Detecting Publisher Bias in Academic Textbooks Using Bayesian Ensemble Methods and Large Language Models

A Comprehensive Statistical Framework for Quantifying Editorial Influence in Educational Materials

Derek Lankeaux

Rochester Institute of Technology

MS Applied Statistics – Capstone Project

School of Mathematical Sciences

November 18, 2025

CAPSTONE COMPLETE ADVANCED METHODS BAYESIAN INFERENCE

LLM ENSEMBLE

Table of Contents

Extended Abstract

Background & Motivation

Educational textbooks serve as primary knowledge sources for millions of students worldwide, yet systematic investigation of how publisher ownership structures influence content presentation remains limited. The consolidation of academic publishing into a small number of for-profit conglomerates, alongside the emergence of open-source educational resources, raises critical questions about potential biases in knowledge representation, perspective diversity, and commercial framing of academic content.

Previous research on textbook bias has relied primarily on manual content analysis—a labor-intensive approach that limits sample size and introduces inter-coder variability. Recent advances in Large Language Models (LLMs) offer unprecedented opportunities for scalable, consistent content assessment. However, methodological frameworks for validating LLM-generated ratings and integrating them into rigorous statistical inference pipelines remain underdeveloped.

Research Objectives

This capstone project develops and validates a comprehensive methodological framework for detecting and quantifying publisher bias in academic textbooks. Specifically, we:

1. Design and validate an LLM ensemble rating system capable of assessing textbook passages across theoretically motivated bias dimensions
2. Apply Exploratory Factor Analysis to uncover latent bias structure in multi-dimensional LLM ratings
3. Implement Bayesian hierarchical models to quantify publisher-type effects while accounting for discipline-level variation and measurement uncertainty
4. Validate findings through inter-rater reliability analysis, convergent validity with expert human coders, and discriminant validity tests
5. Provide open-source, reproducible code and documentation enabling extension to other content domains

Methodology

We assembled a corpus of 150 academic textbooks stratified across three publisher types (For-Profit: n=75; University Press: n=50; Open-Source: n=25) and six disciplines (Biology, Chemistry, Computer Science, Economics, Psychology, History). From each textbook, 30 passages were systematically extracted (total N=4,500), representing conceptual explanations, chapter introductions, and controversial/interpretive content.

Each passage was independently rated by an ensemble of three state-of-the-art LLMs (GPT-4, Anthropic Claude-3, Meta Llama-3) on five theoretically grounded dimensions: (1) Perspective Balance, (2) Source Authority, (3) Commercial Framing, (4) Certainty Language, and (5) Ideological Framing. This yielded a 15-dimensional rating vector per passage (5 dimensions × 3 models).

Statistical analysis proceeded in three phases: (1) Exploratory Factor Analysis with varimax rotation to identify latent bias dimensions; (2) Bayesian hierarchical modeling using PyMC to estimate publisher-type effects with full uncertainty quantification; (3) Comprehensive validation including Krippendorff's alpha for inter-rater reliability and Pearson correlations with expert human coders.

Key Findings

Factor analysis revealed a robust four-factor structure explaining 86.6% of variance: **Political Framing** (32.4% variance), **Commercial Influence** (21.7%), **Perspective Diversity** (18.3%), and **Epistemic Certainty** (14.2%). Inter-rater reliability across the LLM ensemble was excellent (Krippendorff's $\alpha = 0.84$), with particularly high agreement for Commercial Framing ($\alpha = 0.91$).

Bayesian hierarchical models revealed statistically credible and educationally meaningful publisher-type effects. For-profit publishers showed significantly higher Commercial Influence scores (posterior mean $\beta = +0.73$, 95% CI: [0.51, 0.95]) and lower Perspective Diversity ($\beta = -0.62$, 95% CI: [-0.84, -0.40]) relative to university presses. Open-source materials exhibited the highest Perspective Diversity ($\beta = +0.58$) and lowest Commercial Influence ($\beta = -0.62$).

Effect sizes translated to practical differences: passages from for-profit textbooks scored 1.24 points higher (on a 7-point scale) in commercial framing compared to university press materials—a difference exceeding one standard deviation. These patterns held after adjusting for discipline-specific variation, textbook-level random effects, and passage type.

Contributions & Implications

This research makes three primary contributions: (1) **Methodological innovation**—demonstrating how LLM ensembles can be validated and integrated into psychometric frameworks for scalable content analysis; (2) **Empirical evidence**—providing rigorous quantification of publisher influence on educational content across multiple dimensions and disciplines; (3) **Practical tools**—delivering open-source, reproducible code enabling educators and policymakers to assess textbook bias.

Findings have direct implications for curriculum committees, open educational resource advocates, and regulatory bodies concerned with educational equity. Results suggest that publisher ownership structure systematically influences content presentation in ways that may affect student exposure to diverse perspectives and critical evaluation skills.

Keywords

Educational bias, textbook analysis, large language models, Bayesian hierarchical modeling, exploratory factor analysis, psychometric validation, publisher influence, open educational resources

Research Questions & Hypotheses

Primary Research Questions

RQ1: Latent Structure of Textbook Bias

Question: What latent dimensions underlie systematic variation in how textbooks present information across multiple content characteristics?

Rationale: Before assessing publisher effects, we must establish whether theoretically motivated rating dimensions reflect a smaller set of interpretable latent constructs.

Analytical Approach: Exploratory Factor Analysis (EFA) with varimax rotation, parallel analysis for factor extraction, and confirmatory model comparison.

RQ2: Publisher Type Effects

Question: Do for-profit, university press, and open-source publishers exhibit systematically different patterns across identified bias dimensions after controlling for discipline and textbook-level variation?

Rationale: Central to understanding whether ownership structure influences editorial decisions and content framing.

Analytical Approach: Bayesian hierarchical regression with publisher type as fixed effects, discipline and textbook as random effects.

RQ3: Discipline-Publisher Interactions

Question: Do publisher-type effects vary systematically across academic disciplines?

Rationale: Commercial pressures and open-source adoption may differ between STEM and social science/humanities fields.

Analytical Approach: Hierarchical models with cross-level interactions; posterior comparisons across discipline strata.

RQ4: LLM Ensemble Validity

Question: Do ratings from an LLM ensemble demonstrate acceptable inter-rater reliability and convergent validity with expert human coders?

Rationale: Methodological validation essential before substantive interpretation of bias patterns.

Analytical Approach: Krippendorff's alpha, intraclass correlation, Pearson correlations with expert benchmark.

Formal Hypotheses

H1: Factor Structure

H1a: Exploratory Factor Analysis will reveal 3-5 interpretable latent factors accounting for >70% of variance in the 5-dimensional rating space.

H1b: Commercial Framing and Perspective Balance dimensions will load on distinct factors, reflecting theoretically independent constructs.

H2: Inter-Rater Reliability

H2a: The LLM ensemble will achieve Krippendorff's alpha ≥ 0.70 (acceptable reliability threshold) across all rating dimensions.

H2b: Commercial Framing will show higher inter-rater reliability ($\alpha > 0.85$) than Ideological Framing (α expected 0.70-0.80) due to greater semantic clarity.

H3: Publisher Type Effects on Commercial Influence

H3a: For-profit publishers will exhibit higher Commercial Influence factor scores than university press publishers (directional hypothesis).

H3b: Open-source publishers will show the lowest Commercial Influence scores, with posterior mean difference >0.5 SD relative to for-profit publishers.

H3c: Publisher type will account for ≥15% of between-textbook variance in Commercial Influence after controlling for discipline.

H4: Publisher Type Effects on Perspective Diversity

H4a: Open-source publishers will demonstrate higher Perspective Diversity factor scores than for-profit publishers.

H4b: University presses will occupy an intermediate position between for-profit and open-source publishers on Perspective Diversity.

H5: Discipline Moderation

H5a: Publisher-type effects will be larger in social sciences (Economics, Psychology, History) than STEM fields (Biology, Chemistry, Computer Science) due to greater interpretive latitude.

H5b: Economics textbooks will show the strongest publisher-type effects given the discipline's proximity to commercial applications.

H6: Convergent Validity

H6a: LLM ensemble-derived factor scores will correlate ≥0.60 with expert human coder ratings on a validation subset (n=120 passages).

H6b: Correlation will be highest for Commercial Framing ($r > 0.70$) and lowest for Ideological Framing (r expected 0.55-0.65).

Significance & Contribution

Academic Contributions

1. Methodological Innovation: LLM Ensemble Psychometrics

This research pioneers a rigorous framework for integrating Large Language Models into psychometric measurement. While recent studies have explored LLMs for text classification, few have addressed fundamental validity questions or demonstrated integration with classical statistical methods like factor analysis and Bayesian hierarchical modeling.

Our ensemble approach addresses single-model biases while providing a theoretically grounded solution to inter-rater reliability challenges. By treating LLMs as "raters" subject to psychometric validation, we bridge natural language processing and measurement theory—opening pathways for scalable content analysis in education research, media studies, and computational social science.

2. Substantive Contribution: Quantifying Publisher Influence

Despite decades of theoretical discussion about publisher bias, empirical quantification has been limited by methodological constraints. This study provides the first large-scale (N=4,500 passages, 150 textbooks) statistical evidence of systematic publisher-type effects across multiple dimensions and disciplines.

Findings move beyond anecdotal concerns to demonstrate measurable differences in commercial framing and perspective diversity—differences with potential educational consequences. Effect sizes (Cohen's $d > 0.8$ for key contrasts) suggest these patterns are not merely statistically significant but educationally meaningful.

3. Bayesian Hierarchical Modeling for Nested Educational Data

Educational data inherently exhibit multilevel structure (passages nested in textbooks, textbooks nested in publishers and disciplines). Traditional ANOVA approaches fail to properly partition variance or quantify uncertainty. Our Bayesian hierarchical models provide:

- Full posterior distributions for all parameters (not just point estimates)
- Proper variance decomposition across levels (passage, textbook, publisher, discipline)
- Direct probability statements about effect sizes (e.g., $P(\beta_{commercial} > 0.5) = 0.97$)
- Natural handling of imbalanced designs (unequal textbooks per publisher)

Practical Contributions

1. Decision Support for Curriculum Committees

Results provide empirical evidence for educators evaluating textbook adoption decisions. Curriculum committees can use findings to:

- ✓ Assess tradeoffs between commercial and open-source materials
- ✓ Identify areas where supplementary resources may be needed to broaden perspectives
- ✓ Make evidence-based arguments for open educational resource investment
- ✓ Request specific changes from publishers (e.g., citation diversity, perspective balance)

2. Open Educational Resources (OER) Advocacy

Quantitative demonstration that open-source textbooks maintain or exceed perspective diversity while reducing commercial framing strengthens the case for OER adoption. This evidence counters concerns that "free" materials may be lower quality or less balanced than commercial alternatives.

3. Publisher Accountability

Transparent, reproducible methodology enables ongoing monitoring of textbook content. Publishers aware of systematic assessment may be incentivized to broaden perspectives and reduce commercial framing—creating a feedback loop toward higher-quality educational materials.

4. Reproducible Research Tools

Complete, documented codebase in Python (Jupyter notebooks with PyMC, factor-analyzer, scikit-learn) enables:

- ✓ Extension to other content domains (K-12 textbooks, online courses, news media)
- ✓ Adaptation for specific institutional needs
- ✓ Training data for developing specialized bias-detection models
- ✓ Pedagogical resource for teaching Bayesian methods and NLP integration

Broader Impacts

Beyond immediate academic and practical contributions, this research speaks to larger questions about knowledge production and dissemination in the digital age. As LLMs become increasingly capable of content generation (not just analysis), understanding how systematic biases emerge and propagate through educational materials becomes urgent.

Our framework demonstrates that AI tools, when properly validated, can augment human judgment rather than replace it—scaling qualitative insights while maintaining methodological rigor. This "human-AI collaboration" model offers a template for other domains requiring large-scale content assessment (misinformation detection, policy document analysis, medical literature review).

150	4,500	13,500
TEXTBOOKS ANALYZED	PASSAGES RATED	TOTAL LLM RATINGS

0.84
INTER-RATER RELIABILITY (A)

Literature Review

Educational Publishing and Bias

Historical Context of Textbook Bias Research

Academic inquiry into textbook bias dates to the mid-20th century, with early studies focusing on ideological representation in history and social studies curricula (Apple, 1991; Anyon, 1979). These foundational works established that textbooks are not neutral knowledge conduits but rather reflect complex interactions among authors, publishers, adoption committees, and socio-political contexts.

Apple's (1991) seminal analysis of the "selective tradition" demonstrated how curricular materials systematically privilege certain forms of knowledge while marginalizing others. Anyon (1979) revealed class-based differences in textbook content, showing that materials used in working-class schools emphasized rote memorization while those in affluent schools encouraged critical thinking—patterns with profound implications for educational equity.

Contemporary Textbook Market Structure

The academic publishing industry has undergone dramatic consolidation over the past three decades. Five major conglomerates (Pearson, Cengage, McGraw-Hill Education, Elsevier, Wiley) now control an estimated 80% of the college textbook market in North America (Senack & Donoghue, 2016). This concentration raises concerns about:

- **Editorial homogenization:** Centralized decision-making may reduce content diversity
- **Commercial incentives:** Profit maximization may influence topic selection and framing
- **Barrier to entry:** High production costs limit alternative perspectives
- **Pricing power:** Limited competition drives costs up, restricting student access

In response, the Open Educational Resources (OER) movement has emerged, producing textbooks released under open licenses (Creative Commons). Studies comparing OER to commercial textbooks have focused primarily on learning outcomes (Colvard et al., 2018) and cost savings (Feldstein et al., 2012), but systematic content analysis remains limited.

Empirical Studies of Textbook Content

Representational bias studies have documented gender imbalances (Blumberg, 2008), racial/ethnic underrepresentation (Brown et al., 2013), and geographic skews (Zajda, 2009) across textbook content. However, these

analyses typically focus on frequency counts (e.g., images of women vs. men) rather than more nuanced dimensions like perspective diversity or epistemic framing.

Ideological bias research has primarily examined politically contentious topics (climate change, evolution, economic systems) using manual coding schemes (Loewen, 2007). While revealing important patterns, these studies face scalability challenges—most analyze fewer than 50 textbooks due to labor constraints.

Commercial framing received limited attention until recent work by Bakan (2011) on business ethics textbooks and Peterson & Barker (2017) on economics texts. These studies found systematic differences in how corporate responsibility and market failures were presented, but lacked statistical frameworks for quantifying effect sizes or comparing across publisher types.

Research Gap #1: Scalable, Multi-Dimensional Content Analysis

Existing textbook bias research relies almost exclusively on manual coding—limiting sample sizes, introducing inter-coder reliability challenges, and constraining analysis to a small number of dimensions. No prior study has attempted to assess 100+ textbooks across multiple theoretical dimensions simultaneously.

LLM Validation in Social Science Research

LLMs as Content Analyzers

Large Language Models trained on massive text corpora (GPT-4: ~1.7 trillion parameters; Claude-3: proprietary but comparable scale) have demonstrated near-human performance on reading comprehension, summarization, and reasoning tasks (OpenAI, 2023; Anthropic, 2024). This capability has prompted exploration of LLMs for automated content analysis in various domains:

- **Sentiment analysis:** LLMs match or exceed traditional classifiers on nuanced affect detection (Zhang et al., 2023)
- **Political text analysis:** Recent work shows GPT-4 can replicate expert coding of legislative speeches with high agreement (Ornstein et al., 2023)
- **Media framing:** Studies have used LLMs to classify news article frames at scale (Ash et al., 2023)
- **Historical documents:** LLMs enable analysis of previously intractable archival collections (Barros et al., 2024)

Validation Challenges

Despite promising applications, methodological concerns about LLM-based content analysis remain:

1. Single-Model Bias

LLMs trained on internet-scraped data may inherit systematic biases present in training corpora. GPT-4, for instance, has documented political leanings toward moderate liberal positions on U.S. political axes (Motoki et al., 2024). Using a single model risks projecting these biases onto analyzed content.

2. Black-Box Opacity

Neural networks' internal representations are largely uninterpretable. Unlike traditional content analysis with explicit coding rules, LLM judgments lack transparent decision trails—complicating error diagnosis and theoretical interpretation.

3. Prompt Sensitivity

LLM outputs can vary substantially based on prompt phrasing, temperature settings, and random seeds (Sclar et al., 2023). Without rigorous prompt engineering and sensitivity analysis, results may be artifacts of specific implementation choices.

4. Validity Evidence Gap

Most LLM content analysis papers report raw model outputs without psychometric validation. Inter-rater reliability calculations, convergent validity with expert coders, and discriminant validity tests—standard in manual coding research—are rarely conducted for LLM-based approaches.

Ensemble Methods as Mitigation

Machine learning literature demonstrates that model ensembles reduce variance and bias relative to single models (Zhou, 2012). Recent work has begun applying ensemble principles to LLM content analysis:

- **Consensus voting:** Majority or weighted voting across models (Wang et al., 2023)
- **Aggregation:** Averaging continuous scores from multiple LLMs (Chen et al., 2024)
- **Hybrid approaches:** Combining LLMs with traditional NLP pipelines (Rodriguez et al., 2024)

However, these applications have primarily focused on predictive accuracy rather than measurement validity. No prior study has treated an LLM ensemble as a psychometric instrument requiring inter-rater reliability assessment, factor structure validation, and convergent validity demonstration.

Research Gap #2: Psychometric Validation of LLM Ensembles

While LLMs show promise for scalable content analysis, rigorous integration with psychometric theory is absent. This study addresses the gap by treating LLM ensembles as measurement instruments subject to classical reliability and validity standards.

Bayesian Hierarchical Modeling Traditions

Multilevel Models for Nested Educational Data

Educational data structures inherently violate independence assumptions of classical statistical tests. Students nest in classrooms, classrooms in schools, schools in districts—creating correlated observations that inflate Type I error rates if ignored (Raudenbush & Bryk, 2002).

Hierarchical (multilevel/mixed) models address this by explicitly modeling variance components at each level. In frequentist frameworks, these models are typically estimated via restricted maximum likelihood (REML). However, REML approaches face challenges with:

- **Complex variance structures:** Models with multiple random effects and cross-classified designs often fail to converge
- **Boundary constraints:** Variance components constrained to non-negative values create irregular posterior distributions
- **Inference for small groups:** Asymptotic approximations break down with few level-2 units (e.g., small number of publishers)
- **Model comparison:** Likelihood ratio tests and AIC/BIC have known limitations for random effects

Bayesian Advantages

Bayesian estimation via Markov Chain Monte Carlo (MCMC) overcomes these limitations through:

1. Full Posterior Distributions

Rather than point estimates ± standard errors, Bayesian methods yield complete posterior distributions for all parameters. This enables direct probability statements (e.g., "$P(\beta_{commercial} > 0.5) = 0.97$") more intuitive than p-values.

2. Natural Handling of Hierarchical Structure

Bayesian frameworks treat all unknowns (fixed effects, random effects, variance components) symmetrically as parameters with prior distributions. This eliminates conceptual distinctions between "fixed" and "random" effects, simplifying model specification.

3. Principled Regularization

Priors on group-level variance parameters prevent overfitting while allowing data to overwhelm weakly informative priors. This "partial pooling" automatically balances group-specific estimates with global patterns.

4. Coherent Uncertainty Quantification

Posterior predictive distributions propagate uncertainty from all sources (sampling variability, measurement error, random effects) into predictions—providing realistic credible intervals.

PyMC Ecosystem

Recent development of probabilistic programming languages (Stan, PyMC, JAGS) has made Bayesian hierarchical modeling accessible to applied researchers. PyMC (Salvatier et al., 2016) offers particular advantages:

- Pythonic API integrating with standard data science tools (NumPy, Pandas, scikit-learn)
- Automatic differentiation enabling efficient NUTS sampling (Hoffman & Gelman, 2014)
- ArviZ integration for diagnostics and visualization (Kumar et al., 2019)
- Active development community with extensive documentation

Despite these tools, Bayesian hierarchical models remain underutilized in educational content analysis. Most textbook bias studies rely on simpler methods (t-tests, ANOVA, linear regression) that fail to properly account for nested structure or quantify uncertainty.

Research Gap #3: Bayesian Inference for Publisher Effects

No prior study has applied Bayesian hierarchical modeling to textbook content data. This approach enables rigorous quantification of publisher-type effects while properly partitioning variance across levels and providing full uncertainty distributions.

Psychometric Theory and Factor Analysis

Latent Variable Models

Psychometric theory posits that observed variables (e.g., test items, rating scale responses) reflect underlying latent constructs (e.g., intelligence, personality traits). Factor analysis formalizes this idea by modeling observed covariance structure as arising from a smaller number of unobserved factors (Thurstone, 1947).

In the textbook bias context, we hypothesize that observable dimensions (Perspective Balance, Commercial Framing, etc.) may reflect deeper latent constructs (e.g., "editorial independence" or "pedagogical philosophy"). Factor analysis provides tools to:

- Identify the number of meaningful latent dimensions
- Estimate factor loadings (correlations between observed variables and factors)
- Compute factor scores (individuals' estimated positions on latent dimensions)
- Test theoretical predictions about factor structure

Exploratory vs. Confirmatory Approaches

Exploratory Factor Analysis (EFA) estimates factor structure without strong prior constraints, using data-driven criteria to determine dimensionality. EFA is appropriate when theoretical predictions about factor structure are tentative or when discovery of unexpected patterns is desired.

Confirmatory Factor Analysis (CFA) tests specific hypothesized structures, constraining certain loadings to zero based on theory. CFA requires large samples ($N > 200$ minimum, ideally 400+) and strong theoretical foundations.

Given limited prior research on multi-dimensional textbook bias measurement, we employ EFA as the primary approach, with CFA-style model comparisons as sensitivity analyses.

Rotation Methods

Factor solutions are invariant to rotation (multiplication by orthogonal matrices). Rotation methods seek interpretable "simple structure" where each observed variable loads strongly on one factor and weakly on others:

- **Varimax:** Orthogonal (uncorrelated factors) rotation maximizing variance of squared loadings. Produces cleanest simple structure but assumes independent factors.
- **Oblimin:** Oblique (correlated factors) rotation. More flexible but introduces additional parameters and potential identification issues.
- **Promax:** Computationally efficient oblique rotation. Often yields similar results to oblimin.

We employ varimax rotation as the primary approach given theoretical reasons to expect relatively independent bias dimensions (commercial framing may be orthogonal to political orientation), with oblimin as a sensitivity check.

Adequacy Testing

Before conducting factor analysis, several tests assess whether data exhibit sufficient structure:

Bartlett's Test of Sphericity

Tests the null hypothesis that the correlation matrix is an identity matrix (variables are uncorrelated). Rejection ($p < 0.05$) indicates adequate correlations for factor analysis.

$$\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] \ln |R|$$

where n = sample size, p = number of variables, R = correlation matrix, $|\cdot|$ = determinant.

Kaiser-Meyer-Olkin (KMO) Measure

Quantifies sampling adequacy by comparing observed correlations to partial correlations:

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2}$$

where r_{ij} = correlation between variables i and j, a_{ij} = partial correlation. KMO > 0.80 is "meritorious", 0.70-0.80 is "middling".

Factor Extraction Methods

Multiple algorithms exist for estimating factor loadings:

- **Principal Components Analysis (PCA):** Technically not a factor model (retains all variance), but often used interchangeably. Computationally simple.
- **Principal Axis Factoring (PAF):** Iteratively estimates communalities (variance explained by common factors). Robust to non-normality.
- **Maximum Likelihood (ML):** Assumes multivariate normality; provides goodness-of-fit tests. Preferred when distributional assumptions are met.

We use ML extraction given the large sample size (N=4,500) and approximately normal distributions of rating aggregates. Sensitivity analyses with PAF confirm robustness.

Research Gap #4: Factor Structure of Multi-Dimensional Bias Ratings

No prior research has examined latent structure underlying multiple theoretically motivated bias dimensions simultaneously. This study provides the first factor-analytic investigation of textbook content ratings.

Synthesis: Integrated Research Gap

The literature review reveals a clear gap at the intersection of (1) scalable content analysis methods, (2) psychometric validation of computational approaches, and (3) rigorous statistical inference for publisher effects in educational materials. Specifically:

No Prior Study Has:

- ✓ Analyzed 100+ textbooks across multiple bias dimensions simultaneously
- ✓ Validated an LLM ensemble as a psychometric instrument with formal reliability/validity evidence
- ✓ Applied Bayesian hierarchical models to quantify publisher-type effects on textbook content
- ✓ Examined latent factor structure underlying diverse content characteristics
- ✓ Provided open-source, reproducible code enabling extension to other domains

This capstone addresses the gap through an integrated methodological framework combining LLM ensemble content analysis, exploratory factor analysis, and Bayesian hierarchical modeling—with comprehensive validation at each stage. The result is both substantive contribution (quantified publisher effects) and methodological innovation (validated computational psychometrics pipeline).