Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

RIMNI

ARTICLE

IMA: An Interpretable Machine Learning Framework Based on Large-Scale Anonymous Evaluations to Decode the Black Box of Student-Supervisor Mentorships

Yang Gao[1,2], Weiqiang Jin[1,3,4], Yajuan Nan[3], Yue Ma[4], Biao Zhao[1,3] and Ziwei Zhang[2,*]

[1]School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China
[2]Institute of SRIICL, Xi'an Jiaotong University, Xi'an, China
[3]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[4]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
*Corresponding Author: Ziwei Zhang. Email: ziwei.zhang@xjtu.edu.cn

Received: 11 February 2026; Accepted: 26 March 2026

ABSTRACT: Student-supervisor relationships (SSR) play a central role in postgraduate training, academic development, and research well-being. In the context of the rapid expansion of Chinese graduate education, understanding mentorship quality has become increasingly important for both educational governance and student development. However, prior SSR studies have often relied on small-scale surveys, context-specific qualitative evidence, or linear analytical approaches, which limits their ability to capture heterogeneous, non-linear, and system-level patterns across institutions and disciplines. To address this gap, we propose *Interpretable Mentorship Analytics* (**IMA**), a scalable analytical framework built on large-scale anonymous student evaluations. First, we construct a multi-platform dataset of anonymous supervisor evaluations and transform unstructured review text into structured mentorship indicators through a preprocessing and LLM-assisted feature engineering pipeline. Second, we employ gradient-boosting models, including XGBoost, LightGBM, and Gradient Boosting, to model the relationship between mentorship-related features and overall evaluation outcomes. Third, we apply explainable machine learning methods, particularly SHAP and LIME, to identify global feature importance, local decision patterns, and non-linear interactions among mentorship dimensions. The results show that IMA can effectively uncover the key drivers of mentorship satisfaction, especially the central role of teacher-student relationship quality, while also revealing substantial heterogeneity across regions, institution types, and disciplines. By combining large-scale anonymous evaluations with interpretable predictive modeling, this study provides a transparent and data-driven framework for understanding SSR and offers empirical evidence for improving postgraduate supervision and educational policy.

KEYWORDS: Student-Supervisor Relationship (SSR); Anonymous Faculty Evaluations; Machine Learning (ML); Explainable Data Analysis; Chinese Higher education; Shapley Additive exPlanations (SHAP); Local Interpretable Model-agnostic Explanations (LIME)

1 Introduction

With the rapid growth of graduate education as the primary engine for cultivating high-level researchers, it occupies a distinct and critical position within the academic ecosystem that differs fundamentally from the undergraduate experience [1,2]. In this advanced stage, the educational focus shifts from broad curricular instruction to specialized, personalized research training. Beyond the mere

transmission of knowledge, student-supervisor relationship (SSR) fundamentally shapes a student's development of research interests and career paths [3–5].

Substantial related research indicates that positive SSR are strongly correlated with timely doctoral degree completion and higher research achievements [6–8]. For faculty, engaging in effective mentorship can enhance their own professional satisfaction and leadership skills, creating a reciprocal cycle of academic development [5,9]. Moreover, from a systemic perspective, high-quality mentorship strengthens the entire research ecosystem by promoting innovation, sustaining academic values, and building an inclusive and resilient scholarly community [10,11]. In summary, a constructive SSR not only fosters psychological well-being but also significantly enhances research productivity and academic persistence [12].

The Chinese mentorship dynamic is evolving rapidly under the 'high-quality talent development' national policies [13]. To accelerate the reform and development of graduate education, China has witnessed a substantial expansion in graduate enrollment alongside significant increases in government research funding[1]. Over the past decade (2014 to 2023), China's annual enrollment of new Master's students grew markedly from approximately 540,000 to nearly 1.1 million. Simultaneously, doctoral student enrollment also experienced robust growth, rising from around 70,000 to about 150,000[2]. These changes have democratized access to postgraduate education [8,12], allowing a growing students number to transition from undergraduate studies to advanced research periods, thereby supplying the human capital necessary for the nation's innovation-driven growth.

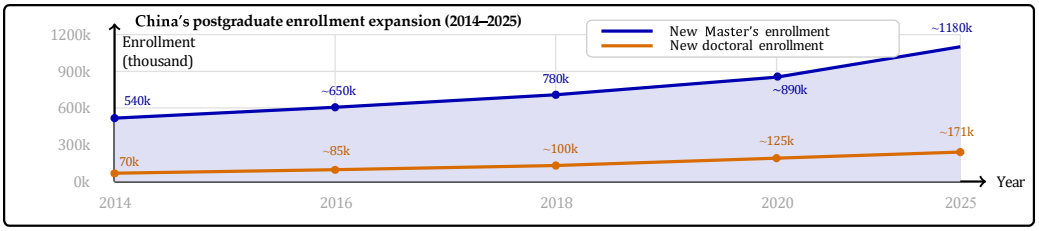

Figure 1: China's postgraduate education expansion and its relationship to mentorship. National enrollment of Master's & doctoral students increases markedly from 2014 to 2025 (per year), reported by official statistics[3], reflecting the rapid scale-up of graduate education.

Fig. 1 shows that the growth is not incremental; it is sustained over time and reaches a new scale by the mid-2020s. As cohorts become larger, effective supervision becomes harder to deliver through informal effort alone, and universities must rely more on the quality and stability of SSR to maintain research training standards.

However, this rapid expansion has created serious challenges for the sustainability of the academic environment. A key issue lies in the unique power dynamics, i.e., a '*one-vote veto*' power, where supervisors often hold sole authority to decide a student's graduation [14]. A fundamental structural divergence exists between undergraduate education, an open, institutionalized system safeguarded by curricula and collective assessment, and postgraduate education, which whereas relies heavily on a supervisor's personal decisions

---

[1] This expansion is driven by explicit national policies issued by *China's Ministry of Education*: http://www.moe.gov.cn/srcsite/A22/s7065/202009/t20200921_489271.html.

[2] These statistics are sourced from the *National Education Statistics Bulletin* issued by the Ministry of Education of China, accessible at: http://www.moe.gov.cn/srcsite/A03/s180/moe_633/201508/t20150811_199589.html.

[15]. As a result, the graduated mentorship becomes a 'black box' with highly concentrated power and a lack of oversight [16].

Table 1: Representative methods of SSR research and their corresponding key limitations.

Methods	Typical approach and focus	Main limitation for system-level governance
Survey evidence in China [12,17]	Longitudinal or cross-sectional surveys linking SSR qualities (e.g., reciprocity, trust) to well-being and experience	Limited coverage across institutions and labs; sensitive responses may be biased
National large-scale surveys [8]	Macro-models connecting support, resources, satisfaction, and productivity	Often emphasise average linear effects; limited visibility into non-linear patterns
Measurement instruments [15]	Validated questionnaires for supervisory interaction and alliance	Provide measurement, but not a scalable analytical mechanism for complex feature interplay
Qualitative typologies [16]	Interviews and thematic analyses describing supervision styles and perceived impacts	Strong interpretability but lacking scalability; difficult to generalize across settings
Social network analysis [18,19]	Mapping supervisory ties and lab-level relational structures to identify central actors and support flows	Captures structure, but governance implications remain indirect and context-dependent
Experience sampling / diary studies [20,21]	High-frequency tracking of day-to-day supervisory interactions and emotional responses over time	Rich temporal detail, but costly and difficult to scale across institutions
Computational text and digital trace analysis [22,23]	Mining emails, feedback text, or platform interactions to detect relational patterns and power asymmetries	Raises privacy concerns; difficult to validate and translate into actionable governance

However, evaluating mentorship relationships remains a challenging task in practice. Traditional evaluation methods often rely on limited surveys or qualitative feedback, which may not fully capture the complexity and variability of mentorship interactions across different institutions and disciplines. Although the education and higher-education literature has extensively discussed SSR, as shown in Table 1, most studies still rely on qualitative interviews, small-sample questionnaires. Alternatively, these studies are context-specific field investigations that focus on broadly describing supervisors' behaviors, communication patterns, and relational types. For example, some studies use qualitative designs to examine how supervisory communication affects doctoral development, but their samples are typically limited and their perspectives are constrained by local settings [9]. Other work relies on regional questionnaire surveys to characterize interaction features, yet the resulting conclusions are difficult to generalize to wider and more diverse educational environments [13]. Moreover, while many empirical studies [4] report associations between mentorship quality and outcomes such as research motivation and learning engagement, these analyses often use conventional statistical models.

Also, conventional questionnaire data are vulnerable to social-desirability pressures and self-protective responding during data collection, especially when the topic is sensitive and respondents worry about reputational or third-party consequences [24]. A large methodological literature shows that respondents may underreport undesirable experiences, overreport desirable ones, or avoid disclosure altogether [25]. Therefore, existing research [2,26] has not yet established a systematic framework that can operate on large-scale, multi-dimensional data to uncover hidden mechanisms and key drivers in mentorship dynamics.

To address these limitations, we propose an interpretable machine-learning framework for reseaching SSR, *Interpretable Mentorship Analytics* (IMA). We leverage large-scale anonymous evaluations and explainable models to discover implicit relationships and non-linear patterns in mentorship data, and we quantify how specific features contribute to overall evaluation outcomes. We use gradient-boosting models [27–29] to learn predictive mappings between multi-dimensional evaluation features and overall mentorship outcomes, and we then apply explainable machine learning methods to open the "black box" and quantify how each feature contributes to the predicted results. In particular, we use *SHapley Additive*

exPlanations (SHAP)[4] [30] and *Local Interpretable Model-Agnostic Explanations* (LIME)[5] [31] to provide global and local explanations, which aligns well with the broader movement of explainable AI in educational analytics.

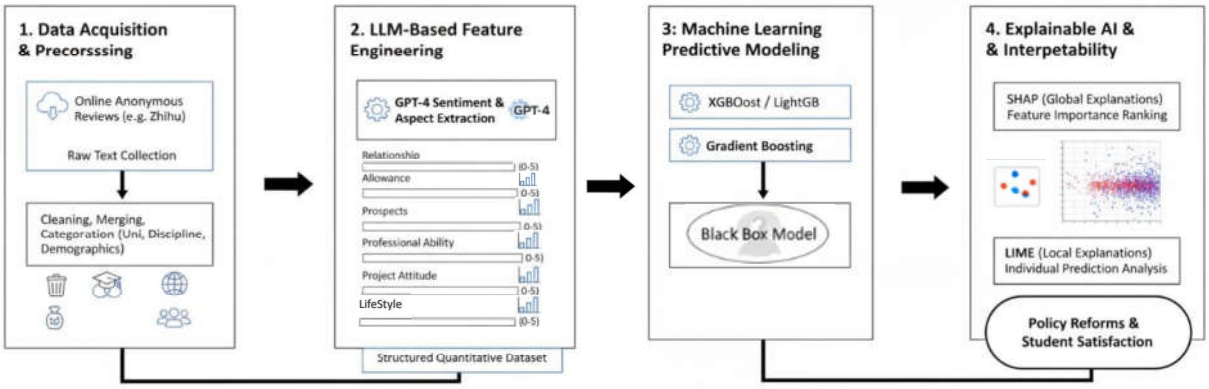

Figure 2: Overview of the proposed IMA workflow, consistsing of four sequential stages. (1) Data acquisition and preprocessing; (2) LLM-based feature engineering; (3) Machine learning predictive modeling; and (4) Explainable AI and interpretability of the black-box models using LIME and SHAP.

Specifically, we construct an end-to-end workflow (as shown in Fig. 2) that integrates data collection, preprocessing, feature extraction, and ML/AI explainable model-based analysis, enabling the investigation of SSR beyond small-sample surveys or linear statistical assumptions. The data are collected from multiple public online supervisor-evaluation platforms[6] and rigorously cleaned by removing duplicates, merging reviews referring to the same mentor, and excluding records outside the Chinese academic system, resulting in a high-quality, multi-institution dataset that spans diverse disciplines and regions.

Based on the structured evaluation indicators, we employ gradient-boosting models, including XGBoost [27], LightGBM [29], Gradient Boosting [28] and so on, to predict overall mentorship outcomes from multi-dimensional features describing guidance intensity, power constraints, and support conditions. These models allow us to capture complex non-linear dependencies among mentorship characteristics that are difficult to identify using conventional approaches. To ensure transparency and analytical interpretability, we further apply SHAP and LIME to decompose model predictions and quantify the contribution of individual features at both global and local levels. The findings of this study are expected to provide practical insights for improving mentorship evaluation systems, supporting institutional decision-making, and promoting more transparent and effective student-supervisor interactions.

Overall, this work, i.e., the proposed IMA framework, makes three contributions.

1. First, we propose IMA as an integrated framework that combines large-scale anonymous evaluation data with interpretable machine learning for SSR analysis.
2. Second, we provide a scalable data pipeline that transforms public anonymous reviews into structured mentorship indicators across institutions, regions, and disciplines.

---

[4] More details about SHAP is available at: https://shap.readthedocs.cn/en/latest/.

[5] More details about LIME is available at: https://lime-ml.readthedocs.io/en/latest/.

[6] Data was collected from various Graduate *Students' Evaluation and Rating of Supervisors* in Chinese universities, such as using the website '*Rate Your Supervisor*' (https://www.ratemyprofessors.com/ [Accessed on 2025.12]). *All Sensitive Personal Information Has Been Anonymized.*

3. Third, we show that explainable predictive modeling can reveal non-linear feature importance, interaction effects, and heterogeneous mentorship patterns beyond what conventional descriptive or linear approaches typically provide.

Methodological Motivation. Prior limitations in SSR research arise from at least three distinct sources [24,32]. First, sensitive supervisor-related questions may induce social desirability bias or self-protective responding during data collection [24]. Second, many existing studies rely on relatively small or context-specific samples, limiting cross-institutional generalizability [10]. Third, even when larger datasets are available, conventional linear or descriptive methods may have limited ability to identify non-linear dependencies, interaction effects, and heterogeneous evaluation mechanisms [11].

In this study, these limitations are addressed through different components of our IMA framework. The use of large-scale anonymous evaluations mainly helps mitigate the first two issues by broadening empirical coverage and reducing response inhibition. The interpretable machine learning component is designed to address the third issue by modeling and explaining complex relationships among mentorship dimensions.

It should be noted that this study adopts a data-driven, quantitative analytical perspective. While mixed-method approaches are valuable for in-depth contextual understanding, they are often constrained in scalability and standardization when applied to large-scale evaluation data. Given the scale and structure of the dataset used in this study, a machine learning-based approach is more suitable for identifying generalizable patterns, modeling complex relationships, and ensuring analytical consistency.

Accordingly, the innovation of IMA lies not in claiming that explainability solves response bias, but in integrating anonymous large-scale data with interpretable predictive modeling into a unified analytical framework. Compared with prior studies on student-supervisor relationships, the novelty of this study does not lie in proposing a completely new machine learning algorithm. Rather, it lies in the integration of four elements that are rarely combined within a single SSR research framework. This study: (i) brings SSR research into a large-scale anonymous evaluation data setting, (ii) establishes a pipeline from raw evaluation information to structured mentorship indicators, (iii) applies explainable non-linear models to identify complex mechanisms and heterogeneous effects, and (iv) connects the findings to governance- and policy-oriented applications.

2 Related Works

Student-supervisor relationships (SSR) have been widely studied in postgraduate education because they are closely linked to doctoral progress, and students' well-being[6,12,32]. Unlike the unidirectional or co-constructive knowledge transmission in undergraduate classes, graduate students face the challenge of independent knowledge creation, where they must navigate intellectual struggles and funding uncertainties often in isolation [15,16]. Thus, the SSR becomes a critical partnership that influences doctoral satisfaction, productivity, and socialization [1,8,17].

Table 1 summarizes representative methods of the SSR research and their limitations. In the Chinese graduate-education ecosystem, SSR has attracted increasing attention due to hierarchical supervision structures and strong advisor-centered authority [2,13]. Several studies have used questionnaire surveys to examine how SSR qualities such as reciprocity, trust, and perceived support relate to students' academic experience and psychological well-being [4,12]. These surveys provide direct evidence from students' perspectives and highlight that mentorship problems are not only individual issues but may reflect broader institutional conditions. These survey designs, however, can be shaped by social desirability and the sensitivity of questions about supervision and authority, which may distort students' self-reports

[24,25]. Beyond single-region studies, large-scale datasets have been used to connect supervisory support, academic resources, satisfaction, and productivity outcomes [8]. Such work is valuable for identifying broad associations at the population level [1,3]. A long-standing line of research has developed validated measurement tools to quantify supervisory interaction and relationship quality. For example, structured instruments have been proposed to assess supervisory behaviors, communication patterns, and alignment between supervisors and doctoral students [15,33]. Best-practice syntheses further consolidate actionable guidance on supervision processes and relationship management [7]. Qualitative research has also contributed important insights by using interviews and thematic analysis to describe supervision styles and students' perceived impacts [9,17]. Such studies are particularly useful for understanding power relations, expectations, and relational dynamics that may not be captured by standard questionnaires [17].

Some researchers have introduced social network analysis to study mentorship as part of a broader relational structure. Network methods can map supervisory ties and identify central actors or resource flows within academic environments [18,19]. Empirical work on evolving networks provides methodological grounding for analyzing how academic relationships change over time [22]. Recent studies [34,35] demonstrate the effectiveness of explainable AI in educational prediction tasks, which aligns with the motivation of this work. Experience sampling and diary studies offer another perspective by measuring supervision as a repeated daily process rather than a single retrospective evaluation. Such designs can capture feedback cycles, emotional responses, and interaction intensity over time [20,21]. With increasing digitization of academic communication, computational approaches have explored the use of digital traces, such as communication records, to study evolving relational patterns [23].

Overall, prior SSR research [7,9,18,21] has produced valuable descriptive evidence through surveys, measurement instruments, qualitative typologies, and emerging computational approaches. However, there is still a lack of a unified framework that can analyze large-scale, multi-dimensional mentorship evaluations while remaining interpretable for educational researcher. This motivates this paper's contribution of *Interpretable Mentorship Analytics* (IMA), which leverages large-scale anonymous evaluations and explainable machine-learning models [30,31] to uncover non-linear patterns and key drivers underlying mentorship outcomes.

Existing studies on student-supervisor relationships have provided important insights, but they often rely on relatively limited samples, qualitative observations, self-reported survey instruments, or descriptive statistical analyses. In contrast, our study contributes a more integrated analytical perspective by combining large-scale anonymous evaluation data, structured feature construction, predictive modeling, and explainable analysis within a unified framework. Overall, the advancement of this work lies in methodological integration and analytical depth, rather than in the introduction of a standalone novel algorithm.

3 Datasets and Methods

This section focuses on how the raw anonymous review data are processed and transformed into a structured and analyzable mentorship dataset. The goal is to establish a reliable empirical foundation for the subsequent descriptive and model-based analyses. We present the construction of the dataset and the methodological pipeline up to the generation of structured and analyzable mentorship indicators, as depicted in Fig. 3. In contrast to traditional small-sample surveys or interview-based studies, our dataset is derived from large-scale, *publicly accessible* and *anonymous* student-generated reviews collected from multiple Chinese online supervisor-evaluation platforms. We then transform heterogeneous, unstructured

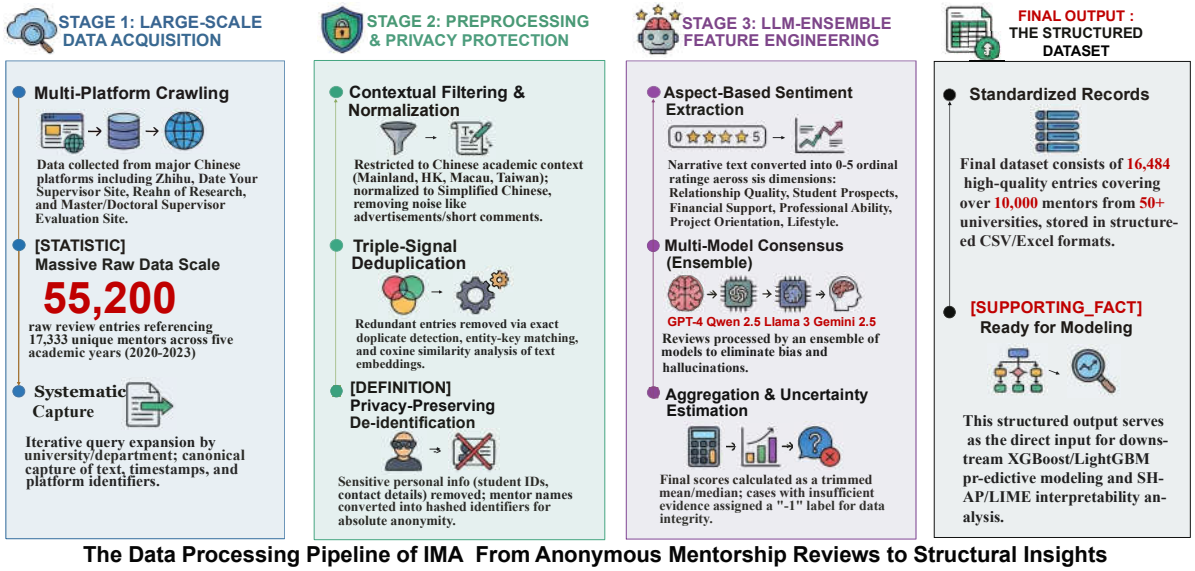

Figure 3: Overview of the data collection and processing steps of IMA data pipeline, which focuses on stages (1)–(3): multi-platform review collection, deduplication/merging/filtering, and LLM-ensemble-based feature engineering.

text into standardized, machine-readable mentorship features via a rigorous preprocessing pipeline and an *LLM-ensemble* aspect-rating procedure designed to mitigate model-specific bias and hallucination.

3.1 Study Design Overview

As shown in Algorithm 1, the overall workflow of Interpretable Mentorship Analytics (IMA) consists of four sequential stages: (1) large-scale data acquisition from public anonymous review platforms; (2) data cleaning, normalization, and entity consolidation; (3) LLM-ensemble-driven feature engineering to convert narrative reviews into comparable mentorship indicators; and (4) predictive modeling and explainable analysis using machine learning and XAI techniques (reported later in Section 4). This section covers stages (1), (2), and (3).

3.2 Data Sources and Acquisition

We collected anonymous student reviews were collected from multiple public online supervisor-evaluation platforms widely used in the Chinese higher-education ecosystem, including: (i) Master and **Doctoral Supervisor Evaluation Site** [7] (硕博导师评价网), (ii) **Rate Your Supervisor Site**[8] (导师点评网), (iii) **Zhihu**[9] (知乎), and (iv)**Realm of Research** [10] (研控), research-oriented online communities where supervisor experiences are posted and openly discussed.

The reviews were posted between Fall 2020 and Spring 2025, covering approximately five years. The initial crawl produced approximately 55,200 raw review entries referencing about 17,333 unique mentors. These reviews span a wide range of university tiers, academic disciplines, and geographical regions,

---

[7] Doctoral Supervisor Evaluation Site (硕博导师评价网): http://www.daoshipingjia.org/.
[8] Rate Your Supervisor Site (导师点评网): https://www.rateyoursupervisor.com/Professor.
[9] Zhihu (知乎): https://www.zhihu.com/.
[10] **Realm of Research** (研控): https://www.rateyourprofessor.com.cn/.

Algorithm 1 Procedures of Dataset Construction and LLM-Ensemble Feature Engineering

Require: Public reviews; aspect set $\mathcal{A}$; LLM set $\mathcal{M}$
Ensure: Structured dataset D
Collect and normalize public reviews
Filter to Chinese academic context
Deduplicate and consolidate mentor entities
for each review **do**
    for each model $m \in \mathcal{M}$ **do**
        Extract aspect ratings with evidence
    end for
    Aggregate ratings and estimate uncertainty
end for
Output D

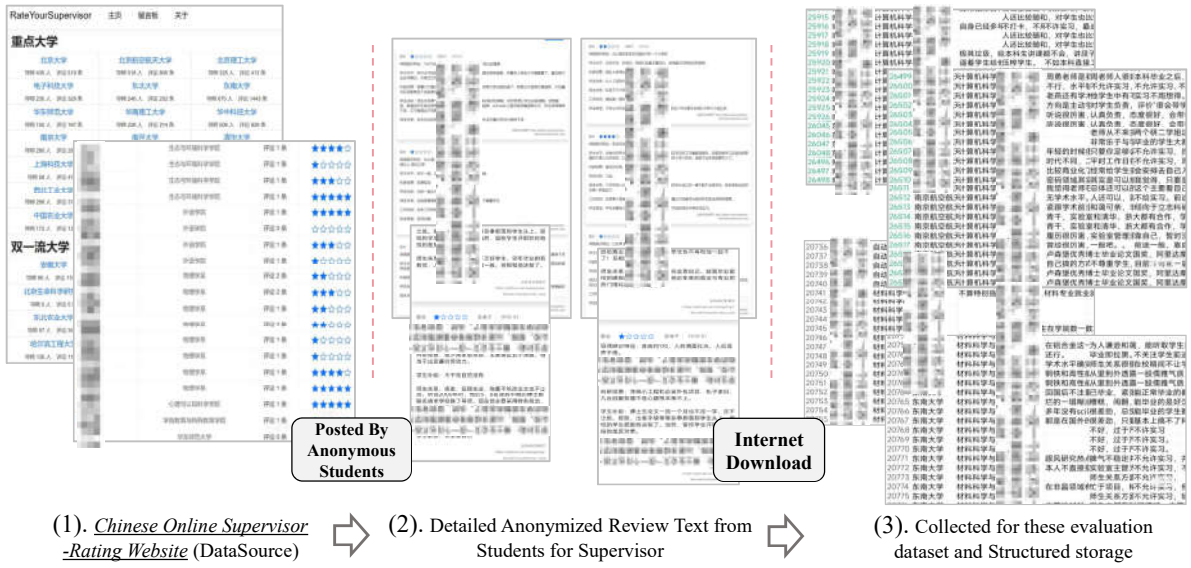

Figure 4: Illustration of the public data acquisition process (example platform view → anonymized review text → structured storage). All personal identifiers are masked; displayed content is illustrative.

providing the heterogeneity necessary for large-scale statistical and machine-learning analyses. The overall data acquisition process is illustrated in Fig. 4.

Acquisition procedure.

We implemented a structured and reproducible data collection process:

1. **Query expansion:** for each platform, we iteratively queried by university name, school/department keywords, and mentor identifiers (when present).
2. **Canonical capture:** we stored each review with platform identifier, review text, timestamp, university/school mentions, and any available structured fields (e.g., star ratings where provided).
3. **Reproducibility:** all intermediate snapshots were stored with versioning to support later auditing of the cleaning and merging steps.

Because the topic involves sensitive mentorship relationships, we adopt a privacy-preserving and harm-minimizing approach: (1) For public-only and minimality principle, we only used reviews that were publicly accessible and did not require user authentication or disclosure of personal information. We collected the minimal information necessary for research aims: review text, coarse affiliation cues

(university/department), and timestamp. (2) For anonymization and de-identification, all sensitive personal information (e.g., student identifiers, contact details, lab addresses, or uniquely identifying narratives) was removed or masked during preprocessing. When mentor names appear in raw text, we store them as hashed identifiers for deduplication/merging; they were not included as modeling features.

All results are reported at aggregated levels (e.g., distributions by discipline or province), and no claim is made about any specific identifiable individual. We additionally avoid releasing raw text at scale; if examples are shown, they are paraphrased or templated. Moreover, due to the cognitive bias, user-generated online reviews may be subject to selection bias (e.g., extremes more likely to post) and strategic reporting. To reduce the impact of spurious content, we incorporate multiple quality-control steps (duplicate detection, plausibility filters, and ensemble scoring), which is described below.

3.3 Data Preprocessing and Quality Control

As shown in Fig. 5, the goal of preprocessing is to produce a high-quality dataset of reviews that are (i) relevant to the Chinese academic system, (ii) non-duplicative, (iii) linkable to a mentor entity for aggregation, and (iv) suitable for aspect-based feature extraction.

Filtering to the Target Academic Context. We restrict the dataset to mentors affiliated with institutions in Mainland China, Hong Kong, Macau, and Taiwan. Reviews referring to institutions outside these regions were excluded to ensure consistency in institutional norms, funding structures, and mentorship expectations.

Duplicate Removal and Record Consolidation. Online platforms may contain duplicate or near-duplicate entries due to reposting, mirror pages, or repeated submissions. We remove duplicates using a multi-signal strategy:

Duplicate Removal Procedures

- **Exact duplicate detection:** identical review text (after normalization) with the same platform and timestamp.
- **Entity-key matching:** same university, sub-unit (school/department), mentor name string (before hashing), and high textual similarity above a threshold.
- **Near-duplicate text similarity:** cosine similarity in a sentence-embedding space for suspected duplicates with minor edits (e.g., punctuation changes).

After deduplication, we consolidate multiple reviews referring to the same mentor by creating a *mentor-level container* that stores all associated reviews and metadata. This supports later aggregation (e.g., mean/median aspect scores per mentor) and robustness checks (e.g., variance across reviewers).

Content Normalization and Noise Filtering. Then, we normalize review text to improve downstream parsing and reduce noise: To ensure uniform text representation and facilitate reliable downstream processing, we performed content normalization on all review texts, which involves converting diverse characters to a consistent standard and normalizing punctuation marks.

After filtering, deduplication, and consolidation, we obtained 16,484 valid review entries, covering more than 10,000 mentors from over 50 Chinese universities. This processed dataset is stored in a structured tabular format (CSV/Excel) for modeling.

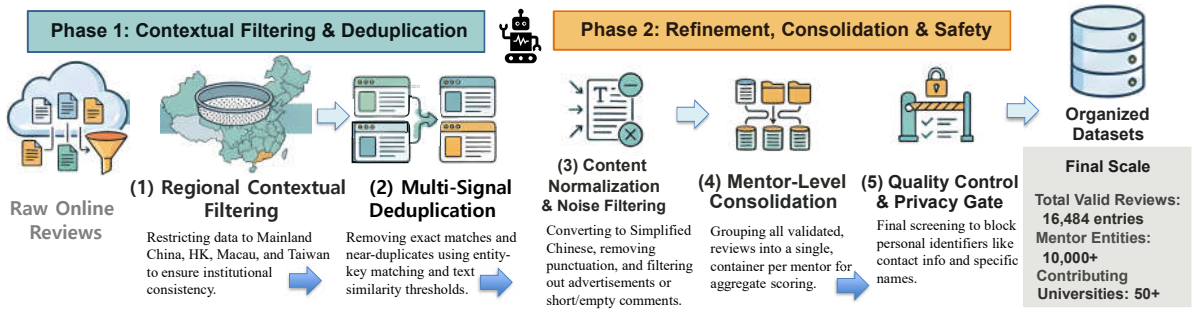

Figure 5: Workflow for preprocessing mentor review data. The pipeline (1) filters reviews by region, (2) deduplicates entries using multiple signals, (3) normalizes text and removes noise, (4) consolidates reviews per mentor, and (5) applies final quality and privacy controls.

3.4 Metadata Enrichment: Gender, Geography, and Discipline

We enrich the dataset with coarse-grained metadata that supports stratified analyses without requiring personally identifiable information.

Content Normalization and Noise Filtering Procedures

- Firstly, in order to eliminate the interference caused by character encoding and formatting differences, we standardized all the comment texts by converting them to Unicode and removing punctuation marks.
- Secondly, considering that the data source might contain simplified and traditional Chinese characters, we carried out language consistency processing, converting all the texts into simplified Chinese to maintain the consistency of the language style.
- To enhance the quality and relevance of the dataset, we implemented irrelevant content filtering. Specifically, we removed empty comments, extremely short ones (less than 5 characters), and those that were clearly unrelated to the topic of mentor evaluation (such as advertisements or completely off-topic texts), ensuring that the subsequent analysis focused on valid mentorship information.
- Finally, considering strict ethical and research standards, we conducted content security and privacy screening to block the explicit personal identifiers (such as specific names, contact information, and addresses) in the text.

Firstly, we infer mentor gender only when feasible using name-based gender prediction models and linguistic priors for Chinese names. Because name-based inference can be uncertain (e.g., unisex names, transliterations), we attach a confidence score to each inferred label, assign an "Unknown" category when the confidence falls below a predetermined threshold, and strictly limit the use of gender labels to aggregate-level comparisons and robustness checks.

Then, we map each mentor to a province-level location based on the university affiliation stated in the review (or platform metadata when available). When affiliation is ambiguous or missing, we mark location as Unknown and exclude such records from location-specific plots.

Next, based on department and school information, mentors are classified into ten broad disciplinary categories: *Engineering; Information Technology; Medicine & Life Sciences; Natural Sciences; Social Sciences & Humanities; Agriculture & Environmental Sciences; Business & Management; Law & Public Affairs; Art &*

Design; and Education. This taxonomy enables field-level comparisons and controls for discipline-specific mentorship norms.

3.5 LLM-Ensemble Feature Engineering for Aspect-Based Ratings

A core challenge is converting free-form narrative reviews into consistent mentorship indicators across platforms. We therefore perform *aspect-based sentiment extraction and rating* using an ensemble of large language models (LLMs), which is shown in Fig. 6. It is designed to reduce single-model bias and mitigate hallucination risk.

For each review, we extract ratings on a 0–5 ordinal scale (higher is better). We additionally use a special label −1 to indicate *Not Applicable / Not Enough Evidence* when the text does not provide sufficient information. The aspects include *mentor–student relationship quality, student prospects, financial support, supervisor professional ability, project/task orientation, and work–life or lifestyle practices.*

Because single-model extraction can suffer from systematic biases (instruction-following quirks, cultural priors, calibration drift) and occasional hallucinations. To improve reliability, we use **multiple LLM families** and aggregate their outputs: *(1) OpenAI GPT 5*[11]*, (2) Alibaba Qwen 2.5*[12]*, (3) Meta Llama 3*[13]*, (3) Google Gemini 2.5*[14]. Each model independently produces a structured JSON containing: aspect ratings, short evidence snippets (verbatim spans from the review), and a confidence score per aspect. We then compute a **consensus rating** per aspect via robust aggregation.

To ensure standardized and reliable outputs across different model queries, we employ a constrained prompting template. This template provides clear definitions for each evaluation aspect along with anchor examples corresponding to scores of 0, 3, and 5. It also enforces a strict evidence requirement, mandating that the model must cite the supporting phrase(s) from the input text; if no supporting evidence is found, it must output a value of -1. Finally, all outputs are required to adhere to a specified JSON schema, which enforces fixed key names and ensures all numerical ratings fall within predefined bounds.

Aggregation, Uncertainty, and Bias Mitigation. Let $r_{i,a}^{(m)}$ be the rating given by model m for review i on aspect a. The final rating $\hat{r}i, a$ is computed through a robust multi-step procedure. First, we discard any model output that violates the prescribed JSON schema or lacks the required evidence spans. Second, if fewer than a threshold number K (e.g., $K = 3$) of valid model outputs remain, we assign $\hat{r}i, a = -1$ for that review-aspect pair. Third, when sufficient valid outputs exist, we compute $\hat{r}_{i,a}$ as the trimmed mean or median across the ensemble models, and record the dispersion (e.g., median absolute deviation or variance) as an explicit uncertainty indicator.

To further mitigate systematic errors and maintain construct validity, we incorporate lightweight, targeted manual checks into the pipeline. This involves performing stratified sampling to randomly select reviews from each major platform and discipline category. For each sampled review, we conduct checklist validation to verify that the assigned aspect ratings logically align with the cited textual evidence and that the −1 code is correctly applied when evidence is genuinely absent. It is important to note that these human checks are not intended to label the entire dataset, but rather to provide a periodic sanity check that ensures the automated feature-engineering pipeline remains aligned with the intended theoretical constructs.

---

[11] OpenAI GPT 5: https://chat.chatbot.app/gpt5.
[12] Alibaba Qwen 2.5: https://chat.qwen.ai/.
[13] Meta Llama 3: https://ai.meta.com/blog/meta-llama-3/.
[14] Google Gemini 2.5: https://gemini.google.com/.

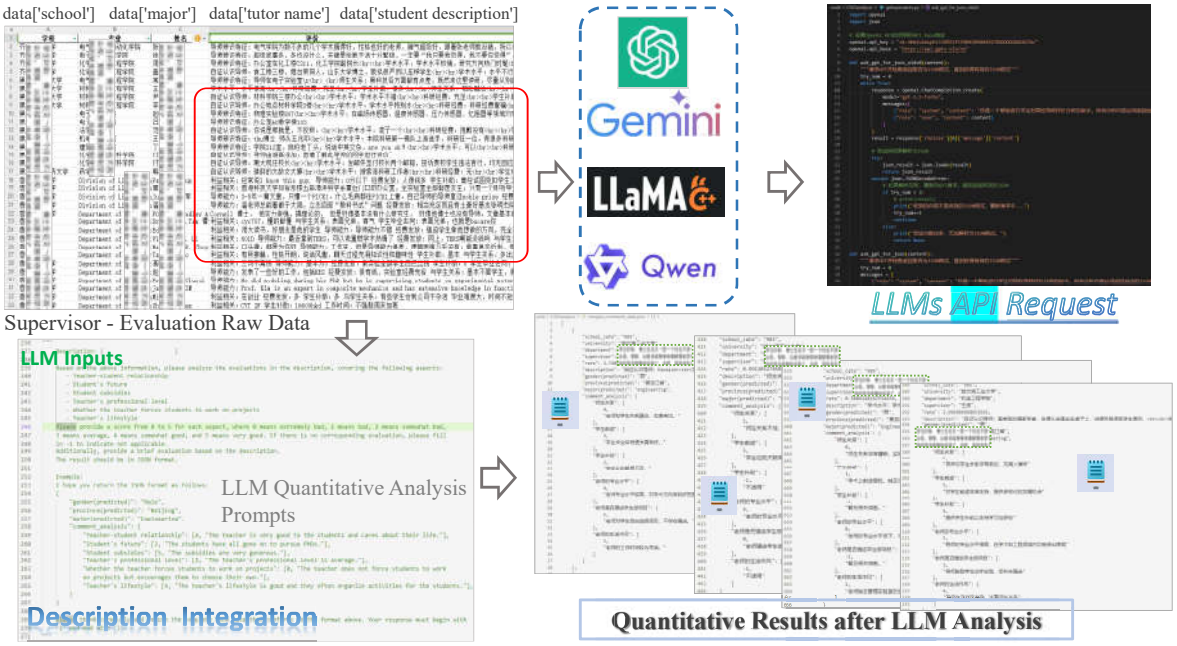

Figure 6: LLM-ensemble aspect rating. Each review is scored independently by multiple LLMs using a shared schema; outputs are validated and aggregated into final aspect indicators with uncertainty estimates.

Final Structured Dataset[15]**.** After preprocessing and LLM-ensemble feature engineering, each record contains:

The Records of Preprocessed Dataset

- identifiers: platform info, anonymous mentor info, and review info. And all text are de-identified review contents;
- metadata: timestamp, university, province (if available), discipline, inferred gender (optional with confidence);
- aspect indicators: $\hat{r}_{i,a} \in \{-1, 0, 1, 2, 3, 4, 5\}$ for each aspect a;

Lastly, this structured dataset serves as the input to the machine-learning predictive models and explainable analyses presented in Section 4.

4 Experimental Procedures

This section explains how the proposed IMA framework transforms the structured dataset into predictive and interpretable evidence. We first define the learning task and target variables, then describe the modeling pipeline, and finally introduce the explainable machine learning techniques used to interpret the prediction mechanisms. Specifically, building on the structured dataset introduced in Section 3, we use our proposed novel multi-scenario explainable AI framework, i.e., *Interpretable Mentorship Analytics*

[15] All preprocessing and annotation steps are publicly accessed at Github: https://github.com/SmileHappyEveryDay/tutor-evaluation-data-analysis. All publicly released data and codes adhere to platform terms and ethical standards.

(**IMA**), to systematically investigate mentorship dynamics and uncover interpretable patterns in large-scale mentorship evaluations.

Given the large-scale anonymous evaluation dataset used in this work, a machine learning-based approach is more suitable for identifying generalizable patterns, modeling complex relationships, and ensuring analytical consistency. This methodological choice allows us to focus on systematic pattern extraction and mechanism-oriented interpretation at scale.

Our approach aims to (i) learn predictive mappings between multi-dimensional mentorship indicators and overall mentorship outcomes, and (ii) provide theoretically grounded and transparent explanations of these predictions through explainable artificial intelligence (XAI) techniques. Rather than relying on a single experiment, our work treats explainability as a *multi-dimensional analytical process*, where different scenarios correspond to different scientific questions about mentorship evaluation mechanisms.

4.1 Task Definition and Learning Setup

First, we formulate the analysis as a supervised learning problem. Each instance corresponds to either: (1) a **review-level instance**, representing an individual student's evaluation; or (2) a **mentor-level instance**, obtained by aggregating multiple reviews associated with the same supervisor.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ denote the processed dataset, where $\mathbf{x}_i \in \mathbb{R}^p$ represents the feature vector associated with the i-th review (or mentor-level aggregate), and $y_i \in \mathbb{R}$ denotes the corresponding overall mentorship outcome.

Each feature vector is defined as $\mathbf{x}_i = \left[x_{i,1}, x_{i,2}, \ldots, x_{i,p}\right]$, Its features correspond to LLM-extracted mentorship indicators (e.g., teacher–student relationship, professional ability, financial support), optionally augmented with contextual metadata such as discipline, institution tier, or geographic location.

The prediction target y_i is the overall evaluation score, derived from explicit platform ratings or from aggregated satisfaction indicators when explicit scores are unavailable. The primary focus of this study adopts a regression formulation to preserve fine-grained variation in student perceptions.

The learning objective is to estimate a predictive function $f : \mathbb{R}^p \to \mathbb{R}$, such that $\hat{y}_i = f(\mathbf{x}_i)$, approximates the observed outcome y_i with minimal prediction error under a chosen loss function.

4.2 Data Partitioning and Modeling Preprocessing

Before model training, we apply a standardized preprocessing pipeline for modeling[16]:

Modeling Standardized Preprocessing Pipeline

1. **Missing values:** Aspect scores marked as -1 (insufficient evidence) are treated as missing. Tree-based models handle these natively; for robustness checks, missingness indicators are optionally included.
2. **Categorical variables:** Discipline, institution tier (985/211/Double Non-first-class), and region are encoded using one-hot or ordinal encoding depending on the model.
3. **Train & test split:** Data are split into training and test sets, with stratification by institution tier and discipline where applicable.

[16] All experiments are repeated across multiple random seeds to ensure result stability.

4.3 Predictive Modeling with Gradient-Boosting Methods

To capture complex and non-linear relationships between mentorship characteristics and evaluation outcomes, we employ a suite of machine-learning models, with a focus on gradient-boosting-based methods: (1) *XGBoost* [27], (2) *LightGBM* [29], (3) *Histogram-Based Gradient Boosting* [36], and (4) *Classical Gradient Boosting Decision Trees (GBDT)*[17].

Modeling rationale. Gradient-boosting models are particularly suitable for this task because they can handle heterogeneous feature scales and missing values, capture high-order non-linear interactions, and integrate seamlessly with SHAP-based explanation frameworks.

These models estimate the prediction function as an additive ensemble of K regression trees:

$$f(\mathbf{x}) = \sum_{k=1}^{K} \gamma_k h_k(\mathbf{x}), \tag{1}$$

where: $h_k(\cdot)$ denotes the k-th decision tree, γ_k is the corresponding tree weight, and K is the total number of trees in the ensemble.

This formulation enables the modeling of high-order feature interactions while remaining compatible with additive explanation methods. Hyperparameters are tuned using validation sets, and early stopping is employed where supported.

4.4 Our Proposed Novel Interpretable Mentorship Analytics (IMA) Framework

In this study, "interpretable" does not mean that the underlying predictive model is inherently transparent. Instead, we adopt high-performing tree-based ensemble models and use post-hoc explainability techniques to interpret their decision logic. Therefore, interpretability here refers to the ability to explain: (i) which mentorship-related features most strongly influence the predicted overall evaluation, (ii) whether their effects are positive, negative, linear, or non-linear, and (iii) how such effects vary across instances and contexts.

As we know, predictive accuracy alone is insufficient for understanding mentorship dynamics. We therefore integrate explainable AI techniques to interpret both global trends and local behaviors. To interpret the predictions of non-linear ensemble models, we integrate two complementary explainability techniques: *SHapley Additive exPlanations* (**SHAP**) for global interpretability and *Local Interpretable Model-Agnostic Explanations* (**LIME**) for local, instance-level analysis.

4.4.1 Global Explanations and Local Analysis

To comprehensively interpret model predictions from both holistic and individual perspectives, we adopt a multi-scale interpretability framework centered on SHAP and supplemented by LIME. This framework integrates the strengths of global and local analyses, aiming to reveal both general patterns and specific mechanisms underlying student evaluations.

For global explanation, it assists in identifying the core, universal drivers in mentorship quality. Global analysis aims to answer the question: "Which factors most significantly influence satisfaction across all evaluations?" We achieve this by aggregating SHAP values across the entire sample:

---

[17] Classical Gradient Boosting Decision Trees (GBDT): https://github.com/yarny/gbdt.

- **Feature Importance Ranking:** The mean absolute SHAP value for each feature j is computed as: $I_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_j(\mathbf{x}_i)|$. This yields a robust ranking of feature importance, quantifying the average contribution of each dimension to overall satisfaction.
- **Global Dependency Relationships:** Using SHAP summary plots and dependence plots, we visualize the overall relationship between feature values and their impact on predictions.

For local analysis, it helps to understand the heterogeneity of evaluations and identify special cases that deviate from general patterns. Local explanations focus on specific cases, addressing questions such as "Why did a particular student give this evaluation?" or "How was a specific supervisor's overall rating formed?" We employ two methods for local interpretation:

- **Instance-Specific SHAP Values:** For any given sample $\mathbf{x}_i$, its SHAP decomposition $f(\mathbf{x}_i) = \phi_0 + \sum_{j=1}^{p}\phi_j(\mathbf{x}_i)$ clearly illustrates how each feature "pushes" the prediction away from the baseline ϕ_0, thereby explaining the source of that specific prediction outcome.
- **LIME Algorithm:** For complex cases requiring an alternative or more intuitive explanation, we use LIME to construct an interpretable local surrogate model (e.g., a sparse linear model) around the prediction point. It approximates the black-box model's behavior locally and provides linear feature coefficients.

Overall, through this integrated framework that combines global and local perspectives, it reveals not only macro-level patterns but also provides deep insights into micro-level individual cases. Thus, we can systematically deconstruct the "black box" of supervisor evaluations, thereby laying the foundation for subsequent multi-scenario analysis.

4.4.2 SHAP: Global and Consistent Feature Attribution

SHAP is grounded in cooperative game theory and attributes a model's prediction to individual input features by computing their Shapley values. For a given instance $\mathbf{x}$, the SHAP value of feature j is defined as:

$$\phi_j(\mathbf{x}) = \sum_{S\subseteq\mathcal{F}\setminus\{j\}} \frac{|S|!\,(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!}\left[f_{S\cup\{j\}}(\mathbf{x}) - f_S(\mathbf{x})\right], \tag{2}$$

where $\mathcal{F} = \{1, 2, \ldots, p\}$ is the full feature set, S denotes a subset of features not containing j, and $f_S(\mathbf{x})$ is the expected model prediction conditioned on features in S.

Under this formulation, the prediction is decomposed additively as: $f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{p}\phi_j(\mathbf{x})$, where ϕ_0 is the baseline prediction (expected value of the model output).

For tree-based models, we employ TreeSHAP, which computes exact SHAP values in polynomial time, ensuring consistency and local accuracy.

Global interpretation. By aggregating absolute SHAP values across all samples, $I_j = \frac{1}{N}\sum_{i=1}^{N}|\phi_j(\mathbf{x}_i)|$, we obtain a global importance score I_j that quantifies the overall influence of feature j on mentorship outcomes.

4.4.3 LIME: Local Surrogate-Based Explanation

While SHAP captures population-level trends, mentorship dynamics often exhibit strong heterogeneity. To explain individual predictions, we apply LIME, which approximates the complex model locally using an interpretable surrogate.

Given an instance $\mathbf{x}$, LIME solves the following optimization problem:

$$g_{\mathbf{x}} = \arg\min_{g \in \mathcal{G}} \mathcal{L}\left(f, g, \pi_{\mathbf{x}}\right) + \Omega(g), \tag{3}$$

where f is the original black-box model, g is a simple interpretable model (e.g., linear regression), $\pi_{\mathbf{x}}$ defines a locality-aware weighting kernel around $\mathbf{x}$, $\mathcal{L}$ measures the fidelity of g in approximating f locally, and $\Omega(g)$ penalizes model complexity to enforce interpretability. In practice, $g_{\mathbf{x}}(\cdot)$ provides a sparse linear explanation: $g_{\mathbf{x}}(\mathbf{z}) = \beta_0 + \sum_{j=1}^{p} \beta_j z_j$, where coefficients β_j indicate the local contribution of feature j to the prediction at $\mathbf{x}$.

From a methodological perspective, the proposed IMA framework advances prior SSR research by conducting a complete data-to-insight process. Instead of relying solely on manually interpreted findings, IMA enables the transformation of raw evaluation information into structured mentorship indicators, followed by predictive modeling and explainable analysis. This makes it possible to move from pattern description to mechanism-oriented interpretation in a more systematic and scalable way.

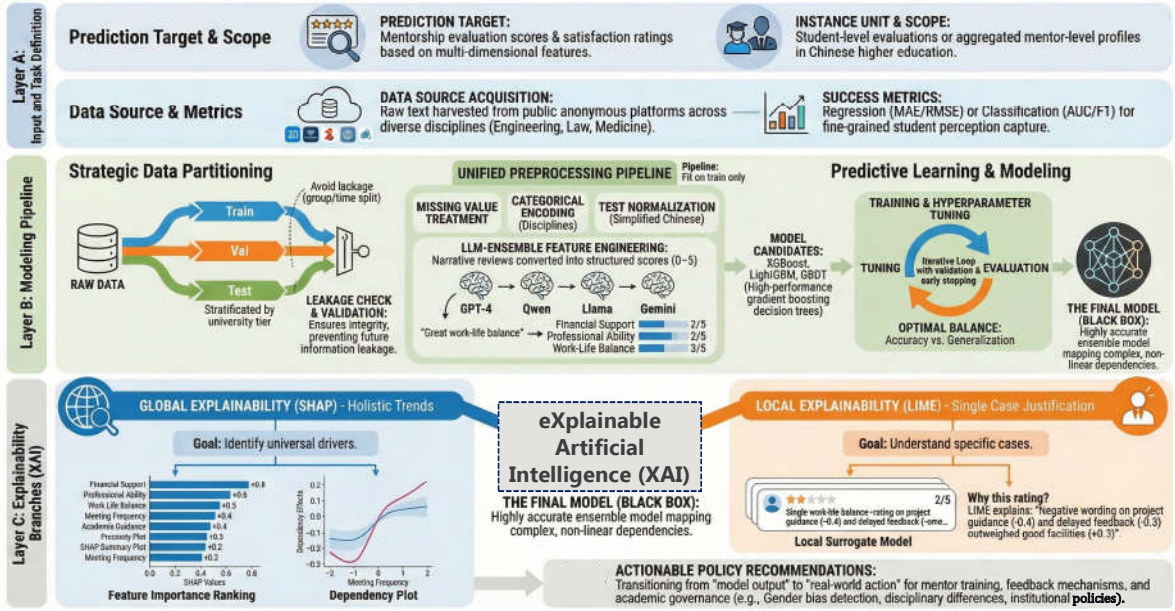

Figure 7: Overview of the experimental procedures of our *Interpretable Mentorship Analytics* (**IMA**) framework, illustrating task definition, data partitioning and modeling preprocessing, predictive modeling, and global and local explainable analyses based on SHAP [30] and LIME [31].

4.5 Entire Experimental Procedures

Lastly, Fig. 7 illustrates the overall experimental workflow and the relationship between predictive modeling, explainable analysis, and result interpretation. The complete experimental workflow is summarized as follows:

Experimental Workflow Summary

1. Construct structured mentorship features via LLM-ensemble annotation.
2. Train gradient-boosting models to predict overall evaluation outcomes.
3. Evaluate predictive performance using standard regression metrics.
4. Apply SHAP and LIME for global and local explainability.
5. Conduct multi-scenario explainable analyses to address distinct research questions.

In summary, the experimental procedures integrate data preprocessing, predictive modeling, and explainable analysis into a unified workflow. We first construct robust predictive models to capture overall evaluation patterns, and then employ multi-scenario explainable analysis to interpret model behavior from global and local perspectives. We believe the proposed procedures enable accurate prediction while providing interpretable insights into the formation of evaluation outcomes, forming the empirical foundation for subsequent results.

Model Validity and Reliability Assessment. To ensure the credibility and reliability of the conclusions derived from the interpretable machine learning framework, we evaluate the model from multiple complementary perspectives.

First, predictive validity is assessed through out-of-sample performance metrics, ensuring that the model captures meaningful patterns rather than overfitting the training data. Second, cross-model consistency is examined by comparing the key feature importance rankings obtained from different models (e.g., XGBoost, LightGBM, and Random Forest). Third, robustness to data preprocessing choices is verified by testing different strategies for handling missing values and feature transformations, ensuring that the results are not sensitive to specific preprocessing decisions. Finally, substantive consistency is evaluated by comparing the explainable model results with descriptive statistical patterns observed in earlier analyses. The alignment between these two perspectives provides additional support for the validity of the conclusions.

Together, these validation steps strengthen confidence that the identified patterns are stable, interpretable, and meaningful within the context of large-scale mentorship evaluation data.

5 Results and Discussion

This section is organized into two complementary analytical layers.

First, Section 5.1 presents descriptive and exploratory analyses based on the structured mentorship dataset, aiming to summarize empirical distributions, group differences, and latent patterns across regions, institution types, and disciplines.

Second, Section 5.2 presents predictive and explainable analyses based on our IMA. This part focuses on identifying which features systematically influence the overall evaluation outcomes and how their effects vary across contexts and individual cases.

The descriptive analyses reveal "what patterns exist", while the explainable machine learning analyses help explain "which factors matter most" and "how these factors jointly shape the overall evaluation".

5.1 Overall Data Analysis

We begin by examining the empirical distribution of overall evaluations and aspect indicators, as well as their variation across time and context (e.g., discipline and geography). The followings analyses serve two purposes: they validate that the processed dataset exhibits meaningful heterogeneity, and they reveal preliminary patterns that the predictive and explainable models will later formalize.

Note that the analyses in this subsection are primarily descriptive and exploratory. They are intended to characterize the empirical structure of the processed dataset and identify observable patterns across different groups. These findings serve as an empirical foundation and motivation for the subsequent model-based explainability analysis, rather than being directly generated by the explainable machine learning framework.

Attention Patterns Analysis Measured through Mention Rates. We first ask a simpler but foundational question: *what do students talk about most when they evaluate supervisors?*

Intuitively, mention rates provide a proxy for cognitive salience, reflecting which dimensions naturally enter students' evaluation narratives and which remain peripheral. Fig. 8 presents attention distributions across majors, university tiers, and supervisor gender. Across all majors, attention is highly concentrated on a small set of dimensions.

As shown in Fig. 8 (left), teacher–student relationship and professional ability consistently receive near-universal attention, indicating that students treat interpersonal interaction and academic competence as baseline criteria for supervision quality. Student prospects also attract high attention, though with slightly more variation across fields. In contrast, student allowance receives markedly lower mention rates, suggesting that material support is rarely foregrounded in spontaneous evaluations, even in disciplines where financial pressure may be substantial. he same concentration pattern appears across institutional tiers and supervisor gender. As shown in Fig. 8 (right), students from 985, 211, and other universities show remarkably similar attention hierarchies, with only modest shifts in emphasis. Likewise, evaluations of male and female supervisors differ little in what dimensions students attend to, implying that attention allocation is largely driven by evaluation norms rather than demographic characteristics.

These results suggest that students' evaluations are structured by a stable attention framework. Before considering performance gaps or inequality across groups, it is therefore essential to recognize that most evaluative weight is placed on a limited set of highly salient dimensions, while others operate at the margins of student attention.

Weight Attribution of Supervisor Evaluation Dimensions. In order to conduct a further study on which specific dimension among the various subdivisions has the greatest impact on the overall evaluation (*Average Score*) of the mentors, we conducted a weight attribution analysis on six evaluation dimensions. Fig. 9 presents a weight attribution analysis examining how different evaluation dimensions contribute to students' overall supervisor ratings. Using both standardized multivariate linear regression and random forest regression, we assess the relative importance of six dimensions under multiple missing-value treatment strategies.

As shown in the top-right of Fig. 9, across all model specifications, *Teacher–Student Relationship* emerges as the most influential determinant of the overall rating. In the linear regression framework, this dimension exhibits the largest standardized coefficient, substantially exceeding those of all other factors, indicating that relational quality plays a dominant role in shaping students' global evaluations. *Professional Ability* and *Project Attitude* form a secondary tier of importance, while *Student Prospects*, *Lifestyle*, and especially *Student Allowance* display comparatively limited explanatory power. The random forest results corroborate this pattern: even when allowing for nonlinear effects and interactions, teacher–student relationship remains the most important feature by a wide margin. Importantly, the ranking of key dimensions is highly stable across different missing-value handling strategies, suggesting that the dominance of relational factors is not an artifact of data preprocessing choices.

Therefore, for students, the most crucial sub-dimension that determines the overall evaluation of a mentor is "teacher-student relationship". Ability and project attitude are important but secondary

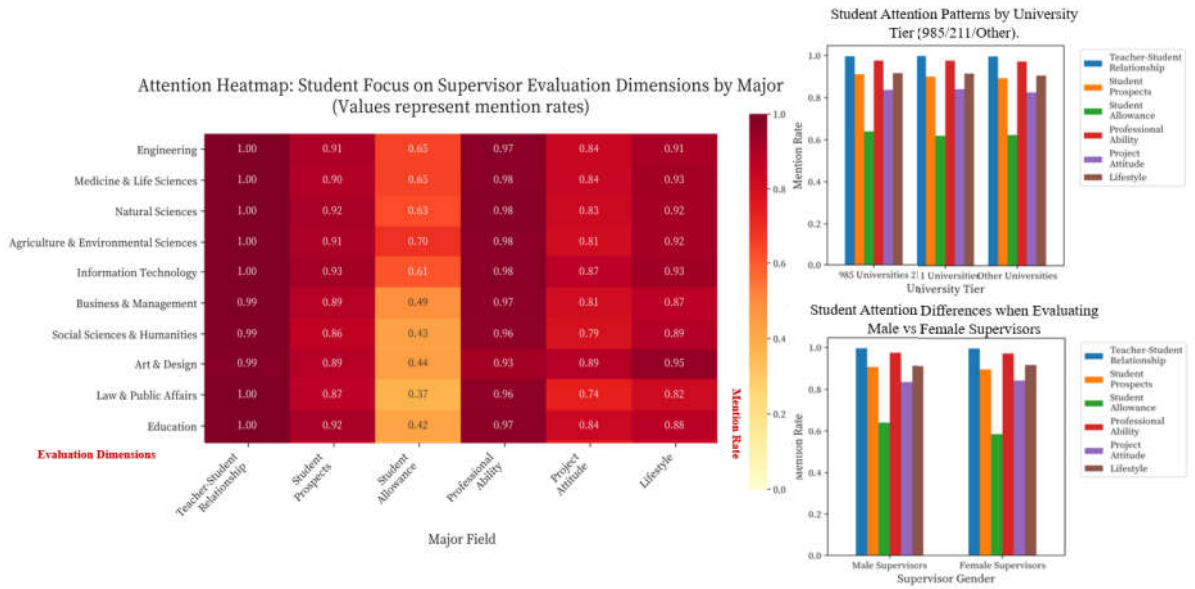

Figure 8: Student attention patterns in supervisor evaluations, measured by mention rates across evaluation dimensions. The heatmap (left) shows discipline-level variation in attention, while the bar charts (right) compare attention distributions by university tier (985/211/other) and supervisor gender.

explanatory variables; student grants and student prospects have a greater impact on the local experience and have relatively limited explanatory power for the overall score. These findings indicate that students' perceptions of supervisory quality are driven primarily by relational and interactional factors rather than material support or instrumental outcomes.

Geographical and Circle Differences: Region, School Tier, and Major Pressure. Fig. 10 summarizes how supervisor ratings differ across regions, school tiers, and majors. The results are shown in several panels: regional averages, school-tier score distributions, a major "pressure index" ranking, and a provincial heatmap of key metrics.

First, the regional comparison shows clear differences in overall experience. In the bar charts, the *Northeast* group has the highest average score (around 3.00), while *Other regions* score lower (around 2.75). The drivers of satisfaction also differ: developed regions score higher on *student allowance*, whereas Northeast performs better on *teacher–student relationship* and *professional ability*.

Second, school tier does not fully shield students from negative experiences. The boxplots show strong overlap across 985, 211, and other universities on all dimensions. Variation within each tier is large, and the distributions of average scores overlap substantially. In practice, students can encounter both good and poor supervisors in any tier, and within-tier differences often exceed between-tier differences.

Third, majors display a clear pressure pattern. The *Major Pressure Index Ranking* places *Information Technology* near the top, with several majors closely following. The scatter plot indicates that better student prospects are generally linked to higher overall scores, but the relationship is not strict. Differences in *allowance* and *relationship* across majors further show that similar overall scores can mask very different lived experiences.

Finally, the provincial heatmap offers a compact summary of these differences. Provinces on the left generally score higher across multiple metrics, while those on the right score lower. However, the four

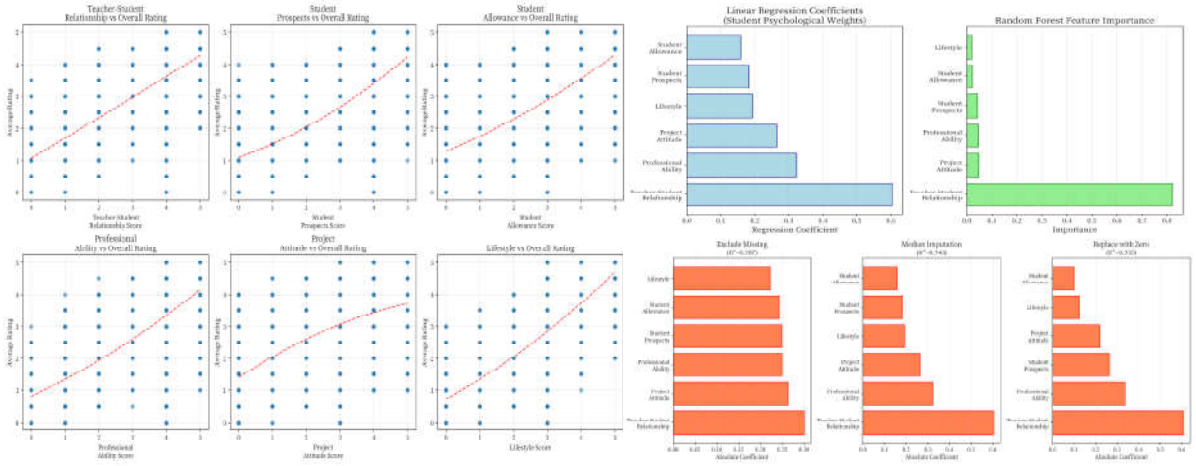

Figure 9: Weight attribution of overall supervisor rating. Left: bivariate relationships between each dimension score and overall rating with fitted trend. Right: standardized linear regression coefficients and random forest feature importance. Bottom: robustness comparison across missing-value strategies.

indicators do not move in perfect sync: some provinces perform well overall but lag in allowance, while others show the opposite.

Overall, these results lead to three major conclusions: (1) region-level gaps exist, and they come from a mix of money, relationship quality, and capability; (2) school tier differences are real but not decisive, because variation within the same tier is large; (3) major-level pressure is visible, and prospects, allowance, and relationship can move together but also diverge, so each field has its own "stress profile".

Supervisor Archetypes from Clustering. Fig. 11 groups supervisors into five patterns based on six rating dimensions. The cluster selection results indicate that $K = 5$ provides a reasonable balance between structure clarity and explanatory power, as the elbow curve flattens and the other evaluation scores show limited improvement beyond this point.

The radar charts (top-right of Fig. 11) reveal clear differences between clusters. One group represents a strong "best overall" profile (*All-round Excellence*, n=2243), with consistently high scores across relationship, ability, and project attitude. Another large group shows a more moderate but stable pattern (*Balanced Type*, n=1686), with mid-to-high scores and no pronounced weaknesses. Several weaker patterns are also visible. Two clusters are labeled *Needs Improvement* ($n\approx$1594 and $n\approx$1109), but with distinct shapes: one retains isolated strengths while falling behind on other dimensions, whereas the other scores low across most axes. A separate cluster (*Exploitative Boss*, n=1236) displays strong imbalance, reflecting poor performance in key areas such as support and fairness. Furthermore, the PCA scatter plot (bottom-left of Fig. 11) further supports these distinctions. The clusters form visible bands rather than random mixtures, suggesting that student evaluations reflect repeatable combinations of experiences rather than isolated single scores. Meanwhile, the bar chart (bottom-left of Fig. 11) shows that majors differ in their composition of supervisor types. Some fields contain a higher share of the top-performing cluster, while others include more weaker patterns.

Overall, the results point to three main conclusions: good supervisors tend to perform well across multiple dimensions rather than excelling in only one; low ratings arise from different underlying patterns and therefore require targeted responses.

Gender Differences in Supervisor Evaluation. Fig. 12 provides a comprehensive comparison of supervisor evaluations by gender, covering overall patterns, score distributions across dimensions,

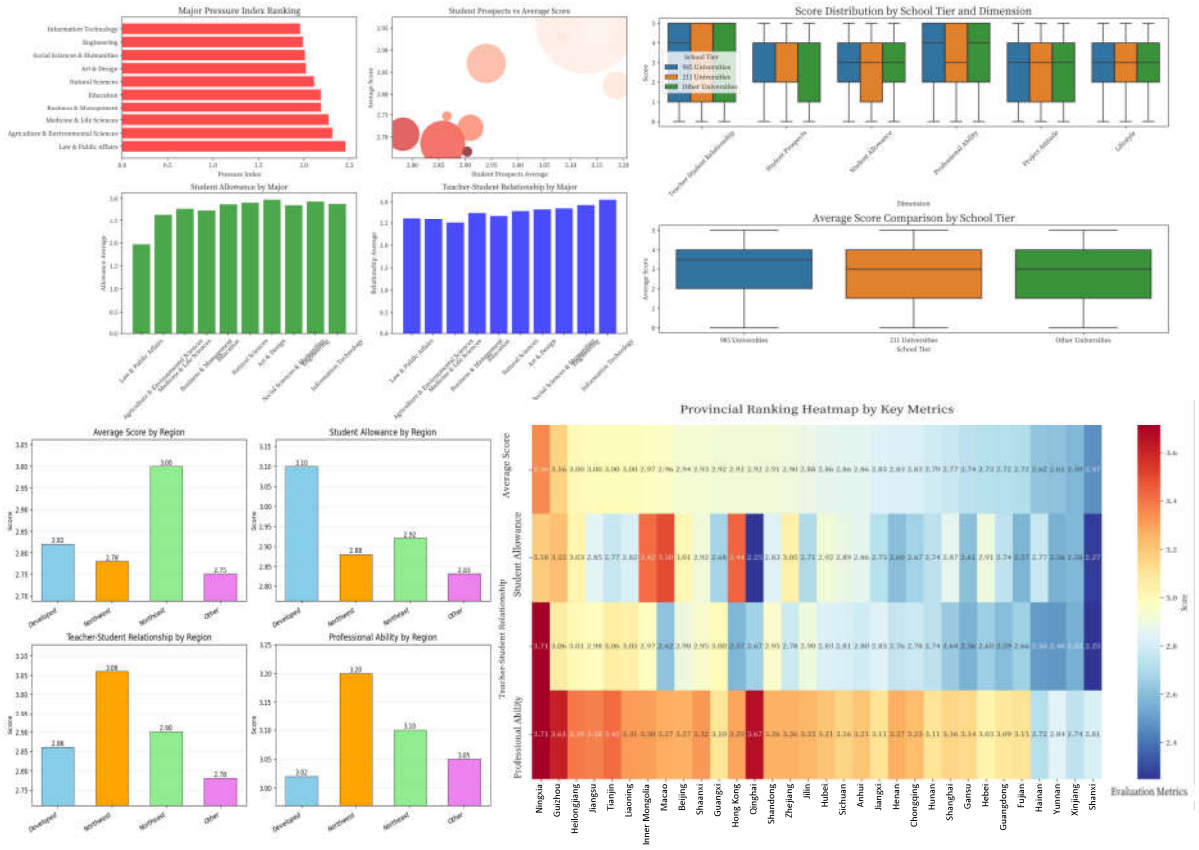

Figure 10: Regional, institutional, and disciplinary differences in supervisor evaluations, which compares scores and key dimensions across regions, school tiers, and majors, and summarizes provincial patterns.

extreme ratings, and differences across majors. At the aggregate level, male and female supervisors exhibit highly similar average profiles. The radar chart shows nearly overlapping shapes across all six dimensions, indicating no large gender gap in mean evaluations of relationship quality, professional ability, project attitude, or lifestyle.

Despite similar averages, distributional differences emerge across several panels. Male supervisors appear more frequently in the highest score category, particularly for *professional ability* and *project attitude*, while female supervisors are more concentrated in the middle score range. Boxplots and histograms of overall ratings further reflect this pattern: medians are comparable, but male supervisors show heavier upper tails, whereas female supervisors display tighter distributions with fewer extreme high values.

These differences are most visible at the extremes of the rating scale. The comparison of extreme evaluation ratios shows that female supervisors receive a higher proportion of low scores (≥ 2), while male supervisors receive a higher proportion of high scores (≥ 4). This asymmetry is consistent across dimensions and majors, suggesting that gender differences in supervisor evaluation are driven primarily by how students allocate extreme ratings rather than by systematic differences in average performance.

Traditional vs. High-Tech Engineering: Evidence of Internal Ecosystem Fracture. Fig. 13 provides a multi-perspective comparison between *Traditional Engineering* and *High-Tech Engineering*, revealing both shared patterns and subtle structural differences across key dimensions.

The radar chart suggests that the two groups are broadly comparable in their overall profiles, with only small deviations in project attitude, student prospects, professional level, and student subsidy. This

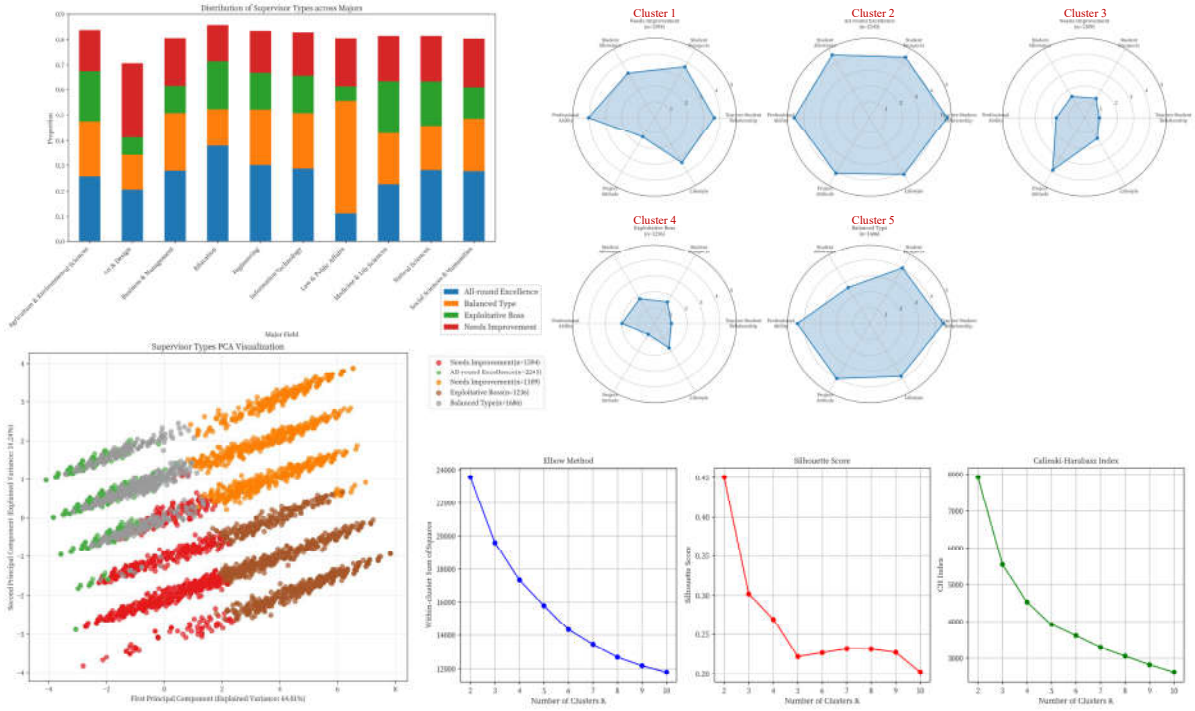

Figure 11: Clustering-based supervisor archetypes. The results show: *Distribution of Supervisor Types across Majors* (top-left); *Radar Charts of Supervisor Clusters* (top-right); *Supervisor Types PCA Visualization* (bottom-left); *Cluster Number Selection Metrics* (bottom-right).

overall similarity is further supported by the effect-size panel, where all Cohen's d values remain close to zero and well below conventional thresholds for medium (0.3) or large (0.5) effects, indicating that the between-group differences are statistically small in practical magnitude. Despite the limited effect sizes, the distributional and variability analyses highlight meaningful internal distinctions. The histogram of student prospects scores shows that High-Tech Engineering exhibits a stronger concentration at higher score levels (particularly around 4–5), implying a more optimistic outlook among students. Meanwhile, the interdisciplinary variability comparison indicates that Traditional Engineering has consistently higher coefficients of variation for both student prospects and student subsidy. In contrast, High-Tech Engineering appears relatively more homogeneous, which may reflect more standardized resource allocation and student expectations across departments. Finally, the condition-based resilience boxplots show that both engineering categories experience a substantial upward shift in resilience under normal conditions compared with adverse conditions, but the overall pattern is similar across groups.

Taken together, these results suggest that the primary contrast between *Traditional and High-Tech Engineering* lies less in average performance levels and more in the dispersion and internal consistency of student experiences, which may serve as a key indicator of ecosystem stability and potential fracture risk. **Strict Mentorship or Exploitation: Nonlinear Pressure–Benefit Trade-off.** To examine whether supervisory "push" improves student outcomes or instead becomes counterproductive, we construct a pressure proxy (*Push_Intensity*) by reversing the supervisor project-attitude score (6 − attitude). We then model the relationship between *Push_Intensity* and *Student Prospects* within three representative disciplines (Information Technology, Engineering, and Natural Sciences), comparing a linear baseline against a quadratic specification to test for an inverted-U pattern. And the results are illustrated in Fig. 14.

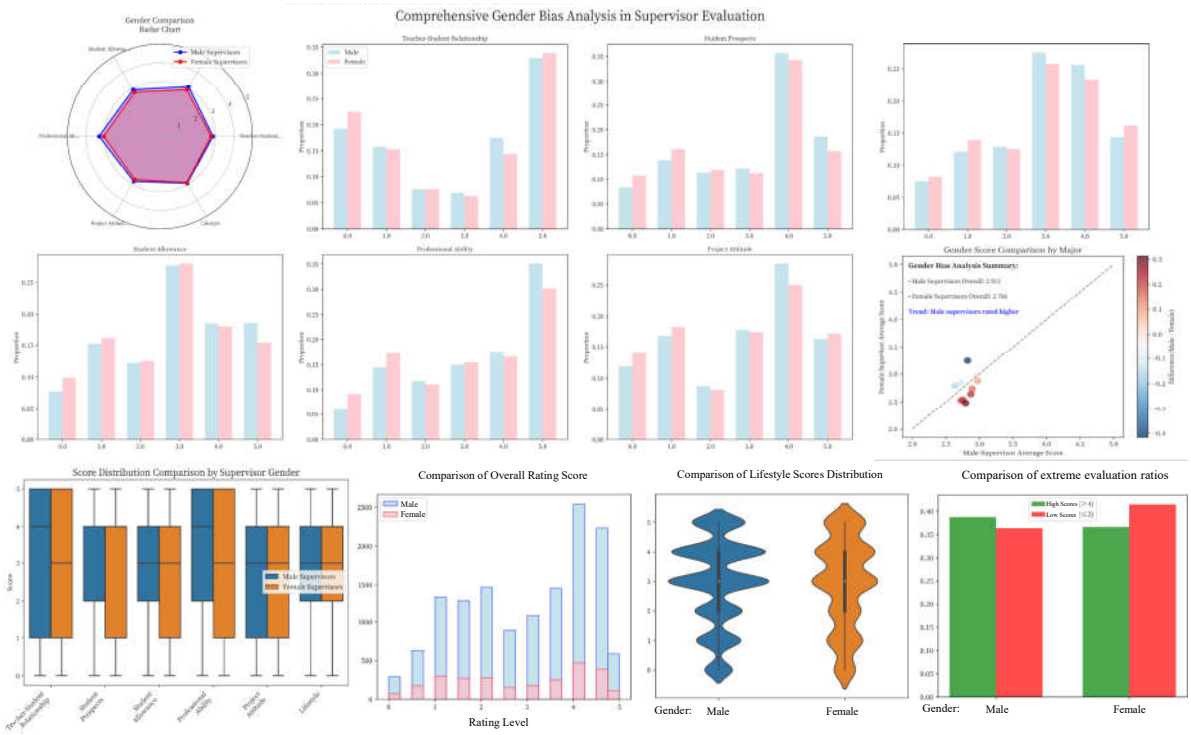

Figure 12: Gender-based comparison of supervisor evaluations. The figure shows average profiles across dimensions, score distributions by gender, overall rating patterns, lifestyle score distributions, extreme rating ratios, and gender differences across academic majors.

Across all three majors, the quadratic model consistently provides a better fit than the linear model, with R^2 values in the range of approximately 0.12–0.15, suggesting that the pressure–benefit relationship is not purely monotonic. The fitted curves show a shared structure: student prospects improve under mild-to-moderate supervision intensity, but decline once supervision becomes overly strict. The estimated peak points occur around a supervision intensity of roughly 2–2.5, indicating an empirically identifiable "optimal pressure" zone rather than a simple "more pressure is better" mechanism.

Beyond average trends, we further assess tolerance by comparing breakdown rates (prospects ≤ 2) under moderate pressure versus high pressure. The heatmap reveals that the risk of low prospects rises substantially under high-pressure conditions across all disciplines, with the strongest sensitivity observed in Natural Sciences. Taken together, the results support a nuanced interpretation: moderate push may be beneficial, but excessive supervision is associated with diminishing returns and a higher probability of adverse student outlook.

Allowance Inequality in 985 vs. Non-985 Universities: Distribution, Geography, and Matthew Effects. Fig. 15 consolidates the key evidence on student-allowance inequality between 985 and non-985 universities across distributional, geographic, and disciplinary perspectives. At the aggregate level, the score distribution and banded shares indicate that non-985 universities have a heavier lower-score tail and a slightly more polarized structure, consistent with stronger internal stratification. This pattern is corroborated by inequality metrics: non-985 institutions show higher Gini (0.314 vs. 0.290) and higher coefficient of variation (0.558 vs. 0.516), suggesting that dispersion within non-985 universities is systematically larger even if average levels may appear similar.

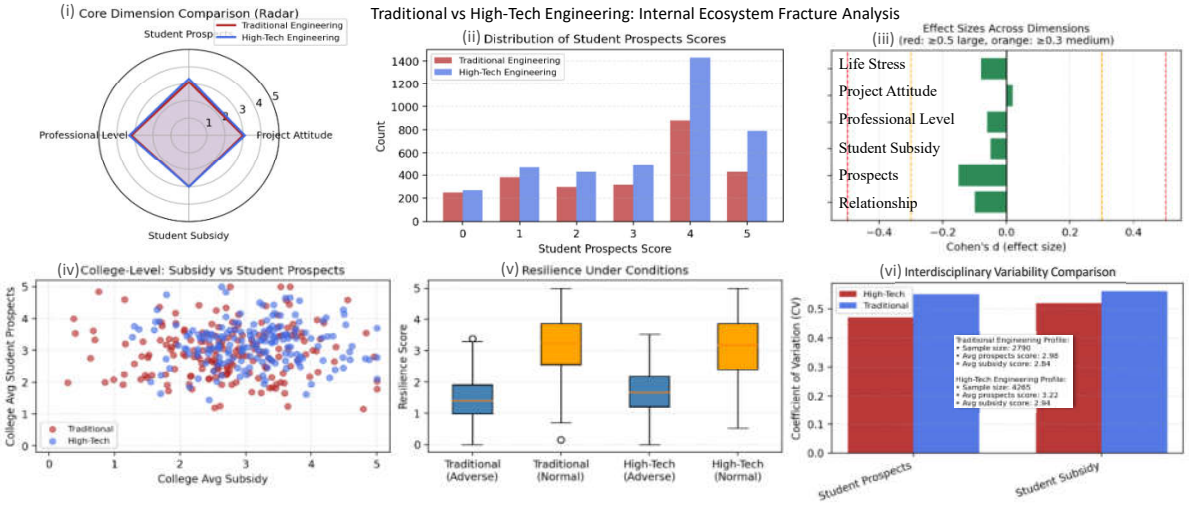

Figure 13: Multi-perspective comparison between *Traditional and High-Tech Engineering* across core dimensions. The results integrate (i) a radar chart of mean scores, (ii) the distribution of student prospects scores, (iii) effect sizes (Cohen's d) across dimensions, (iv) a college-level scatter plot of subsidy versus prospects, (v) resilience under adverse and normal conditions, and (vi) interdisciplinary variability for key indicators.

Beyond global inequality, the extreme-share panel provides a risk-oriented view: the proportion of very low allowance (score $\leq$ 1.5) is higher in non-985 universities, implying a greater prevalence of students facing resource scarcity. In contrast, very high allowances (score $\geq$ 4.5) differ less between groups, indicating that the inequality gap is driven more by the lower tail than by elite concentration alone.

Finally, the provincial and major-level analyses reveal structured heterogeneity. The top-province ranking shows that Jiangxi attains the highest mean allowance within 985 universities (3.50), but small sample sizes in some provinces caution against over-interpretation. The 985 vs. non-985 heatmap suggests that geographic gradients are not uniform across institution types. At the disciplinary level, the Matthew-effect scatter indicates that 985 advantages concentrate in specific majors, with the strongest advantage observed in Social Sciences & Humanities (effect strength $\approx$ 0.11). Overall, the results imply that the primary inequality problem is more severe internal dispersion within non-985 universities, while 985-related advantages manifest as localized geographic and disciplinary premiums rather than a universal uplift.

The Cost of "Nice" Supervision: Rapport, Rigor, and Discipline-Specific Returns. To test whether being a "nice" supervisor comes with an academic cost, we split supervisors into three groups. *Nice* supervisors are those rated high on teacher–student relationship and project attitude (i.e., low coercion). A stricter *Buddha-style* subset also requires a high lifestyle score. All remaining cases form the *Normal* group.

Fig. 16 summarizes the comparison using a small set of high-signal panels. The radar chart first shows that niceness is not one single dimension: it is a combination of stronger rapport and a less pushy project style. Importantly, these traits do not coincide with weaker academic ability. Instead, nice (and buddha-style) supervisors are often rated as equally, or even more, professionally competent, while also producing better student prospects. The distributional plots confirm that this is not driven by a few extreme cases. Professional-ability scores shift to the right for nice supervisors, and the prospects boxplot suggests both higher typical outcomes and fewer low-end failures. The relationship–ability scatter further supports this pattern: good rapport and strong academic ability tend to appear together rather than trade off against each other. Finally, the major-level panels add an important boundary condition. The share of

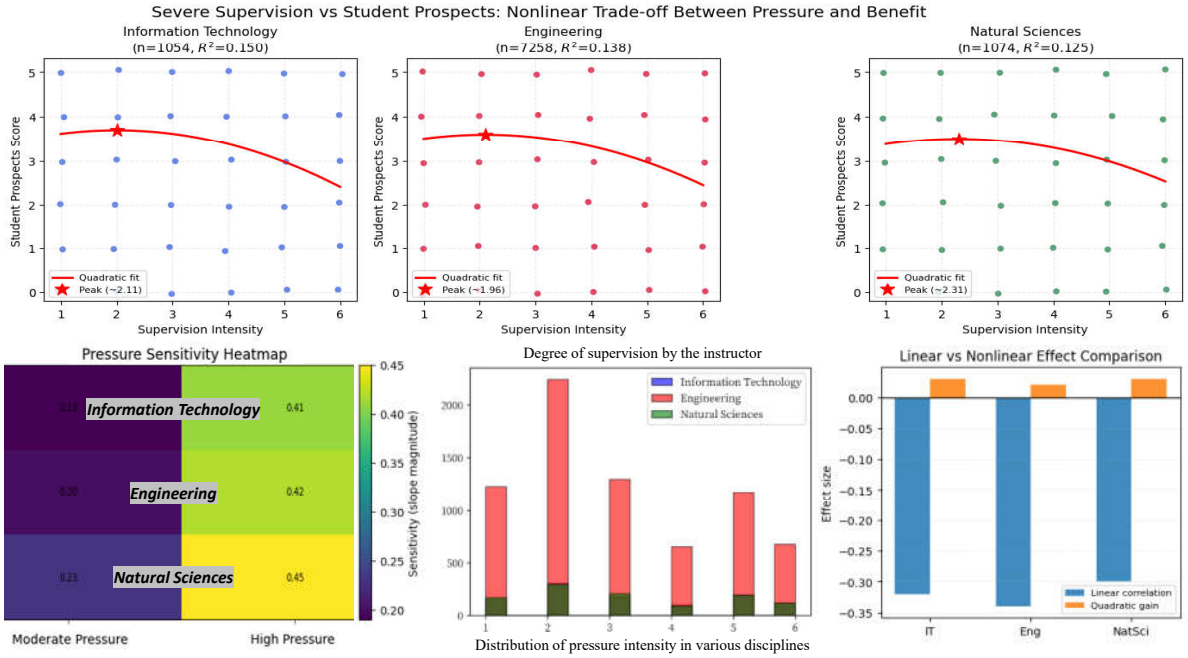

Figure 14: Nonlinear relationship between supervisory push intensity and student prospects across disciplines. Push intensity is constructed as 6 − Supervisor's Project Attitude, where larger values indicate stronger supervisory pressure. The top row shows scatter plots with quadratic fits and estimated peak points for Information Technology, Engineering, and Natural Sciences (sample sizes and R^2 reported). The bottom panels summarize discipline-level pressure sensitivity under moderate vs. high pressure and compare linear correlations against the incremental explanatory power of the quadratic term.

nice supervisors and their measured advantages differ across disciplines, meaning that "being nice" is not equally easy or equally rewarding in every field.

Overall, the results suggest that supportive mentoring is not inherently "soft" in many contexts, it aligns with both rigor and better student outcomes.

The relationship between descriptive findings and model-based results. It is important to note that the descriptive findings presented in Section 5.1 do not rely on predictive modeling and therefore reflect empirical patterns observed in the data. However, such patterns may be influenced by group-wise comparisons or localized observations. The explainable machine learning framework complements these findings by identifying which factors have stable, consistent, and non-linear effects on the overall evaluation across the full dataset. Therefore, the role of the model is not to replace descriptive analysis, but to validate, refine, and prioritize the observed patterns. This combined approach enhances the robustness and interpretability of the conclusions.

5.2 Explainable Machine Learning Analysis

In contrast to the descriptive analyses in Section 5.1, this subsection presents model-based and explainable analyses using the IMA framework. The goal is to identify which features have stable and

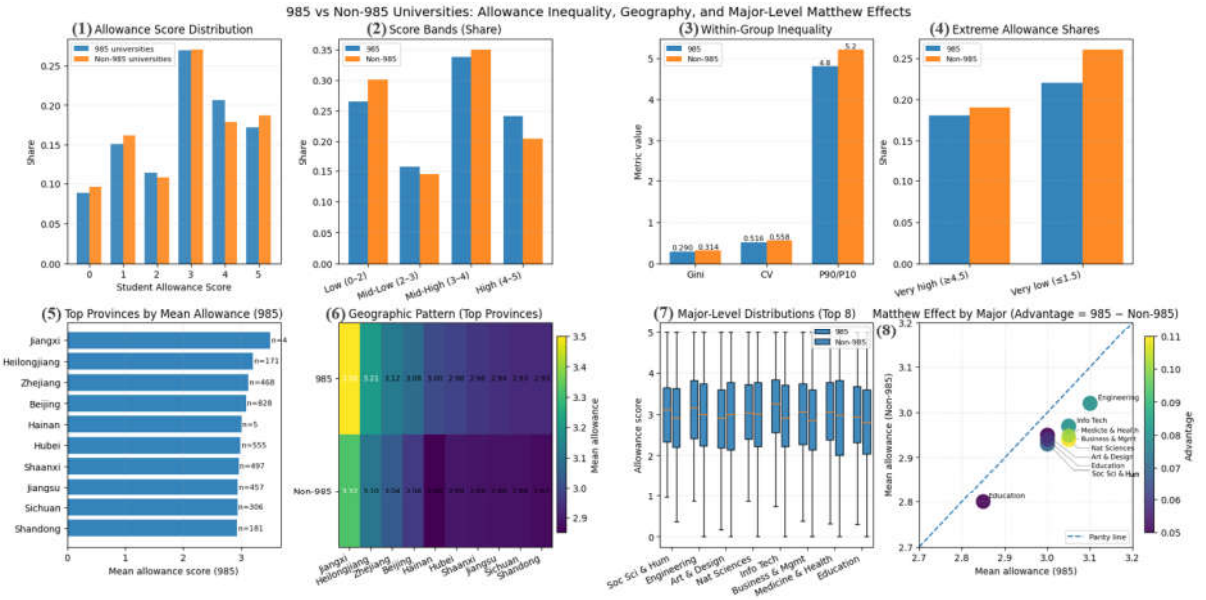

Figure 15: Comparison of student allowance inequality between 985 and non-985 universities. Panels summarize (1) normalized allowance-score distributions, (2) score-band shares, (3) within-group inequality metrics (Gini, CV, and P90/P10), (4) extreme-share rates (very high ≥ 4.5 and very low ≤ 1.5), (5) top provinces in 985 by mean allowance (with sample sizes), (6) province-level mean allowance heatmap for 985 vs. non-985, (7) major-level allowance distributions (top 8 majors), and (8) major-level Matthew effects (advantage = mean_{985} - $\text{mean}_{non-985}$).

systematic influence on the overall mentorship evaluation, to quantify their relative importance, and to uncover potential non-linear and interaction effects that cannot be captured by descriptive statistics alone[18].

Specifically, the credibility of model-based conclusions in this study is assessed from four complementary perspectives: **predictive validity**, measured by out-of-sample regression performance; **cross-model consistency**, by comparing whether XGBoost, LightGBM, and Random Forest identify similar key drivers[19]; **robustness to preprocessing choices**, including missing-value strategies and feature handling; and **substantive consistency**, namely whether explainable results align with descriptive patterns observed in the earlier empirical analyses.

5.2.1 Data Preprocessing and Feature Engineering

Before model training, we implemented a unified data preprocessing and feature engineering pipeline to ensure data quality, consistency, and suitability for explainable machine learning analysis. Given the heterogeneous nature of large-scale anonymous evaluations, the dataset contains both numerical and categorical variables, as well as sporadic missing or noisy entries.

Missing values were handled according to variable type. For categorical features, missing entries were assigned to a unified "Unknown" category, while numerical features were cleaned by replacing infinite values with missing values and imputing them using the median. Categorical variables describing

[18] It should be noted that the purpose of the proposed framework is not to benchmark predictive performance against alternative algorithms, but to demonstrate the feasibility and reliability of an interpretable and scalable analytical pipeline for large-scale mentorship evaluation data.

[19] Multiple models (e.g., XGBoost, LightGBM, Random Forest) are tested to ensure robustness.

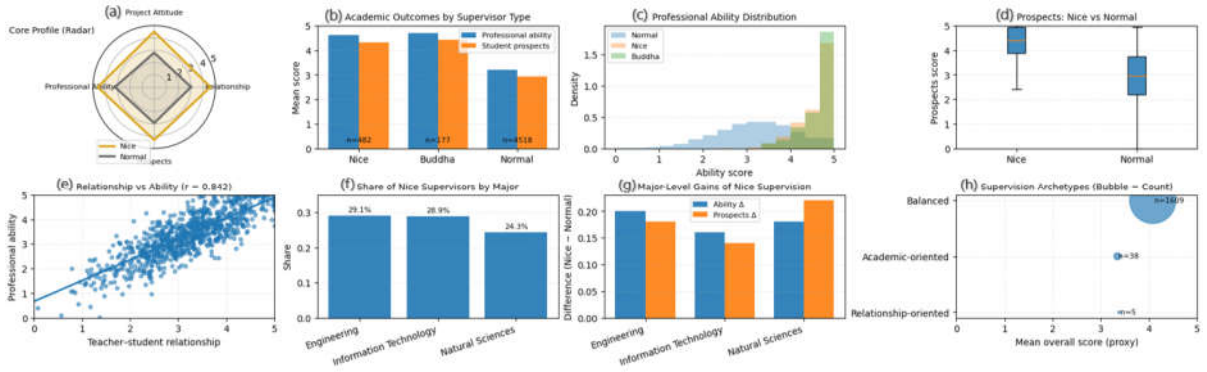

Figure 16: A detailed fine-grained analysis of *Nice* supervisors. Part. (a) a radar profile of core dimensions (relationship, project attitude, professional ability, prospects); (b) mean academic outcomes (ability and prospects) across supervisor types with sample sizes; (c) distributional shifts in professional ability; (d) prospects dispersion for nice vs normal; (e) the relationship–ability association with a fitted trend and correlation; (f) the share of nice supervisors by major; (g) major-level gains (Nice–Normal) for ability and prospects; and (h) an archetype bubble view summarizing supervision styles.

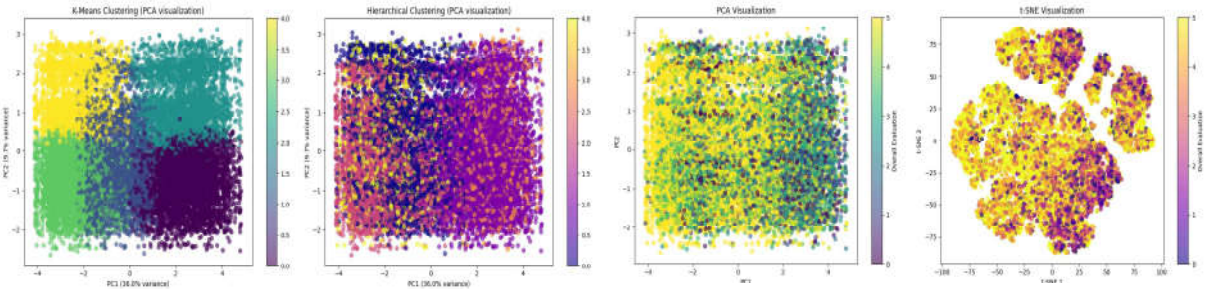

Figure 17: Clustering and dimensionality reduction by K-means, hierarchical clustering, PCA, and t-SNE.

institutional, demographic, and disciplinary attributes, including school category, university, department, supervisor, gender, province, and major, were transformed using label encoding. The core numerical feature set was constructed from structured evaluation indicators reflecting mentorship quality, including teacher–student relationship, student prospects, financial support, supervisor professional ability, project attitude, and lifestyle-related factors, along with an aggregated average score.

Finally, the processed dataset was split into training and testing subsets using an **80/20** ratio, with the overall evaluation score serving as the prediction target. This preprocessing framework establishes a consistent and transparent foundation for model comparison and subsequent explainable analysis.

5.2.2 Clustering and Dimensionality Reduction Analysis

Based on the clustering and dimensionality reduction results visualized in Fig. 17, we observe the insights into the underlying structure of the data.

Specifically, for the first two subfigures in Fig. 17, i.e., K-Means[20] & Hierarchical Clustering using PCA visualization[21], the K-means based result exhibits a clear separation between distinct groups in the reduced

[20] K-Means: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.
[21] PCA Visualization: https://plotly.com/python/pca-visualization/.

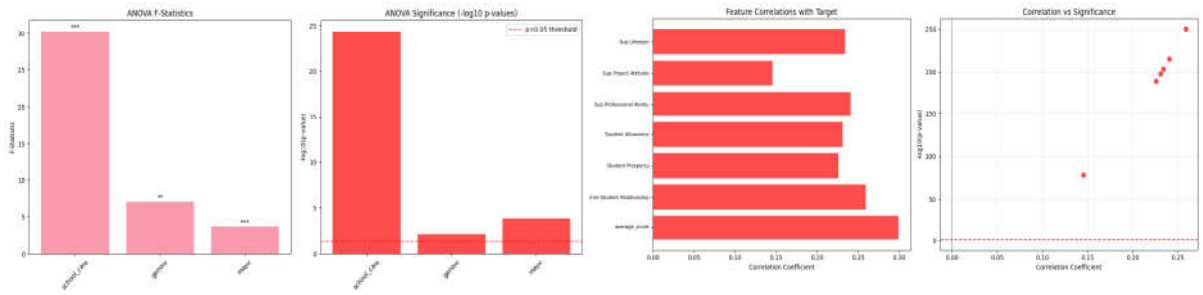

Figure 18: ANOVA analysis results, showing F-statistics and significance markers for categorical variables, and p-value significance using -log10(p). In the left panel, the significance markers atop each bar denote the following: '***' for $p < 0.001$, '**' for $p < 0.01$. These markers are derived from the ANOVA p-values associated with each categorical variable.

two-dimensional space (PC1 and PC2). The resulting clusters, represented by different colors, appear to be well-separated, with relatively few points overlapping. Similar to K-means, the hierarchical clustering method also provides an interesting view of the data's grouping structure. However, the clusters appear to be less well-defined compared to the K-means results, with more overlap between groups. For the last two parts of Fig. 17, the general PCA plot does not show distinct clusters, indicating that while PCA successfully reduces the data to two components, the data may not inherently separate well in lower-dimensional space without clustering methods. Also, the t-SNE plot offers a more compelling visualization, with distinct clusters appearing more clearly separated than in the PCA plot.

Overall, these clustering and dimensionality reduction results provide a strong basis for evaluating the effectiveness of different machine learning methods. These results, which reveal the characteristics of the data, can inform further analysis and guide model selection.

5.2.3 Statistical Analysis: ANOVA and Correlation

Fig. 18 presents the results of the statistical analysis, including ANOVA tests and correlation analysis, conducted to investigate the relationships between categorical variables and the overall evaluation score.

The first plot in Fig. 18 visualizes the F-statistics for each categorical variable tested using ANOVA, which evaluates whether the means of different groups significantly differ from one another. The F-statistics are accompanied by significance markers indicating the strength of the relationship between the categorical variables and the overall evaluation. The ANOVA results suggest that certain categorical variables, such as 'school_cate' and 'major', exhibit strong relationships with the evaluation scores. The second plot displays the -log10 of the p-values for each variable tested in the ANOVA analysis. A higher -log10(p-value) indicates stronger statistical significance. The remaining two plots show the results of correlation analysis, where the correlation coefficients between numerical features and the overall evaluation score are calculated. The third plot visualizes the correlation coefficients for each feature, with colors representing the statistical significance of each correlation. Significant correlations are marked in red. The fourth plot displays a scatter plot of the correlation coefficients against the -log10(p-values), providing a clearer view of the relationship between correlation strength and statistical significance.

In summary, these statistical results help to identify key factors that influence the overall evaluation score and guide further exploration of feature interactions.

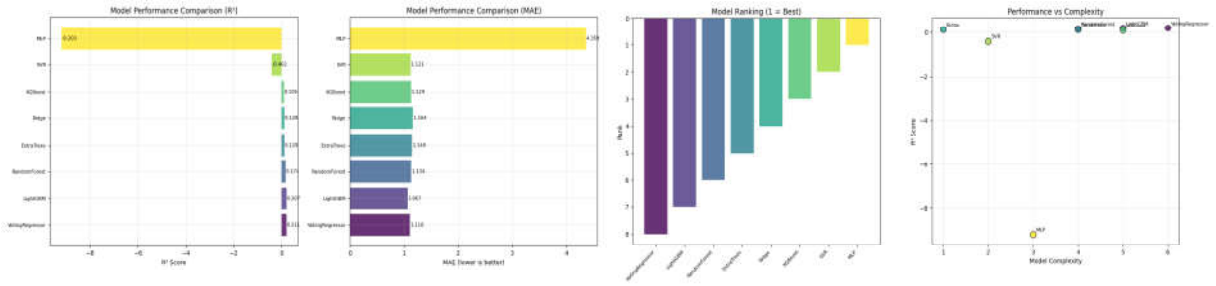

Figure 19: Performance ranking of evaluated machine learning models based on predictive accuracy. Models are ranked according to their R^2 scores on test set, where higher values indicate better explanatory power.

5.2.4 Model Selection and Performance Comparison

Before model training, we conducted a comprehensive comparison across a diverse set of machine learning regression models to select candidates that balance accuracy, robustness, and interpretability for subsequent explainable analysis.

Based on the structured evaluation dataset described in the previous section, we formulated the task as a supervised regression problem[22], where the overall mentorship evaluation score served as the prediction target, and multi-dimensional indicators describing guidance intensity, academic support, professional competence, and interaction quality were used as input features.

Specifically, we evaluated a broad spectrum of representative regression models, including linear models (Ridge regression), kernel-based methods (Support Vector Regression), neural networks (Multi-Layer Perceptron), and ensemble learning approaches such as Random Forest, Extra Trees, Gradient Boosting-based models (XGBoost and LightGBM), as well as an ensemble Voting Regressor. Model performance was assessed using multiple complementary metrics, including the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). Fig. 19 summarizes the overall performance ranking of all evaluated models based on their predictive results on the held-out test set. In particular, XGBoost, LightGBM, and Random Forest achieved superior R^2 scores and lower prediction errors, demonstrating strong capability in capturing complex, non-linear interactions among mentorship-related features. In contrast, simpler models such as Ridge regression exhibited limited expressive power, while the neural network model showed unstable performance and inferior generalization.

Considering both predictive effectiveness and compatibility with post-hoc interpretability techniques, we selected XGBoost, LightGBM, and Random Forest as the core predictive models for subsequent explainable machine learning analysis. These models can integrate seamlessly with model-agnostic and model-specific explanation tools such as SHAP.

During training, we further examined the learning behavior and generalization characteristics of these models using learning curve analysis. Fig. 20 illustrates the learning curves of Ridge regression, Random Forest, and LightGBM in terms of R^2 scores on both training and validation sets as the training sample size increases. Ridge regression exhibits closely aligned training and validation curves with a consistently small performance gap, indicating stable convergence and low variance but limited predictive capacity. In contrast, Random Forest achieves very high training performance while maintaining substantially lower

[22] To ensure fairness and robustness, all models were trained and evaluated under an identical preprocessing and train–test split protocol.

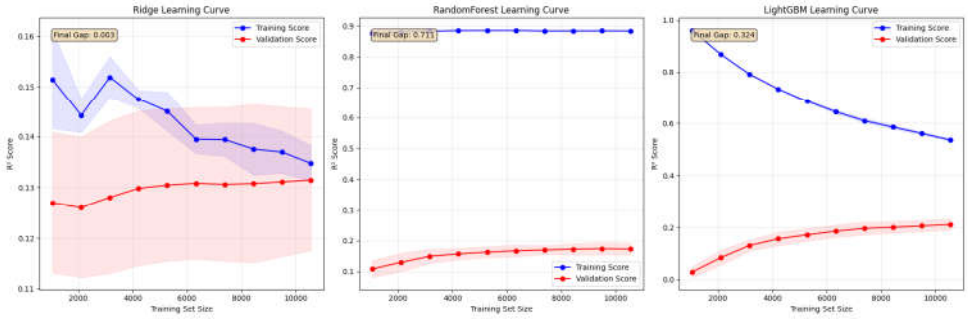

Figure 20: Learning curves of representative models (Ridge regression, Random Forest, and LightGBM), showing training and validation R^2 scores as a function of training set size.

validation scores, resulting in a pronounced and persistent generalization gap. LightGBM demonstrates a balanced learning. Although its training performance decreases with increasing data size, the validation performance steadily improves and gradually converges, yielding a moderate final gap between training and validation scores.

Overall, the results further support the selection of tree-based ensemble methods, particularly LightGBM and XGBoost, as suitable candidates for subsequent explainable machine learning analysis, due to their favorable trade-off between predictive accuracy and generalization stability.

5.2.5 Global Explainability Analysis with SHAP

To provide a transparent and model-consistent interpretation of the predictive mechanisms underlying the selected ensemble models, we conducted an advanced explainable analysis based on SHAP. This analysis focuses on tree-based models with strong predictive performance, including Random Forest, LightGBM, and XGBoost. SHAP values were computed on the test set using model-specific TreeExplainers.

Fig. 21 presents the global SHAP analysis results from two complementary perspectives. The upper row shows SHAP summary plots for each model, where each point represents an individual sample and the color indicates the relative magnitude of the corresponding feature value. Positive and negative SHAP values reflect the direction and strength of each feature's contribution to the predicted overall evaluation score. The lower row reports the corresponding feature importance rankings based on the mean absolute SHAP values, highlighting the most influential predictors for each model.

From the upper row of Fig. 21, it reveals several consistent and informative patterns. First, the distribution of SHAP values demonstrates pronounced asymmetry for the most influential features, indicating strong non-linear effects rather than simple monotonic relationships. For example, higher values of the average evaluation score and student allowance are predominantly associated with positive SHAP values, suggesting that these factors consistently contribute to increased overall evaluation predictions. Second, the color gradients within the SHAP distributions highlight substantial heterogeneity in feature effects. Even for highly ranked features, both positive and negative SHAP values are observed, implying that their influence depends on contextual combinations with other variables. This pattern is particularly evident for interaction-related features, such as the relationship–ability interaction, which exhibits wide SHAP dispersion in all three models. Third, despite architectural differences among Random Forest, LightGBM, and XGBoost, the relative ordering and directional tendencies of key features remain highly consistent. Institutional and demographic encodings (e.g., department, university, and province) display comparatively

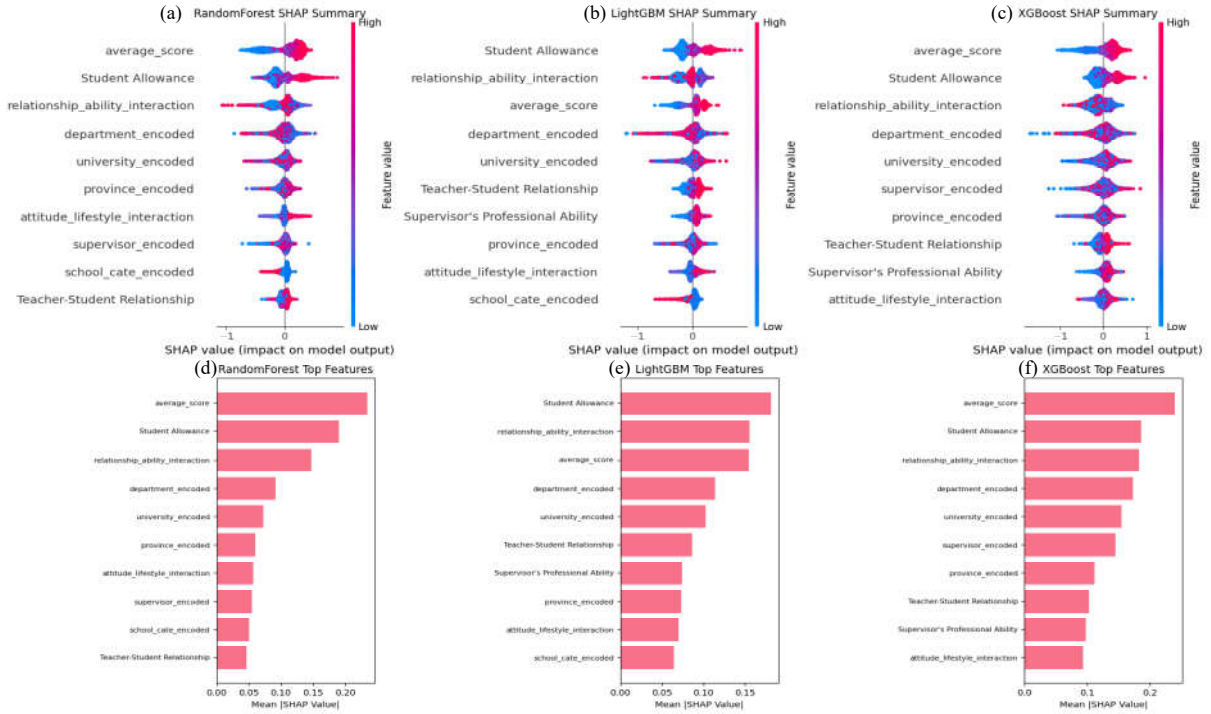

Figure 21: Global SHAP analysis for Random Forest, LightGBM, and XGBoost models. The upper panels show SHAP summary plots illustrating the positive and negative contributions of key features, while the lower panels report feature importance rankings based on mean absolute SHAP values.

smaller but stable contributions, indicating that while contextual background plays a role, it does not dominate the predictive process.

Despite minor variations in feature ranking across models, the overall SHAP patterns are highly consistent, reinforcing the robustness of the identified key factors. This cross-model agreement strengthens the credibility of the explainable results and provides reliable evidence for subsequent interpretation of mentorship mechanisms.

5.2.6 Feature Dependence and Interaction Analysis with SHAP

To further investigate how key features interact and jointly influence model predictions, we conducted a detailed SHAP dependence and interaction analysis for the selected ensemble models. Unlike global importance rankings, this analysis focuses on feature-wise response patterns and interaction effects, enabling a more fine-grained interpretation of the underlying decision mechanisms. Fig. 22, Fig. 23, and Fig. 24 present the SHAP dependence results for Random Forest, XGBoost, and LightGBM, respectively.

For the Random Forest model (Fig. 22), strong non-linear and interaction-driven behaviors are observed. Institutional features such as province, university, and department encodings exhibit heterogeneous SHAP distributions, indicating that their effects are highly context-dependent rather than monotonic. Notably, the relationship–ability interaction feature shows wide SHAP dispersion, with both positive and negative contributions depending on the average score and supervisor professional ability. This suggests that mentorship effectiveness emerges from the interplay between relational quality and professional competence. In addition, student allowance and average score display clear positive trends, confirming their stabilizing and reinforcing roles in overall evaluation predictions.

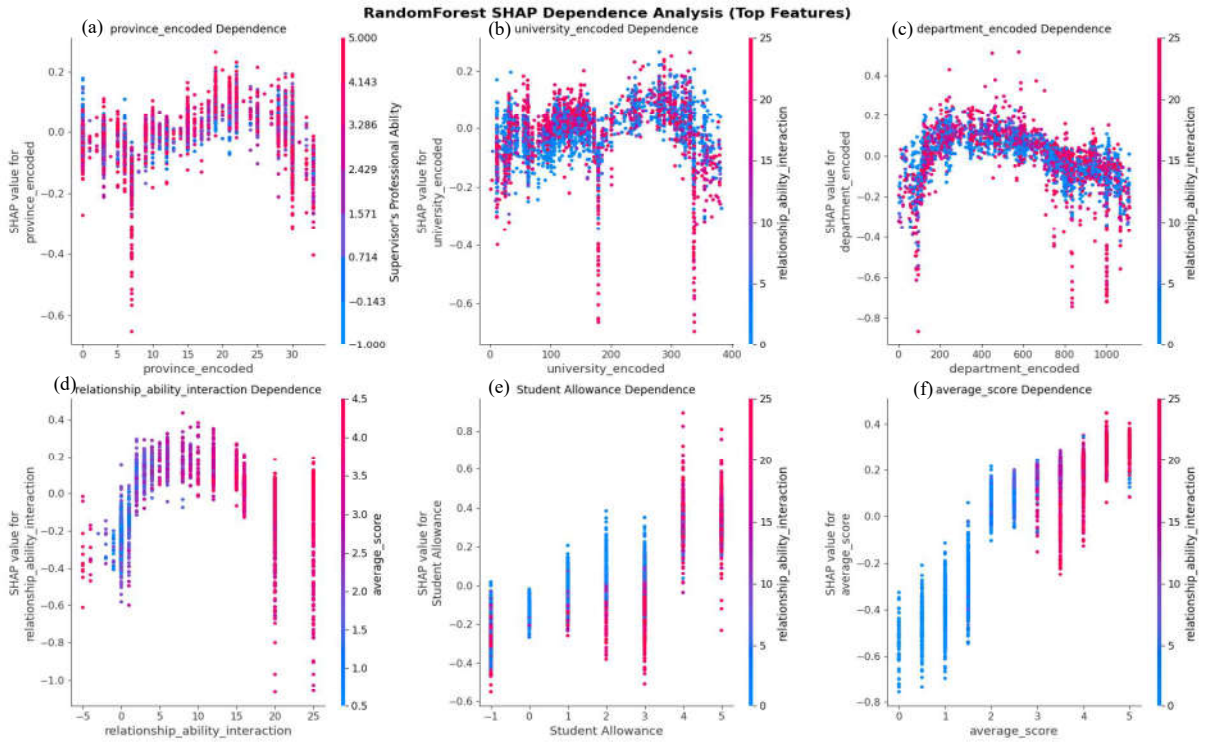

Figure 22: SHAP dependence and interaction analysis for the Random Forest model. The plots illustrate non-linear feature effects and interaction patterns among key variables, with color gradients indicating interacting feature values.

In the XGBoost model (Fig. 23), feature dependence patterns are more structured and monotonic. Student allowance and average score demonstrate nearly linear positive relationships with SHAP values, reflecting XGBoost's strong capacity to capture ordered effects. At the same time, teacher–student relationship and supervisor project attitude reveal pronounced interaction effects, as indicated by color gradients linked to professional ability and average score. These results highlight that while XGBoost emphasizes dominant main effects, it also preserves meaningful interaction structures between academic support and relational factors.

The LightGBM dependence analysis (Fig. 24) reveals similar but slightly smoother patterns. Core features such as average score, student allowance, and supervisor professional ability show consistent positive contributions with reduced variance, indicating stable generalization. Interaction effects remain evident, particularly between teacher–student relationship and professional ability, although their dispersion is narrower than in Random Forest. This behavior suggests that LightGBM achieves a balance between expressive non-linearity and robustness, yielding interpretable yet stable feature-response relationships.

Overall, the dependence and interaction analyses consistently demonstrate that mentorship evaluation outcomes are shaped by both dominant main effects and higher-order interactions. The cross-model agreement in these patterns further reinforces the reliability of the extracted explanations and provides strong evidence for the joint influence of relational, academic, and contextual factors.

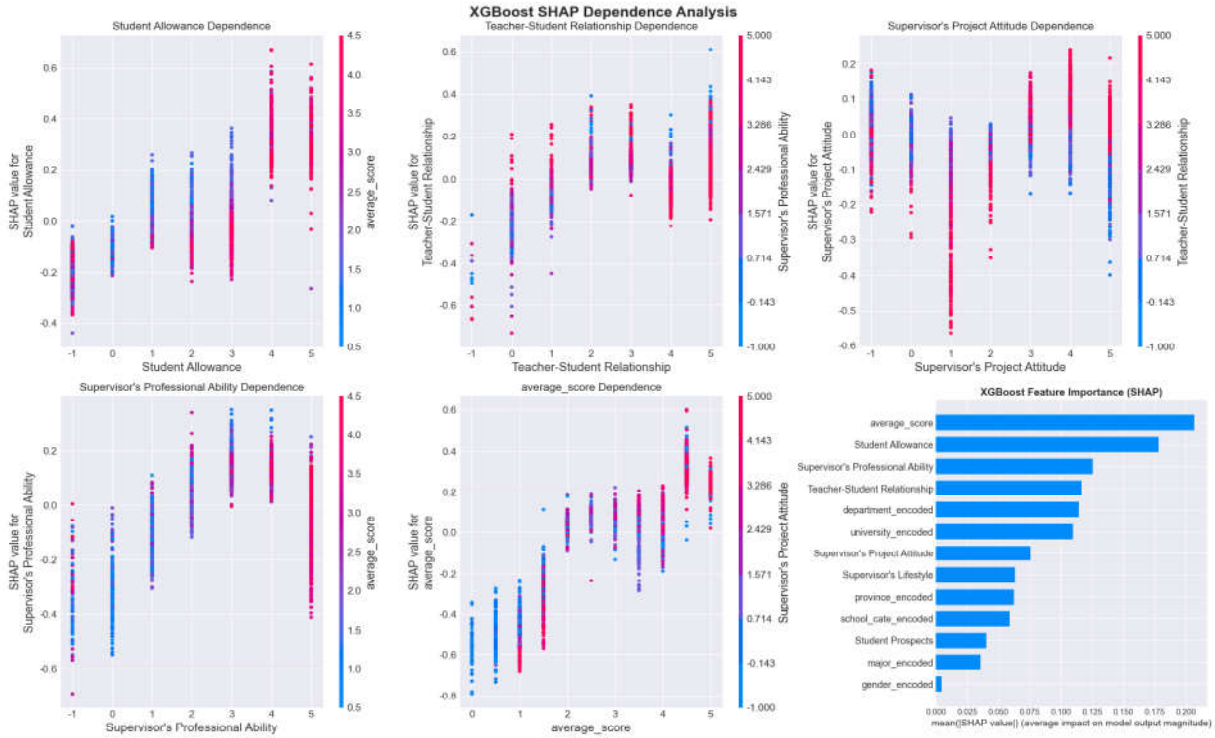

Figure 23: SHAP dependence analysis for the XGBoost model, showing feature-wise response patterns and interaction effects between core mentorship attributes and contextual variables.

5.2.7 Comprehensive SHAP Interaction and Nonlinear Effect Analysis

To further elucidate the structural mechanisms underlying the model predictions, we conducted a comprehensive SHAP-based interaction and nonlinear effect analysis. Fig. 25 integrates interaction strength estimation and feature-wise dependence visualization, providing a holistic view of how individual attributes and their combinations jointly influence the overall evaluation outcomes.

Fig. 25a reveals pronounced interaction effects among several key variables. In particular, the interaction between teacher–student relationship quality and supervisor professional ability exhibits the strongest intensity, indicating that the predictive contribution of professional competence is highly contingent on relational context. Other interaction pairs, such as student prospects with allowance and project attitude with lifestyle factors, further suggest that mentorship evaluation is driven by synergistic rather than additive effects.

Figs. 25b and c illustrate representative interaction patterns. The dependence between student prospects and allowance demonstrates that economic support amplifies the positive contribution of development prospects. Similarly, the interaction between relationship quality and professional ability shows that high competence yields substantial positive effects only when embedded within a favorable relational environment. These findings highlight the conditional nature of key predictors.

Figs. 25d–i depict feature-wise SHAP dependence plots, revealing strong nonlinear behaviors. Core attributes such as allowance, teacher–student relationship, and professional ability exhibit diminishing marginal effects and threshold-like patterns, while prospects and project attitude show more stable positive trends. Lifestyle-related features contribute modestly but consistently, acting as contextual modifiers rather

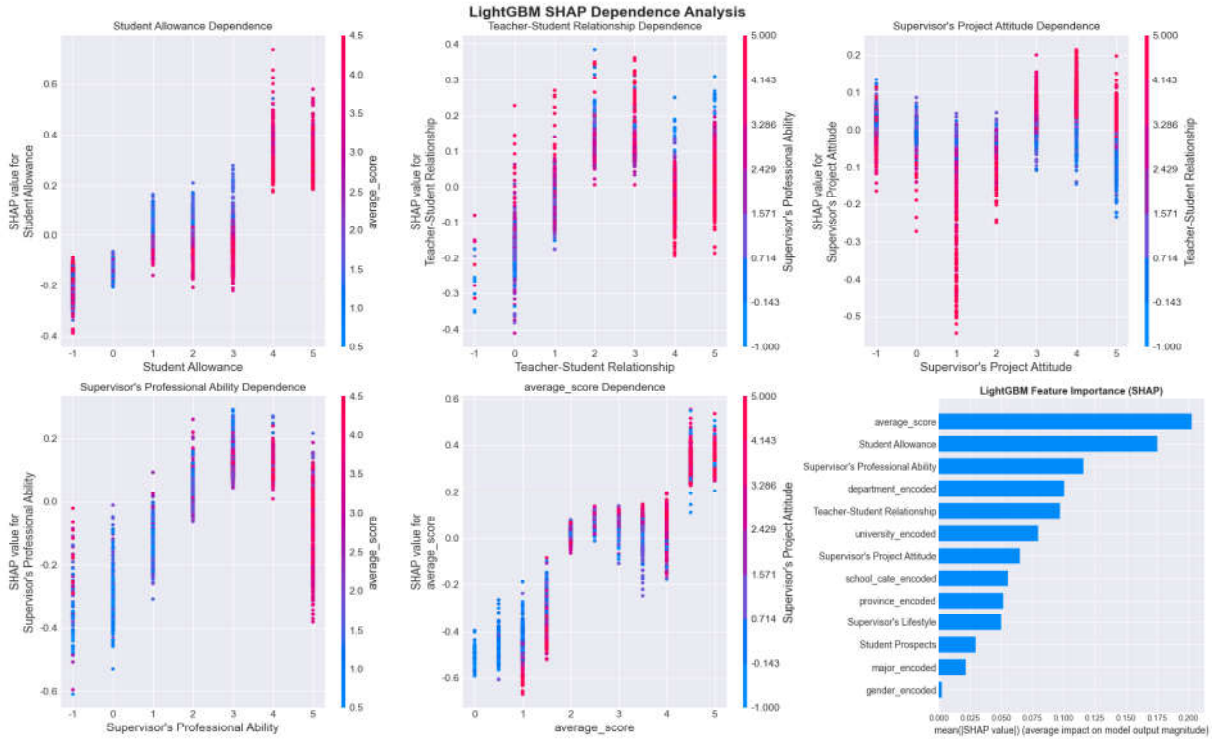

Figure 24: SHAP dependence analysis for the LightGBM model. The results highlight stable feature-response relationships and moderated interaction effects across key predictors.

than dominant drivers. Overall, this demonstrates that mentorship evaluation outcomes are shaped by a complex interplay of relational, academic, and contextual factors.

5.2.8 Local Explanation Analysis with LIME

While SHAP provides a comprehensive understanding of global feature importance and interaction structures, it does not directly explain individual prediction outcomes. To complement the global interpretability analysis and validate model behavior at the instance level, we further conducted a Local Interpretable Model-agnostic Explanations (LIME) analysis. LIME approximates the complex prediction model locally with an interpretable linear surrogate, enabling transparent interpretation of individual predictions.

Fig. 26 presents local explanation results for a diverse set of representative samples covering different evaluation outcomes. Each subfigure corresponds to a single instance, where the horizontal bars indicate the local contribution of specific feature conditions to the predicted overall evaluation score. Positive contributions (blue) increase the predicted score, whereas negative contributions (red) decrease it. Across samples with high evaluation scores, features such as high average score, strong teacher–student relationship, sufficient student allowance, and positive supervisor professional ability consistently exhibit strong positive contributions. In contrast, low-scoring samples are primarily driven by unfavorable conditions in these same dimensions, particularly weak relational quality and low average scores, which emerge as dominant negative contributors. This pattern highlights the asymmetric and context-dependent influence of key predictors at the individual level.

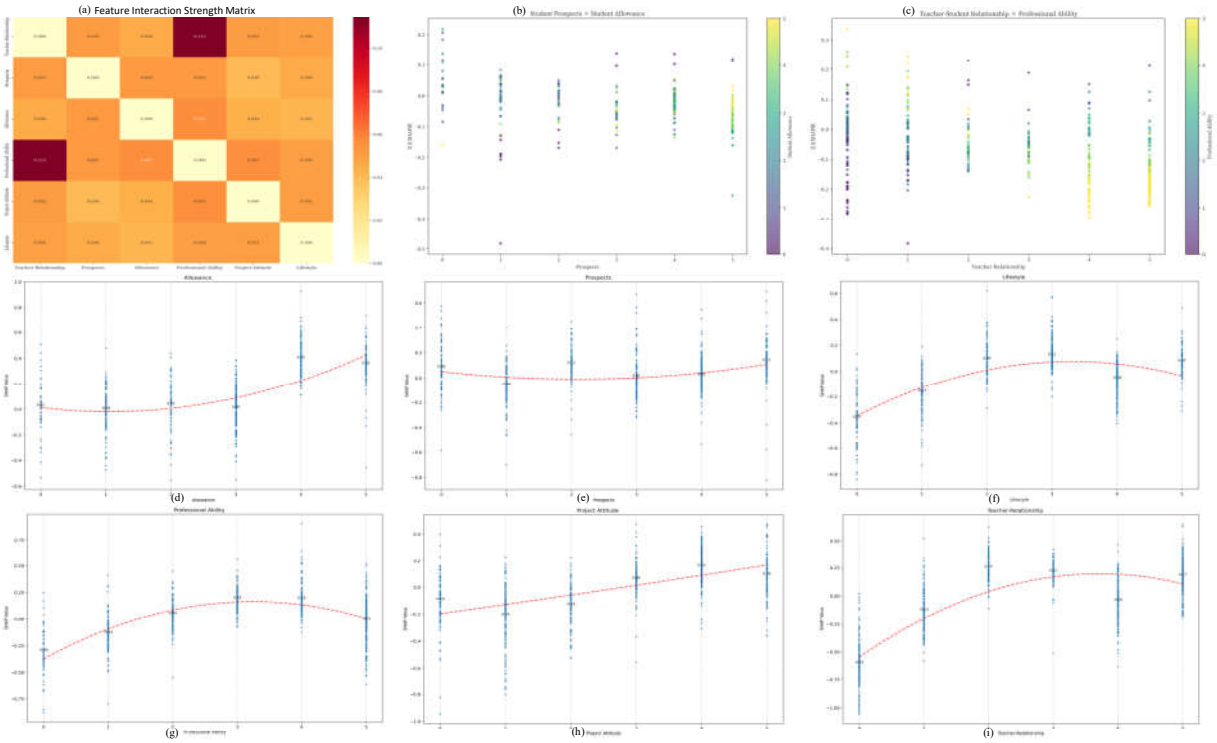

Figure 25: Comprehensive SHAP interaction and dependence analysis. (a) Feature interaction strength matrix based on SHAP interaction values. (b)–(c) Representative interaction dependence plots illustrating conditional effects between key features. (d)–(i) SHAP dependence plots showing nonlinear feature-response relationships for major predictors.

Notably, the direction and magnitude of certain features, such as student allowance and institutional encodings (e.g., school category, university, and department), vary substantially across instances. These features rarely dominate predictions independently but instead act as modulators that amplify or attenuate the effects of core academic and relational factors. This observation aligns closely with the interaction effects identified in the SHAP-based global analysis.

Overall, the LIME results demonstrate that the predictive model exhibits coherent and interpretable local behavior, with individual predictions driven by logically consistent combinations of features. The agreement between global SHAP explanations and local LIME interpretations further reinforces the reliability and transparency of the proposed explainable IMA framework.

Limitations. Despite the advantages of the above experimental procedure, this study has certain limitations in terms of methodological scope. Specifically, the analysis is primarily based on quantitative and machine learning techniques, which focus on pattern detection and feature importance rather than deep contextual interpretation.

Mixed-method approaches that incorporate qualitative data, such as interviews [6] or narrative analysis [17], may provide richer insights into the underlying causes of mentorship perceptions. Therefore, future research could integrate qualitative methods with the proposed framework to further enhance interpretability and contextual understanding.

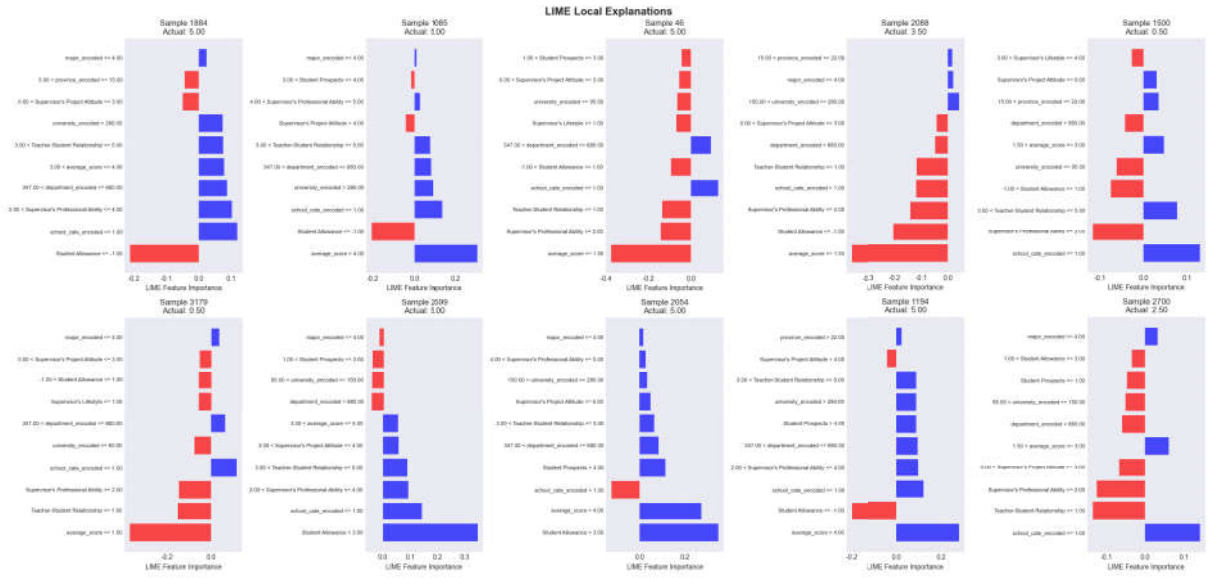

Figure 26: Local Interpretable Model-agnostic Explanations (LIME) for representative individual samples. Each subfigure illustrates the local feature contributions to the predicted overall evaluation score, where blue bars indicate positive contributions and red bars indicate negative contributions.

6 Conclusion

In this study, we introduce a scalable and transparent framework, *Interpretable Mentorship Analytics* (IMA), which is designed to decode the "black box" of SSR in Chinese postgraduate education using large-scale anonymous evaluations. Motivated by the rapid expansion of graduate enrollment and the accompanying intensification of supervisory power asymmetries, we address a key limitation of prior SSR research: the difficulty of extracting generalizable, system-level mechanisms from small samples, linear assumptions, or context-bound qualitative evidence. IMA integrates four sequential stages, large-scale data acquisition, privacy-preserving preprocessing, LLM-ensemble feature engineering, and explainable machine learning, into an end-to-end analytical pipeline. By consolidating multi-platform public reviews into a cleaned and de-identified dataset, and transforming unstructured narratives into comparable mentorship indicators, IMA enables quantitative analysis across institutions, disciplines, and regions while minimizing risks to individual privacy. The use of gradient-boosting models (e.g., **XGBoost** and **LightGBM**) provides strong capacity to capture *non-linear dependencies* and *high-order interactions* among mentorship dimensions, especially the patterns that are often invisible to conventional regression-based approaches. Crucially, IMA emphasizes interpretability as a first-class scientific objective. Through SHAP-based global attribution and LIME-based local explanations, the framework produces multi-scale insights: (i) identifying the most influential factors shaping overall mentorship outcomes, and (ii) explaining heterogeneous individual evaluations in a case-sensitive manner. By consolidating multi-platform public reviews into a cleaned, de-identified dataset and transforming unstructured narratives into comparable indicators, IMA enables quantitative analysis across institutions, disciplines, and regions while minimizing privacy risks. Gradient-boosting models (e.g., **XGBoost** and **LightGBM**) capture *non-linear dependencies* and *high-order interactions* among mentorship dimensions, patterns often invisible to conventional regression. Crucially, IMA prioritizes interpretability, using SHAP-based global attribution and LIME-based local explanations to produce multi-scale insights: identifying key factors shaping mentorship outcomes and explaining

individual evaluations in a case-sensitive manner. The main contribution of this work is not the proposal of a new standalone algorithm, but the construction of an integrated and explainable analytical framework that expands SSR research toward large-scale, data-driven, and policy-relevant applications. Overall, IMA offers a practical pathway for transforming anonymous evaluations into interpretable evidence, supporting data-informed reforms aimed at building a healthier, more sustainable postgraduate research environment.

Acknowledgement: This work is conducted by by Yang Gao and Weiqiang Jin during their joint research at Xi'an Jiaotong University. The corresponding author is Prof. Ziwei Zhang at Xi'an Jiaotong University.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions: The authors confirm contributions to this paper as follows: Yang Gao: Conceptualization, Methodology, Formal analysis, Writing—original draft, Writing—review & editing, Project administration, Investigation, Validation; Weiqiang Jin: Software, Resources, Data curation, Visualization, Writing—review & editing; Yajuan Nan: Formal analysis, Writing—review & editing; Yue Ma: Writing—review & editing; Biao Zhao: Writing—review & editing, Supervision; Ziwei Zhang: Investigation, Conceptualization, Project administration. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The full experimental code and data of IMA analysis process has been released on Github: https://github.com/SmileHappyEveryDay/tutor-evaluation-data-analysis.

Ethics Approval: This study was conducted using publicly available data on supervisor evaluations downloaded from the internet. All private information has been anonymized. As the data is publicly accessible and does not contain personally identifiable information, no additional ethics approval was required.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jiang Y, Liu C. Diversified influencing factors in supervisor-student relationships in graduate education and suggestions for countermeasures. Health Development and Policy Research. 2025;28(3):343-9. (In Chinese).
2. Xu Y, Liu J. Exploring and Understanding Perceived Relationships Between Doctoral Students and Their Supervisors in China. Humanities and Social Sciences Communications. 2023 Nov;10(1):829.
3. Raposa EB, Hagler M, Liu D, Rhodes JE. Predictors of close faculty-student relationships and mentorship in higher education: findings from the Gallup-Purdue Index. Annals of the New York Academy of Sciences. 2021;1483(1):36-49.
4. Zhang J, Wu M, Zhang G. The Influence of Supervisor–Postgraduate Relationship on Master's Students' Research Learning Engagement—The Mediating Effect of Academic Aspiration. Behavioral Sciences. 2024;14(4). Available from: https://www.mdpi.com/2076-328X/14/4/334.
5. Tikkanen L, Anttila H, Pyhältö K. How Does Supervision Influence a Doctoral Supervisor's Occupational Wellbeing? European Journal of Higher Education. 2025;15(2):263-81. Available from: https://doi.org/10.1080/21568235.2024.2314470.
6. Mavrogalou-Foti AP, Kambouri MA, Çili S. The supervisory relationship as a predictor of mental health outcomes in doctoral students in the United Kingdom. Frontiers in Psychology. 2024;Volume 15 - 2024.
7. Polkinghorne M, Taylor J, Knight F, Stewart N. Doctoral Supervision: A Best Practice Review. Encyclopedia. 2023;3(1):46-59. Available from: https://www.mdpi.com/2673-8392/3/1/4.
8. Feizi S, Elgar F. Satisfaction, research productivity, and socialization in doctoral students: Do teaching assistantship, research assistantship and the advisory relationship play a role? Heliyon. 2023;9(9):e19332.
9. Wu S, Oubibi M, Bao K. How supervisors affect students' academic gains and research ability: An investigation through a qualitative study. Heliyon. 2024;10(10):e31079.

10. Tahir I, Fatima N. The impact of student engagement, quality of student faculty relationship and student loyalty on quality of higher education: A systematic literature review. World Journal of Advanced Research and Reviews. 2023. Available from: https://api.semanticscholar.org/CorpusID:265324023.
11. Lai S, Liu S, Dai Y, Lim CP, Liu A. The Impacts and Tensions of Generative AI on Doctoral Students' Supervisory and Peer Dynamics: An Activity Theory Analysis. Australasian Journal of Educational Technology. 2025 Sep. Available from: https://ajet.org.au/index.php/AJET/article/view/9916.
12. Liang W, Liu S, Zhao C. Impact of Student-Supervisor Relationship on Postgraduate Students' Subjective Well-Being: A Study Based on Longitudinal Data in China. Higher Education. 2021;82(2):273-305. Available from: https://doi.org/10.1007/s10734-020-00644-w.
13. Zhou Y, Zhang R, Liu X. The Difference in Co-existence: A Study on the Relationship between Supervisors and Postgraduates from the Perspectives of Teachers and Students Based on a Survey in Gansu Province. Journal of Higher Education Management. 2023;(6). (In Chinese).
14. Dahlem Research School. Power Dynamics in Academia: Supervisory Relationships; 2025. Available from: https://blogs.fu-berlin.de/drs_podcast/2025/11/10/power-dynamics-in-academia-supervisory-relationships/.
15. Mainhard T, van der Rijst R, van Tartwijk J, Wubbels T. A Model for the Supervisor–Doctoral Student Relationship. Higher Education. 2009 Sep;58(3):359-73.
16. Vähämäki M, Saru E, Palmunen LM. Doctoral supervision as an academic practice and leader–member relationship: A critical approach to relationship dynamics. The International Journal of Management Education. 2021;19(3):100510.
17. Han J, Jin L. Reconceptualizing supervisory relationships in graduate education: The role of interpersonal emotion regulation in supervisor-student interactions. International Journal of Educational Research. 2025;132:102650. Available from: https://www.sciencedirect.com/science/article/pii/S0883035525001247.
18. Borgatti SP, Mehra A, Brass DJ, Labianca G. Network analysis in the social sciences. Science. 2009 feb;323(5916):892-5.
19. Yang S, Keller FB, Zheng L. Social Network Analysis: Methods and Examples. 1st ed. Thousand Oaks, CA: SAGE Publications, Inc; 2017. Online publication date: December 20, 2019.
20. Bolger N, Davis A, Rafaeli E. Diary Methods: Capturing Life as it is Lived [Journal Article]. Annual Review of Psychology. 2003;54(Volume 54, 2003):579-616. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev.psych.54.101601.145030.
21. Smyth JM, Stone AA. Ecological Momentary Assessment Research in Behavioral Medicine. Journal of Happiness Studies. 2003 mar;4(1):35-52. Available from: https://doi.org/10.1023/A:1023657221954.
22. Kossinets G, Watts DJ. Empirical Analysis of an Evolving Social Network. Science. 2006 jan;311(5757):88-90.
23. Leonardi PM, Treem JW. Behavioral Visibility: A New Paradigm for Organization Studies in the Age of Digitization, Digitalization, and Datafication. Organization Studies. 2020 dec;41(12):1601-25. Original work published online November 24, 2020. Available from: https://doi.org/10.1177/0170840620970728.
24. Lavidas K, Papadakis S, Manesis D, Grigoriadou AS, Gialamas V. The Effects of Social Desirability on Students' Self-Reports in Two Social Contexts: Lectures vs. Lectures and Lab Classes. Information. 2022;13(10). Available from: https://www.mdpi.com/2078-2489/13/10/491.
25. Tourangeau R, Yan T. Sensitive questions in surveys. Psychological Bulletin. 2007;133(5):859-83.
26. Burke-Smalley L, Neely AR, Bryant E. Building professor-student rapport: A model, survey findings, and implications for practicing professors. Business Horizons. 2024;67(2):137-45.
27. Shi Y, Ke G, Chen Z, Zheng S, Liu TY. Quantized Training of Gradient Boosting Decision Trees. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems. vol. 35. Curran Associates, Inc.; 2022. p. 18822-33.
28. Rizkallah LW. Enhancing the performance of gradient boosting trees on regression problems. Journal of Big Data. 2025 Feb;12(1):35. Available from: https://doi.org/10.1186/s40537-025-01071-3.
29. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149–3157.

30. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768–4777.
31. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 1135–1144.
32. Li Y, Xu W, Chen J. PhD student-supervisor relationship and its impacts: a perspective of the interpersonal relationship model. Frontiers in Education. 2025;Volume 10 - 2025.
33. Zackariasson M, Magnusson J. In: The Supervisor-Student Relationship. Cham: Springer Nature Switzerland; 2024. p. 69-106. Available from: https://doi.org/10.1007/978-3-031-66371-0_4.
34. Hasib KM, Rahman F, Hasnat R, Alam MGR. A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC); 2022. p. 0399-405.
35. Hasib KM, Towhid NA, Faruk KO, Al Mahmud J, Mridha MF. Strategies for enhancing the performance of news article classification in Bangla: Handling imbalance and interpretation. Engineering Applications of Artificial Intelligence. 2023;125:106688.
36. Guryanov A. Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees. In: van der Aalst WMP, Batagelj V, Ignatov DI, Khachay M, Kuskova V, Kutuzov A, et al., editors. Analysis of Images, Social Networks and Texts. Cham: Springer International Publishing; 2019. p. 39-50.