



Santa Clara
Markkula Center
for Applied Ethics

LLM Performance Benchmark for Journalistic Source Annotations

Evaluating 13 Large Language Models from
Anthropic, OpenAI, Google, Meta, and Deepseek

Subramaniam Vincent, Zhan Shi, Yi Fang
Santa Clara University
V1.0: February 2026

Full dataset available at:

<https://huggingface.co/datasets/subbuvincent/mcae-llms-benchmark-report/tree/main>

EXECUTIVE SUMMARY.....	3
WHY BENCHMARK LLMs FOR JOURNALISTIC SOURCING	4
METHODOLOGY	7
Five Elements of Sourcing Annotations	7
Sourced Statement.....	7
Type of Source	7
Name of Source	7
Title of Source.....	7
Source Justification.....	7
Example of a Sourcing Annotation as Data	8
Ground Truth Dataset Creation	9
Ground Truth Dataset (annotations).....	11
The benchmarking system: three moving parts.....	13
Models tested.....	14
PERFORMANCE RESULTS & SCOREBOARD.....	15
Model Performance Summary.....	15
Interpretation of Findings.....	16
Sourced Statement Accuracy: The Critical Baseline	16
Source Justification Annotation Accuracy: The Significant Challenge	19
Strong Performance: Structured Elements In Sourcing Annotations	23
What can change in these scores and why?.....	27
HOW TO USE THIS BENCHMARKING SYSTEM.....	28
CONCLUSION: THE NEED FOR ONGOING BENCHMARKING.....	29
ABOUT THE AUTHORS	30
ACKNOWLEDGMENTS.....	31
ABOUT THE MARKKULA CENTER FOR APPLIED ETHICS.....	31
APPENDIX 1: Vocabulary for training students and LLMs (system prompt)	31
APPENDIX 2: LLMs user prompt.....	38
APPENDIX 3: A full LLM annotation example for a news story in JSON format.....	41

EXECUTIVE SUMMARY

This benchmark report presents the first systematic evaluation of Large Language Models (LLMs) for annotating journalistic sourcing in news articles. We tested 13 state-of-the-art models (as of Dec 2025)—variants from Anthropic, OpenAI, Google, Meta, and Deepseek—on their ability to identify and categorize source attributions in news stories using a five-element schema. Our ground truth dataset comprises 645 sourcing annotations from 43 professionally published news articles, developed by trained student annotators with inter-coder reliability scores exceeding 0.88.

Key findings reveal significant variance in model performance across annotation tasks. Only two models—Gemini 2.5 Pro and Claude 3.7 Sonnet—achieved 80% accuracy for identifying sourced statements. Accuracy in pulling all statements attributed to sources is a critical baseline metric for any practical application in sourcing analysis. All models performed well (80%+) on structured elements like source type, name, and title. However, every model struggled with source justification annotation, with the highest score reaching only 56%. This capability gap has important implications for ethics-based applications in news analysis and distribution.

This benchmark provides news feed product designers, journalism researchers, news organizations, and AI model developers with performance metrics for evaluating LLM capabilities in analyzing sourcing of news stories at scale. Our benchmarking framework is built around three moving parts: a story corpus, a detailed prompt set (system and user), and the LLMs themselves and their settings. The fourth part is the ground truth dataset, serving as the reference against which model outputs are scored.

We expect and would like the scores in this benchmark to shift, preferably towards greater accuracy for the models. Shifts in any of the three moving parts—expanding the story corpus, refining prompts, or testing newer model variants—could move accuracy numbers up or down. A dedicated section explains what drives these changes and why. The publicly available dataset, prompts, and scoring code support four practical use cases: sourcing analysis of journalists’ own stories, large-scale content research, ground truth data for independent experiments, and full reproduction of the benchmarking system. The complete news article corpus, ground truth annotations, LLM-generated outputs, and prompts are publicly available for ongoing benchmarking as new models emerge.

WHY BENCHMARK LLMs FOR JOURNALISTIC SOURCING

Without *sources*, there is no journalism. Even news influencers, much talked about, do not have a factual basis to jump off into commentary without referring to primary news reporting by journalists themselves. Journalistic sourcing starts with the practice by which reporters decide who or what to source from, why, and how. It includes attributing statements, claims, and information to the people, organizations, documents, or other entities from whom they obtained that content. The sourcing practice manifests in who is included and how they quoted or paraphrased or used for backgrounding. It also refers to how sources are introduced, characterized, justified and portrayed, and how they accorded authority for their knowledge, expertise, experiences and witnessing. Not surprisingly journalistic sourcing involves ethical decision-making, to account for diverse moral pressures: rights of stakeholders, fairness and justice, harm minimization, the common good and so forth.

There is a narrative about the ethics of journalistic sourcing that emerges from critical studies. It shows that the nature of power relations between journalists and sources drives both the strengths and weaknesses of reporters' success with truth-telling. Journalistic power is at the heart of how sources are included, portrayed and legitimated in news stories¹, while journalistic authority itself draws from reporters mediating between sources.² Yet, journalistic sourcing has often defaulted to prioritize what scholars call a “normative order of authorized knowers”—conferring upon sources with social and political power the epistemic authority to characterize social reality, while everyday people offer witnessing authority.³ Over decades, scholars and news practitioners have called for journalists to expand their range of sources beyond the elites and the powerful, as a way to democratize journalistic practice.⁴

Both in response to these calls and to address long-running diversity deficits in story sourcing, centering and inclusion of vulnerable community and marginalized voices, there have been shifts in some newsrooms. In particular journalistic ventures (niche, mission-oriented, social justice-centric)

¹ Vincent, S., Heider Don. (2026). Norms to reconstruct boundaries between sources and storytellers in local news. *Handbook of Media Ethics Research*. Edwin Elgar (upcoming).

² Carlson, M. (2017). *Journalistic Authority: Legitimizing News in the Digital Era*. Columbia University Press.

³ The phrase “normative order of authorized knowers” originates with Fishman, M. (1980). *Manufacturing the News* (pp. 144–145). University of Texas Press. The concept of “witnessing authority” for everyday people is developed in Edy, J. A., Anderson, C., & Tapia, M. H. (2024). News sources and the epistemic authority to represent reality: How journalism constructs an order of authorized knowers in local television news. *Journalism*. <https://doi.org/10.1177/14648849241291727>

⁴ Wenzel, A. D. (2020). Sourcing Diversity, Shifting Culture: Building “Cultural Competence” in Public Media. *Digital Journalism*, 9(4), 461–480. <https://doi.org/10.1080/21670811.2020.1810585>; Thompson, M. E. (2024). Multicultural journalism: Critical reflexivity in news practice (p. 206). Routledge; Varma, A., Limov, B., & Cabas-Mijares, A. (2023). “They Always Get Our Story Wrong”: Addressing Social Justice Activists’ News Distrust Through Solidarity Reporting. *Media and Communication*, 11(4), 286–296. <https://doi.org/10.17645/mac.v11i4.7006>

in the long-tail of news cycles do produce stories that ethics scholars would highlight as more inclusive. But even if there is such supply, it begs the question “so what”, will people see it on their feeds, i.e. the fair distribution⁵ of such stories. That brings up question of reliably making sourcing-related measurements across thousands of stories in news cycles at scale. Consider a major news cycle—thousands of stories published by hundreds of news outlets over days or weeks—covering a systemic issue. Even if a meaningful percentage of stories employed diverse, norms-inspired sourcing approaches, how would news consumers affirmatively discover *those stories* while others using legacy norms dominate attention? Equally, how might third parties monitor news cycles at scale to study the prevalence of diversity in sourcing?

In today's digital news distribution marketplace, ripe with algorithmic aggregators, social media platforms, search and AI-based answer systems, there is little differentiation or labeling of stories by types of sources relied upon, authenticity of sources' connection to story issues, or justifications for source inclusion as stakeholders and impacted groups. This gap in being to distinguish stories from one another around their sourcing represents an opportunity with distinctive facets for both for media ethicists and technologists.

This is where the opportunity for large language models or LLMs comes in. While LLMs have demonstrated substantial language processing capacities, they have not been substantively evaluated for one prospect: using them at scale to generate source-based statistics and signals for news stories clustered in large buckets of content competing for news cycle attention. Potential applications include:

- News aggregator and distribution algorithms that want to differentiate themselves by surface stories based on source diversity and stakeholder representation
- Real-time news cycle monitoring connecting sourcing practices with journalistic norms and epistemics
- Source diversity audits for news organizations seeking to evaluate their sourcing practices
- Critical journalism studies examining sourcing patterns across outlets, topics, and time periods

However, without rigorous benchmarks, it remains unclear which LLMs perform adequately for these applications. Benchmarks may help decide which types of real-time news analysis applications may be low-hanging fruit and which may require LLMs to advance in capacities further before reliable deployment. General-purpose LLM evaluations in areas like law, medicine, and mathematics

⁵ News Distribution Ethics Use Cases, documented as part of a roundtable series of meetings convened at Santa Clara University to bring together product and policy leaders in tech companies engaged in news distribution. <https://www.scu.edu/ethics/focus-areas/journalism-and-media-ethics/resources/the-ethical-distribution-of-news-roundtable-use-cases/>

exist, but journalistic sourcing—a culturally complex, ethically significant, yet everyday practice—has received insufficient attention in benchmark scenario development.⁶

This report makes a first step in addressing that gap by offering a proof-of-concept for a benchmark system. We provide a real-world framework to evaluate LLMs for accuracy in news sourcing annotations by comparing LLM-generated annotations with human ground truth data validated by inter-coder reliability scoring. Our approach demonstrates that systematic benchmarking is possible for this domain, and our publicly available dataset enables ongoing evaluation as new models are released.

⁶ Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1), 140–146. <https://doi.org/10.1111/nyas.15007>; Holistic Evaluation of Language Models (HELM). (2023). Stanford CRFM. <https://crfm.stanford.edu/helm/classic/latest/>; Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12, 39–57. https://doi.org/10.1162/tacl_a_00632

METHODOLOGY

Our benchmark evaluates LLM performance on a fundamental journalistic practice: identifying and categorizing source attributions in news stories. This section describes the annotation schema, ground truth dataset creation, model testing procedures, and comparison methods used to generate accuracy scores.

Five Elements of Sourcing Annotations

We developed a five-element schema for annotating source attributions in journalistic texts. Each sourced statement identified in an article receives annotations across these categories:

Sourced Statement

Every statement in a story that the reporter would not have been able to include without drawing or receiving content from one or more sources. These are actual text lines where reporters make attributions using verbs like "said," "according to," "stated," "explained," or implied attributions.⁷

Type of Source

The category of source being attributed: Named Person, Named Organization, Document, Anonymous Source (known to reporter but identity withheld), Unnamed Person (not granted anonymity), or Unnamed Group of People.

Name of Source

The full name when applicable (for Named Person or Named Organization types). For documents, the publishing organization or author.

Title of Source

Formal positions of power, authority, or expertise held by the source (e.g., "Director of Marketing," "Mayor," "Professor," "Chief of Police"). Titles signal authorized knowledge, power and/or expert status.

Source Justification

⁷ The everyday journalistic practice of attribution — signaled by verbs such as “said,” “according to,” “stated,” and “explained” — is analyzed as a foundational feature of news discourse in: Bednarek, M. (2006). Evaluation in Media Discourse: Analysis of a newspaper corpus. <https://ci.nii.ac.jp/ncid/BA80202804>. These verbs are also referred to in journalism linguistics as “reporting verbs” or “speech-act verbs.”

This term is the unstructured text reporters use to explain in why a source is included—their connection to the story, lived experiences, stakeholder status, or witnessed events. This text often appears closer to where the source is first introduced, but not always. It may appear anywhere in the article.

Added note: Source Justification is a new and less understood term and we have conceptualized it as a new data element to capture an aspect of journalistic sourcing that is crucial for understanding source portrayal and ethical treatment. The practice of justifying source inclusion has been present implicitly in journalistic writing about sources. But it has not been formally defined as a term for structured data purposes, for instance in news source analysis, and hence is not part of any sourcing standardization convention. Yet, it is in the justification text where the exercise of source portrayal, characterization and depiction of people⁸ and communities takes place. We are proposing this term as a new concept to deepen both sourcing annotations datasets for journalism ethics research as well as to benchmark LLMs for spotting them at scale.

For a more detailed description of these and all associated terms, see Appendix 1 – Vocabulary.

Example of a Sourcing Annotation as Data

This example (Table 1) taken from our ground truth dataset shows a fully annotated statement of attribution from a news report. In this story about the impact of a new Florida law on transgender people, a statement is attributed to a named person, with their title. The source justification data field has the text that captures why that person is being quoted.

Sourced Statement	Type of Source	Name of Source	Title of Source	Source Justification (taken from story text)
“Our ability to participate in this democracy in Florida, it’s been on the line for a while,” Diaz said. Transgender people, they said, are being “forced to choose between participating in our greatest civic	Named Person	Quinn Diaz	Public policy associate at advocacy group Equality Florida	Diaz, who is transgender, moved to Florida from Massachusetts, where they had an “X” gender marker on their driver’s license. Because they didn’t have all the necessary documentation and weren’t established yet with a local provider when they

⁸ As journalists adopt recent advancements in journalism ethics frameworks, the opportunity to make source justifications more explicit in stories increases. Also see footnote 14.

right or just sitting out because the state doesn't see us for who we are."				transferred their license to Florida, they forfeited that marker and defaulted to a license with their sex assigned at birth.
---	--	--	--	---

Table 1: Example of a fully annotated attribution from a real news story.

Ground Truth Dataset Creation

Currently, our ground truth dataset comprises 645 sourcing annotations extracted from 43 professionally published news articles. Articles were selected from award-winning local, regional, and national outlets including both for-profit and non-profit publishers. The corpus includes stories from publications such as *Cal Matters*, *MLK50*, *The Oaklandside*, Associated Press, New York Times, and 18 other outlets. See the table below.

News Publisher	Type of site	Number of stories in sample (total=43)	Awards Won Y/N
Regional/Local News			
Cal Matters	CA news	3	Y
Documented NY	Immigration News	1	Y
LAIst	Southern California Public Media	2	Y
MLK50	Tennessee regional	4	Y
The Oaklandside	Bay Area Local News	3	Y
Sacramento Observer	CA regional news, Black-owned	3	Y
Gazetteer SF	SF Local News	2	New publication
The San Francisco Standard	SF Local News	2	Y
VTDigger	Vermont Local News	2	Y
Mercury News	Bay Area Local News	1	Y
The Central Virginian	Virginia regional	1	Y
KOMO News	Seattle, WA local	2	Y
Chicago Sun Times	Chicago Local	1	Y

Non-local/Mission/National/Wire			
19th News	Mission oriented, gender and politics	1	Y
Capital B News	Black-owned	2	Y
Native News Online	Indigenous American	1	Y
Associated Press	National/Syndicated News	5	Y
New York Times	National/International	3	Y
Reuters	National/International	1	Y
Salon.com	National/International	1	Y
Reason.com	News and Opinion	1	Y
Word in Black	News and Opinion	1	Y
Total number of articles		43	

Table 2a: List of publishers and articles in our news corpus.

Undergraduate and graduate students at Santa Clara University were trained to apply the five-element schema through an iterative process. Students first annotated practice articles, which were reviewed collectively to correct misunderstandings. Following training, students annotated 43 production articles over three quarters to reach the final 645-row dataset.

To validate annotation reliability, we subjected 100 rows to inter-coder reliability testing using Krippendorff's Alpha⁹. Two independent student annotators coded the same 6 articles (100 rows of sourcing), and their work was compared using semantic distance for unstructured elements (Sourced Statement, Title, Source Justification) and fuzzy distance¹⁰ for structured elements (Type and Name). All ICR scores exceeded 0.88, with Sourced Statements at 0.9483, Type of Source at

⁹ We use Krippendorff's Alpha to compute inter-coder reliability scores for our annotators. Our use of trained student annotators to produce ground truth data follows a similar approach to Spangher et al. (2023). See: Spangher, A., Peng, N., Ferrara, E., & May, J. (2023). Identifying Informational Sources in News Articles. Proceedings of EMNLP 2023 (pp. 3626–3639).

¹⁰ Semantic distance comparisons use sentence embeddings from: Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP 2019. Similarity is measured via cosine similarity, with a threshold of 0.8 for sourced statements and 0.55 for titles and justifications. Fuzzy distance comparisons use the Levenshtein-based algorithm from: Mouselimis, L. (2021). fuzzywuzzyR: Fuzzy String Matching. R package version 1.0.5. <https://CRAN.R-project.org/package=fuzzywuzzyR>, with a threshold of 0.8.

0.9558, Name of Source at 0.9798, Title at 0.9714, and Source Justification at 0.8810—indicating high reliability in the human annotation.

Ground Truth Dataset (annotations)

Descriptive statistics of various aspects of sourcing in the ground truth annotations dataset are in the table below.

N = 645 sourced statements across 43 news stories. Percentages in sections 1–4 are of the full dataset. Section 5a percentages are of Named Person rows (n=384). Section 5b shows intersectional breakdown across all 645 rows.

Category / Breakdown	n	% of N=645
1. DIRECT QUOTES IN SOURCED STATEMENTS		
Contains direct quote(s)	280	43.4%
Other types of attribution	365	56.6%
2. TYPE OF SOURCE		
Named Person	384	59.5%
Named Organization	109	16.9%
Document	71	11.0%
Unnamed Group of People	58	9.0%
Anonymous Source	17	2.6%
Unnamed Person	6	0.9%
3. TITLE OF SOURCE PRESENCE (ALL 645 ROWS)		
Has a Title of Source	303	47.0%
No Title of Source	342	53.0%
4. SOURCE JUSTIFICATION PRESENCE (ALL 645 ROWS)		
Has Source Justification	319	49.5%
No Source Justification	326	50.5%

Category / Breakdown	n	% of N=645
5A. TITLE PRESENCE \cap NAMED PERSON SOURCES (N = 384)		
Named Person WITH title <i>% of 384 Named Person rows</i>	302	78.6%
Named Person WITHOUT title <i>% of 384 Named Person rows</i>	82	21.4%
5B. TITLE PRESENCE \cap SOURCE JUSTIFICATION PRESENCE (ALL 645 ROWS)		
Has BOTH title and justification	152	23.6%
Has title only (no justification)	155	24.0%
Has justification only (no title)	167	25.9%
Has NEITHER title nor justification	171	26.5%
6. NAMED PERSON SOURCES \cap SOURCE JUSTIFICATION PRESENCE (N = 384)		
Named Person WITH Source Justification <i>% of 384 Named Person rows</i>	214	55.7%
Named Person WITHOUT Source Justification <i>% of 384 Named Person rows</i>	170	44.3%
7. TYPE OF SOURCE DISTRIBUTION — ROWS WITHOUT SOURCE JUSTIFICATION (N = 326)		
Named Person	170	52.1%
Named Organization	83	25.5%
Document	35	10.7%
Unnamed Group of People	35	10.7%
Unnamed Person	3	0.9%
Anonymous Source <i>All 17 Anonymous Sources have a justification</i>	0	0.0%

Table 2b: Sourcing Statistics from the Ground Truth Annotations Data

The benchmarking system: three moving parts

Our benchmarking system that led to this report has three main moving parts.

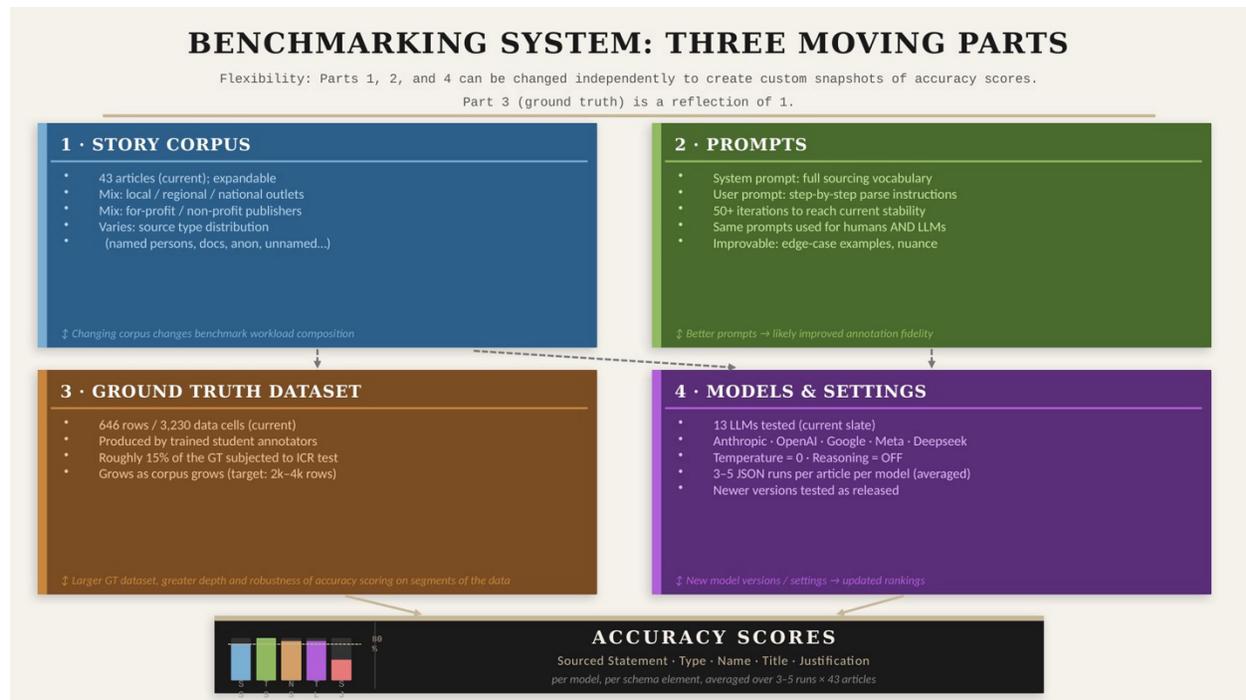


Figure 1: Benchmark System Diagram

Part 1 - Story corpus: We currently have 43 articles selected from various news outlets (Table 2) for annotations. This corpus can be independently expanded to more news stories with a more diverse distribution of the types of sources in them: people, organizations, documents, unnamed groups of people, anonymous sources, and so on.

Part 2 - Prompts: Currently we have a detailed set of prompts (system and user) shared with both human annotators and LLMs. These prompts took 50 iterations to build to stability, and yet, they could still be improved further with more examples for nuanced edge cases in annotations.

Part 4 - Models and settings: Currently, we have tested 13 off-the-shelf variants (release versions) with reasoning set to OFF and temperature to Zero. Moving to newer and newer models (e.g. Claude Sonet 4.5 and 4.6 are next) and possibly tweaking settings and/or trying custom models is a separate moving part that will move the scores around – higher or lower. This is the goal: testing model performance and comparing them.

.....

Part 3 - Ground truth (GT) dataset: This is a critical part of the benchmarking system, but is not a moving part. It holds the human (trained student) annotations done for the story corpus of news articles, producing 645 rows of data. The is GT dataset is a direct reflection of the corpus through the sourcing vocabulary lens. Our ICR scores show that annotator reliability and consistency is high. This means a given story corpus is broadly going to produce the same GT dataset, subject to minor variations, even if the annotator teams change in their composition.

The project will keep subjecting the 15-20% of the ground truth dataset to inter-code-reliability analysis as it grows to 2k-3k-4k rows. (15-20% is the recommended ICR analysis sample size for scoring.)

Models tested

We tested the following 13 off-the-shelf LLMs for their capacity to parse news story texts, and spot and annotate the five elements of sourcing annotations outlined above.

1. Anthropic: Claude 3.5 Sonnet, Claude 3.7 Sonnet, Claude 3.7 Sonnet-Thinking, and Claude 4 Sonnet.
2. OpenAI: ChatGPT-4o-latest and GPT 4.1, and GPT 5.
3. Google: Gemini Pro 1.5, Gemini Pro 2.5.
4. Meta: Llama 3.1 405B Instruct and Llama 4 Maverick.
5. Deepseek R1-0528, Deepseek-Chat-V3-0324

Released Data: Structure and Organization

The complete benchmark dataset¹¹ is publicly available at Hugging Face and includes:

- 43 news article texts (news articles corpus folder). The text files are numbered 1-43 (e.g., "11-Restoring_rights_felons.txt", "25-Article_title.txt")
- 645 rows of ground truth annotations (students_annotations_GTs folder). These are CSV files with matching names containing student-validated annotations for each article. Associated inter-coder reliability scoring code and data.

¹¹ Dataset: <https://huggingface.co/datasets/subbuvincent/mcae-llms-benchmark-report>

- Annotations from 13 LLMs, 3-5 iterations each (Llm_generated_annotations folder), Organized by model, with 3 iteration CSV files per article per model.
- System and user prompts for LLM testing; this has all the definitions for the sourcing vocabulary in the system prompt, and instructions to annotate in the user prompt.

Our file naming convention (articles numbered 1-43 with matching annotation files) enables cross-referencing between article texts, ground truth data, and model outputs.

PERFORMANCE RESULTS & SCOREBOARD

This section presents comprehensive performance metrics for all 13 evaluated models across the five annotation elements. Scores represent averages, with each article annotated at least three times per model.

Model Performance Summary

Table 1 displays accuracy scores for each model across all annotation elements. Green highlighting (80%+) indicates strong performance, yellow (60-80%) shows moderate capability, and red (<60%) signals areas requiring improvement.

Model	Sourced Statement	Type of Source	Name of Source	Title of Source	Source Justification
Gemini 2.5 Pro	83.66%	88.60%	79.94%	79.81%	44.79%
Claude 3.7 Sonnet-Thinking	80.48%	90.01%	84.79%	84.57%	49.00%
Claude 3.7 Sonnet	80.01%	88.25%	82.67%	82.61%	51.95%
GPT 4.1	78.49%	91.07%	86.61%	85.61%	38.91%
Claude 4 Sonnet	76.76%	91.43%	85.47%	86.03%	56.17%
Gemini 1.5 Pro	76.25%	91.89%	89.81%	89.48%	49.79%
GPT 5	73.72%	90.49%	85.25%	86.66%	41.46%
Deepseek Chat-V3-0324	73.37%	88.83%	84.83%	85.24%	43.99%
ChatGPT 4o-latest	72.72%	93.45%	87.69%	87.53%	41.56%

Deepseek R1-0528	67.46%	92.95%	84.81%	83.97%	40.09%
Llama 4 Maverick	66.59%	90.59%	86.16%	86.09%	27.86%
Llama 3.1-405B	64.90%	90.55%	85.32%	85.26%	37.35%
Claude 3.5 Sonnet	61.82%	92.49%	87.93%	87.34%	49.33%

Table 3. Performance scores for 13 LLMs across five sourcing annotation elements. Green cells (80%+) indicate strong performance; yellow (60-80%) shows some capability but unusable for demanding applications; red (<60%) signals areas requiring substantial improvement before any applications can be considered. Bold values highlight best-in-class scores.

Interpretation of Findings

Sourced Statement Accuracy: The Critical Baseline

The most critical finding from our benchmark concerns sourced statement identification—the foundation upon which all other annotations depend. Only two models achieved the 80% accuracy: Gemini 2.5 Pro (83.66%) and Claude 3.7 Sonnet (80.01%), with Claude 3.7 Sonnet-Thinking close behind at 80.48%.

An 80% accuracy score is a rough threshold we assert could be minimum viable point for practical applications that require enumeration of all sourced statements from a complex story with a variety of different types of attributions and source types. Reliable enumeration of attributions is key for usability in downstream analysis. The performance gap between the top two models and the remainder is substantial—several models scored below 70%, with Claude 3.5 Sonnet at 61.82% and Llama 3.1-405B at 64.90%.

The challenge lies in journalistic language’s complexity. Reporters use varied attribution styles—direct quotes, indirect speech, paraphrasing, and implied attributions using speech-act verbs beyond simple “said” statements.¹² Models must distinguish between a reporter’s own analysis and content drawn from sources, even when attribution verbs are subtle or contextually dependent. Our results suggest that current off-the-shelf models vary widely in this enumeration capacity.

¹² Journalists’ use of varied attribution forms—direct quotes, indirect speech, paraphrasing, and implied attributions via speech-act verbs—is documented in: Bednarek, M. (2006). Evaluation in Media Discourse: Analysis of a newspaper corpus. <https://ci.nii.ac.jp/ncid/BA80202804>; and Jaakkola, M. (2018). Journalistic writing and style. Oxford Research Encyclopedia of Communication. <https://doi.org/10.1093/acrefore/9780190228613.013.884>. Direct quotes represent only one form of manifest sourcing, with journalists drawing on a wide range of attribution types. An example of quoting data patterns analysis is in: Vincent, S., Wu, X., Huang, M., & Fang, Y. (2023). Could Quoting Data Patterns Help in Identifying Journalistic Behavior Online? International Symposium on Online Journalism, 13(1).

We present some examples comparing LLM annotations that identify sourced statements with the ground truth data. As noted earlier, accuracy in identifying sourced statements is the foundational metric—models must correctly identify which text constitutes an attribution to a source before any downstream annotations (source type, name, title, justification) can be used as data for other analyses. We show two examples where models successfully captured the exact statement, and two where models completely missed the statement.

Comparative Examples showing accurate Sourced Statement annotations and failures

Example 1: Article 11 (Restoring Voting Rights)

Ground Truth	Gemini 2.5 Pro	Claude 4 Sonnet
Voting, King said in an interview, gives "a little bit of your strength back and a little bit of your voice back. Being able to vote, being able to have a say in what happens in your society, in your state, is extremely important."	Voting, King said in an interview, gives "a little bit of your strength back and a little bit of your voice back. Being able to vote, being able to have a say in what happens in your society, in your state, is extremely important."	Voting, King said in an interview, gives "a little bit of your strength back and a little bit of your voice back. Being able to vote, being able to have a say in what happens in your society, in your state, is extremely important."

Table 4

In this case the model data match the ground truth exactly. The statement uses the explicit attribution verb "said" with clear quote marks, making it straightforward to identify. Our overall observation is that models do not struggle with identifying explicit, properly punctuated attributions with standard reporting verbs. Both Gemini 2.5 Pro and Claude 4 Sonnet were the higher-performing models for this metric.

Example 2: Article 11 (Restoring Voting Rights)

Ground Truth	Gemini 2.5 Pro	Claude 4 Sonnet
More than 4.6 million people are disenfranchised in the United States because of felony convictions, according to the Sentencing Project, which studies the issue and advocates for restoration of voting rights for former felons.	More than 4.6 million people are disenfranchised in the United States because of felony convictions, according to the Sentencing Project, which studies the issue and advocates for restoration of voting rights for former felons.	More than 4.6 million people are disenfranchised in the United States because of felony convictions, according to the Sentencing Project, which studies the issue and advocates for restoration of voting rights for former felons.

Table 5

This is similar to example 1. Both models correctly identified this statement using the "according to" attribution phrase. The attribution signal ("according to") followed by the source name is a common pattern in journalistic writing where statistical data or factual claims are attributed to sources – people or organizations. This makes it straightforward for models to identify as a sourced statement.

Inaccurate Examples

The following examples demonstrate cases where the models completely failed to identify sourced statements that trained human annotators recognized. These failures therefore also cascade into missing all downstream annotations (source type, name, title, justification) for these statements.

Example 3: Article 11 (Restoring Voting Rights)

Ground Truth	GPT 4.1		Claude 3.5 Sonnet		Llama 3.1-405B	
Several of them said they were confused by the arrests because they had been allowed to register to vote.	STATEMENT FOUND	NOT	STATEMENT FOUND	NOT	STATEMENT FOUND	NOT

Table 6

All three models completely missed this sourced statement. The attribution uses "Several of them said"—this is an attribution to an “unnamed group of people” without explicit quote marks. This type of attribution is common in journalism--attributing to groups of people as one source collectively, and we created a type of source in our vocabulary called “Unnamed Group of People” for this. This is an area where the models need to improve.

Example 4: Article 11 (Restoring Voting Rights)

Ground Truth	GPT 4.1		Claude 3.5 Sonnet		Llama 3.1-405B	
Wilcot stressed she was speaking as an individual and not on behalf of the university.	WRONG STATEMENT: Kathy Wilcot, a member of the University of Nebraska Board of Regents, was the lone dissenter...		STATEMENT FOUND	NOT	STATEMENT FOUND	NOT

Table 7

This is another failure pattern. Two models (Claude 3.5 Sonnet and Llama 3.1-405B) completely missed this statement. GPT 4.1 found a different statement about Wilcot but not this specific one. The Ground Truth statement uses the reporting verb "stressed" combined with indirect speech (no quotes)—a pattern that appears to confuse the models. This is an indication that models struggle with variations in attribution verbs beyond the most common ones ("said," "according to").

Summary: Sourced Statement Accuracy Findings

We expected find high accuracy for explicit, properly punctuated attributions. When statements use standard reporting verbs ("said," "according to"), and quotation marks. We see challenges for other situations that need a larger dataset and further exploration. For e.g. models miss sourced statements without quotation marks, using indirect speech, or attributing to unnamed groups ("Several of them said").

When a sourced statement is missed, it means zero downstream annotations (no source type, name, title, or justification) for that attribution. Relatively high-performing models like Gemini 2.5 Pro (83.66% sourced statement accuracy) and Claude 3.7 Sonnet (80.01%) are still missing some sourced statements (roughly 17% and 20% statements respectively). If those sources represent communities with less formal but significant lived experiences of exclusion or struggle or success, the models may miss their presence in otherwise excellent storytelling.

Source Justification Annotation Accuracy: The Significant Challenge

Every model struggled significantly with source justification annotation. The highest-performing model, Claude 4 Sonnet, achieved only 56.17% accuracy—barely better than chance for this task. Most models scored in the 40–50% range, with Llama 4 Maverick notably lower at 27.86%.

Source justification represents reporters' explanations of why sources are included—their lived experiences, stakeholder status, witnessed events, or community standing. This unstructured text may appear anywhere in an article and requires understanding implicit connections between sources and story issues. More recently, journalists practicing solidarity journalism or multicultural approaches may include additional contextual information about non-expert or non-official sources, explaining their authentic connection to systemic issues.¹³ Current models largely fail to identify this crucial ethical dimension.

¹³ Solidarity Journalism, a framework to help journalists stand for basic human dignity and against suffering, and improve coverage of marginalized communities is developed in: Varma, A. (2025). A Solidarity Framework for Representing Suffering: Resisting a Desensitizing Status Quo of Normalizing Social Injustice. *Journal of Communication Inquiry*, 49(4), 403–420. <https://doi.org/10.1177/01968599251348905>. The

Standard deviation for source justification scores (0.21 to 0.32) exceeded that of sourced statements (0.09 to 0.17), indicating higher swings in detection capability for this attribute. Some stories with explicit source characterization may be easier to annotate, while others with implicit or scattered justifications prove more challenging. This variability has important implications: applications requiring source justification analysis cannot yet rely on off-the-shelf models and will need substantial fine-tuning or alternative approaches.¹⁴

Comparative Examples: Accurate Source Justification Annotations

We cite three examples where the models do a better of annotating source justification text.

Example 1: Article 11 (Restoring Voting Rights)

Sourced Statement:

More than 4.6 million people are disenfranchised in the United States because of felony convictions, according to the Sentencing Project, which studies the issue and advocates for restoration of voting rights for former felons.

Ground Truth	Gemini 2.5 Pro	Claude 4 Sonnet
which studies the issue and advocates for restoration of voting rights for former felons.	which studies the issue and advocates for restoration of voting rights for former felons.	which studies the issue and advocates for restoration of voting rights for former felons

Table 8

This annotation match represents ideal performance. The high-performing models (Gemini 2.5 Pro and Claude 4 Sonnet) captured the source justification exactly as it appears in the ground truth. The justification explains the Sentencing Project's organizational mission and expertise, showing the reader why the reporter chose to source a disenfranchisement statistic from their work.

It is possible that this type of mission statement embedded in attribution to organizations as sources is easier for models to identify, than contextual or lived experience justifications for unnamed groups of people – this warrants further study.

Example 2: Article 31 (Vermont Hair Discrimination Bill)

multicultural journalism framework for centering sources' lived experiences and community standing is developed in: Thompson, M. E. (2024). *Multicultural journalism: Critical reflexivity in news practice* (p. 206). Routledge.

¹⁴ On the potential of fine-tuning to improve LLM performance on annotation tasks, see: Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2024). Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1). <https://doi.org/10.1007/s42001-024-00345-9>.

Source:

Aaliyah Wilburn, leader with the Vermont Student Anti-Racism Network

Sourced Statement:

"A senior at North Country Union High School in Newport, Wilburn, 17, recalls being in line for the bathroom when the girl in front turned around and reached for her hair. Despite telling her not to touch it, Wilburn said the girl 'grabbed' her hair."

Ground Truth	Gemini 2.5 Pro	Claude 3.7 Sonnet
Aaliyah Wilburn moved to Vermont two years ago and has faced more hair discrimination here than ever before.	Aaliyah Wilburn moved to Vermont two years ago and has faced more hair discrimination here than ever before. A senior at North Country Union High School in Newport, Wilburn, 17, recalls being in line for the bathroom when the girl in front turned around and reached for her hair.	moved to Vermont two years ago and has faced more hair discrimination here than ever before; A senior at North Country Union High School in Newport, Wilburn, 17

Table 9

This is a statement attribute to a student (Black American) who has the lived experience of hair discrimination. The ground truth source justification establishes this as key reason for her inclusion. Both high-performing models are able to pull this from the text. The difference is that sometimes models will add a part of the sourced statement as well, likely because the source may cite an experience there, and the model may be unable to separate justification text outside of the sourced statement from similar text in the statement itself (Gemini 2.5 Pro).

Example 3: Article 25 (Ballot Access for Transgender Voters)

Sourced Statement:

"I can imagine that a lot of trans folks in Florida who might not have access ... to an accurate and affirming ID might not want to engage in that process at all, especially in such a heightened political environment," Diaz said.

Ground Truth	Claude 4 Sonnet
Diaz, who is transgender, moved to Florida from Massachusetts, where they had an "X" gender marker on their driver's license. Because they didn't have all the necessary documentation and weren't established yet with a local provider when they transferred their license to Florida, they forfeited that marker and defaulted to a license with their sex assigned at birth.	who is transgender, moved to Florida from Massachusetts, where they had an "X" gender marker on their driver's license; they didn't have any problems voting in person in the 2022 midterms, but they plan to vote by mail this November

Table 10

Claude 4 Sonnet captured some source justification text that is present in the GT. It also added extra context from elsewhere in the story.

Inaccurate Source Justification Examples

The following examples demonstrate common failure patterns: The models may completely miss the justification (return “null”), or mistakenly capture text from elsewhere in the story.

Example 4: Article 11 (Restoring Voting Rights)

Sourced Statement:

Voting, King said in an interview, gives a little bit of your strength back and a little bit of your voice back. Being able to vote, being able to have a say in what happens in your society, in your state, is extremely important.

Ground Truth	GPT 4.1	Llama 3.1-405B	Claude 3.7 Sonnet
King's first chance to vote will be in the 2024 presidential election season -- unless a legislative proposal introduced in January that would remove the two-year requirement passes and becomes law.	WRONG TEXT: An outreach specialist with the Nebraska AIDS Project, King came off probation in August	NULL	WRONG TEXT: former felon who had his voting rights restored

Table 11

All three models fail to find the source justification here. GPT 4.1 and Claude 3.7 Sonnet captured different text entirely (title/characterization rather than justification), while Llama returned null. In the ground truth for this, the justification pulled from the reporter’s story provides critical context about the source’s upcoming voting opportunity and the legislative proposal that could affect it—none of which the models captured.

Example 5: Article 32 (OpenAI Board Restructuring)

Sourced Statement:

Mr. Altman complained that the research paper seemed to criticize OpenAI's efforts to build artificial general intelligence, or A.G.I., a machine that can do anything the human brain can do.

Ground Truth	GPT 4.1	Claude 3.7 Sonnet	Llama 3.1-405B
according to an email that Mr. Altman wrote to colleagues and that was viewed by The New York Times.	NULL	NULL	NULL

Table 12

A source (anonymous to the public, but not to the reporter) may send an authentic document (email) to a reporter. The reporter may then include a statement from that email with the attribution. In our system prompt, we show the LLMs that they need to pull this kind of text as the reporter’s justification for running the statement. All three models returned null, missing this context of the reporter’s authentic document access entirely.

Summary of Source Justification Performance Patterns

1. Closely tied justifications is easier: Two higher-performing models (Claude 4 Sonnet, Gemini 2.5 Pro) succeed primarily when source justification closely follows the sourced statement, embeds organizational mission statements in attribution, or provides identity/experience context clearly tied to the source. When source justification appears in different paragraphs or requires understanding implicit connections between sources and story issues, all models—including higher performers—return null values.

2. Lived experience justifications are a challenge: Sources without formal titles whose inclusion is justified through lived experience, stakeholder status, or community connection are areas where the models have to improve significantly. This is precisely the type of sourcing that more inclusive and non-hegemonic journalistic sourcing prioritize.

3. Document source justifications are tricky: When reporters explain how they accessed documents or establish document authenticity (ethically important-transparency) to justifying including claims or statements from them, models may not catch them.

These patterns confirm that source justification annotation remain a tough challenge for LLMs. Even the best-performing model (Claude 4 Sonnet at 56.17%) succeeds in the simpler cases and struggles when the reporter’s language to justify source inclusion involves contextual nuance. Making progress with LLM accuracy on this attribute is key for building more sophisticated journalism ethics analysis applications.

Strong Performance: Structured Elements In Sourcing Annotations

In contrast to source justification, all models performed well on the three more structured annotation elements: Type of Source (88–93%), Name of Source (80–90%), and Title of Source (80–89%). This reveals current LLM strengths in extracting categorical information and proper nouns from text. On the whole, this is not surprising because previous generation AI technologies such as NLP (e.g., the Stanford CoreNLP package) were also able to pull named entities with high accuracy.¹⁵

Type of Source classification—distinguishing Named Persons from Named Organizations, Documents, Anonymous Sources, Unnamed Persons, and Unnamed Groups—proved straightforward for models once they correctly identified sourced statements. ChatGPT 4o-latest achieved 93.45% accuracy here, while even lower-performing models exceeded 88%.

Title extraction—identifying formal positions like “Mayor,” “Professor,” or “Chief of Police”—similarly showed strong performance. Models excel at recognizing these formal markers of power and expertise, suggesting that title-based source analysis is immediately viable with current technology.

However, this strength also highlights a limitation: models identify formal authority markers for sources more readily than the contextual justifications that might explain non-expert sources’ authentic standing.

Comparative Examples For Accurate Title Annotations

We show examples where models captured titles accurately, and three cases illustrating common edge case challenges: handling sources without formal titles, distinguishing titles from characterizations, and managing complex organizational affiliations.

Example 1: Article 11 (Restoring Voting Rights)

Source:

Nicole Porter, Director of advocacy at the Sentencing Project

Ground Truth	Gemini 2.5 Pro	Claude 4 Sonnet
--------------	----------------	-----------------

¹⁵ The Stanford CoreNLP-based system for quote extraction and attribution in journalistic texts is documented in: Muzny, G., Fang, M., Chang, A., & Jurafsky, D. (2021). A Two-stage Sieve Approach for Quote Attribution. ACL Anthology. <https://aclanthology.org/E17-1044/>; and augmented with a machine learning model in: Shang, X., Peng, Z., Yuan, Q., Khan, S., Xie, L., Fang, Y., & Vincent, S. (2022). DIANES. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. <https://doi.org/10.1145/3477495.3531660>. On the capacities of NLP-based technology for annotation and attributions see: Full list of annotators. (n.d.). Stanford CoreNLP. <https://stanfordnlp.github.io/CoreNLP/annotators.html>

Director of advocacy, Sentencing Project	director of advocacy, The Sentencing Project	the Sentencing Project's director of advocacy
--	--	---

Table 13

Expectedly this is near-perfect match between the models annotations of the title of the source and the ground truth data. Both models correctly identified Nicole Porter's formal title ("director of advocacy") and organizational affiliation ("Sentencing Project") where the title is held. Minor variations in phrasing ("director of advocacy, Sentencing Project" vs. "Sentencing Project's director of advocacy") are semantically equivalent and demonstrate that models handle standard organizational titles with high accuracy. This pattern holds across our dataset for formal positions clearly stated in the attribution.

While models expectedly perform well on standard formal titles, here are some examples with three types of edge cases present challenges: (1) sources without formal titles where models must correctly return null, (2) distinguishing titles from characterizations, and (3) handling unnamed groups where title assignment is inappropriate.

Example 2: Article 11 (Restoring Voting Rights) - Sources Without Formal Titles

Source:

Steven Scott

Story text that introduces the source: "Steven Scott, 33, was paroled in 2015 after serving more than four years on assault and other charges. He also has regained voting rights and cast ballots for Republican candidates in his first elections, including 2020. He sees the two-year wait period as one link in a long chain of barriers for those trying to reenter society."

Ground Truth	GPT 4.1	Claude 3.5 Sonnet	Llama 3.1-405B
NULL (no formal title)	NULL ✓	NULL ✓	NULL ✓

Table 14

This example shows correct handling across all models. The source, Steven Scott, is included for his lived experience as someone who regained voting rights after felony conviction, and he is not introduced with a formal title. All three models correctly returned NULL, demonstrating they can distinguish between story text that is a candidate for source justification (why the person is included) and title (formal position).

Example 3: Article 25 (Anti-trans laws and ballot access) - Sources Without Formal Titles

Source:

Hazel Krebs

Story text that brings in the source: "a 42-year-old transgender woman living in Kansas, one of the states with a new anti-trans law"

Ground Truth	GPT 4.1	Claude 3.5 Sonnet	Llama 3.1-405B
NULL (no formal title)	NULL ✓	NULL ✓	NULL ✓

Table 15

This is another example showing accuracy. Hazel Krebs is included as a transgender woman with lived experience of anti-trans laws—her inclusion is justified by her identity and experience, not a formal title. The characterization "42-year-old transgender woman living in Kansas" is not a title signifying positions of power or authority as we define in our prompts. The models correctly distinguished between identity-based characterization and formal title, returning NULL as appropriate.

Example 4: Article 25 (Anti-trans laws and ballot access) - Unnamed Group of People

Source:

"Politicians"

Story text attributing speech for politicians: "Like with other laws targeting transgender people, including bans on transgender student-athletes from participating in women's sports, politicians supporting these policies frame them as protecting cisgender women."

Ground Truth	Llama 3.1-405B
Type of Source: Unnamed group of people Name: Politicians Title: NULL	Type of Source: Unnamed group of people Name: politicians Title: NULL ✓

Table 16

This the correct handling of the source type "unnamed group of people" which in this case is "Politicians". When an attribution is to an unnamed group, individual titles are not applicable—the "Name" field captures the group designation ("politicians") but Title remains NULL. Llama 3.1-405B correctly recognized this pattern. This demonstrates that models are able to apply the definition Unnamed Group of People as a type of source, where individual titles are not applicable, even when the group designation might itself reference occupations or roles.

Summary of Title Annotation Performance

Source title annotation represents one of the strongest performing elements for models in our benchmark (80-89% accuracy across models):

1. High accuracy for formal organizational titles: Models consistently capture standard titles with organizational affiliations ("Director," "Counsel," "CEO"). Minor phrasing variations ("director of" vs. "director, ") do not affect semantic accuracy.

2. Successful handling for sources without formal titles: Models correctly distinguish between source justification (lived experience, stakeholder status) and formal title. Sources included for their experience rather than institutional authority appropriately are annotated as having no title.

3. Proper handling of unnamed groups: Models recognize that Unnamed Group of People sources do not have individual titles, correctly returning NULL even when the group designation references occupations or roles.

Implications for sourcing analysis: The high accuracy for title annotation will likely be useful for applications tracking expert citations, analyzing source authority patterns, and measuring organizational representation. It will also be useful for statistical analysis where we want to know, for example, the proportion of named sources in a news cycle stream that are non-titled voices vs. titled voices.¹⁶ (Diversity along the expert vs. lived experience sourcing fault line.) Unlike source justification (which remains challenging), title extraction appears production-ready for many use cases.

What can change in these scores and why?

One way to understand this LLM benchmarking project is look at the moving parts-diagram (Figure 1) of the benchmarking system we have included earlier in this report. The three critical moving parts are the story corpus, prompts, and the models (and their settings) themselves. The fourth is the ground truth dataset which is the human annotation output of the story corpus, and it could be subject to a little variation. Here is how the score can change – go up or down.

¹⁶ Measuring source diversity proportions—including expert vs. non-expert sourcing—as an analytical use case is discussed in: Vincent, S., Wu, X., Huang, M., & Fang, Y. (2023). Could Quoting Data Patterns Help in Identifying Journalistic Behavior Online? *International Symposium on Online Journalism*, 13(1), p. 9 (using source-diversity proportions data from direct quote annotations across over 800 news sites). For recent empirical work on tracking source diversity in U.S. news organizations, see: Grannan, C., Toff, B., & Searles, K. (2025). Counting on Diversity: Approaches to Tracking Sources in U.S. News Organizations. *Journalism Practice*, 1–20. <https://doi.org/10.1080/17512786.2025.2453630>. For quantitative assessment of sourcing in issue-based coverage, see: Moorhead, L. (2024). Who Gets a Voice in Homelessness? A Content Analysis of Quotations Used by Journalists. *Journalism Practice*, 1–28. <https://doi.org/10.1080/17512786.2023.2300281>

Story corpus shifts: As the corpus is expanded to more news stories, the type of sources in the stories (people, organizations, documents, unnamed groups of people, anonymous sources, and so on) could be changed to deliberately include more non-expert sources with substantial lived experience-type source justifications. Since source justification annotations are the weakest area, there could be swings in accuracy scores if more stories are included where the justification texts are harder for the LLMs to parse accurately. Currently there is also a low proportion of anonymous sources (and likewise, their justifications), and increasing that proportion in an expanded corpus makes for a richer benchmark workload. This could impact the scores.

Also, as we revise the vocabulary to handle source documents attributed implicitly by links, the accuracy score for documents as a type of source accuracy could change. HTML links are in the structure of the stories, so we don't expect link detection itself to be complex. But at the other end of the link could a document issued by a government authority or scientific body, or a secondary sourcing (press) news article. These are new areas for accuracy scoring.

Prompts: While our current prompts are very detailed and took 50 iterations to build to stability, as noted earlier, we are seeking ways to improve them further. The goal there would be to see if can get less-expensive LLMs to score higher.

Models and settings: Our primary goal is to test with newer models and newer variants of existing models and figuring whether some settings need to be changed, or trying custom models later is another moving part. This part is necessary because we are testing model performance to compare them.

HOW TO USE THIS BENCHMARKING SYSTEM

We list below a few early stage examples of how to use our system.

1. **Journalists:** If you are newsroom or reporter, you can use our use our prompts to analyze your own story drafts and published texts for sourcing. For instance you could a 10-20 of recent stories and see how often you have included different types of sources or what your justifications look like for sources without title.
2. **Journalism/ethics research:** If you are a researcher on news content analysis and are looking to extract all quotes, source names and their titles from your collection, this is possible with our system. Use our prompts on any high-scoring LLM from our chart, and create your sourcing dataset itself for the stories.

In addition for use cases 1 and 2 above, once you have a dataset, you can run basic statistical analysis on the source data to calculate proportions. For instance you can compute: a) percentage of titled sources vs non-titled sources b) percentage of organization sources vs people c) top quoted people, etc. Going further, if you have access to gender detection API (or can manually code the gender for sources who are named persons), you can compute the proportion of male vs female sources in your story collection, and intersect that with experts and non-expert sources.

3. **Deeper data analysis:** Some people are interested in ground truth datasets for their own experiments. Our ground truth data has stories and all the sourcing data for them. So if you want to do a content analysis on all the source justifications texts and correlate their presence with say another column's values, you can use the dataset itself for that.
4. **Reproduce the whole benchmarking system:** You can download the entire project from Hugging face or GitHub. Use your API access to models and produce the annotations for the same story corpus for your choice of LLM and settings, and score them. If you plan to do this you could re-use our code that uses OpenRouter to get the annotations done and our scoring code. (All in python at GitHub.)
5. **Send us your use case:** Or, send us your use case and we could show you how to leverage our system to meet that need.

CONCLUSION: THE NEED FOR ONGOING BENCHMARKING

Our benchmark reveals that current LLM capabilities for journalistic sourcing annotation are currently selective rather than comprehensive.

- Two models—Gemini 2.5 Pro and Claude 3.7 Sonnet—demonstrate fair accuracy (80%+) for identifying sourced statements, a critical baseline for enabling downstream analysis.
- All evaluated models perform well on structured elements like source type, name, and title extraction.
- However, every model fails to reach acceptable accuracy on source justification, the element most relevant to ethical questions about source portrayal, stakeholder inclusion, and lived experience representation.

This mixed performance profile suggests immediate applications are possible within defined scopes. News organizations could deploy top-performing models to track expert vs. non-expert sourcing ratios, monitor quoted vs. paraphrased attribution patterns, or analyze source diversity by formal titles across coverage areas. Conversely, applications requiring source justification analysis—stakeholder portrayal and representation audits, solidarity journalism assessments—remain beyond currently beyond off-the-shelf LLM capabilities. These use cases will likely require further advancements in the models themselves and an architecture for layered building of capabilities using the models. (Better prompts, agents, stacking architecture, etc.)

The rapid pace of LLM development makes ongoing benchmarking essential. Our evaluation used models current as of Fall 2025, but new variants with improved linguistic review capacities emerge regularly. Major releases merit re-evaluation on this benchmark to track progress. The publicly available dataset enables precisely this kind of continuous assessment. We plan to expand the story corpus and ground truth data in 2026.

Future benchmark expansions also need to address several limitations in the current dataset. First, increasing representation of anonymous sources, unnamed sources, and document-based attribution would test model performance across the full spectrum of journalistic sourcing practices. Second, expanding beyond text to include video and audio source attribution would address emerging formats in news influencer content. Third, developing comparative benchmarks across languages would reveal whether current performance patterns hold globally or reflect English-specific capacity (training) biases.

Most fundamentally, this benchmark demonstrates that the systematic evaluation of LLMs for journalistic tasks is both feasible and necessary. While general-purpose language benchmarks provide valuable baseline metrics, our domain-specific evaluation reveals capabilities and limitations invisible in broader assessments. Journalistic sourcing—routine, linguistically complex and ethically consequential—exemplifies the kind of specialized task requiring dedicated benchmarking.

ABOUT THE AUTHORS

Subramaniam Vincent is Director of Journalism and Media Ethics, Markkula Center for Applied Ethics, Santa Clara University. His research focus is the intersection of journalism ethics and news distribution algorithms on sourcing.

Zhan Shi is a PhD student in Computer Science and Engineering at Santa Clara University. Her research focuses on responsible AI and AI Agents.

Yi Fang is Professor with the Department of Computer Science and Engineering and Director of responsible AI with the School of Engineering, Santa Clara University. His research interests include information retrieval, natural language processing, and machine learning.

ACKNOWLEDGMENTS

The LLMs benchmarking project started with the MS thesis work of CS/Engineering Student Phoebe Wang (MS, 2024) at Santa Clara University. Sahas Koka helped with initial coding for the annotation generation.

ABOUT THE MARKKULA CENTER FOR APPLIED ETHICS

The mission of the Markkula Center for Applied Ethics is to engage individuals and organizations in making choices that respect and care for others. The center works to empower people and organizations to make better decisions for a more caring world. The Center's Framework for Ethical Decision Making is widely recognized by individuals and organizations as a clear and helpful process to make better decisions. The center offers additional materials on its website; customized engagements with organizations; and training to Santa Clara University students, faculty, and staff in applied ethics.

APPENDIX 1: Vocabulary for training students and LLMs (system prompt)

The following definitions were provided to both human annotators and LLMs to ensure consistent application of the sourcing annotation schema. These are reproduced verbatim from the system prompt used in the benchmark evaluation.

Source: A source in journalism is a person, organization, document, or another news article from whom a journalist takes viewpoints, experiences, claims, expertise, positions, insights, knowledge, data or documents. Reporters may directly quote their sources or use indirect speech to paraphrase a source's views or claims. Journalists attribute statements to sources using a range of speech-act verbs. While "said" and "according to" are often used and neutral-sounding verbs, there are many others. Here are some examples: said, stated, reported, wrote, told, added, according to, claimed,

alleged, purported, explained, pointed out, noted, revealed, confirmed, demonstrated, shouted, warned, urged, lamented, blasted, praised, insisted, demanded, pleaded.

Sometimes the source is the person who sent sensitive material such as emails or documents or other internal organizational correspondence to the reporter. The source is a person who may have been present at meetings where sensitive deliberations or discussions took place, and that person then shared material from the meetings with the journalist. For e.g., when a reporter attributes a claim or statement to a person or a group of people with words such as "according to people familiar with..", or "according to people who were present at the meeting..", it means that person or those people are the source. When the reporter attributes claims or a statement using words like "according to a copy of emails reviewed by this newspaper" or "according to a copy of the document.." it means the source could be a document. If the person who sent the emails or documents to the reporter is granted anonymity by the report, then that person is an anonymous source. (See also, our definition of Anonymous Source.)

A source is an organization when the reporter cites the organization by contacting its spokesperson or other representative or officials. The reporter may cite such spokespersons or officials from a direct interaction or from a press release or social media or corporate blog.

A source can also be another news organization itself. When the journalist has attributed something in the text to another news organization that reported it earlier, then that organization is the source. Sometimes, the reporter will name that organization directly in the text. At other times, the text may simply carry a web link directly to that news article, similar to a document link.

Types of Sources: a) Named Person b) Named Organization c) Document d) Anonymous Source e) Unnamed Person and f) Unnamed Group of People.

Named Person: A source who is a named human being in the story. Note that documents like emails, audio, video, meeting minutes, or other material internal to an organization sent by a human source may contain named people expressing views, taking actions, decisions, etc. Those named people are not sources, even if the reporter included those names and reported their viewpoints or actions in the news article. The person sharing the emails or documents with the journalist is the source, or the documents themselves, if they are authenticated, are a source. (See also, our definition of Document as a type of source.)

Note: When a named person is quoted (directly or indirectly) citing or referring to the contents of a document, the named person is the source, not the document.

Named Organization: An organization named as a source. This includes cases where reporters attribute a statement or statements to unnamed "officials" or unnamed "spokesperson" of a named organization.

Note: When a statement or statements are attributed to a named spokesperson of an organization, the Type of Source is Named Person. But when the spokesperson is not named, and the organization is named, (e.g. "an Exxon Mobil spokesperson said..") it is a Named Organization type of source. It is not an Unnamed Person type of source.

Document: An original or authentic document issued by an authoritative organization or person for that field, jurisdiction or expertise. It is usually accessible by the public online or may be retrieved in return for a public records request. The document may itself be a full webpage or PDF or other portable format. It may be attributed through a URL inline in the article. However, when documents sent to a reporter are part of internal material in an organization, such as emails, correspondence, presentations, meeting recordings, and data, the person sending the documents is the source. Also, the reporter may have included statements from a person who is citing or referring to the contents of some document or study or research. That does not make the document the source. The person is the source. The document is the source only if the author of the story has accessed the document itself and used its contents directly for the story.

Anonymous Source: An anonymous source is a person, known by name and identity to the reporter and often the reporter's editor, but the name is being withheld from inclusion in the article as a named source before its publication. You may see the reporter disclose this in the story, that they spoke to the person "on condition of anonymity", or that they were offered anonymity to discuss sensitive matters. This source may have been present at meetings where sensitive deliberations or discussions took place then shared material from the meetings with the journalist. Or this source may be a key player or witness in internal decisions or proceedings at an organization. For e.g., when a reporter attributes a claim or statement to a person or a group of people with words such as "according to people familiar with..", or "according to people who were present at the meeting..", it means that person or those people are the source. This source may have sent sensitive documents, emails, audio, video, meeting notes, or other internal correspondence from inside an organization to the reporter, and the person's identity needs to be protected. So they are not named.

In journalism ethics, the reporter may offer anonymity to the source if the reporter believes or concludes that the source's life or career/job or family members lives, etc., might be in jeopardy or that they might face retaliation or retribution. An unnamed spokesperson or official of a named organization is not an anonymous source.

Unnamed Person: Unlike the type Anonymous Source, in this case a journalist sometimes attributes statements, claims, utterances, comments, or other information to an unnamed individual, but there is no language in the text disclosing that the person is "known" to the reporter and/or asked for anonymity (i.e. to not be named). Examples may be a police officer saying or doing something during a law enforcement, but the reporter has not named the person, just attributing the statement to "police officer". Another example could be a hearing where a judge ruled or said something and the reporter did not name the judge. Another example maybe a commenter on an online platform whose views are being included in the story, but the reporter does not attempt to identify the person, and simply refers to the person as "commenter".

Unnamed Group of People: Sometimes a reporter may attribute a statement or statements to a group of people who are saying or expressing or advocating for something as a group. We call this type of source, Unnamed Group of People. An Unnamed Group of People source is a group the journalist has witnessed on the ground or online or at a meeting or had access to the group of people otherwise. The group may be a professional group, like "consultants", or "investigators", with a specific connection to the issue in the story, or a group involved in some action. For example: players, teachers, children, protestors, attendees, advocates, activists, participants, consultants, onlookers, commenters, etc. The reporter does not name all of them individually and simply refers to them as a group, and may quote or paraphrase what the group of people is saying using words like "the protestors said this", or "the teachers chanted.."

Note 1: Do not confuse them with the Anonymous Source type. Anonymous Sources are individuals who ask for anonymity from the journalist and are granted that. An Unnamed Group of People-type source are not people who asked for anonymity from the reporter.

Note 2: A document sent by a source to a journalist may contain references to groups of people saying something. That is not an example of an Unnamed Group of People source. The document is the source since the reporter is citing the document.

Note 3: When reporters attribute a statement or statements to unnamed "officials" or a "spokesperson" or "lawyers" of a named organization, the type of source is Named Organization. It is not an Anonymous Source or Unnamed Person or Unnamed Group of People. The Name of Source is the named organization. e.g. if "San Mateo County officials" are cited, the type of source is not Unnamed Group of People, it is Named Organization and the Name of Source would be "San Mateo County".

Sourced Statement: Every statement in a story that the reporter would NOT have been able to put in without drawing or receiving the content or part of the content from one or more sources. Sourced Statements are actual text lines in the article that reporters have attributed to a source. Sourced statements may contain quotes or indirect speech. They may include viewpoints, experiences, criticism, questioning, advising, support or other expressions from a source. They are often but not always signaled using explicit attributions like the source "said", "says", "opined", "shouted", "agreed", "stated", "wrote", "criticized", "questioned", etc. "According to" is a very common attribution wording. Here is a fuller list of examples: said, stated, reported, wrote, told, added, according to, claimed, alleged, purported, explained, pointed out, noted, revealed, confirmed, demonstrated, shouted, warned, urged, lamented, blasted, praised, insisted, demanded, pleaded. Attributions may also be implied. Sourced Statements also include all other attributions the reporter has made to a source, referring to the source's positions, or expressed attitudes. Those lines may be present in addition to direct or indirect quotes from that person. But when journalists report out what they are witnessing first hand or what they conclude about a source or a source's actions, and they are not attributing it, those lines are not sourced statements.

Note 1: A common practice in news writing is a reporter attributes the first statement in a paragraph to a source. But one or more lines after that will draw from the same source without citing it each time. It's implicit. See these two examples:

Example 1: "At least 14 states have introduced proposals this year focused on restoration of voting rights, according to the Brennan Center for Justice. An Oregon proposal would allow felons to vote while incarcerated. A Tennessee bill would automatically restore voting rights once a sentence is completed, except for a small group of crimes. Texas legislation would restore voting rights to those on probation or parole." - In this example the first line is attributed to the Brennan Center for Justice. The next two line are follow-up reporting and it's highly likely that source for that is the Brennan Center too.

Example 2: "In Nebraska, nearly 18,000 people are unable to vote because of felony convictions, said The Sentencing Project's director of advocacy, Nicole Porter. That includes 7,072 who fall under the two-year wait requirement and are currently unable to cast a ballot. The rest have not completed their full sentences." - Here too, the second line is not being explicitly attributed to Nicole Porter, but she is the source for it.

Note 2: A ethical practice (transparency) in news writing is the reporter discloses that source was contacted for comment, but "did not return calls", or "did not respond by deadline", or "declined comment", or "did not respond to emails". When these in a story, they are not sourced statements, but disclosures. Do NOT list them as sourced statements.

Note 3: Occasionally a reporter may have a single statement with attributions to multiple valid sources. When analyzing such a sentence (statement), it helpful to create multiple (copied) instances of Sourced Statements, and assign one each to each of the sources. See this example: "During the meeting, Fulton County Commissioner Lisa Wheeler and City of Atlanta Public Works Department director Tom Nicol endorsed it and two restaurants association representatives spoke about its steep costs for businesses." For the above example, an analyst will create three instances of the same statement, and assign one each to each of the three sources: Lisa Wheeler, Tom Nicholas, and restaurants association representatives.

Name of Source: Name of Source applies to the following types of sources: Named Person, Named Organization and Document. For Named Person and Named Organization types, it is the person or organization's full name. This is identifiable as a proper noun in the article. When a source is referred to using a generic designation such as "spokesperson" or "representative" or "officials" from a named organization, the name of the source is the name of the organization. When the source is a public document, the name of the source is the document's publishing organization or person, if available. For Anonymous Sources, the Name of Source does not apply, i.e. it has a 'null' value. For Unnamed Group of People source, the name of source is the term being used to refer to the group. For e.g., it might be "teachers", "participants", "parents", "attendees", "rallygoers", "protestors", etc.

Title of Source: The words defining or designating the formal position of power, authority, responsibility, expertise and/or leadership held by the Named Person source, including where the position is held. E.g. "Assistant Director, Marketing, XYZ Foundation" is a valid Title of Source. Terms like Director, Mayor, Vice President, President, Secretary, Treasurer, Board Member, Professor, Provost, Principal, Congresswoman, Senator, Chief of Police, General Counsel, Spokesperson, Analyst, etc., are all valid titles for a source, signifying power, authority and/or expertise. If the person is a legislator, the Title of Source includes the constituency. Experts are often Named Person sources and their expertise is signaled by their title, designating their position/role usually held at some organization, with domain specialization. Sometimes people in leadership roles will simply be referred to as "leader". That is also a Title of Source.

Note 1: Characterizations or designations like team member, player, senior, sophomore, junior, freshman, activist, protestor, attendee, participant, eyewitness, etc. are NOT titles. They are meant to factually portray the source and how they are connected to an issue, organization or community by virtue or enrollment, membership, actions taken, etc. It does not mean they are in a position of power or authority or expertise. Likewise, occupations in the trades like carpenter, plumber, janitor, installer, coal miner, etc., are not titles. They represent people in a type of trade.

Note 2: Titles are often held at organizations, but the organization names are not Source Justifications. They are part of the Title of Source. For instance, "Jane Doe, director, the Brennan Center", "John Done, Executive Director, The Sentencing Project", or "Radha Kumar, the student anti-racism network" are all valid full Title of Source values. The organization names are not valid Source Justifications in and off themselves. See the Source Justification definition below.

Note 3: For anonymous sources, the Title of Source may or may not be in the story and only the association or justification for inclusion in the story may be mentioned.

Note 4: For Named Organizations and Document sources, the title of source does not apply.

Note 5: Journalism ethics encourages reporters to include the voices of everyday people and people who, in that story's context are NOT being quoted for their formal positions of authority or power. Such Named Person sources will usually not have a formal Title Of Source.

Source Justification: This refers to any additional source characterization, context, explanation that justifies to the reader why the source is in the story or that section of the story, how they are connected to the story and/or to other sources in the story. Any of the six defined types of sources may have such justifications and explanations present. It may be a few words, a part of a sentence, multiple sentences, or a full paragraph. Sometimes the source justification is part of the source's introduction, but it might also come later in the story after some statements have been attributed to the source. For named persons, the source justification text may portray the lived experience of the source. When sources are people who are stakeholders to the issue being reported on, who are taking action on an issue, witnessed something happen, or have a lived experience related to the issue of the story, or a co-litigant in a lawsuit, etc. When a reporter narrates this, it is the context demonstrates their significance in the story for readers. For e.g., someone who went through a period of homelessness may be quoted for their lived experience and opinion about solutions. Someone else may have spent four years waiting to get a job or to get their voting rights back because of a prior felony conviction. A lawmaker said something noteworthy while sponsoring or debating a bill in the legislature. Note that while source justification is NOT the same as title of the source, it may include the title of the source. Also, Named Persons or Anonymous Sources or Unnamed Persons without a title may still have source justification present in the text. Note, mere text of attribution is NOT source justification. For e.g. "Jane Doe said by email", "Ram Kumar posted on X", "ABC Company said on its website", etc., are not valid source justifications for the statements being attributed. The source justification maybe earlier in the text or in some other sentence. "Said", "said by email", "posted on Instagram", etc., are simply the reporter's transparency at work: indications of how (pathway) the reporter was able to access the source.

APPENDIX 2: LLMs user prompt

(Note: Instructions given to human annotators are exactly the same in content and steps, except they directly populate spreadsheets with the data extracted.)

INSTRUCTIONS

Read the attached news article carefully. Apply the instructions below only on the attached news article. (Ignore the subtitle text if it exists). Follow ALL the steps below to extract the full list of Sourced Statements in the article. For each Sourced Statement, extract the Name of Source, Type of Source, Title of Source, and Source Justification.

1. Start with the body of the article, ignore the Subtitle (if there is one), go line by line from beginning to end, identify each line or lines together that fit the definition of a Sourced Statement. Remember to read the three notes and examples in the system_prompt (Note 1, Note 2, Note 3) that accompany the definition of Sourced Statement. If it's a sourced statement, extract the exact full sentences from the article. Do not paraphrase or summarize. Copy and paste the exact text from the article. Sometimes, in the same paragraph, there may be multiple contiguous statements, one after another, attributed to the same source. In this case, they are one instance of a sourced statement. Do not separate out the lines as multiple, separate, sourced statement instances. On the other hand, there may be sourced statements attributed to the same source in other (i.e. different) paragraphs. These are new instances of Sourced Statements, do not merge them into one. No need to print anything yet.

2. For each sourced statement, identify the attributed Type of Source. Use the six available definitions to do this: Anonymous Source, Unnamed Person, Document Source, Unnamed Group of People, Named Person, or Named Organization. Remember that the types of source "Unnamed Person", "Unnamed Group of People" and "Anonymous Source" are all distinct from each other, so apply the definition carefully. If the Type of Source is not identifiable, set the Type of Source as 'Unresolved'. No need to print anything yet.

3. Next, for that sourced statement, if the Type of Source is Named Person or Named Organization, identify the name using the Name of Source definition. Usually names of individuals or organizations are stated in full when they are first referred to. No need to print or write anything yet.

4. If the Type of Source is Anonymous Source or Unnamed Person, the Name of Source has a 'null' value. No need to print or write anything yet. If the Type of Source is Document, the name of the

document may or may not be in the text. If the name of the document is present, set the Name of Source to that, otherwise it has a 'null' value.

5. If the type of source is Unnamed Group of People, the Name of Source is the term the reporter has used to refer to the group. For e.g., it might be "teachers", "participants", "parents", "attendees", "rallygoers", "protestors", "advocates", "activists", "students", etc. If not present, set Name of Source to the 'null' value. No need to print or write anything yet.

6. At this point you have a Sourced Statement, and Type of Source and Named of Source. No need to write or print anything yet.

7. Next, if your Type of Source is Named Person or Anonymous Source, apply the definition of Title of Source and extract the title if provided. Remember that Named Persons in positions of power, formal authority, leadership, management, or expertise will usually be quoted with a title. When the voices of everyday people and people without structural power are included in the story, they may not have a title. Note, when the Type of Source is Document or Named Organization or Unnamed Group of People, there is no Title of Source. Set Title of Source to 'null' value. No need to write or print anything yet.

8. Next, find the Source Justification for the source in this sourced statement using the definition. Directly extract the text that matches this definition from the story. Do not synthesize or generate or summarize the Source Justification in your own words. Extract only the actual text. If the same source is attributed in earlier sourced statements, even if there is no source justification text in this instance, you may copy the Source Justification from the source's earlier sourced statement to this one. Remember that source justification refers to words or sentences doing additional characterization or context or explanation for inclusion of that source in the story. If more than one source justification is present in the text, concatenate them as one single string separated by ';'. No need to write or print anything yet.

9. Organize and add your data to a JSON data object as follows, but do not output this JSON yet:

```
{  
  "Sourcing Table": [  
    {  
      "Sourced Statement": "string",  
      "Type of Source" : "string",  
      "Name of Source": "string",  
      "Title of Source": "string",  
      "Source Justification": "string"    }  
  ]  
}
```

```
}  
]  
}
```

10. Continue and repeat this exercise for the whole article, finding one sourced statement after another. For each sourced statement, identify the Type of Source, Name of Source (if applicable), Title of Source (as applicable), and Source Justification. Add to the JSON object. Continue until no sourcing statements are left. Remember each Sourced Statement in your findings list must only occur once. You cannot have two or more replicas of the Sourced Statement being attributed to different Types of Sources. Once a Sourced Statement is identified and attribution data annotated, it cannot be reinstated afresh.

11. If you have reached this step, that means no sourced statements are left. Now output the full JSON Sourcing Table, which is array of entries, each entry has a Sourced Statement, Type of Source, Name of Source, Title of Source, and Source Justification. Your JSON should take this form.

```
{  
  "Sourcing Table": [  
    {  
      "Sourced Statement": "string",  
      "Type of Source" : "string",  
      "Name of Source": "string",  
      "Title of Source": "string",  
      "Source Justification": "string"  
    },  
    {  
      "Sourced Statement": "string",  
      "Type of Source" : "string",  
      "Name of Source": "string",  
      "Title of Source": "string",  
      "Source Justification": "string"  
    },  
    {  
      "Sourced Statement": "string",  
      "Type of Source" : "string",  
      "Name of Source": "string",  
      "Title of Source": "string",  
      "Source Justification": "string"  
    }  
  ]  
}
```

```
]
}
```

Do not print or provide any other analysis. We need a valid JSON format that can be parsed to convert to CSV. Give me no other text but the JSON.

END OF INSTRUCTIONS.

APPENDIX 3: A full LLM annotation example for a news story in JSON format.

(Article 31 in our dataset: [Vermont House overwhelmingly backs bill prohibiting race-based hair discrimination](#), Jan 2024)

Claude Sonet 4 output:

```
{
  "Sourcing Table":[
    {
      "Sourced Statement": "Despite telling her not to touch it, Wilburn said the girl \"grabbed\" her hair.",
      "Type of Source" : "Named Person",
      "Name of Source": "Aaliyah Wilburn",
      "Title of Source": "null",
      "Source Justification": "A senior at North Country Union High School in Newport, Wilburn, 17; a leader with the Vermont Student Anti-Racism Network"
    },
    {
      "Sourced Statement": "\"She then told me that I felt like a sheep and she imitated the noises of one,\" said Wilburn, a leader with the Vermont Student Anti-Racism Network.",
      "Type of Source" : "Named Person",
      "Name of Source": "Aaliyah Wilburn",
      "Title of Source": "null",
      "Source Justification": "A senior at North Country Union High School in Newport, Wilburn, 17; a leader with the Vermont Student Anti-Racism Network"
    },
    {
```

```

    "Sourced Statement": "Once, when she was wearing a head covering because she was in the
midst of braiding her hair, Wilburn said a student pulled it off and ran away in the school hall.",
    "Type of Source" : "Named Person",
    "Name of Source": "Aaliyah Wilburn",
    "Title of Source": "null",
    "Source Justification": "A senior at North Country Union High School in Newport, Wilburn, 17; a
leader with the Vermont Student Anti-Racism Network"
  },
  {
    "Sourced Statement": "\"Later in the day, someone took it off me,\" she said.",
    "Type of Source" : "Named Person",
    "Name of Source": "Aaliyah Wilburn",
    "Title of Source": "null",
    "Source Justification": "A senior at North Country Union High School in Newport, Wilburn, 17; a
leader with the Vermont Student Anti-Racism Network"
  },
  {
    "Sourced Statement": "\"I wouldn't be scared of going out with a bonnet or a head wrap if this
bill was in place at the time of these incidents,\" she said in an email.",
    "Type of Source" : "Named Person",
    "Name of Source": "Aaliyah Wilburn",
    "Title of Source": "null",
    "Source Justification": "A senior at North Country Union High School in Newport, Wilburn, 17; a
leader with the Vermont Student Anti-Racism Network"
  },
  {
    "Sourced Statement": "Addie Lentzner, executive director of the student anti-racism network,
said she has seen hair discrimination in schools — not just in policies and norms but also in the
form of microaggressions and racial bias.",
    "Type of Source" : "Named Person",
    "Name of Source": "Addie Lentzner",
    "Title of Source": "executive director of the student anti-racism network",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "\"I think it's a huge problem in our schools and this bill would help
address it,\" Lentzner said.",
    "Type of Source" : "Named Person",
    "Name of Source": "Addie Lentzner",

```

```

    "Title of Source": "executive director of the student anti-racism network",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "Modeled after language promoted as the CROWN Act and already
enacted in 23 other states, including New York and Maine, the bill aims to prevent \"an erasure and
minimization of the experiences of marginalized people,\" particularly Black and brown people,
Rep. Saudia LaMont, D-Morristown, said during a preliminary vote on the bill Tuesday.",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Saudia LaMont",
    "Title of Source": "Rep., D-Morristown",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "She cited a 2023 study that found 66% of Black girls in predominantly
white schools and 44% of Black girls in all schools report experiencing hair discrimination, and that
the experiences typically happen before they are 10.",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Saudia LaMont",
    "Title of Source": "Rep., D-Morristown",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "For her daughter, she said, it first occurred in pre-school. While picking
up her daughter one afternoon she said she was dismayed to find her wearing a stop sign on her
chest that read, \"Please ask to touch my hair.\",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Saudia LaMont",
    "Title of Source": "Rep., D-Morristown",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "A teacher explained that her daughter was reacting to several instances
of students touching and petting her hair without her consent, LaMont continued. \"So it was in fact
my 4-year-old's idea to make a sign as a signal to her peers to stop violating her body and personal
space,\" she said.",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Saudia LaMont",
    "Title of Source": "Rep., D-Morristown",

```

```

"Source Justification": "null"
},
{
  "Sourced Statement": "In another incident, her daughter's dance teacher asked to have her recently installed braids taken out so the child's hair could be put up in a bun to look like the majority of the other white kids in the class.",
  "Type of Source": "Named Person",
  "Name of Source": "Rep. Saudia LaMont",
  "Title of Source": "Rep., D-Morristown",
  "Source Justification": "null"
},
{
  "Sourced Statement": "LaMont outlined to her House colleagues on Tuesday the ways in which caring for Black hair is a process. It can take a day to wash, dry and detangle and protective styling, such as box braids or twists, can involve 8 to 22 hours of work.",
  "Type of Source": "Named Person",
  "Name of Source": "Rep. Saudia LaMont",
  "Title of Source": "Rep., D-Morristown",
  "Source Justification": "null"
},
{
  "Sourced Statement": "The bill is important, she said then, because \"hair type, textures, styles, protective hairstyles\" are \"a symbol of one's identity and ability to show up authentically,\" which includes \"the freedom to take pride in one's appearance in ways people choose that feels in alignment with the traits that are associated with their race (and) improves confidence, self esteem, dignity and respect.\"
  "Type of Source": "Named Person",
  "Name of Source": "Rep. Saudia LaMont",
  "Title of Source": "Rep., D-Morristown",
  "Source Justification": "null"
},
{
  "Sourced Statement": "Black women's hair is more likely to be perceived as unprofessional, she said, citing surveys. Many of them feel they must wear their hair straight in a job interview and report being sent home or denied a job because of their hair. The same kind of assumptions are used to prevent students with certain hairstyles from engaging in sports and other activities, she added.",
  "Type of Source": "Named Person",
  "Name of Source": "Rep. Saudia LaMont",

```

```

    "Title of Source": "Rep., D-Morristown",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "\"Hair and head coverings do not determine behavior, capacity or competency and people should not be penalized, discriminated against or treated as such,\" she said.",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Saudia LaMont",
    "Title of Source": "Rep., D-Morristown",
    "Source Justification": "null"
  },
  {
    "Sourced Statement": "Hair discrimination is degrading, disrespectful and invades people's privacy in deep ways, said one of the bill's primary sponsors, Rep. Kevin \"Coach\" Christie, D-Hartford. He pointed to incidents of it nationwide that have challenged athletes, students and teachers.",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Kevin \"Coach\" Christie",
    "Title of Source": "Rep., D-Hartford",
    "Source Justification": "one of the bill's primary sponsors"
  },
  {
    "Sourced Statement": "\"It seems like a small bill but in actuality, it's another big block in the foundation of equity in the state,\" Christie said, adding that it pains him when families choose to leave Vermont after facing racial bullying, hazing and harassment.",
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Kevin \"Coach\" Christie",
    "Title of Source": "Rep., D-Hartford",
    "Source Justification": "one of the bill's primary sponsors"
  },
  {
    "Sourced Statement": "Representing Vermont's most diverse community, Rep. Daisy Berbeco, D-Winooski, also a sponsor, told VTDigger the bill is an important step in advancing equity and \"a step in the right direction towards people feeling safe and welcome in our state, no matter where they are in Vermont.\"
    "Type of Source" : "Named Person",
    "Name of Source": "Rep. Daisy Berbeco",
    "Title of Source": "Rep., D-Winooski",

```

```

"Source Justification": "Representing Vermont's most diverse community; also a sponsor"
},
{
  "Sourced Statement": "Rep. Mike Mrowicki, D-Putney, another bill sponsor, said he has heard
from constituents who have faced discrimination due to their hair.",
  "Type of Source" : "Named Person",
  "Name of Source": "Rep. Mike Mrowicki",
  "Title of Source": "Rep., D-Putney",
  "Source Justification": "another bill sponsor"
},
{
  "Sourced Statement": "With the population and demographic of Vermont fast changing, this bill
is a common sense effort to make sure that residents are not ostracized or mistreated because of
who they are, he said.",
  "Type of Source" : "Named Person",
  "Name of Source": "Rep. Mike Mrowicki",
  "Title of Source": "Rep., D-Putney",
  "Source Justification": "another bill sponsor"
},
{
  "Sourced Statement": "\"The reality that people of color are often ostracized or mistreated
because of their hair. It can be as seemingly innocent as people wanting to touch it — which is
really an invasion of people's space,\" he said.",
  "Type of Source" : "Named Person",
  "Name of Source": "Rep. Mike Mrowicki",
  "Title of Source": "Rep., D-Putney",
  "Source Justification": "another bill sponsor"
}
]
}

```