# Misinformation with Legal Consequences (`MisLC`): A New Task Towards Harnessing Societal Harm of Misinformation

**Anonymous ACL submission**

## Abstract

*Misinformation*, defined as false or inaccurate information, can result in significant societal harm when it is spread with malicious or even innocuous intent. The rapid online information exchange necessitates advanced detection mechanisms to mitigate misinformation-induced harm. Existing research, however, has predominantly focused on assessing veracity, overlooking the legal implications and social consequences of misinformation. In this work, we take a novel angle to consolidate the definition of misinformation detection using legal issues as a measurement of societal ramifications, aiming to bring interdisciplinary efforts to tackle misinformation and its consequence. We introduce a new task: Misinformation with Legal Consequence (`MisLC`), which leverages definitions from a wide range of legal domains covering 4 broader legal topics and 11 fine-grained legal issues, including hate speech, election laws, and privacy regulations. For this task, we advocate a two-step dataset curation approach that utilizes crowd-sourced checkworthiness and expert evaluations of misinformation. We provide insights about the `MisLC` task through empirical evidence, from the problem definition to experiments and expert involvement. While the latest large language models and retrieval-augmented generation are effective baselines for the task, we find they are still far from replicating expert performance.[1]

## 1 Introduction

Artificial intelligence is advancing with an unprecedented speed, and many emerging problems with profound societal impact need multi-disciplinary research efforts and solutions. Misinformation, broadly defined as *false* or *inaccurate information,* has had a widespread harmful impact. If unaddressed, it will persist and exacerbate systemic problems in our daily life as well as many critical areas (Budak et al., 2024). For instance, conflicting information during the COVID-19 pandemic significantly influenced people's attitudes and behaviors toward preventing viral spread (Enders et al., 2020). In significant economic or political events, misinformation also proves extremely detrimental (e.g., in the form of *fake news*), where malicious actors are motivated to purposely spread false information to manipulate public opinion while an event unfolds (Nyilasy, 2019).

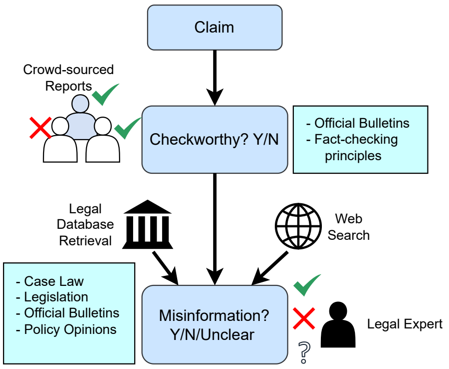


Figure 1: An overview of our proposed task framework for legal misinformation. We obtain crowd-sourced labels of checkworthiness. If a claim is checkworthy, we use legal annotators to annotate potential legal issues of misinformation.

The growing menace of online misinformation underscores the urgent need for regulation, as exemplified by the European Commission's recent action plans.[2] We believe that NLP enabled solutions will play a critical role in mitigating the adversarial affects of misinformation. These solutions require a human-centric approach, with the basic design to ensure the *alignment* between humans and AI, centring on the values and interests of humans (Bai et al., 2022; Pyatkin et al., 2023).

[1] Our code and data are available at [PLACEHOLDER] for replicability.

[2] https://digital-strategy.ec.europa.eu/en/policies/online-disinformation

Collaborations with experts in the social sciences are essential to achieve this goal.

In this work, we take a novel angle to define misinformation with its outcome that can be regularized by laws or regulations, building on legal issues as a measurement of societal ramifications, and aiming to bring interdisciplinary efforts to tackle misinformation and its consequence. Unlike previous work that has focused on factual accuracy or checkworthiness as potential controversy of a topic (Das et al., 2023), we ground our definition in legal literature and social consequence. Our main contributions are summarized as follows:

- We introduce a new task: Misinformation with Legal Consequence (`MisLC`), which leverages definitions from a wide range of legal domains covering 4 broader legal topics and 11 fine-grained legal issues, including hate speech, election laws, and privacy regulations. We advocate a two-step dataset curation approach, utilizing crowd-sourced checkworthiness and expert evaluations of misinformation. We expect our process and discussions could help other similar tasks that need to involve costly domain experts to jointly solve problems with significant societal impact.
- We evaluate the state of the art of the most recent large language models (LLMs) on `MisLC`, by performing a comprehensive study on a wide range of open-source and proprietary LLMs that covers a broad parameter spectrum and varying training data. Two advanced Retrieval-Augmented Generation (RAG) architectures are investigated to detect legally consequential misinformation, involving retrieval from legal document databases and web search, mimicing expert techniques.
- We provide insights about the `MisLC` task through empirical evidence, from the problem definition to experiments and domain expert involvement. After thorough empirical study, we find the existing LLMs perform reasonably well at the task, achieving non-random performance without external resources. Their performance also increases consistently with RAG. However, LLMs are still far from matching human expert performance. Through this work, we urge further exploration in this challenging task with significant societal impact.

## 2 Related Work

Misinformation is a serious issue with significant societal impact, as factual dissonance can cause disorder in peoples' worldviews (Nyilasy, 2019). There have been various works that address separate components of the fact-checking pipeline: identifying checkworthy claims, gathering sources on those claims, and predicting veracity (Das et al., 2023). There is growing interest in addressing the problem with LLMs (Chen and Shu, 2023), and emerging works proposing new methodologies for fact-checking (Pelrine et al., 2023; Pan et al., 2023). However, these works do not consider the legal concept of misinformation.

Generative, or auto-regressive models have recently demonstrated strong proficiency in a wide variety of tasks such as relevance, stance, topics, and frame detection in tweets (Gilardi et al., 2023; Bang et al., 2023). Large Language Models (LLMs) have also demonstrated the ability to capture and memorize a vast amount of world knowledge during pretraining (Guu et al., 2020). However, this knowledge is stored implicitly within their parameters, leading to a lack of transparency for the facts and information generated in their outputs (Rashkin et al., 2023; Manakul et al., 2023). One viable strategy for factual accuracy is giving explicit knowledge from external corpora, or Retrieval-Augmented Generation (Du and Ji, 2022). Some approaches prepend retrieved documents in the input (Guu et al., 2020; Shi et al., 2023). For further related work, please refer to Appendix A.

## 3 Misinformation with Legal Consequences (`MisLC`)

### 3.1 Definition

The `MisLC` dataset $\mathbf{D} \equiv \{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ is composed of $N$ instances, each $\mathbf{d}_i \in \mathbf{D}$ being a tuple $\langle \mathbf{t}_i, E_i, L_i, y_i \rangle$, where $\mathbf{t}_i$ is a piece of text (e.g., a social media article). $E_i$ is a set of external *evidence* documents that can be used to support or refute the text $\mathbf{t}_i$. $L_i \subset \mathcal{L}$ is a subset of *legal issues* from a predefined issue set $\mathcal{L}$. Each legal issue refers to an area of law that can be used to indict or punish misinformation, e.g., *Election Laws*, *Public Mischief*, or *Cyberbullying*. We will discuss the details in this section.

The label of `MisLC` is $y_i \in \{-1, 1, 2\}$, where '2' represents *Misinformation with Legal Consequence* (`MisLC`), '1' for *Unclear*, and '$-1$' for the cases that are not `MisLC` (`Non-MisLC`). *Unclear* is reserved for cases that are impossible to determine a classification when there is insufficient context to make the decision. This label is crucial because in real-life applications, we need to separate them for further legal processing, including collecting more evidence. The details will be further discussed in Section 3.2. The MisLC evaluation is organized in two settings: (1) a binary task, with MisLC as the only positive class of interest, and the other two as negative, and (2) a 3-way classification task, where MisLC and unclear are separate positive classes.

The evidence $E_i$ and legal issues $L_i$ are used by legal professionals to obtain the ground-truth labels for MisLC. A necessary condition of a span of text $\mathbf{t}_i$ being *misinformation with legal consequence* is that it makes a claim, where a claim is defined as "stating or asserting that something is the case, typically without providing evidence or proof."[3] If a expert annotator assesses a claim in $\mathbf{t}_i$ is associated with a legal issue, i.e. it passes its relevant legal tests and does not pass possible defences, this will trigger the MisLC label. Formally, the set of legal issues $L_i = \{\text{test}_j(\mathbf{t}_i, E_i) \wedge \neg\ \text{def}_j(\mathbf{t}_i, E_i)\ \forall\ (\text{test}_j(\mathbf{t}, E),\ \text{def}_j(\mathbf{t}, E)) \in L\}$.

$$y_i = \begin{cases} 2 & |E_i| > 0\ \wedge\ |L_i| > 0 \\ -1, & |E_i| > 0\ \wedge\ |L_i| = 0 \vee \mathbf{t}_i\ \text{ not a claim} \end{cases} \tag{1}$$

**Legal Resources.** Defining misinformation from a legal standpoint is challenging. Misinformation is an umbrella term to capture the act of publishing any form of false or misleading information in a public space. This is reflected in current legal practices; the issue of false or misleading information may fall under multiple distinct area of law. For example, misinformation aimed at a target group can be punished under hate speech laws. We note that there are very few jurisdictions with provisions that directly address misinformation as a separate legal issue. Since the definitions of misinformation are broad, they better serve as an indication of a policy domain rather than a legal category (van Hoboken et al., 2019). Despite concerns about regulating misinformation (Ó Fathaigh et al., 2021), the existing laws have been crafted through extensive discussion to mitigate the harm caused by misinformation to society, reflecting a deliberate and thoughtful approach to a complex issue.

We collaborate with legal annotators to build a text database on the legal definition of misinformation. Our search spans diverse legal areas, including hate speech, consumer protection, election laws, defamation, food and drug safety, and privacy regulations. For specific citations, please refer to Appendix E. We consider the following sources:

- **Legislation** — Written laws that provide rules of conduct regarding misleading or false information. This encompasses both criminal laws, which can lead to incarceration, and civil laws, which can result in fines.
- **Case Law** — The written decisions of judges in higher court cases that have been made on misinformation issues. We focus on Canadian and European cases.
- **Official Bulletins** — Publications from global organizations, including Unicef, the European Union, the United Nations High Commissioner for Refugees (UNHCR), and the Canadian Government, regarding the general definition and identification of misinformation.
- **Policy Opinions** — Publications from reputable policy makers on misinformation and how legal policy should be applied to prevent its harm.

**Legal Issues.** From the legal resources , we compile 11 major **legal issues** to form $\mathcal{L}$. The full set is listed in Appendix E. Each issue has two components: a **legal test** to determine potential violations, and **defences** that can counter allegations . Our legal issues $l \in L$ are mostly differentiated by the topic/nature of the post, as well as a contention between the post's intent, consequences, and truthfulness. One defence is proving a statement as fact — if the actor can establish their statement is true, their post is no longer punitive. However, this largely depends on the legal issue. The *defamation* issue, for example, focuses on the *defamatory* nature that has the ability to lower someone's reputation, as well as the context the claim was uttered. All relevant legal tests $\text{test}(\mathbf{t}, E)$ and defences $\text{def}(\mathbf{t}, E)$ were compiled into a comprehensive annotation guide.

### 3.2 Creation of the MisLC Dataset

As illustrated in Figure 1, we advocate a two-stage data curation process. First, non-legal crowd-sourced annotators discover checkworthy misinformation samples that arouse their suspicion. Second, legal experts annotate the samples and decide relevant legal issues.

**Crowd-sourced Misinformation Checkworthiness.** We first want to assess a layperson's understanding of misinformation. This component does not require legal expertise, but builds the dataset on which legal practitioners can operate. We curated

[3] https://languages.oup.com/google-dictionary-en/

social media data from (Chen and Ferrara, 2023), a large public domain dataset with Twitter data (tweets) regarding the *Russia-Ukraine conflict*. For more details on data processing and data samples, please refer to Appendix B.1.

We collected crowd-sourced annotations on this data for *checkworthiness,* in order to filter samples that are likely to contain legal consequences. The crowd-sourced workers could choose between three labels: Checkworthy, Not Checkworthy, or Invalid/No Claim. We source our definition of checkworthiness from (Das et al., 2023). Additionally, we incorporated indicators of disinformation from an official bulletin released by the Government of Canada. [4] To ensure data quality, we conducted a pre-screening test with a pool of 100 samples using the same instructions as the main task. This pool was labelled by two members of the research team given the annotation instructions. Details of annotator instructions and the prescreening procedure are contained within Appendix B.3. After this screening process, we obtained a pool of eleven Turk workers for annotations. We randomly sample an additional 1,500 tweets from the 4,000 that we had collected, and provided these to our Turk workers in batches of 500 over one month.

**Adversarial Filtering.** We performed a secondary adversarial data filtering step to ensure the data is sufficiently consistent. Compared to previous works (Sakaguchi et al., 2021), we replaced cross-entropy loss with KL divergence over the annotation distribution to model annotator disagreement. We score each sample by its training loss as defined in Equation 2 and Algorithm 1. We perform this filtering three times with $k = 1000$ and retain a set of 711 samples that is consistently retained in each trial. The filtering process biases the label distribution to Checkworthy samples, as shown in Table 1. This complements our intended pipeline where a sample is flagged by laypeople and further investigated by legal annotators, and indicates strongly Checkworthy samples are likely more consistent than ambiguous agreement. Further details are discussed in Appendix B.4.

$$\mathcal{L}(\mathbf{t}_i, y_i) = D_{KL}(f(\mathbf{t}_i), \text{softmax}(y_i)) \qquad (2)$$

**Algorithm 1** Our adversarial filtering process.

```
Require: Dataset $\{\mathbf{t}_i, y_i\}_{i=1}^{n} \in X$, target dataset size $k$
Encode all samples as the last embedding layer $f(\mathbf{t}_i) = E_{LM}(\mathbf{t}_i)[-1]$
Apply softmax to all $y_i \in X$
Initialize $X' = X$
while True do
    Train a linear classifier $f(t)$ on $X'$
    for $(\mathbf{t}_i, y_i) \in X'$ do
        $s_i = \mathcal{L}(\mathbf{t}_i, y_i)$
    end for
    $\tau_\mu = \mu(score)$
    $S = \{(\mathbf{t}_i, y_i) \in X' | s_i > \tau_\mu\}$
    if $|X' \setminus S| < k$ then
        break
    else
        $X' = X' \setminus S$
    end if
end while
```

**Legal Annotations.** We collaborated with eleven law researchers to annotate our dataset. The law researchers are graduate students in their first and second year . They performed 2-3 hours of annotations per week as part of a practicum course, and received credits as compensation. Each expert is provided a document summarizing the legal tests and defences $\in L$. The legal experts first determined whether any claims in a sample, or the sample in its entirety, qualify as misinformation by selecting one of the following three options: `yes`, `no`, or `unclear`. After this preliminary step, the legal experts identify whether the sample raises any potential legal issues. If it does not, the annotators can then specify whether this is due to an available defence (`defence`) or a lack of factual claims (`noClaim`). Each sample is annotated three times, and we obtain an overall label via majority voting. We also decide the legal issues $L_i$ for a sample $\mathbf{t}_i$ via majority voting. The nominal Krippendorff's Alpha for the legal annotators is 0.441, while the minimum recommended threshold for reliable data is 0.667 (Krippendorff, 2018). However, this Krippendorff's Alpha is consistent with previous works in legal task datasets (Thalken et al., 2023). This indicates greater subjectivity in legal tasks, possibly due to their complexity and opportunity for interpretation.

As shown in Figure 2, the most relevant legal issues for our data to be Freedom of Expression, followed by closely by Defamation. Next, there are Election laws, the criminal offenses of Cyberbullying and Public Mischief, and Hate Speech.

[4] https://www.canada.ca/en/campaign/online-disinformation.html

| Label | Count (%) |
|---|---|
| Total Dataset Size | 711 (100) |
| Checkworthy | 650 (91.4) |
| MisLC | 93 (13.1) |
| Non-MisLC | 540 (75.9) |
| Unclear | 78 (11.0) |
| Evidence available | 363 (51.0) |
| noClaim | 304 (42.8) |
| defence | 242 (34.0) |

Table 1: Statistics on our dataset, including total dataset size, the number of crowd-sourced checkworthy samples, label distribution for MisLC, and special labels from legal annotations. Each sample can have one ground truth label (MisLC, Non-MisLC, or Unclear).

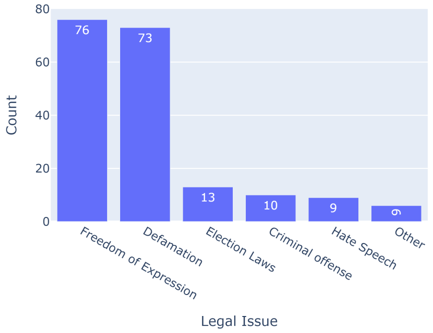


Figure 2: The most frequent legal issues that appear in our dataset.

Our label distribution is summarized in Table 1. While checkworthiness had a positive rate of 91.4% (650), only 13.1% (93 samples) of the dataset has some possible legal violation for misinformation. Additionally, there were a substantial number of Unclear samples (11.0%, or 78). These are samples with unclear context or implications that annotators felt could not be fact-checked, e.g. “we all know what he did.” In the context of our formal definition, this implies the evidence $E_i$ is non-existent, or $|E_i| = 0$. Examining the samples that were checkworthy but not a legal violation, there are a few recurring themes:

*The claim is supported by a reputable source after fact-checking.* We explicitly instructed the crowd-sourced workers to ignore truthfulness of a statement, so this is an expected outcome. This also demonstrates the importance of identifying $E_i$.

*The claim was deemed to be an opinion.* A key component of the crowd-sourced annotator instructions, sourced from a bulletin by the Canadian Government on disinformation, was whether or not a claim invoked an emotional reaction. There appears to be a subtle distinction between an outrageous claim and a personal/political opinion not captured in the crowd-sourced annotations. After manually inspecting some annotations, we found that the annotators sometimes did not investigate a claim if it was combined with opinion.

**Human Expert Performance.** We also calculate an approximation of human performance on our task as an upper bound. First, we assign a random number to each annotator and retrieve all of their individual annotations. For statistical significance, we only retain experts that performed more than 50 annotations. Next, treating their annotations as predictions, we calculate their performance against the majority vote label. We include the mean expert performance for reference in Table 3.

## 4 Models

### 4.1 Experimental Setup

Along with our dataset, we present a comprehensive set of baselines evaluating the performance of state-of-the-art LLMs on detecting misinformation with legal consequences. We examine a wide range of both proprietary and open-source LLMs: GPT-4o[5], GPT-3.5-turbo (Ouyang et al., 2022), Llama2-(7b, 13b, 70b) (Touvron et al., 2023), Llama3-(8b, 70b)[6], Mistral-7b (Jiang et al., 2023a), and Solar-10b (Kim et al., 2023). We choose Llama 2 and 3 to isolate the effect of model size, since these suites of models are trained with the same method at various parameter counts. We compare this suite to three open-source models trained on various combinations of fine-tuning, instruction tuning, and Proximal Policy Optimization (PPO) or Direct Policy Optimization (DPO). The Solar-10b we test combines two checkpoints of Solar (Kim et al., 2023): Solar-Instruct, trained with instruction tuning, and OrcaDPO. Please refer to Appendix C and Appendix C.1 for further details on the models and experimental settings.

**Evaluation Method.** The models are first prompted to classify misinformation based on $\mathbf{t}_i$ without any external knowledge, purely based on their understanding of misinformation along with some evidence $E_i$ potentially available in their parametric knowledge. Intuitively, this should be equivalent to the crowd-sourced annotators, and *we do not expect good performance.*

Our prompt template can be found in Appendix C.2. We ask the model to only output one of three keywords: misinformation, factual, or

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

[6] https://ai.meta.com/blog/meta-llama-3/

`unsure`. Then, we search the generated text for one of these keywords. If none of these keywords are present, we count the generation as an *error*, and report Error Rate (ER) for each model. Errors are converted into a `Not MisLC` prediction, i.e. label -1. We also report Binary F1 (Bin-f1) as performance in the binary task setting, and Macro- and Micro-f1 (Ma-f1, Mi-f1) for 3-way classification.

### 4.2 Retrieval Enhanced Pipeline

LLMs have a significant amount of world knowledge, but our task of misinformation with legal consequences relies on legal material that likely does not exist in their pre-training data. As discussed in Section 3.1, our ground truth labels are not just determined by the input text $\mathbf{t}_i$, but also the relevant legal issues $L_i$ and evidence $E_i$. We use RAG to introduce knowledge from our legal literature, as well as to retrieve potential evidence via web search, in order for the model to receive the same information as our legal annotators.

**RAG Methods.** We employ a retrieval-augmented approach for our misinformation detection pipeline. Generally speaking, given a document corpus $\mathcal{C}$ and a retrieval system $\mathcal{R}_\mathcal{C}$ that can retrieve most related documents to the input query $q$ from corpus $\mathcal{C}$, RAG can be formulated as $p(w_1, ..., w_n) = \prod_{i=1}^{n} p\left(w_i | w_{<i}, \mathcal{R}_\mathcal{C}\left(w_{<i}\right)\right)$, where $w_{<i}$ is the sequence of tokens preceding $w_i$, i.e. $\mathbf{t}_i$ (Ram et al., 2023). In this work, we experiment with two state-of-the-art RAG methods. We choose these methods as they do not require pretraining or fine-tuning LMs, which can be expensive due to the large LM sizes. These methods also do not require access to the LMs layers and weights.

In-Context RALM (***IC-RALM***) (Ram et al., 2023) uses the given input $w_{<i}$ as a query to retrieve a document, and prepends the document to the prompt to generate the output. In this approach, the retrieval is triggered at fixed generation intervals, or retrieval strides $\delta$. To avoid information dilution with long queries, the query is limited to the last $\ell$ tokens of the $w_i$. More formally, IC-RALM is formulated in Equation 3, where $q_j^{\delta,\ell} = w_{\delta.j-\ell+1}, ..., w_{\delta.j}$ and $[a; b]$ denotes the concatenation of strings a and b.

$$p(w_1, ..., w_n) = \prod_{j=0}^{n_\delta - 1} \prod_{i=1}^{\delta} p\left(w_{\delta.j+i} \middle| \left[\mathcal{R}_\mathcal{C}(q_j^{\delta,\ell}); w_{<i}\right]\right) \tag{3}$$

***FLARE*** (Jiang et al., 2023b) generates a temporary sentence $\hat{s}$, where $p(\hat{s}) = \prod_{i=1}^{m} p\left(w_i | w_{<i}\right)$, and then chooses whether to regenerate the sentence with retrieval based on model confidence, i.e. the minimum token probability in the sentence. This is formulated in Equation 4, where $\theta$ is the threshold parameter. Moreover, FLARE formulates the regenerated sentence $s'$ as $p(s') = \prod_{i=1}^{m} p\left(w_i | \left[\mathcal{R}_\mathcal{C}(\mathrm{qry}(w_{<i})); w_{<i}\right]\right)$. The query formulation function $\mathrm{qry}(\cdot)$ generates retrieval queries based on the lowest confidence token spans and by masking low confidence tokens. We adapt their implementation to share the same BM25 indexing and retrieval as IC-RALM. Please refer to Appendix C.3 for further implementation details.

$$s = \begin{cases} \hat{s} & \text{if all tokens of } \hat{s} \text{ have probs } \geq \theta \\ s' & \text{otherwise} \end{cases} \tag{4}$$

**Legal Database.** To align language models to our legal issues, we build a database using the full text of the documents compiled in Section 3.1. We collect 27 documents with an average length of ≈ 24,000 tokens and the maximum being ≈ 96,000 tokens. Having such long documents in the database might cause a few problems: (i) the text chunks are significantly longer than the context window of some LLMs, and (ii) most parts of the text chunk are irrelevant to the query. To this end, we perform a process to split the database into small, yet coherent, text chunks. Please refer to Appendix B.2 for further processing steps.

**Web Search.** A crucial component of our legal tests is the availability of evidence $E_i$ for a piece of text $\mathbf{t}_i$. We query the Google Custom Search API[7] set to retrieve from the entire internet, using the same query we use for our legal database retrieval. One issue is that web search does not return results if there are no sufficiently relevant findings — we test 100 samples of our dataset and find this occurs for 26.5% of FLARE queries and 37.9% of RALM queries. The web search returns various metadata such as the website link, the title, and the most relevant snippet from the webpage. We concatenate the snippets of the first result for each query and insert them into the prompt. We acknowledge this is not the most effective method — there are many works on algorithms to iteratively retrieve evidence (Das et al., 2023). We urge further exploration of evidence gathering pipelines for future work.

[7] `https://developers.google.com/custom-search/v1/overview`

| Model | No Retrieval | | | | IC-RALM (Legal) | | | | FLARE (Legal) | | | | FLARE (Web) | | | | FLARE (Legal+web) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1 ↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1 ↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ |
| `GPT-3.5-trb` | 30.4 | 19.3 | 45.8 | 0.0 | 24.1 | 12.0 | 39.8 | 0.0 | 29.7 | 16.1 | **48.7** | 0.0 | 30.5 | 17.6 | **49.1** | 0.0 | 31.1 | 17.9 | **49.6** | 0.0 |
| `GPT-4o` | 28.7 | 23.2 | 43.5 | 0.0 | **35.8** | **28.5** | **50.9** | 0.0 | 32.3 | **25.8** | 46.7 | 0.0 | **34.5** | **26.2** | 47.4 | 0.0 | **37.7** | **28.0** | 46.7 | 0.0 |
| `Mistral-7b` | 27.9 | 17.2 | 42.8 | 6.8 | 25.9 | 19.1 | 39.3 | 0.0 | 24.5 | 21.2 | 41.0 | 0.0 | 21.7 | 21.6 | 44.7 | 0.0 | 16.7 | 18.3 | 42.2 | 0.0 |
| `Llama2-7b` | 21.0 | 11.7 | 34.7 | 24.1 | 23.1 | 11.6 | 38.8 | 0.0 | 23.2 | 12.8 | 38.8 | 0.0 | 22.9 | 11.5 | 38.6 | 0.0 | 23.2 | 12.8 | 38.8 | 0.0 |
| `Llama3-8b` | 30.7 | 18.0 | 48.2 | 0.0 | 0.0 | 9.9 | 38.8 | 0.0 | 27.2 | 13.6 | 43.0 | 0.0 | 31.1 | 18.0 | 48.1 | 0.0 | 25.3 | 13.8 | 41.6 | 0.0 |
| `Solar-10b` | 27.7 | 14.9 | 31.1 | 32.5 | 28.6 | 22.8 | 39.1 | 3.8 | 27.1 | 21.7 | 41.8 | 1.7 | 32.6 | 22.9 | 44.5 | 4.4 | 28.5 | 21.3 | 40.8 | 2.7 |
| `Llama2-13b` | 22.0 | 11.0 | 31.8 | 56.1 | 21.7 | 17.6 | 39.6 | 0.1 | 22.4 | 17.3 | 38.9 | 0.0 | 23.4 | 19.2 | 39.0 | 0.0 | 23.0 | 15.6 | 39.0 | 0.0 |
| `Llama2-70b` | 23.1 | 11.5 | 38.9 | 0.0 | 23.2 | 11.6 | 38.8 | 0.0 | 25.0 | 13.3 | 39.9 | 0.0 | 25.2 | 12.6 | 41.7 | 0.0 | 25.4 | 12.7 | 42.0 | 0.0 |
| `Llama3-70b` | **34.8** | **26.5** | **49.8** | 0.0 | 0.0 | 9.9 | 38.8 | 0.0 | **33.3** | 22.6 | 46.6 | 0.0 | 34.0 | 23.0 | 48.3 | 0.0 | 35.1 | 24.0 | 48.5 | 0.0 |

Table 2: Summary of our results across nine autoregressive LLMs, open- and closed-source, organized by different classes of model size. Bin-f1 refers to the f1 score in the binary classification setting, where we only consider label 2 (`MisLC`) as the positive class. Ma-f1 and Mi-f1 are the macro- and micro-f1 for the 3-way classification task, where label 1 and 2 (`MisLC`, `Unclear`)are both positive classes.

| Setting | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ |
|---|---|---|---|
| `Random class` | 18.4 ±2.8 | 17.4 ±1.8 | 35.2 ±1.7 |
| `All label 2` | 23.1 | 11.6 | 38.8 |
| `All label 1` | 0.0 | 9.9 | 38.8 |
| `Mean Expert Performance` | 71.1 ±16.8 | 64.9 ±16.7 | 73.1 ±13.0 |

Table 3: Point-of-reference values for our binary and 3-way classification settings. Random class is a classifier where we sample predictions from a random distribution. The random class performance is taken over 100 runs. ± indicates standard deviation.

## 5 Experiment Results

We perform experiments on a wide range of publicly available LLMs. Considering its importance for the legal domain and our task here, we extend our investigation to include Retrieval Augmented Generation (RAG).

### 5.1 The State of the Art of LLMs on MisLC

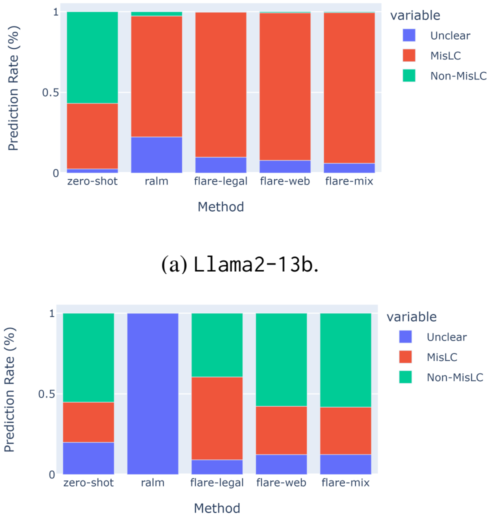
Our results are summarized in Table 2. We also provide additional reference performances, including the lower bound random classifier and upper bound human performance, in Table 3. Bin-f1 refers to the f1 score in the binary classification setting, where we only consider label 2 as the positive class. Ma-f1 and Mi-f1 represent the macro- and micro-f1 for the 3-way classification task, where label 1 and 2 are separate positive classes, as defined earlier in this paper. Overall, the experiments show that the `MisLC` task is challenging for current large language models, even when augmented with retrieval, and they do not achieve human performance. This finding emphasizes the need to develop sophisticated methods to solve `MisLC`.

**`MisLC` performance scales with general domain performance.** In general, the performance trends observed in `MisLC` align with models' general performance. For example, the open-source models `Mistral-7b` and `Solar-10b` are known to perform better than the default Llama-2 models, but the more recent Llama-3 models generally exhibit higher performance than others at similar sizes. The best performing closed-source model in the binary setting is `GPT-3.5-turbo`, performing +12 points f1 better than random guessing, while the best performing open-source model is `Llama3-70b` (+14.4 f1 score). For the 3-way classification, all the models except `GPT-4o` and `Llama3-70b` performed close to the random classifier baseline. It is more challenging to predict the `unsure` class than `MisLC`.

**Larger models follow legal instructions more effectively.** Older language models, especially the Llama 2 series, show high error rates (ER), i.e., failing to provide an expected keyword for 20-60% of the answers. Upon inspecting the generations, we find they often refuse to answer the prompt despite instructions asking otherwise. We also perform experiments without this constraint, allowing the model to generate freely and performing more extensive post-processing for evaluation. While the error rate decreases, the trends in performance are inconsistent. Please refer to Appendix D.1 for further discussion.

### 5.2 Effect of Retrieval

Our task is heavily reliant on external data, evidence $E_i$ and legal issues $L_i$, so a language model should be able to effectively retrieve and parse relevant knowledge. We retrieve from two sources: the legal resources used to create our definition, as described in Section 3.1, as well as web search. Similar to the above *no-retrieval* setting, the models that have the best general domain performance benefit



(a) Llama2-13b.

(b) Llama3-70b.

Figure 3: Label distribution of the model predictions in our five settings for the best- and worst-performing models with retrieval.

| GPT-4o | Bin-f1↑ | Ma-f1↑ | Mi-f1↑ | Ablation | Bin-f1↑ | Ma-f1↑ | Mi-f1↑ |
|---|---|---|---|---|---|---|---|
| FLARE(legal) | 32.3 | 25.8 | 46.7 | Random(legal) | 34.5 | 27.3 | 48.2 |
| FLARE(legal) | 32.3 | 25.8 | 46.7 | Oracle(legal) | 32.3 | 25.6 | 46.5 |
| FLARE(web) | 34.5 | 26.2 | 47.4 | Oracle(web) | 36.4 | 28.5 | 46.5 |
| FLARE(legal+web) | 37.7 | 28.0 | 46.7 | Oracle(legal+web) | 35.9 | 27.6 | 46.2 |

i

Table 4: Summary of our ablations with GPT-4o using FLARE pipeline.

the most from retrieval. In particular, GPT-4o is the only model with a significant increase in performance (+9.0 Bin-f1) compared to other models. In the smaller models, combining the two sources *hinders* performance. Compared to the *no-retrieval* setting, Mistral-7b has a decrease in performance (-11.2 Bin-f1). Its 3-way classification performance remains constant, due to the model's improved performance on the Unsure class.

We also note some models are not responsive to the retrieval methods combined with our task. For example, the Llama 3 series predict only the Unsure class with the IC-RALM method, scoring 0.0 points on Bin-f1. The label distributions of the best and worst models are shown in Figure 3, and other models can be found in Figure 6. The worst-performing model Llama2-13b predicts the majority of our samples as MisLC. The major difference between IC-RALM and FLARE is the frequency of retrieval; FLARE chooses when to retrieve adaptively based on the minimum model perplexity in a generated sentence. This indicates that retrieving too often can harm performance — even in general domain tasks, FLARE's dynamic retrieval is found to perform better than static methods (Jiang et al., 2023b). We perform additional experiments to explore this hypothesis in Appendix D.2, and we urge further study in this direction.

### 5.3 Detailed Analysis and Ablations

While retrieval is important due to the broad range of knowledge required to detect and classify misinformation, we also examine the effectiveness of the models when directly given the legal issues $L_i$ and evidence $E_i$. We present two ablations with the FLARE pipeline: **Random-legal**, where we retrieve a random document from the legal dataset as a lower bound, and an **Oracle** setting as an upper bound. In the oracle setting, we provide the *definition* of the ground truth legal issues $L_i$ as shown in Table 11. If there are no legal issues, we perform retrieval as per our normal pipeline. We also consider the ground truth evidence $E_i$, where we download the sources provided by legal annotators as HTML files, extract the first 500 characters of text, and concatenate all sources as the retrieved document. We present results with GPT-4o, our best-performing model, as well as Llama3-70b (in Appendix D.3).

As shown in Table 4, the random document does not confuse the model, with performance increasing consistently by approximately 2 points f1 across all metrics. The oracle setting demonstrate improvement when only performing web search. We observe a decrease in performance when utilizing the ground truth definitions of our legal issues. This indicates the context afforded by the legal resources benefits model performance more than just a definition, but the retrieval algorithm does not necessarily choose the most relevant documents.

## 6 Conclusion

We introduce a new task: Misinformation with Legal Consequence (MisLC) built on a body of literature spanning 4 broader legal topics and 11 fine-grained legal issues. A comprehensive study is performed on a wide range of open-source and proprietary LLMs that covers a broad parameter spectrum and varying training data. We also adapt existing works in Retrieval-Augmented Generation (RAG), retrieving from the web as well as our curated body of legal documents. We show the task remains challenging for the existing state-of-the-art large language models. We hope our work can enable future research on this important task with significant societal impact.

## Limitations

We recognize there are several limitations in our work. As alluded to in various sections of the paper, misinformation is not its own legal issue. There are many historical cases where legal solutions to misinformation have been misused for censorship, and then repealed.[8] Some argue the government should not be the arbitrator of the truth (Ó Fathaigh et al., 2021)[9]. However, the growing menace of online misinformation and disinformation underscores the urgent need for policy intervention. Regulation is an increasingly viable strategy, exemplified by the European Commission's recent action plan aimed at combating online disinformation. [10]

There are minor details in our work that rely on closed-source API solutions, such as OpenAI Chat Completions API and Google Search, that reduces the reproducibility. Additionally, the adversarial filtering method we used has significant variance in the chosen samples every run. We did implement the AFLITE method used in WinoGrande (Sakaguchi et al., 2021) and found the difference between the two methods to be negligible after inspecting the samples manually. We will also posit that many models such as OpenAI specify they are not meant for domain-specific applications — our results are meant to benchmark current performance and demonstrate there is continued room for improvement. Additionally, there is no legal LLM currently released, despite previous works calling for its development (Dahan et al., 2023).

## Ethics Statement

**Intended Use.** Some applications include:

- **Content Moderation for Social Media Platforms:** Social media platforms can use such a system to moderate content and identify misinformation that could potentially lead to legal liabilities. This can help platforms comply with regulations related to illegal content, defamation, hate speech, or other forms of harmful content.

- **Compliance Monitoring for Regulatory Bodies:** Regulatory bodies responsible for overseeing social media activities can utilize such a system to audit compliance with laws and regulations related to online content. For instance, it can help identify posts that violate consumer protection laws, election regulations, or intellectual property rights.

- **Journalistic Integrity Verification:** News organizations can use the system to verify the accuracy of social media content before reporting on it. This can help uphold journalistic integrity and avoid publishing false information that could lead to defamation lawsuits or damage the credibility of the news outlet.

This paper defines a new task for harnessing misinformation societal harms and encourages researchers to develop more advanced algorithms to mitigate this.

## References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Acl 2023 tutorial: Retrieval-based language models and applications. *ACL 2023*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tony Barrera, Anders Hast, and Ewert Bengtsson. 2004. Incremental spherical linear interpolation. In *The Annual SIGRAD Conference. Special Theme-Environmental Visualization*, 013, pages 7–10. Linköping University Electronic Press Linköping, Sweden.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

[8] R. v. Zundel, [1992] 2 S.C.R. 731

[9] https://www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/summary-session-eight.html

[10] https://digital-strategy.ec.europa.eu/en/policies/online-disinformation

Ceren Budak, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts. 2024. Misunderstanding the harms of online misinformation. *Nature*, 630(8015):45–53.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Emily Chen and Emilio Ferrara. 2023. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between ukraine and russia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1006–1013.

Samuel Dahan, Rohan Bhambhoria, David Liang, and Xiaodan Zhu. 2023. Lawyers should not trust ai: A call for an open-source legal language model. *Available at SSRN 4587092*.

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.

Yasmin Dawood. 2020. Protecting elections from disinformation: A multifaceted public-private approach to social media and democratic speech. *Ohio St. Tech. LJ*, 16:639.

Michiel de Jong, Yury Zemlyanskiy, Nicholas Arthur FitzGerald, Fei Sha, and William Weston Cohen. 2022. Mention memory: incorporating textual knowledge into transformers through entity mention attention. In *10th International Conference on Learning Representations, ICLR 2022, Virtual Conference , April 25-29, 2022*.

Xinya Du and Heng Ji. 2022. Retrieval-augmented generative question answering for event argument extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4649–4666, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adam M Enders, Joseph E Uscinski, Casey Klofstad, and Justin Stoler. 2020. The different forms of covid-19 misinformation and their consequences. *The Harvard Kennedy School Misinformation Review*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5703–5714, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Konstantinos Katevas, Timo Steidle, Max Winter, et al. 2022. Legal foundation–do legal remedies work? *Central and Eastern European eDem and eGov Days*, 342:127–153.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Chu Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Legally enforceable hate speech detection for public forums. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10948–10963, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Greg Nyilasy. 2019. Fake news: When the dark side of persuasion takes over. *International Journal of Advertising*, 38(2):336–342.

Ronan Ó Fathaigh, Natali Helberger, and Naomi Appelman. 2021. The perils of legally defining disinformation. *Internet policy review*, 10(4):2022–40.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Kellin Pelrine, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–64.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Edson C. Tandoc Jr. 2019. The facts of fake news: A research review. *Sociology Compass*, 13(9):e12724. E12724 SOCO-1481.R1.

Rosamond Thalken, Edward H Stiglitz, David Mimno, and Matthew Wilkens. 2023. Modeling legal reasoning: Lm annotation at the edge of human agreement. *arXiv preprint arXiv:2310.18440*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Joris van Hoboken, Naomi Appelman, Ó Fathaigh Ronan, Paddy Leerssen, Tarlach McGonagle, Nico van Eijk, and Natali Helberger. 2019. The legal framework on the dissemination of disinformation through internet services and the regulation of political advertising. *Institute for Information Law, University of Amsterdam*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

# A Detailed Related Work

Misinformation is a serious issue with significant societal impact, as factual dissonance can cause disorder in peoples' worldviews (Nyilasy, 2019). One option to minimize the effect of misinformation is automatic regulation or content filtering. Automatic methods play an important role in detecting misinformation, as they can reduce manual labour costs in searching for emerging rumours (Das et al., 2023). In practice, many such automatic systems often result in a poor user experience due to their lack of transparency (Gorwa et al., 2020). There have been various works that address separate components of the fact-checking pipeline: identifying checkworthy claims, gathering sources on those claims, and cross-checking the sources to confirm veracity (Das et al., 2023). There is growing interest in how to address the problem with LLMs (Chen and Shu, 2023; Bang et al., 2023), and emerging works proposing new methodologies for fact-checking (Pelrine et al., 2023; Pan et al., 2023). However, these works do not consider issues in the law. While there are concerns with regulating misinformation with the law (Ó Fathaigh et al., 2021), we argue it is because of this discourse that the laws that currently exist have undergone rigorous vetting processes and are balanced to reduce societal harm. The most similar work to ours in objective is (Luo et al., 2023), which finds there are discrepancies between hate speech detection works and the law.

Generative models have recently demonstrated strong proficiency in a wide variety of tasks such as relevance, stance, topics, and frame detection in tweets (Gilardi et al., 2023). Many new methods have emerged following the success of RLHF, including Direct Preference Optimization (DPO) to train a policy directly into a language model (Rafailov et al., 2023). There is also a wide breadth of literature on improving the reasoning of an LM. (Wei et al., 2022) introduced few-shot chain-of-thought (CoT) prompting, which prompts the model to generate intermediate reasoning steps before reaching the final answer. Due to the success of CoT prompting and the quality of the reasoning, several newer models incorporate step-by-step demonstrations in the training process (Lightman et al., 2023). This can act as a form of knowledge distillation when a larger language model generates higher quality demonstrations for a smaller model (Mukherjee et al., 2023).

Large Language Models (LLMs) have also demonstrated the ability to capture and memorize a vast amount of world knowledge during pretraining (Guu et al., 2020). However, this knowledge is stored implicitly within their parameters, leading to a lack of transparent source attribution for the facts and information generated in their outputs (Rashkin et al., 2023; Manakul et al., 2023). LLMs are also susceptible to hallucinations, potentially fabricating facts and sources in their responses (Ye et al., 2023). While some previous works refer to these errors as hallucinations (Luo et al., 2023), more recent works clarify hallucinations as a plausible answer with fabricated facts (Ye et al., 2023). One viable strategy to address factual accuracy is Retrieval-Augmented Generation (RAG), where the language model is given explicit knowledge from external corpora (Du and Ji, 2022). Broadly speaking, various RAG strategies differ in three aspects: i) retrieval as text chunks, tokens, or other text snippets, ii) how to integrate the retrieved text with the LM, and iii) when to trigger retrieval (Asai et al., 2023). Some approaches prepend retrieved documents in the input layer of the LM, leaving the LM architecture unchanged (Guu et al., 2020; Shi et al., 2023). In this category, In-Context RALM (IC-RALM) (Ram et al., 2023) and FLARE (Jiang et al., 2023b) methods do not require pretraining or fine-tuning LMs, which can be expensive due to the large LM sizes. RAG can also be done by incorporating the retrieved text in intermediate layers (Borgeaud et al., 2022; Févry et al., 2020; de Jong et al., 2022), or the output layers (Khandelwal et al., 2020; He et al., 2021). These approaches require access to the intermediate layers of the models, changes to the LM architecture, and/or further training in order for the model to use the data effectively.

# B Additional Dataset Details

Please refer to Table 5 for example social media posts from our dataset.

## B.1 Data Processing

The dataset contains a year's worth of tweet metadata from February 2022 to February 2023, collected to facilitate further research in misinformation. We hydrated 1 million English-language tweets, from which 10,000 tweets are randomly sampled. This was performed in February 2023, before Twitter's API policy changes were enacted.

| | |
|---|---|
| Checkworthy & MisLC | “We can deploy troops halfway around the world, in the middle of the Iraq desert, and feed them lobster on Sunday night. The Russians can’t even supply their troops 50 miles from their homeland with unexpired MREs.” |
| Checkworthy but not MisLC | Some Russian performing artists are speaking out against Putin - NPR |
| Not Checkworthy, not MisLC | “Ukraine is my home. Every street, corner, alleyway, nook and cranny all over the country have made me what I am today. If all of that is lost I have no idea who I’ll be.” |

Table 5: Cherry-picked samples from our dataset comparing crowd-sourced labels of Checkworthiness to our expert annotations of MisLC.

We then used Google Translate’s language detection function[11] as a secondary filter for tweets exclusively in English. All usernames (words starting with an @ symbol) in the tweets are replaced with <user>, and we remove unicode characters by encoding to ASCII. Finally, we identify social media posts with claims using a fine-tuned version of DeBERTa for claim detection[12], stopping once we have 4,000 samples. We also tested ChatGPT, but find DeBERTa is better aligned to Claim vs. No Claim annotations by our research team, which was performed based on previous definitions from (Das et al., 2023).

### B.2 Retrieval Database Preprocessing

We convert the text from PDF to HTML format using Adobe Acrobat, and then split each document into paragraphs by searching for two consecutive newline characters. Next, we rejoin the paragraphs in chunks of 2048 tokens with a 50% sliding window context to preserve one paragraph’s context and relationships with its immediate neighbours. After the splitting process, we obtain 590 text chunks. We build a positional BM25 index upon them using Pyserini (Lin et al., 2021).

### B.3 Crowd-sourced Annotations

Please refer to Table 10 for crowd-sourced annotation instructions. We first chose workers on Mechanical Turk through a prescreening process. We sampled 100 tweets and collected a set of annotations from two researchers given the instructions in Table 10. The researchers’ labels had a fourth option of “ambiguous” — that is, these samples appeared to be too subjective to indicate good understanding of the worker’s performance. This “ambiguous” label is automatically assigned where the researchers disagreed, or if one researcher preemptively assigns a sample as ambiguous. Then, we scored all workers with the researcher annotations as a ground truth. A worker needed to have a 70% agreement with researcher annotations, excluding ambiguous samples, and they needed to have completed at least 10 HITs in the prescreening to be considered for further annotation. Among the annotators that met all requirements, two of them only labelled ambiguous samples — for them, we sent a secondary test to obtain a fair assessment. We compensated the workers at $0.18 per HIT.

### B.4 Adversarial Filtering

We use embeddings from RoBERTa-Large, and train the linear classifier with a KL divergence loss objective as shown in Equation 2. Since this is a different task from binary classification, we do not set a fixed $\tau$ — instead, we take $\tau$ to be the mean loss over the entire dataset, which we indicate with $\tau_\mu$. During our filtering process, we find $\tau_\mu$ to be approximately 0.1 across three rounds.

## C Experiment Details

### C.1 Model Choice

- GPT-3.5-turbo (Ouyang et al., 2022) — A closed-source model trained with Reinforcement Learning with Human Feedback (RLHF). We performed experiments in June of 2024.

- GPT-4o[13] — A closed-source model trained with Reinforcement Learning with Human Feedback (RLHF). We performed experiments in June of 2024.

- Llama2-(7b, 13b, and 70b) (Touvron et al., 2023) — A suite of open-source models trained using RLHF, as well as safety fine-tuning to enhance helpfulness.

[11] https://cloud.google.com/translate/docs/reference/rest

[12] https://huggingface.co/Nithiwat/mdeberta-v3-base_claim-detection

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

- `Llama3-(8b, and 70b)`[14] — A suite of open-source models trained using a combination of supervised fine-tuning (SFT), rejection sampling, proximal policy optimization (PPO), and direct preference optimization (DPO), with a focus on safety fine-tuning to enhance helpfulness.

- `Mistral-7b` (Jiang et al., 2023a) — A model trained with instruction tuning; rather than reinforcement learning, they fine-tune directly on instruction data.

- `Solar-10b` — This is a merged model that combines the instruction-tuned version of Solar, Solar-Instruct (Kim et al., 2023), and OrcaDPO, another checkpoint trained with Direct Preference Optimization (Rafailov et al., 2023). These two models are merged with Spherical Linear Interpolation (SLERP) (Barrera et al., 2004).

We choose Llama 2 and 3 to isolate the effect of model size, since these suites of models are trained with the same method at various parameter counts. We compare this suite to three open-source models trained on various combinations of fine-tuning, instruction tuning, and Proximal Policy Optimization (PPO) or Direct Policy Optimization (DPO). We compare them to Mistral and Solar as they are the best-performing models on the Huggingface Open LLM leaderboard.[15] We searched for the best model in three sizes (7B parameters, ≈13B parameters, and 60B+ parameters) as of January 2024. The `Solar-10b` we test combines two checkpoints of Solar (Kim et al., 2023): Solar-Instruct, trained with instruction tuning, and OrcaDPO.

Additionally, we compare these against the closed-source `GPT-3.5-turbo` and `GPT-4o` as an alternative for limited computational resources. We utilize OpenAI's Chat Completions API for ChatGPT, and Hugging Face's text generation pipeline for all other models. For all models, we choose a sampling temperature of 0.3 after a hyperparameter search, testing temperatures of 0.1, 0.3, 0.5, 0.8, and 1.

## C.2 Prompting

Please refer to Table 6 for our prompting format. While individual model prompts might vary based on their specific template formatting requirements, the core text is held constant throughout all of our experiments.

| | |
|---|---|
| Retrieval | Here is some relevant legal context on "misinformation": [doc]<br><br>Web search results for the claim: [snippets] |
| Classification | Claim: [claim]<br>Classify the claim as either "factual" or "misinformation." |
| Constraints | Do not refuse to answer. Do not engage in explanations and politeness. Only respond with the words "misinformation", "factual", or "unsure". Do not add further context. |

Table 6: Prompt template used in our experiments. We use a simple sentence to indicate the context of our retrieved document [doc] and/or web search results [snippets], and a keyword 'Claim' to indicate the target input text within the prompt. In the main results, we also add some instructions to constrain the output to a single keyword.

## C.3 Retrieval

Please refer to Figure 4 for an illustration of the two architectures. We take the implementation of IC-RALM directly from their Github,[16] and take most of FLARE's original implementation[17] except for the generation. We use ChatGPT for query generation in FLARE — we also tested using the model itself and found performance comparable. At the time these experiments were conducted (January 2024), the OpenAI Chat Completions API did not return log probabilities, so we were not able to conduct experiments with `gpt-3.5-turbo` and FLARE.

## C.4 Hyperparameter tuning and Hardware Specifications

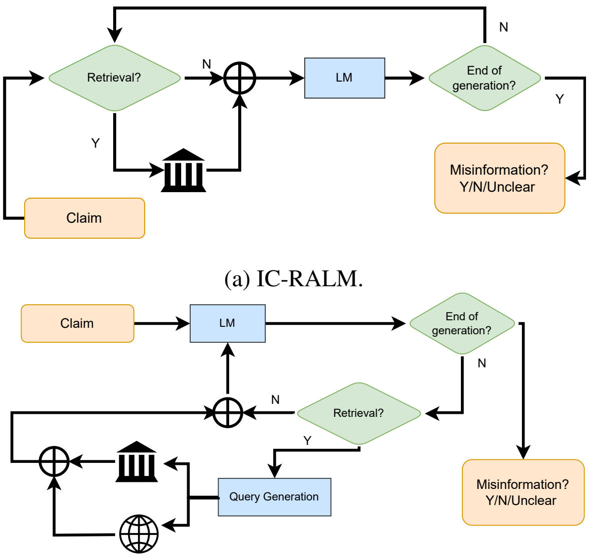
For the IC-RALM experiments, we set the *stride* parameter to the $s = 4$ tokens that was used in most of (Ram et al., 2023) experiments, as it keeps a balance between performance and efficiency. This parameter is the frequency of which the retrieval is triggered. In FLARE experiments, we set the $\beta$ (the

[14] https://ai.meta.com/blog/meta-llama-3/

[15] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

[16] https://github.com/ai21labs/in-context-ralm

[17] https://github.com/jzbjyb/flare



(a) IC-RALM.

(b) FLARE.

Figure 4: Illustrations of the IC-RALM and FLARE retrieval architectures.

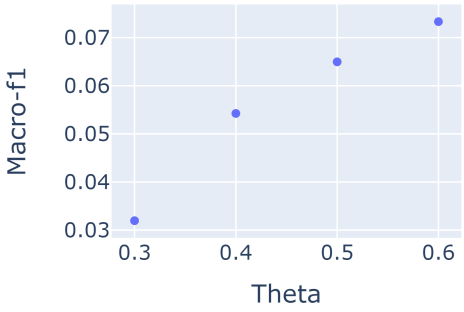


Figure 5: Change in macro-f1 as we increase $\theta$ over the first 100 samples.

confidence threshold for query generation) value to be 0.4 and did a grid search for $\theta$ (the confidence threshold for triggering retrieval) with 100 samples of our dataset to find the best-performing value. We found performance scales consistently with $\theta$, as shown in Figure 5, and we choose $\theta = 0.5$ to balance performance with throughput.

We generate outputs with the vLLM library[18], setting a maximum generation length of 1024. Experiment run times depended largely on the model size and experimental setting; smaller models took approximately 1.5 hours on our full dataset in the Flare (legal+web) setting, while larger models could take 3 hours. This equates to 1.5 GPU hours for smaller models, or 12 GPU hours for larger

| Model | FLARE ($\theta = 0.5$) | | | | FLARE ($\theta = 1$) | | | | IC-RALM (Legal) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1 ↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1 ↑ | Mi-f1 ↑ | ER ↓ |
| GPT-4o | 32.3 | 25.8 | 46.7 | 0.0 | 30.7 | 25.0 | 46.7 | 0.0 | 35.8 | 28.5 | 48.9 | 0.0 |
| Llama3-8b | 27.3 | 13.6 | 43.0 | 0.0 | 25.3 | 13.9 | 41.4 | 0.0 | 0.0 | 9.8 | 38.7 | 0.0 |
| Llama2-13b | 22.2 | 17.8 | 38.8 | 0.0 | 22.6 | 16.0 | 39.3 | 0.0 | 22.6 | 18.2 | 39.7 | 0.1 |

Table 7: A comparison of the RALM retrieval method with FLARE set to retrieve at every possible step (i.e. $\theta = 1$). We conducted experiments for all models but only present results for these three to illustrate the effect of retrieval.

| Llama3-70b | Bin-f1↑ | Ma-f1↑ | Mi-f1↑ | Ablation | Bin-f1↑ | Ma-f1↑ | Mi-f1↑ |
|---|---|---|---|---|---|---|---|
| FLARE (legal) | 33.3 | 22.6 | 46.6 | Random (legal) | 32.2 | 22.7 | 47.9 |
| FLARE (legal) | 33.3 | 22.6 | 46.6 | Oracle (legal) | 32.0 | 21.1 | 47.5 |
| FLARE (web) | 34.0 | 23.0 | 48.3 | Oracle (web) | 32.2 | 22.7 | 48.7 |
| FLARE (legal+web) | 35.1 | 24.0 | 48.5 | Oracle (legal+web) | 34.2 | 21.6 | 48.0 |

Table 8: Summary of our ablations with Llama3-70b using FLARE pipeline.

models. We conducted experiments with open-source models on a server cluster with a combination of Nvidia RTX6000-48GB and A100-40GB GPUs.

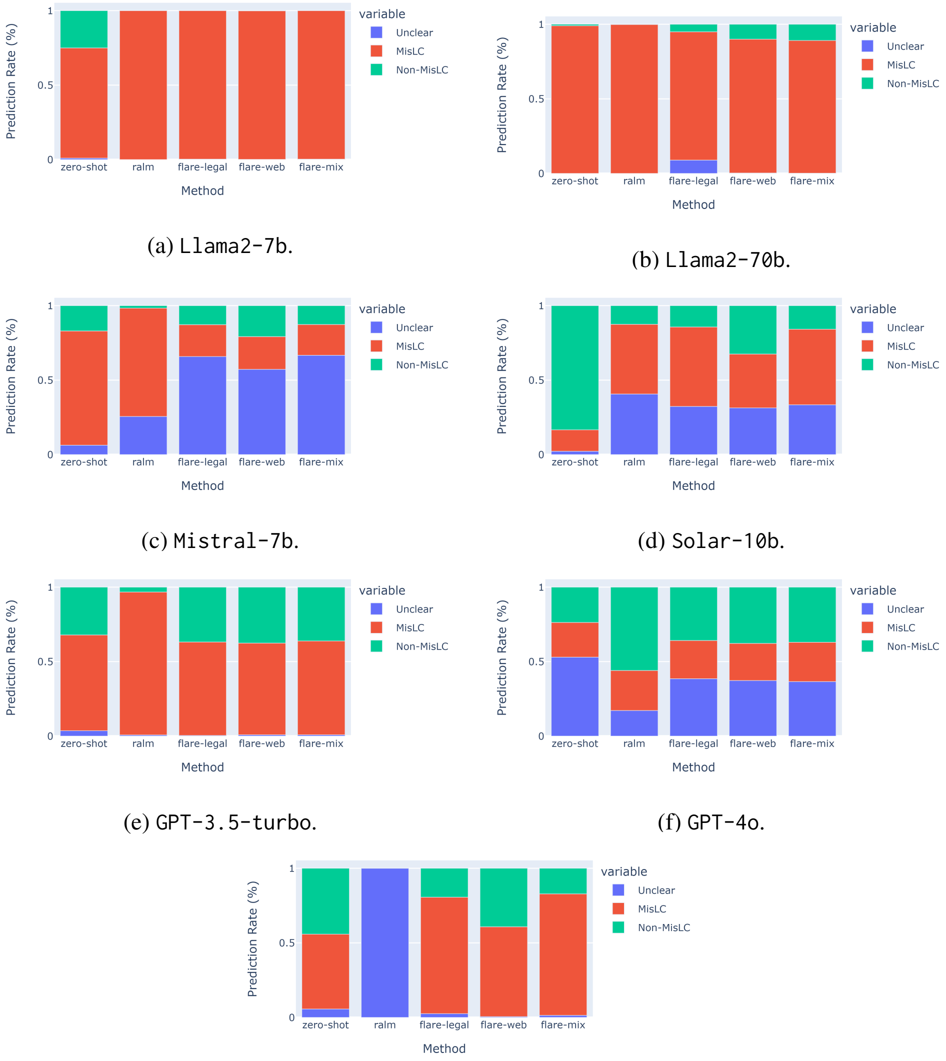
## D Additional Experiments

### D.1 Prompt Constraint and Error Rate

Previous works have observed legal tasks with long contexts often lead to a model being more "decisive" (Luo et al., 2023). In our experiments, we note that adding retrieved text to the input context significantly reduces the error rate. This suggests there is some trade-off between the instruction complexity and the safety fine-tuning performed for the Llama 2 models. Llama 2's safety fine-tuning has been noted to be unstable and easily reversed with a few steps of parameter-efficient fine-tuning (Lermen et al., 2023), and we hypothesize this instability is also causing the fluctuations in error rate.

In addition to the stricter prompting instructions reported in the main body, we also evaluate the models without constraining the outputs — i.e. simply asking for a classification, as shown in Table 6. We evaluate the generations by searching the entire generated text for the keywords "factual" or "misinformation." We first check for the keywords in quotes (" "), as that is the format given in the prompt, and then we check for all other mentions if quotes do not exist. If a model's generated text contains both of these keywords, we count this as an unclear prediction. For any generation without either keyword, we first filter over all model responses to analyse the responses. Many of these answers are non-answers, such as "As an AI language model, I am unable to provide a response." is a non-answer, or an error in the generation. We

[18] https://docs.vllm.ai/en/latest/



(a) `Llama2-7b`.

(b) `Llama2-70b`.

(c) `Mistral-7b`.

(d) `Solar-10b`.

(e) `GPT-3.5-turbo`.

(f) `GPT-4o`.

(g) `Llama3-8b`.

Figure 6: Label distribution of the model predictions in our five settings for the remaining models tested.

report the Error Rate (ER) alongside macro- and micro-f1 score.

### D.2 Retrieval Extended

We conducted an ablation with FLARE where we always performed retrieval on the legal dataset (i.e. set $\theta$ to 1) and observed similar performance as RALM across all models, summarized in Table 7. While we conducted experiments with all of the models, we only present three key results. First, `Llama3-8b` had a Bin-f1 score of 0.0 with the IC-RALM retrieval method. However, FLARE even at the highest retrieval level does not exhibit this behaviour.

### D.3 Ablations Extended

Please refer to Table 8 for results with `Llama3-70b`. As shown, the trend is similar to what was observed with `GPT-4o`.

## E Additional Legal Details

Please refer to Tables 11 for a comprehensive list of legal issues considered in our annotations.

| Model | No Retrieval | | | | IC-RALM (Legal) | | | | FLARE (Legal) | | | | FLARE (Web) | | | | FLARE (Legal+web) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1 ↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1 ↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ | Bin-f1↑ | Ma-f1↑ | Mi-f1 ↑ | ER ↓ |
| `GPT-3.5-turbo` | **30.9** | 22.0 | **44.2** | 0.3 | 0.0 | 15.0 | 26.7 | 0.0 | 30.4 | 19.9 | 45.0 | 0.1 | 31.8 | 25.2 | 44.3 | 0.1 | 30.4 | 16.4 | 48.2 | 0.0 |
| `Mistral-7b` | 21.1 | **22.5** | 41.9 | 0.3 | 21.0 | **21.3** | **43.1** | 0.1 | 23.7 | 22.6 | 42.4 | 0.1 | 12.2 | 16.4 | 42.4 | 0.0 | 11.8 | 15.1 | 41.1 | 0.0 |
| `Llama2-7b` | 21.1 | **22.5** | 45.5 | 0.1 | 23.0 | 20.9 | 40.5 | 0.1 | 16.5 | 18.5 | 40.4 | 0.7 | 23.3 | 22.6 | 41.4 | 0.4 | 18.9 | 20.1 | 40.1 | 0.9 |
| `Solar-10b` | 19.2 | 18.3 | 39.4 | 0.3 | **26.6** | **21.3** | 36.5 | 2.1 | 25.3 | 21.7 | 39.5 | 0.7 | 25.3 | 21.7 | 39.5 | 0.7 | 26.2 | 20.7 | 38.8 | 1.2 |
| `Llama2-13b` | 18.0 | 17.7 | 40.0 | 0.0 | 13.2 | 15.0 | 41.3 | 0.3 | 17.7 | 19.5 | 41.4 | 1.5 | 17.5 | 19.6 | 41.2 | 0.9 | 18.7 | 20.0 | 41.5 | 1.5 |
| `Llama2-70b` | 24.1 | 21.0 | 43.5 | 0.1 | 23.6 | 21.2 | 42.0 | 0.0 | 22.8 | 20.7 | 42.1 | 3.1 | 21.3 | 21.8 | 43.4 | 3.0 | 21.9 | 21.1 | 43.5 | 3.3 |

Table 9: Summary of the unconstrained results across seven LLMs, open- and closed-source, organized by model size. Since the size of ChatGPT is unknown, we present it at the top.

We aim to identify checkworthy claims. A claim is defined as "stating or asserting that something is the case, typically without providing evidence or proof."

Examples of claims: "The Eiffel Tower is the tallest tower in the world" "Michael Jackson was seen at the department store last week" "My wife can't eat chocolate because she's allergic"

Not claims: opinions, emotions, exclamations. For example:
"I think Snow White was an idiot,"
"My wife is so nice and I love her,"
"Wow! Amazing!"
If there is no claim, please indicate "Empty/no claim. (3)"

Please choose "Checkworthy" (1) if you consider at least one claim in the statement to be checkworthy. Checkworthy is defined as: Having the potential to influence/mislead people, organizations and countries. If you read this statement, it would influence your opinion of the topic. Discussing a topic, or quoting a person capable of signficant social impact.**

It might be checkworthy if you can answer "yes" to any of these questions:
Does it provoke an emotional response?
Does it make a bold statement on a controversial issue?
Is it an extraordinary claim?
Does it contain clickbait?
Does it have topical information that is within context?
Does it use small pieces of valid information that are exaggerated or distorted?

For example: "Biden's Climate Requirements: Cut 90% of Red Meat From Diet; Americans Can Only Eat One Burger Per Month" is a checkworthy claim because it suggests the President of the United States wants to regulate peoples' diets. Some might feel angry because it is outside Biden's jurisdiction, so it is important to fact-check this statement.

Choose Not Checkworthy (2) if the claim is not checkworthy. Not checkworthy claims are at least one of the following:
Innocuous (eg. Ryan Renolds has six fingers on his right hand)
Based on common knowledge (eg. water is wet, a cough makes your throat sore)
Made solely based on private information (eg. I had a sandwich for lunch yesterday)

Table 10: Instructions provided to crowd-sourced (Mechanical Turk) workers for identifying checkworthiness.

| **Broad Legal Topic** | **Legal Issue** | **Key legal tests** | **Defences** |
|---|---|---|---|
| Defamation | Defamation | 1. Defamatory in Nature (in the sense that the things in question would tend to lower the plaintiff's reputation in the eyes of a reasonable person)<br>2. Publication (communicated to a third party) | 1. Qualified Privilege<br>2. Responsible Communications<br>3. Fair Comment (assuming (a) the comment is on a matter of public interest; (b) the comment is based on fact; (c) the comment, though it can include inferences of fact, is recognizable as comment; and (d) any person could honestly express that opinion on the proved facts) |
| Freedom of Expression | Freedom of Expression | 1. The activity must be an expressive, i.e. must "convey meaning" ("It might be difficult to characterize certain day-to-day tasks, like parking a car, as having expressive content.")[25]<br>2.Is the government's purpose, or otherwise effect, to restrict expression of this meaning? | 1. Can establish the "truth," eg. clinical evidence<br>2. Non-intent, i.e. published misinformation without intent[26] |
| Criminal Laws | Cyberbullying | If false/inaccurate information is being spread to harass or harm others, the spreader could face cyberbullying or harassment charges | N/A |
|  | Public Mischief | Every one commits public mischief who, with intent to mislead, causes a peace officer to enter on or continue an investigation by<br>(a) making a false statement that accuses some other person of having committed an offence;<br>(b) doing anything intended to cause some other person to be suspected of having committed an offence that the other person has not committed, or to divert suspicion from himself;<br>(c) reporting that an offence has been committed when it has not been committed; or<br>(d) reporting or in any other way making it known or causing it to be made known that he or some other person has died when he or that other person has not died. | N/A |
| Consumer Protection Laws | Food and Drugs Act | Spreading false and private information about someone without their consent can lead to privacy violation claims because it infringes upon their right to control their personal information and keep it private. | N/A |
|  | Data Privacy | Under the Food and Drugs Act, Health Canada is tasked with (among other things) monitoring misleading health claims and regulatory enforcement to address health risks<br>Among other things, food in Canada shall not be sold or advertised in a manner that is false, misleading or deceptive | N/A |

| **Broad Legal Topic** | **Legal Issue** | **Key legal tests** | **Defences** |
|---|---|---|---|
| Consumer Protection Laws | Federal Competition Act | The Commissioner of the Competition and the Department of Public Prosecutions can initiate actions to address misleading claims using either of the criminal [section 52(1)] or civil tracks<br>- All representations that are false or misleading in a material respect, in any form, are subject to the Competition Act<br>- If a representation could influence a consumer to buy or use the product or service advertised, it is material<br>- NOTE: Propaganda and advertising are usually based on real accounts, with an incomplete focus on parts that are favourable to a campaign (Tandoc Jr., 2019).<br>- To determine whether a representation is false or misleading, the courts consider the "general impression" it conveys, as well as its literal meaning | N/A |
| Other (i.e., not fitting within one of the broad legal topics above) | Election Laws | The Canada Elections Act has prohibited false or misleading statements, since 2018, about electoral candidates if they are expressed during the election period with the intention of affecting the results of the election. The Election Modernization Act sets out important transparency and disclosure requirements for political advertising (Dawood, 2020) | N/A |
| | Intentional Infliction of Mental Suffering | This common law tort involves intentionally inflicting emotional distress through acts or words which results in emotional harm as visible, provable illness.<br>- The plaintiff must prove 1) Act (Statement need not be false, but speech must be extreme), 2) Intent (i.e. calculated to produce harm), 3) Injury (i.e. the plaintiff must have suffered actual harm; some injury in the form of psychological harm) | N/A |
| | Hate Speech | Fake news affects society as a whole, whereas hate speech harms individuals or members of a specific group (Katevas et al., 2022) | N/A |
| | Intellectual Property | Trademarks Act provides that no person shall "make a false or misleading statement tending to discredit the business, goods or services of a competitor", nor "make use, in association with goods or services, of any description that is false in a material respect and likely to mislead the public as to" the character, quality, quantity or composition, the geographical origin, or the mode of the manufacture production or performance of the goods or services. | N/A |

Table 11: Areas of law that can be used to indict misinformation published online.