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 unintentional intent. The rapid online information exchange necessitates advanced detection mechanisms to mitigate misinformation-induced harm. Existing research predominantly focuses on the veracity of information, overlooking the legal implications and 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. To enable the study, we introduce a two-stage annotation process where we first identify checkworthiness with crowd-sourced workers, and then potential legal issues with legal annotators. Next, we utilize a Retrieval-Augmented Generation (RAG) architecture to address the task with large language models (LLMs). We present empirical baseline results over seven open- and closed-source LLMs, with and without RAG. Our findings contribute to the interdisciplinary efforts to combat misinformation by leveraging legal perspectives. [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 online communications. For instance, conflicting information during the COVID-19 pandemic significantly influenced people's attitudes and behaviors toward preventing viral spread (Enders et al., 2020). Misinformation proves extremely detrimental in the context of current events, often labeled as 'fake news.' Malicious actors are particularly motivated to maliciously spread false information to manipulate public opinion while an event unfolds and the truth is yet to be widely acknowledged (Nyilasy, 2019). Recent years have also seen a rise in the deliberate sharing of false information intended to mislead, spreading "fear and suspicion" among the population. This deliberately misleading misinformation is referred to as *disinformation*. [2]

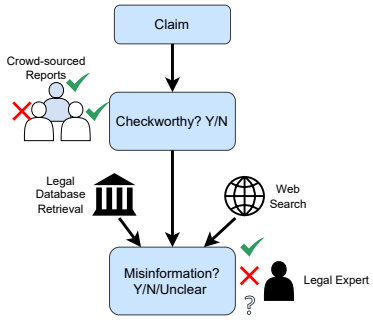


Figure 1: An overview of our proposed task framework for legal misinformation. We obtain crowd-sourced labels of checkworthiness. If a claim is checkworthy, we use legal annotators to annotate potential legal issues of misinformation.

The growing menace of online misinformation and disinformation underscores the urgent need for policy intervention. Regulation emerges as a viable strategy, exemplified by the European Commission's recent action plan aimed at combating online disinformation. [3] Nevertheless, the mechanisms through which regulators intend to enforce such regulations on individuals and social media companies remain ambiguous. Given the dynamic

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

[2] https://www.unhcr.org/innovation/wp-content/uploads/2022/02/Factsheet-4.pdf

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

nature of online information flow, we posit that AI-enabled solutions could play a critical role in mitigating this challenge. Such legal AI solutions requires a human-centric approach, one that augments human capabilities and prioritises human well-being, values, equity, and collaboration between humans and AI systems. Human-centric design is important to the growing interest in AI alignment, or aligning artificial intelligence systems to the values and interests of humans (Bai et al., 2022; Pyatkin et al., 2023). However, these values are often determined by the researchers and developers of these models. Previous task definitions for automatic misinformation detection focus on factual accuracy or checkworthiness as potential controversy of a topic (Das et al., 2023). Within the legal domain, while factual accuracy is important, efforts to regulate misinformation need to be carefully balanced with a person's right to freedom of expression (Ó Fathaigh et al., 2021). This subtle difference in values can have a substantial impact on the goals of a language model.

Another challenge is introducing trustworthy information to validate a claim. Large Language Models (LLMs) are limited to their pretraining knowledge. If there was an unprecedented event such as a global pandemic, a model trained before the event would have no knowledge to judge a claim. LLMs are also susceptible to hallucinations, potentially fabricating facts and sources in their responses (Ye et al., 2023). To bridge this gap in knowledge, retrieval-augmented generation (RAG) can introduce new information to a language model without further training (Gao et al., 2023). Our main contributions are 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 12 fine-grained legal issues, including hate speech, election laws, and privacy regulations. We use a novel dataset curation approach based on a two-step process, comparing crowd-sourced checkworthiness to expert evaluations of legal misinformation issues.
- We provide comprehensive baselines based on state-of-the-art LLMs, both open- and closed-source, to benchmark the task, covering a broad parameter spectrum and varying training data. We utilize two advanced Retrieval-Augmented Generation (RAG) architectures to detect legally consequential misinformation, involving retrieval from legal document databases and web search, which can mimic expert techniques. We show the task remains challenging for the existing models, inviting research efforts to tackle it.
- We provide insights about the `MisLC` task through empirical evidence, from the problem definition to experiments and to how domain experts are involved. We hope some of the insights and discussions could generalize and help with other across-disciplinary tasks.

2 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. 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). However, these works do not consider the legal concept of misinformation.

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). 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). One viable strategy to address 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). 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). For further related work, please refer to Appendix B.

3 Misinformation with Legal Consequences (`MisLC`)

3.1 Definition

Our task is a multi-category classification task with output labels: label 1 (`mislc`: misinforma-

tion with legal consequence), label 0 (`non-mislc`), and label 2 (`unclear`) for when it is impossible to determine a classification. This occurs when there is insufficient context to perform fact-checking, as further discussed in Section 3.2. Formally, our dataset is defined as a set of tuples of the input sample x_i, label y_i, legal issues $l_i \subset L$, and evidence e_i $\{(x_i, y_i, l_i, e_i) \in X\}$. Our legal issues L is a set of tuples derived from an extensive body of literature, with each tuple containing logical **legal tests** and **defences** $(\text{test}(x, e), \text{def}(x, e))$ for an input x and evidence e. A legal issue is associated with a claim if the claim passes its relevant legal tests and does not pass possible defences, i.e. $l_i = \{\text{test}_j(x_i, e_i) \wedge \neg\ \text{def}_j(x_i, e_i)\ \forall\ (\text{test}_j(x, e),\ \text{def}_j(x, e)) \in L\}$. The final label y_i is determined by the set of legal issues l_i and evidence e_i as shown in Equation 1. It should be noted that evidence e_i can either support or refute the claim. The `unclear` class (label 2) is a special case that will be discussed more in Section 3.2.

$$y_i = \begin{cases} 1, & |e_i| > 0\ \wedge\ |l_i| > 0 \\ 0, & |e_i| > 0\ \wedge\ |l_i| = 0 \vee x_i \text{ is not a claim} \end{cases} \tag{1}$$

Legal Literature Defining misinformation from a legal standpoint is challenging—in general, 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, particularly within the Canadian federal and provincial contexts. Research was gathered from 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 cases, as well as European cases which may assist future decision-making by Canadian judges.
- **Official Bulletins** — Publications from global organizations, such as 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 or should not be applied to prevent its harm.

Defining Legal Issues. From the legal documents, we compile eleven major legal issues pertaining to false information, as listed in Appendix E. Each legal 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. Defamation, for example, focuses on the *defamatory* nature, i.e. has the ability to lower someone's reputation, as well as the context the claim was uttered. All relevant legal tests $\text{test}(x, e)$ and defences $\text{def}\ x, e$ were compiled into a comprehensive annotation guide.

3.2 Creation of the `MisLC` Dataset

As illustrated in Figure 1, we construct a pipeline to maximize label certainty such that our legal annotators annotate the most consistent portions of the data. First, non-legal crowd-sourced annotators discover checkworthy samples that arouse their suspicion. Second, legal annotators check these samples for relevant legal issues.

Crowd-sourced Checkworthiness. First, we wish to source samples that are likely to contain misinformation. This component does not require legal expertise, but builds the basis/foundation on which legal practitioners can operate. We collected 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, please refer to Appendix C.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 included indications of disinformation as provided by the Government of Canada. [4] First, 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 C.2. 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 finally 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 670 samples that is consistently retained in each trial. The filtering process biased the label distribution to Checkworthy samples, as shown in Table 1. This complements our intended pipeline where a sample is flagged and further investigated by legal annotators, and indicates strongly Checkworthy samples are likely more consistent than ambiguous agreement. Further details are discussed in Appendix C.3.

$$\mathcal{L}(x_i, y_i) = D_{KL}(f(x_i), \text{softmax}(y_i)) \quad (2)$$

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

Algorithm 1 Our adversarial filtering process.

Require: Dataset $\{x_i, y_i\}_{i=1}^{n} \in X$, target dataset size k

Encode all samples as the last embedding layer $f(x_i) = E_{LM}(x_i)[-1]$
Apply softmax to all $y_i \in X$
Initialize $X' = X$
while True **do**
  Train a linear classifier $f(x)$ on X'
  for $(x_i, y_i) \in X'$ **do**
    $s_i = \mathcal{L}(x_i, y_i)$
  end for
  $\tau_\mu = \mu(score)$
  $S = \{(x_i, y_i) \in X' | s_i > \tau_\mu\}$
  if $|X' \setminus S| < k$ **then**
    break
  else
    $X' = X' \setminus S$
  end if
end while

Legal Annotations. We worked with eleven law researchers to annotate our dataset. Our annotators are law students in their first and second year. They perform 2-3 hours of annotations per week as part of a practicum course, and receive credits as compensation. Each expert is provided a 15-page 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 x_i via majority voting. The nominal Krippendorff's Alpha for the legal annotators is 0.44, 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 the complexity and opportunity for interpretation.

As shown in Figure 2, we find that the most relevant legal issues for our data to be freedom

Label	Count (%)
Total Dataset Size	670 (100)
Checkworthy	626 (93.4)
MisLC	93 (13.7)
Not MisLC	513 (76.6)
Unclear	64 (9.5)
Evidence available	353 (52.7)
No Claim	288 (43.0)
Defence	233 (34.8)

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.

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

of expression, followed by closely by defamation. Next, there are criminal offenses of Cyberbullying and Public Mischief, Hate Speech, and Election laws. Our label distribution is summarized in Table 1. While checkworthiness had a positive rate of 93.6% (626), only 13.7% (93 samples) of the dataset has some possible legal violation for misinformation. Additionally, there were a substantial number of unclear samples (9.5%, or 64). 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 correct evidence 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.

3.3 Prompting Without Retrieval

To evaluate the misinformation detection capabilities of LLMs, we prompt model predictions without any retrieval. Our prompt template is shown in Table 4 of Appendix D. The models are instructed to classify the samples in our dataset, without any retrieval context, as either *factual* or *misinformation*. They are given no external knowledge, and simply classify based on the contents of the input x_i and potential evidence e_i they draw from their own parametric knowledge. 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/ambiguous" 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 report the Error Rate (ER) alongside macro- and micro-f1 score.

3.4 RAG Pipeline

LLMs have a significant amount of world knowledge, but our task of misinformation with legal consequences likely does not exist in their pre-training data. Also, as discussed in Section 3.1, a significant portion of our legal issues is often decided based on the factual truth of a claim. To this end, we wish to introduce information about our task in order to improve their alignment to the thought process of legal annotators. We use RAG to introduce knowledge from our legal literature, as well as to retrieve potential evidence similar to 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 (Ram et al., 2023). In this work, we experiment with two state-of-the-art RAG methods.

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 δ. To avoid information dilution with long queries, the query is limited to the last ℓ 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.

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

FLARE (Jiang et al., 2023b) generates a temporary sentence $\hat{s}_t$, where $p(\hat{s}_t) = \prod_{i=1}^{m} p\left(w_i | w_{<i}\right)$, 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 θ is the threshold parameter. Moreover, FLARE formulates the regenerated sentence s'_t as $p(s'_t) = \prod_{i=1}^{m} p\left(w_i | \left[\mathcal{R}_{\mathcal{C}}(\mathrm{qry}(w_{<i})); w_{<i}\right]\right)$. 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 D.1 for further implementation details.

$$s_t = \begin{cases} \hat{s}_t & \text{if all tokens of } \hat{s}_t \text{ have probs } \geq \theta \\ s'_t & \text{otherwise} \end{cases} \quad (4)$$

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. We find that performance between IC-RALM and FLARE is comparable, as shown in Table 2. Therefore, we only present results with FLARE on our full pipeline.

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. We split each document into paragraphs where there are 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).

Web Search. A crucial component of our legal tests is the availability of evidence e_i for a claim x_i. We query the Google Custom Search API[5] 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 results in batches of 10, and contains 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 fact-checking method — there are many works on algorithms to systematically collect evidence (Das et al., 2023). However, since FLARE generates multiple queries per retrieval step, we believe our method is an effective baseline.

4 Experimental Settings

We present experiments with the following decoder-only generative models: `GPT-3.5-turbo` (Ouyang et al., 2022), `Llama2-(7b, 13b, and 70b)` (Touvron et al., 2023), `Mistral-7b` (Jiang et al., 2023a), `Solar-10b`, and `Mixtral-60b`. We choose Llama 2 chat version as this suite of models is trained with safety RLHF, and to isolate the effect of model size holding everything else constant. We compare this suite to three open-source models trained on vari-

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

Model	No Retrieval			IC-RALM (Legal)			FLARE (Legal)			FLARE (Web)			FLARE (Legal+web)		
	Ma-f1↑	Mi-f1 ↑	ER ↓	Ma-f1 ↑	Mi-f1 ↑	ER ↓	Ma-f1 ↑	Mi-f1 ↑	ER ↓	Ma-f1↑	Mi-f1 ↑	ER ↓	Ma-f1↑	Mi-f1 ↑	ER ↓
`GPT-3.5-turbo`	**37.2**	**52.5**	1.4	**37.2**	**49.0**	7.3	-	-	-	-	-	-	-	-	-
`Mistral-7b`	34.5	46.9	0.3	30.0	35.2	0.6	29.4	33.7	0.6	29.6	40.6	0.4	28.3	33.9	1.6
`Llama2-7b`	33.5	44.2	0.1	21.3	21.2	0.1	28.5	33.0	0.1	22.0	21.9	0.4	18.6	18.8	0.9
`Solar-10b`	31.5	40.6	0.6	32.3	37.0	6.0	**33.5**	**41.3**	1.0	**33.7**	**41.5**	0.7	**32.7**	**34.8**	2.2
`Llama2-13b`	16.0	15.7	0.0	19.2	20.1	0.3	19.7	20.6	1.5	19.3	19.6	0.9	20.0	20.7	1.5
`Mixtral-60b`	34.5	46.9	0.3	25.2	26.3	0.4	24.9	30.0	1.8	30.0	43.6	0.8	29.4	34.1	0.9
`Llama2-70b`	29.7	32.2	0.3	25.6	26.7	0.0	29.6	34.0	3.1	29.3	33.1	3.0	31.4	37.0	3.6
`Majority class`	26.8 ±1.7	33.2 ±1.9	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2: Summary of our results across seven generative LLMs, open- and closed-source, at different model sizes. Since the size of ChatGPT is unknown, we present it at the top. The `Majority class` setting is the score with randomly generated predictions between 0 and 2. ± indicates standard deviation for 100 runs — any result without standard deviation is a single run with greedy decoding or the most deterministic sampling settings.

ous combinations of fine-tuning, instruction tuning, and policy optimization. Most of these models, except for `Solar-10b` and `Mixtral-60b`, are associated with a previous publication. `Solar-10b` merges two checkpoints of Solar (Kim et al., 2023): Solar-Instruct, trained with instruction tuning, and OrcaDPO. `Mixtral-60b` is a mixture-of-experts (MoE) model trained from two Yi-34b checkpoints[6]. Please refer to Appendix D for further details on the models and experimental settings.

5 Results and Discussion

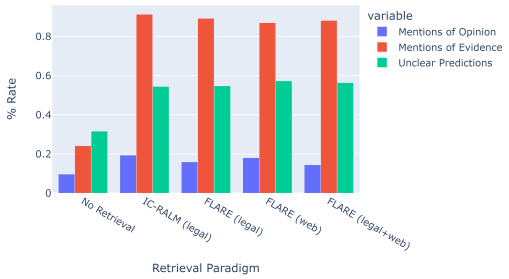
We perform experiments on a wide range of publicly available LLMs. To extend our investigation into the applicability of emerging paradigms for our task, we include baselines of Retrieval Augmented Generation (RAG). This formulation also mimics the human-centric methods for MisLC identification. Our results are summarized in Table 2. The overarching observation from our experiments indicates that **the task is still challenging for current language models**, even when augmented with retrieval. This finding emphasizes the need to develop sophisticated methods to advance performance on MisLC. Additionally, we highlight some key observations:

Trends in task performance vary by model, with no pattern in size or pre-training strategy. Performance without retrieval is mostly comparable, if slightly better than the majority class baseline. The only exception to this trend is `Llama2-13b`, which performs signficantly worse. Performance then *decreases* when we add retrieval from our legal database with IC-RALM. Considering the size of our dataset, there is no significant difference between the two retrieval methods for `Mistral-7b`, `Llama2-13b`, and `Mixtral-60b`. However, FLARE outperforms IC-RALM by a significant margin for `Llama2-7b` (+7.2% macro-f1, +8.8% micro-f1), `Solar-10b` (+1.2% macro-f1, +4.3% micro-f1), and `Llama2-70b` (+4.0% macro-f1, +7.3% micro-f1). Both retrieval methods do not benefit the performance, and occasionally performs worse than No retrieval. `Mistral-7b`, `Solar-10b`, `Llama2-13b`, and `Llama2-70b` have relatively consistent performance between the two retrieval sources. However, `Llama2-7b` performance decreases significantly comparing Legal db → web search (-6.5% macro-f1, -11.1 micro-f1), and `Mixtral-60b` experiences an increase (+5.1% macro-f1, +14.1% micro-f1).

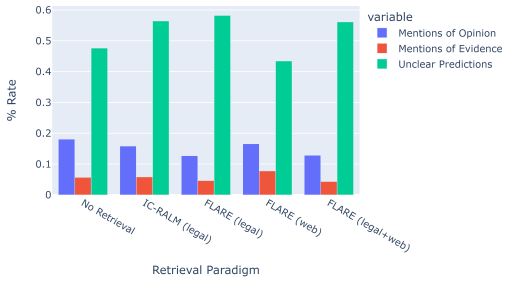
Smaller models drastically decrease in performance when we combine web search and legal retrieval, with `Llama2-7b` performance declining -15.2 points macro-f1 and -21.4 points micro-f1 comparing our full pipeline to zero-shot. The large-sized `Mixtral-60b` sees a similar reduction in performance (-5.1 macro-f1, -22.8 micro-f1), while `Llama2-70b` sees a positive improvement (+1.7 macro-f1, +4.8 micro-f1). `Solar-10b` seems to maintain constant performance, showing marginal improvements in macro- and micro-f1, while `Llama2-13b` has a +4.0 improvement in macro-f1 and +5.0 improvement in micro-f1. It is difficult to determine how significant the `Llama2-13b` result is, considering the baseline is -10.8 macro-f1 and -18.5 micro-f1 worse than the `Random` majority class baseline. However, it is clear that the smallest models have the greatest difficulty utilizing the retrieval context.

Language models struggle to align with our legal context. We analysed the per-class prediction frequency for all of our models in all experiment settings and found most models tended to have

[6] https://github.com/01-ai/Yi



(a) `Llama2-7b`.



(b) `Solar-10b`.

Figure 3: Per-class prediction frequency in our five settings for the best- and worst-performing models with retrieval. We measure the % rate where model generations mention opinion (via the keyword "opinion") and the retrieved evidence (via the keywords "based on").

more `unclear` predictions as they were provided more context. The most striking example of this is `Llama2-7b`, as shown in Figure 3. `Solar-10b`, our best-performing model, does not seem to change significantly in prediction label distribution, which also indicates retrieval is not necessarily decreasing uncertainty. The error rate also increases 0.6-3.3% for the open-sourced models when comparing zero-shot to FLARE (legal + web search), and there is a 5.9% increase for `GPT-3.5-turbo`.

Our legal annotators make a clear distinction between claim and opinion, so we wish to see if our legal database increases the model's mentions of the keyword "opinion." We also examine whether or not the model response utilizes the retrieved evidence by looking for mentions of the keyphrase "based on" in the model response. Both are also plotted in Figure 3 as Mentions of Opinion and Mentions of Evidence respectively. There is a sharp increase in Mentions of Evidence in retrieval architectures for `Llama2-7b`, but the mentions of Opinion do not increase. For `Solar-10b`, the mentions of evidence and opinion both increase slightly when we incorporate web search results, and decrease again when we combine web search with our legal retrieval. This evidence suggests the current state-of-the-art LLMs struggle to incorporate the legal context into their reasoning.

Ablation Setting	Ma-f1 ↑	Mi-f1 ↑
`Random-legal`	29.0	32.7
`Oracle-legal`	30.6	34.9
`Oracle-web`	33.2	38.9
`Oracle-legal+web`	32.9	37.6

Table 3: Summary of our ablations with `Solar-10b`.

5.1 Ablations

We present two ablations with the FLARE pipeline: **Random-doc**, where we retrieve a random document from the legal dataset, and an **Oracle** setting. In the oracle setting, based on the ground truth legal issues l_i provided by our annotators, we retrieve the most relevant document(s) for an input x_i. If there are no legal issues, we perform retrieval as per our normal pipeline. For evidence e_i, we download the sources provided by legal annotators as HTML files, extract the first 500 characters of text, and concatenate all sources into the prompt. These estimate the lower and upper bound of the effect of retrieval on performance. We experiment with `Solar-10b`, our best-performing model.

As shown in Table 3, the random document does confuse the model, and performance decreases by -4.5% macro-f1 and -8.5% micro-f1. The oracle results also demonstrate improvements in micro-f1 for the legal + web search setting, but we observe a decrease in performance everywhere else. This is likely because, due to our relatively low positive rate, there is not a significant amount of ground truth legal documents to observe. However, the model does not improve significantly with the ground truth evidence.

6 Conclusion

We introduce a new task: Misinformation with Legal Consequence (`MisLC`), building on a body of legal documents from 4 broader legal topics and 12 fine-grained legal issues. We curate a dataset for our task using a two-stage annotation pipeline, first with crowd-sourced workers and then with legal annotators. We adapt existing works in Retrieval-Augmented Generation (RAG) to our task, retrieving from the web as well as our curated body of legal documents. We provide comprehensive baselines based on state-of-the-art LLMs, both open- and closed-source, to benchmark the task, covering a broad parameter spectrum and varying training data. We show the task remains challenging for the existing models, which invites future research efforts to tackle it.

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*.

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.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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.

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.

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 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.[7] Some argue the government should not be the arbitrator of the truth (Ó Fathaigh et al., 2021)[8]. 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. [9]

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).

B 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). 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

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

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

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

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.

C Additional Dataset Details

C.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. We then used Google Translate's language detection function[10] 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[11], 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).

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

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

C.2 Crowd-sourced Annotations

Please refer to Table 5 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 5. 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. There were two annotators that met all requirements, except all samples they labelled were ambiguous — for them, we sent a secondary test to obtain a fair assessment.

C.3 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 τ — instead, we take τ to be the mean loss over the entire dataset, which we indicate with τ_μ. During our filtering process, we find τ_μ to be approximately 0.1 across three rounds.

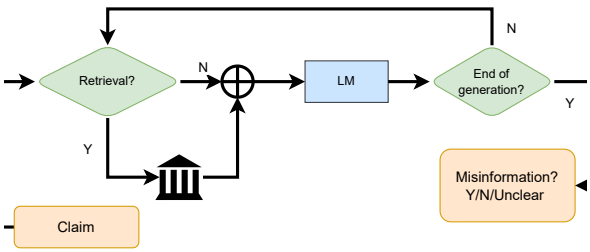
D Additional Experimental Details

D.1 Retrieval

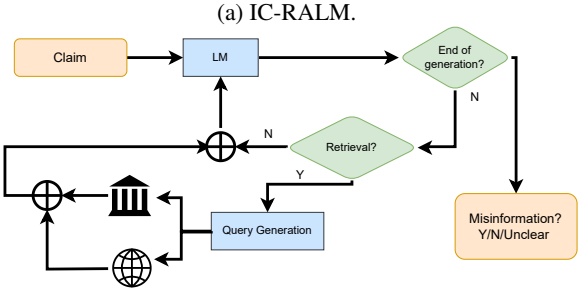
Please refer to Figure 4 for an illustration of the two architectures. We take the implementation of IC-RALM directly from their Github,[12] and take most of FLARE's original implementation[13] except for the generation. We use ChatGPT for query generation in FLARE — we also tested using the model itself 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.

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

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



(a) IC-RALM.



(b) FLARE.

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

Retrieval	Here is some relevant legal context on "misinformation": [doc] Web search results for the claim: [snippets]
Classification	Claim: [claim] Classify the claim as either "factual" or "misinformation."

Table 4: 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.

D.2 Prompting

Please refer to Table 4 for our prompting format. We combine all components of our pipeline through the input prompt, indicating the context of all of our text with simple sentences.

D.3 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 January 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.

- `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).

- `Mixtral-60b` — A mixture-of-experts (MoE) model trained from two Yi-34b fine-tuned checkpoints[14]. The architecture is the same as Llama 2, but the model is fine-tuned with a mix of supervised fine-tuning data and instruction-tuning data.

We choose Llama 2 chat version as this suite of models is trained with strong moral alignments for AI safety, and they are a suite of models trained with the same method at various model sizes. We compare them to Mistral, Solar, and Mixtral 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. Additionally, we compare these models against the closed-source `GPT-3.5-turbo` 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 open-sourced models, we use greedy decoding for the most deterministic prediction, while for OpenAI we use the recommended hyperparameter settings for ChatGPT (temperature = 1, top-p = 1).

D.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

[14] https://github.com/01-ai/Yi, https://github.com/jondurbin/bagel

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

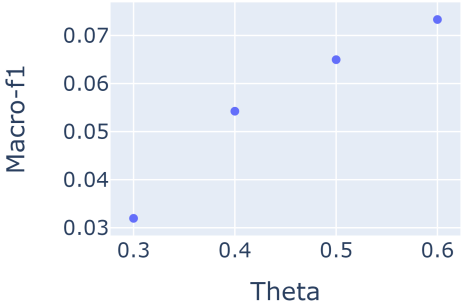


Figure 5: Change in macro-f1 as we increase θ over the first 100 samples.

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

We generate outputs with a maximum generation length of 1024. Experiment run times depended largely on the model size; smaller models took approximately 6 hours on our full dataset, while larger models could take 36-100 hours. This equates to 12 GPU hours for smaller models, or 180-500 GPU hours for larger models. We conducted experiments with open-source models on a server cluster with a combination of Nvidia RTX6000-48GB and A100-40GB GPUs.

E Additional Legal Details

Please refer to Tables 6 for a comprehensive list of legal issues considered in our annotations. We convert the text from PDF to HTML format using Adobe Acrobat.

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 5: Instructions provided to Mechanical Turk workers.

Legal Issue	Key legal tests	Defences
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) 2. Publication (communicated to a third party)	1. Qualified Privilege 2. Responsible Communications 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	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.")[22] 2.Is the government's purpose, or otherwise effect, to restrict expression of this meaning?	1. Can establish the "truth," eg. clinical evidence 2. Non-intent, i.e. published misinformation without intent[23]
Cyberbullying	If false/inaccurate information is being spread to harass or harm others, the spreader could face cyberbullying or harassment charges	
Public Mischief	Every one commits public mischief who, with intent to mislead, causes a peace officer to enter on or continue an investigation by (a) making a false statement that accuses some other person of having committed an offence; (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; (c) reporting that an offence has been committed when it has not been committed; or (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.	
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.	
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 Among other things, food in Canada shall not be sold or advertised in a manner that is false, misleading or deceptive	

Legal Issue	Key legal tests	Defences
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 - All representations that are false or misleading in a material respect, in any form, are subject to the Competition Act - If a representation could influence a consumer to buy or use the product or service advertised, it is material - 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). - To determine whether a representation is false or misleading, the courts consider the "general impression" it conveys, as well as its literal meaning	
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)	
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. - 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)	
Hate Speech	Fake news affects society as a whole, whereas hate speech harms individuals or members of a specific group(Katevas et al., 2022)	
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.	

Table 6: Areas of law that can be used to indict misinformation published online.