Framing Analytics with Large Language Models in Congressional Transcripts

Sebastián Concha-Macías[1][0009−0003−1857−6964] and Brian Keith-Norambuena[1][0000−0001−5734−8962]

Department of Computing & Systems Engineering, Universidad Católica del Norte, Chile
sebastian.concha@ucn.cl,brian.keith@ucn.cl

Abstract. Large Language Models (LLMs) enable detailed analysis of thematic structures in political text, but reproducible validation of frame extraction remains limited. This study examines Chilean congressional debates and develops a validation methodology based on frame-evidence coherence. An LLM extracts frames with supporting evidence quotes, mapping them to a nine-category taxonomy, while a rule-based regex extractor provides a baseline. Frame quality is validated with an LLM-as-a-judge approach that rates coherence between frames and evidence on a 1–5 scale. This approach has a high inter-rater reliability with $\text{ICC}(2,1) = 0.927$ for single judges and $\text{ICC}(2,k) = 0.975$ for averaged scores. The mean coherence score of 4.06 ± 0.82 indicates a high alignment between frames and their corresponding evidence. Perturbation tests confirm the validity of the coherence score, as this metric drops significantly when evidence is mismatched ($\Delta = 2.15$, $p < 0.001$) or paired with wrong frames ($\Delta = 2.57$, $p < 0.001$). Finally, a multi-model comparison between LLMs shows that generally larger models achieve higher extraction consistency with a Jaccard similarity of 0.91 for GPT-5.

Keywords: Frame-evidence coherence · Framing analysis · Frame extraction · Information retrieval · Large language models

1 Introduction

Legislative debates are key arenas in which political actors define problems, priorities, and justify policy choices. *Framing* is a key part of these communication processes, emphasizing specific aspects of an issue to shape how it should be interpreted or evaluated [7]. In particular, the framing process [19] organizes the political meaning by emphasizing specific causes, moral evaluations, and potential solutions [7], with frames playing strategic roles in political mobilization [23]. Thus, analyzing the emergence and evolution of frames between sessions of congress offers insights into political agenda-setting, polarization, and the formation of collective interpretations.

Large Language Models (LLMs) are transforming computational social science [33], offering new capabilities for large-scale analysis of political documents

[37] through tasks such as classification, topic identification, and summarization. However, their application to framing analysis introduces several methodological challenges. In general, computational approaches to frame detection [22], including both LLM-based and classifier-based methods, face disagreement and consistency issues that complicate validation [30].

Recent studies have demonstrated that LLMs can match or exceed human annotation quality for political text classification tasks. Gilardi et al. [9] showed that ChatGPT outperforms crowd workers by 25 percentage points in frame detection tasks, while Törnberg [34] found that GPT-4 achieves higher accuracy than expert coders across multiple languages with zero-shot learning. However, these capabilities come with reproducibility concerns: González-Bustamante [10] argues that only open-source LLMs ensure full reproducibility in social science.

Unconstrained LLM output frequently produces labels that are overly specific, insufficiently comparable across documents, or unstable across runs. Evaluating the coherence of LLM-generated summaries is also non-trivial, as LLM evaluators have been shown to rate candidates inconsistently [28]. In addition, the effects of specific design decisions (fixed taxonomies, rule-based cues, and prompt structures) remain inadequately examined in framing analytics.

This study investigates these challenges by analyzing 18 Chilean congressional transcripts held after October 2019, contributing to the emerging literature on computational analysis of parliamentary discourse [16, 8, 32]. The objective is to evaluate how different methodological components influence three key dimensions of framing analysis: **(i)** the stability of frame assignments across sessions, **(ii)** the coherence of LLM-generated summaries, and **(iii)** the reproducibility of corpus-level statistics. To this end, we implement a hybrid pipeline that integrates a fixed frame taxonomy with two extraction modules: a rule-based detector of lexical indicators and an LLM-based classifier that maps candidate labels onto the predefined categories.

To assess how methodological decisions shape outputs, we compare rule-based extraction with LLM-based extraction under the fixed taxonomy and evaluate extraction consistency across multiple LLM models. In particular, this study provides a case study of framing analytics [21, 4] in legislative transcripts using large language models [8] in the context of the Chilean social outbreak [20].

2 Materials and Methods

This section describes the corpus of congressional debates, the preprocessing steps applied to the transcripts, the fixed frame taxonomy used to ensure comparability, the extraction procedures based on rule-based cues and LLM-based classification, and the evaluation of coherence-evidence pairs with independent LLM-based judges [11]. The complete prompts and regex patterns are available in the GitHub repository.[1]

The analysis focuses on 18 transcripts of Chilean congressional debates recorded between late 2019 and early 2020, obtained from publicly accessible parliamen-

[1] https://github.com/briankeithn/worldcist-framing-analytics

tary records.[2]. Sessions address issues related to social unrest, public security, institutional reforms, and economic recovery. Each session is treated as a single document. Two short sessions with nearly identical agendas were merged, resulting in 18 units. All transcripts were processed as plain text. Basic cleaning removed headers, footers, and repeated formatting marks. Sentences were segmented using a rule-based Spanish sentence splitter; no stopword removal or filtering was applied. The sessions have between about 180 and 400 sentences.

A fixed frame taxonomy was constructed to reduce label proliferation and improve consistency across documents. The categories were informed by research on political communication and social movements [7, 23, 27, 31] and adapted to the Chilean context of the 2019 "Estallido Social" period [24, 5, 3]. Thus, these categories integrate established media framing dimensions [27]—conflict, human interest, economic consequences, morality, and responsibility—with social movement framing concepts [31, 2].

This constrained approach addresses the "under-specified label space" problem identified in recent work [35], where unconstrained LLM outputs produce labels that are overly specific or inconsistent across documents. The taxonomy consists of the following nine categories:

- *Crisis/Urgency*: Emergency situations requiring immediate action with critical deadlines.
- *Human Rights*: Fundamental rights, dignity, and denouncing state violations.
- *Equity/Inequality*: Distributive justice, gaps between rich/poor, and elite abuse.
- *Institutionality/Legality*: Proper procedures, legal frameworks, and constitutional order.
- *Responsibility/Accountability*: Assigning blame and demanding explanations.
- *Participation/Dialog/Unity*: Collective action, social pacts, and bridging divides.
- *Social Protection*: State safety nets, including impacts on citizens' basic welfare.
- *Security/Public Order*: Public safety, policing, military, and emergency states.
- *Democracy/Constitution*: Constitutional reform, plebiscites, and democratic renewal.

Two extraction methods were used to extract frames. The first method is a rule-based baseline that uses manually defined lexical indicators associated with each frame category. Regular expressions capture expressions related to each frame of the taxonomy. Although more limited in coverage, this approach provides a reference against which to compare the labels extracted with LLMs.

For the second method, an LLM receives the complete transcript and extracts frames along with supporting evidence. Each detection includes: **(i)** a frame label mapped to one of the categories, **(ii)** an evidence quote consisting of 1–2 sentences that directly support the frame detection, and **(iii)** a confidence rating (high, medium, or low). The prompts explicitly instruct the model to

[2] Cámara de Diputadas y Diputados de Chile: https://www.camara.cl

avoid introducing new frame categories. Furthermore, low confidence detections are excluded from the evaluation.

We note that possible alternatives to extraction include supervised and semi-supervised machine learning approaches [6] that require labeled data. However, these alternatives are cost-intensive because of their supervised nature, as they require having labeled data available. Thus, we opted to only compare against rule-based approaches as a non-LLM baseline.

To assess extraction consistency across models, we conducted a multi-model comparison using four LLM variants (GPT-5.1, GPT-5, GPT-5-mini and GPT-5-nano). Each model performed three independent extraction runs on all sessions. Consistency was measured using pairwise Jaccard similarity of frame sets across runs, allowing us to identify which models produce stable assignments.

Frame detection quality is validated through frame-evidence coherence scoring. This approach assesses whether the extracted evidence actually supports the detected frame. For each frame-evidence pair, $K = 3$ independent LLM judges rate coherence on a scale of 1 to 5, where 1 indicates no connection between evidence and frame, and 5 indicates direct support. This approach is based on the LLM-as-a-judge paradigm established by Zheng et al. [36], who demonstrated that GPT-4 achieves over 80% agreement with human preferences.

We adopt a coherence scoring framework similar to G-Eval [18], which uses structured evaluation criteria to improve alignment with human judgments, and narrative coherence evaluation with LLM-as-a-judge paradigms for adjacent text analytics tasks, such as structured narrative extraction [14]. Formally, let (f_i, e_i) denote a frame-evidence pair, where f_i is the frame label and e_i is the supporting evidence quote. Each of K judges provides a coherence rating $y_{ik} \in \{1, 2, 3, 4, 5\}$. The mean coherence for pair i is $C_i = \frac{1}{K} \sum_{k=1}^{K} y_{ik}$. Thus, the overall coherence across all N frame-evidence pairs is computed as the average $\bar{C} = \frac{1}{N} \sum_{i=1}^{N} C_i$.

Inter-rater reliability is assessed using the intraclass correlation coefficient for a two-way random effects model [29]. ICC(2,1) measures the reliability of a single judge, while ICC(2,K) measures the reliability of the mean of the judges.

We conduct perturbation tests to make sure that the coherence metric captures meaningful frame-evidence relationships. Two perturbation conditions were assessed: (i) *random evidence*, in which evidence quotes are randomly rearranged between frames, and (ii) *wrong frame*, where each frame is associated with evidence from an alternate frame category. A valid coherence metric should show lower under both conditions in comparison to the original frame-evidence pairs. We use t-tests to figure out if this difference is statistically significant.

This validation strategy is based on the behavioral testing principles [26] and counterfactual evaluation methods [13], which show that perturbation-based tests can determine whether models identify significant relationships or depend on spurious correlations. To evaluate the quality of frame extraction, we compare rule-based and LLM-based extraction methods and assess extraction consistency between four LLM variants of different sizes. This addresses broader concerns about reproducibility in LLM-based research, where run-to-run variability and model versioning can threaten scientific replication [1].

3 Results

We now present the resulting frame distribution, agreement between extraction methods, frame-evidence coherence evaluation, inter-rater reliability, perturbation validation, and multi-model comparison. We note that all primary analyzes use GPT-5.1, while the multi-model comparison also evaluates GPT-5, GPT-5-mini, and GPT-5-nano to assess extraction consistency across model sizes.

3.1 Frame Frequencies and Distribution Across Sessions

Across the 18 debates, the taxonomy achieves full coverage with all nine frame categories detected at least once. The most frequent category is *Crisis / Urgency*, appearing in 16 of 18 sessions (88.9%), followed by *Institutionality / Legality* (14/18; 77.8%) and *Security / Public Order* (12/18; 66.7%). This pattern reflects the context of the period, marked by social unrest and institutional uncertainty. The high presence of urgency, institutional procedure, and public security is consistent with crisis-oriented framing in political communication research.

The less frequent categories include *Human Rights*, *Responsibility / Accountability*, *Equity / Inequality*, and *Democracy / Constitution*, each appearing in approximately one-third of sessions. Rather than indicating an absence of these themes, this distribution reflects the priorities of legislators during the crisis period, with institutional and security concerns dominating the discourse.

Table 1. Frame frequencies across the 18 legislative sessions.

Frame	Sessions Present	Share (%)
Crisis/Urgency	16	88.9
Institutionality/Legality	14	77.8
Security/Public Order	12	66.7
Participation/Dialog/Unity	10	55.6
Social Protection	10	55.6
Human Rights	7	38.9
Responsibility/Accountability	7	38.9
Equity/Inequality	6	33.3
Democracy/Constitution	6	33.3

To visualize how frames vary between documents, Figure 1 presents a frame-by-session heatmap. Most sessions contain the *Crisis / Urgency* and *Institutionality / Legality* frames, whereas others display a more mixed pattern involving security and social protection frames.

3.2 Comparison Between Rule-Based and LLM-Based Extraction

We compared the categories detected by the rule-based (regex) method with those assigned by the LLM. The regex baseline uses conservative phrase-based

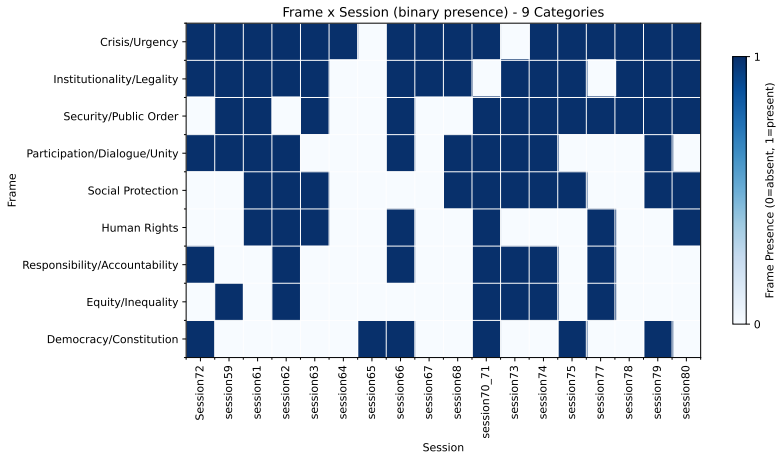

Fig. 1. Frame-by-session binary heatmap of the presence of each frame in the debates.

Table 2. Overlap between rule-based and LLM-based frame extraction. Note that sessions 70 and 71 were merged due to their short length and nearly identical agenda, as defined by the official transcriptions.

Session	#LLM	#Regex	Overlap	Jaccard
Session 72	5	9	5	0.56
Session 59	5	8	5	0.63
Session 61	6	5	4	0.57
Session 62	7	9	7	0.78
Session 63	5	9	5	0.56
Session 64	1	9	1	0.11
Session 65	1	9	1	0.11
Session 66	7	7	6	0.75
Session 67	2	9	2	0.22
Session 68	4	9	4	0.44
Session 70/71*	8	7	7	0.88
Session 73	6	9	6	0.67
Session 74	7	8	6	0.67
Session 75	5	9	5	0.56
Session 77	5	9	5	0.56
Session 78	3	8	3	0.38
Session 79	6	8	5	0.56
Session 80	5	7	4	0.50

patterns designed for reliable counting, detecting frames only when specific multi-word expressions appear (e.g., "seguridad social" rather than just "seguridad"). The LLM typically detects 1–8 frames per session. The mean Jaccard similarity between the two methods is 0.53, indicating moderate agreement.

3.3 Frame-Evidence Coherence

Table 3 reports the frame-evidence coherence scores by frame category. The overall mean coherence across 203 frame-evidence pairs is 4.06 ± 0.82, indicating that the extracted evidence strongly supports the detected frames. Per-frame coherence ranges from 3.67 (*Equity / Inequality*) to 4.38 (*Democracy / Constitution*), with all categories achieving scores above the midpoint of the scale.

Table 3. Frame-evidence coherence scores by frame category.

Frame	Mean	Std	n
Democracy/Constitution	4.38	0.59	15
Institutionality/Legality	4.24	0.80	45
Security/Public Order	4.18	0.73	20
Human Rights	4.10	0.89	16
Responsibility/Accountability	4.04	0.63	8
Crisis/Urgency	3.99	0.70	39
Social Protection	3.91	1.02	27
Participation/Dialog/Unity	3.85	0.70	24
Equity/Inequality	3.67	0.99	9
Overall	**4.06**	**0.82**	**203**

3.4 Inter-Rater Reliability

The three LLM judges show high agreement in their coherence ratings. Inter-rater reliability, estimated via the intraclass correlation coefficient, indicates excellent agreement with $\mathrm{ICC}(2,1) = 0.927$ (single-judge reliability) and $\mathrm{ICC}(2,k) = 0.975$ (three-judge mean reliability). These values indicate that any single judge provides a reliable estimate of frame-evidence coherence, and that the averaged score is highly stable. The high ICC values suggest that the coherence construct is well-defined and that judges apply consistent criteria when evaluating frame-evidence relationships. Furthermore, these ICC values substantially exceed typical inter-annotator agreement in political text classification, where human coders often achieve lower reliability [9]. The high agreement aligns with recent findings that LLM evaluators can provide consistent judgments when given well-defined evaluation criteria [15].

3.5 Perturbation Test Validation

To confirm that the coherence metric captures meaningful frame-evidence relationships, we conducted perturbation tests on a sample of 30 frame-evidence pairs (see Table 4). We randomly selected subsamples to balance statistical power with computational cost, as each pair requires evaluation under three conditions

Table 4. Perturbation test results for frame-evidence coherence.

Condition	Mean	Std	p-value
Original pairs	4.06	0.79	—
Random evidence	1.94	1.28	< 0.001
Wrong frame	1.52	0.71	< 0.001

by three independent judges. This sample size provides sufficient power to detect the large effect sizes observed while maintaining efficiency.

When evidence is randomly shuffled between frames, coherence drops by 2. points ($p < 0.001$). When frames are systematically paired with evidence from a different category, coherence drops by 2.54 points ($p < 0.001$). Both conditions have scores lower than the original pairs, validating that the coherence metric captures frame-evidence relationships rather than superficial text features.

3.6 Multi-Model Comparison

To assess extraction consistency, we compared four LLM variants across three independent runs per model. Table 5 reports the mean Jaccard similarity of frame sets. GPT-5 achieves the highest consistency (Jaccard = 0.91), indicating that it provides nearly identical frame assignments across independent runs. The smaller variants (GPT-5-mini and GPT-5-nano) show lower consistency, with higher variability. GPT-5.1 falls between these extremes.

Table 5. Extraction consistency across LLM models (mean Jaccard similarity).

Model	Mean Jaccard	Std
GPT-5	0.908	0.206
GPT-5.1	0.752	0.283
GPT-5-nano	0.601	0.476
GPT-5-mini	0.554	0.414

These patterns suggest that larger models provide more stable frame extraction, which is important for reproducibility in framing analytics pipelines. These findings corroborate concerns about LLM reproducibility in social science research [10]. While Renze and Guven [25] showed that temperature does not significantly affect consistency, model versioning poses challenges for replication.

4 Discussion

The findings of this study primarily contribute to the understanding of *framing* in legislative discourse and the validation of LLM-based frame extraction. Our

results extend recent work on LLM-based annotation of political texts [9, 34] to the domain of framing analysis in legislative discourse, demonstrating that constrained extraction with evidence validation can produce reliable frame assignments. The fixed taxonomy, combined with evidence-based extraction and coherence validation shows that LLMs can produce reliable frame assignments.

The consistent appearance of categories such as *Crisis / Urgency*, *Institutionality / Legality*, and *Security / Public Order* reflects stable framing patterns that legislators repeatedly activate to contextualize proposals, justify interventions, or assign responsibility. The full taxonomy coverage (all nine categories detected across the corpus) indicates that the constrained taxonomy is well-aligned with the thematic content of the debates, while the distribution of frame frequencies reflects the crisis-oriented priorities of the period. The prevalence of *Crisis / Urgency* (88.9%) and *Institutionality / Legality* (77.8%) suggests that legislators prioritized procedural legitimacy and emergency response over redistributive or rights-based arguments; the relative underrepresentation of *Equity / Inequality* and *Human Rights* frames indicates that despite widespread social grievances, dominant legislative framing focused on restoring institutional order.

Frame-Evidence Coherence. The frame-evidence coherence methodology validates our approach. The high overall coherence score (4.06 ± 0.82) indicates that the LLM successfully identifies evidence that genuinely supports each detected frame. The variation across frame categories (from 3.67 for *Equity / Inequality* to 4.38 for *Democracy / Constitution*) suggests that some frames may be more clearly articulated in legislative discourse than others. For instance, a *Participation / Dialog / Unity* frame was detected with evidence: "*promover un pacto social que involucre a los sectores de la sociedad civil, a los empresarios, a los trabajadores, a los sectores políticos*" ("promote a social pact involving civil society, employers, workers, political sectors"). All three judges rated this pair 5/5, confirming the explicit invocation of collective action language.

Reliability. The high inter-rater reliability ($\text{ICC}(2,1) = 0.927$) is higher than the typical values found in human annotation studies, suggesting that the coherence metric is well-defined and that LLM judges consistently apply the evaluation criteria. This finding aligns with previous findings in the literature, such as the Prometheus framework [15], which achieved a 0.897 Pearson correlation with human assessors using a rubric-based assessment.

Perturbation Analysis. The perturbation tests provide strong validation of the coherence metric, following the behavioral testing paradigm [26]. Kaushik et al. [13] showed that models trained on original data often fail on counterfactually revised inputs, our perturbation results demonstrate that the coherence metric does not exhibit such brittleness. The drops in coherence when the evidence is mismatched (random: $\Delta = 2.12$) or paired with incorrect frames ($\Delta = 2.54$) show that the metric captures the frame-evidence relationships rather than the superficial textual features, supporting the use of the frame-evidence coherence as a quality indicator for frame extraction. The frame-evidence coherence approach addresses what Jacovi and Goldberg [12] term the distinction between *plausibility* and *faithfulness* in model explanations.

Multi-Model Comparison. The multi-model comparison shows differences in extraction consistency between LLM variants. GPT-5 achieves near-perfect consistency (Jaccard = 0.91), while smaller models show more variability. This finding has practical implications: reproducible framing analytics requires models that produce stable outputs, and the choice of model can affect the reliability of downstream analyzes. The moderate agreement between LLM and regex extraction (mean Jaccard = 0.53) reflects their complementary nature. The regex baseline uses conservative phrase-based patterns, while the LLM can identify frames from contextual cues and paraphrased expressions.

Limitations. The use of proprietary models raises concerns about reproducibility [10]. Future studies could employ open-source alternatives to ensure complete reproducibility, though this can involve trade-offs in extraction quality [17]. Furthermore, we note that the primary analyzes use GPT-5.1, which the multi-model comparison revealed to have a lower extraction consistency (Jaccard = 0.75) than GPT-5 (Jaccard = 0.91). This model choice was made prior to conducting the comparative analysis. However, the high frame-evidence coherence scores and high inter-rater reliability suggest that GPT-5.1 produces valid extractions despite its lower run-to-run consistency.

5 Conclusion

This study examined how fixed taxonomies, evidence-based extraction, and LLMs can be combined to analyze and validate framing patterns in congressional debate transcripts. By applying a constrained set of frame categories to Chilean legislative sessions, the analysis shows stable thematic structures dominated by *Crisis / Urgency*, *Institutionality / Legality* and *Security / Public Order*.

The frame-evidence coherence methodology validates our frame extraction method. The high overall coherence (4.06 ± 0.82 indicates that the extracted evidence supports the detected frames. The three-judge panel achieved high inter-rater reliability. Perturbation tests confirmed that the metric captures frame-evidence relationships, with coherence dropping when evidence is mismatched.

The multi-model comparison shows that larger LLMs produce more consistent frame assignments than smaller variants, with practical implications for reproducibility. The moderate agreement between LLM and regex extraction reflects their complementary strengths, suggesting that combining both methods could provide useful cross-validation.

Acknowledgments. This study is funded by the ANID FONDECYT 11250039 Project. The corresponding author is also supported by Project 202311010033-VRIDT-UCN.

References

1. Barrie, C., Palmer, A., Spirling, A.: Replication for language models problems, principles, and best practice for political science. URL: https://arthurspirling.org/documents/BarriePalmerSpirling TrustMeBro. pdf (2024)

2. Benford, R.D., Snow, D.A.: Framing processes and social movements: An overview and assessment. Annual Review of Sociology **26**, 611–639 (2000). https://doi.org/10.1146/annurev.soc.26.1.611
3. Bonner, M., Dammert, L.: Constructing police legitimacy during protests: Frames and consequences for human rights. Policing and Society **32**(5), 629–645 (2022)
4. Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., Jurafsky, D.: Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. Proceedings of the National Academy of Sciences **119**(31), e2120510119 (2022)
5. D'Ottone, S., Varela, M., Castro, D., Carvacho, H.: From war to crime rhetoric: The evolution in the presidential framing of the 2019 chilean social uprising. Journal of Language and Politics **24**(3), 505–527 (2025)
6. Eisele, O., Heidenreich, T., Litvyak, O., Boomgaarden, H.: Capturing a news frame – comparing machine-learning approaches to frame analysis with different degrees of supervision. Communication Methods and Measures **17**(3), 205–226 (2023)
7. Entman, R.M.: Framing: Toward clarification of a fractured paradigm. Journal of Communication **43**(4), 51–58 (1993). https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
8. Ghafouri, V., McNeil, R., Yankov, T., Sumption, M., Rocher, L., Hale, S.A., Mahdi, A.: Framing migration: A computational analysis of uk parliamentary discourse. arXiv preprint arXiv:2509.14197 (2025)
9. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30), e2305016120 (2023). https://doi.org/10.1073/pnas.2305016120
10. González-Bustamante, B.: Benchmarking llms in political content text-annotation: Proof-of-concept with toxicity and incivility data. arXiv preprint arXiv:2409.09741 (2024)
11. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al.: A survey on llm-as-a-judge. arXiv preprint arXiv:2411.15594 (2024)
12. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205 (2020)
13. Kaushik, D., Hovy, E., Lipton, Z.C.: Learning the difference that makes a difference with counterfactually-augmented data. In: International Conference on Learning Representations (2020)
14. Keith, B.: Llm-as-a-judge approaches as proxies for mathematical coherence in narrative extraction. Electronics **14**(13), 2735 (2025)
15. Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., et al.: Prometheus: Inducing fine-grained evaluation capability in language models. In: International Conference on Learning Representations (2024)
16. Kostikova, A., Beese, D., Lieberum, B., Sterner, A., Lange, L., Pado, S., Klinger, R.: Fine-grained detection of solidarity for women and migrants in 155 years of german parliamentary debates. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5884–5097 (2024)
17. Le Mens, G., Gallego, A.: Positioning political texts with large language models by asking and averaging. Political Analysis **33**(3), 274–282 (2025)
18. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522 (2023)
19. Matthes, J.: Framing politics: An integrative approach. American Behavioral Scientist **56**(3), 247–259 (2012). https://doi.org/10.1177/0002764211426324

20. Molina, I., Morales, J., Keith, B.: Web scraping chilean news media: A dataset for analyzing social unrest coverage (2019–2023). Data **10**(11), 174 (2025)
21. Otmakhova, J., Frermann, L.: Narrative media framing in political discourse. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 9167–9196 (2025)
22. Otmakhova, J., Khanehzar, S., Frermann, L.: Media framing: A typology and survey of computational approaches across disciplines. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15407–15428 (2024)
23. Polletta, F., Ho, M.: Frames and their consequences. The Oxford Handbook of Contextual Political Analysis (2006)
24. Proust, V., Saldaña, M.: Another violent protest? new perspectives to understand protest coverage. Media and Communication **10**(4), 18–29 (2022)
25. Renze, M., Guven, E.: The effect of sampling temperature on problem solving in large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024 (2024)
26. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: Behavioral testing of nlp models with checklist. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4902–4912 (2020)
27. Semetko, H.A., Valkenburg, P.M.: Framing european politics: A content analysis of press and television news. Journal of Communication **50**(2), 93–109 (2000). https://doi.org/10.1111/j.1460-2466.2000.tb02843.x
28. Shen, C., Cheng, L., Nguyen, X.P., You, Y., Bing, L.: Large language models are not yet human-level evaluators for abstractive summarization. In: Findings of EMNLP (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.278
29. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability. Psychological bulletin **86**(2), 420–428 (1979)
30. Sinelnik, A., Hovy, D.: Narratives at conflict: Computational analysis of news framing in multilingual disinformation campaigns. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop). pp. 131–143 (2024)
31. Snow, D.A., Rochford, E.B., Worden, S.K., Benford, R.D.: Frame alignment processes, micromobilization, and movement participation. American Sociological Review **51**(4), 464–481 (1986). https://doi.org/10.2307/2095581
32. Suter, S., Haim, M., Kessler, S.H.: When politicians talk ai: Framing artificial intelligence in parliamentary debates across four democracies. Policy & Internet (2025). https://doi.org/10.1002/poi3.70010
33. Thapa, S., et al.: Large language models in computational social science: Prospects and challenges. Social Network Analysis and Mining **15**(1) (2025)
34. Törnberg, P.: Large language models outperform expert coders and supervised classifiers at annotating political social media messages. Social Science Computer Review **43**(6), 1181–1195 (2025)
35. Wan, M., Safavi, T., Jauhar, S.K., Kim, Y., Counts, S., Neville, J., Suri, S., Shah, C., White, R.W., et al.: Tnt-llm: Text mining at scale with large language models. arXiv preprint arXiv:2403.12173 (2024)
36. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
37. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? Computational Linguistics **50**(1), 237–291 (2024)