GLM_REVIEW_v7

Peer Review: Governance or Competition? Divergent Frames in AI Policy Discourse Across the US and Global South

Overall Assessment

This manuscript presents a novel comparative analysis of AI policy framing between the United States and three Global South nations (South Africa, Brazil, India). The research question—how nations construct AI as a policy problem—is theoretically significant and understudied. The finding of divergent framing patterns (competition vs. governance) contributes valuable insights to the emerging literature on AI governance and comparative technology politics.

The methodological approach—using LLMs as content coders following CommDAAF framework—is innovative and addresses important methodological challenges in automated content analysis. The authors demonstrate awareness of key validity concerns, including document-type bias and cross-validation between models from different cultural contexts.

However, the manuscript has significant methodological weaknesses that limit the validity and credibility of its findings. These include: (1) fundamental comparability issues between document types that are not adequately addressed; (2) underdeveloped statistical analysis lacking inferential rigor; (3) insufficient detail for replicability; (4) conflation of correlation with causal explanation in discussion sections; and (5) limited consideration of selection bias and sample representativeness.

While the conceptual contribution is promising, the methodological limitations are substantial enough that I cannot recommend publication in its current form. Major revisions are required to address the concerns outlined below.

Major Concerns

1. Document Type Comparability Problem

Issue: The study compares fundamentally different document types—US congressional hearings (testimony transcripts with structured Q&A) versus Global South mixed policy documents (committee reports, written submissions, policy analyses)—without adequate methodological justification or controls for this difference.

Why this matters: Congressional hearings have distinctive rhetorical conventions (witness testimony, adversarial questioning, political posturing) that systematically differ from committee reports, legislative analyses, and written policy submissions. The authors acknowledge this limitation but do not address it methodologically. The observed divergence in framing patterns may reflect document-type effects rather than genuine national differences. For example:

- Hearings may naturally elicit more competitive framing due to their adversarial nature and political theater dynamics
- Committee reports may naturally elicit more governance framing due to their institutional and policy-focused purpose
- Written submissions may differ in their rhetorical structure from

spoken testimony

Required action: The authors must either (a) address this limitation through statistical controls, matched sampling, or robustness checks; (b) restrict analysis to comparable document types; or (c) explicitly model document type as a confounding variable in a multivariate analysis. The current “acknowledgement” of this limitation is insufficient.

2. Statistical Analysis and Inference

Issue: The statistical analysis is severely underdeveloped. The manuscript reports chi-square tests of frame distribution differences but provides no confidence intervals, no multivariate analysis, no adjustment for multiple comparisons, and no discussion of effect size beyond percentage differences.

Specific deficiencies:

- **Chi-square test limitations:** The chi-square tests reported do not account for the nested structure of the data (documents nested within countries, frames potentially correlated within documents). No multilevel modeling or cluster-robust standard errors are used.
- **Sample size disparities:** With N=192 US documents vs. N=102 Global South documents, statistical power differs substantially between comparisons. No power analysis or discussion of Type II error risk is provided.
- **No multivariate analysis:** The authors do not test whether observed differences persist after controlling for potential confounders (document type, year, length, speaker characteristics). The bivariate comparison is insufficient for establishing that national context drives framing patterns independent of other factors.
- **Multiple comparisons:** With 8 frames × 2 contexts = 16 comparisons, no adjustment for multiple testing is reported. The p-values shown ($p < .05$, $p < .01$) may be inflated.
- **Missing confidence intervals:** No confidence intervals are reported for frame proportions or differences. Readers cannot assess precision of estimates or uncertainty around claims.
- **Effect size interpretation:** While percentage differences are reported (e.g., 22.9% difference in governance framing), no formal effect size measures (Cramer’s V for chi-square, odds ratios) are provided to assess practical significance.

Required action: The authors should provide (a) confidence intervals for all frame proportions and differences; (b) multivariate analysis (logit or multinomial logit) controlling for document type, year, and other confounders; (c) multilevel modeling if appropriate given data structure; (d) adjustment for multiple comparisons; (e) formal effect size measures with interpretation; (f) power analysis discussion given sample size disparities.

3. Inter-Rater Reliability and Coding Validity

Issue: While the authors report Cohen’s kappa values, several validity concerns remain unaddressed:

- **Rights frame reliability ($\kappa = .52$):** Below conventional threshold of .60, yet the manuscript repeatedly highlights the rights framing difference (US 9% vs. Global South 18%) as a key finding without sufficient caution about potential coding artifacts. This differential may reflect noise rather than signal.
- **Adjudication transparency:** For South Africa, 15 disagreements were resolved through author adjudication with $\kappa = 1.0$ post-adjudication. This raises concerns about inflated reliability—adjudication should produce a third coder rating, not force agreement to 1.0. The method suggests the authors effectively coded the disagreements themselves rather than using independent adjudication.
- **No human-coded validation set:** The study relies entirely on LLM coding without a human-coded gold standard for validation. While cross-validation between two LLMs is valuable, it does not establish that either model is accurate relative to ground truth.

- **Prompt engineering transparency:** The authors describe prompt revisions addressing “document-type bias” but do not provide the final prompts, making replication impossible. The critical intervention—explicit instructions distinguishing document type from substantive framing—is central to the study’s validity but opaque to readers.

Required action: (a) Provide a human-coded validation subset (minimum 20-30 documents) to establish LLM coding accuracy against ground truth; (b) Use proper adjudication methodology (independent third coder) rather than author-driven resolution that produces κ = 1.0; (c) Provide complete prompts in appendix or supplementary materials; (d) Either recast rights findings as exploratory given reliability concerns or conduct targeted recoding with improved reliability.

4. Replicability and Data Transparency

Issue: The manuscript provides insufficient detail for replicability. Several critical information gaps exist:

- **Search queries:** The specific search terms and API queries used for data collection are not provided. What exactly was searched for in GovInfo? What date ranges? What inclusion/exclusion criteria beyond the AI term density threshold?
- **Document identification:** No list of included documents is provided (e.g., document IDs, titles, URLs). Without this, replication is impossible.
- **LLM specifications:** While model names are provided (Kimi K2.5, Claude Opus 4.5), no information is given about temperature settings, system prompts, version control, or reproducibility parameters. LLMs are stochastic; without these details, results cannot be reproduced.
- **Density threshold details:** The “AI term frequency” threshold used to distinguish substantive hearings from incidental mentions is described but not operationalized. What count? What normalization? What cutoff?
- **Data cleaning:** No description of text preprocessing, cleaning, or preparation for LLM analysis is provided.

Required action: Provide (a) complete search query strings and API parameters; (b) a document registry with IDs and metadata; (c) complete LLM prompts, temperature settings, version information; (d) detailed operationalization of the density threshold; (e) description of text preprocessing pipeline.

5. Selection Bias and Representativeness

Issue: The sampling strategy introduces significant selection bias that is not adequately addressed:

- **India sample (N=7):** This sample is too small for meaningful analysis, yet the manuscript reports country-specific findings for India (Governance 57.1%, Rights 28.6%) without appropriate caveats. These percentages are based on 4 and 2 documents respectively—insufficient for generalization.
- **Document source bias:** US data comes from GovInfo (official government records) while Global South data comes from PMG, Chamber of Deputies, and unspecified Indian sources. These sources have different selection criteria, accessibility, and completeness.
- **Temporal bias:** US data is 90% post-ChatGPT (Nov 2022), but no comparable temporal distribution is reported for Global South data. If Global South data predates ChatGPT, observed differences may reflect temporal rather than national factors.
- **Language bias:** The manuscript does not specify language of documents. Were Brazilian documents analyzed in Portuguese or translated to English? If translated, how? Translation can introduce framing artifacts.

Required action: (a) Explicitly exclude or flag India findings as exploratory given inadequate sample; (b) Provide temporal distribution comparison across datasets; (c) Discuss and address source bias; (d) Specify language(s) of original documents and translation methodology if applicable; (e) Consider sensitivity analysis excluding India to assess robustness.

6. Causal Inference Limitations

Issue: The Discussion section engages in causal explanation ("Why do the US and Global South frame AI so differently?") despite acknowledging that "causal claims require longitudinal analysis beyond this study's scope." The manuscript offers speculative explanations (geopolitical positioning, developmental priorities, adopter vs. developer status) without empirical testing or alternative explanation consideration.

Why this matters: Cross-sectional observational data cannot support causal inference. The observed correlation between national context and framing patterns may be confounded by numerous factors: committee composition, speaker demographics, document timing, topic-specific considerations, institutional context. Without a causal identification strategy, these explanations remain untested speculation.

Required action: Either (a) Remove causal explanations and recast as descriptive findings with hypotheses for future testing; or (b) Provide some empirical basis for explanations (e.g., variation in framing within countries correlated with hypothesized drivers). Current speculation exceeds what the data support.

Minor Concerns

1. Theoretical Framework Integration

The manuscript references Ulnicane et al. (2021) extensively but does not systematically map their frame categories onto the eight-frame typology used here. This mapping would strengthen theoretical integration. A table showing correspondence between categories would be helpful.

2. Frame Category Validity

The eight-frame typology is presented without clear justification or theoretical grounding. Why these eight? Are there missing frames (e.g., environmental sustainability, labor, public interest)? Are some frames too broad (Governance captures both institutional capacity AND regulatory frameworks)? The typology needs clearer definition and conceptual justification.

3. Statistical Significance Reporting

Table 1 reports p-values from chi-square tests but does not specify whether the tests compare US vs. Global South across all 8 frames (global chi-square) or frame-by-frame. The asterisks suggest the latter, but this is unclear. Additionally, p-values are reported inconsistently (some frames marked with asterisks, others without).

4. Missing References

Key literature on LLM-based content analysis is missing. Recent work by Nelson et al. (2021) is cited, but more recent developments in LLM prompting for content analysis (e.g., Liu et al., 2023; Zhou et al., 2023) should be referenced given the method's centrality to the study.

5. Repetitive Limitations Section

The "Limitations" section appears twice (lines 255-268 and 269-275) with nearly identical content. This should be consolidated.

6. Missing Effect Size Interpretation

The manuscript reports a 22.9 percentage point difference in governance framing but does not interpret what this means substantively. Is this a large effect in the context of framing research? Effect size conventions for this domain should be referenced.

7. Abstract Claims

The abstract makes strong claims about "starkly divergent framing patterns" and "fundamental divergence" without qualifying these as descriptive findings from a limited sample with methodological limitations. Abstract should accurately reflect scope of claims.

8. Temporal Analysis

No temporal trend analysis is provided despite data spanning 2019-2026. Did framing patterns shift over time? Did the post-ChatGPT period show different distributions? This analysis would strengthen the study's contribution.

Recommendations

For Revision (Conditional on Addressing Major Concerns):

1. **Address document type comparability:**
    - Conduct multivariate analysis controlling for document type
    - Or restrict analysis to comparable document types
    - Or explicitly model document type effects
2. **Expand statistical analysis:**
    - Add confidence intervals for all proportions
    - Implement multivariate analysis (logit/multinomial logit)
    - Use multilevel modeling for nested data structure
    - Adjust for multiple comparisons
    - Report formal effect sizes (Cramer's V, odds ratios)
    - Conduct power analysis given sample size disparities
3. **Improve validity assessment:**
    - Add human-coded validation subset (20-30 documents)
    - Implement proper adjudication methodology
    - Provide complete prompts in appendix
    - Either improve rights frame reliability or qualify findings
4. **Enhance replicability:**
    - Provide search queries and API parameters
    - Create document registry with IDs and metadata
    - Specify LLM parameters (temperature, version, etc.)
    - Operationalize density threshold precisely
    - Describe text preprocessing pipeline
5. **Address selection bias:**
    - Exclude or flag India findings as exploratory
    - Provide temporal distribution comparison
    - Discuss source bias implications
    - Specify document languages and translation methods
    - Conduct sensitivity analysis
6. **Reframe causal explanations:**
    - Remove untested causal claims
    - Present explanations as hypotheses for future research
    - Or provide empirical basis for explanations
7. **Minor improvements:**
    - Consolidate duplicate limitations section
    - Add frame typology justification and theoretical mapping
    - Include LLM content analysis citations
    - Add temporal trend analysis if feasible
    - Qualify abstract claims appropriately
    - Clarify statistical test reporting in tables

For Future Research:

1. Consider longitudinal design to establish temporal precedence and causal direction
2. Incorporate speaker-level analysis to examine institutional and

individual variation
3. Compare legislative discourse to executive documents to assess venue effects
4. Develop causal identification strategy (e.g., natural experiments, difference-in-differences)
5. Expand Global South sample beyond BRICS members to assess generalizability
6. Include more granular temporal analysis to track framing evolution

---

Decision: Major Revision Required

The manuscript addresses an important theoretical question and presents intriguing findings. However, the methodological limitations are substantial enough that the findings' validity cannot be established with confidence. The authors must address the six major concerns—particularly document type comparability, statistical underdevelopment, and replicability gaps—before the manuscript can be considered for publication.

I encourage the authors to undertake these revisions. The conceptual contribution is significant, and with improved methodological rigor, this could make a strong contribution to the literature on AI governance and comparative technology politics.

Reviewer's Note: This review was prepared as an anonymous methodological assessment focusing on research design validity, statistical methods, replicability, and causal inference. Comments are intended to be constructive while maintaining rigorous standards for empirical social science research.