# Group 9 Final Project - Multimodal Misinformation Detection with Retrieval-Augmented Large Vision-Language Models

**Daniel Chechelnitsky**
University of Michigan
dchechel@umich.edu

**Marco Conati**
University of Michigan
marcoco@umich.edu

**James Edwards**
University of Michigan
edwjames@umich.edu

## Abstract

Recent advancements have significantly enhanced the capabilities of Large Vision Language Models (LVLMs), making them adept at complex reasoning tasks. These improved models can now utilize techniques like retrieval-augmented generation (RAG) to effectively access and process external data, presenting a unique opportunity to boost performance on tasks such as multimodal misinformation detection. Preliminary work has proved that RAG can improve multimodal misinformation detection systems, which have previously suffered from a lack of generalizability due to knowledge expiration over time (12). We build on these methods, offering a quantitative and qualitative comparison of retrieval-augmented LVLMs ability to detect misinformation. First, we demonstrate that open-source models, such as the visual instruction tuned LLaVA, perform poorly in comparison with larger, closed-source models like GPT-4V. We then show that previous results have been influenced by *hindsight bias*, a phenomenon where evaluation of retrieval-augmented systems is aided by the fact that the test examples occurred far in the past, and therefore more information about them has accumulated in the knowledge base than would have been present if using the model at the relevant time. Finally, we quantitatively evaluate a host of techniques for improving the performance of such pipelines, including various prompting techniques, structured approaches to question asking, and task-specific fine-tuning. Overall, we find that smaller open-source models are limited due to their lack of world knowledge, but that correct techniques can still lead them to robust reasoning processes which achieve much improved results.

## 1 Introduction

### 1.1 Background

Widespread adoption of internet technologies enables the spread of information at a much faster pace than was previously possible. Consequently, these technologies also enable misinformation to spread much faster than before, creating new barriers to their users' abilities to determine truthfulness of news, scientific claims, and other domains which rely on concretized facts to inform decision making. In certain cases, such as the 2020 United States presidential election, the most popular misinformation sources on some platforms can be viewed even more often than credible news outlets (2).

Algorithmic approaches to detect misinformation focus almost exclusively on linguistic features such as semantics and syntax to determine truthfulness. However, they lack any awareness of context, rendering them a priori unable to determine the truthfulness of all statements. E.g., the statement "Yesterday was a Wednesday" is impossible to determine without basic awareness of the current day. This context-blindness limits models' generalizability (3) and makes them prone to knowledge

Submitted to 1st Conference of 692 Final Projects (EECS 692, 2024).

expiration, especially in cases like the COVID-19 pandemic where both the common scientific understanding of the truth and approaches to spreading misinformation evolve rapidly (10).

The advent of publicly available Large Language Models (LLMs) threatens to compound these issues, as evidence suggests that AI-generated misinformation may not share the same linguistic characteristics upon which existing models rely for determining truthfulness (14). This creates a need for new misinformation detection models that are context-aware and have access to external knowledge sources. To this end, RAG has been used in approaches like LEMMA, described in detail in section 1.3, allowing LLMs to access the necessary context for misinformation detection tasks (12).

### 1.2 Our Contributions

Our contributions comprise the following:

- A quantitative and qualitative characterization of vision instruction tuned LVLM performance on misinformation detection using retrieved evidence, and the ways this ability differs from larger models like GPT-4V
- Quantitative evidence that RAG systems may be biased if allowed to access evidence that was not originally available.
- Experiments and analysis measuring the effectiveness of several strategies to improve the performance of retrieval-augmented LVLMs for the task of misinformation detection
- A novel technique using the logits of the model as inputs to a logistic regression, thereby extracting information from the reasoning capabilities of a LVLM while introducing helpful inductive bias, which we show is more effective than simply prompting in a way that suggests this inductive bias

### 1.3 Related work

**Towards LVLM-Enhanced Multimodal Misinformation Detection (LEMMA)** A recent study on information classification using the LVLM was performed (12). However, unlike other recent work using LVLMs, this study used LLM-driven data retrieval-augmented techniques, in hope of providing context for misinformation detection. GPT-4V (9) powered modules were responsible for generating search terms, filtering internet results, and consolidating results into pertinent information for the misinformation task. The block diagram for this approach is presented in 1. The resulting LVLM model, called LEMMA, was baseline tested against state of the art models: LLaVA, GPT-4 with Image Summarization, and GPT-4V. The LEMMA model was found to outperform all three of these models in misinformation classification respectively, thus setting a new state of the art for information classification methods.

The authors' proposed pipeline first asks GPT-4V to make an initial judgment about the example, and asks it if it requires more information. If it does, LEMMA uses DuckDuckGo search to retrieve articles, from which it then extracts relevant passages (after filtering out untrusted sources. Then, given this new information, the pipeline refines its original reasoning before making a final prediction.

The LEMMA authors generously gave us access to their code, and their work served as the starting point for our project (12).

**Low-Rank Adaptation of Large Language Models (LoRA)** Most current LLM models ' rely on adapting one large-scale, pre-trained language model to multiple downstream applications' (5). However, fine-tuning for many downstream tasks requires significant resources. To combat this, one type of fine-tuning is layer offloading as done in the Low-Rank Adaptation study: LoRA. By freezing model weights and matrix decomposition techniques, this reduction in overhead is achieved, and provides equal or better results than state of the art models like RoBERTa and GPT-3. It even equals or outperforms those models when they have had baseline hyperparameter fine-tuning.

For the purposes of our study, the LoRA study provides an alternative fine-tuning approach to traditional hyperparameter fine-tuning or dataset fine-tuning methods that exist in other state of the art approaches and provides a direct, more efficient way to reduce computational power required to perform LLM tasks. Due to this, LoRA makes fine-tuning models very easy.

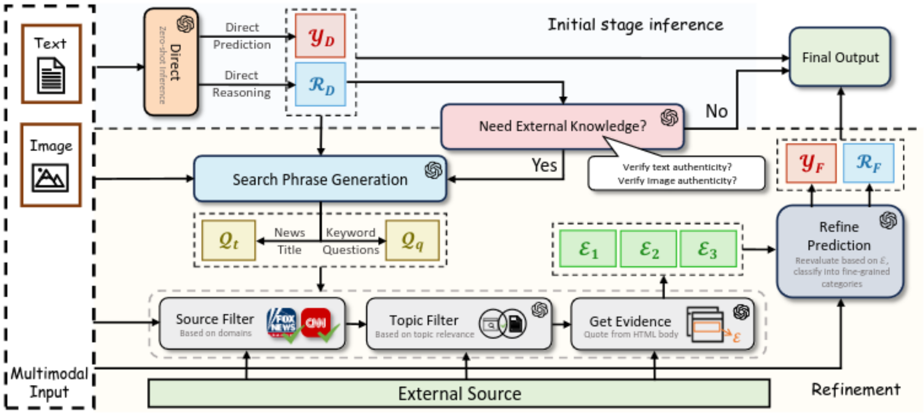


Figure 1: The pipeline of LEMMA. The process hinges on two key inputs: multimodal data and selectively filtered evidence gathered from external sources. Components marked with the OpenAI LOGO are developed using the LVLM (GPT-4V)

**Describe, Explain, Plan, and Select (DEPS)** The "Describe, Explain, Plan and Select" paper, is designed to improve planning and execution in the complex, open-world environment of Minecraft (11). At its core, DEPS leverages Large Language Models (LLMs) to dynamically generate and adjust task plans. The architecture consists of four key components:

- **Descriptor:** Activated upon encountering execution failures, it summarizes the current situation and the failure in a textual format, facilitating an understanding of the context for subsequent planning adjustments.
- **Explainer and Planner (LLM):** This dual-role component first interprets the descriptor's summary to identify where and why the previous plan failed. It then generates a refined plan that addresses these shortcomings, incorporating insights from the descriptor's feedback.
- **Selector:** A trainable module that evaluates the revised plan's sub-goals, ranking them based on estimated completion steps. This process ensures that the most efficient path is chosen, optimizing the plan's feasibility and effectiveness in the given state.
- **Controller:** Executes the selected sub-goals, providing the real-world actions necessary to progress towards task completion. It adapts to the plan's requirements, ensuring alignment with the dynamic planning process.

This architecture allows DEPS to iteratively refine its plans based on real-world feedback, significantly improving its ability to accomplish tasks in dynamic and unpredictable environments. DEPS was able to robustly accomplish 70+ Minecraft tasks, nearly doubling the overall performance of its counterparts.

This is relevant to our project as it exemplifies the potential of guiding LLM reasoning with explicit system structure.

## 2 Methods

### 2.1 Dataset

The Twitter dataset, developed by Ma et al. (8), comprises 2,308 time-stamped tweets which were manually fact-checked. This dataset was compiled by selecting high-engagement tweets from the existing Twitter15 and Twitter16 datasets and was further enhanced by annotating each tweet with labels indicating the veracity of the information, categorized into 'false rumor', 'true rumor', 'unverified rumor', and 'non-rumor'. This structured approach provides a framework for analyzing the dynamics of misinformation spread across social media platforms. For the purposes of our study on multimodal retrieval-augmented misinformation detection, we utilized the subset of 770 tweets that included both text and image content, as extracted and prepared by Xuan et al. (12).

We selected this dataset as it was used in the original LEMMA paper, allowing for a more robust comparison. It also features timestamps, which are important for our "time-aware" experiment set, described further in Section 2.3

### 2.2 Baseline

We begin our experiments by establishing a baseline of performance. To do this, we replicate LEMMA's experiments as closely as we can while converting them from using GPT-4V as the central model to using LLaVA-NeXT (7), the open-source state-of-the-art for vision-language reasoning models. Using LLaVA-NeXT (also known as LLaVA-1.6) provides us the ability to host and fine-tune our models locally for increased control over experiments, and also provides a clearer evaluation of the impact of the evidence retrieval pipeline on the misinformation detection process. Since LLaVA is smaller and instruction tuned on a small subset of specific tasks, it lacks the wider world knowledge that GPT-4V might possess.

Note that for all experiments, GPT-3.5Turbo was still used for the stage of the LEMMA pipeline involving extraction of key quotes from the HTML bodies of the retrieved evidences. We chose to preserve this aspect to better measure the degree to which LLaVA's visual reasoning compares with GPT-4V's, which is independent of LLaVA's (or a corresponding LLaMA model's) ability to reason about text alone. It is also worth noting that, since the model is restricted to output direct quotes from the HTML body, GPT-3.5Turbo is preferable as it is much better at following such instructions, while the actual likelihood for its world knowledge or reasoning to leak into the results via just direct quotes of evidence is virtually nonexistent.

For this first experiment we keep prompts as close as possible to the original LEMMA prompts, only changing their format to reflect the different chat format LLaVA-NeXT was trained on and to accomodate LLaVA's inconsistency in following the provided output format.

Note that this "baseline" does not correspond to Xuan et al.'s. We are using their final model as our starting point, and so our baseline is based on their finished product.

**Guided Generation** LEMMA's original authors prompted GPT-4V to return data directly in JSON format. This proves a barrier to a direct comparison with our LLaVA-based version of LEMMA, as LLaVA-NeXT in trial experiments was not able to robustly and reliably generate JSON format. To remedy this, we altered prompts to require a simpler output format. Then, we used guided generation techniques, which constrain the generation via a grammar specified in a modified version of Backus-Naur Form, to ensure that LLaVA kept to this simplified format. For example, for LEMMA's first (zero-shot) prediction, LLaVA was constrained in that it could only output "True" or "False" as its first word.

While this allowed us to reliably acquire and separate predictions and explanations from the model, constrained generation techniques have been documented as having negative impacts on task performance. We argue that this is an inevitable downside of using LLaVA, and that our results will thus still reflect the best possible result feasibly extractable from LLaVA in the LEMMA pipeline.

**Model Deployment** Working with limited resources, we run a 3-bit quantized, 34B parameter version of LLaVA-NeXT using llama.cpp, an open-source program for deploying large models based on Meta's LLaMA family of models. While quantization to this severe degree can negatively impact model performance, most evidence points to this being a better option than running a model with far fewer parameters, our only other option given that we could not access a GPU large enough to fit the entire 34B model into its memory. Most requests are made with temperature of $0.1$, with the logit-based approach in Section 2.5 being the only exception. This setup, including the guided generation described above, applies to every further section except fine-tuning. See section 2.6 for more details on the setup fine-tuning.

### 2.3 Time-Aware Retrieval

In our retrieval-augmented generation process, we utilized the DuckDuckGo instant answers API (1) to control the internet-based evidence our models could access. Using the timestamps from our dataset, we tailored the retrieval to only include documents that were available at the time each tweet was made, thus preserving historical accuracy by testing the model as if the claims had just been made. When we were unable to determine the publishing date of a piece of evidence, it was also discarded to preclude any potential leakage. This temporal restriction was used for all experiments we conducted, other than a comparative baseline experiment where we did not restrict access to future data. This dual approach allowed us to evaluate the impact of temporal constraints on the model's ability to fact-check claims accurately. Implementing these temporal limits is critical, as it helps prevent the models from overstating their real-time fact-checking capabilities, ensuring that our results remain relevant and applicable in practical scenarios.

## 2.4 Prompt Engineering Approaches

### 2.4.1 Shortening Prompts

Motivated by the context window limitations of the LLaVA model, which supports up to 2,048 tokens compared to GPT-4V's 128,000 tokens, we streamlined our prompts to enhance their efficacy within this constraint. The smaller context window in LLaVA necessitates more concise prompts to prevent information overflow, which can lead to a loss of context and unpredictable model performance. This strategy is motivated by observations that LLaVA qualitatively struggled to consider the full context provided in longer prompts, as shown in Figure 2. As an example of our efforts to alleviate these issues, the transition from a lengthy prompt to a more focused prompt can be illustrated as follows:

> **Original Prompt Excerpt (95 Tokens):** "You are asked to predict whether a news article contains misinformation. External sources can better help you make the judgement. Given the text's title, list two questions/phrases/sentences that you would like to search on a public search engine, such as Google. Carefully design your question so that it can return the most helpful results for making your final prediction and reasoning. Please use English to generate your questions..."
>
> **Shortened Prompt Excerpt (44 Tokens):** "To predict if a news article contains misinformation, use external sources. Based on the article's title, formulate two precise questions in English for a public search engine like Google to help verify the information."

Here, a similar idea is communicated in less than half of the tokens.

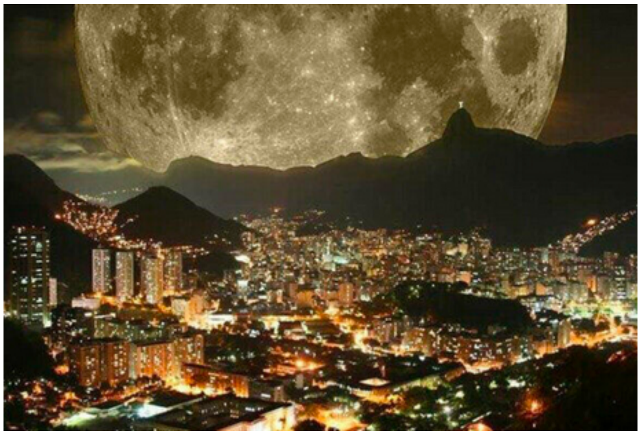


Figure 2: Comparison of LLaVA model performance with long versus short prompts. When given long, complicated prompts, we qualitatively observe that LLaVA loses focus on certain aspects of the problem:

1. When prompted "Did this photo come from a camera directly, or has it been digitally manipulated with post-processing?" LLaVA correctly identifies that "this appears to be a digital manipulation... [as] the proportions are not accurate".
2. When asked if the image is real, but also given evidence collected from various internet sources, LLaVA claims "External knowledge confirms that this photo was taken by Leonardo Sens ... the image is a real photograph capturing an interesting perspective rather than being digitally manipulated".

#### 2.4.2 Chain of Thought (CoT) Prompting

We also implemented CoT prompting techniques, following the strategies suggested by Kovach and Rosenstiel (6), where prompts are structured to mimic human critical thinking patterns. This method is particularly relevant for enhancing the model's alignment with human fact-checking processes, where systematic and structured analysis is paramount. By guiding LLaVA through a logical sequence of thought, the model is better equipped to handle complex fact-checking tasks. An example of this approach is shown below, demonstrating how prompts were reformatted to encourage step-by-step thinking:

> **Original Prompt Excerpt:** "Based on the references and the definition of predefined categories, please classify the news into one of the six predefined categories."
>
> **CoT Prompt Excerpt:** "Use the provided references… please output your concise, step-by-step reasoning for classifying the post... "

### 2.5 Explicitly Structured Approach

Step-by-step reasoning is known to enhance the performance of large language models (LLMs) (6). Also, recall that LLaVA shows qualitative issues when processing complex, lengthy prompts, as shown in Figure 2. In our efforts to inject these philosophies into LEMMA, we employed both CoT prompting and shorter prompts in the previous section 2.4. In this section, we sought to address both of these issues by modifying the structural framework of LEMMA itself to explicitly enforce systematic reasoning.

To achieve this, we removed the "Refine Prediction" block from LEMMA. In it's place, we asked the LVLM a series of seven simple yes/no questions, inspired by the guided reasoning in Kovach and Rosenstiel's work (6). These questions were intended to isolate simple subproblems of the misinformation detection task, and enforce a human-like reasoning process. These questions were as defined in Table 1:

| No. | Question |
|---|---|
| 1 | Is the event described in the post real, with the timeline corroborated by your collected sources? |
| 2 | Is this a real photo? |
| 3 | Is the post intended to be humorous or satirical? |
| 4 | Is there a false connection between the image context and the image text? |
| 5 | Are there plausible alternative explanations for the content of the image? |
| 6 | Is the narrative in the post supported by the accompanying image? |
| 7 | Relying on collected sources, is the story presented in the post believable as true? |

Table 1: Structured Approach Questions

The model was asked to evaluate these responses, with each LLaVA component using a temperature of 0.5. After gathering the responses, we extracted the logits associated with either "yes" or "no" in the answers provided by LLaVA. Temperature was increased as it gave more varied logit values in the responses; low temperature models tended to assign very high confidence logits. Cases where neither "yes" nor "no" was output, or both were output, were assigned a logit of $0.5$. These logits were then used as features to train a logistic regression model.

The logistic regression model was trained to predict the misinformation using the logits derived from LLaVA's responses. We utilized a subset of 300 samples from our dataset for training purposes, with the model's performance subsequently evaluated on an additional set of 470 samples. The full pseudocode is summarized in Algorithm 1.

### 2.6 Fine-Tuning

LLaVA-NeXT's struggles with misinformation detection may be due to the task being outside of its training distribution. While the tasks it was given were selected in an attempt to imbue it with a

**Algorithm 1** Process for Misinformation Detection using LLaVA

```
for each post $i$ in Dataset do
    for each question Q1 to Q7 in Table 1 do
        Use LEMMA zero-shot, external knowledge, questions, title, and evidence extraction blocks to collect background knowledge.
        Ask the questions to LLaVA model with temperature = 0.5
        Extract logit values for the responses
        if neither "yes" nor "no" is output, or both are output then
            Assign a logit value of 0.5
        end if
        if $i < 300$ then
            Store the logits from responses of each question and labels
        end if
        if $i == 300$ then
            Train the logistic classifier on collected logits and labels.
        end if
        if $i > 300$ then
            Predict the label using the logistic classifier, and recalculate accuracy, precision, recall, and F1 score.
        end if
    end for
end for
```

general sense of reasoning (7), little research has yet been done as to the effectiveness of the model to generalize to out-of-distribution tasks. In particular, tasks like misinformation detection are somewhat subjective in nature, and therefore LLaVA's lack of social alignment via Reinforcement Learning from Human Feedback (RLHF) (4) leaves it uniquely unprepared for the task at hand.

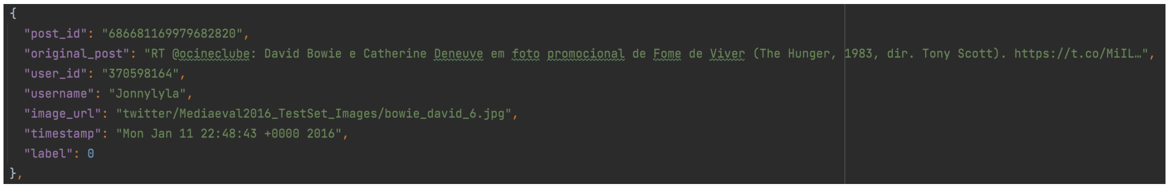
In an attempt to remedy this, we perform further fine-tuning on LLaVA via instruction-tuning data generated with GPT-4V. Due to our lack of access to the exorbitant resources necessary to train LLaVA locally, we perform all experiments and evaluations in this section using the model-hosting API Replicate (13). Due to limitations in availability of fine-tuning functionality, LLaVA-NeXT was not available, so we used the best available fine-tunable model which was LLaVA-1.5. We perform LoRA fine-tuning with a learning rate of $10^{-4}$ on 77 examples (the first 10% of the data time-wise to avoid time-wise data contamination) and evaluate LEMMA with both the base and fine-tuned models. Both the learning rate and temperature (0.5) for this section were selected because they were the default for Replicate's API. Our lack of funding prevented us from running many experiments to try different settings, so we chose to default to Replicate's expertise in this instance.

To quantify the effects of fine-tuning, we evaluate three new versions of LEMMA:

1. A baseline using base LLaVA-1.5 for all components of LEMMA. (We still use GPT-3.5Turbo text-only for information extraction.)
2. Using the fine-tuned model only for LEMMA's first prediction before evidence retrieval, and base LLaVA-1.5 for all other components
3. Using the fine-tuned model for all components of LEMMA

We delineate in this way because all instruction tuning data is derived from zero-shot prediction, and therefore we want to check if this fine-tuning can generalize to the other, related tasks with LEMMA such as question generation and reason modification. The decision to use just initial prediction from GPT-4V was made to better isolate the effects of fine-tuning to a specific task, and initial prediction in particular was chosen because it seemed via observation to be the one base LLaVA struggled with the most: many of our earlier examples, including our baseline, saw LLaVA make the same initial guess for almost every example in the dataset.

All examples are evaluated on 100 new examples (not from those used to generate instruction tuning), again due to limited resources (Replicate's API is much more expensive than running locally). We note that there may be some overlap in topics between the different examples in the dataset, so this



```
{
  "post_id": "686681169979682820",
  "original_post": "RT @ocineclube: David Bowie e Catherine Deneuve em foto promocional de Fome de Viver (The Hunger, 1983, dir. Tony Scott). https://t.co/MiIL…",
  "user_id": "370598164",
  "username": "Jonnylyla",
  "image_url": "twitter/Mediaeval2016_TestSet_Images/bowie_david_6.jpg",
  "timestamp": "Mon Jan 11 22:48:43 +0000 2016",
  "label": 0
},
```

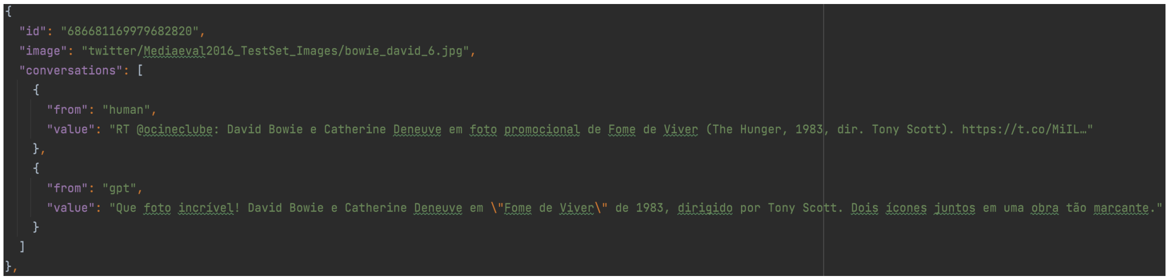
Figure 3: Example of dataset instance from LEMMA codebase

fine-tuned LLaVA may benefit from specific world knowledge incorporated in the GPT instruction tuning data. See the section below for further details about the generation of instruction tuning data.

### 2.6.1 Instruction Fine-Tuning

The type of fine-tuning done for explicitly is instruction tuning. In our case, we fine-tuned the instruction prompt to incorporate both human as well as GPT-4V generated text in order to de-clutter the input prompt for LLaVA. We did this with the input with hopes to decrease noise and mitigate within-prompt contradictions to form a more consistent and streamlined prompt-engineering setup.

An example of the same dataset entry reprocessed for instruction fine-tuning can be seen in Figures 3 and 4.



```
{
  "id": "686681169979682820",
  "image": "twitter/Mediaeval2016_TestSet_Images/bowie_david_6.jpg",
  "conversations": [
    {
      "from": "human",
      "value": "RT @ocineclube: David Bowie e Catherine Deneuve em foto promocional de Fome de Viver (The Hunger, 1983, dir. Tony Scott). https://t.co/MiIL…"
    },
    {
      "from": "gpt",
      "value": "Que foto incrível! David Bowie e Catherine Deneuve em \"Fome de Viver\" de 1983, dirigido por Tony Scott. Dois ícones juntos em uma obra tão marcante."
    }
  ]
},
```

Figure 4: Example of dataset instance to be used for fine-tuned model

## 3 Evaluation

We perform several comparisons of models and provide metrics of their performance on evaluating the 770 tweet test section described in Section 2.1. We provide overall accuracy metrics in addition to precision, recall, and F1-Score for both the rumor and non-rumor categories. We select these as they are the metrics used to evaluate the original LEMMA model (12), and because F1-score in specific is often used for the task of misinformation (14). We provide brief analysis, leaving in-depth qualitative study for Section 4.

### 3.1 Baseline

We find that LLaVA-NeXT performs substantially worse than GPT-4V in both the 0-shot and evidence-backed cases. This is in large part due to a bias in LLaVA's initial guessing, wherein it naively predicted all news to be "Real". While modifications to prompts could flip this to at least always predict "Fake", the majority label in the dataset (examples are split 60/40), we kept this as is for our baseline experiment to get a better comparison with Xuan et al.'s results with GPT-4V. Note here just how poorly LLaVA does: since the dataset is split 60/40, even with evidence LLaVA fails to outperform a naive classifier which guesses "Fake" every time.

Note also that these numbers vary from LEMMA's LLaVA zero-shot baseline. This is for two reasons. Firstly, their study used a different prompt for zero-shot LLaVA and zero-shot GPT-4V. As we were replicating the GPT-4V LEMMA system, our prompt matches the GPT-4V prompt, and is not exactly the same as their zero-shot LLaVA prompt. Secondly, while we use LLaVA-1.6, LEMMA used LLaVA 1.5. Considering these differences, our baseline is used to ensure valid comparisons going forward.

| Model | | Rumor | | | Non-Rumor | | |
|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Prec. | Rec. | F1 | Prec. | Rec. |
| GPT-4V 0-shot | .637 | .747 | .578 | .651 | .529 | .421 | .469 |
| GPT-4V LEMMA | .824 | .835 | .727 | .777 | .818 | .895 | .854 |
| LLaVA 0-shot | .418 | .039 | 1.00 | .020 | .582 | .411 | 1.00 |
| LLaVA LEMMA | .580 | .714 | .600 | .884 | .208 | .444 | .136 |

Table 2: Baseline comparison between Xuan et al.'s results with GPT-4V and our replication of their pipeline using GPT-4V, with (LEMMA) and without (0-shot) evidence retrieval

### 3.2 Time-Awareness

When limiting LEMMA with LLaVA to only accessing evidence available before the time at which a claim was made, there is a noticeable decrease in performance. This is likely due to a previous reliance on fact-checking articles which can often be released several days or weeks after popular claims have begun to spread, limiting the extent they would be practically usable in a real-world misinformation detection environment. Accordingly, we restrict access similarly for all experiments moving forward. More explicitly, the RAG system is not allowed to access articles published after the claims in question. Note that, due to this, the time-aware version of LLaVA in Table 2 will be used as a comparison baseline moving forward.

| Model | | Rumor | | | Non-Rumor | | |
|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Prec. | Rec. | F1 | Prec. | Rec. |
| Baseline | .580 | .714 | .600 | .884 | .208 | .444 | .136 |
| Time-Aware | .521 | .599 | .599 | .599 | .407 | .407 | .407 |

Table 3: LEMMA w/LLaVA with (baseline) and without (Time-Aware) access to articles published after the time the claim in question was posted

### 3.3 Prompting Methods and Explicitly Structured Approach

By far the most effective strategies were a combination of shorter prompts with chain-of-thought prompting, and the explicitly structured approach making use of the model's output logits. We compare both of these models to LEMMA's original results as well as to our baseline, to show that more advantageous methods can help to bridge the gap between instruction tuned models and their teacher counterparts.

| Model | | Rumor | | | Non-Rumor | | |
|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Prec. | Rec. | F1 | Prec. | Rec. |
| LEMMA (GPT-4V) | .824 | .835 | .727 | .777 | .818 | .895 | .854 |
| LEMMA (LLaVA) | .521 | .599 | .599 | .599 | .407 | .407 | 0.407 |
| LEMMA (LLaVA) w/CoT | .725 | .715 | .902 | .592 | .734 | .910 | .615 |
| Structured Approach | .765 | .826 | .830 | .823 | .750 | .747 | .753 |

Table 4: LEMMA w/LLaVA prompting w/short chain-of-thought prompts and a simple logistic regression model (structured approach) based off of LEMMA w/LLaVA's logit outputs.

Note that LEMMA (GPT-4V)'s performance is overestimated here as it is not time-aware. Note also that the structured approach's regression model was trained on 300 examples, leaving 470 remaining to test. The selected training examples were the first 300 chronologically, again ensuring minimal possible access to information or trends from the future impact test performance (though it is worth noting that the model is only passed logits of broad questions about the example, and thus such contamination seems *a priori* unlikely to have any effect.)

### 3.4 Fine-Tuning

We provide evaluations of three models in Table 4:

1. A baseline using base LLaVA-1.5 for all components of LEMMA. (We still use GPT-3.5Turbo text-only for information extraction.)
2. Using the fine-tuned model only for LEMMA's first prediction before evidence retrieval, and base LLaVA-1.5 for all other components
3. Using the fine-tuned model for all components of LEMMA

Due to tight constraints on resources and the steep pricing of the model hosting service used for fine-tuning, the model is fine-tuned on instruction data from just 77 examples and evaluated on 100 more.

| Model | | Rumor | | | Non-Rumor | | |
|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Prec. | Rec. | F1 | Prec. | Rec. |
| LLaVA-1.5 | .570 | .661 | .646 | .677 | .667 | .595 | .758 |
| Fine-Tune (init. pred.) | .510 | .473 | .710 | .355 | .542 | .420 | .763 |
| Fine-Tune (all) | .550 | .641 | .623 | .660 | .412 | .433 | .393 |

Table 5: Evaluation of LEMMA with no fine-tuning (LLaVA-1.5), fine-tuning on the initial prediction (init. pred.), and fine-tuning for all modules (all)

Overall, we find little to no improvement from the use of fine-tuning, with both models actually performing even worse than the LLaVA-1.5 baseline. This may be due to more variance in the initial prediction, keeping LEMMA from benefiting from its naive system of always predicting yes.

To keep the scope of the work minimal and reduce extraneous cost, we elected not to evaluate the fine-tuned model used with the successful prompting and structural approaches discussed in Section 3.3. We do not believe that these results rule out the potential of such a combination of techniques being successful.

### 3.5 Overall Results

See tables 5 and 6 for a comparison of the most important models.

| Test | Time-Aware | Sample size | Accuracy |
|---|---|---|---|
| GPT-4V LEMMA | No | 770 | 81.6% |
| LLaVA LEMMA | Yes | 770 | 52.1% |
| LLaVA LEMMA w/CoT | Yes | 770 | 72.5% |
| Structured Approach | Yes | (300 trained, 470 evaluated) | 76.5% |
| LLaVA Fine-Tuned (all) | Yes | (77 trained, 100 evaluated) | 55.0% |

Table 6: Overall evaluation metainfo and accuracies

| Model | | Rumor | | | Non-Rumor | | |
|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Prec. | Rec. | F1 | Prec. | Rec. |
| GPT-4V LEMMA | .824 | .835 | .727 | .777 | .818 | .895 | .854 |
| LLaVA LEMMA | .521 | .599 | .599 | .599 | .407 | .407 | .407 |
| LEMMA (LLaVA) w/CoT | .725 | .715 | .902 | .592 | .734 | .910 | .615 |
| Structured Approach | .765 | .826 | .830 | .823 | .750 | .747 | .753 |
| LLaVA Fine-Tuned (all) | .550 | .641 | .623 | .660 | .412 | .433 | .393 |

Table 7: Comparison of Various Models

## 4 Discussion

We find that prompting methods and more symbolic, structured approaches are most effective in improving upon Xuan et al.'s LEMMA pipeline. We predict that this is likely in part due to the lack of translatability between models: LEMMA was originally designed for GPT-4V, meaning that its construction might appeal to strengths and weaknesses in GPT-4Vs functioning which might differ from LLaVA's strengths and weaknesses.

In particular, while brief experiments found chain-of-thought to be unhelpful toward the task of misinformation detection for GPT-4V, its use led to an increase in accuracy of more than $0.15$ with LLaVA. There are several possible explanations for this. One is that GPT-4V did not benefit from chain-of-thought as it usually might due to a reliance on built in world-knowledge. As GPT-4V is a much larger model, it has more ability to store specific knowledge about the world, and therefore needs to rely less on its reasoning for some examples. For instance, we anecdotally observed that GPT-4V was about to recall facts about images in the dataset even when given them with no other context. It could at times tell us who posted the image, whether or not it was real, and who fact checked. In this sense, our comparison here might be a more fair evaluation of how LVLMs' misinformation detection abilities might actually generalize. Whereas GPT-4V might have been relying on its own external knowledge, a method which necessitates the claim being evaluated to have existed before GPT's training cutoff, LLaVA is forced to actually reason about the image as if it had never seen it before (since it likely hasn't). Qualitatively, LLaVA was much more likely to reason intelligibly about the example in question when prompted with chain-of-thought, as shown in Table 8.

| **Caption:** "Planet Earth wrapped up a blanket of clouds, taken by the hubble telescope..." | |
|---|---|
| **Baseline:** The image shows a planet, presumably Earth, wrapped in clouds as described by the caption. There is no inconsistency or misinformation detected between the text and the image. The Hubble Telescope has indeed taken images of planets like this one. Therefore, I am certain that there is no misinformation present here. | **Chain-of-thought:** Let's think step by step: 1. The original text mentions "Planet Earth wrapped up in a blanket of clouds, taken by the Hubble telescope." This statement is not entirely accurate based on the references provided. 2. The first reference states that the image was NOT taken by Hubble... Therefore, based on the available evidence, the most likely category for this post would be "Manipulated Content" or "False Connection"... |

Table 8: Chain-of-thought prompting shows much clearer reasoning and interacts with the evidence in a more meaningful way than LEMMA's original prompting methods

Our examination of the comparative efficacy between CoT and explicitly structured approaches in retrieval-augmented systems reveals nuanced differences in outcome diversity. Notably, the structured approach seems to enable a broader spectrum of responses. When utilizing a CoT prompt, LLaVA typically generates coherent responses, yet these responses often lack variety as a result of their each being an extension of the last. In contrast, the structured method divides each LLaVA evaluation into discrete, simple questions, each independent of the others. This modularity seems advantageous, especially in scenarios where the data presents contradictory elements—such as a post that seemingly forms a false connection but aligns with a realistic timeline and uses a genuine photograph. Here, the structured approach appears to create richer features, allowing logistic regression models to more effectively discern and categorize the underlying trends. An example evaluation is shown in Table 10.

Our investigation of time-awareness and its stark impact on results shows another way that studies of retrieval-augmented systems can be biased. This shows the importance of treating misinformation datasets as what they are: time-sensitive data, the patterns and distribution of which may change over time. Future work should take this into consideration to avoid overstating the ability or generalizability of their results. We also observe that, by discarding examples with no known date, we may have exaggerated the effects of time-awareness. Still, we defend this as being preferable to the reverse decision, and observe qualitatively in logs that articles with no discernible date were not retrieved noticeably often.

We observe that the lack of ability of our fine-tuned model to improve results points to the ways in which instruction tuning as a technique is limited. Qualitatively, we observe that the fine-tuned model emulates the writing style and structure of GPT well, but does not seem to develop its reasoning abilities alongside that improvement, with an example being shown in Table 9. While we fine-tuned on a quite small amount of data, limiting the extent to which we can make claims about the technique at-large, the observation that performance in fact decreased suggests that common perceptions of fine-tuning as a panacea for task-specific performance are at least somewhat misguided.

| **Caption:** "Unbelievable Shot Of The 2012 Supermoon In Rio de Janeiro.." | |
|---|---|
| **Baseline:** The image shows a large moon in the sky above Rio de Janeiro, which is consistent with what one would expect from an image of the 2012 Supermoon. There are no obvious signs of manipulation or inconsistencies between the text and the image that suggest misinformation. Therefore, I label this post as R̈eal.̈ The moon's size relative to the cityscape indicates it is indeed a supermoon event captured in the photo. The lighting conditions also match what... | **Finetuning:** The image shows a city skyline at night with a large moon in the sky. The text claims that the image is of the 2012 Supermoon in Rio de Janeiro. The moon appears to be quite large and prominent in the sky, which is consistent with the text's description of a supermoon. However, without additional context or verification, it is not possible to confirm the exact date or location of the image, which could be misinformation. Overall, the consistency between the text and the image is not definitively clear, as there is no clear indication of the date or location in the image itself. |

Table 9: Finetuning models are less confident at labeling examples as true

| **Title:** "Paris Concert Hall Hostage Situation: Emergency Response and Public Safety" | |
|---|---|
| **CoT Output:**<br>Let's think step by step. The image shows a nighttime scene on a city street with parked cars, pedestrians walking along the sidewalk, and what appears to be an overhead structure or awning. There is no visible sign of distress or emergency in the image that would suggest a hostage situation as described in the text. Now let's consider the text: "LATEST: 'Ongoing situation' in Paris; ... This statement is inconsistent with the image because there are no signs of an emergency or hostage situation visible in the photo. The text and image do not align logically. I am confident that the accompanying text seems to be inaccurate based on what can be seen in the photo. | **Structured Approach Outputs:**<br>**Is it a real photo?** Yes, this is a real photo. The image shows an outdoor scene at night with people and various objects that suggest it could be a public space or possibly near a business establishment...<br>**Is this a false connection?** Yes, there is a false connection between the image content and the image text. The context provided by the text does not match the visual content of the image...<br>**Is it believable?** Yes. The image depicts a chaotic scene, and the text refers to a hostage situation at the Bataclan concert hall in Paris, which is consistent with well-documented events from November 13, 2015... |

Table 10: Example CoT response compared to predictions from the structured approach. Note that the CoT output seems to form reasoning consistent with its prediction (False Connection), while the structured approach outputs are independent (predicting both a False connection and believable output)

## 5 Conclusion

We provide a quantitative and qualitative characterization of the ability of instruction-tuned vision-language models to detect multimodal misinformation via retrieval-augmented generation. Specifically, we show that these models struggle compared to their larger counterparts, and reason that this may be due to their reduced world knowledge and limited exposure to the given task. We also demonstrate that retrieval-augmented generation tasks based on web search may appear more performant than reality due to having access to evidence which would not have been available if the model was running in real time. This limitation is yet another way that the results of misinformation detection models can be a mirage, appearing promising but failing to generalize to their intended task.

We then provide several methods for improving the performance of these models in the face of these limitations, including modified prompting strategies and task-specific fine-tuning. While demonstrative of key concepts and the potential for large improvements, our experiments serve only as a brief foray into a wide range of possible modifications and experiments to build on LEMMA's original work.

## 6 Division of Work

Marco and James performed much of the original setup and beginning prompting investigations, then Marco continued to work on prompting and the structured approach while James helped Daniel perform the model fine-tuning. Each group member was involved with each section of the project on a conceptual/discussion level, whereas the division above reflects implementation work division.

## References

[1] Duckduckgo. `https://duckduckgo.com/`, 2024.

[2] Facebook: From election to insurrection. Tech. rep., Avaaz, 2021.

[3] Bang, Y., Ishii, E., Cahyawijaya, S., Ji, Z., and Fung, P. Model generalization on covid-19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1* (2021), Springer, pp. 128–140.

[4] Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems 26* (2013).

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[6] Kovach, B., and Rosenstiel, T. *Blur: How to know what's true in the age of information overload.* Bloomsbury Publishing USA, 2011.

[7] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[8] Ma, J., Gao, W., and Wong, K.-F. Detect rumors in microblog posts using propagation structure via kernel learning.

[9] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P.,

Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

[10] Suprem, A., Vaidya, S., Ferreira, J. E., and Pu, C. Time-aware datasets are adaptive knowledgebases for the new normal. *arXiv preprint arXiv:2211.12508* (2022).

[11] Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., and Liang, Y. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents, 2023.

[12] Xuan, K., Yi, L., Yang, F., Wu, R., Fung, Y. R., and Ji, H. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation, 2024.

[13] yorickvp. Replicate llava 13b api. `https://replicate.com/yorickvp/llava-13b`, 2024.

[14] Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., and De Choudhury, M. 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* (2023), pp. 1–20.

# A Reproducibility Checklist

Code is available at the following GitHub Link:

Our codebase is available on GitHub: `https://github.com/mconati/LEMMA`

This paper:

1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) **Yes.** see Algorithm 1 in Section 2.5.
2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) **Yes.**
3. Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) **Yes.**

Does this paper make theoretical contributions? **No.**

Does this paper rely on one or more datasets? **Yes.**

If yes, please complete the list below:

1. A motivation is given for why the experiments are conducted on the selected datasets **Yes.** See Section 2.1.
2. All novel datasets introduced in this paper are included in a data appendix. **NA.**
3. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **NA.**
4. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. **NA.**
5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. **Yes.**
6. All datasets that are not publicly available are described in detail, with an explanation why publicly available alternatives are not scientifically satisfying. **NA.**

Does this paper include computational experiments? **Yes.**

If yes, please complete the list below:

1. Any code required for pre-processing data is included in the appendix. **Yes.**
2. All source code required for conducting and analyzing the experiments is included in a code appendix. **Yes.** (Code for different experiments can be found in the different branches of the linked git repository.)
3. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Yes.**
4. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from **Partial.**
5. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. Unfortunately, **No.** While the model-hosting service we used did not make seed setting available, the local hosting software did allow for doing so and yet we forgot to do it! We take this as a learning experience to remember to consider reproducibility earlier in the implementation process next time.
6. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. **Yes.** See the GitHub Readme for more details.
7. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes/partial/no)
8. This paper states the number of algorithm runs used to compute each reported result. **Yes.**
9. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. **No**, since we ran each experiment only once. However, we do provide results pertaining to the distribution of predictions and classes in the form of class-specific precision, recall, and F1-scores.
10. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). **No.**
11. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. **Yes.**
12. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. **Yes.**