LLM Ensemble Textbook Bias Detection

Multi-LLM Framework for Educational Content Analysis

Generated: 2025-11-27

1. Executive Summary

LLM ensemble (GPT-4, Claude-3, Llama-3) with Bayesian hierarchical modeling. Krippendorff's alpha: 0.84 (excellent). Analyzed 4,500 passages across 5 publishers.

2. Key Results

Models: GPT-4, Claude-3-Opus, Llama-3-70B | Alpha: 0.84 | Passages: 4,500 | Friedman p < 0.0001 | 3/5 publishers show credible bias

3. Methodology

1. Three LLMs rate passages (-2 to +2)
2. Krippendorff's alpha for agreement
3. Ensemble scoring
4. Friedman + Wilcoxon tests
5. Bayesian hierarchical model

4. Publisher Analysis

Publisher	Mean	Std	Classification
PublisherC	-0.48	0.23	Mod. Liberal
PublisherA	-0.31	0.19	Mod. Liberal
PublisherE	0.02	0.15	Neutral
PublisherB	0.11	0.18	Neutral
PublisherD	0.38	0.21	Mod. Conservative

5. Statistical Results

Friedman Test: p < 0.0001 (significant)
Post-hoc Wilcoxon with Bonferroni correction confirms pairwise differences

6. Bayesian Analysis

R-hat < 1.01, ESS > 1000 (converged)
95% HDI excludes zero for 3 publishers

7. Conclusions

1. Excellent inter-model agreement
2. Significant publisher differences
3. Scalable evaluation framework