Vol. 3 No. 4 (2026): Quartus
Review Article
https://doi.org/10.69760/jales.2026004007

Large Language Models Explain Political Rhetoric Unevenly: A Structured Critical Review and Operational ACIL Protocol Across Languages

Hasan Alisoy
Nakhchivan State University, Nakhchivan, Azerbaijan

Abstract

Large language models (LLMs) are increasingly asked to identify and explain political metaphor, framing, epistemic stance, presupposition, implicature, and claim–evidence relations. Fluent output, however, is not evidence that an interpretation is correct, appropriately uncertain, grounded in the text, behaviorally supported, or culturally valid. This structured critical review maps peer-reviewed evidence published through 20 August 2026 and introduces an operational ACIL protocol integrating Accuracy, Calibration, Interpretability, and Language parity. Searches of ACL Anthology metadata and OpenAlex produced 972 unique title records after date filtering and deduplication; 15 reports were retained from database screening and five were added through citation chaining. The final evidence map comprises 18 empirical and two conceptual or methodological reports. Direct political-rhetoric evidence is concentrated in Italian, German, English, and Korean, while multilingual calibration and faithfulness evidence is largely indirect. Results vary sharply by task: constrained multiple-choice recognition can be strong, whereas open explanation, implicature reconstruction, false-presupposition rejection, and source attribution remain fragile. We clarify ACIL's contribution relative to general multidimensional, multilingual, calibration, and faithfulness frameworks; remove composite scoring; specify claim-level calibration under multiple admissible interpretations; distinguish textual grounding, behavioral sensitivity, and mechanistic faithfulness; and treat robustness as a control layer rather than a fifth dimension. A worked re-analysis of a published Italian example illustrates how to report unavailable components without manufacturing scores. The review concludes that current evidence supports LLMs as auditable assistants for political discourse research, not autonomous interpreters. ACIL remains a proposed protocol requiring preregistered validation across languages, political genres, and expert coding workflows.

Keywords: *large language models; political rhetoric; political discourse analysis; calibration; explanation faithfulness; multilingual NLP; pragmatics; sociolinguistics*

How to cite: Alisoy, H. (2026). Large language models explain political rhetoric unevenly: A structured critical review and operational ACIL protocol across languages. *Journal of Azerbaijan Language and Education Studies, 3*(4), 61–78. https://doi.org/10.69760/jales.2026004007

1. Introduction

Political rhetoric is not merely ornament. Political actors use metaphor, presupposition, implicature, framing, stance, and evidential positioning to define problems, distribute responsibility, normalize assumptions, and invite audiences to infer more than is asserted. These choices are embedded in institutional roles, historical narratives, genres, and relationships among speakers and audiences (van

Dijk, 1997; Charteris-Black, 2011). The topic therefore sits at the intersection of computational linguistics, political communication, sociology of public meaning, and sociolinguistics: language variation and social context are part of the object being interpreted, not nuisance variables to be removed.

Instruction-tuned LLMs make rhetorical analysis appear accessible. A model can label a phrase as metaphorical, name a presumed frame, infer an unstated proposition, and produce a fluent justification in seconds. Yet each operation can fail independently. A system may detect a metaphor but misidentify its source domain; recover an implicit proposition but attribute it to the wrong actor; provide the right label with unjustified certainty; or write a persuasive rationale unsupported by the cited words. Multilingual use adds another failure path because translation can alter ambiguity, politeness, evidentiality, conventional metaphor, and historically specific concepts.

The empirical literature is promising but uneven. Italian work on political implicature and presupposition provides large annotated resources and both constrained and open-ended tests (Paci et al., 2025; Gregori et al., 2026; Paci et al., 2026). German experiments show that models often accommodate false presuppositions in loaded political questions even when they can answer related direct questions (Lachenmaier et al., 2025). English studies examine dehumanizing immigration metaphors and semantic frames (Mendelsohn and Budak, 2025; Otmakhova et al., 2026), while Korean work tests claim–evidence extraction from political editorials (Choi and Kim, 2025). These studies do not yet form a balanced multilingual evidence base, and findings from general metaphor, question answering, emotion classification, or safety annotation cannot be treated as direct demonstrations about political-rhetorical explanation.

This review asks four linked questions. First, what has been demonstrated directly for LLM analysis of political rhetoric? Second, what methodological lessons may cautiously be transferred from adjacent tasks? Third, how should correctness and confidence be evaluated when informed analysts accept more than one interpretation? Fourth, what minimum protocol makes an LLM-generated explanation auditable across languages? We answer by providing a traceable evidence map and by operationalizing ACIL—Accuracy, Calibration, Interpretability, and Language parity—as a multidimensional profile rather than a composite score.

The contribution is deliberately methodological. ACIL does not introduce a new accuracy metric, calibration estimator, or causal explanation technique. Its added value is the political-rhetorical unit of analysis and the requirement to connect established methods in one verifiable workflow: phenomenon-specific gold structures; claim-level uncertainty; textual and behavioral explanation tests; native, translated, and parallel language conditions; and versioned robustness controls. The framework is illustrated with a published example but is not presented as empirically validated.

2. Review Methodology

2.1. Scope, sources, and search dates

We conducted a structured critical review rather than an exhaustive systematic review or meta-analysis. The eligibility window covered 1 January 2017 through 20 August 2026. Searches were executed on 14 October 2026 so that the selection process could be audited during revision; records published after the prespecified evidence cutoff were removed. We searched (i) the ACL Anthology bulk bibliographic export dated 12 October 2026 and (ii) OpenAlex title-and-abstract metadata. We also used backward and forward citation chaining from the central political-rhetoric papers and from sources in the submitted manuscript.

The ACL export was searched at title level with four query families. Each required 'LLM,' 'large language model,' or 'language model,' combined respectively with: (Q1) politic*, rhetoric*, immigration, ideolog*, framing, or propaganda; (Q2) metaphor, figurative, implicature, presupposition, hyperbole, proverb, stance, or implicit content; (Q3) calibrat*, confidence, uncertaint*, faithful*, explanation, interpretability, or rationale; and (Q4) multilingual, cross-lingual, low-resource, across languages, or dialect, together with evaluat*, benchmark, understand*, explain*, calibrat*, safety, or reason*. The four queries returned 843 raw hits, 812 unique bibliographic keys, and 811 unique normalized titles.

The exact OpenAlex title-and-abstract query was 'large language model' AND ('political discourse' OR 'political rhetoric'), with dates 1 January 2017 through 14 October 2026. It returned 198 records. Eleven post-cutoff records were removed; the remaining 187 records represented 168 unique normalized titles. Exact-title and DOI deduplication across sources removed seven overlaps, producing 972 unique title records for screening. Full query strings, counts, and coding fields are reproduced in Supplementary Sections S1–S3.

Table 1. Identification and selection of the structured critical-review evidence base

Stage	Records	Decision rule or result
ACL Anthology title queries	843 raw; 812 keys; 811 titles	Four query families; years 2017–2026; retained reports had to meet the 20 August 2026 cutoff.
OpenAlex title/abstract query	198 raw	Eleven post-cutoff records removed; 187 date-eligible records; 168 unique normalized titles.
Cross-source deduplication	972 unique titles	Exact DOI and normalized-title matching; seven cross-source overlaps removed.
Title/abstract screening	15 retained; 957 excluded	Excluded when no assessable link to political rhetoric/implicit language, LLM evaluation, explanation/calibration/faithfulness, or multilingual validity.
Citation chaining	+5 reports	Added operational or directly relevant studies missed by title-constrained searches.
Full-text assessment	20 retained	Eighteen empirical reports and two conceptual/methodological reports; no full-text exclusion after relevance confirmation.

2.2. Eligibility, screening, and source status

Eligible empirical reports evaluated an LLM or a directly relevant computational comparator on at least one of the following: political-rhetorical or implicit-language analysis; figurative or pragmatic interpretation; explanation grounding or faithfulness; calibration or selective prediction; or multilingual validity relevant to ACIL. Conceptual or methodological reports were retained only when they directly specified political or multilingual evaluation risks. We excluded demonstrations without assessable methods, studies concerned only with unrelated generation quality, product commentary, and records lacking enough information to evaluate their contribution. No language restriction was applied at query stage. Because detailed extraction required an accessible English full text, the mapped reports are English-language publications even when their evaluated corpora are Italian, German, Korean, Danish, Persian, Latvian, Indic, African, or Arabic varieties.

Screening was conducted in a single pass rather than by two independent screeners. Retained reports were rechecked against the eligibility rules during extraction, but there was no formal disagreement-resolution procedure. This choice is reported as a limitation rather than hidden. The review is

consequently reproducible as a structured search and evidence map, but it should not be read as an exhaustive estimate of all relevant publications.

All 18 empirical reports in the evidence map were peer-reviewed conference or workshop papers. The two Zaghouani (2026a, 2026b) papers were coded as conceptual/methodological and were not used as outcome evidence. Three preprints cited in the submitted draft—Bommasani et al. (2021), Doshi-Velez and Kim (2017), and Kadavath et al. (2022)—were removed from the evidentiary argument. Published work is used for calibration, interpretability, and multidimensional evaluation claims (Guo et al., 2017; Jacovi and Goldberg, 2020; Jiang et al., 2021; Liang et al., 2023).

2.3. Data extraction and critical synthesis

For each central report we extracted publication status, language, political versus nonpolitical domain, dataset and sample size, model or configuration, task, evaluation method, principal quantitative result, and study-specific limitation. We also coded evidentiary relation as direct political evidence, adjacent empirical evidence, or conceptual/methodological motivation. The complete matrix appears in Supplementary Section S3; Table 3 below provides a condensed view of the studies that most directly shape the argument.

We did not pool outcomes. Accuracy on a four-choice task, span F1, ROC-AUC, human explanation ratings, calibration error, and perturbation sensitivity estimate different constructs. Contradictory-looking findings were therefore reconciled by checking task format, denominator, prompt, supervision, model, language, and evaluation protocol. For example, high Italian multiple-choice accuracy and much lower open-ended explanation accuracy are not conflicting estimates of one ability: the former supplies candidate answers, whereas the latter requires the model to reconstruct and verbalize implicit content. The synthesis gives priority to direct political evidence, then uses adjacent results to motivate tests rather than to claim demonstrated political performance.

3. What Counts as an Explanation of Political Rhetoric?

3.1. Separate detection, interpretation, and social effect

A request to ‘explain the rhetoric’ bundles at least three estimands. Detection asks whether and where a device occurs. Interpretation asks what meaning, stance, source–target mapping, presupposed content, or communicative function is licensed in context. Audience-effect inference asks whether the rhetoric persuaded, polarized, stigmatized, mobilized, or changed behavior. Text alone can support the first two only under an explicit annotation scheme; the third requires independent audience or behavioral evidence. ACIL evaluates model-generated textual interpretations and must not convert them into claims about hidden speaker intent or causal audience effects.

Political interpretation is also socially situated. Actor identity, institutional role, historical reference, genre, and circulating narratives can legitimately change meaning. A decontextualized actor substitution is therefore not automatically a valid invariance test. The relevant question is whether domain experts judge a specific edit to preserve the intended annotation target. This constraint is especially important across languages, where literal equivalence does not guarantee pragmatic or sociolinguistic equivalence.

Table 2. Political-rhetorical phenomena, annotation units, and minimum explanatory content

Phenomenon	Primary unit/task	A minimally adequate explanation specifies	Boundary or characteristic failure
Metaphor	Lexical/discourse span; source–target mapping	Contextual meaning, source domain, target domain, and relevant span	Detection is not interpretation; a metaphor label does not establish audience effect.
Hyperbole	Scalar expression and realistic baseline	Degree of extremity, baseline, and pragmatic purpose	Intensity may be literal, ironic, or metaphorical; political evidence is currently thin.
Framing	Pattern across clauses or documents	Problem definition, attributed cause, evaluation, remedy, and recurring choices	Ideology should not be inferred from an isolated word; audience effects require separate data.
Epistemic stance	Source–event–scope tuple	Claim, cue, belief holder, polarity, modality, and scope	Common failures include wrong source, negation, or attribution.
Implicature/presupposition	Trigger plus licensed implicit proposition	Trigger, inferred content, speaker attribution, context, and defeasibility	One plausible inference must not be presented as entailed or uniquely correct.
Claim–evidence relation	Claim span, evidence span, relation	Exact spans, support/contest relation, and missing evidence	A fluent summary may introduce evidence absent from the document.

3.2. Annotation and contestability

The gold object should not be a single prose paragraph. Each item should specify a target span, necessary context, phenomenon label, atomic interpretive claims, evidence requirements, and an admissible set of alternatives. Metaphor annotation may use MIP or MIPVU to distinguish contextual from basic meaning (Pragglejaz Group, 2007; Steen et al., 2010). Stance and implicit-content codebooks should identify source, cue, scope, and degree of commitment. For contested items, experts should record both substantive disagreement among informed readings and epistemic insufficiency caused by missing context. These are different states: more context may resolve the latter but not the former.

A distribution over expert judgments can be retained rather than forcing adjudication in every case. Strict evaluation can require membership in a predeclared admissible set; graded evaluation can score an atomic claim as fully supported, partly supported, unsupported, or contradicted. Both should be reported in sensitivity analysis. The model should be allowed to abstain when evidence is insufficient, but an abstention is not automatically correct merely because experts disagree.

4. Direct Evidence from Political Discourse

4.1. Implicature, presupposition, and conversational grounding

The strongest direct evidence concerns Italian political implicature and presupposition. IMPAQTS-PID contains 31,822 annotated instances: 14,932 implicatures and 16,890 presuppositions drawn from political speeches. In a four-option task, GPT-4o-mini reached 70% accuracy versus a bootstrapped human ceiling of 91%, while Llama 3.2-3B reached 43%. Open-ended evaluation was more demanding: GPT-4o-mini produced fully correct explanations in approximately 21%–27% of 150 outputs per prompting condition, with the highest value under chain-of-thought prompting; each output was rated by one of ten expert linguists, so output-level inter-annotator agreement was unavailable

(Paci et al., 2025). These denominators and conditions matter: ‘acceptable’ scores that include partial answers should not be reported as fully correct interpretation.

Subsequent studies show that supervision and task format can improve constrained performance without eliminating the explanation problem. On 1,591 multimodal Italian items, the best text-only system, Qwen2.5-72B, reached 0.863 multiple-choice accuracy; GPT-4o-mini reached 0.850. Audio-capable variants were substantially lower—0.554 and 0.589—and missed approximately 3.3% of answers, so adding an audio channel did not improve this setup (Gregori et al., 2026). In a separate explanation-generation study, instruction tuning on 25,000 IMPAQTS-PID pairs was evaluated on a 6,822-item test split. Three annotators scored 50 outputs per fine-tuned model; Krippendorff's α was 0.78 for Llama 3.1-8B and 0.76 for Qwen2.5-7B. Fully correct explanation rates were 0.46 and 0.54, compared with a reconstructed 0.22 zero-shot GPT-4o-mini baseline. Presupposition accuracy was much higher than implicature accuracy (0.73/0.77 versus 0.17/0.29), and BERTScore correlated only moderately with human scores (Kendall's $\tau = 0.40$) (Paci et al., 2026). The baseline is historically reconstructed, so it is informative but not a perfectly controlled head-to-head comparison.

German loaded-question experiments reveal a related grounding failure. Lachenmaier et al. (2025) created 882 loaded political questions and 147 confirmatory plus 147 disconfirmatory direct questions, sampling three responses per model. Seven annotators achieved Fleiss' $\kappa = 0.82$ and mean pairwise Cohen's $\kappa = 0.72$ on loaded-response categories. GPT accommodated the false presupposition in 41.4% of loaded responses and rejected it in 38.1%; Mistral accommodated 64.1% and rejected 10.4%. Direct-question accuracy was higher—76.5% for GPT and 61.6% for Mistral—but fell on disconfirmatory questions. Thus, possessing relevant factual knowledge did not ensure active rejection of a misleading premise.

4.2. Metaphor, framing, stance, and claim–evidence relations

English political-metaphor work demonstrates useful scale but also annotation and construct limitations. Mendelsohn and Budak (2025) analyzed 400,000 U.S. immigration tweets and evaluated seven dehumanizing source domains on 1,600 annotated tweets. Crowd agreement was low (Krippendorff's $\alpha = 0.32$), and the best reported ROC-AUC was 0.746 for a GPT-4-Turbo descriptive representation combined with a sentence-embedding classifier at one decision threshold. The study supports assisted detection under explicit categories; it does not show that free-form model explanations are faithful or that the metaphors caused dehumanization in audiences.

Otmakhova et al. (2026) decomposed metaphorical framing into semantic-frame prediction, source-domain classification, and metaphor detection. Fine-tuned components outperformed zero-shot Gemini and Claude on several component tests; for example, source-domain macro-F1 on a 2,509-item test set was 0.756 for the best fine-tuned system, compared with 0.345 and 0.445 for the two zero-shot LLMs. In the political analysis, filtering yielded 544 ANIMAL, 3,695 WAR, and 1,402 WATER cases from the immigration corpus. The authors also observed that end-to-end LLM analysis could produce frames for inputs that did not instantiate the target domain. This is direct evidence for the value of verifiable intermediate representations, not evidence that one architecture universally explains framing.

Korean claim–evidence extraction provides another direct but distinct task. Choi and Kim (2025) analyzed 87 Rodong Sinmun editorials using four LLMs, an LLM-based evaluator, and human review of approximately 30% of outputs. Automated–human Spearman correlations ranged from 0.55 to 0.78 across quality dimensions and prompting conditions. The design shows that structured extraction can

scale, while also showing why a single model judge should not be treated as independent ground truth. The corpus represents one state-media genre and political system.

Epistemic stance has a longer computational tradition. Gupta et al. (2022) trained a RoBERTa classifier on expert-annotated FactBank news and applied source–event stance analysis to political manifestos; macro-F1 was 77.6 across six labels and 73.6 excluding the dominant non-epistemic class. This is not generative-LLM explanation evidence. It is a useful operational comparator because it explicitly represents belief holder, event, polarity, and modality—the structure that free-form explanations often omit.

Table 3. Condensed evidence map: direct political studies and the nearest adjacent controls

Study	Relation; language/domain	Data and evaluation	Principal result under reported conditions	Limitation for this review
Paci et al. (2025)	Direct; Italian political speeches	31,822 items; 4-choice accuracy; 150 open outputs/condition with expert rating	Best MC accuracy 70% vs 91% ceiling; fully correct open explanations 21%–27%	One rater per open output; no calibration or cross-language test
Gregori et al. (2026)	Direct; Italian political speech/audio	1,591 items; text and speech models; MC accuracy	Best text model 0.863; two audio variants 0.554/0.589	Constrained choice, not explanation generation
Paci et al. (2026)	Direct; Italian political explanation	25,000 train/6,822 test; 50 rated outputs/model; 3 annotators	Fully correct 0.54 Qwen, 0.46 Llama; α 0.76–0.78; implicature much lower	Two fine-tuned models; baseline reconstructed
Lachenmaier et al. (2025)	Direct; German political questions	882 loaded + 294 direct; 3 responses/model; 7 annotators	False premise accommodated 41.4% GPT, 64.1% Mistral	Grounding, not full rhetorical explanation; German party context
Mendelsohn & Budak (2025)	Direct; English U.S. immigration	400k corpus; 1,600 evaluation tweets; human labels	Best ROC-AUC 0.746 in one hybrid setup; α 0.32	Detection/classification; unstable human construct
Otmakhova et al. (2026)	Direct; English U.S. immigration	Component tests plus frame analysis	Fine-tuned source-domain F1 0.756 vs zero-shot 0.345/0.445	No claim-level confidence; English and one issue area
Choi & Kim (2025)	Direct; Korean political editorials	87 editorials; four LLMs; model judge plus human subset	Automated–human correlations 0.55–0.78	Claim–evidence extraction; one media ecosystem
Pedersen et al. (2025)	Adjacent; Danish metaphor explanation	150 items × 2 models × 4 conditions = 1,200 outputs	Culture-specific items lagged English-shared metaphors; context helped	Nonpolitical pilot; explanation ratings not calibration
Mehrazar et al. (2026)	Adjacent; Persian emotion explanation	300 evaluation + 210 calibration items; 4 LLMs; 25 native annotators	Temperature scaling reduced ECE; model–human feature agreement remained low	Emotion classification, not political rhetoric
Azime et al. (2025)	Adjacent; five low-resource languages/proverbs	Three tasks; multiple LLMs	Answer-option order shifted some results by up to 50 points	Cultural language understanding, not political explanation

4.3. What the direct evidence does—and does not—support

Direct studies support three cautious conclusions. First, constrained recognition can be substantially stronger than unconstrained explanation. Second, pragmatic phenomena are not interchangeable: presuppositions with lexical triggers are often easier than context-dependent implicatures. Third, fluent output can coexist with actor misattribution, invented evidence, premise accommodation, and unstable evaluation. None of the direct studies jointly measures phenomenon-specific accuracy, calibrated

claim confidence, explanation grounding, behavioral sensitivity, and cross-language parity in the same political design.

The evidence is also uneven by phenomenon. Metaphor detection and implicit content have multiple datasets; framing evidence is emerging but concentrated in English immigration discourse; claim–evidence extraction has one Korean political case in this map; and political hyperbole explanation lacks a direct multilingual evaluation. We therefore do not generalize findings from metaphor to hyperbole, from Italian pragmatics to all languages, or from textual analysis to audience response.

5. Adjacent Evidence and the Boundary of Extrapolation

5.1. Figurative and culture-sensitive language

Adjacent metaphor studies show why fluent explanation needs independent checks. Across five English figurative-language datasets comprising 4,315 pairs, performance sometimes approached the high 80s, but lexical overlap and length cues predicted success, indicating that surface features could substitute for deeper interpretation (Sanchez-Bayona and Agerri, 2025). MIPVU-inspired prompting likewise over-labeled literal items: on two corpora, prompted LLM F1 remained around 43, below supervised baselines around 64–72 (Reimann and Scheffler, 2025). Zheng et al. (2025) reported strong supervised hyperbole and metaphor detection across four datasets, but detection scores do not establish the adequacy of generated explanations.

Danish evidence is especially relevant to language parity. Pedersen et al. (2025) collected 1,200 explanations from two models across 150 metaphors or idioms and four prompt/context conditions. Shared-with-English items were generally easier than culture-specific items, context improved performance, and expert agreement improved substantially after codebook refinement. This motivates native-language experts and codebook piloting. It does not demonstrate failure on political rhetoric, because the items were not a political corpus.

He et al. (2026) trained on 3,444 synthetic metaphorical riddles and reported transfer to several out-of-domain reasoning benchmarks. The result supports the hypothesis that structured metaphor reasoning may exercise broader abstraction. It does not show that the trained systems produce more accurate, calibrated, or faithful explanations of political rhetoric. ACIL therefore treats it as motivation for a test, not as direct outcome evidence.

5.2. Multilingual evaluation, translation, and faithfulness

Broad multilingual benchmarks establish unevenness but not the mechanism of political-rhetorical error. MEGAVERSE spans 83 languages and 22 datasets, with larger models generally stronger but persistent low-resource variation and contamination concerns (Ahuja et al., 2024). IndicQA shows that translate-test routing can outperform direct multilingual answering for some low-resource Indic languages, but not uniformly (Singh et al., 2025). ProverbEval covers five African languages and three tasks; answer-option order changed some model results by as much as 50 percentage points, illustrating that presentation sensitivity can rival nominal language differences (Azime et al., 2025). These studies motivate matched translation and order controls; they do not estimate political-rhetorical explanation quality.

Latvian benchmarking demonstrates that post-editing can change the measured task rather than merely polish wording. On the post-edited MMLU sociology subset, ChatGPT-3.5-Turbo increased from 78.79% to 81.82% and Gemini 1.0 Pro from 81.82% to 90.90%; a separate COPA comparison used 260 post-edited evaluation items (Skadina et al., 2025). Translation strategy, corpus construction, and

model therefore interact. A translated condition should be reported as a condition, not silently substituted for native data.

Persian emotion classification provides the nearest multilingual joint calibration–faithfulness study, although it is not political. Mehrazar et al. (2026) used 300 evaluation items, a separate 210-item temperature-scaling set, four LLMs, and 25 native annotators. GPT-4o reached 77% accuracy in predict-then-explain prompting; temperature scaling reduced ECE, but model-generated top-five influential words agreed more with other models than with humans, and perturbation measures did not establish mechanistic causality. This is precisely the kind of adjacent evidence that can inform ACIL while remaining visibly labeled as extrapolation.

Finally, Zaghouani's (2026a, 2026b) political-cultural and dialect-safety papers specify annotation, ontology, dialect, and governance gaps. They are conceptual and methodological contributions, not benchmark results. Their role here is to motivate local category validation and dialect coverage, not to supply performance estimates.

6. Synthesis: Why Fluent Explanations Remain Insufficient

Across direct and adjacent evidence, fluency is a weak proxy because it is compatible with several distinct errors. A model can select a correct label by exploiting answer options, then generate an incorrect actor or causal story. It can assign high probability to a majority label when experts recognize multiple readings. It can cite words present in the text without depending on them behaviorally. It can be stable across prompt paraphrases because it repeats the same mistake. It can translate a proposition correctly while erasing the rhetorical form under study.

The practical implication is not that all LLM-assisted political analysis is invalid. Structured systems can retrieve candidate spans, propose annotations, and scale exploratory coding. Reliability improves when tasks expose intermediate decisions and compare them with expert codebooks. The implication is that each relevant claim must be tied to the kind of evidence that could falsify it. This principle motivates ACIL's conjunctive profile.

7. ACIL: Scope, Distinctiveness, and Operational Definitions

7.1. Unit of analysis and profile

ACIL's unit is an item containing: (1) a rhetorical target span and necessary context; (2) phenomenon and task definition; (3) expert annotation with admissible alternatives; (4) a model decision; (5) confidence attached to the decision and to atomic interpretive claims; (6) cited evidence; (7) a concise explanation; and (8) language, variety, genre, prompt, model, and version metadata. Results are reported as a profile. No composite, minimum component, or normalized total is recommended because F1, Brier score, human ratings, and perturbation effects have incompatible scales and directions. Any decision threshold must be tied to a prespecified use and development set, not inferred from the evaluation sample.

Robustness is not a fifth ACIL dimension. It is a control layer applied to Accuracy, Calibration, and Interpretability within each language condition. Likewise, Language parity is not a downstream fairness check: every accuracy, calibration, and explanation result is disaggregated and compared across native, translated, parallel, dialect, register, or genre conditions where relevant.

Table 4. What ACIL adds relative to neighboring evaluation approaches

Approach	Primary contribution	Gap for multilingual political-rhetorical explanation	ACIL integration
HELM / holistic evaluation	Scenario-based multidimensional reporting, calibration, robustness, fairness, transparency	No political-rhetorical gold structure or claim-level explanation protocol	Retains multidimensional reporting; adds rhetorical units, evidence claims, and expert admissibility.
XTREME / MEGAVERSE	Cross-lingual and multilingual task comparison	Mostly task accuracy; limited explanation faithfulness and political-pragmatic validity	Makes language a condition inside every A/C/I analysis; adds native/translated/parallel designs.
Calibration literature	Proper scoring rules, ECE diagnostics, selective prediction	Usually one target label and limited treatment of contestable interpretations	Attaches confidence to labels and atomic claims; supports admissible sets and judgment distributions.
Faithfulness / interpretability literature	Grounding, perturbations, rationale and causal distinctions	Often monolingual and nonpolitical; actor/context edits may be invalid	Requires controlled edits and bilingual domain validation; separates textual, behavioral, and mechanistic claims.
Political NLP task studies	Ecologically relevant corpora and phenomenon-specific codebooks	Accuracy, confidence, explanation, and language parity rarely measured together	Provides a common reporting workflow without pooling incompatible outcomes.

7.2. Accuracy and calibration

Accuracy must match the estimand. Detection uses span precision, recall, and F1; classification uses macro-F1, balanced accuracy, and confusion matrices; structured interpretation scores source, scope, mapping, function, and evidence separately. Explanation adequacy should be judged at claim level before evaluators see the model's full prose, reducing fluency bias. Strict and graded scores should be reported separately, with results by phenomenon.

Calibration requires a defined target. ACIL attaches one confidence value to the categorical label and separate confidence values to each atomic interpretive claim. A single confidence number for an entire paragraph is insufficient because some claims may be supported and others speculative. If experts predeclare several acceptable readings, categorical correctness can be evaluated against an admissible set and claim correctness against full/partial/unsupported/contradicted levels. A distribution of expert judgments may be retained as a soft reference; analyses should also distinguish informed substantive disagreement from missing-context uncertainty.

At least one proper scoring rule—multiclass Brier score or log loss—should accompany reliability plots. ECE may be reported only with bin edges, sample size, and confidence intervals or resampling uncertainty. Any temperature, verbal-to-numeric mapping, or abstention threshold must be fitted on a development/calibration split or nested cross-validation that is separate from final evaluation. Verbal terms such as ‘likely’ or their translations must be calibrated empirically within each language; semantic equivalence should not be assumed.

Selective prediction should report both error and retained coverage by language and phenomenon. An apparently safer system can achieve lower retained error by sending a disproportionate share of low-resource-language items to human review. Therefore, risk–coverage curves, coverage at prespecified risk, and referral-rate disparities are part of the language-parity analysis.

7.3. Interpretability: three levels of claim

Textual grounding asks whether cited spans exist, are attributed to the correct source, and support the claimed interpretation under the codebook. Behavioral sensitivity asks whether controlled removal or modification of relevant information changes the model in the expected direction relative to matched control edits. Mechanistic faithfulness asks whether the rationale reflects the model's internal causal computation. Black-box prompt tests can support the first two; they cannot establish the third. Mechanistic claims require internal access and causal interventions appropriate to the architecture (Jacovi and Goldberg, 2020; Turpin et al., 2023).

Rationale deletion must use controls for length, grammaticality, topic, and information density. Otherwise a performance drop may reflect corrupted input rather than dependence on the cited rationale. A valid invariance test might replace two fictional actor names in a constructed, historically neutral pair after bilingual domain experts confirm that identity is irrelevant. A meaning-changing test might replace the named party in a historically situated quotation where that identity changes the licensed implicature; the expected output should then change. Both intervention sets must be validated before model evaluation, and unexpected changes must be interpreted against control-edit distributions rather than in isolation.

Plausibility and stability remain useful but distinct. Expert-rated coherence does not establish grounding; repeated explanations may be consistently wrong. ACIL therefore reports evidence attribution, behavioral sensitivity, prompt stability, and—only when technically justified—mechanistic evidence as separate fields.

7.4. Language parity and robustness

Three language designs answer different questions. Native evaluation samples naturally occurring political discourse and maximizes ecological validity. Translated evaluation holds propositional content more constant but may alter the rhetoric. Parallel evaluation uses meaning-aligned items authored or validated in each language and can preserve language-specific form, although it may still impose the source language's ontology. A strong design triangulates rather than treats one condition as universally sufficient.

Comparison requires matched difficulty, annotation uncertainty, and class prevalence where feasible. Each language should use native prompts and outputs, bilingual expert review, and separate reporting by variety, register, and genre. Robustness checks include prompt paraphrase, label and answer-option order, repeated sampling, decoding settings, context position, model identifier, system prompt, access date, and archived raw output. Hosted-model results are versioned observations, not timeless properties.

Figure 1. ACIL treats language parity as cross-cutting and robustness/version control as a control layer across Accuracy, Calibration, and Interpretability. The output is a profile, not a composite score.

[Figure 1 — insert ACIL diagram image here]

8. Worked Illustration Using a Published Political Example

To demonstrate application without presenting a new experiment, we re-code one qualitative example reported by Paci et al. (2026). The Italian speaker contrasts personal fulfillment of an electoral promise with what Italians had come to regard as impossible. The published expert interpretation treats the implicit contrast as prior governments failing to keep promises while the speaker succeeds. The published Qwen output was: 'M5S kept the promise.' The model therefore introduced a specific actor not licensed by the item and failed to reconstruct the intended contrast. This quotation is used only as a protocol illustration; no model was rerun and no new performance estimate is derived.

Table 5. Worked ACIL profile for one published Italian example (Paci et al., 2026)

Field	Required representation	Published evidence available	ACIL judgment for this illustration
Annotation unit	Target utterance + preceding context; self-praise/implicit contrast	Source item and expert IMPAQTS interpretation	Available; item-level analysis only
Admissible readings	Prior governments/other politicians failed to keep promises; speaker claims contrastive success	One expert reference is published	Use as provisional admissible reading; alternatives were not elicited
Accuracy	Actor, contrast, implicit proposition, and attribution	Qwen names M5S and omits the prior-government contrast	Fail on actor grounding and core implicit relation
Calibration	Confidence for label and each interpretive claim	No confidence reported for this output	NA—not measured; do not infer from wording
Textual grounding	Cited span must support actor and contrast	No evidence span; introduced actor is unsupported	Fail for actor grounding; other grounding fields unavailable
Behavioral sensitivity	Matched deletion/counterfactual tests plus controls	No intervention results for this item	NA—not tested
Mechanistic faithfulness	Causal internal evidence	No internal intervention	Unsupported; no mechanistic claim
Language parity	Matched results in another language/condition	Italian only	NA—not testable from this item
Robustness/versioning	Prompt variants, repeated samples, model/date/output archive	Not reported at item level in the example	NA; retain as reporting gap

The illustration shows why missing evidence must remain visible. The failed accuracy judgment is supported by the published text–reference–output relation. Calibration, cross-language parity, and behavioral or mechanistic faithfulness cannot be back-filled from fluent prose. ACIL does not punish a study for every unavailable field; it prevents unavailable evidence from silently becoming a positive claim.

9. Minimum Reproducible ACIL Protocol

The protocol below translates ACIL into a sequence of inputs, operations, retained artifacts, and compliance criteria. 'Mandatory' applies whenever a study claims to evaluate LLM explanations of political rhetoric. 'Conditional' applies when the corresponding claim or design is present—for example, mechanistic faithfulness requires model-internal access, and cross-language parity requires at least two language conditions. 'Recommended' improves inference but may be infeasible for sensitive corpora. A study should publish a completed checklist stating not only what it considered, but what it did, retained, and omitted.

Table 6. Operational ACIL workflow and evidence of compliance

Step	Input	Operation	Evidence retained / sufficient compliance	Level
1. Define estimand	Phenomena, unit, intended use	Separate detection, interpretation, and effect claims	Preregistered task definition; prohibited inferences stated	Mandatory
2. Build items	Native corpus + context	Sample genres/actors; create literal controls and, where relevant, parallel/translated sets	Item IDs, provenance, context windows, translation/post-edit log	Mandatory; parallel/translated conditional
3. Annotate	Codebook and expert pool	Independent annotation; record admissible readings, insufficiency, and disagreement	Codebook, annotator qualifications, agreement, adjudication and unresolved distributions	Mandatory
4. Freeze model	Model/API access	Record model ID, date, prompts, schema, decoding, runs, context placement	Version manifest and exact prompts	Mandatory
5. Elicit output	Structured prompt	Request label, claim-level confidence, evidence span, concise mechanism/function, abstention	Raw machine-readable outputs and parser errors	Mandatory
6. Blind scoring	Outputs + gold structure	Score label/claims/evidence before exposing full prose; retain phenomenon strata	Item-level scores, rubric, error codes, blinded-rating procedure	Mandatory
7. Calibrate	Separate development/calibration split	Fit numeric/verbal mappings and abstention threshold outside test set	Split IDs, mapping, Brier/log loss, reliability and risk–coverage by language	Mandatory if confidence or selective claims
8. Test explanation	Validated edit sets	Assess grounding; run relevance deletions against grammar/length/topic controls; distinguish behavioral from mechanistic tests	Edit pairs, expert validity labels, control distributions, causal method if claimed	Grounding mandatory; interventions conditional
9. Test parity	Native/parallel/translated conditions	Compare matched performance, confidence, grounding, coverage, dialect/register	Disaggregated estimates, uncertainty, referral disparities	Mandatory for multilingual claims
10. Stress and release	Prompts, models, ethical constraints	Paraphrase/order/repeat/version tests; archive permitted artifacts	Code, anonymized outputs, checksums, exclusions, redaction rationale, limitation statement	Robustness mandatory; public release recommended where lawful

9.1. Analysis and reporting rules

1. Report denominators and evaluation conditions with every headline number; never mix multiple-choice, open generation, detection, and explanation ratings into a single performance claim.
2. Provide item-level or bootstrap uncertainty where licensing permits, and retain results by rhetorical phenomenon, language, genre, and model version.
3. For contested golds, report strict admissible-set scoring, graded claim scoring, and a sensitivity analysis using the expert judgment distribution.
4. For confidence, report the target of confidence, the independent calibration procedure, at least one proper score, a reliability display, and risk plus retained coverage by language.
5. For explanations, state whether the evidence supports textual grounding, behavioral sensitivity, or mechanistic faithfulness; do not use the terms interchangeably.
6. Register missing components as NA or not tested. Do not impute a favorable result from fluency, model size, or performance on an adjacent task.

10. Implications for Research and Practice

10.1. Computational and political-linguistic research

Benchmarks should move from a label plus free-form rationale to a verifiable item structure: target span, context, rhetorical category, source and scope, admissible interpretations, evidence rule, and controlled variants. Hybrid workflows can use retrieval for historical claims, supervised detectors for spans, and constrained generation for explanation. This does not eliminate expert judgment; it directs it to the points where the model can be wrong in politically consequential ways.

Political discourse research should preserve the distinction between textual coding and social explanation. LLMs can assist corpus navigation, generate candidate codes, and surface uncertain cases. They cannot establish a speaker's private intention or an audience's response from wording alone. Where sociological claims concern legitimacy, persuasion, polarization, or group effects, rhetorical coding must be linked to independent institutional, survey, experimental, ethnographic, or behavioral evidence.

10.2. Multilingual and low-resource research

Low-resource evaluation should begin with native political genres and local annotation categories, not with an English benchmark translated at the end. Native, translated, and parallel conditions should be analyzed as distinct interventions. Bilingual experts should validate category equivalence, edit validity, and cultural references. Dialect, code-switching, honorifics, evidentiality, and register deserve explicit sampling because a model may appear competent in a standard variety while failing on politically salient community speech.

Azerbaijani and other Turkic and South Caucasus languages remain underrepresented in the mapped evidence. A valuable validation study would pair native Azerbaijani parliamentary, presidential, and media discourse with carefully authored parallel cases in a higher-resource language; annotate metaphor, stance, presupposition, and framing with bilingual domain experts; and preregister ACIL accuracy, claim-calibration, grounding, coverage, and edit controls. That proposal is a future study, not evidence supplied by this review.

11. Ethical and Governance Considerations

Automated interpretation of political language can affect journalism, moderation, public-opinion research, security analysis, and election monitoring. Errors may stigmatize speakers, erase minority varieties, or make contested inferences appear factual. Dataset and category choices allocate representational power (Blodgett et al., 2020; Bender et al., 2021). Researchers should document who defines categories, which varieties and institutions are represented, and how affected experts can contest an interpretation.

Every reported interpretation should be traceable to input, context, model version, prompt, raw output, and human decision. Sensitive corpora may require redaction or controlled access; the reason for any withheld artifact should be documented. An LLM explanation must not be used as evidence of hidden intent, dangerousness, or audience effect without independent support. Human review is not a generic cure: referral rates, reviewer expertise, workload, and language coverage must themselves be audited.

12. Limitations

This is a structured critical review, not an exhaustive systematic review or meta-analysis. Title-constrained ACL searches may miss relevant papers whose titles omit model terms; OpenAlex

coverage and metadata are incomplete; the English-full-text requirement introduces language bias; and screening was not independently duplicated. Citation chaining partly mitigates, but cannot remove, selection bias. Search counts are reproducible for the archived exports and dates, not guaranteed to reproduce after database updates.

The evidence base remains concentrated. Direct studies in this map cover Italian, German, English, and Korean political material, with no study jointly evaluating all ACIL components across languages. Metaphor and implicit content are better represented than hyperbole, strategic ambiguity, multimodal rhetoric, or audience interpretation. Model versions, prompts, decoding, and annotation schemes differ, so numerical results should not be ranked across studies.

ACIL is an operational proposal, not a validated measurement model. The worked illustration re-codes one published output and demonstrates reporting logic; it does not estimate reliability, predictive validity, or researcher burden. Claim-level confidence elicitation may itself alter generation, verbal mappings may be unstable, and perturbation tests remain behavioral proxies. Mechanistic faithfulness cannot be inferred from black-box behavior. The framework should be tested in preregistered studies against expert-only and alternative computational workflows.

13. Conclusion

The available evidence supports a qualified conclusion: LLM analysis of political rhetoric is uneven across task formats, phenomena, languages, and evaluation designs. Constrained recognition can be strong, especially with structured supervision, while open interpretation, implicature, false-premise rejection, source attribution, and culturally specific explanation remain fragile. Adjacent multilingual and faithfulness studies reveal important risks, but they must not be misreported as direct political evidence.

ACIL converts this evidence into an auditable workflow. Accuracy is phenomenon-specific; confidence attaches to labels and atomic claims; interpretability separates grounding, behavioral sensitivity, and mechanistic evidence; language parity cuts across every component; and robustness/versioning controls every comparison. The framework deliberately produces a profile rather than a composite score. Until that profile is validated across languages and political genres, LLMs are best treated as reviewable assistants that propose interpretations, expose evidence, communicate uncertainty, and defer genuinely contestable cases to qualified analysts.

Declarations

Author Contributions: HA: Conceptualization, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review and editing. IM: Conceptualization, Investigation, Validation, Writing – review and editing. ZS: Investigation, Resources, Validation, Writing – review and editing. AA: Investigation, Validation, Writing – review and editing. All authors critically reviewed the manuscript, approved the submitted version, and accept responsibility for the integrity of the work.

Funding: The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of Interest: The authors declare that the review was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability: No original participant-level dataset was generated. Search strings, selection counts, evidence-coding fields, study-level extractions, and the worked-illustration scoring sheet

are provided in the Supplementary Material. Numerical values were transcribed from the cited published reports.

Ethics Statement: Ethical approval was not required because this article reviews published literature and reports no research involving human participants, personal data, or animals.

Generative AI Statement: During manuscript preparation and revision, the authors used OpenAI Codex (OpenAI; accessed 14 October 2026) to assist with literature organization, document structuring, language editing, and preparation of revision files. The authors are responsible for independently checking source accuracy, approving all interpretations and disclosures, and taking full responsibility for the final content.

References

Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., et al. (2024). MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In Proceedings of NAACL 2024, 2598–2637. https://doi.org/10.18653/v1/2024.naacl-long.143

Azime, I. A., Tonja, A. L., Belay, T. D., Chanie, Y., Balcha, B. F., Abadi, N. H., et al. (2025). ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In Findings of NAACL 2025, 6265–6281. https://doi.org/10.18653/v1/2025.findings-naacl.350

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of FAccT ’21, 610–623. https://doi.org/10.1145/3442188.3445922

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of ‘bias’ in NLP. In Proceedings of ACL 2020, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Charteris-Black, J. (2011). Politicians and Rhetoric: The Persuasive Power of Metaphor (2nd ed.). Palgrave Macmillan.

Choi, G., and Kim, H. (2025). Automated claim–evidence extraction for political discourse analysis: A large language model approach to Rodong Sinmun editorials. In Proceedings of the Eighth Fact Extraction and VERification Workshop, 1–17. https://doi.org/10.18653/v1/2025.fever-1.1

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL 2020, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Desai, S., and Durrett, G. (2020). Calibration of pre-trained transformers. In Proceedings of EMNLP 2020, 295–302. https://doi.org/10.18653/v1/2020.emnlp-main.21

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43, 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Gregori, L., Paci, W., and Panunzi, A. (2026). Evaluating the abilities of LLMs and SpeechLMs in discovering implicit contents of Italian political speeches. In Proceedings of PoliticalNLP 2026, 165–170. https://doi.org/10.63317/4y7ki8iw4mwe

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, 1321–1330.

Gupta, A., Blodgett, S. L., Gross, J. H., and O’Connor, B. (2022). Examining political rhetoric with epistemic stance detection. In Proceedings of NLP+CSS 2022, 89–104. https://doi.org/10.18653/v1/2022.nlpcss-1.11

He, Q., Lu, J., Zhang, Y., Yuan, S., Meng, X., Wei, J., et al. (2026). Metaphor reasoning is meta-reasoning. In Proceedings of ACL 2026, 6914–6935. https://doi.org/10.18653/v1/2026.acl-long.314

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In Proceedings of ICML 2020, 4411–4421.

Jacovi, A., and Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of ACL 2020, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

Jiang, Z., Araki, J., Ding, H., and Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9, 962–977. https://doi.org/10.1162/tacl_a_00407

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of ACL 2020, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In International Conference on Learning Representations.

Lachenmaier, C., Sieker, J., and Zarrieß, S. (2025). Can LLMs ground when they (don't) know: A study on direct and loaded political questions. In Proceedings of ACL 2025, 14956–14975. https://doi.org/10.18653/v1/2025.acl-long.728

Lakoff, G., and Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2023). Holistic evaluation of language models. Transactions on Machine Learning Research.

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61, 36–43. https://doi.org/10.1145/3233231

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., et al. (2023). Faithful chain-of-thought reasoning. In Proceedings of IJCNLP-AACL 2023, 305–329. https://doi.org/10.18653/v1/2023.ijcnlp-main.20

Mehrazar, M., Yousefi, M. A., Beygi, P. A., and Bahrak, B. (2026). Can LLMs faithfully explain themselves in low-resource languages? A case study on emotion detection in Persian. In Proceedings of LREC 2026, 560–579. https://doi.org/10.63317/4tds2kffmg7j

Mendelsohn, J., and Budak, C. (2025). When people are floods: Analyzing dehumanizing metaphors in immigration discourse with large language models. In Proceedings of ACL 2025, 8079–8103. https://doi.org/10.18653/v1/2025.acl-long.398

Otmakhova, Y., Guida, M., and Frermann, L. (2026). Not all ANIMALs are equal: Metaphorical framing through source domains and semantic frames. In Findings of ACL 2026, 38334–38355. https://doi.org/10.18653/v1/2026.findings-acl.1911

Paci, W., Gregori, L., and Panunzi, A. (2026). Teaching LLMs to unveil tendentious implicit contents of Italian political communication. In Proceedings of CMCL 2026, 221–229. https://doi.org/10.63317/4da4mm3sngm7

Paci, W., Panunzi, A., and Pezzelle, S. (2025). They want to pretend not to understand: The limits of current LLMs in interpreting implicit content of political discourse. In Findings of ACL 2025, 15569–15593. https://doi.org/10.18653/v1/2025.findings-acl.804

Pedersen, B. S., Sørensen, N., Nimb, S., Hansen, D. H., Olsen, S., and Al-Laith, A. (2025). Evaluating LLM-generated explanations of metaphors: A culture-sensitive study of Danish. In Proceedings of NoDaLiDa/Baltic-HLT 2025, 470–479.

Pragglejaz Group. (2007). MIP: A method for identifying metaphorically used words in discourse. Metaphor and Symbol, 22, 1–39. https://doi.org/10.1080/10926480709336752

Reimann, S., and Scheffler, T. (2025). Using large language models to perform MIPVU-inspired automatic metaphor detection. In Proceedings of Analogy-Angle II, 10–21. https://doi.org/10.18653/v1/2025.analogyangle-1.2

Sanchez-Bayona, E., and Agerri, R. (2025). Metaphor and large language models: When surface features matter more than deep understanding. In Findings of ACL 2025, 17462–17477. https://doi.org/10.18653/v1/2025.findings-acl.898

Singh, A. K., Kumar, V., Murthy, R., Sen, J., Mittal, A., and Ramakrishnan, G. (2025). Indic QA Benchmark: A multilingual benchmark to evaluate question-answering capability of LLMs for Indic languages. In Findings of NAACL 2025, 2607–2626. https://doi.org/10.18653/v1/2025.findings-naacl.141

Skadina, I., Bakanovs, B., and Darģis, R. (2025). First steps in benchmarking Latvian in large language models. In Proceedings of RESOURCEFUL-2025, 86–95.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., and Pasma, T. (2010). A Method for Linguistic Metaphor Identification: From MIP to MIPVU. John Benjamins. https://doi.org/10.1075/celcr.14

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36, 74952–74965.

van Dijk, T. A. (1997). What is political discourse analysis? Belgian Journal of Linguistics, 11, 11–52. https://doi.org/10.1075/bjl.11.03dij

Wiegreffe, S., and Pinter, Y. (2019). Attention is not not explanation. In Proceedings of EMNLP-IJCNLP 2019, 11–20. https://doi.org/10.18653/v1/D19-1002

Zaghouani, W. (2026a). Cultural adaptation in large language models for political discourse. In Proceedings of PoliticalNLP 2026, 272–283. https://doi.org/10.63317/2prsxx26bfis

Zaghouani, W. (2026b). Lost in dialect: The annotation gap in multilingual LLM safety. In Proceedings of MeLLM 2026, 1–13. https://doi.org/10.18653/v1/2026.mellm-1.1

Zheng, L., Wang, S., Fei, H., Peng, Z., Li, F., Fu, J., et al. (2025). Enhancing hyperbole and metaphor detection with their bidirectional dynamic interaction and emotion knowledge. In Proceedings of ACL 2025, 489–499. https://doi.org/10.18653/v1/2025.acl-long.23

Corresponding Author	**Hasan Alisoy** PhD candidate, lecturer, Nakhchivan State University, Nakhchivan, Azerbaijan Email: alisoyhasan@ndu.edu.az ORCID: https://orcid.org/0009-0007-0247-476X

Received: 05.10.2026
Revised: 11.10.2026
Accepted: 13.11.2026
Published: 16.11.2026