An LLM-Assisted Study of Attitude Construction in News Discourse: Evidence from the 2025 Sino-U.S. Trade Disputes

Tianran Wang
Jiangsu Food and Pharmaceutical Science College, Huai'an, Jiangsu, China

Abstract: **Drawing on the Attitude subsystem of Appraisal Theory, this study compares how evaluative meaning is constructed in early English-language reporting on the 2025 Sino-US trade disputes by two influential outlets: China Daily (CD) and The New York Times (NYT). Using Factiva, a comparable corpus of 82 reports (42 CD; 40 NYT) was compiled to represent the initial stage of the dispute. Methodologically, the study operationalizes Affect, Judgement, and Appreciation as an annotation scheme and employs a large language model (LLM) as an automated coder to extract attitude resources, label polarity (positive/negative), and identify the speaking actor in quoted (citizen/official/expert). Prompt design was refined through a pilot calibration against manual coding, yielding high agreement (precision 0.91; recall 0.88). On the full corpus, the LLM identified 473 attitude resources (CD 215; NYT 258). Results show convergence in attitude type distribution: both outlets rely most heavily on Judgement, indicating that trade actions are primarily moralized through responsibility and fairness claims. However, sourcing and polarity patterns diverge: CD attributes evaluation mainly to officials, whereas NYT foregrounds citizen voices; and NYT exhibits a stronger negative tilt, while CD displays a comparatively more constructive orientation. The paper argues that institutional role, audience design, and genre constraints jointly shape how trade disputes are personalized and framed. It also outlines a replicable LLM-assisted workflow and discusses reliability and error auditing for appraisal research.**

Keywords: Large Language Model; Appraisal Theory; Attitude Resources; China Daily; The New York Times; Trade Disputes

1. Introduction

Evaluative meaning is a key mechanism through which news discourse positions events, orchestrates reader alignment, and naturalizes ideology. Rather than functioning as neutral mirrors of reality, news reports select, foreground, and legitimate particular interpretations of events through patterned choices of lexis, attribution, and narrative structure. Within the systemic functional linguistic tradition, Appraisal Theory provides a principled framework for describing interpersonal meaning and, in particular, for mapping how writers construe Attitude (feelings, moral evaluations, and valuations of phenomena) in ways that invite readers to share or resist a proposed stance.

Research on high-salience issues, such as disasters, conflicts, and public-health crises, has repeatedly shown that media from different national contexts differ in their evaluative repertoires and in the discursive resources they mobilize to assign responsibility, justify action, and sustain legitimacy. In such reporting, attitude resources do not simply describe events; they continuously construct boundaries of meaning (e.g., what counts as "reasonable," "responsible," or "harmful") and boundaries of belonging (e.g., who is positioned as a cooperative partner versus a threatening other). These evaluative boundaries are particularly consequential in international disputes, where public understanding is often formed through mediated representations rather than direct experience.

Against the backdrop of shifting geopolitical and economic configurations, reporting on Sino-U.S. trade disputes is especially representative of discursive competition. Trade disputes are not only economic events; they are also symbolic contests over fairness, reciprocity, and national interest. News texts therefore become sites where policy moves are narrativized as either legitimate protection, unjustified coercion, prudent negotiation, or reckless escalation. Comparing Chinese and U.S. media coverage

can thus illuminate how distinct media systems and communicative objectives shape the interpretive horizons offered to readers.

Methodologically, appraisal-informed discourse analysis has long faced a tension between theoretical granularity and empirical scalability. Fine-grained identification of Attitude resources-especially implicit or evoked evaluation-often requires labor-intensive manual coding and iterative calibration among annotators [1]. Recent advances in AI large language models (LLMs) make it possible to automate parts of this workflow, potentially enabling larger corpora and more replicable results [2]. Yet the use of LLMs also raises methodological questions about taxonomy fidelity, recall, stability, and transparency, all of which are critical for journal-ready discourse studies.

To connect theoretical concerns in Appraisal Theory with empirical questions in comparative media studies, the study addresses three research questions:

RQ1: How do China Daily and The New York Times differ in the distribution of Attitude types (Affect, Judgement, Appreciation) in early reporting on the 2025 trade disputes?

RQ2: How do the two outlets differ in attribution patterns-i.e., which news actors (citizens, officials, experts) are positioned as sources of evaluative meaning?

RQ3: How do the outlets differ in evaluative polarity (positive vs. negative), and what institutional and communicative factors plausibly motivate these differences?

Conceptually, the paper contributes to comparative studies of international-issue reporting by specifying how different sourcing strategies may structure evaluative authority. Methodologically, it contributes a transparent LLM-assisted annotation pipeline for appraisal analysis, including prompt design, pilot calibration, and quantitative robustness checks that can be adopted or adapted in future research.

Section 2 reviews appraisal-informed media studies and recent computational approaches to attitude identification. Section 3 details corpus construction, annotation design, and LLM-based coding procedures. Section 4 reports quantitative distributions and interprets cross-outlet differences with reference to sourcing norms and communicative objectives. Section 5 concludes with implications, limitations, and directions for future work.

2. Literature Review

Appraisal Theory, most commonly associated with Martin and White's: The Language of Evaluation: Appraisal in English, offers a principled account of how writers encode evaluation to negotiate solidarity, authority, and legitimacy [3]. Within the framework, Attitude is typically operationalized as Affect (emotions), Judgement (normative evaluation of people and behavior), and Appreciation (evaluation of things, events, and processes), while Graduation scales intensity and Engagement manages dialogic positioning. For news discourse research, the Attitude system has been particularly productive because it allows analysts to link micro-linguistic choices to macro-social processes such as moralization, othering, legitimation, and the management of institutional credibility.

Trade disputes are reported not only as economic events but also as contests over legitimacy. Tariffs, export controls, and retaliatory measures are routinely framed through a moral vocabulary of fairness, reciprocity, and responsibility [4], which makes the Attitude system especially suitable for analysis. In this context, Judgement resources can function as a shorthand for assigning blame or credit (e.g., construing a move as 'unjustified' or 'necessary'), while Appreciation can stabilize broader valuations of policies and outcomes (e.g., 'damaging' consequences or 'productive' talks). Because trade policy is technical and temporally extended [5], journalists often rely on attributed voices to translate abstract measures into everyday stakes. Comparing outlets therefore requires attention not only to what is evaluated but also to who is licensed to evaluate and what kinds of experience or expertise are treated as persuasive.

Analytically, Appraisal studies also distinguish between inscribed evaluation (overt attitude lexis) and evoked evaluation, where stance is inferred from co-text, framing, and shared cultural knowledge [6]. Evoked attitudes are common in trade reporting, for instance when numerical losses, disrupted supply chains, or 'tit-for-tat' sequences invite readers to infer harm or irresponsibility without explicit labeling. This distinction matters for LLM-assisted workflows because recall tends to be higher for inscribed triggers than for subtle evocation,

reinforcing the need for targeted auditing of implicit cases in addition to overall count comparisons.

However, the analytic value of Appraisal Theory is frequently constrained by the costs and subjectivities of manual coding [7]. Although corpus methods can robustly quantify overt evaluative lexis, implied or context-dependent attitudes (evoked evaluation) are often under-represented because they require inference from co-text, socio-cultural knowledge, and rhetorical cues [8]. Manual annotation improves interpretability and can be disciplined through guidelines and reliability calibration, but it remains labor-intensive and still vulnerable to disagreement, particularly when deciding whether a span is evaluative, where its boundaries lie, and how implicitness should be treated.

Recent work has converged on computational solutions that aim to preserve the theoretical granularity of Appraisal Theory while enabling larger-scale analysis. One line of research develops supervised appraisal taggers trained on expert-annotated corpora and treats appraisal identification as a sequence-labeling problem [9]. Such models can be reproducible once trained, but their generalization is constrained by training-data availability, domain shift, and cross-lingual transfer challenges [10]. Another line of research deploys LLMs as automated coders or analytic assistants. LLMs can flexibly interpret context and produce structured labels without task-specific training, which is attractive for discourse studies that examine emergent events and rapidly shifting vocabularies [11]. At the same time, appraisal-specific evaluation studies caution that general-purpose chatbots may show uneven recall, hallucinated spans, and output instability across runs-risks that can undermine methodological accountability if not systematically managed.

For comparative news discourse, a further motivation for LLM-assisted workflows lies in quotation and attribution. News texts frequently distribute stance through reported speech, indirect quotation, and strategic attribution [12]. This distribution complicates Appraisal analysis because the analyst must decide whether an attitude is authored by the journalist or attributed to a quoted actor, and how the selection and framing of quotes still functions as a journalistic stance strategy. An automated workflow that records actor categories (citizen/official/expert) alongside Attitude type and polarity can therefore help connect appraisal analysis with media-sociological questions about sourcing and institutional authority.

Taken together, the literature suggests a pragmatic methodological position for the present study: LLMs are well suited for accelerating the first pass of attitude identification in large corpora, but journal-ready analysis benefits from prompt transparency, pilot calibration against human coding, recall auditing, and interpretable quantitative checks. In this sense, LLM outputs are treated not as an unquestionable ground truth but as structured evidence that supports discourse-analytic interpretation and facilitates replicability.

3. Methodology

To achieve the research objectives, this study draws on the Factiva database to compile a corpus of early reporting on the 2025 Sino-U.S. trade dispute from two authoritative outlets: China Daily (CD) from China and The New York Times (NYT) from the United States. CD was selected because it is a major English-language national newspaper in mainland China and a key conduit through which Chinese official perspectives are communicated to international audiences. NYT was selected as a high-profile U.S. newspaper with strong agenda-setting influence and an established reputation for international coverage.

In corpus collection, “China and United States” and “trade war” were used as the keyword string, and the time span was set to March 1 to May 31, 2025, corresponding to the initial stage of the renewed dispute when bargaining and policy signaling were most visible. To increase topical relevance, searches were limited to headlines and leading paragraphs, and duplicated articles were removed. The final dataset consists of 82 reports (42 CD; 40 NYT), a comparable volume that supports contrastive analysis while keeping manual checks feasible.

The unit of analysis is the attitude resource, i.e. a lexical or phrasal span that realizes Affect, Judgement, or Appreciation, whether explicitly inscribed or inferentially evoked. Prior to annotation, articles were cleaned to remove metadata (e.g., datelines) while preserving quotation structure. This preserves the central news-writing feature relevant to the study: the distribution of evaluation across attributed voices.

The study uses ChatGPT with the GPT-5.1 Thinking model (introduced in ChatGPT in November 2025, per OpenAI's product release notes). Given the compute demands of local deployment, an online model was selected to ensure stable performance and to facilitate replication by other researchers with similar access. To reduce randomness and improve reproducibility, the annotation was conducted with a consistent role specification, fixed output schema, and (where configurable) deterministic settings (e.g., low sampling variability).

To enhance replicability, annotation was conducted in a standardized two-step routine. Each article was processed as a standalone unit to preserve local quotation structure, and outputs were saved in a machine-readable table recording (a) the minimal attitude trigger, (b) category and polarity, (c) actor label, and (d) surrounding sentence context for later qualitative checking. Where the interface allowed, randomness was minimized through conservative generation settings (e.g., low temperature) and a fixed output schema. After the initial pass, a second pass was used on flagged cases to correct common boundary problems (over-long spans) and to disambiguate actor labels when attribution chains were complex (e.g., 'analysts said officials argued'). This procedure keeps the LLM's role explicit: it accelerates extraction and first-pass labeling, while interpretation and final inclusion remain under researcher control.

Before large-scale annotation, one article from each outlet was used for a pilot test to calibrate model behavior through prompt refinement. The two pilot articles were also manually annotated, and the results were compared with the LLM outputs using precision, recall, and weighted F-score. Precision was 0.91, recall was 0.88, and the weighted F-score was 0.895, indicating high agreement and supporting formal large-scale annotation. The final prompt used for annotation is presented below for transparency and replicability.

Role specification: You are a trained attitude-analysis expert who can identify different types of Attitude in news reporting.

Concept specification: First, learn Martin and White's definitions and examples of Affect, Judgement, and Appreciation, and clarify the scope of each attitude category [3]. Second, following the examples, determine the polarity categories of Positive and Negative. Finally, learn Bell's definitions of news actors [13] and clarify the categories of Citizen, Official, and Expert.

Task specification: Based on the above attitude categories, identify attitude resources in the given news report on the new round of the Sino-U.S. trade dispute; determine whether each attitude is positive or negative; and label the category of the news actor who voices the attitude. Note that a single text may contain multiple attitudes; identify all of them.

Output specification: List the attitudes in the order in which they occur in the text and use a coded schema that pairs each attitude category with its polarity (positive/negative).

In the annotation prompt, Attitude was defined following Appraisal Theory, and polarity was labeled as Positive or Negative. Actor categories were adapted from classic news-discourse work on quotation and sourcing and operationalized as: (i) citizen (non-elite individuals, business owners, consumers, workers), (ii) official (government actors and institutional representatives), and (iii) expert (academics, analysts, industry specialists). Actor labeling is an analytic simplification, but it enables a systematic comparison of how evaluative authority is distributed across voices in each outlet.

Beyond the initial pilot, two additional checks were used to increase methodological accountability. First, a random subset of automatically annotated instances was manually audited to identify common error types (missed implicit attitudes, over-extended span boundaries, and occasional actor misclassification in complex attributions). Second, a small subset of texts was re-run with the same prompt to assess output stability; highly variable cases were inspected to refine boundary guidelines. These checks do not eliminate all risk but help ensure that patterns reported in Section 4 are not driven by obvious systematic annotation artifacts.

Counts were aggregated by outlet to produce distributions of attitude type, actor source, and polarity. To test whether cross-outlet differences exceed what would be expected by chance, chi-square tests were conducted on contingency tables (outlet × category). Effect sizes are reported using Cramér's V, which is interpretable across different table sizes. These inferential statistics are used as supportive evidence alongside discourse-analytic

interpretation, rather than as substitutes for close reading.

4. Results and Discussion

Across the 82 news reports, the LLM identified a total of 473 attitude resources, including 215 in China Daily and 258 in The New York Times. Table 1 summarizes the distributions of attitude types, sources, and polarity.

Table 1. Distribution of Attitude Resources in Trade-Dispute Reporting by CD and NYT

Feature	Category	CD (n)	CD (%)	NYT (n)	NYT (%)
Attitude type	Affect	62	29	96	37
	Judgement	130	60	142	55
	Appreciation	23	11	20	8
	Total	215	100	258	100
Source	Citizens	45	21	116	45
	Officials	102	47	62	24
	Experts	68	32	80	31
Polarity	Positive	89	41	82	32
	Negative	126	59	176	68

To complement descriptive percentages, chi-square tests were conducted to assess whether differences between outlets are statistically reliable. Table 2 reports test statistics and effect sizes.

Table 2. Chi-Square Tests of Cross-Outlet Differences (Based on Table 1 Counts)

Comparison	Chi-square (df)	p-value	Cramér's V	Interpretation (approx.)
Attitude types	4.18 (2)	0.124	0.094	not statistically robust
Source/actor	38.45 (2)	<0.001	0.285	robust difference
Polarities	4.29 (1)	0.038	0.095	modest difference

4.1 Attitude Types: Why Judgement Dominates

In both outlets, Judgement accounts for the largest share of attitudes (60% in CD; 55% in NYT), while Appreciation is least frequent (11% and 8%). This convergence is plausibly driven by the event type: trade disputes are narrated as sequences of institutional actions (tariff hikes, restrictions, exemptions, negotiations), and evaluations therefore cluster around responsibility, legitimacy, and strategic competence-semantic domains typically realized as Judgement. Affect is present but secondary, and when it appears it often functions to dramatize consequences (e.g., uncertainty, concern, frustration) rather than to frame the dispute as an emotionally intimate story.

Statistically, the distribution of Attitude types does not differ strongly across outlets ($\chi^2(2)=4.18$, $p=0.124$), suggesting that the broad appraisal "menu" is constrained by the genre and the policy-centric nature of the topic. Nevertheless, the descriptive pattern that NYT uses Affect more frequently (37% vs. 29%) is consistent with the expectation that a market-oriented outlet may employ more affective cues to sustain reader engagement, while an outlet with a stronger institutional communication role may prefer a more restrained affective profile.

The Judgement-heavy profiles arise from recurrent evaluative scripts in trade-dispute narration. In CD, Judgement frequently supports claims about propriety and diplomatic responsibility, often pairing criticism of coercive measures with calls for dialogue and 'mutual respect'. In NYT, Judgement more often foregrounds accountability and consequences, construing actions as aggressive, risky, or politically motivated. Across both outlets, Judgement is commonly realized through responsibility predicates (e.g., 'to blame', 'to pressure'), deontic evaluations ('should', 'must'), and labels that imply moral standing (e.g., 'reasonable', 'hardline'). These realizations show how trade disputes are moralized: policy moves are turned into judgments about actors rather than only descriptions of market effects.

4.2 Sources of Evaluation: Attribution as a Stance Strategy

The most striking cross-outlet difference concerns who is positioned as the source of evaluation. CD attributes attitudes primarily to officials (47%), followed by experts (32%) and citizens (21%). NYT shows the opposite pattern for the two most frequent sources: citizens

account for 45% of attitudes, officials 24%, and experts 31%. This difference is statistically robust ($\chi^2(2)=38.45$, $p<0.001$) with a moderate effect size (Cramér's V=0.285), indicating that sourcing is not a minor stylistic variation but a central dimension of contrast.

From an Appraisal perspective, attribution matters because it distributes evaluative responsibility: journalists can advance a stance while maintaining an appearance of objectivity by locating evaluation in the mouths of quoted actors [14, 15]. In CD, foregrounding officials effectively elevates policy legitimacy and aligns reader interpretation with institutional priorities. In NYT, foregrounding citizens can personalize the dispute by anchoring evaluation in everyday experience (e.g., perceived costs, anxieties, or skepticism), thereby producing psychological proximity and narrative vividness. The near-expected distribution of expert voices in both outlets suggests that "expertise" functions as a shared credibility resource across media systems, even as the balance between institutional authority and experiential voice differs.

A robustness check based on standardized residuals (not shown in the table for brevity) indicates that the largest contributors to the sourcing difference are CD's under-representation of citizen voices and over-representation of official voices, with the reverse pattern in NYT. This aligns with the interpretation that sourcing is a communicative choice shaped by institutional role and target audience: CD's international readership may be expected to treat official positions as newsworthy cues, whereas NYT's domestic readership may be more responsive to individualized impact narratives.

The attribution contrast can also be read as a difference in how evaluative authority is staged. When officials dominate as sources, evaluation tends to align with policy rationales and institutional objectives, and the report can present stance as public-interest governance [16]. When citizens dominate, evaluation is more readily anchored in lived experience, uncertainty, and affective impact, which can intensify reader involvement even when journalists avoid overt editorializing [17]. Importantly, attribution does not eliminate authorial stance: journalists still shape alignment by selecting which voices enter the report, positioning them early or late, and framing quotes with reporting verbs that can amplify, soften, or distance evaluation. Therefore, differences in sourcing should be interpreted as differences in stance distribution strategies rather than as simple differences in 'bias'.

4.3 Polarity: Negativity, Escalation, and Constructive Framing

Both outlets predominantly construct negative attitudes, a pattern consistent with the news value of negativity and with the adversarial framing of a trade dispute. At the same time, CD exhibits a higher proportion of positive attitudes (41% vs. 32%), while NYT exhibits a higher proportion of negative attitudes (68% vs. 59%). The difference is statistically significant ($\chi^2(1)=4.29$, $p=0.038$) but the effect size is small (V=0.095), suggesting that the two outlets share a broadly negative evaluative orientation while differing modestly in the extent to which they highlight de-escalatory or solution-oriented frames.

Negative attitudes tend to cluster around evaluations of trade-restrictive measures (e.g., construing actions as harmful, unfair, coercive, or reckless), whereas positive attitudes more often attach to the prospect of negotiation, stabilization, and reduced uncertainty. In CD, positive attitudes frequently orient toward collective benefit (e.g., mutual gains, global economic stability), which is consistent with a constructive communication stance that foregrounds cooperation as an evaluative horizon. In NYT, the stronger negativity may reflect the narrative salience of conflict and the journalistic practice of highlighting costs, risks, and grievances to sustain public attention.

It is a common feature of news reporting that negative attitudes predominate. The renewed round of the Sino–US trade war is, in itself, a negative event, and factual reporting on such an event often entails the articulation of negative stances. In a broad sense, de Hoog and Verboon [18] argue that news consumption is positively associated with negative affect in news discourse; that is, the more negative attitudes a news text contains, the more likely it is to attract readers. This tendency, in turn, implies that commercially oriented outlets such as *The New York Times*, which operate under audience and market pressures, are structurally incentivized to construct negative attitudes more frequently. By contrast, the comparatively higher proportion of positive attitudes in *China Daily* may be

attributed to the "constructive approaches" commonly associated with Chinese media practice. Under this model, news coverage is more likely to foreground positive elements of events, cultivate hope, and encourage readers, while also orienting the reporting toward identifying solutions to the problems at issue.

4.4 Interpreting Cross-Outlet Patterns: Institutional Role, Readership, and Genre Constraints

Taken together, the results suggest a layered explanation. At the level of genre constraints, both outlets converge on Judgement-heavy evaluation because trade disputes are construed as sequences of strategic actions and counter-actions. At the level of institutional role, CD's reliance on official sources positions the state as the primary legitimate narrator of the dispute, while NYT's reliance on citizen sources positions the dispute as an issue with everyday consequences, thereby increasing personalization. At the level of readership and market dynamics, the comparatively higher Affect and negativity in NYT can be interpreted as consistent with engagement-oriented news values, whereas CD's comparatively higher positivity can be interpreted as consistent with solution-oriented or de-escalatory framing.

Importantly, these interpretations do not assume that attribution eliminates journalist stance. Rather, selection of quotes, ordering of voices, and the framing of reported speech are themselves stance resources. From an Appraisal Theory perspective, such practices can be understood as creating "evaluative prosody" across the text: even when individual attitude tokens are attributed, the cumulative distribution of who gets to evaluate and what gets evaluated constructs a coherent orientation toward the dispute.

4.5 Methodological Reflection: What LLM-Assisted Appraisal Analysis Adds

Methodologically, the study demonstrates that LLM-assisted annotation can support appraisal-informed discourse analysis in three ways. First, it increases scalability: 473 attitude instances across 82 reports can be identified and structured rapidly, enabling patterns to be examined across a corpus rather than inferred from a small set of illustrative texts. Second, it increases transparency when prompts, schemas, and pilot benchmarks are reported, allowing other researchers to replicate or challenge the procedure. Third, it supports mixed-method interpretation: quantitative distributions guide where close reading should focus (e.g., attribution patterns), while qualitative interpretation explains why these distributions matter for ideology and alignment.

At the same time, LLM outputs should be treated as fallible. The most consequential risks for appraisal analysis are (i) missed evoked attitudes (recall limitations), (ii) boundary inflation where a model labels a broad span rather than the minimal evaluative trigger, and (iii) occasional misclassification between Judgement and Appreciation in policy discourse where actions and policies blur. Accordingly, the workflow adopted here-pilot calibration, manual auditing of samples, and stability checks should be seen as minimal good practice for future LLM-assisted appraisal research.

5. Conclusion and Implication

Adopting an LLM as an annotator, this study identified attitude resources in early reporting on the 2025 Sino-U.S. trade disputes in China Daily and The New York Times. Descriptively, the two outlets display similar distributions of attitude types, with Judgement most frequent and Appreciation least frequent. Inferential checks suggest that this similarity is not a robust point of divergence, highlighting the constraining effect of genre and event structure on appraisal choices.

Clear differences emerge in attribution patterns and, to a smaller extent, evaluative polarity. China Daily attributes attitudes more often to officials, while The New York Times attributes attitudes more often to citizens; this contrast is statistically robust and suggests different strategies for distributing evaluative authority and constructing credibility. Both outlets predominantly construct negative attitudes, but China Daily shows a relatively higher share of positive attitudes, consistent with a somewhat more constructive or de-escalatory framing orientation.

These findings carry two implications. Substantively, they contribute to understanding how international economic disputes are moralized and personalized across media systems, showing that "who gets to evaluate" is as important as "what is evaluated." Methodologically, they support a pragmatic, transparent approach to LLM-assisted appraisal

analysis in which automated coding is paired with pilot benchmarking and targeted manual auditing.

Limitations should be noted. The corpus focuses on English-language reporting and on an early time window; evaluative patterns may shift in later stages as policy outcomes and domestic politics change. Only two outlets are examined; future work can expand to additional Chinese and U.S. media as well as to multilingual corpora that include Chinese-language reporting. Finally, LLM annotation remains sensitive to prompt design and model updates, which underscores the importance of releasing prompts, documenting settings, and reporting reliability checks. Future studies can also explore richer modeling of Engagement (e.g., hedging, concession, denial) to capture how attribution and dialogic positioning interact with Attitude in the construction of trade-dispute narratives.

Future research can extend the present design in several directions. First, adding additional outlets and expanding the timeline would test whether the observed sourcing asymmetry is specific to the initial bargaining stage or persists as negotiations evolve. Second, incorporating the Engagement and Graduation subsystems would capture how writers manage dialogic space (e.g., hedging, countering, evidentiality) and scale intensity, which likely interacts with polarity and attribution. Third, a mixed-method validation strategy could combine LLM annotation with targeted human double-coding of difficult cases (especially evoked attitude) and with cross-model comparison to estimate sensitivity to model choice. Finally, sharing prompts, coding guidelines, and de-identified annotation outputs would strengthen cumulative research by enabling reanalysis and methodological benchmarking across studies that apply Appraisal Theory to international economic conflict.

The results point to several actionable recommendations for future LLM-supported discourse analysis. First, researchers can improve auditability by separating the tasks of (a) span extraction and (b) label assignment. In practice, this means prompting the model to mark minimal triggers (the smallest phrase that carries evaluation), then running a second pass that labels only the extracted triggers. Such decomposition reduces boundary inflation and makes manual adjudication faster because analysts check shorter units. Second, recall auditing is essential for implicit evaluation: analysts can sample texts that contain high densities of quotation, metaphor, or irony and explicitly ask the model to list any "likely evoked" attitudes, then compare these suggestions with human judgments to estimate where the model systematically under-detects stance.

Third, stability should be treated as a measurable property rather than an assumption. Even with the same prompt, small variations can occur in actor labeling or in whether a borderline expression is counted as Judgement versus Appreciation. One practical strategy is to re-run a fixed subset of texts (e.g., 10% of the corpus) and compute agreement across model runs; instances with low stability can then be flagged for human review. In the present dataset, the strongest substantive finding, cross-outlet differences in sourcing, would likely remain robust under such checks because it is driven by large count differences. Nevertheless, routine stability reporting would strengthen the credibility of LLM-assisted appraisal studies in peer review.

Finally, for practice-oriented international communication, the sourcing pattern suggests that "voice design" is a strategic resource. If an outlet aims to increase affective resonance without abandoning institutional credibility, it may consider diversifying quoted citizens and small-business stakeholders while maintaining a backbone of official and expert voices. Conversely, an outlet that primarily emphasizes citizen experience may consider how expert and official voices can contextualize impacts and reduce the risk of presenting trade disputes as purely emotive conflicts. In this sense, Appraisal Theory combined with scalable LLM-assisted annotation offers not only an analytic lens but also a diagnostic toolkit for news organizations seeking to balance credibility, engagement, and constructive framing in reporting international economic disputes.

Acknowledgements

This study was carried out as a part of the 2025 Jiangsu Provincial Social Science Applied Research Boutique Project (Foreign Languages), No. 25SWC-61.

References

[1]Read, J. and J. Carroll, *Annotating expressions of Appraisal in English.*

Language Resources and Evaluation, 2012. 46(3): p. 421–447.

[2]Pang, T.T.-Y., *Leveraging large language models to supplement corpus-based inductive learning of Chinese as a second language.* Applied Corpus Linguistics, 2026. 6(1).

[3]Martin, J.R. and P.R.R. White, *The language of evaluation: Appraisal in English*. 2005, London: Palgrave Macmillan.

[4]Ferrari, M., F. Kurcz, and M. Pagliari, *Do words hurt more than actions? The impact of trade tensions on financial markets.* Journal of Applied Econometrics, 2022. 37(6): p. 1138–1159.

[5]Brown, S., *Free trade, yes; ideology, not so much: The UK's shifting China policy 2010-16.* British Journal of Chinese Studies, 2018. 8(1): p. 92–126.

[6]White, P.R.R., *Praising and blaming, applauding, and disparaging – solidarity, audience positioning, and the linguistics of evaluative disposition*, in *Handbook of interpersonal communication*, G. Antos and E. Ventola, Editors. 2008, De Gruyter Mounton: Berlin. p. 567–594.

[7]Taboada, M. and M. Carretero, *Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in English and Spanish.* Linguistics and the Human Sciences, 2012. 6(1-3).

[8]Thompson, G., *Affect and emotion, target-value mismatches, and Russian dolls: refining the Appraisal model*, in *Evaluation in context*, G. Thompson and L. Alba-Juez, Editors. 2014, John Benjamins: Amsterdam. p. 47–66.

[9]Gao, Q. and D.W. Feng, *Deploying large language models for discourse studies: An exploration of automated analysis of media attitudes.* PLoS One, 2025. 20(1): p. e0313932.

[10]Križan, A. and A. Barbič, *Appraisal Analysis and AI Chatbots: Do We Even Need Humans?* ELOPE: English Language Overseas Perspectives and Enquiries, 2025. 22(1): p. 35–52.

[11]Tian, L., et al., *Task and Sentiment Adaptation for Appraisal Tagging*, in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023. p. 1960–1970.

[12]Pounds, G., *Attitude and subjectivity in Italian and British hard-news reporting: The construction of a culture-specific reporter voice.* Discourse Studies, 2010. 12(1): p. 106–137.

[13]Bell, A., *The language of news media*. 1991, Oxford: Blackwell.

[14]Lee, G., *Verb objectivity and source qualification: Comparison of quotation attributions in offline and online newspapers*. Journalism, 2016. 18(7): p. 890–906.

[15]Sundar, S.S., *Effect of source attribution on perception of online news stories.* Journalism & Mass Communication Quarterly, 2016. 75(1): p. 55–68.

[16]Dür, A. and B. Schlipphak, *Elite cueing and attitudes towards trade agreements: the case of TTIP.* European Political Science Review, 2020. 13(1): p. 41–57.

[17]Kleemans, M., G. Schaap, and L. Hermans, *Citizen sources in the news: Above and beyond the vox pop?* Journalism, 2016. 18(4): p. 464–481.

[18]de Hoog, N. and P. Verboon, *Is the news making us unhappy? The influence of daily news exposure on emotional states.* Br J Psychol, 2020. 111(2): p. 157–173.

