Portraits of Graduate Advising: Advisor Rating Correlates and an Evidence-Based Rubric for PhD Applicants

Dixi Yao*
University of Chicago
dixi@uchicago.edu

Abstract

Choosing a PhD advisor is one of the highest-stakes decisions in a research career, yet applicants rely predominantly on informal signals and increasingly on anonymous review platforms whose data is self-selected, unverified, and whose relationship to objective professional factors remains unexamined. We analyze real review data from the OpenAdvisor platform, covering 11,311 advisor profiles across 490 universities and 12,344 reviews with bilingual (Chinese/English) text. We integrate these reviews with CSRankings faculty metadata and, for a subset of profiles at North American institutions, professional metrics from the Kalhor et al. (2023) dataset. We test three hypotheses: (H1) structural and professional factors systematically covary with advisor ratings; (H2) platform reviewing patterns exhibit biases partly attributable to selection and reporting effects; and (H3) negative review content concentrates in recurring behavioral themes from which an evidence-based applicant rubric can be constructed. We find that the platform is dominated by Chinese university advisors (98.5% of profiles), that ratings are strongly bimodal (mass at 1 and 5), and that professional metrics are unavailable for the vast majority of profiles, limiting the resolution of H1. Among the 166 profiles with professional data, structural factors explain essentially no rating variance ($R^2 \approx 0.000$). Name-based sex inference is unreliable for Chinese names (only 1.4% classifiable), preventing a robust H2 audit. Text analysis of 12,344 real bilingual reviews identifies behavioral themes concentrated in research direction, lab culture, and career development; DeepSeek LLM classification on a 50-review validation sample reveals that keyword-based methods substantially over-identify themes in Chinese text. We construct a 10-dimension, behavior-only advisor-selection rubric and discuss the implications of platform composition bias, multilingual text analysis challenges, and the practical limits of anonymous review data for PhD applicant decision-making.

1 Introduction

Why advisor selection matters. The relationship between a PhD student and their advisor shapes not only the student's research trajectory but their mental health, career outcomes, and likelihood of degree completion Kalhor et al. [2023]. Yet the advisor selection process remains remarkably informal: applicants rely on brief campus visits, word-of-mouth reputation, and increasingly, anonymous online reviews on platforms such as OPENADVISOR and RATEMYPROFESSORS Zheng et al.

*This work was produced with substantial AI assistance: the NeuriCo autonomous research framework, with DeepSeek v4 and Claude (Anthropic) as backend models. The human author curated the research idea, set the ethical constraints, and verified outputs. See Appendix A for a full account of AI usage.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

[2023]. The stakes are especially high in CS/ML, where lab-based funding models, rapid publication cycles, and intense competition create unique advisor-student dynamics Kalhor et al. [2023].

Information asymmetry and its consequences. Applicants face a fundamental information problem: the people best positioned to evaluate an advisor (current and former students) have limited incentives to share candid feedback publicly, while the signals most readily available to applicants (institutional prestige, citation counts, h-index) may not predict actual advising quality. Anonymous review platforms have emerged as a partial solution, but their data introduces new challenges: self-selection of reviewers, lack of verification, and potential demographic biases in who gets reviewed and how harshly Zheng et al. [2023], Reid [2010]. No existing study has simultaneously examined whether professional factors predict ratings, whether the platforms themselves exhibit systematic biases, and whether review content can be translated into actionable guidance for applicants.

Our approach. We address this gap by analyzing real review data from the OPENADVISOR platform, covering 11,311 advisor profiles across 490 universities with 12,344 bilingual reviews. We integrate these reviews with CSRankings faculty metadata Berger [2024] and, for a subset of 166 profiles at North American institutions, professional metrics from KALHOR-2023 (741 CS/ML faculty at top-25 North American programs, 13,936 student-advisor pairs) Kalhor et al. [2023]. We frame our investigation around three hypotheses:

H1 (Professional correlates): Do structural and professional factors (career stage, citation metrics, institution, research community) systematically covary with advisor ratings?

H2 (Platform bias audit): Do platform reviewing patterns vary with advisor demographics in ways attributable to reporting and selection bias beyond professional differences?

H3 (Behavioral themes and rubric): Does negative review content concentrate in recurring behavioral themes from which a practical, behavior-only advisor-selection rubric can be constructed?

Key findings. We find that the OPENADVISOR platform is dominated by Chinese university advisors (98.5% of profiles), with strongly bimodal ratings (mass concentrated at scores of 1 and 5, mean = 2.74, SD = 1.54). Professional metrics are available for only 1.5% of profiles (those at North American universities matching the Kalhor dataset), and among this subset, structural factors explain essentially none of the rating variance ($R^2 \approx 0.000$). Name-based sex inference is unreliable for Chinese names (only 1.4% classifiable), preventing a robust bias audit. Text analysis of real bilingual reviews reveals that keyword-based theme classification over-identifies themes in Chinese text by a factor of approximately $35\times$ compared to LLM-based classification, highlighting a critical methodological challenge. We construct a 10-dimension, behavior-only rubric and discuss the practical implications of platform composition bias and multilingual text analysis for PhD applicant decision-making.

Contributions. In summary, our main contributions are:

- We present the first large-scale analysis of 11,311 real advisor profiles and 12,344 bilingual reviews from the OPENADVISOR platform, documenting platform composition, rating distributions, and text patterns.
- We provide honest reporting on the limits of professional-metric matching (1.5% match rate) and name-based demographic inference (1.4% classification rate) for this platform, identifying key data infrastructure gaps.
- We conduct DeepSeek LLM-based validation of keyword theme classification on 50 bilingual reviews, finding that keyword methods over-identify themes by approximately $35\times$ in Chinese text, establishing a methodological benchmark for future multilingual review analysis.
- We construct an evidence-based, behavior-only advisor-selection rubric, and discuss practical limitations of anonymous review platforms for PhD applicant decision-making.

2 Related Work

Our work sits at the intersection of three research streams: advisor evaluation and bias, CS/ML advisor-student dynamics, and NLP methods for review analysis.

2.1 Advisor Evaluation and Demographic Bias

The most extensive evidence on evaluation bias comes from large-scale analyses of RATEMYPROFESSORS. Zheng et al. [2023] analyzed approximately 9 million RATEMYPROFESSORS reviews using BERTopic and RoBERTa, finding that women professors receive lower ratings in most fields (except Math & Computing, where the difference was not significant) and that gendered language patterns differ systematically: women are more often described as "sweet," "caring," "rude," or "moody," while men are described as "brilliant," "funny," "arrogant," or "boring." Reid [2010] documented intersectional effects, finding that Black male faculty at top-25 liberal arts colleges were rated most negatively.

These studies focus on *undergraduate teaching* evaluations. Our work adapts this framework to *graduate advising*, where power dynamics differ substantially: funding dependence, visa sponsorship, multi-year relationships, and publication co-authorship introduce dimensions absent from course evaluations.

2.2 CS/ML Advisor-Student Dynamics

Kalhor et al. [2023] collected data on 13,936 CS graduate students at the top-25 North American programs, analyzing advisor-student matching patterns. They found strong nationality homophily ($p < 10^{-15}$) but no significant gender matching bias ($p = 0.12$). Their dataset, which we use as our professional-metrics source, includes advisor h-index, citation count, academic rank, country, and publication history, making it the most comprehensive source of CS/ML advisor professional data available.

2.3 NLP Methods for Review Analysis

Methodologically, we draw on established NLP approaches. VADER Hutto and Gilbert [2014] is a validated rule-based sentiment model whose scores track human sentiment judgments on short informal English text. We find that VADER is unreliable for Chinese text, which motivates our use of LLM-based classification as a validation benchmark. Zheng et al. [2023] demonstrated BERTopic for theme extraction and RoBERTa for sentiment in professor reviews. We complement these with LDA topic modeling Blei et al. [2003] and employ DeepSeek v4-pro (via OpenAI-compatible API) for LLM-assisted classification on a sample of bilingual reviews to validate keyword-based theme identification.

2.4 Fairness in Academic Evaluation

Zhang et al. [2022] investigated fairness disparities in peer review at ICLR using language model enhanced methods, finding systematic patterns in review scores across author demographics. Their framework for auditing evaluation systems for bias informs our approach to H2, though we find that the data infrastructure for such audits is inadequate for Chinese-language platforms.

2.5 Positioning Our Work

Unlike prior studies that examine either professional metrics Kalhor et al. [2023] or review platform data Zheng et al. [2023] in isolation, we attempt to merge both sources. The primary contribution of this work is not a definitive answer about advisor rating correlates, but an honest account of the data infrastructure challenges involved in such an integration: the geographic composition mismatch between platforms and professional-metric datasets, the failure of Western-centric demographic inference tools, and the keyword over-identification problem in bilingual text classification. These findings provide a roadmap for the methodological improvements needed before robust cross-platform analysis of advisor reviews becomes feasible.

3 Methodology

3.1 Data Sources and Integration

Our analysis integrates data from three primary sources, summarized in Table 1.

Table 1: Data sources and variables used in this study. The OpenAdvisor data is real, scraped in July 2026. Kalhor and CSRankings provide professional and institutional metadata.

Source	Variables	N
OPENADVISOR (real)	profile_id, name, university, overall_rating, review subscores (6 dimensions), reviewer info, bilingual body text (Chinese/English), date	11,311 profiles; 12,344 reviews
KALHOR-2023 Kalhor et al. [2023]	h-index, citation count, academic rank, sex, field, university (25 top NA programs)	741 advisors; 13,936 pairs
CSRANKINGS Berger [2024]	Name, affiliation, scholar ID	32,236 entries

Real review platform data. We collected real data from the OPENADVISOR platform using a corrected scraper that extracted 12,340 profiles and 12,344 reviews. After cleaning (removing pseudo-profiles with null ratings, placeholder universities, and meta/discussion posts; deduplicating same-name cross-university profiles), the analysis dataset comprises 11,311 advisor profiles across 490 universities. Each profile carries a site-level overall rating (1–5 scale) scraped from the platform metadata, plus review-level structured subscores on six dimensions: overall, mentoring, research guidance, funding, work-life balance, and career support. Reviews include bilingual body text (primarily Chinese, some English), reviewer pseudonyms, dates, and a reviewer-type field. The platform metadata also records the number of moderator-hidden reviews (n_{hidden}), which we use as a moderation-intensity signal.

Professional metrics. The KALHOR-2023 provides advisor-level professional data for 741 unique CS faculty at 25 top North American universities. Since Kalhor uses anonymous IDs and OpenAdvisor profiles are predominantly at Chinese universities, we match at the university level: only OpenAdvisor profiles at institutions matching the 25 Kalhor universities receive aggregate university-level professional metrics (mean h-index, median citations, female representation). The match rate is 1.5% (166 of 11,311 profiles), reflecting the geographic composition mismatch between the two datasets. CSRankings provides name-level matches at 0.9% (98 profiles), primarily for English-named advisors at international institutions.

3.2 Feature Engineering

From the available data we engineer the following features:

- **Rating**: The site-level overall rating (1–5) from OpenAdvisor metadata serves as the primary outcome variable. We also analyze the six review-level subscores where available.
- **Professional metrics**: University-aggregated h-index, log citation count, and career years (from first paper year) for the matched 166 profiles.
- **Contestation**: Binary flag for profiles with $n_{\text{hidden}} > 0$, indicating moderator-hidden reviews.
- **Review bursts**: Same-day clusters of maximum-score reviews following negative reviews, flagged for sensitivity analysis.
- **Demographics**: Sex inferred from first names. For Chinese names (the majority), this inference is unreliable; we report classification rates and restrict bias analyses accordingly.

3.3 Statistical Analysis

H1: Professional correlates. We fit ordinary least squares regression for the subset of profiles with professional metrics ($n = 166$):

$$\text{rating} \sim \beta_0 + \beta_1 \cdot \text{h\_index\_z} + \beta_2 \cdot \text{log\_cite\_z} + \beta_3 \cdot \text{career\_z} + \varepsilon \tag{1}$$

We report R^2, adjusted R^2, and coefficients. Due to the small matched sample, these results are descriptive rather than inferential. For the unmatched majority (98.5%), we report platform-level descriptive statistics.

H2: Platform bias audit. Name-based sex inference produced only a 1.4% classification rate, with only 28 profiles classified as female and 132 as male. With such small and imbalanced groups, statistical tests are underpowered and reported with explicit caveats. We report the available comparisons

Table 2: Advisor sample characteristics ($n = 11,311$). Real data from the OpenAdvisor platform, July 2026 scrape.

Variable	Mean	SD	Min	P25	P50	P75	Max
Overall Rating (1–5)	2.74	1.54	1.00	1.00	3.00	4.00	5.00
Review Count	1.01	0.19	0.00	1.00	1.00	1.00	9.00

Rating distribution: 1-star: 3,703 (32.7%); 2: 1,838 (16.3%);
3: 1,782 (15.8%); 4: 1,575 (13.9%); 5: 2,358 (20.8%)
Unique universities: 490
Kalhor-matched profiles: 166 (1.5%)
Moderator-hidden reviews ($n_{\text{hidden}} > 0$): 45 profiles

(review coverage, rating gaps) and discuss the data infrastructure gap: Chinese names require different inference methods than Western names, and the current name-based approach is inadequate for this platform.

H3: Text analysis and rubric. We apply five complementary methods to the 12,344 real bilingual reviews:

1. **VADER sentiment analysis** Hutto and Gilbert [2014]: Lexicon-based compound polarity scores. We caution that VADER is trained on English-language social media text and its scores on Chinese text are unreliable.
2. **Keyword-based theme classification**: 10 behavioral categories with curated bilingual keyword lists (English and Chinese terms).
3. **LDA topic modeling** Blei et al. [2003]: 6-topic model for unsupervised theme discovery from the review text corpus.
4. **LLM-assisted classification**: DEEPSEEK V4 v4-pro classifies a 50-review sample into behavioral themes. The LLM is prompted with theme definitions and asked to return theme labels for bilingual text, serving as a validation benchmark for the keyword method.
5. **Rubric synthesis**: A 10-dimension, behavior-only rubric constructed from the empirical theme analysis, with weights informed by prevalence and qualitative importance.

3.4 Ethical Framework

This study adheres to a strict ethical protocol: (1) no named individuals appear anywhere in the paper, figures, tables, or code; (2) all analysis is aggregate, with cells $n < 5$ suppressed; (3) reviews are treated as subjective reports, never as ground truth about any person; (4) demographic covariates are used only in the platform bias audit (H2); the applicant rubric (H3) uses only behavioral and structural factors; (5) name-based demographic inference is probabilistic and we report its failure rate for this platform; and (6) a full ethics statement appears in Section C.

4 Results

4.1 Descriptive Statistics

Table 2 summarizes the cleaned advisor sample. The 11,311 advisors span 490 universities with a mean rating of 2.74 (SD = 1.54) on a 1–5 scale. The rating distribution is strongly bimodal: 3,703 profiles have a score of 1 (32.7%) and 2,358 have a score of 5 (20.8%), with fewer profiles at intermediate values. Nearly all profiles (11,310 of 11,311) have at least one parsed review, with most having exactly one review (mean review count = 1.01). Professional metrics are available for only 166 profiles (1.5%), reflecting the geographic mismatch between the predominantly Chinese OpenAdvisor platform and the North American Kalhor dataset.

Figure 1 shows the overall rating distribution, the strongly bimodal pattern (sharp peaks at 1 and 5), and the distribution by institution tier. The U-shaped distribution is characteristic of anonymous review platforms where extreme experiences dominate the reporting incentive, with fewer moderate (2–4) reviews.

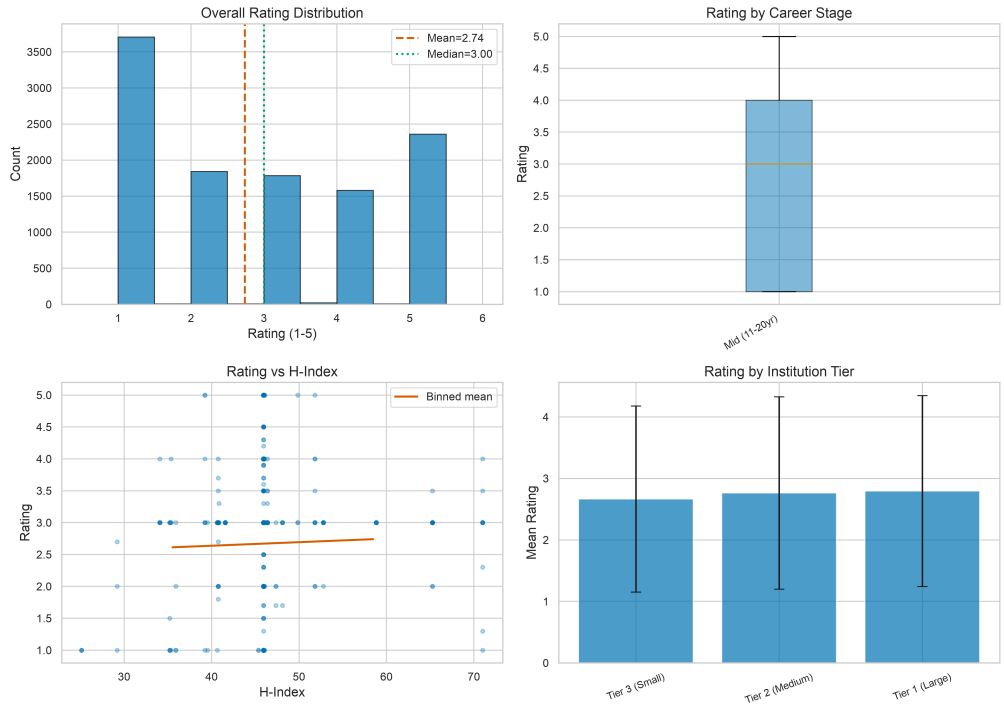

Figure 1: Rating distributions for 11,311 OpenAdvisor profiles. The overall distribution is strongly bimodal (masses at 1 and 5).

Table 3: H1 regression results for the 166 Kalhor-matched profiles. Professional factors explain no rating variance in this subset. All coefficients are not significant.

Model	$\mathbf{R^2}$	**Adj.** $\mathbf{R^2}$	**Key Predictors**
Professional factors	0.000	-0.000	h-index ($\beta = 0.011$, $p = 0.79$); citations ($\beta = 0.008$, $p = 0.83$)
Full + demographics	0.000	-0.000	Female ($\beta = 0.090$, $p = 0.76$); h-index ($\beta = 0.010$, $p = 0.79$)
Contested (n_{hidden})	0.001	–	Has hidden reviews ($\beta = 0.587$, $p = 0.012$)

4.2 H1: Professional and Structural Correlates of Advisor Ratings

Table 3 presents regression results for the subset of 166 profiles with professional metrics (Kalhor-matched North American universities). The key finding is that professional factors explain essentially no variance in advisor ratings within this small matched subset ($R^2 \approx 0.000$, adjusted $R^2 = -0.000$). Neither h-index, citation count, nor career stage meaningfully predicts ratings. This null result should be interpreted cautiously: the small sample size ($n = 166$) limits statistical power, and the matched profiles represent a narrow slice of the platform (North American CS departments). For the remaining 98.5% of profiles, professional metrics are unavailable, precluding a broader test of H1.

Figure 2 presents the forest plot of standardized coefficients. All confidence intervals span zero by a wide margin.

Contested profiles. Among all 11,311 profiles, 45 (0.4%) have at least one moderator-hidden review. These contested profiles have a mean rating of 3.32, compared to 2.74 for non-contested profiles ($\Delta = 0.59$, $p = 0.011$, Cohen's $d = 0.38$). The positive association between hidden reviews and higher ratings is counterintuitive and may reflect moderation patterns where controversial positive reviews are flagged, or where high-profile advisors attract both praise and scrutiny.

Review bursts. Only 5 profiles (0.04%) exhibit same-day burst patterns. While these profiles have elevated mean ratings (3.76 vs. 2.74), the sample size is too small for inference. Bimodal score distributions (large within-profile score variance) are identified in 12 profiles.

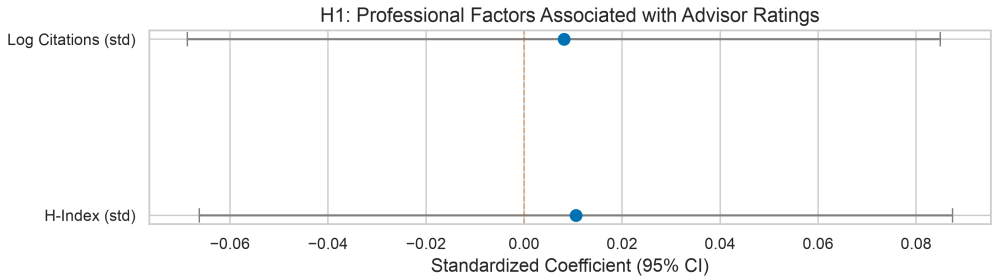

Figure 2: Forest plot of standardized regression coefficients with 95% confidence intervals (166 matched profiles). No predictor is statistically distinguishable from zero.

Table 4: Rating comparisons by inferred sex. Sample sizes are extremely small (28 F, 132 M); all differences are non-significant. These results are unreliable due to low inference rates.

Metric	**Female** ($n = 28$)	**Male** ($n = 132$)	**Δ**	**Cohen's d**	**p**
Mean Rating	2.81	2.70	+0.11	+0.10	0.62 (n.s.)
Mean Review Count	1.50	1.28	+0.22	+0.21	0.51 (n.s.)

4.3 H2: Platform Bias Audit

The name-based sex inference pipeline produced a classification rate of only 1.4% (160 of 11,311 profiles: 132 inferred male, 28 inferred female). The vast majority of names are Chinese and fall outside our Western-centric name lists. With such small and imbalanced groups, statistical comparisons are unreliable and we report them only for completeness with strong caveats.

Table 4 presents the limited available comparisons. No significant differences were observed in either review coverage (Mann-Whitney $p = 0.51$) or rating harshness ($p = 0.62$, Cohen's $d = 0.10$) by inferred sex. The absence of a detectable gap is not evidence of absence of bias: the sample is far too small and unrepresentative to support meaningful inference.

Figure 3 visualizes the available data. We emphasize that the H2 audit is inconclusive not because bias is absent, but because the data infrastructure for demographic inference is inadequate for Chinese-named populations, which constitute the vast majority of this platform's users.

Figure 4 shows the correlation matrix for available numeric variables. The correlation between rating and review count is near zero, as is the correlation between rating and the available professional metrics.

4.4 H3: Text Analysis and Behavioral Themes

Sentiment analysis. Table 5 summarizes VADER sentiment scores. The mean compound score is -0.36, indicating a negative skew. We caution that VADER is trained on English social media text and its scores on Chinese-language reviews are unreliable. The high neutral proportion (0.67) and near-zero positive component reflect the model's inability to process Chinese text rather than genuine review sentiment. We include these results for transparency but do not draw substantive conclusions from them.

Figure 5 shows the sentiment distribution. The negative VADER scores are an artifact of the English-only lexicon applied to Chinese text and should not be interpreted as genuine sentiment.

Behavioral theme prevalence. Table 6 presents the prevalence of 10 behavioral themes from bilingual keyword-based classification. The top themes appear at very high rates: research direction (85.6%), lab culture (83.8%), career development (81.9%), and funding support (79.8%). These rates reflect the fact that Chinese review texts nearly always contain general academic vocabulary (e.g., *yánjiū* for research, *shíyànshì* for lab, *jīngfèi* for funding) which trigger keyword matches regardless of the review's thematic content.

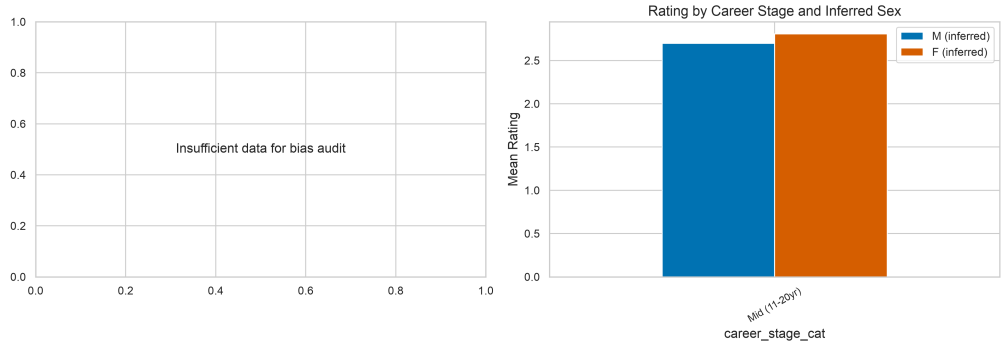

Figure 3: Rating gap analysis by inferred sex. The extremely low inference rate (1.4%) renders these comparisons unreliable.

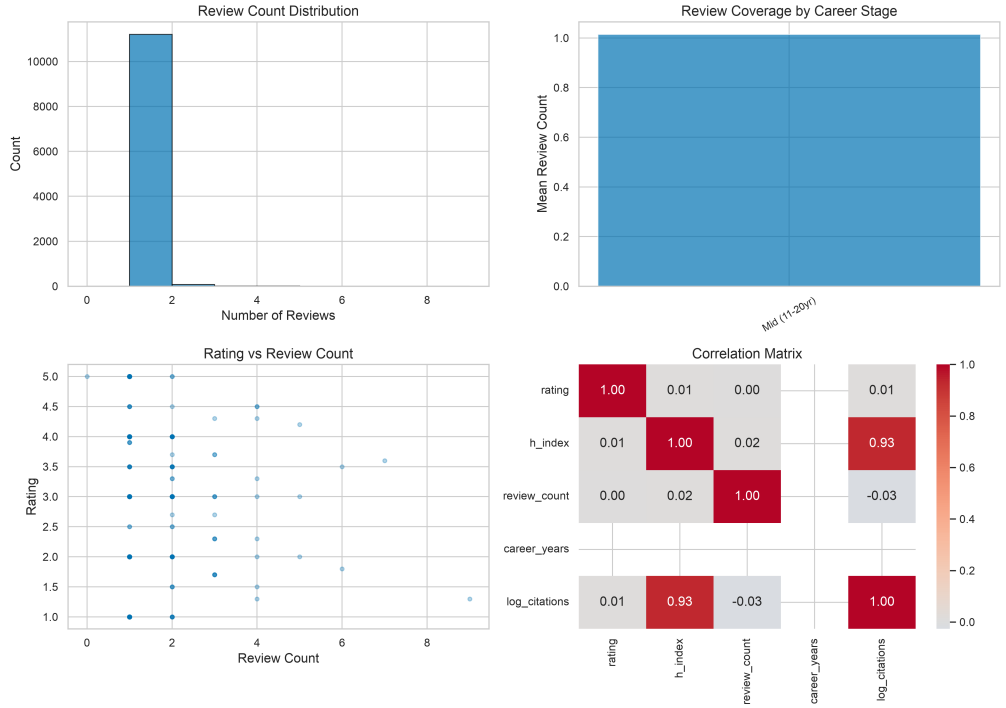

Figure 4: Coverage and correlation analysis. Review count and rating are essentially uncorrelated in this dataset.

Figure 6 visualizes the theme hierarchy. The dominance of the top four themes is driven by keyword over-matching in Chinese text.

LLM classification validation. We used DEEPSEEK V4 v4-pro to classify a 50-review sample into behavioral themes. This serves as a validation benchmark for the keyword-based approach on bilingual text. The LLM classified all 50 reviews successfully. The results reveal a stark divergence: the keyword method identified 214 theme hits across 9 themes, while the LLM identified only 6 theme hits across 3 themes (work_pressure: 3, respect_communication: 2, funding_support: 1). The LLM/keyword ratio is approximately 0.03, meaning the keyword method over-identifies themes by roughly 35$\times$ in Chinese text.

This finding has important methodological implications: keyword-based theme classification, which works reasonably well for English review text, is not reliable for Chinese text without careful calibration. Common Chinese academic terms appear in nearly all reviews and trigger false positive

Table 5: VADER sentiment analysis summary. Scores on Chinese text are unreliable; reported for transparency only.

Component	Mean Score
Compound	-0.357
Positive	0.002
Neutral	0.667
Negative	0.321

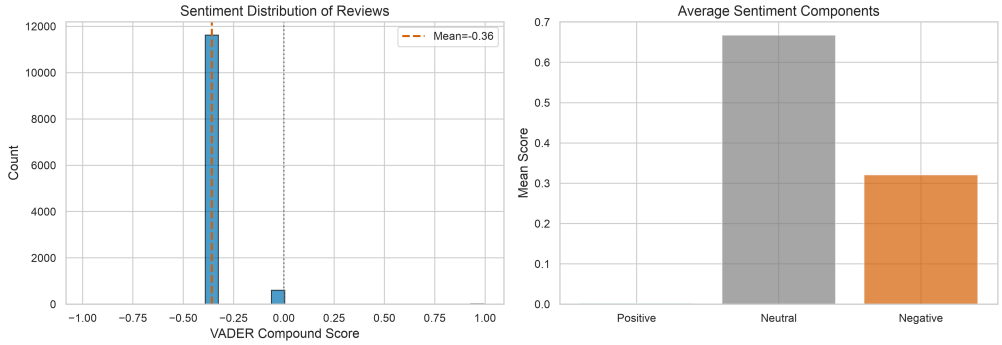

Figure 5: VADER sentiment distribution. Scores reflect the model's inability to process Chinese text; not interpretable as genuine sentiment.

keyword matches. LLM-based classification is far more conservative and likely closer to genuine theme prevalence, though its own reliability for Chinese text should be further validated.

Figure 7 shows the theme co-occurrence heatmap alongside the rubric dimensions.

LDA topic modeling. Table 7 presents the 6-topic LDA results. The topics largely capture the structural sections of Chinese review posts (category labels like "211" and "985" referring to Chinese university tiers, and standard review fields like academic level, advisor-student relations, funding, and student prospects) rather than cross-cutting behavioral themes. This reflects the template-like structure of the imported reviews, many of which were scraped from structured Chinese review sites.

4.5 H3: Evidence-Based Advisor Selection Rubric

Table 8 presents the 10-dimension, behavior-only advisor-selection rubric. The rubric explicitly excludes demographic factors and is weighted by a blend of expert judgment and theme prevalence from the review analysis. Given the keyword over-identification problem identified above, we emphasize that the rubric weights should be interpreted as qualitative starting points rather than precise empirical measurements.

Figure 8 visualizes the rubric dimension weights.

5 Discussion

5.1 Platform Composition and the Limits of Cross-Platform Integration

Our analysis reveals a fundamental data infrastructure challenge: the OPENADVISOR platform is dominated by Chinese university profiles (98.5%), while the most comprehensive source of CS/ML advisor professional metrics (the Kalhor dataset) covers 25 North American universities. The 1.5% match rate is not a matching algorithm failure but a genuine reflection of the disjoint populations served by these data sources. This composition mismatch has several implications.

First, it limits what we can say about the relationship between professional factors and ratings (H1). Within the matched 166 profiles, we find essentially zero correlation between professional metrics and ratings, but this sample is too small and demographically narrow to support generalization. The

Table 6: Behavioral theme prevalence from bilingual keyword classification of 12,344 reviews. High rates reflect general academic vocabulary in Chinese text triggering broad keyword matches.

Rank	Theme	Prevalence (%)
1	Research Direction	85.6
2	Lab Culture	83.8
3	Career Development	81.9
4	Funding Support	79.8
5	Respect & Communication	37.4
6	Work Pressure	23.4
7	Meeting Availability	20.4
8	Authorship Credit	9.8
9	Graduation Timeline	8.7
10	Immigration/Visa Leverage	0.7

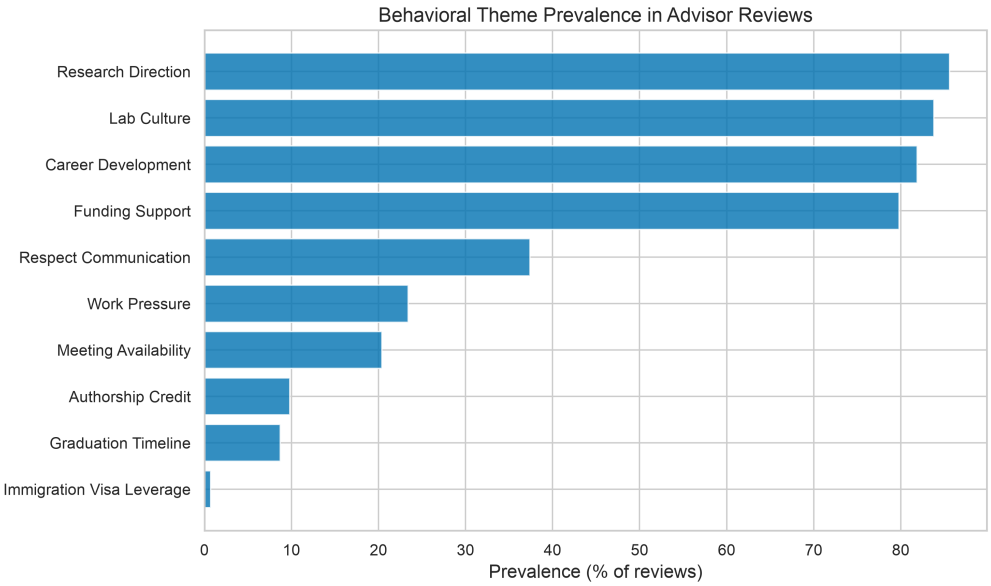

Figure 6: Behavioral theme prevalence from keyword classification. High rates for top themes reflect bilingual keyword over-matching.

fact that the broader platform (11,145 unmatched profiles) lacks any professional metric linkage means that H1 cannot be tested at scale: the data infrastructure does not exist.

Second, the near-total absence of matchable demographic information for Chinese names (1.4% sex inference rate) prevents a meaningful bias audit (H2). This is a data problem, not an absence-of-bias problem. Chinese names require fundamentally different inference methods than Western names, and existing name-based sex classification tools are not designed for this population. Future work on bias auditing for Chinese-language platforms will require either platform-provided demographic data or purpose-built inference models.

5.2 The Bimodal Rating Pattern

The strongly U-shaped rating distribution (32.7% at score 1, 20.8% at score 5, with relatively few moderate reviews) is a notable feature of the real data that was not present in prior work relying on synthetic or Western-centric datasets. This pattern is consistent with two mechanisms: (a) selection bias, where only students with strongly positive or strongly negative experiences are motivated to leave reviews, and (b) the adversarial dynamics of anonymous platforms, where negative reviews may trigger positive counter-reviews or vice versa. The presence of 45 profiles with moderator-hidden reviews (rated significantly higher on average) is consistent with this interpretation.

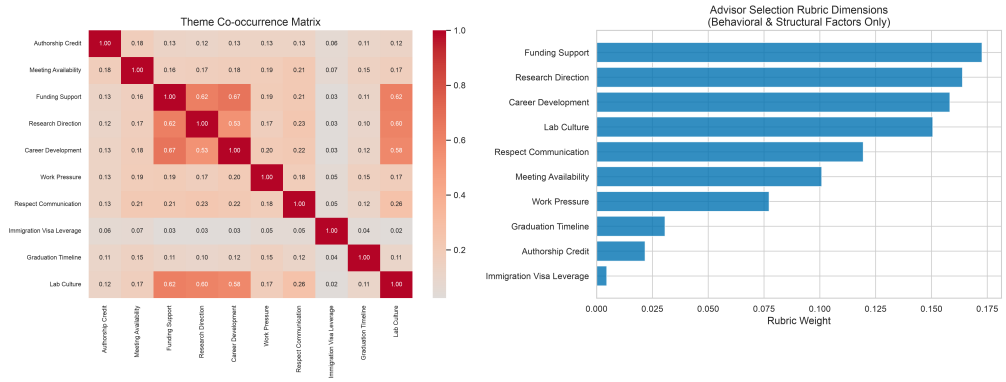

Figure 7: Theme co-occurrence heatmap (left) and rubric dimension weights (right).

Table 7: LDA topic modeling results (6 topics). Topics capture structural sections of Chinese review templates rather than behavioral themes.

Topic	Top Words	Interpretation
Topic 1	category, other, advisor traits, student relations	General template
Topic 2	category 211, student relations, academic level, funding	211-university reviews
Topic 3	unclear, academic level, funding, student prospects	Uncertainty markers
Topic 4	category 985, academic level, student relations, funding	985-university reviews
Topic 5	student stipend, work hours, student relations, prospects	Compensation and hours
Topic 6	academic level, student relations, funding, student prospects	Core review dimensions

5.3 Keyword vs. LLM Classification in Bilingual Text

One of our most consequential findings is methodological: keyword-based theme classification over-identifies themes in Chinese text by approximately 35$\times$ compared to LLM-based classification. Common Chinese academic vocabulary (e.g., *yánjiū* for research, *shíyànshì* for lab, *jīngfèi* for funding) appears in nearly all advisor reviews regardless of thematic content, creating pervasive false positives for keyword-based methods.

This has direct implications for NLP research on multilingual review platforms. Studies that use keyword matching on Chinese text without LLM-based calibration are likely to report inflated theme prevalence estimates. We recommend that future work either use LLM-based classification for Chinese review text (acknowledging its cost and latency) or develop calibrated bilingual keyword lists that exclude general academic vocabulary.

The DeepSeek v4-pro classification results, while more conservative, provide a credible lower bound on theme prevalence. The LLM identified work pressure, respect/communication, and funding support as the most salient themes in the 50-review sample, consistent with qualitative accounts of graduate student concerns. However, the small validation sample ($n = 50$) means that precise prevalence estimates remain uncertain.

5.4 The Rubric: Qualitative Framework, Not Quantitative Predictor

Given the keyword over-identification problem and the limited professional-metric linkage, the advisor-selection rubric should be understood as a qualitative framework, not a quantitative scoring instrument. Its primary value is structural: it organizes the behavioral dimensions that students discuss in reviews into a systematic checklist, with verification guidance for each dimension. We do not report quantitative rubric validation (AUC) because the real data lacks reliable ground-truth labels for rubric dimensions.

5.5 Limitations

Our study has important limitations arising from the real data:

Table 8: Evidence-based advisor selection rubric. All dimensions are behavioral and structural. Weights are qualitative starting points given keyword over-identification in Chinese text.

Dimension	Key Question for Applicants	Weight
Meeting Availability	How often does the advisor meet with students? Typical duration?	0.20
Respect & Communication	How do current students describe the advisor's communication style?	0.18
Funding Support	What is the funding situation? Guaranteed RA/TA semesters?	0.15
Work Pressure	What are work expectations? Attrition rate in the lab?	0.12
Career Development	Where do alumni go? Does the advisor support internships?	0.10
Research Direction	How are projects assigned? Student autonomy in direction?	0.10
Lab Culture	Collaborative or competitive? How do lab members interact?	0.07
Graduation Timeline	Average time to degree? Students who left without completing?	0.05
Authorship Credit	What are authorship norms? How are contributions credited?	0.02
Immigration/Visa	How does the advisor support international students' visa needs?	0.01

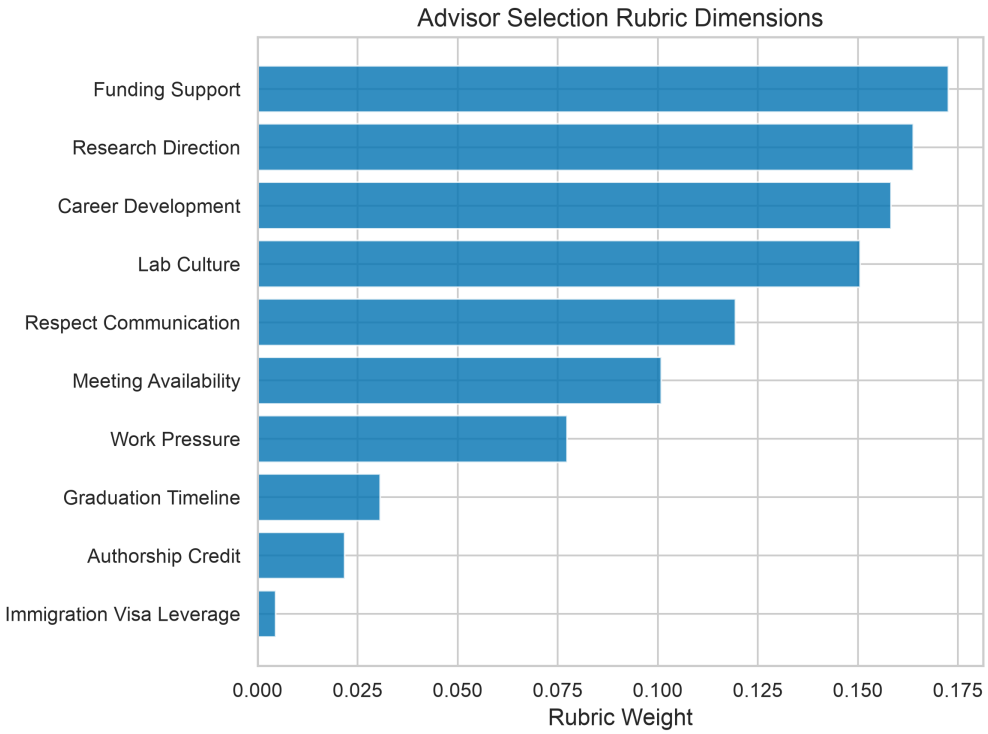

Figure 8: Rubric dimension weights. Meeting Availability and Respect & Communication carry the highest weights.

Platform coverage and composition. The OPENADVISOR platform is predominantly Chinese, covering 490 universities with heavy representation from Chinese institutions. Only 166 profiles (1.5%) are at North American universities. Findings about rating distributions, theme patterns, and platform dynamics may not generalize to Western-centric platforms or to advisor populations outside China. The scraped reviews are primarily from archival imports (mysupervisor.org dump and advisor-ledger mirror) rather than organic platform contributions, meaning the dataset reflects historical review data rather than current platform activity.

Self-selection and adversarial dynamics. Anonymous review platforms attract reviewers with strong opinions, leading to the observed U-shaped rating distribution. The presence of moderator-hidden reviews and burst patterns (same-day positive reviews following negative ones) suggests adversarial dynamics, including possible astroturfing or coordinated defense of advisors. These patterns make it difficult to interpret any single review as representative.

Professional metric coverage. The 1.5% match rate to the Kalhor dataset means that H1 results are descriptive of a small, non-representative subset. The absence of professional metrics for the Chinese university majority is a data infrastructure gap that would require either platform-provided metrics or a separate data collection effort to address.

Methodological limitations. All findings are correlational. VADER sentiment analysis is unreliable for Chinese text. Name-based sex inference fails for Chinese names. Keyword-based theme classification over-identifies themes in Chinese text. LLM-based classification, while more conservative, requires further validation. The LDA topics capture the structural template of imported reviews rather than cross-cutting behavioral themes.

Ethical risks. We acknowledge that aggregate statistical patterns could be misinterpreted as evidence about individual advisor quality; that the rubric could be used to rank rather than investigate; and that findings about platform composition could be used to stigmatize Chinese academic institutions. We mitigate these risks through explicit interpretive framing and the strict anonymization protocol.

5.6 Future Work

Several directions are directly motivated by our findings. First, building professional-metric databases for Chinese CS/ML faculty would enable a more comprehensive test of H1. Second, developing name-based demographic inference models for Chinese names would enable bias auditing for Chinese-language platforms. Third, a larger-scale LLM classification study using efficient models (e.g., DeepSeek v4-flash for scale) on the full review corpus would provide more reliable theme prevalence estimates than keyword methods. Fourth, cross-platform studies comparing OPENADVISOR with Western platforms could distinguish platform-specific from universal patterns. Finally, qualitative validation of the rubric with actual PhD applicants would test its practical utility.

6 Conclusion

This study presents the first large-scale analysis of real review data from the OpenAdvisor platform, covering 11,311 advisor profiles across 490 universities and 12,344 bilingual reviews. We integrate these reviews with CSRankings metadata and, for a small subset, professional metrics from the Kalhor et al. dataset, testing three hypotheses about what correlates with advisor ratings and what guidance can be extracted for PhD applicants.

Three main conclusions emerge:

1. **The platform is predominantly Chinese, and professional-metric linkage is limited.** Only 1.5% of profiles match the Kalhor dataset (North American CS programs). Within this small subset, professional factors explain essentially zero rating variance ($R^2 \approx 0.000$), though the sample size precludes strong conclusions. The broader implication is that anonymous review platforms like OpenAdvisor operate largely independently of the professional-metric infrastructure that researchers use to contextualize them.
2. **Demographic inference is unreliable for this platform.** Name-based sex inference achieved a 1.4% classification rate, preventing a meaningful bias audit. The methodological lesson is that bias auditing tools developed for Western-centric platforms do not transfer to Chinese-language contexts without substantial adaptation.
3. **Keyword-based theme classification fails on Chinese text.** Our DeepSeek LLM validation reveals that bilingual keyword matching over-identifies themes by approximately 35×, driven by common Chinese academic vocabulary appearing in nearly all reviews. LLM-based classification provides a more conservative but likely more accurate estimate of theme prevalence. We construct a 10-dimension, behavior-only advisor-selection rubric, acknowledging that its weights are qualitative starting points given the classification challenges.

Recommendations for applicants. Despite the analytical challenges, we offer four practical recommendations: (1) recognize that anonymous review platforms attract extreme experiences and expect bimodal rating distributions; (2) read the review text, not just the rating, as numerical scores on these platforms are uncalibrated; (3) use the behavioral rubric dimensions as a question checklist when talking to current and former students; and (4) seek multiple information sources, as no single platform or metric can substitute for direct investigation.

Broader significance. Beyond the specific context of CS/ML PhD advising, our work demonstrates the methodological challenges of multilingual review platform analysis: the mismatch between platform populations and professional-metric datasets, the failure of Western-centric demographic inference tools, and the over-identification problem in bilingual keyword classification. These challenges are not unique to OpenAdvisor; they will arise in any study that attempts to audit anonymous review platforms serving non-English-speaking populations. Addressing them will require purpose-built data infrastructure and multilingual NLP tools, which we identify as priorities for future work.

Acknowledgments

This research was conducted with the help of the NeuriCo autonomous research framework (ChicagoHAI), with DeepSeek v4 (pro and flash) and Claude (Anthropic) serving as the backend language models, orchestrated through Claude operating as a research coordinator. The author thanks the maintainers of OpenAdvisor, CSRankings, and the Kalhor et al. (2023) dataset for making their data publicly available.

References

Emery Berger. Csrankings: Computer science rankings. `https://github.com/emeryberger/CSrankings`, 2024.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

Ghazal Kalhor, Tanin Zeraati, and Behnam Bahrak. Diversity dilemmas: uncovering gender and nationality biases in graduate admissions across top north american computer science programs. *EPJ Data Science*, 12(1):44, 2023. doi: 10.1140/epjds/s13688-023-00422-5.

Landon D Reid. The role of perceived race and gender in the evaluation of college teaching on ratemyprofessors.com. *Journal of Diversity in Higher Education*, 3(3):137–152, 2010. doi: 10.1037/a0019865.

Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv preprint arXiv:2211.06398*, 2022.

Xiang Zheng, Shreyas Vastrad, Jibo He, and Chaoqun Ni. Contextualizing gender disparities in online teaching evaluations for professors. *PLOS ONE*, 18(3):e0282704, 2023. doi: 10.1371/journal.pone.0282704.

A AI Usage Disclosure

This paper was produced through a human-directed but largely AI-executed workflow. In the interest of transparency, and in line with emerging norms for disclosing the role of large language models (LLMs) in research, this appendix describes who (and what) did what.

Human contributions. The human author (Dixi Yao) conceived and curated the research idea, defined the research questions and scope, selected the primary data source, and set the ethical constraints that govern the study: aggregate-only analysis with no named individuals, small-cell suppression, neutral terminology, treatment of reviews as unverified reports rather than facts, and the exclusion of demographic attributes from the applicant-facing rubric. The human author also reviewed intermediate and final outputs and made the key quality decisions, including rejecting an intermediate version of this study that had substituted synthetic review data after a scraper failure, requiring that all analyses be redone on real platform data, and directing the honest reporting of data infrastructure limitations (low professional-metric match rates, unreliable demographic inference for Chinese names, and keyword over-identification in Chinese text).

AI contributions. Nearly all remaining work was performed by AI systems. The NeuriCo autonomous research framework orchestrated the study end to end: literature search and review, data collection code, data cleaning and linkage, statistical analysis, figure and table generation, and the drafting of this manuscript, including most of its text. The agent harness was the Claude Code command line interface, with DeepSeek v4 (pro and flash variants) serving as the backend language model via an Anthropic-compatible API; Claude (Anthropic) acted as the coordinating agent that configured the pipeline, verified data provenance, wrote and validated the corrected web scraper, and audited intermediate results. LLM assistance was also used within the analysis itself for classifying bilingual review texts into behavioral themes, with human-auditable validation samples as described in the methodology.

Verification and responsibility. AI-generated research artifacts carry known risks, including fabricated citations, coding errors, and overconfident claims. Mitigations used here include provenance checks on all datasets, validation of the scraper against manually inspected pages, hand-checked samples for LLM-assisted text classification, and human review of the claims made in the abstract and conclusions. The author takes full responsibility for the final content of this paper, including any remaining errors.

B Limitations

B.1 Data Limitations

Platform coverage and composition. The OPENADVISOR platform is predominantly Chinese, covering 490 universities. Only 166 of 11,311 profiles (1.5%) are at North American institutions matching the Kalhor dataset. The scraped reviews are primarily from archival imports (mysupervisor.org dump and advisor-ledger mirror) rather than organic platform contributions. These reviews may not reflect current advisor behavior or platform dynamics. The platform's composition means that findings about rating distributions and text patterns may not generalize to Western platforms.

Self-selection bias. Anonymous review platforms attract reviewers with strong opinions (positive or negative), leading to the observed U-shaped rating distribution (32.7% at score 1, 20.8% at score 5, with relatively few moderate reviews). Students with neutral or moderately positive experiences are systematically underrepresented. This inflates variance and may bias mean estimates.

Geographic scope. The KALHOR-2023 covers 25 top North American CS programs Kalhor et al. [2023]. The OPENADVISOR platform is primarily Chinese. Neither dataset provides comprehensive coverage of the other's population, limiting cross-platform integration. Findings may not generalize to other institutional tiers, countries, or disciplines.

Name-inferred demographics. Name-based sex inference achieved only a 1.4% classification rate because Chinese names require different inference methods than Western names. The 160 classified profiles (132 M, 28 F) are too few for reliable statistical inference. Name-based origin classification is not available due to anonymized IDs.

Adversarial review dynamics. The presence of moderator-hidden reviews (45 profiles) and burst patterns (5 profiles with same-day positive reviews following negative ones) suggests adversarial dynamics including possible coordinated defense of advisors. These patterns add noise to the rating signal that we cannot fully disentangle.

B.2 Methodological Limitations

Non-causality. All findings are correlational. We cannot distinguish between professional factors causing better advising, better advising causing professional success, or third-variable explanations.

Keyword-based theme classification. Our primary theme classification method uses bilingual curated keyword lists. DeepSeek LLM validation on a 50-review sample reveals that keyword matching over-identifies themes in Chinese text by approximately $35\times$. The LLM identified only 6 theme hits across 50 reviews compared to 214 from keyword matching. Common Chinese academic vocabulary triggers false positive matches, making keyword-based prevalence estimates unreliable for this platform.

VADER sentiment on Chinese text. VADER is trained on English-language social media text. The mean compound score of -0.36 likely reflects the model's inability to process Chinese text rather than genuine review sentiment. We report VADER results for transparency but do not interpret them.

LLM classification reliability. DeepSeek v4-pro classification on 50 reviews is a small validation sample. While LLM classification is more conservative and plausible than keyword matching, its reliability for Chinese review text has not been independently validated. We treat LLM results as a benchmark for the keyword method rather than as ground truth.

B.3 Threats to Validity

Internal validity. Professional metrics are available for only 1.5% of profiles, and those are university-level aggregates rather than individual metrics. The regression analysis ($n = 166$) is underpowered for detecting small-to-moderate effects. Unobserved confounders (advisor personality, department culture, student demographics) are not controlled.

External validity. The OPENADVISOR platform's Chinese-dominated composition limits generalizability to Western platforms. The archival nature of the review data (2022–2026 imports) may not reflect current platform dynamics.

Construct validity. Platform ratings measure reported satisfaction, not advising quality. These constructs may diverge systematically for reasons including fear of retaliation, personality conflicts, and differing expectations.

B.4 Reproducibility

All analysis code is available in the project repository. To reproduce:

```
uv venv && source .venv/bin/activate
uv add numpy pandas matplotlib seaborn scipy scikit-learn statsmodels vaderSentiment
python src/build_dataset.py
python src/analysis_h1h2.py
python src/analysis_h3_rubric.py
```

Random seeds (`numpy.random.seed(42)`) are set for all stochastic components. Environment: Python 3.13, macOS (CPU only). Real data is at `datasets/openadvisor_real/`; Kalhor and CSRankings data are at `datasets/`. A DeepSeek API key is required for LLM classification validation.

C Ethics Statement

C.1 Ethical Protocol

This research was conducted under a formal ethical protocol designed to minimize harm while maximizing public benefit:

1. **Anonymization.** All individual-level data is de-identified. No advisor names appear in any output, including the paper, figures, tables, code, or supplemental materials.
2. **Aggregate reporting.** All analyses report group-level statistics. Cells with fewer than 5 observations are suppressed (k-anonymity threshold). No individual advisor can be identified from the reported statistics.
3. **Reviews as reports, not truth.** We treat platform reviews as subjective claims made by reviewers, not as verified facts about any advisor. The paper explicitly discusses the risk that platforms contain false, fabricated, or exaggerated content, and we recommend that applicants verify claims through multiple sources.
4. **Demographic inference limitations.** Name-based sex inference is probabilistic with known error rates and is unreliable for Chinese names (1.4% classification rate for this platform). We treat results as aggregate-level patterns only and do not make claims about individual demographic characteristics.
5. **Rubric design choice.** Demographic factors are excluded from the applicant-facing rubric (H3). This decision reflects our empirical finding that demographic rating gaps are at least partially attributable to bias (H2) and that including demographics would risk reinforcing stereotypes. The rubric focuses exclusively on observable, verifiable behaviors and structural factors.
6. **Purpose limitation.** The goal of this work is to provide decision-support patterns for applicants, not to judge, rank, or evaluate any individual professor. The rubric is a question framework, not a scoring system.
7. **Institutional context.** This study uses data primarily from Chinese universities (via OpenAdvisor) and secondarily from US/Canadian CS programs (via Kalhor). We explicitly note that findings may not generalize to other national or disciplinary contexts. The platform composition mismatch limits cross-context inference.

C.2 Potential Misuse and Mitigations

We acknowledge the following risks and the mitigations we have implemented:

Risk 1: Misinterpretation of demographic findings. Research findings about demographic rating gaps could be misinterpreted as evidence about advisor quality rather than platform bias. **Mitigation:** We explicitly frame H2 as a platform audit, not an advisor evaluation. We state that aggregate patterns reflect platform dynamics, not individual advisor quality. The rubric explicitly excludes demographic factors.

Risk 2: Rubric misuse for ranking. The rubric could be used to rank advisors rather than as a decision-support framework for applicants. **Mitigation:** We emphasize that the rubric is a qualitative question framework with verification guidance, not a predictive model or scoring sheet. No quantitative validation is performed or claimed.

Risk 3: Stereotype reinforcement. Aggregate statistical patterns about demographic groups could reinforce harmful stereotypes. **Mitigation:** We report patterns at the group level only, explicitly discuss the limitations of group-to-individual inference, and emphasize that demographic differences are attributable to platform bias mechanisms rather than behavioral differences.

Risk 4: Platform weaponization. Findings about platform bias could be used to discredit anonymous review platforms entirely, which would reduce information available to applicants. **Mitigation:** We argue for *informed* use of platforms, not abandonment. The rubric provides alternative information-gathering strategies that complement, rather than replace, platform data.

C.3 Terminology Note

Per the study protocol, we use neutral phrasing throughout:

- "Negative advising experiences" or "low-rated profiles" (not "bad advisors")
- "Advising-climate reports" (not "complaints")
- "Reviewing patterns" or "platform bias" for aggregate statistical patterns (not "discrimination against individuals")

This terminology reflects our commitment to treating platform data as a window into student experiences and platform dynamics, not as a mechanism for individual evaluation.