



ISSN (Print) 2596 – 7517

ISSN (Online) 2597 – 307X

FULL PAPER

Epistemic Possibility and Textual Source Voice in AI Technology Journalism: A Corpus-Assisted Case Study of May, Might, Could, and Likely, 2022–2026

Prepared by

Asjad Ahmed Saeed Balla
Department of English Language
and Literature
College of Languages
and Humanities
Qassim University, Saudi Arabia
A.Balla@qu.edu.sa
ORCID ID: <https://orcid.org/0000-0001-9387-8653>

Abstract

AI journalism often couples statements about current technical capability with predictions whose realization is uncertain. This exploratory, corpus-assisted case study investigates may, might, could and likely in 20 criterion-based TechCrunch AI articles, chosen as two eligible articles from two fixed 14-day windows each year, 2022–2026. On October 4, 2026 a script was executed to audit the captured canonical pages, raw html, cleaned editorial text and verification logs. Case-insensitive, whole-word searches in headlines and editorial bodies yielded 40 target tokens. Was was the most frequent form ($n = 16$, 40.0%) and occurred in nine articles, but only 7 of the 16 occurrences (43.8%) were epistemic; the other nine were capacity or affordance. Epistemicity was found in 10 of the 12 may tokens (83.3%) and all the might and likely tokens. Of these, 29 (72.5 %) were epistemic. One watermarking article in 2023 accounted for 8 of the 40 tokens, so removing it dropped the proportion of all target tokens in 2023 from 22.5% to 3.1%. Two blinded machine-assisted coding passes agreed on 35 of 40 semantic codes (87.5%; $\kappa = .679$). This is an algorithmic audit diagnostic, not human inter-rater reliability. The findings are descriptive, entirely auditable, and limited to this small, purposively bounded corpus.

Keywords: epistemic modality, uncertainty, artificial intelligence, technology journalism, corpus-assisted discourse studies, modal verbs, source attribution

Introduction

Public prognostication has entered the realm of artificial intelligence (AI). Technology reporting is about systems whose capabilities, commercial trajectories, regulatory consequences, and social effects are often unknown at the time of writing. This means journalists should report what might be important without claiming it already is. For this task, one resource is epistemic modality. Epistemic modality encodes the degree of commitment to a proposition and distinguishes between observation, inference, prediction, and scenario building.

Lab studies suggest that expressing uncertainty doesn't have to destroy trust. The five experiments, including a BBC News field experiment, conducted by Van der Bles et al. (2020) found that explicit uncertainty increased perceived uncertainty but led to small reductions in trust, mostly for verbal rather than numerical formats. Figure 1. Map showing the research area. In a precision medicine case study, Ratcliff (2021) found that press releases tended to omit research limitations, while some news reports included caveats. Capurro et al. (2021) found uncertainty to be a master frame in coverage of the early COVID-19 pandemic in Canadian newspapers, as readers sought to make sense of emerging evidence and policy responses (p. These results do not offer a one-size-fits-all rule for AI journalism, but they do point to uncertainty being conveyed through identifiable textual choices.

AI reporting has a peculiar semantic problem. Often claims about what a system can do include predictions about what could happen as a consequence of that system. Thus, a surface search for possibility markers can conflate epistemic assessment with ability, affordance, and circumstantial enablement. Carretero (2023) manually distinguished epistemic from non-epistemic forms and tagged the source voice in newspaper discourse, demonstrating the need for contextual checking in lexical retrieval. Hollmann et al. (2024) also found that, in a corpus of Japanese undergraduates, could was frequent but overwhelmingly expressed ability rather than epistemic possibility.

The AI news study has largely focused on frames, actors, sentiment, headlines, or narratives. Artificial Intelligence: Potential and Dangers for Meta-Journalistic Reporting, Shaikh Moran. 2022. Roe and Perkins (2023) argue that the UK's stories about ChatGPT and AI create a duality between systemic/immediate threat and social opportunity. Nguyen and Hekman (2024) analyzed a sample of 3,098 items and found that debate about AI news grew more critical over time. After the release of Chatbot, Ryazanov et al. (2025) observed a tenfold increase in attention and a stronger association between AI and risk/hazard. Cross-national and actor-centered studies also noted variation in frames, sentiment, and source standing (Ittefaq et al., 2025; Zai et al., 2025). After analyzing 25,595 news articles about AI, they found a strong future orientation and a slight preference for opportunity over risk.

These macro-level studies do not consider whether a given prospective statement is grammatically qualified, whether a modal refers to technical capability rather than probability, or whether an assessment is part of reporter narration or attributed source discourse. This more targeted measurement problem is tackled in the present study on a small, time-stratified sample of specialist technology journalism. Rather than estimating language change across technology journalism as a population, it is explicitly exploratory, testing the validity of four candidate markers and how sensitive descriptive patterns are to article composition.

Research Questions

RQ1. What are the frequencies, shares, and article-level incidence of may, might, could, and likely in the sampled AI technology journalism?

RQ2. To what extent does each marker express epistemic uncertainty rather than ability, technical capacity, circumstantial possibility, or a fixed expression?

RQ3. How does the distribution of the four markers vary across the 2022–2026 sampling periods, and how sensitive is the pattern to document-level concentration?

RQ4. Among epistemic occurrences, how are assessments distributed between journalist-text and attributed-text loci, and how sensitive is that distribution to quotation-rich articles?

Literature Review

Contextual Analysis of Epistemic Modality

Epistemic modality is concerned with the speaker or writer's evaluation of the possibility, probability, evidence or truth of a proposition. Speakers and writers express it through modal auxiliaries and a wider range of lexical and grammatical resources. For instance, Carretero (2023) examined epistemic modal and evidential expressions in 400 articles published in The Guardian in two stages of the COVID-19 pandemic. The study manually distinguished between epistemic and non-epistemic instances and between writer, direct-report, and attributed voices. The procedure is relevant to the present study because the same surface form can be assigned different semantic functions. The four focal forms occupy different positions in this system. May and might tend to describe a situation as possible without commitment. Likely describes probability in adjectival or adverbial constructions. Could is more polyfunctional. In a corpus of Japanese undergraduates, Hollmann et al. (2024) report that only 3 of 148 instances were epistemic, while 143 expressed ability. Technology journalism has a similar pressure to classify. Its reports regularly speak to what software, models, devices or users can technically do. Thus a dictionary of uncertainties with only a form may contain capacity claims as epistemic hedges. Contextual analysis also requires transparent choices of methods. Curry et al. (2024) found that ChatGPT altered input and produced false concordance inferences in replication tests, showing that model-generated interpretation cannot replace direct corpus evidence. Gillings and Jaworska (2025) likewise show that, in discourse-topic identification, the analytical method and available textual context can affect the topics identified and labeled. Consequently, the present design separates deterministic retrieval from contextual classification and retains a token-level audit trail.

Uncertainty, Attribution and Textual Origin

Uncertainty in journalism can be deployed to buttress claims and signal the limits of evidence. The result depends on the way it is communicated: numerical and verbal expressions do not evoke the same trust reactions (van der Bles et al., 2020) and limitations can be lost or restored as information passes through dissemination stages (Ratcliff, 2021). Uncertainty can also be an organizing frame rather than simply a lack of knowledge in fast-moving events (Capurro et al., 2021).

Source attribution is a textual difference which is useful for annotation. Reporter narration or direct or explicit quotations of source discourse may contain uncertain propositions. Attribution can foreground the source's assessment and background the journalist's evaluative role (White, 2012), but does not necessarily transfer pragmatic responsibility. Thus, the textual-locus variable codes for where an assessment occurs in the text, not for reporter endorsement, belief, or legal or moral responsibility.

Actor-focused AI-media research encourages attention to provenance. In AI reporting, Zai et al. (2025) found that economic and scientific actors had high standing, and that journalists themselves provided considerable contextualization. Dominant were progress and economic-consequence frames; underrepresented were ethics and morality. One way to study how journalists encode such source visibility in the grammar of single forecasts is through token-level attribution analysis.

AI News: Potential, Danger and Forward Focus

AI news often shows discursive ambivalence. Moreover, Cools et al. (2024) found that between 1985 and 2020, two elite U.S. newspapers provided more hopeful than pessimistic coverage, with ethical dilemmas prominently featured in dystopian frames. Early headlines about ChatGPT in the UK, according to Roe and Perkins (2023), 'mixed warnings of imminent, systemic danger with societal promise'. Nguyen and Hekman (2024) analyzed 3,098 articles from The New York Times, The Guardian, Wired, and Gizmodo and found that discourse around AI became increasingly critical over time, particularly regarding risk references to surveillance, data bias, cyberwar or cybercrime, and information disorder.

The evidence base grew after ChatGPT arrived, but researchers' methods have changed. Ryazanov et al. (2025) reviewed 5846 articles and found that interest in AI increased tenfold in the six months after ChatGPT's release; fears about AI and perceptions of it as a threat also increased. Ittefaq et al. (2025) examined framing, sentiment, and trends in newspapers from 12 countries from 2010 to 2023. To identify possible future directions, Jhan and Kuriakose (2026) reviewed 25,595 AI-related headlines published from August 2024 through December 2025. Wang et al. (2026) identified two general perspectives on AI in journalism—optimistic and pessimistic—in American, French, and Chinese media.

These studies show the benefits of prospective evaluation but tend to use documents, headlines, topics, or frames as units of analysis. They do not determine whether a given modal evaluates a proposition or characterises an affordance or occurs in reporter and not source discourse. This larger project can be enhanced by a corpus-based study of local form-meaning relations.

Method

Design and Scope

The study used a corpus-assisted case study design consisting of a mixture of deterministic lexical retrieval, contextual semantic coding and descriptive quantitative analysis. The retrieval unit was a single instance of one of the four target forms. The unit of analysis was a news story. This is an exploratory study and makes no claims that 20 articles are representative of TechCrunch, or English language technology journalism, or AI discourse more broadly.

Sampling and Corpus

The sampling frame was TechCrunch's public AI category (category ID 577047203; <https://techcrunch.com/category/artificial-intelligence/>) enumerated via the publisher's WordPress REST endpoint. The observation period was from 1 January 2022 to 31 August 2026; 2026 was thus a partial-year stratum. Article text and metadata were provided by canonical article pages. Appendix A lists all 20 retained sources, and Appendix B lists the archived source, extraction, and verification artifacts.

All strata used two fixed inclusive 14-day windows: Window 1, March 1–14, and Window 2, August 1–14. The windows were established before token collection and are about 5.5 months apart, both within the partial 2026 stratum. For each window, we queried the category endpoint in ascending order of publication date and recorded the exact query URL and the result position for each candidate.

An article was considered AI-focused if its primary subject was an artificial intelligence (AI) technology, an AI system or model, machine learning, generative AI, autonomous robotics, or the governance, business, research, or deployment of such a system. It had to be identifiable in the headline, dek, or lead and developed substantively in the editorial body; a passing reference to AI, or to a company known only in passing to use AI, didn't qualify. Eligible items were English editorial news, analysis, feature, interview, or roundup posts published by TechCrunch with a stable editorial body. Sponsored/partner material, event promotion, newsletters, podcasts, profile pages, video-only or gallery-only pages, duplicates, and pages without an accessible canonical article page were omitted.

The four target forms were not reviewed for eligibility. The first two articles identified by the AI-focus and article-type checks within a one-year window were selected in ascending order of displayed-publication-date; ties on the same date were resolved by canonical URL. If fewer than two eligible articles were available, we moved the window forward one day at a time up to seven days, noting the departure. No such extension was needed. All 20 final canonical URLs are stored in the source register along with selection ranks.

Table 1

Corpus composition and audited target-token yield by sampling period

Year	Article	Coverage	Target tokens
2022	4	Mar. 1–14 & Aug. 1–14	15
2023	4	Mar. 1–14 & Aug. 1–14	9
2024	4	Mar. 1–14 & Aug. 1–14	3
2025	4	Mar. 1–14 & Aug. 1–14	7
2026	4	Mar. 1–14 & Aug. 1–14	6
*			
Total	20	—	40

Note. The sample contains four articles per stratum. The equal article count supports only descriptive, article-balanced comparison. *2026 is partial-year coverage through August 31.

Retrieval and Source Reconstruction

All retrieved canonical pages were retrieved on October 4, 2026. Each page returned HTTP 200 on the first request and on one reload; the final canonical URL, redirect chain, retrieval timestamp, raw-HTML filename, and content hash are recorded in `source_register.csv` and the per-page access logs. The `REST` field `post.content.rendered` was used by the publisher to identify editorial content. The cleaned corpus kept the headline and the `p`, `h2`, `h3`, `li`, and `blockquote` text from that field. Scripts, styles, figures, asides, navigation, advertising, subscription prompts, recommendation modules, author bios, boilerplate, and comments were filtered out. The raw HTML and cleaned text were archived in the supplementary reproducibility package; thus, this study is reconstructing the live-page version as captured, not claiming the wording seen at first publication.

Token retrieval was scripted using Python 3.12. The code employed whole-word, case-insensitive regular expressions (`\bmay\b`, `\bmight\b`, `\bcould\b`, and `\blikely\b`) to match each cleaned headline and editorial body. For each result, Supplementary Data S1 provides the exact retrieval pattern, paragraph number, character offsets, sentence snippet, three-block context, canonical URL, capture time, and text hash. The script generated 40 valid tokens: 12 may, 7 might, 16 could, and 5 likely. An instance of ‘likely’ within ‘unlikely’ would not satisfy the word-boundary expression; calendar-month uses of ‘May’ would be preserved by the script but coded as non-epistemic if present. We report raw counts and article incidence, not normalised rates, because this small corpus is for contextual auditing, not rate estimation.

Coding in Context

The binary semantic and textual-locus rules were completed prior to coding the regenerated corpus. Each token was read in relation to its sentence and the surrounding editorial blocks. An epistemic token evaluates the truth, probability, inference, prediction or possible realisation of a proposition. Ability/capacity, affordance, and circumstantial or practical enablement are not separate count categories, but one umbrella code: ability/capacity refers to what an actor or system can do; affordance refers to a function made available by a design; and circumstantial or practical enablement refers to what conditions make an action feasible. A token may fit more than one of these examples but receives one non-epistemic code. The umbrella also includes fixed/conventional expressions. During manuscript revision, the relation between those examples was clarified, and no subcategory was added, and no token code changed. The distinction was particularly important for could: a claim that a system could generate images was coded non-epistemic when it meant “was capable of generating,” while a claim that an outcome could occur was coded epistemic when it meant “it was possible that. The fixed expression, as the case may be, would be retained in lexical retrieval but coded as non-epistemic.

Textual locus is the only operational term for the textual location of an epistemic assessment. Journalist-text locus is reporter narration without explicit external attribution. Attributed-text locus: a direct quotation or an explicitly attributed paraphrase of an external source. Textual locus was coded only for epistemic tokens; it does not measure endorsement, responsibility, or private belief on the part of a writer. The broader phrase textual source voice is used only as a conceptual description, not as a distinct coding variable.

Table 2

Operational coding framework

	Do	C	Rule	Reading
main	ode			
antic	Sem	E	Grades a proposition’s truth, probability, inference, prediction, or possible realization.	Possible or probable proposition.
antic	Sem	N	One umbrella: ability/capacity, affordance, enablement, permission, or fixed expression; examples are not separate subcounts.	Capacity or feasibility, not probability.
ual locus	Text	J	Epistemic assessment in reporter narration without explicit attribution.	Journalist-text locus.
ual locus	Text	A	Epistemic assessment in quotation or explicitly sourced paraphrase.	Attributed-text locus.

Note. E = epistemic; N = the single non-epistemic umbrella code; J = journalist-text locus; A = attributed-text locus. Textual locus is assigned only to epistemic tokens.

Audit, Coding, and Machine-Assisted Duplicate Check

Candidate enumeration, page retrieval, HTML capture, editorial-text extraction, hash generation, token retrieval, and aggregate calculations were all scripted. For contextual coding, we used the pre-finalized protocol in Table 2. All 40 token records were coded in two independent, blinded machine-assisted passes: gpt-5 as the primary pass and gemini-3.1-pro-preview as the duplicate pass. They received the

same protocol and context, but not the other pass’s labels. The supplement includes all decision notes, both coding files and the full prompt, and unprocessed model outputs.

The duplicate pass is not human second coding; its agreement statistics are not human interrater reliability. As an algorithmic audit diagnostic, the two passes agreed on 35 of 40 semantic classifications (87.5%; Cohen’s $\kappa = .679$). Of the 27 tokens both passes tagged as epistemic, textual-locus labels agreed for all 27 (100.0%; $\kappa = 1.000$). Five disagreements were adjudicated against the published protocol by inspecting the archived context; each preliminary label, final decision, and rationale is shown in adjudication_log.csv. A substantive limitation is that we still lack independent human coding, so we present the results as exploratory machine-assisted classification, rather than as a validated human-coded estimate.

Analysis

In the analysis raw frequency, share of all target tokens, article incidence, yearly distribution, semantic status and textual locus among the epistemic tokens are reported . Significance tests are not reported, as tokens are nested within articles that are purposively bounded, and the corpus is not a probability sample. We conduct a document-concentration sensitivity analysis, removing the single 2023 watermarking article with eight target tokens. This shows the impact of one document on descriptive totals.

Results

General Distribution and Incidence of Articles

There are 40 valid target tokens on the audited pages. Could is the most frequent form (16 tokens, 40.0%), may is second (12, 30.0%), might is third (7, 17.5%), and likely is fourth (5, 12.5%). Could is the most common document, appearing in 9 of the 20 articles (45.0%). These are counts of raw forms and not measures of epistemic function.

Table 3

Frequency, share, and article incidence of the four target forms

Marker	Tokens, n	Share	Articles	Incidence
may	12	30.0%	6/20	30.0%
might	7	17.5%	5/20	25.0%
could	16	40.0%	9/20	45.0%
likely	5	12.5%	4/20	20.0%
Total	40	100.0%	—	—

Note. The incidence of an article is the number of sampled documents that contained it at least once, not a rate per word, normalized or otherwise.

Distribution over sampling time

The five strata are descriptively different: 2022 contains 15 target tokens, 2023 contains 9, 2024 contains 3, 2025 contains 7 and the partial 2026 stratum contains 6. These numbers do not support a temporal trend at the population level. The 2023 total is document-sensitive: one watermarking article accounts for 8 of its 9 tokens. The effect of excluding that document is shown in Figure 1 and Table 7.

Table 4

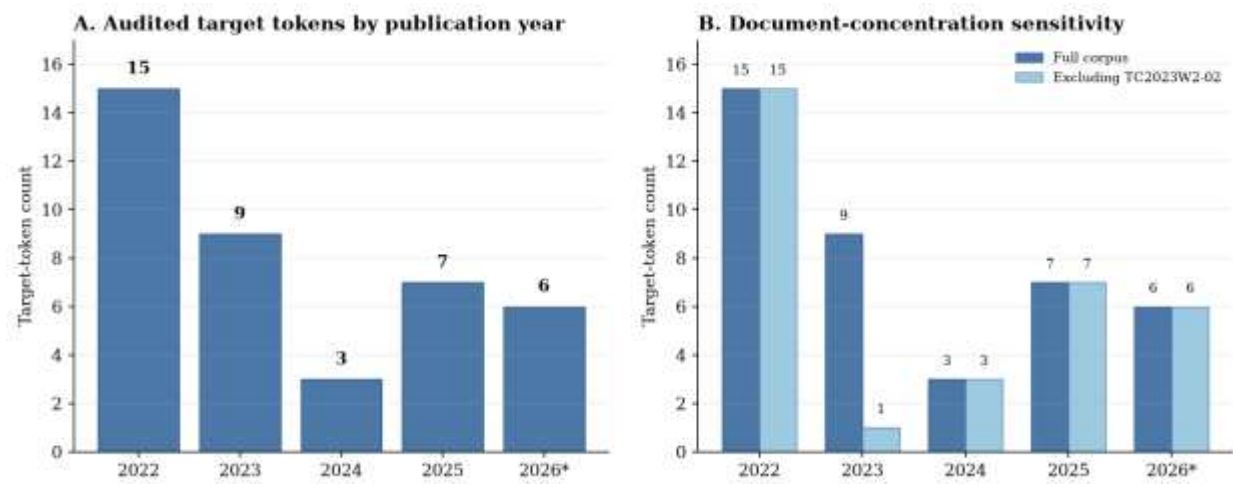
Distribution of target forms by publication year

Year	may	might	could	likely	Total	Token ns/article
2022	5	3	5	2	15	3.75
2023	1	3	4	1	9	2.25
2024	0	1	1	1	3	0.75
2025	3	0	4	0	7	1.75
2026*	3	0	2	1	6	1.50
Total	12	7	16	5	40	2.00

*Note. *2026 covers January 1 through August 31. Tokens per article is an article-balanced descriptive index, not a word-normalized rate.*

Figure 1

Temporal Distribution and Document-Concentration Sensitivity



Note. Panel A shows raw audited counts for four sampled articles per stratum. Panel B excludes the eight-token 2023 watermarking article (TC2023W2-02). *Reflects partial-year coverage for 2026.

Epistemic and Non-Epistemic Usage

Contextual coding showed that 29 of the 40 tokens (72.5%) were epistemic and 11 (27.5%) were non-epistemic. All might and likely tokens were epistemically. Of the 12 may tokens, 10 (83.3%) were epistemic; Could was less specialised: 7 of 16 tokens (43.8%) were epistemic, with 9 being ability, technical-capacity, or affordance uses. Thus, a form-only approach would more than double the observed epistemic count for could in this corpus, from 7 to 16.

Table 5

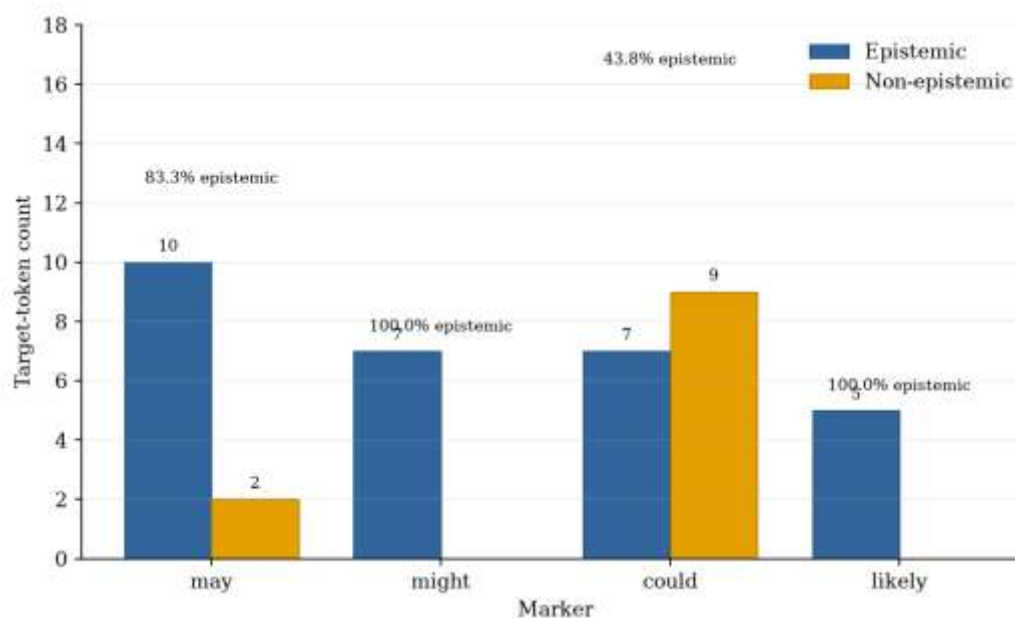
Semantic classification of the four target forms

Marker	Epistemic	Non-epistemic	Total	Epistemic %
may	10	2	12	83.3%
might	7	0	7	100.0%
could	7	9	16	43.8%
likely	5	0	5	100.0%
Total	29	11	40	72.5%

Note. Semantic decisions derive from the finalized protocol and the adjudicated 40-row S1 dataset. The reported machine-assisted agreement is an audit diagnostic, not human interrater reliability.

Figure 2

Epistemic and Non-Epistemic Uses by Target Form



Note. Counts are based on the 40-row adjudicated token dataset. Percentages above the bars are epistemic shares within each form.

Epistemic Assessments: Textual Locus

Of the 29 epistemic tokens, 20 (69.0%) are of journalist-text locus and 9 (31.0%) are of attributed-text locus. The 2025 stratum has the largest share of attributed text (4 of 5 epistemic tokens), but the pooled distribution is still descriptive and sensitive to document composition. These categories are for locating the textual site of the assessment, not the endorsement, belief, or responsibility.

Textual locus of epistemic assessments by sampling period

Year	Journalist-text	Attributed-text	Total epistemic
2022	9	1	10
2023	6	1	7
2024	1	1	2
2025	1	4	5
2026*	3	2	5
Total	20	9	29

Note. *2026 is partial-year coverage. Textual locus is assigned only to epistemic tokens.

Document concentration sensitivity

8 of 40 (20.0%) target tokens provided by the 2023 Steg.AI watermarking article (TC2023W2-02). Without it, the 2023 share of all tokens moves from 22.5% (9/40) to 3.1% (1/32). The change in the textual locus is smaller: journalist-text tokens decrease from 20 of 29 epistemic tokens (69.0%) to 15 of 23 (65.2%), and attributed-text tokens increase from 9 of 29 (31.0%) to 8 of 23 (34.8%). The article is judged as valid. The sensitivity analysis shows document dependence, not a reason for exclusion.

Table 7

Document-concentration sensitivity analysis

Indicator	Full corpus	Watermarking article excluded
All target tokens	40	32
2023 target tokens	9	1
2023 share of target tokens	22.5%	3.1%
Epistemic tokens	29	23
Journalist-text epistemic	20 (69.0%)	15 (65.2%)
Attributed-text epistemic	9 (31.0%)	8 (34.8%)

Note. The sensitivity analysis describes dependence on one eight-token document; it does not treat that document as invalid.

Discussion

Form Frequency Does Not Determine Epistemic Function

The main finding is a mismatch between frequency and contextual function. The most common and most widely distributed form is could, but 9 of its 16 uses are non-epistemic. This pattern fits the construct-validity problem discussed by Hollmann et al. (2024): could may be frequent because it refers to ability or technical affordance, not epistemic uncertainty. That ambiguity is structurally plausible in AI journalism: the beat repeatedly describes what models, devices, infrastructure, and users can do.

The four markers are, therefore, not interchangeable entries in a dictionary of uncertainty. Might and likely are purely epistemic in this corpus, May is largely epistemic (10 of 12), and could requires contextual disambiguation. This is a local descriptive result, but it supports validating could in context before aggregating it as an uncertainty marker.

Capacity and prediction are two different items of analysis.

Descriptions of current system capability that slide into predictions about future consequences are common in many AI reporting articles. A capacity statement and an uncertain forecast may share the same modal form, but make different claims. A form-only count thus mixes technological affordance (what an actor or system is able to do) with epistemic commitment (how likely a proposition is presented as being). Contextual coding preserves the separation of those two analytical objects.

The Document Effect of the 2023 Watermarking Article

The raw annual totals do not show a monotonic trend, but the 2023 value is still document sensitive. Eight of the nine 2023 tokens are in a single watermarking paper. Removing that document drops 2023 from 9 of 40 tokens to 1 of 32. The defensible conclusion is not that 2023 was different from other years in the population, but that one quotation-rich or technically detailed article can dominate a small corpus's time-stratified descriptive total.

This conclusion does not conflict with larger studies showing changes in AI news after ChatGPT. Ryazanov et al. (2025) find a tenfold increase in attention and a stronger danger/risk association; Ittefaq et al. (2025) present a longitudinal comparison of frames and sentiment across 12 countries; and Jhan and Kuriakose (2026) find strong future orientation in headlines from 2024–2025. Those studies have used different sampling frames and units. The present case shows that the frequency of four individual forms can depend heavily on article type, topic, and source structure.

Textual Provenance Is Not an Epistemic Attribute

The textual-locus code indicates whether the journalist-text or the attributed-text receives an epistemic assessment. In the full corpus, journalist-text tokens outnumber attributed-text tokens (20 vs 9), and removing the eight-token article only slightly shifts the balance. The code is a record of textual location, not a measure of author belief or endorsement.

The interpretation should be restrictive. White (2012) demonstrates how attribution can foreground source evaluation and background reporter voice, whereas Stenvall (2012) shows responsibility can remain shared or ambiguous. Here the code cannot reveal a journalist's private belief or show endorsement. But it provides a clear measure of textual provenance that future work can use to distinguish industry executives, financial analysts, researchers, regulators, civil-society actors, and anonymous sources.

Methodological Contributions and Limitations

This work presents an auditable workflow between local modal semantics and descriptive AI-media analysis. It begins with a fixed source frame, deterministic candidate selection, canonical URLs and page captures, a specified editorial-text layer, target forms fetched with a documented regular expression, and one contextual record per token. The workflow then separates semantic status from textual locus, stores duplicate machine-assisted codes and adjudications, and reconstructs each table reported in the released token file. These steps make the results inspectable rather than asserted.

There are major limitations. Twenty purposively bounded articles from one specialist outlet cannot provide population-level estimates, and 2026 coverage is partial. Publisher pages are mutable; the archive keeps the captured version, not necessarily the original wording on the day of publication. The study reports raw counts instead of word-normalized rates. Most importantly, the authors used machine-assisted, not independently human, duplicate coding; the agreement statistics are diagnostic only. A confirmatory extension should preregister the sampling frame, archive historical page versions, use consistent word denominators, and obtain independent human double coding with reported agreement, before testing population-level claims.

Conclusion

This exploratory case study demonstrates that may, might, could, and likely are not interchangeable measures of uncertainty in AI journalism. The most common form in this 20-article corpus is could, but 9 of its 16 uses refer to ability, capacity, or affordance rather than epistemic possibility. One 2023 document also accounts for a fifth of all target tokens, suggesting the need for a document-concentration check. The accompanying source register, page images, extraction logs, 40-row token file, blinded coding outputs, adjudication log, and aggregate scripts allow for an exact audit of these descriptive findings. These results are preliminary due to the small size of the corpus and the lack of human interrater reliability.

References

1. Capurro, G., Jardine, C. G., Tustin, J., & Driedger, M. (2021). Communicating scientific uncertainty in a rapidly evolving situation: A framing analysis of Canadian coverage in early days of COVID-19. *BMC Public Health*, 21(1), Article 2181. <https://doi.org/10.1186/s12889-021-12246-x>
2. Carretero, M. (2023). Knowledge and belief in the times of COVID-19: A comparative analysis of epistemicity in English newspaper discourse of two stages of the pandemic. *Applied Corpus Linguistics*, 3(2), Article 100054. <https://doi.org/10.1016/j.acorp.2023.100054>
3. Cools, H., Van Gorp, B., & Opgenhaffen, M. (2024). Where exactly between utopia and dystopia? A framing analysis of AI and automation in US newspapers. *Journalism*, 25(1), 3–21. <https://doi.org/10.1177/14648849221122647>
4. Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, 4(1), Article 100082. <https://doi.org/10.1016/j.acorp.2023.100082>
5. Gillings, M., & Jaworska, S. (2025). How humans and machines identify discourse topics: A methodological triangulation. *Applied Corpus Linguistics*, 5(1), Article 100121. <https://doi.org/10.1016/j.acorp.2025.100121>
6. Hollmann, W. B., Fujimoto, K., & Kuroda, M. (2024). Japanese EFL undergraduate students' use of the epistemic modal verbs may, might, and could in academic writing. *Language Learning in Higher Education*, 14(1), 21–40. <https://doi.org/10.1515/cercles-2023-0014>
7. Ittefaq, M., Zain, A., Arif, R., Ala-Uddin, M., Ahmad, T., & Iqbal, A. (2025). Global news media coverage of artificial intelligence (AI): A comparative analysis of frames, sentiments, and trends across 12 countries. *Telematics and Informatics*, 96, Article 102223. <https://doi.org/10.1016/j.tele.2024.102223>
8. Jhan, P. J., & Kuriakose, R. (2026). How is artificial intelligence framed in online news media? A data-driven analysis of news trends, sentiment, and media framing. *Quality & Quantity*, 60(5), 15345–15364. <https://doi.org/10.1007/s11135-026-02783-1>
9. Mak, M. H. C. (2024). Corpus linguistics will benefit from greater adoption of pre-registration: A novice-friendly split-corpus approach to pre-registration. *Applied Corpus Linguistics*, 4(3), Article 100111. <https://doi.org/10.1016/j.acorp.2024.100111>
10. Moran, R. E., & Shaikh, S. J. (2022). Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism*, 10(10), 1756–1774. <https://doi.org/10.1080/21670811.2022.2085129>
11. Nguyen, D., & Hekman, E. (2024). The news framing of artificial intelligence: A critical exploration of how media discourses make sense of automation. *AI & Society*, 39(2), 437–451. <https://doi.org/10.1007/s00146-022-01511-1>
12. Ratcliff, C. L. (2021). Communicating scientific uncertainty across the dissemination trajectory: A precision medicine case study. *Science Communication*, 43(5), 597–623. <https://doi.org/10.1177/10755470211038335>
13. Roe, J., & Perkins, M. (2023). ‘What they’re not telling you about ChatGPT’: Exploring the discourse of AI in UK news media headlines. *Humanities and Social Sciences Communications*, 10, Article 753. <https://doi.org/10.1057/s41599-023-02282-w>

14. Ryazanov, I., Öhman, C., & Björklund, J. (2025). How ChatGPT changed the media's narratives on AI: A semi-automated narrative analysis through frame semantics. *Minds and Machines*, 35(1), Article 2. <https://doi.org/10.1007/s11023-024-09705-w>
15. Stenvall, M. (2012). Responsibility and the conventions of attribution in news agency discourse. *Journal of Applied Linguistics and Professional Practice*, 9(3), 405–423. <https://doi.org/10.1558/japl.v9i3.20845>
16. van der Bles, A. M., van der Linden, S., Freeman, A. L. J., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14), 7672–7683. <https://doi.org/10.1073/pnas.1913678117>
17. Wang, Z., Downey, J., & Zhuang, X. (2026). Shared utopias, distinct anxieties: Sociotechnical imaginaries of AI-in-journalism in American, French and Chinese newspapers (2019–2025). *Journalism*. Advance online publication. <https://doi.org/10.1177/14648849261466375>
18. White, P. R. R. (2012). Exploring the axiological workings of ‘reporter voice’ news stories—Attribution and attitudinal positioning. *Discourse, Context & Media*, 1(2–3), 57–67. <https://doi.org/10.1016/j.dcm.2012.10.004>
19. Zai, F., Rohrbach, T., & Hänggli Fricker, R. (2025). Voices and media frames in the public debate on artificial intelligence: Comparing results from manual and automated content analysis. *Frontiers in Communication*, 10, Article 1599854. <https://doi.org/10.3389/fcomm.2025.1599854>

Appendix A. Data Sources and Retained Article Register

The primary discovery source was TechCrunch's public AI category (category ID 577047203), accessed through the publisher's WordPress REST posts endpoint. The category's public description identifies coverage of artificial intelligence, machine learning, generative AI, large language models, text-to-image/video systems, speech technologies, and predictive analytics. The category endpoint was queried separately for each fixed window in ascending publication-date order. The canonical article page is the primary source for identity and page-level verification. The supplementary source_register.csv retains the exact category query URL, result position, visible and JSON-LD date, final canonical URL, retrieval timestamp, two HTTP status checks, author, text hash, and archive file paths for every record.

Table

A1

Retained article sources (full retrieval timestamps and verification fields are in source_register.csv)

Sample	Sa	Publication date	Author	Article title and final canonical URL
C2022W1-01	T	2022-03-01 00:13 (site)	Ingrid Lunden	Starship Technologies raises another \$42M to fuel the growth of its fleet of self-driving delivery robots https://techcrunch.com/2022/03/01/starship-technologies-raises-another-42m-to-fuel-the-growth-of-its-fleet-of-self-driving-delivery-robots/
C2022W1-02	T	2022-03-01 07:18 (site)	Ron Miller	Zero Systems gets \$12M Series A to bring automation to professional services https://techcrunch.com/2022/03/01/zero-systems-gets-12m-series-a-to-bring-automation-to-professional-services/
C2022W2-01	T	2022-08-01 06:47 (site)	Paul Sawers	UiPath acquires London-based NLP startup Re:infer https://techcrunch.com/2022/08/01/uipath-acquires-london-based-nlp-startup-reinfer/
C2022W2-02	T	2022-08-02 05:58 (site)	Haje Jan Kamps	It's paintbrushes at dawn as artists feel the pressure of AI-generated art https://techcrunch.com/2022/08/02/ai-art-generated/
C2023W1-01	T	2023-03-01 09:32 (site)	Paul Sawers	Sesamm bags \$37M to give corporates ESG insights using natural language processing https://techcrunch.com/2023/03/01/sesamm-bags-37m-to-give-corporates-esg-insights-using-natural-language-processing/
C2023W1-02	T	2023-03-01 10:00 (site)	Kyle Wiggers	Addressing criticism, OpenAI will no longer use customer data to train its models by default https://techcrunch.com/2023/03/01/addressing-criticism-openai-will-no-longer-use-customer-data-to-train-its-models-by-default/
C2023W2-01	T	2023-08-01 00:01 (site)	Manish Singh	Ola's Aggarwal sets eyes on AI, semiconductor design https://techcrunch.com/2023/08/01/bhavishh-aggarwal-ola-ai-semiconductor/

Sample	Sa on date	Publicati	Author	Article title and final canonical URL
C2023W2-02	T 01 04:01 (site)	2023-08- Coldewey	Devin	Steg.AI puts deep learning on the job in a clever evolution of watermarking https://techcrunch.com/2023/08/01/steg-ai-puts-deep-learning-on-the-job-in-a-clever-evolution-of-watermarking/
C2024W1-01	T 01 02:30 (site)	2024-03- Singh	Manish	Elon Musk sues OpenAI and Sam Altman over ‘betrayal’ of nonprofit AI mission https://techcrunch.com/2024/03/01/elon-musk-openai-sam-altman-court/
C2024W1-02	T 01 09:32 (site)	2024-03- Wiggers	Kyle	AI chip startup Groq forms new business unit, acquires Definitive Intelligence https://techcrunch.com/2024/03/01/ai-chip-startup-groq-forms-new-business-unit-acquires-definitive-intelligence/
C2024W2-01	T 01 00:53 (site)	2024-08- Mehta	Ivan	Zuckerberg says Meta will need 10x more computing power to train Llama 4 than Llama 3 https://techcrunch.com/2024/08/01/zuckerberg-says-meta-will-need-10x-more-computing-power-to-train-llama-4-than-llama-3/
C2024W2-02	T 01 04:33 (site)	2024-08- Lomas	Natasha	The EU’s AI Act is now in force https://techcrunch.com/2024/08/01/the-eus-ai-act-is-now-in-force/
C2025W1-01	T 01 06:00 (site)	2025-03- Temkin	Marina	OpenAI’s startup empire: The companies backed by its venture fund https://techcrunch.com/2025/03/01/openais-startup-empire-the-companies-backed-by-its-venture-fund/
C2025W1-02	T 01 09:59 (site)	2025-03- Ha	Anthony	DeepSeek claims ‘theoretical’ profit margins of 545% https://techcrunch.com/2025/03/01/deepseek-claims-theoretical-profit-margins-of-545/
C2025W2-01	T 01 04:00 (site)	2025-08- Zeff	Maxwell	Google rolls out Gemini Deep Think AI, a reasoning model that tests multiple ideas in parallel https://techcrunch.com/2025/08/01/google-rolls-out-gemini-deep-think-ai-a-reasoning-model-that-tests-multiple-ideas-in-parallel/
C2025W2-02	T 01 06:21 (site)	2025-08- Bellan	Rebecca	Vast Data in talks with Alphabet’s CapitalG, Nvidia to fund round at up to \$30B valuation https://techcrunch.com/2025/08/01/vast-data-in-talks-with-alphabets-capitalg-nvidia-to-fund-round-at-up-to-30b-valuation/
C2026W1-01	T 01 06:00 (site)	2026-03- Madori Davis	Dominic-	SaaS in, SaaS out: Here’s what’s driving the SaaS apocalypse https://techcrunch.com/2026/03/01/saas-in-saas-out-heres-whats-driving-the-saasapocalypse/

Sample	Sa	Publicati	Author	Article title and final canonical URL
	T	2026-03-	Anthony	Anthropic's Claude rises to No. 1 in the App Store following
C2026W1-02	01 06:48 (site)	Ha		Pentagon dispute https://techcrunch.com/2026/03/01/anthropics-claude-rises-to-no-2-in-the-app-store-following-pentagon-dispute/
	T	2026-08-	Anthony	Sam Altman is still making the case for parenting via ChatGPT
C2026W2-01	01 10:07 (site)	Ha		https://techcrunch.com/2026/08/01/sam-altman-is-still-making-the-case-for-parenting-via-chatgpt/
	T	2026-08-	Anthony	YouTuber Hank Green says his AI usage is 'not healthy'
C2026W2-02	01 12:45 (site)	Ha		https://techcrunch.com/2026/08/01/youtuber-hank-green-says-his-ai-usage-is-not-healthy/

Appendix B. Verification, Editorial-Text Extraction, and Audit Package

Accessibility was assessed on the canonical article page in a non-authenticated HTTP session. A record passed only if it returned an HTTP 200 both initially and after one reload, and exposed a readable headline, author, publication date, and editorial body without login, subscription, or payment. The 20 selected records met those conditions. The access log for each page preserves the requested URL, final resolved URL, redirect chain, timestamps, status codes, and content hash.

Editorial content was separated from page furniture by using the publisher REST post.content.rendered field and retaining the headline plus paragraph, heading, list-item, and blockquote blocks. Scripts, styles, figures, asides, navigation, advertising, subscription prompts, recommendations, author biographies, boilerplate, and comments were excluded. Raw canonical-page HTML and the resulting clean text are both archived. Token retrieval, aggregate calculations, and figure generation were scripted; contextual coding was conducted through two blinded machine-assisted passes followed by protocol-based adjudication. The supplement preserves every intermediate and final output.

Table B1

Reproducibility files supplied with the manuscript

Artifact	Contents and verification use
source_register.csv	20 article-level records: discovery queries, final canonical URLs, capture times, status checks, authors, AI-focus rationale, filenames, and SHA-256 hashes.
raw_html/ and clean_text/	One direct-page HTML capture and one cleaned headline-plus-editorial-body text file per article.
access_logs/ and api_json/	Per-page HTTP/reload logs and publisher API records supporting date, text, and extraction checks.
S1_tokens.csv	40-row token dataset with canonical URL, retrieval time, text hash, offsets, sentence snippets, three-block context, both preliminary codes, final decision, and adjudication note.
coding_protocol.md, primary_coding.csv, duplicate_coding.csv, adjudication_log.csv, coding_agreement.json	Finalized rules, blinded machine-assisted passes, all five resolved disagreements, and agreement diagnostics explicitly not presented as human interrater reliability.
aggregate_tables.csv, results_summary.json, scripts/	Recomputed manuscript tables, result summary, and executable collection, retrieval, coding, adjudication, and figure scripts.

Note. The complete reproducibility package is released with the manuscript. Because live publisher pages can change, re-analysis should begin from the supplied captures and hashes; a changed live page must be reported as a changed version rather than silently substituted.