Toward Analytical Calibration of Large Language Models in Data-Driven News Reporting

Keonvin Park[1][0009−0007−0626−0080]

Interdisciplinary Program in Artificial Intelligence, Seoul National University
kbpark16@snu.ac.kr

Abstract. Generative AI and large language models (LLMs) are increasingly integrated into data-driven journalism workflows, enabling automated report generation from structured data sources. While prior work has emphasized fluency, coherence, and reasoning performance, significantly less attention has been paid to the analytical calibration of the claims produced by LLM-based systems. In data-centric news contexts, miscalibrated interpretations—such as overstated statistical significance or unsupported causal claims—can distort public understanding and reduce trust in AI-assisted journalism. This extended abstract proposes a research framework for evaluating and improving the analytical calibration of LLM-generated news reports. Rather than presenting empirical findings, we outline a structured evaluation methodology that measures alignment between numerical evidence (e.g., statistical test outputs, effect sizes, confidence intervals) and the corresponding natural-language claims. We introduce proposed metrics to quantify overstatement, underconfidence, and statistical inconsistency, and describe an experimental design for controlled evaluation under varying data conditions. The goal of this work is to establish a principled foundation for assessing and improving calibration in LLM-assisted news generation systems, contributing to the responsible deployment of generative AI in journalism.

Keywords: Large Language Models, Generative AI in Journalism , Data-Driven News Reporting , Analytical Calibration , Responsible AI

1 Introduction

Generative artificial intelligence and large language models (LLMs) are rapidly reshaping the end-to-end news ecosystem, influencing news production, summarization, personalization, and audience engagement. Recent advances in foundation models have demonstrated strong capabilities in long-form generation, structured reasoning, and knowledge synthesis [1]. These capabilities have accelerated the integration of LLMs into data-driven journalism pipelines, where structured datasets are translated into narrative reports. While prior research has extensively examined factual consistency and hallucination in language models [3], relatively limited attention has been devoted to the calibration of analytical claims produced by LLM-assisted news systems. In journalistic contexts, numerical evidence such as statistical tests, polling results, or economic indicators

often forms the basis of public interpretation. Misalignment between quantitative evidence and the strength of generated claims may lead to overstated conclusions, implied causality, or misleading interpretations. Existing work in computational journalism has explored misinformation detection [4], media bias analysis [5], and automated news generation [6]. However, the question of whether generative systems appropriately calibrate their analytical language relative to structured evidence remains underexplored. In this extended abstract, we propose a research framework for evaluating *analytical calibration* in LLM-driven news reporting systems. Rather than presenting completed empirical findings, we outline a principled methodology, define measurable calibration metrics, and describe an experimental design for systematic evaluation. Our objective is to establish foundations for responsible deployment of generative AI in data-driven journalism.

2 Methods

This work proposes a structured evaluation framework for studying analytical calibration in LLM-assisted data-driven news reporting systems. As this extended abstract outlines a research proposal rather than completed empirical findings, we describe the methodological components and planned evaluation protocol.

2.1 System Overview

We consider a three-stage pipeline:

1. **Structured Evidence Generation**: Numerical outputs are derived from controlled datasets, including statistical test results (e.g., p-values), effect sizes, confidence intervals, and summary statistics. These structured signals simulate common data journalism scenarios such as polling analysis, economic reporting, and comparative trend analysis.
2. **LLM-Based Report Generation**: Given structured evidence, a large language model generates a narrative report. We consider multiple prompting strategies, including:
   - Direct numerical input (raw statistical outputs),
   - Template-grounded prompts that explicitly describe statistical interpretation rules,
   - Minimal-context prompts where the model infers interpretation implicitly.
3. **Calibration Assessment**: The generated textual claims are evaluated for alignment with the structured evidence using proposed calibration metrics.

2.2 Definition of Analytical Calibration

Let E denote structured numerical evidence extracted from data, and let C denote the natural-language claim generated by the LLM. Analytical calibration

concerns whether the strength and type of claim in C are consistent with the statistical support in E.

We categorize potential miscalibration into three types:

- **Overstatement**: The generated claim implies stronger statistical support than justified (e.g., implying significance when $p > 0.05$).
- **Under-confidence**: The generated claim is overly cautious despite strong statistical evidence.
- **Statistical Inconsistency**: The narrative contradicts explicit numerical values.

2.3 Proposed Calibration Metrics

We propose the following evaluation metrics:

- **Claim Strength Alignment Score (CSAS)**: A rule-based or model-assisted score measuring consistency between statistical strength (e.g., significance level) and linguistic intensity (e.g., "strong evidence," "suggests," "may indicate").
- **Overclaiming Rate**: The proportion of generated outputs that imply statistical significance or causality beyond predefined thresholds.
- **Consistency Error Rate**: A binary or graded measure indicating contradictions between numerical evidence and textual interpretation.

2.4 Planned Experimental Design

Although experiments have not yet been conducted, we outline a structured evaluation protocol:

1. Construct controlled synthetic datasets where statistical strength is systematically varied (e.g., different effect sizes and sample sizes).
2. Generate corresponding LLM reports under different prompting strategies.
3. Automatically extract claim strength indicators from generated text using rule-based linguistic patterns or secondary LLM evaluation.
4. Compute calibration metrics and analyze sensitivity to statistical variation.

This design enables controlled comparison of model behavior under distributional shifts in numerical evidence, providing insights into the robustness and reliability of generative systems in journalistic contexts.

2.5 Scope and Limitations

This proposed framework focuses specifically on analytical calibration rather than factual hallucination or misinformation detection. While related, calibration emphasizes the alignment between quantitative evidence and narrative strength, which constitutes a distinct reliability dimension in data-driven journalism.

3 Data

As this work outlines a proposed evaluation framework, we describe the planned data sources and construction strategy for studying analytical calibration in LLM-assisted news reporting systems.

3.1 Structured News Analytics Scenarios

We focus on data-driven journalism settings in which numerical evidence plays a central role. Planned scenarios include:

- **Polling and Election Analysis**: Synthetic and semi-real polling datasets with varying sample sizes, margins of error, and statistical significance levels.
- **Economic Indicators**: Time-series data representing unemployment rates, inflation, or GDP growth with controlled effect sizes and trend shifts.
- **Comparative Studies**: Group comparisons with systematically varied effect sizes and confidence intervals.

These scenarios simulate realistic reporting contexts where statistical interpretation is critical.

3.2 Public News Corpora

To contextualize generated reports within realistic journalistic language, we plan to incorporate publicly available news datasets, such as:

- Large-scale news article corpora (e.g., general news collections used in computational journalism research),
- Datasets used in misinformation detection and media bias studies,
- Open news aggregation datasets for headline and summary analysis.

These corpora will be used to extract linguistic patterns and claim intensity markers for calibration analysis.

3.3 Synthetic Calibration Benchmark

A key component of the proposed framework is the construction of a controlled synthetic benchmark. In this benchmark:

1. Statistical parameters (e.g., p-values, effect sizes, confidence intervals) are systematically varied.
2. Each structured evidence instance is paired with a prompt requesting a news-style analytical summary.
3. Generated outputs are evaluated for alignment with the known statistical ground truth.

This synthetic design enables controlled experimentation under varying statistical strength and distributional conditions, facilitating precise measurement of overstatement and under-confidence behaviors.

3.4 Data Characteristics and Limitations

The proposed datasets emphasize structured numerical evidence rather than unstructured misinformation content. As such, the focus is not on fact-checking real-world claims but on evaluating alignment between quantitative inputs and generated interpretations.

Future extensions may incorporate real-world newsroom workflows and longitudinal reporting archives to validate ecological robustness.

4 Expected Results

As this work outlines a proposed research framework, empirical results are not yet available. However, based on prior observations of LLM behavior in analytical and reasoning tasks, we formulate several expected outcomes.

4.1 Baseline Calibration Behavior

We expect that, under minimal prompting conditions, LLM-generated news reports will exhibit measurable levels of analytical miscalibration. In particular:

- Moderate rates of **overstatement**, especially when statistical evidence is marginal (e.g., p-values close to conventional thresholds).
- Occasional **causal language** emerging from correlational inputs.
- Increased inconsistency under low-signal or noisy statistical conditions.

Such behaviors are consistent with previously observed tendencies of LLMs to produce confident and fluent outputs even under uncertainty.

4.2 Impact of Structured Grounding

We hypothesize that explicitly grounding the model with structured statistical cues and interpretation guidelines will reduce calibration errors. Specifically:

- A decrease in the proposed Overclaiming Rate.
- Improved alignment between statistical strength and linguistic intensity.
- Lower Consistency Error across controlled perturbations.

We further anticipate that template-based prompting strategies will outperform minimal-context prompting in calibration robustness.

4.3 Sensitivity to Statistical Variation

Under controlled synthetic variation (e.g., systematically changing effect sizes and sample sizes), we expect LLM outputs to demonstrate partial but imperfect sensitivity to quantitative shifts. While models may reflect directional changes in evidence strength, calibration may remain coarse-grained without explicit numerical grounding.

4.4 Broader Implications

If these hypotheses are supported, the findings would suggest that calibration deficiencies in generative news systems are measurable and potentially mitigable through structured grounding mechanisms. This would provide actionable guidance for responsible deployment of generative AI in data-driven journalism workflows.

5 Conclusion

This extended abstract proposes a structured framework for evaluating analytical calibration in LLM-assisted data-driven news reporting systems. As generative AI becomes increasingly embedded in journalistic workflows, ensuring alignment between quantitative evidence and narrative interpretation is critical for maintaining public trust and informational integrity. Rather than presenting empirical findings, this work outlines a principled methodology, defines measurable calibration criteria, and proposes an experimental design for systematic evaluation. We argue that analytical calibration constitutes a distinct and underexplored reliability dimension in generative news systems, complementary to existing work on factual hallucination, misinformation detection, and bias analysis. By formalizing calibration assessment and proposing controlled evaluation protocols, this framework aims to support the responsible deployment of generative AI in journalism. Future work will involve implementing the proposed benchmark, conducting empirical validation across multiple model architectures, and exploring mitigation strategies through structured grounding and prompt design. We hope this research direction contributes toward more transparent, accountable, and evidence-aligned generative systems in the evolving news ecosystem.

References

1. Brown, T.B., Mann, B., Ryder, N., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 1877–1901 (2020)
2. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv:2302.13971 (2023)
3. Ji, Z., Lee, N., Frieske, R., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
4. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations **19**(1), 22–36 (2017)
5. Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., Nakov, P.: We can detect your bias: Predicting the political ideology of news articles. In: Proceedings of EMNLP, pp. 4982–4991 (2020)
6. Graefe, A.: Automated journalism: Toward new frontiers in journalism. Digital Journalism **6**(6), 745–762 (2018)