cursor during the project.

References

[1] T. Aljrees. Fake news stance detection using selective features and fakenet. *PLOS ONE*, 18(6), 2023. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0287298. 1, 2

[2] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna/, March 2023. Blog post. 2

[3] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. 1, 2, 7

[4] Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2, 6, 7

[5] I. Gur, H. Furuta, A. Huang, M. Saber, Y. Matsuo, D. Eck, and A. Fishi. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*, 2024. https://arxiv.org/abs/2307.12856. 2

[6] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Zhang, J. Li, Y. Dong, M. Ding, and J. Tang. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2024. https://arxiv.org/abs/2312.08914. 2

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 2, 4, 7

[8] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. https://arxiv.org/abs/2310.03744. 1, 2, 3

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. https://arxiv.org/abs/2304.08485. 1, 2

[10] G. Luo, T. Darrell, and A. Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. https://aclanthology.org/2021.emnlp-main.545/. 1, 2, 3, 6

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. https://arxiv.org/abs/2103.00020. 1, 2, 3