Dataset of News Articles with Provenance Metadata for Media Relevance Assessment

Tomas Peterka
Gymnazium Jana Keplera
xpetto01@gjk.cz

Matyas Bohacek
Stanford University
maty@stanford.edu

Abstract

Out-of-context and misattributed imagery is the leading form of media manipulation in today's misinformation and disinformation landscape. The existing methods attempting to detect this practice often only consider whether the semantics of the imagery corresponds to the text narrative, missing manipulation so long as the depicted objects or scenes somewhat correspond to the narrative at hand. To tackle this, we introduce *News Media Provenance Dataset*, a dataset of news articles with provenance-tagged images. We formulate two tasks on this dataset, location of origin relevance (LOR) and date and time of origin relevance (DTOR), and present baseline results on six large language models (LLMs). We identify that, while the zero-shot performance on LOR is promising, the performance on DTOR hinders, leaving room for specialized architectures and future work.

1 Introduction

Over the last few years, the use of manipulated imagery for disinformation and misinformation has grown steadily (Dufour et al., 2024; Shen et al., 2021; Weikmann and Lecheler, 2023; Wang et al., 2024). Many believe this is largely due to the abundance of AI-powered tools that allow users to edit or generate media from scratch, including images (text-to-image (Baldridge et al., 2024; Bie et al., 2024; Ramesh et al., 2021), in-painting (Liu et al., 2023; Lee et al., 2021)), audio (text-to-speech (Eskimez et al., 2024; Chen et al., 2024; Łajszczak et al., 2024), voice cloning (Qin et al., 2023; Luong and Yamagishi, 2020)), and video (deepfakes (Pei et al., 2024; Stanishevskii et al., 2024; Croitoru et al., 2024), text-to-video (Singer et al., 2022; Zhang et al., 2025)). These tools have not only become easily accessible online but also increasingly intuitive to use, often requiring only textual descriptions (Rombach et al., 2022). Consequently, a large body of work has emerged focusing on the detection of AI-manipulated or AI-generated content (Nguyen et al., 2022; Farid, 2022).

However, despite the proliferation of AI tools, a simpler form of image-based manipulation remains prevalent in misinformation and disinformation (Garimella and Eckles, 2020): the use of out-of-context or misattributed imagery to frame events in misleading ways (Fazio, 2020). For example, in April 2020, images of body bags from Ecuador were falsely presented as deceased COVID-19 patients in New York hospitals (News Literacy Project, 2025), sparking confusion and controversy online. Studies indicate that this type of manipulation appears in over 40% of online misinformation containing images, whereas AI-generated media is used in approximately 30% (Dufour et al., 2024).

Despite this, the literature has not responded to the threat of out-of-context and misattributed imagery with the same urgency as AI-manipulated and AI-generated content. As a result, there is a scarcity of specialized resources—methods, tools, datasets, and benchmarks—for studying this phenomenon from the perspective of natural language processing (NLP). Some existing work evaluates whether an image is relevant to the article in which it appears, it primarily considers whether the depicted object or scene aligns with the textual narrative (Aneja et al., 2021). While this analyzes one aspect of media-based manipulation, it misses cases where the imagery and text appear semantically consistent but were captured at times or places that may be irrelevant or outright deceptive.

Peterka and Bohacek (2025), therefore, suggest a new formulation of this task. Rather than asking *"Is this image relevant to the news story?"*, they instead ask *"Was this image captured at a time and place that is relevant to the news story?"*. To this end, they hypothesize that provenance metadata—a record of a file's existence from its creation through edits to distribution—could help answer this ques-

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 114–127

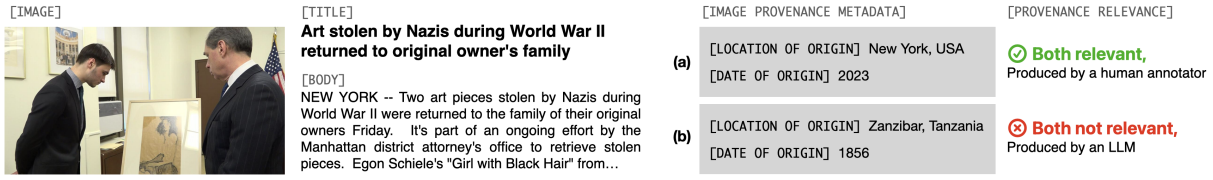

Figure 1: Representative example of a news article from the *News Media Provenance Dataset* with a structured title, body, and image. This article appears in the dataset multiple times with alternative image provenance metadata, shown on the right. (a) One data point contains provenance metadata that was produced by a human annotator to match the relevance of the article. (b) Another data point contains provenance metadata that was randomly produced by an LLM to not match the relevance of the article. The article is sourced from CBS News.

tion. Hence, they conduct some exploratory experiments with large language models (LLMs) to analyze the metadata of images used in news articles. However, they identify two major limitations: (1) the absence of a benchmark dataset for this task and (2) the early-stage adoption of provenance metadata among news outlets, restricting robust evaluation.

In response, we introduce a dataset of news articles with provenance-tagged images and annotations regarding their relevance to the article. Since the news outlets from which the articles were sourced do not yet incorporate provenance metadata (consistent with the limitation identified above), we simulate it. Specifically, we gather annotations for relevant locations and dates and embed them into the images using C2PA (Rosenthol, 2022), a widely used provenance metadata library. We then use an LLM to generate alternative, non-relevant dates and locations, constructing a balanced dataset containing relevant, partially relevant, and irrelevant images based on provenance.

While provenance metadata is not limited to images, our dataset and evaluations focus exclusively on news articles with images. Other modalities, such as video or audio, are not included, since the modality of the file from which provenance metadata is extracted does not affect the included information.

The primary contributions of this paper can be summarized as follows:

- We introduce the first news dataset with provenance metadata-equipped images, *News Media Provenance Dataset*, and open-source[1] it for research use.
- We propose two provenance-based tasks with applications beyond news and authenticity analysis: (1) location of origin relevance (LOR) assessment and (2) date and time of origin relevance (DTOR) assessment.
- We report baseline results of six LLMs and detail a qualitative assessment of their shortcomings, with fully open-sourced[2] experimental scripts and prediction data.

2 Related Work

This section reviews existing NLP literature connected to image and video relevance assessment in news articles. First, we provide an overview of the broader area of study, which involves news articles in NLP. We then proceed specifically to existing work on image and video relevance and data provenance.

2.1 News-Specific Tasks and Datasets

News articles have become a productive subject of study in the NLP community, as they are largely abundant, reflective of current discourse, and invite many direct applications of NLP technology. We categorize some of the most prominent works in this domain by the nature of their task.

2.1.1 Text Classification

There is a robust body of work pertaining to news article classification—spanning topic categories, sentiment analysis, political tendencies, and more. Prominent datasets for this task category include AG News (Gulli, 2005) with 120,000 articles, 20 Newsgroups (Lang, 1995) with 18,000 articles, Reuters-21578 with 21,000 articles focused on finance, News Category Dataset (Misra, 2022) with 210,000 articles from HuffPost, Multilabeled News Dataset (MN-DS) (Petukhova and Fachada, 2023) with 10,000 articles across 215

[1] https://huggingface.co/datasets/matybohacek/News-Media-Provenance-Dataset

[2] https://news-provenance.github.io

news sources, and KINNEWS/KIRNEWS (Niyongabo et al., 2020) with 3,000 tailored for low-resource African languages.

2.1.2 Summarization

Another prominent task involving news articles is summarization, attempting to reduce the full article body into a concise abstract while preserving the core information value. Prominent datasets for this task category include CNN/DailyMail (Hermann et al., 2015) with 287,000 article-highlight pairs, NEWSROOM (Grusky et al., 2018) with 1.3M articles across 38 news sources, CCSUM (Jiang and Dreyer, 2024), with 1.3M articles, and SumeCzech (Straka et al., 2018) with 1M Czech articles.

2.1.3 Disinformation Detection

In the last few years, disinformation detection (also referred to as fake news detection) has emerged as a productive area of study in the literature. The framing of the problem varies both on the side of category definitions (what constitutes disinformation and how to categorize its severity) and on the side of modeling (approaches range from classification to feature detection to question answering).

Prominent datasets for this task category include the LIAR benchmark (Wang, 2017) with over 12,000 articles, the Verifee dataset (Bohacek et al., 2023) with over 10,000 articles spanning 60 news sources, NELA-GT (Gruppi et al., 2021) with 713,000 articles, and FNC-1 (Slovikovskaya, 2019) with 49,972 articles.

2.2 Image and Video Relevance in News

Next, we review previous work specifically targeting the relevance of imagery in news articles.

Cheema et al. (2023) were among the first to explore computational approaches to modeling this relationship between imagery and news articles with modern NLP techniques. Their work, however, primarily set out to review the landscape of existing methods at the time and assess the overall feasibility of future methods in the area; the paper is, hence, primarily descriptive and does not present a specific dataset or architecture.

Tonglet et al. (2024) materialized many of the dynamics described by Cheema et al. (2023) by using a VLM to ask questions about the thumbnail image, deriving its relevance to the rest of the article. However, these inferences are based purely on LLM predictions, and so imagery presenting semantically relevant events may pass the test even when taken at an irrelevant time or place.

Later, Yoon et al. (2024) proposed CFT-CLIP, a framework evaluating the relevance of thumbnail images with respect to the remaining text based on multimodal embeddings. To that end, they also introduced a curated dataset called NewsTT, which contains 1,000 annotated news image-text pairs with relevance labels. This method, however, only reflects the relevance of an image based on its semantic distance from the text, disregarding when and where the image was taken.

Finally, Aneja et al. (2021) introduced the COSMOS dataset for out-of-context thumbnail image detection, enriched by captions with named entity labels. The authors also proposed a self-supervised architecture tailored to this task. While this dataset is concerned with the relevance of media in news articles, as are we, it is, yet again, based on semantical consistency or divergence between the semantics of the image and its caption.

2.3 Data Provenance

Moving beyond semantics inferred from pixels, data provenance can offer information about the origin, evolution, and ownership of a piece of data. While specific implementations of data provenance metadata vary in the covered scope of information, underlying transaction mechanisms, and security guarantees, most existing frameworks include the location and date/time of origin of the data. The framework that has recognized the most adoption by social media platforms, newspapers, and tech companies to date, as compared to alternatives, is C2PA (Rosenthol, 2022), which we adopt in this paper.

While C2PA offers advantages such as guarantees of cryptographic security and unstrippable metadata technology, it has multiple limitations (Longpre et al., 2024; Coalition for Content Provenance and Authenticity (C2PA), 2023). The primary limitation hindering adoption is that most digital content today lacks C2PA provenance metadata. As a result, any analysis dependent on C2PA remains infeasible for the majority of online content. While this may be prohibitive for existing consumer-facing applications, the adoption of C2PA and similar frameworks has been increasing, and so we can expect that, in the future, such analysis will be feasible.

Given the cryptographic guarantees for establishing the trace of an image or a video, which prove-

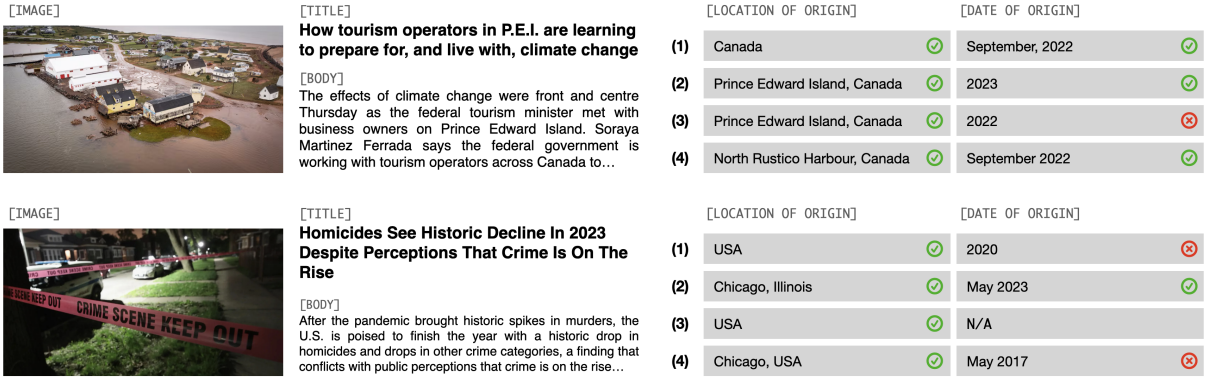

Figure 2: Examples of images from the *News Media Provenance Dataset* used to evaluate annotator reliability. All four annotators provided the location and date of origin for each image, with their accuracy indicated on the right. The article at the top is sourced from CBC and the article at the bottom is sourced from Forbes.

nance metadata enables, it seems highly desirable for relevance assessment of imagery in news articles. To the best of our knowledge, no datasets or resources currently exist for evaluating provenance in news articles.

3 News Media Provenance Dataset

This section presents the *News Media Provenance Dataset*, comprising 637 news articles with simulated image provenance metadata, which is labeled either as *relevant* or *not relevant*. The provenance is inserted into the images using the C2PA library (Rosenthol, 2022) by us: the relevant information is provided by annotators and the not relevant is generated using an LLM. Two example data points are shown in Figure 1.

3.1 Dataset Construction

This section reviews the dataset construction including data sourcing, filtering, and annotations management. The code used for these tasks is fully open-sourced[3]. Any modifications to default library behavior mentioned below are further expanded upon in the documentation of the code release.

3.1.1 Data collection

A list of news article URLs was obtained from the the Webz.io News Dataset Repository (Webhose.io, 2024) in November 2024. Newsarticle4k (Ou-Yang, 2013; AndyTheFactory, 2023) with custom extensions was then used to loop over these article URLs (in randomized order), extracting structured information from the website: the title, body, main image, and its caption. This loop terminated once 200 news articles were successfully scraped.

3.1.2 Annotation Procedure

Four annotators were recruited through Prolific to simulate relevant image provenance metadata for the 200 scraped articles. Out of these annotators, two were male and two were female, ranging in age from 23 to 31. All were based in the United States and we paid them 12 USD per hour.

Each annotator was assigned 55 articles. The first five were shared across all annotators for annotator reliability evaluation; the remaining 50 were unique to the annotator.

The annotations were facilitated through the Argilla[4] tool. Representative screenshots of the tool are presented in Figures 6-8 (Appendix D). It took the annotators, on average, 60 minutes to annotate all the assigned articles. This excludes the time spent familiarizing themselves with the annotation instructions and set up the interface.

3.1.3 Annotation Reliability

The annotator reliability was evaluated on the first five articles which were assigned to all annotators. The annotators provided the correct location of origin in 80% of the cases and the correct date of origin in 56% of the cases.[5]

Examples of these articles alongside annotator responses are shown in Figure 2. The article at the top had an solid annotator performance; the article at the bottom had a somewhat poor annotator performance on the date of origin. Note that the

[3] https://news-provenance.github.io

[4] https://argilla.io

[5] This discounts cases in which the user deemed the attribute as ambiguous and responded with N/A. We allowed a ± 1 buffer for the date of origin units.

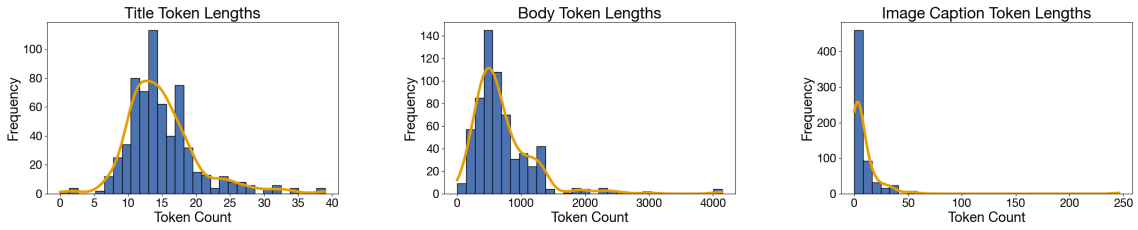

Figure 3: Distribution of the title, body, and image caption token length in the *News Media Provenance Dataset*. A fitted Kernel Density Estimation (KDE) is shown in orange. Outliers were manually reviewed to prevent scraping issues.

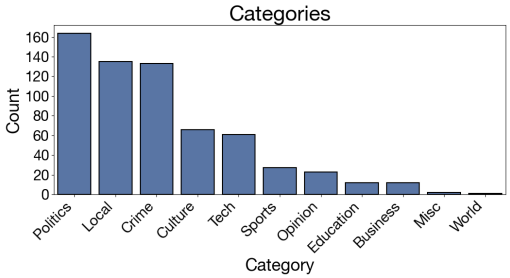

Figure 4: Distribution of categories in the *News Media Provenance Dataset*. A single news article (data point) is represented only once by its primary category.

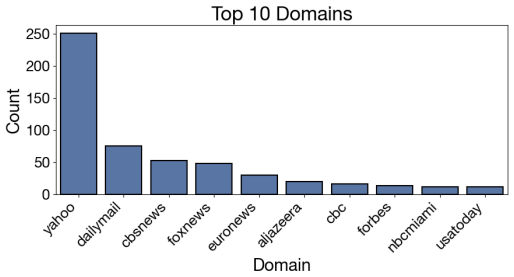

Figure 5: Distribution of source domains in the *News Media Provenance Dataset*, showing the top 10 domains.

level of detail of both the provided location and date differ; as long as all components match, the response is deemed as correct.

3.2 Alternative Provenance Generation

While the annotations served to simulate provenance metadata where both the location and date and time of origin are relevant to the articles, ChatGPT-4o (Hurst et al., 2024) was used to simulate additional provenance metadata that were not relevant to the article. With the prompt presented in Appendix B, the model was asked to generate three additional data points:[6] two data points where one of the provenance metadata fields is not relevant but the other is kept intact, and one data point where both provenance metadata fields are not relevant.

3.3 Dataset Statistics

In total, the dataset contains 637 news articles. Their length statistics are shown in Figure 3. The average length of the headline, body, and image caption, calculated with NLTK (Bird, 2006), are 15, 705, and 9 tokens, respectively.

[6]If either annotation was N/A, then the generation of respective matches (that are not relevant to the article) was skipped.

The top-10 domains by absolute article count are `yahoo`, `dailymail`, `cbsnews`, `foxnews`, `euronews`, `aljazeera`, `cbc`, `forbes`, `nbcmiami`, and `usatoday`, as shown in Figure 3. There appears to be an imbalance of `yahoo`-domain articles. We investigated this, but found that it is because `yahoo` republishes news articles from other domains, and that the actual source distribution among these articles is diverse. We, hence, did not pursue any balancing remedies.

The category statistics, as predicted by a one-shot text classification model (Lewis et al., 2019), are shown in Figure 4. The majority of articles in the dataset fall within the category of Politics, Local, and Crime news.

3.4 Proposed Tasks

We propose two tasks on the dataset: *Location of Origin Relevance (LOR)* assessment and *Date and Time of Origin Relevance (DTOR)* assessment. Note that, while the image was presented to the annotators, these tasks do not assume access to the image. The purpose of these tasks is not to assess whether the semantics of the image (inferred from the pixel space) are relevant to the topic, but rather whether the circumstances, in which the image was captured, are relevant to the presented article.

Model	Feature-level		Article-level		
	LOR	DTOR	2 corr	1 corr	0 corr
ChatGPT-4o	**0.81**	0.57	**0.45**	0.47	0.08
DeepSeek V3	0.69	0.56	0.36	0.54	0.10
Gemma 2 27B Instruct	0.77	**0.58**	0.41	0.53	0.06
Llama 3.1 8B Instruct	0.64	0.42	0.24	0.57	0.19
Mistral 7B Instruct v0.3	0.73	0.47	0.32	0.56	0.12
Phi 3.5 Vision Instruct	0.64	0.48	0.30	0.53	0.17

Table 1: Accuracy of baseline LLMs on the newly proposed LOR and DTOR (feature-level) tasks using the *News Media Provenance Dataset*. The article-level statistics indicate the proportion of articles where both LOR and DTOR predictions were correct (2 corr), one of the predictions was correct (1 corr), and no prediction was correct (0 corr).

3.4.1 Location of Origin Relevance (LOR)

The LOR task comprises the following: given the main image's location of origin found in the provenance metadata, determine whether the image is relevant to the article (represented as title and body).

3.4.2 Date and Time of Origin Relevance (DTOR)

The DTOR task comprises the following: given the main image's date and time of origin found in the provenance metadata, determine whether the image is relevant to the article (represented as title and body).

4 Baseline Models

We evaluate the following off-the-shelf LLMs to establish baseline results: ChatGPT-4o (Hurst et al., 2024), DeepSeek V3 (Liu et al., 2024), Gemma 2 27B Instruct (Team et al., 2024), Llama 3.1 8B Instruct (Dubey et al., 2024), Mistral 7B Instruct (Jiang et al., 2023), and Phi 3.5 Vision Instruct (Abdin et al., 2024). These are some of the most prominent models in the community, chosen based on their popularity on Hugging Face Transformers (Wolf, 2020) and overall benchmark performance at the time of writing.

Note that the parameter size and training scope of these models vary, and one can, of course, expect the larger models to outperform the smaller ones. For example, it is reasonable to expect that ChatGPT-4o or Deepseek V3 will outperform the much smaller Llama 3.1 8B Instruct. The results of this analysis should serve as a baseline for future work investigating methods designed specifically for LOR and DTOR.

The ChatGPT-4o inference was performed using OpenAI's API. The remaining models were implemented using the Hugging Face Transformers (Wolf, 2020) library. To preserve some comparability across models, all inference parameters were left at their default values, and thus mimicking off-the-shelf use. The full prompt is presented in Appendix B.

The binary responses to LOR and DTOR were converted from text to corresponding boolean representations. Whenever the LLM returned a response that did not conform to the JSON format specified in the prompt, the inference was repeated. The inference code and prediction data is open-sourced[7] to maximize reproducibility.

5 Evaluation

This section presents both quantitative and qualitative results of the baseline models evaluated on the *News Media Provenance Dataset*.

5.1 Quantitative Evaluation

Table 1 presents the LOR and DTOR accuracy for all evaluated models. LOR performance ranges from 64% to 81%, with the highest accuracy achieved by ChatGPT-4o. Close behind are Gemma 2 27B Instruct at 77%, Mistral 7B Instruct v0.3 at 73%, and DeepSeek V3 at 69%. Llama 3.1 8B Instruct and Phi 3.5 Vision Instruct both attain an accuracy of 64%.

DTOR performance ranges from 42% to 58%, with the highest accuracy achieved by Gemma 2 27B Instruct. Compared to LOR, accuracies on this task are generally lower, indicating that determining the relevance of date and time of origin is more challenging than assessing location relevance. While the three larger models—ChatGPT-4o, DeepSeek V3, and Gemma 2 27B In-

[7] https://news-provenance.github.io

struct—performed comparatively on this task (with ChatGPT-4o and DeepSeek V3 achieving 57% and 56%, respectively), the smaller models—Llama 3.1 8B Instruct, Mistral 7B Instruct v0.3, and Phi 3.5 Vision Instruct—scored at or below 50%. This suggests that their representations or architectures either fail to capture relevant signals for this task or even lead to counterproductive predictions.

Table 1 also reports aggregate article-level performance for the evaluated models. The best-performing model in this regard is ChatGPT-4o, which correctly predicted both LOR and DTOR for 45% of the articles, predicted only one correctly 47% of the time, and failed on both 7% of the time. Overall performance across models is relatively low, primarily due to poor DTOR performance, which significantly impacts the aggregate metrics.

5.2 Qualitative Evaluation

For ChatGPT-4o, the best-performing model in the above quantitative evaluation, we manually reviewed all article analyses produced during the evaluation. In two follow-up questions, we asked the model to support its reasoning. This approach aimed to better understand the model's ability to reason about LOR and DTOR.

To illustrate the overall trend we observed across many articles, consider an article titled *"N.Y. sushi restaurant owner out of coma after dine-and-dash attack over $425 bill"*, published on NBC News (see Appendix C). When presented to ChatGPT-4o for evaluation with relevant location metadata (Albany, USA) and date and time metadata (May, 2024), it predicted yes for LOR (which is correct) and no for DTOR (which is incorrect).

We asked the model to provide reasoning on the LOR analysis as follows:

LOR Follow-up Prompt

Provide reasoning for your LOR assessment.

The model responded with the following:

LOR Reasoning Response

The image was taken in Albany, USA, which is directly relevant to the news story. The incident occurred in Albany, specifically outside Shogun Sushi, making the location pertinent to the events described.

This suggests that the model directly compared the location described in the article to the provenance metadata, as intended. This example reflects a broader pattern we observed in the data: the model was usually able to extract relevant locations and compare them to the provenance metadata according to the instructions.

We followed with another question:

DTOR Follow-up Prompt

Provide reasoning for your DTOR assessment.

The model responded with the following:

DTOR Reasoning Response

The image was taken in May 2024, which is a bit ambiguous in terms of relevance. While the attack occurred on May 29, the story also mentions events happening after that—Su Wen waking from a coma and being moved out of ICU occurred later in June 2024. Since it is unclear whether the image was taken before or after the attack, and the article focuses heavily on recovery updates after the incident, the timing is not conclusively relevant to the most significant developments in the article.

This indicates that the model made additional assumptions about the task, while lacking the ability to distinguish between the article's temporal perspective and that of the analysis.

This reflects a broader tendency we observed: the model was largely unable to reason soundly about dates and timelines, leading to many incorrect inferences. It also often conflated days of the week with months and years in a way that is not relevant to DTOR analysis.

6 Discussion

This section first discusses the significance of the results attained and then shifts to examining the broader implications of our dataset and the newly proposed tasks.

The quantitative results presented in Section 5.1 suggest that current LLMs can perform LOR assessment in a zero-shot setting and achieve strong baseline performance. This is not the case, however, for DTOR, where the evaluated models struggled to achieve accuracy above 50%. These findings

are supported by our qualitative results, reported in Section 5.2, which show that while the models can reason soundly about the location presented in the article, they struggle with reasoning about dates and timelines. This highlights a broader limitation of LLMs and underscores the need for further research into improving temporal reasoning capabilities.

In addition to challenges with representing time, we also observed that more recent news articles were often more difficult for the models to reason about. We hypothesize that this may stem from the nature of the models' training, as the most recent events are typically not included in their training datasets, making it harder for them to process or contextualize such information.

As expected, larger models outperformed smaller models in our evaluation. The performance of each model could likely be improved by optimizing its parameters and customizing the instruction prompts. We, however, chose to pursue minimal optimization to maintain a level of comparability necessary for measuing baseline results. The relatively low baseline performance nonetheless reinforces the need for developing new architectures tailored to the LOR and DTOR tasks.

We expect our dataset to play a critical role in this effort, as, to the best of our knowledge, there are no other datasets explicitly designed for the tasks of LOR and DTOR. Expanding the dataset to include non-Western news contexts and additional languages will also be essential to ensure inclusive support for underserved communities, who are often at greater risk of media manipulation.

7 Limitations

Despite the benefits of provenance metadata for assessing the relevance of media in news articles, some limitations remain. One major issue is that, even when an image or video presented alongside an article matches the scope and timeline of the story, the article can still be inaccurate or outright manipulative. We, therefore, see our method as just one tool that should be a part of a broader suite of techniques aimed at discerning problematic practices in news articles.

C2PA, the employed provenance metadata framework, also has some drawbacks. Older photos usually lack provenance data, limiting the use of our method on historical images. Moreover, there are articles in which the presence of time- and location-matched media is not necessarily an indicator of relevance. An example of this would be articles reporting on events without clearly bounded locations and/or time frames, such as natural disasters, which often span broad regions and extended periods. Additionally, certain media can be used for illustrative purposes, where strict provenance alignment is less critical to the integrity of the article (e.g., historical illustrations or generic portraits). In such cases, assessing metadata relevance requires a more flexible, nuanced approach. Future work could explore automatic methods for detecting when precise alignment is necessary. Furthermore, as C2PA is still a new technology, its adoption among media organizations is still limited. With many outlets pledging to join, however, its use is expected to grow.

8 Ethical and Societal Implications

The use of provenance metadata for assessing the relevance of media in news articles raises ethical concerns pertaining privacy. Embedding provenance metadata includes potentially sensitive information, such as location and device information, that could put journalists and activists reporting from unsafe regions at risk. Sharing any information that could reveal identity or location of individuals in such contexts may be undesirable and, we believe, should take priority over establishing trustworthy news channels.

This also leads to a broader point, which we wish to highlight. Even though we gathered feedback on our approach from both practitioners and scholars of journalism, there may be additional implications for journalists and their readers. We, therefore, recommend that before this method (or its derivatives) are put in use at a news organization, they should be first extensively scrutinized by its staff to uncover any additional concerns.

Simultaneously, we remain optimistic that this method will introduce an effective tool to support individuals in an increasingly less credible and transparent information ecosystem. To that end, we believe our dataset will serve as a critical tool to improve and evaluate approaches to LOR and DTOR moving forward.

9 Conclusion

This paper defined the tasks of Location of Origin Relevance (LOR) and Date and Time of Origin Relevance (DTOR) for media (images and videos)

presented alongside news articles, based on their provenance metadata. Since no suitable datasets existed for these tasks, we introduced the *News Media Provenance Dataset*—a collection of news articles with provenance-tagged images—containing both human-annotated relevant metadata and irrelevant metadata generated by a large language model (LLM). We presented baseline zero-shot results for six prominent LLMs and found that, while out-of-the-box LOR performance is strong, DTOR performance remains limited, as models struggle to reason about time relevance and temporal relationships.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AndyTheFactory. 2023. Newspaper4k: Article scraping & curation.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. 2024. Imagen 3. *arXiv preprint arXiv:2408.07009*.

Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A Clifton, et al. 2024. Renaissance: A survey into ai text-to-image generation in the era of large model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.

Matyas Bohacek, Michal Bravansky, Filip Trhlík, and Václav Moravec. 2023. Czech-ing the news: Article trustworthiness dataset for czech. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 96–109.

Gullal S Cheema, Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A Bateman, and Ralph Ewerth. 2023. Understanding image-text relations and news values for multimodal news analysis. *Frontiers in artificial intelligence*, 6:1125533.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Coalition for Content Provenance and Authenticity (C2PA). 2023. Harms, Misuse, and Abuse: Initial Adoption Assessment.

Florinel-Alin Croitoru, Andrei-Iulian Hiji, Vlad Hondru, Nicolae Catalin Ristea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2024. Deepfake media generation and detection in the generative ai era: A survey and outlook. *arXiv preprint arXiv:2411.19537*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, et al. 2024. AMMeBa: A large-scale survey and dataset of media-based misinformation in-the-wild. arXiv:2405.11697.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. 2024. E2 TTS: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE.

Hany Farid. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4).

Lisa Fazio. 2020. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation*, 14(1).

Kiran Garimella and Dean Eckles. 2020. Images and misinformation in political groups: Evidence from whatsapp in india. *arXiv preprint arXiv:2005.09784*.

Maurício Gruppi, Benjamin D Horne, and Sibel Adalı. 2021. Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Antonio Gulli. 2005. The anatomy of a news search engine. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 880–881.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xiang Jiang and Markus Dreyer. 2024. Ccsum: A large-scale and high-quality dataset for abstractive news summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7299–7329.

Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent Van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. 2024. BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier.

Eunhye Lee, Jeongmu Kim, Jisu Kim, and Tae Hyun Kim. 2021. Restore from restored: Single-image inpainting. *arXiv preprint arXiv:2102.08078*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Anji Liu, Mathias Niepert, and Guy Van den Broeck. 2023. Image inpainting via tractable steering of diffusion models. *arXiv preprint arXiv:2401.03349*.

Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Jad Kabbara, and Sandy Pentland. 2024. Data authenticity, consent, and provenance for ai are all broken: What will it take to fix them?

Hieu-Thi Luong and Junichi Yamagishi. 2020. Nautilus: a versatile voice cloning system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2967–2981.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

News Literacy Project. 2025. Covid-19 video taken out of context. Accessed: 2025-02-23.

Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen. 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525.

Rubungo Andre Niyongabo, Hong Qu, Julia Kreutzer, and Li Huang. 2020. Kinnews and kirnews: Benchmarking cross-lingual text classification for kinyarwanda and kirundi. *arXiv preprint arXiv:2010.12174*.

Lucas Ou-Yang. 2013. Newspaper3k: Article scraping & curation. *Newspaper3k: Article Scraping & Curation-Newspaper 0.0. 2 Documentation*.

Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. 2024. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

Tomas Peterka and Matyas Bohacek. 2025. Large language models and provenance metadata for determining the relevance of images and videos in news stories. *arXiv preprint arXiv:2502.09689*.

Alina Petukhova and Nuno Fachada. 2023. Mn-ds: A multilabeled news dataset for news articles hierarchical classification. *Data*, 8(5):74.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Leonard Rosenthol. 2022. C2pa: the world's first industry standard for content provenance (conference presentation). In *Applications of Digital Image Processing XLV*, volume 12226, page 122260P. SPIE.

Cuihua Shen, Mona Kasra, and James O'Brien. 2021. This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. *arXiv preprint arXiv:2101.07951*.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-A-Video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Valeriya Slovikovskaya. 2019. Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. *arXiv preprint arXiv:1910.14353*.

Georgii Stanishevskii, Jakub Steczkiewicz, Tomasz Szczepanik, Sławomir Tadeja, Jacek Tabor, and Przemysław Spurek. 2024. Implicitdeepfake: Plausible face-swapping through implicit deepfake generation using nerf and gaussian splatting. *arXiv e-prints*, pages arXiv–2402.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtskỳ, Vojtěch Hudeček, and Jan Hajic. 2018. Sumeczech: Large czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. "image, tell me your story!" predicting the original meta-context of visual misinformation. *arXiv preprint arXiv:2408.09939*.

Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, and Ximing Li. 2024. Harmfully manipulated images matter in multimodal misinformation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2262–2271.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Webhose.io. 2024. Free news datasets.

Teresa Weikmann and Sophie Lecheler. 2023. Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, 25(12):3696–3713.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yejun Yoon, Seunghyun Yoon, and Kunwoo Park. 2024. Understanding news thumbnail representativeness by counterfactual text-guided contrastive language-image pretraining. *arXiv preprint arXiv:2402.11159*.

Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. 2025. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. *arXiv preprint arXiv:2502.05179*.

A Annotator Instructions

These annotator instructions were posted both in the Prolific participant sourcing interface and in the Argilla annotation tool. The participants reviewed these instructions during paid response time.

Annotator Instructions

This study involves reading short news articles and answering questions about the main images featured in these articles. The questions will ask you to identify the time and location of capture, based on the context provided in the article. The collected dataset will be open-sourced for use in ethical AI training.
Thank you for participating in our study! You will be presented with short news articles and asked to provide information about the images used in these articles. Specifically, for each image, you are asked to identify the most likely time and location of capture based on the article's context and image caption.

Time of Origin

- Provide the month and year when the image was most likely taken (e.g., "February 2024", "November 2010").
- If the month cannot be inferred, provide only the year (e.g., "2024", "2010").
- If the year cannot be inferred, respond with "N/A".

Location of Origin

- Provide the city and country where the image was most likely taken (e.g., "Boston, USA", "Paris, France").
- If the city cannot be inferred, provide only the country (e.g., "USA", "France").
- If the location cannot be determined, respond with "N/A". Your responses should be based on the context of the article. If you cannot safely infer the time or location, please use "N/A".

Annotate all 55 articles.

B Prompts

Alternative Metadata Generation (System Prompt)

You are a generator of places and locations that are absolutely unrelated to those presented.

Alternative Metadata Generation (Inference Prompt)

Give me a place and a time that are absolutely unrelated to the following: `'{ORIGINAL PLACE}; {ORIGINAL TIME}'`. Give your response in the same format: `'{NEW PLACE}; {NEW TIME}'`, and don't say anything else.

Benchmarking (System Prompt)

You are evaluating the relevance and credibility of images and videos attached to news stories.
Below, you will be presented with:

- The title and the body of the article
- The image caption, as presented in the article
- Provenance metadata indicating source location and time of the image

Benchmarking (Inference Prompt)

Here is the data:

- The title: `TITLE`
- The body: `BODY`
- Image caption: `IMAGE CAPTION`
- (Provenance metadata) Image location: `SOURCE LOCATION`
- (Provenance metadata) Image time: `SOURCE TIME`

Analyze the following:

1. Is the location where the image was taken relevant to the news story? Return yes or no.
2. Is the time (year and month) when the image was taken relevant to the news story? Return yes or no.

Respond in the following comma-separated format: `{yes/no}, {yes/no}`. Possible responses include: `'yes,yes'`, `'no,no'`, `'yes,no'`, or `'no,yes'`. Do not enumerate these or add any extra characters.

C Qualitative Results: Article Example

The following is an excerpt of the article used in the qualitative evaluation (Section 5.2). It was published on June 13, 2024, on www.nbcnewyork.com. We include this excerpt under fair use to demonstrate the reasoning abilities of evaluated LLMs on LOR and DTOR.

Title: N.Y. sushi restaurant owner out of coma after dine-and-dash attack over $425 bill
Body: An Albany sushi restaurant owner is slowly showing signs of recovery after a brutal attack outside his restaurant last month. Su Wen, owner and chef at Shogun Sushi in upstate New York, has woken up from a nearly two-week coma and is experiencing increasing periods of consciousness, said Ray Ren, one of the managers at his restaurant...

Provenance Metadata:
Location of Origin: Albany, USA
Date of Origin: May, 2024

D Annotation Tool Screenshots

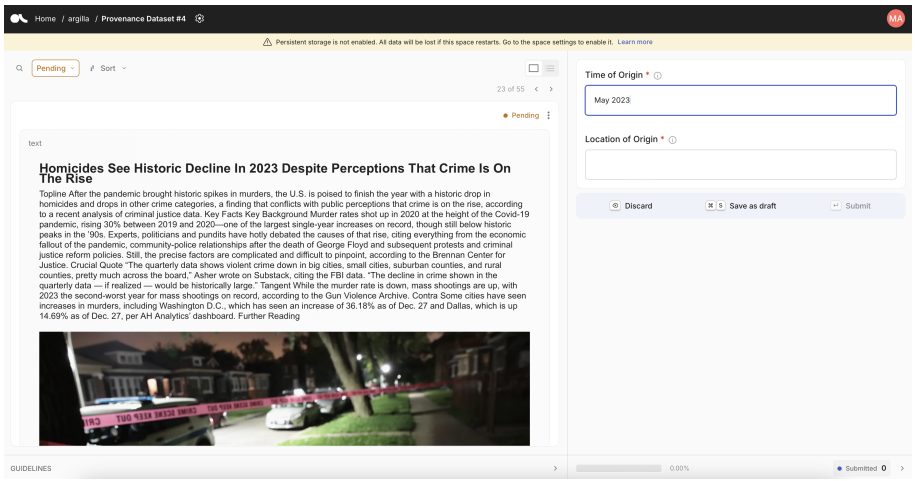

Figure 6: Screenshot of the Argilla annotation tool, focused on an article body.

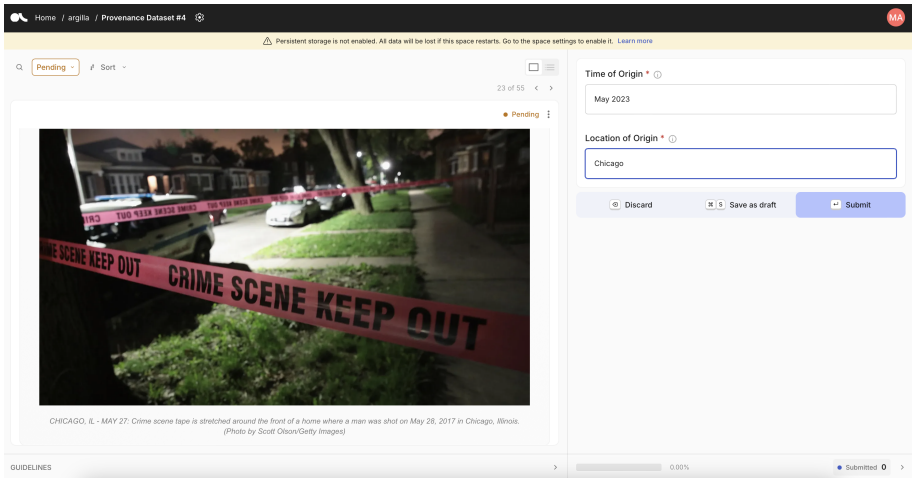

Figure 7: Screenshot of the Argilla annotation tool, focused on an image and its caption.

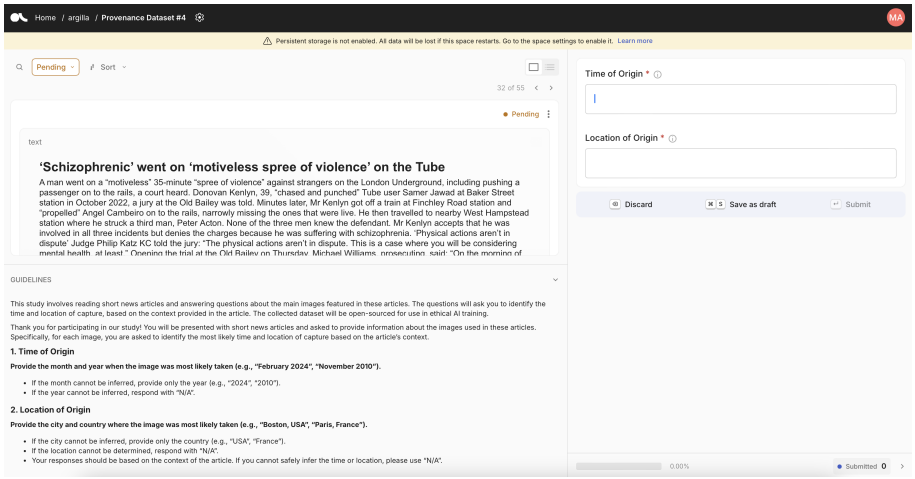

Figure 8: Screenshot of the Argilla annotation tool with the instructions window open.