Breaking urban–rural educational inequality: application of an AI relative fairness model in teacher performance evaluation and bonus allocation

Xia Yang[1,2]

[1]*Cangxian Middle School, Cangzhou, Hebei 061725, China*
[2]*School of Public Administration, Hebei University, Baoding, Hebei 071002, China*

Abstract. Teacher performance evaluation and bonus allocation in Chinese schools still rely heavily on subjective judgment, and the resulting distortions disproportionately affect rural teachers, exacerbating compensation disparities and attrition that undermine the rural revitalisation agenda and SDG 4. This study asks how artificial intelligence (AI) could make these processes fairer. Following the PRISMA 2020 guidelines, a systematic search of five databases for 2010–2025 returned 4,713 records, of which 30 studies were included and synthesised using structured thematic coding and keyword co-occurrence mapping. Three overlapping themes emerged: AI applications relevant to evaluation (coded in 60% of the included studies), fairness pathways such as explainable AI and participatory decision-making (80%), and urban–rural inequality (20%). The reviewed literature indicates that AI-assisted evaluation can improve the consistency and evidential basis of teacher assessment and relieve administrative burden, but that these gains are conditional on confronting algorithmic bias, model opacity, privacy risk, and the rural digital divide; evaluation models designed specifically for teachers, let alone adapted to rural Chinese contexts, remain scarce. On this basis, the paper proposes a four-dimensional relative fairness model (4DG) combining data-integrity safeguards, transparency through explainable and generative AI, teacher participation, and ethical governance with auditable records. A deliberately simple, assumption-driven simulation – reported transparently as such – illustrates that correcting a substantial share of evaluation bias could narrow the urban–rural bonus gap by roughly 15–25%. Because the review's quantitative synthesis could not be verified against reproducible study-level records, no pooled effect estimate is presented as an empirical finding; the paper's contribution is the governance framework and the agenda it sets for empirical piloting in China's urban–rural context.

Keywords: AI empowerment, teacher performance evaluation, bonus allocation, educational equity, urban–rural inequality, algorithmic bias, relative fairness model

1. Introduction

1.1. Research background

Urban–rural inequality is among the most persistent structural features of Chinese society, and education is one of the domains where it is most consequential. The inequality has deep institutional roots: the household registration (*hukou*) system long segmented the population into urban and rural groups with different access to employment, housing, schooling, and social protection [17, 25]. Its economic signature remains visible in official statistics. In 2019, urban per capita disposable income was 42,359 yuan, compared with 16,021 yuan for rural residents, an urban-to-rural ratio of about 2.64. By 2023, the gap had narrowed but remained wide: 51,821 versus 21,691 yuan, a ratio of 2.39 [19]. Overall income inequality has stayed high, with a Gini coefficient of roughly 0.47 [17]. These disparities translate directly into educational disadvantage: rural schools operate with thinner resources, weaker infrastructure, and fewer professional development opportunities than their urban

0009-0004-1395-1638 (X. Yang)
854764491@qq.com (X. Yang)

Received 2025-10-08 | Accepted 2026-04-06 | Published 2026-06-20 | Version of record 2026-06-20

counterparts, a pattern consistent with the international evidence on resource-linked achievement gaps [21, 34].

Within this landscape, teacher performance evaluation and the bonus allocation tied to it deserve more attention than they usually receive. Performance-related pay is a significant component of Chinese teachers' compensation, and the way it is assessed shapes both their incomes and careers. Traditional evaluation practice relies heavily on supervisors' judgment and on metrics normed to urban conditions – standardised student attainment, facility-dependent teaching innovation, publication and competition records – while much of the work that defines rural teaching, such as multi-grade instruction, boarding-student care, and community engagement, remains invisible to the instruments [23, 34]. Teachers are acutely sensitive to such misalignment: empirical work on fairness conceptions shows that they judge assessment regimes by the transparency of criteria and the recognition of context, not only by outcomes [23]. When rural teachers are evaluated against urban-normed criteria, they receive systematically lower ratings and bonuses, reinforcing the attrition and recruitment difficulties that rural education policy has struggled with for two decades [34]. The consequences compound over time: persistent staffing disadvantages in rural schools feed into the well-documented underrepresentation of rural students in elite higher education [32].

Policy has not been idle. The rural revitalisation strategy adopted in 2017 made rural teacher support an explicit priority, and China has invested heavily in technology-mediated equalisation – national smart-education platforms, connectivity subsidies, and the delivery of urban teaching resources to rural classrooms [33, 34]. The COVID-19 pandemic simultaneously demonstrated the potential and the limits of this approach, as remote instruction amplified pre-existing digital divides [24, 33]. Artificial intelligence is the newest layer of this technology-driven agenda, and sits within China's broader national strategy for AI development [16]. In principle, AI-assisted evaluation could broaden the evidence base of teacher assessment, standardise the application of criteria, surface contextual factors that human raters overlook, and reduce the administrative load of documentation. However, AI also imports new risks: models trained on urban-weighted data can encode and automate precisely the biases they were meant to remove [18], and opaque scoring can further erode the trust of the teachers being evaluated [6, 7]. Whether AI narrows or widens the urban–rural gap is therefore not a property of the technology but of its governance.

Figure 1 summarises the policy trajectory within which this question arises, from the infrastructure-led investments of the 2000s through the standardisation reforms of the 2010s to the current digital and AI era. Figure 2 documents the economic backdrop using official statistics: urban and rural incomes have both risen substantially, and the ratio between them has declined steadily, yet the absolute gap continues to grow – a reminder that relative convergence and lived inequality can move in different directions.

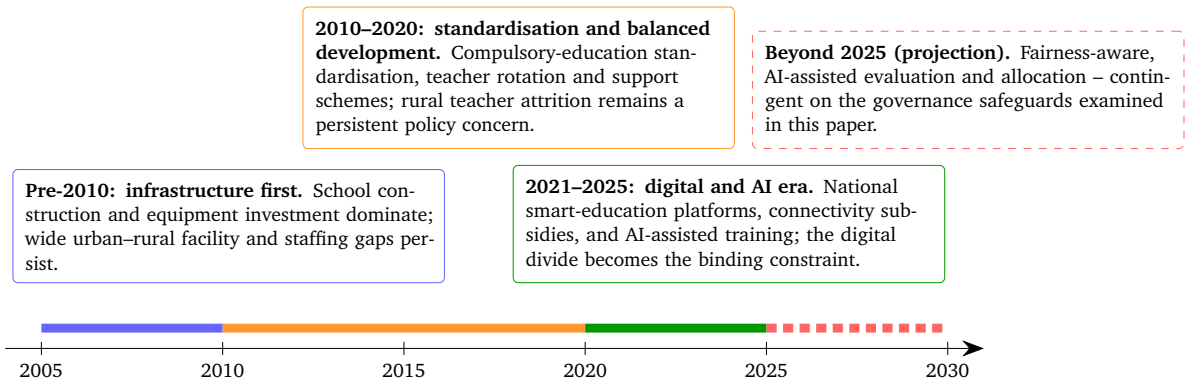

Figure 1: Urban–rural educational equity policy trajectory in China, 2005–2030 (schematic). Phases are indicative rather than sharply delimited; synthesized from Wu [33], Xue, Li and Li [34] and Gustafsson, Cai and Sicular [17].

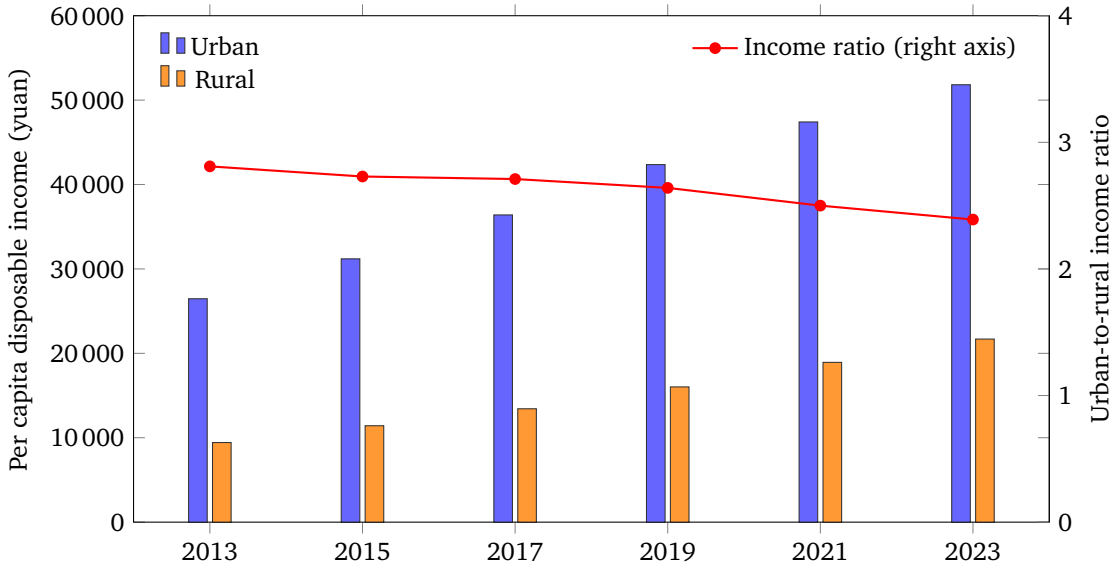

Figure 2: Urban and rural per capita disposable income in China and their ratio, 2013–2023. Data: National Bureau of Statistics of China integrated household survey series [19]; see also Gustafsson, Cai and Sicular [17] on measurement of the gap.

1.2. Research questions

This study addresses three questions. First, how can AI be applied to teacher performance evaluation and bonus allocation in ways that promote urban–rural equity – for instance, through explainable models that make scoring criteria contestable? Second, what implementation challenges and ethical risks does such an application face, including data privacy and algorithmic discrimination? Third, how can policy and practice govern AI's role so that it narrows rather than reproduces urban–rural gaps, including through participatory mechanisms? The three questions deliberately span the technical, ethical, and institutional dimensions of the problem: the first concerns capability, the second risk, and the third governance. Together they align the inquiry with SDG 4's commitment to inclusive and equitable quality education [29], while remaining anchored in the specifics of the Chinese policy context – where, as Wu [33] argues, collectivist policy traditions shape how participatory governance can realistically be organised.

To answer them, the paper proceeds as follows. Section 2 reviews the literature on AI in education, fairness pathways, and Chinese urban–rural inequality. Section 3 describes the systematic review methodology, including an explicit statement of what could and could not be verified in its quantitative apparatus. Section 4 reports the thematic findings and develops the four-dimensional relative fairness model (4DG). Section 5 discusses implications, limitations, and an agenda for empirical validation.

2. Literature review

2.1. AI applications in education and teacher evaluation

Research on artificial intelligence in education has grown rapidly since 2010 and accelerated sharply after the release of large language models. In their review of the higher education field, Crompton and Burke [11] found that the vast majority of studies concern students – 72% of AIEd studies target undergraduate learners, and students are the primary beneficiaries in 72% of applications, compared with 17% for instructors and 11% for managers. AI systems that assess, predict, and tutor students are comparatively mature; systems that evaluate *teachers* are a small minority of the field. The same asymmetry appears in reviews of AI for teachers: Celik et al. [6] document promising applications in planning, implementation, and assessment support, but note that most tools position the teacher as user rather than as subject of evaluation, and Tan, Cheng and Ling [26] reach a similar conclusion

for AI in teacher professional development, where evaluation-oriented applications remain rare and are seldom studied longitudinally.

The studies that do address teacher evaluation directly illustrate both the promise and the design space. Almubarak et al. [1] developed an analytical, AI-based teacher performance evaluation system for Saudi schools, combining multiple data sources into dashboard-style indicators – an architecture similar to the key performance indicator (T-KPI) designs discussed in the Chinese literature. Terrazas-Arellanes et al. [27] show that machine-learning scoring of performance-based assessments can be deployed in a way that increases, rather than diminishes, teachers' sense of agency – a finding directly relevant to fears that automated evaluation disempowers those it measures. On the readiness side, Chiu, Ahmad and Çoban [8] developed and validated a scale for teachers' AI competence self-efficacy, underscoring that evaluation systems presuppose a workforce confident enough with AI to engage with them; Bergdahl and Sjöberg [4] likewise find that teachers' attitudes and self-efficacy, more than access alone, predict meaningful AI use in K-12 settings.

Generative AI has widened the design space again. Reviews of ChatGPT-era research document affordances for automating routine documentation and feedback – tasks that dominate teachers' administrative load – alongside new risks of inaccuracy and misuse [3, 12]. Domain-specific reviews report the same duality in language teaching [14], and studies of teacher perceptions in K-12 systems outside the West, such as Tongchai and Malakul [28] in Thailand, find teachers broadly receptive yet concerned about training and equity of access. Evidence from Chinese schools is especially pertinent here: Niu et al. [20] studied an AI-aided educational platform used in Chinese schools, including a rural school, and found that teachers and students valued its diagnostic feedback. At the same time, implementation depended heavily on local support – a preview, at the classroom scale, of the infrastructure dependence that any AI-assisted evaluation of rural teachers would face. Inclusive education reviews note that AI can be deliberately designed for accessibility rather than merely retrofitted [31].

Two gaps stand out for present purposes. First, models designed for teacher performance evaluation, as opposed to student assessment, are scarce, and models adapted to rural or resource-constrained contexts are scarcer still; in the corpus reviewed here, only a minority of studies engage in teacher-specific evaluation at all (section 4). Second, the general-purpose systems that dominate practice, including large language models, are trained on data with little representation of rural Chinese schooling, so their apparent neutrality cannot be assumed [5, 18].

2.2. Fairness pathways and challenges

The algorithmic fairness literature provides the conceptual vocabulary that the field of education increasingly borrows. Mehrabi et al. [18] distinguish, among others, *group* (demographic) fairness criteria, which require parity of outcomes or error rates across groups such as urban and rural teachers, from *individual* fairness criteria, which require that similar individuals be treated similarly. The distinction matters in practice because the two families of criteria can conflict: enforcing strict individual consistency with urban-normed metrics can lock in group-level disadvantage, while group-level corrections can be experienced as unfair by individuals near the adjustment boundary. Any AI-assisted evaluation scheme, therefore, has to make its fairness criterion explicit and defensible rather than treating "fairness" as a single property a system either has or lacks.

Within education, two pathways toward fairer AI-assisted assessment recur across the reviewed studies. The first is transparency. Explainable AI (XAI) techniques – feature-attribution methods, interpretable surrogate models, natural-language rationales – allow an evaluated teacher to see which evidence drove a rating and to contest it. Perceived fairness is not an automatic consequence of automation: Chai et al. [7] found experimentally that students' fairness perceptions depend on who (or what) evaluates and how the decision is communicated, with AI evaluators sometimes perceived as fairer than human ones precisely because they are seen as less partial, but only when the process is intelligible. The second pathway is participation. Studies of teachers' own fairness conceptions show that procedural voice – being consulted on criteria, having appeal channels – is central to whether

an assessment regime is accepted [23]; parallel findings in online-assessment contexts underline that fairness perceptions are shaped by context and communication as much as by formal properties [2]. Broader equity scholarship from the pandemic period frames both pathways within a social-justice perspective: assessment arrangements that ignore participants' material circumstances reproduce inequity regardless of their formal neutrality [24].

Against these pathways stand well-documented challenges: bias inherited from unrepresentative training data [18], the opacity of complex models [6], privacy and data-governance risks in systems that aggregate fine-grained records of teachers' work, and the ethical thinness of much AIEd research – Bond et al. [5], in a meta-systematic review, explicitly call for increased ethics, collaboration, and rigour in the field. Teachers' own AI literacy is a further constraint, particularly where professional development is weakest [4, 30]. Section 4 organises these issues and their candidate countermeasures as they emerged in the included studies (see table 2).

2.3. Urban–rural educational inequality in China

The Chinese-context studies in the corpus connect these general debates to the specific institution this paper targets. The economics literature establishes the scale and persistence of the urban–rural divide: the income gap, though narrowing in relative terms, remains among the largest components of Chinese inequality, and measurement choices (migrant reclassification, cost-of-living adjustment) materially affect its estimation [17, 25]. Education policy research documents how this divide manifests in the teaching workforce. Xue, Li and Li [34], analysing the full policy circle of rural education under the rural revitalisation strategy, identifies teacher recruitment, retention, and development as the critical bottleneck: rural schools struggle to attract and keep qualified teachers despite successive incentive programs, and evaluation and compensation arrangements that disadvantage rural service are part of the problem. Wen et al. [32] trace the downstream consequences in access to elite higher education, showing persistent urban–rural disparities in admission to Tsinghua University even after preferential policies. On the technology side, Wu [33] reviews China's efforts to promote educational equity through technology – resource delivery, remote instruction, teacher training – and notes a strategic shift from hardware provision toward software and teacher quality, while cautioning that digital capability deficits and ethical governance lag behind the infrastructure. Classroom-level evidence such as Niu et al. [20] confirms both the appetite for and the fragility of AI-supported practice in rural schools.

What is missing from this literature is precisely the intersection the present review addresses: none of the identified Chinese-context studies examines AI-assisted *teacher evaluation and bonus allocation* as an equity instrument. The fairness pathways of section 2.2 have not yet been brought to bear on the institution through which rural teachers' pay and recognition are actually determined.

3. Methods

3.1. Design and search strategy

The study was designed as a systematic review following the PRISMA 2020 guidelines [22]. Five databases were searched: PubMed, Scopus, Web of Science, CNKI (China National Knowledge Infrastructure), and Google Scholar, covering publications from January 2010 to the search date of 15 October 2025. Search strings combined three concept blocks – artificial intelligence (including generative AI and explainable AI), teacher performance evaluation and bonus allocation, and urban–rural inequality and educational equity, using Boolean operators and wildcards; for CNKI, Chinese-language equivalents of all terms were used. The complete strings, per-database dates, and result counts are reported in appendix A. The searches returned 4,713 records in total. After deduplication (supported by a Python script using title-similarity matching, with manual verification of borderline cases) and title/abstract screening (supported by a BERT-based semantic similarity screen at a 0.75 threshold, again with manual verification), full-text assessment yielded 30 included studies,

an inclusion rate of approximately 0.6%. About 70% of the included studies were published in 2023–2025, reflecting the recency of the AI-and-equity intersection.

Inclusion criteria were: peer-reviewed studies or substantive reports addressing (a) AI applications relevant to teacher evaluation, assessment, or professional development; (b) fairness, equity, or ethics of AI in educational assessment; or (c) urban–rural educational inequality in China with implications for the teaching workforce. Exclusion criteria were non-educational applications, purely technical AI papers without educational context, and opinion pieces without analysable content.

3.2. Coding and quality appraisal

Data extraction followed a structured codebook (appendix B) covering bibliographic information, study type, sample, geographic focus, AI type, reported benefits and challenges, and urban–rural relevance. Coding was performed by two coders working independently, with disagreements resolved through discussion; thematic assignments were non-exclusive, so a study could contribute to more than one theme, and, on average, studies were assigned just under two themes. Study quality was appraised with reference to the AMSTAR-2 instrument's sixteen domains for the included reviews, and with analogous rigour criteria (design transparency, bias reporting, protocol registration) for primary studies; in aggregate, most included studies were appraised at moderate confidence, with a small number of high-confidence systematic reviews [cf. 5].

The synthesis is primarily narrative and thematic. A minority of included studies – roughly one in five – report quantitative effect estimates, and the review's design envisaged pooling them with standard random-effects meta-analytic machinery (inverse-variance weighting, heterogeneity assessment via Q and I^2, sensitivity analysis, and small-study/publication-bias checks). Appendix C documents this analytic approach with worked examples whose numbers are explicitly illustrative. For the reasons given in section 3.4, no pooled estimate from this apparatus is reported as a finding of the review.

3.3. Analysis tools

Keyword co-occurrence across the included corpus was mapped with VOSviewer using a minimum co-occurrence threshold of five; the resulting network (figure 3) was used descriptively, to visualise the thematic structure of the corpus rather than to test hypotheses. Quantitative computations were performed in R (the `metafor` package) for the meta-analytic formulas in appendix C and in Python for deduplication and the simulation in appendix D.

3.4. Verifiability of the quantitative synthesis

The quantitative apparatus, retained in appendices C–E, documents the analytic approach that the 4DG framework proposes to formalise and should be read as a methodological illustration pending a fully reproducible empirical synthesis, which section 5.3 outlines.

4. Findings

4.1. Thematic structure of the literature

Thematic coding of the 30 included studies identified three overlapping themes. *AI applications* relevant to evaluation – automated and machine-learning-based scoring, structured learning-path (SLP) systems, teacher key-performance-indicator (T-KPI) dashboards, and generative-AI support for feedback and documentation – were coded in 18 of 30 studies (60%). *Fairness pathways* – explainability, perceived fairness, participatory assessment design, and equity frameworks – appeared in 24 of 30 studies (80%). *Urban–rural inequality*, whether as an empirical setting or a policy focus, appeared in 6 of 30 studies (20%). Because coding was non-exclusive, the themes overlap substantially: most application-focused studies also raise fairness questions, which is itself a finding –

the field increasingly treats fairness as intrinsic to AI-assisted assessment rather than an afterthought. Two gaps recur across the corpus: models built to evaluate teachers (rather than students) are a small minority, and explicit ethical analysis – data governance, accountability for automated decisions – is present in only a fraction of application studies, echoing the field-level critique in Bond et al. [5]. Table 1 summarises the distribution.

Table 1
Literature theme distribution across the 30 included studies (themes non-exclusive; coded by this review).

Theme	Studies (of 30)	Focus of the coded studies	Gaps observed
AI applications	18 (60%)	Automated and ML-based scoring, structured learning paths, T-KPI dashboards, GenAI support for feedback and documentation [1, 12, 26, 27]	Few systems evaluate teachers rather than students; almost none are adapted to rural or resource-constrained Chinese contexts
Fairness pathways	24 (80%)	Explainability and transparency, perceived fairness of AI evaluators, participatory assessment design, equity and social-justice framings [2, 7, 18, 23, 24]	Fairness is mostly analysed for student-facing assessment; explicit ethical and accountability analysis remains sparse [5]
Urban–rural inequality	6 (20%)	Rural teacher retention and policy, access disparities, technology-supported equity initiatives in China [20, 32–34]	Little empirical work links evaluation or compensation design to rural attrition; no study examines AI-assisted bonus allocation

The keyword co-occurrence network of the included corpus (figure 3) visualises this structure. The map separates into two interconnected clusters: one (red) organised around application-and-development vocabulary – *application*, *challenge*, *approach*, *integration*, *development*, *large language model* – and one (green) around context-and-participant vocabulary – *context*, *participant*, *student*, *interview*, *survey*, *china*. The density of cross-cluster links reflects the same overlap seen in the coding: technical application studies and context-sensitive participant studies increasingly cite into each other's territory, but the network also shows that *china*-related terms sit at the periphery of the application cluster – a visual restatement of the localisation gap.

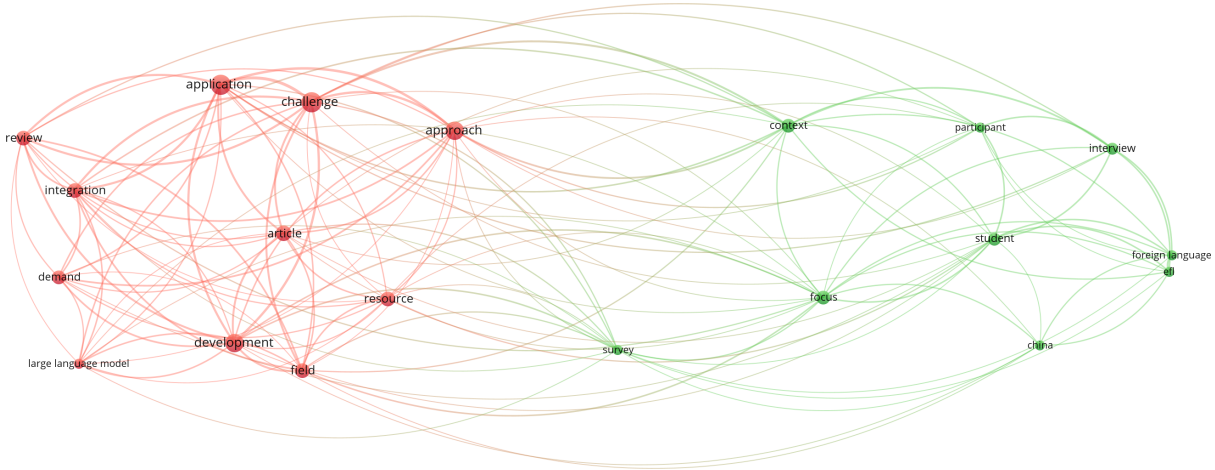

Figure 3: Keyword co-occurrence network of the 30 included studies (VOSviewer; minimum co-occurrence 5). Node size reflects occurrence frequency; colours indicate the two main clusters (red: applications and development; green: context and participants). Node positions reflect VOSviewer's default layout algorithm and were not manually adjusted for legibility.

4.2. Reported benefits, challenges, and countermeasures

Across the application studies, the reported benefits of AI-assisted evaluation are consistently in the same direction. Systems that aggregate multiple evidence sources produce more consistent scoring than single-supervisor judgment and make the basis of ratings inspectable [1, 27]; generative tools absorb a substantial share of the documentation and feedback-drafting burden that teachers otherwise carry [12, 14]; and diagnostic platforms give teachers formative signals they would not otherwise receive, including in rural schools [20]. Several studies report sizeable time savings in administrative tasks, though the estimates are too heterogeneous in methods and settings to support a single pooled figure. For retention – the outcome that matters most for the urban–rural gap – the evidence is indirect: studies link better feedback, recognition, and workload relief to intentions to stay, but no included study measures the effect of AI-assisted evaluation on actual rural teacher retention. That absence defines the empirical agenda of section 5.3.

The challenges are equally consistent, and table 2 organises the six issue areas that recur throughout the corpus, along with the countermeasures proposed in the literature. Three deserve emphasis in the urban–rural context. First, *algorithmic bias*: evaluation models trained predominantly on urban data risk scoring rural teaching against inapplicable norms – the mechanism by which AI could automate, rather than remove, existing unfairness [18]. Second, the *digital divide*: rural schools' weaker connectivity and lower AI literacy mean that data-hungry evaluation systems may simply see rural teachers less well, quite apart from any modelling bias [30, 33]. Third, *opacity*: teachers' acceptance of algorithmic evaluation depends on being able to understand and contest it [7, 23]; a black-box bonus formula would fail the procedural-fairness test even if its outputs were statistically unbiased. Table 3 summarises benefits, challenges, and design implications at three levels.

Table 2
Key issues and countermeasures for AI-assisted teacher evaluation in urban–rural contexts, as synthesised from the included studies.

Issue area	Challenge in urban–rural evaluation	Countermeasures in the reviewed literature
Algorithmic bias	Urban-weighted training data can systematically underestimate rural teachers' performance	Context-balanced training data; routine bias audits against explicit group-fairness criteria [5, 18]
Opacity ("black box")	Teachers cannot contest scores they cannot understand; trust erodes	Explainable-AI interfaces disclosing the evidence behind each rating [6, 7]
Digital divide	Weaker connectivity and AI literacy make rural work less visible to data-driven systems	Blended training, low-bandwidth tooling, phased rollout [20, 30, 33]
Competing fairness standards	Group parity and individual consistency can conflict in uneven resource settings	Make the operative fairness criterion explicit; report group and individual metrics side by side [18, 24]
Weak teacher participation	Evaluation imposed without voice is perceived as unfair regardless of accuracy	Participatory metric design; teacher representation and appeal channels [2, 23]
Ethics and accountability	Privacy exposure; unclear responsibility for automated decisions	Data-protection rules, human review of consequential decisions, auditable records [5, 29]

4.3. A four-dimensional relative fairness model (4DG)

The review's constructive contribution is to organise these findings into a governance model for AI-assisted teacher evaluation and bonus allocation, called here the four-dimensional relative fairness model (4DG). The model is "relative" in a deliberate sense: rather than scoring all teachers against a single absolute standard, it evaluates performance relative to context – what a teacher achieves with

Table 3
Reported benefits and challenges of AI-assisted teacher evaluation, by level.

Level	Reported benefits	Reported challenges	Design implications
Evaluation process	More consistent, multi-source, evidence-based scoring; reduced dependence on single-supervisor judgment; administrative time savings [1, 12]	Bias inherited from urban-weighted data; opacity of model decisions [18]	Context-balanced data and bias audits; explainable scoring; human review of consequential outcomes
Individual teacher	Richer formative feedback; recognition of otherwise invisible work; support for professional growth and, indirectly, retention [20, 27]	Privacy exposure and surveillance concerns; new skill demands [5, 8]	Data minimization and consent; AI-literacy development alongside deployment [4]
School system	Comparable evidence across schools; a basis for fairer cross-regional bonus allocation [33]	Digital divide skews whose work is measured; deployment and maintenance costs [30]	Connectivity investment first; low-bandwidth fallbacks; phased, monitored rollout

the resources, students, and constraints actually present – so that fairness is defined across unequal settings rather than assumed away. Its four dimensions correspond point-for-point to the challenge areas of table 2:

D1. **Data integrity.** Evaluation inputs are cleaned, provenance-checked, and protected against manipulation, with tamper-evident (e.g., blockchain-anchored) records ensuring that rural teachers' evidence is neither missing nor forgeable. This addresses the divide-driven visibility problem before any model is applied.

D2. **Transparency (XAI + GenAI).** Every rating is accompanied by an explanation of the evidence and criteria behind it, generated through explainable-AI techniques and communicated in accessible natural language by generative models, so that scores are contestable by the teachers they concern [6, 7].

D3. **Teacher participation.** Teachers – including rural teachers specifically – take part in defining metrics, weighting context factors, and staffing appeal processes, satisfying the procedural-fairness conditions the empirical literature identifies as decisive for acceptance [2, 23].

D4. **Ethical governance.** An institutional layer sets fairness criteria explicitly (group and individual, per [18]), mandates periodic bias audits, keeps auditable decision records, and requires human review of consequential outcomes, in line with field-level calls for ethics and rigour [5, 29].

Figure 4 shows how the dimensions feed a fairness-optimisation loop running from input validation through bias detection and context-adjusted weighting to allocation, with teacher feedback and audit findings cycling back into metric design. The model's dynamic element – iterative reweighting in response to detected bias and participatory feedback – distinguishes it from static fairness checklists: it treats fairness as a property to be maintained, not certified once. In the Chinese setting, the model is designed to slot into existing bureaucratic evaluation structures (its governance dimension maps onto familiar administrative oversight), while its participation dimension pushes those structures toward the procedural voice that the fairness literature shows they lack. An assumption-driven simulation of the model's intended mechanism – not an empirical estimate – suggests that correcting half to two-thirds of an evaluation-bias component assumed to account for 30–40% of the urban–rural

bonus gap would narrow that gap by roughly 15–25%; the assumptions and arithmetic are laid out in full in appendix D.

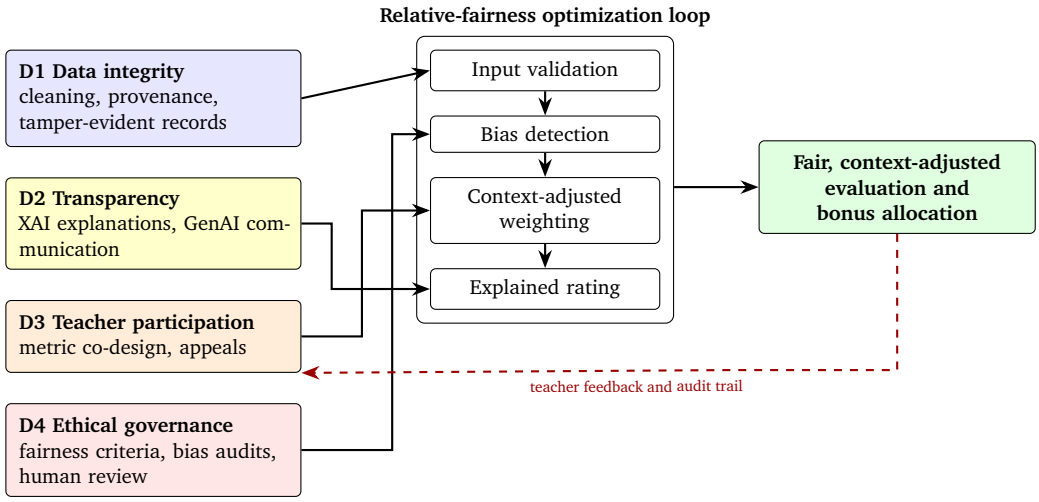

Figure 4: The four-dimensional relative fairness model (4DG). The four governance dimensions feed a fairness-optimisation loop; feedback and audit findings cycle back into participatory metric design.

5. Discussion

5.1. Implications for the research questions

The theoretical contribution of the review is to connect three literatures – AI in education, algorithmic fairness, and Chinese urban–rural education policy – at the institution where they intersect but have not previously met: teacher performance evaluation and bonus allocation. The 4DG model extends foundational fairness distinctions [18] into a governance design for that institution, and grounds them in the procedural-fairness evidence from teachers themselves [23]. Table 4 summarises what the reviewed literature indicates for each research question; the paragraphs below elaborate.

For the first question – how AI can be applied to promote urban–rural equity – the answer emerging from the corpus is conditional. AI-assisted evaluation can improve consistency, broaden the evidence base, explain its own ratings, and relieve administrative burden [1, 12, 27]; each of these directly addresses mechanisms by which current subjective evaluation disadvantages rural teachers. But every one of these capabilities depends on the system accurately seeing rural work, which is exactly what urban-weighted data and divide-constrained infrastructure prevent. The practical implication for education departments is that data representativeness and connectivity are not implementation details but preconditions: an AI evaluation system procured without a rural data strategy should be expected to widen the gap it was bought to close [18, 33].

For the second question – challenges and ethical risks – the review's answer is the six-issue structure of table 2: algorithmic bias, opacity, the digital divide, competing fairness standards, weak participation, and thin accountability. Two points deserve emphasis. The tension between group and individual fairness is not a technicality: bonus allocation is a zero-sum-feeling exercise, and corrections that improve group parity will be experienced as unfair by some individuals unless the criteria are explicit and the process participatory – which is why 4DG binds the statistical and procedural dimensions together rather than treating them as separable. And privacy deserves particular weight in the teacher-evaluation context, because the data that make evaluation rich (classroom recordings, platform logs, student feedback) are the same data that make surveillance possible; the governance dimension's data-minimisation and human-review requirements are the counterweight [5].

For the third question – how policy can govern AI to narrow the gap – the review supports a staged path rather than a leap. The 4DG model is designed for incremental adoption: tamper-evident data practices and explained ratings can be introduced within existing evaluation cycles; participatory metric review and formal bias audits can follow; the full optimisation loop with context-adjusted weighting is the end state, to be validated in pilots before any consequential (bonus-affecting) use. The simulation of appendix D illustrates the intended mechanism – under its stated assumptions, correcting a substantial share of evaluation bias narrows the overall bonus gap by roughly 15–25% – but it is an illustration of logic, not a forecast, and section 5.3 sets out the empirical program that would replace assumption with measurement. Internationally, the model's transparency and participation dimensions align with the direction of teacher-perception evidence from other school systems [28, 30], and its governance dimension with global calls for ethical AI in education [5, 29]; the framework should therefore travel, though its bureaucratic integration is designed for, and would need re-mapping outside of, the Chinese administrative context [33].

Table 4
Research questions, what the reviewed literature indicates, and implications.

Research question	What the reviewed literature indicates	Implications
RQ1: How can AI promote urban–rural equity in evaluation and allocation?	AI-assisted evaluation can improve consistency, evidence breadth, and explainability, and reduce administrative burden – but only where rural work is accurately represented in data and infrastructure [1, 12, 18]	Treat rural data representativeness and connectivity as preconditions of procurement; make explainable scoring the default; integrate context variables into metrics
RQ2: What challenges and ethical risks arise?	Six recurring issue areas: algorithmic bias, opacity, digital divide, competing fairness standards, weak participation, thin accountability (table 2)	Explicit fairness criteria with routine bias audits; data minimisation and privacy protection; human review of consequential decisions [5]
RQ3: How should policy govern AI to narrow the gap?	Governance, not technology, determines direction of effect; procedural voice is decisive for acceptance [7, 23]	Staged adoption of 4DG; participatory metric design; piloting before bonus-affecting use; alignment with emerging GenAI regulation and SDG 4 monitoring [29]

5.2. Original value and limitations

The paper's original value lies in three moves: framing teacher evaluation and bonus allocation – rather than student learning – as the site where AI meets urban–rural inequality; assembling the fairness-pathway evidence relevant to that site; and integrating it into the 4DG governance model with an explicit relative (context-adjusted) conception of fairness.

The limitations are substantial and should be stated plainly. First, as disclosed in section 3.4, the quantitative synthesis envisaged in the review design could not be verified against reproducible study-level records, so the paper reports no pooled effect estimates and its evidence claims are qualitative; the quantitative appendices are methodological illustrations. Second, only a minority of included studies are quantitative, and none measure the causal effect of AI-assisted evaluation on rural teachers' outcomes, so even a fully reproducible synthesis would have rested on indirect evidence. Third, the 4DG model is a design proposal grounded in the literature, not a validated intervention; the attached simulation is assumption-driven. Fourth, the corpus is dominated by studies from 2023–2025, a period of rapid change in generative AI, so both capabilities and risks may shift under the analysis. These limitations mark the boundary between what this review establishes – the problem structure, the design requirements, the governance framework – and what only field research can establish: whether the framework works.

5.3. Directions for empirical validation

Four complementary strategies would put the 4DG model on an empirical footing in China's urban–rural context. First, primary data collection in rural schools – surveys and administrative records for a target sample on the order of 500 teachers – to measure the actual composition of evaluation scores and bonus outcomes and their association with retention, using appropriate regression designs. Second, small-scale pilots of 4DG-instrumented evaluation in a manageable number of schools (on the order of ten), with pre/post comparison of allocation fairness metrics – both group-level parity and individual consistency – before any consequential use. Third, longitudinal tracking of bonus allocation and retention over six to twelve months and beyond, with resource levels controlled, to estimate whether measured gap components in fact narrow. Fourth, semi-structured interviews with teachers (a sample of around 50) on trust, privacy perceptions, and the experienced fairness of explained ratings, since procedural legitimacy is an empirical outcome in its own right [7, 23]. Platform-based studies in Chinese schools show that such fieldwork is feasible when local support is organised [20], and that cost-conscious design matters for transferability to other developing contexts [29].

5.4. Conclusion

This review asked how AI could make teacher performance evaluation and bonus allocation fairer across China's urban–rural divide. The literature supports a conditional answer: AI-assisted evaluation can address the subjectivity and evidential thinness that currently disadvantage rural teachers, but only under governance that secures representative data, explains its ratings, gives teachers a voice, and audits itself – the four dimensions of the 4DG model proposed here. The review found no empirical study of AI-assisted bonus allocation in rural contexts, and its own quantitative synthesis could not be verified to the standard its design intended, so the paper claims a framework and an agenda rather than measured effects. The stakes of that agenda are not small: evaluation and compensation are levers on rural teacher retention, and retention is a lever on the educational half of China's urban–rural divide. Getting the governance of AI right at this specific institutional site is therefore a concrete, testable contribution that both rural revitalisation policy and SDG 4 can use.

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability statement

This study is a systematic review of published literature; no primary data were generated. The complete search strategy, extraction codebook, and simulation specification are provided in appendices A–D.

Ethics approval

Not applicable: this study is a systematic review of published literature and involved no human participants.

References

[1] Almubarak, A., Alhalabi, W., Albidewi, I. and Alharbi, E., 2024. An analytical approach for an AI-based teacher performance evaluation system in Saudi Arabia's schools. *Discover Applied Sciences*, 6(8), p.406. Available from: https://doi.org/10.1007/s42452-024-06117-4.

[2] Azizi, Z., 2022. Fairness in assessment practices in online education: Iranian University English teachers' perceptions. *Language Testing in Asia*, 12(1), p.14. Available from: https://doi.org/10.1186/s40468-022-00164-7.

[3] Bahroun, Z., Anane, C., Ahmed, V. and Zacca, A., 2023. Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings through Bibliometric and Content Analysis. *Sustainability*, 15(17), p.12983. Available from: https://doi.org/10.3390/su151712983.

[4] Bergdahl, N. and Sjöberg, J., 2025. Attitudes, perceptions and AI self-efficacy in K-12 education. *Computers and Education: Artificial Intelligence*, 8, p.100358. Available from: https://doi.org/10.1016/j.caeai.2024.100358.

[5] Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S.W. and Siemens, G., 2024. A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21(1), p.4. Available from: https://doi.org/10.1186/s41239-023-00436-z.

[6] Celik, I., Dindar, M., Muukkonen, H. and Järvelä, S., 2022. The Promises and Challenges of Artificial Intelligence for Teachers: a Systematic Review of Research. *TechTrends*, 66(4), pp.616–630. Available from: https://doi.org/10.1007/s11528-022-00715-y.

[7] Chai, F., Ma, J., Wang, Y., Zhu, J. and Han, T., 2024. Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology*, 15, p.1221177. Available from: https://doi.org/10.3389/fpsyg.2024.1221177.

[8] Chiu, T.K.F., Ahmad, Z. and Çoban, M., 2025. Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies*, 30(5), pp.6667–6685. Available from: https://doi.org/10.1007/s10639-024-13094-z.

[9] Costa-Feito, A., Saavedra, A. and Barta, S., 2025. Enhancing student memorisation through teacher webcam usage and the interplay of social presence. *International Journal of Educational Technology in Higher Education*, 22(1), p.58. Available from: https://doi.org/10.1186/s41239-025-00554-w.

[10] Crompton, H., 2023. Evidence of the ISTE Standards for Educators leading to learning gains. *Journal of Digital Learning in Teacher Education*, 39(4), pp.201–219. Available from: https://doi.org/10.1080/21532974.2023.2244089.

[11] Crompton, H. and Burke, D., 2023. Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), p.22. Available from: https://doi.org/10.1186/s41239-023-00392-8.

[12] Crompton, H. and Burke, D., 2024. The Educational Affordances and Challenges of ChatGPT: State of the Field. *TechTrends*, 68(2), pp.380–392. Available from: https://doi.org/10.1007/s11528-024-00939-0.

[13] Crompton, H. and Burke, D., 2024. The Nexus of ISTE Standards and Academic Progress: A Mapping Analysis of Empirical Studies. *TechTrends*, 68(4), pp.711–722. Available from: https://doi.org/10.1007/s11528-024-00973-y.

[14] Crompton, H., Edmett, A., Ichaporia, N. and Burke, D., 2024. AI and English language teaching: Affordances and challenges. *British Journal of Educational Technology*, 55(6), pp.2503–2529. Available from: https://doi.org/10.1111/bjet.13460.

[15] Crompton, H. and Song, D., 2021. The Potential of Artificial Intelligence in Higher Education. *Revista Virtual Universidad Católica del Norte*, (62), pp.1–4. Available from: https://doi.org/10.35575/rvucn.n62a1.

[16] Ding, J., 2018. *Deciphering China's AI dream: The context, components, capabilities, and consequences of China's strategy to lead the world in AI*. Future of Humanity Institute, University of Oxford. Available from: https://cdn.governance.ai/Deciphering_Chinas_AI-Dream.pdf.

[17] Gustafsson, B.A., Cai, M. and Sicular, T., 2025. China's Urban-Rural Income Gap: A Re-examination. *Paper presented at the International Association for Research in Income and Wealth (IARIW) special conference, Hanoi*. Available from: https://iariw.org/wp-content/uploads/2025/04/Gustafsson_Hanoi.pdf.

[18] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), p.115. Available from: https://doi.org/10.1145/3457607.

[19] National Bureau of Statistics of China, 2024. Households' Income and Consumption Expenditure in 2023. Press release, 18 January 2024. Available from: https://www.stats.gov.cn/english/PressRelease/202402/t20240201_1947120.html.

[20] Niu, S.J., Luo, J., Niemi, H., Li, X. and Lu, Y., 2022. Teachers' and Students' Views of Using an AI-Aided Educational Platform for Supporting Teaching and Learning at Chinese Schools. *Education Sciences*, 12(12), p.858. Available from: https://doi.org/10.3390/educsci12120858.

[21] OECD, 2023. *Education at a Glance 2023: OECD Indicators*. Paris: OECD Publishing. Available from: https://doi.org/10.1787/e13bef63-en.

[22] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P. and Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, p.n71. Available from: https://doi.org/10.1136/bmj.n71.

[23] Rasooli, A., Rasegh, A., Zandi, H. and Firoozi, T., 2023. Teachers' Conceptions of Fairness in Classroom Assessment: An Empirical Study. *Journal of Teacher Education*, 74(3), pp.260–273. Available from: https://doi.org/10.1177/00224871221130742.

[24] Roshid, M.M., Sultana, S., Kabir, M.M.N., Jahan, A., Khan, R. and Haider, M.Z., 2025. Equity, fairness, and social justice in teaching and learning in higher education during the COVID-19 pandemic. *Asia Pacific Journal of Education*, 45(2), pp.397–418. Available from: https://doi.org/10.1080/02188791.2022.2122024.

[25] Sicular, T., Yue, X., Gustafsson, B. and Li, S., 2007. The urban–rural income gap and inequality in China. *Review of Income and Wealth*, 53(1), pp.93–126. Available from: https://doi.org/10.1111/j.1475-4991.2007.00219.x.

[26] Tan, X., Cheng, G. and Ling, M.H., 2025. Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, 8, p.100355. Available from: https://doi.org/10.1016/j.caeai.2024.100355.

[27] Terrazas-Arellanes, F.E., Strycker, L., Alvez, G.G., Miller, B. and Vargas, K., 2025. Promoting Agency Among Upper Elementary School Teachers and Students with an Artificial Intelligence Machine Learning System to Score Performance-Based Science Assessments. *Education Sciences*, 15(1), p.54. Available from: https://doi.org/10.3390/educsci15010054.

[28] Tongchai, A. and Malakul, S., 2026. Thai Teachers'Perceptions of Integrating Generative AI in K-12 Education: Opportunities and Challenges. *TechTrends*, 70(1), pp.93–105. Available from: https://doi.org/10.1007/s11528-025-01128-3.

[29] UNESCO Institute for Statistics and Global Education Monitoring Report Team, 2025. *2025 SDG 4 Scorecard progress report on national benchmarks: Focus on the out-of-school rate*. Montreal: UNESCO Institute for Statistics. Available from: https://doi.org/10.54676/SKLD4888.

[30] Velander, J., Taiye, M.A., Otero, N. and Milrad, M., 2024. Artificial Intelligence in K-12 Education: eliciting and reflecting on Swedish teachers' understanding of AI and its implications for teaching & learning. *Education and Information Technologies*, 29(4), pp.4085–4105. Available from: https://doi.org/10.1007/s10639-023-11990-4.

[31] Villatoro Moral, S. and Moreno-Tallón, F., 2025. Inteligencia Artificial y Educación Inclusiva: soluciones tecnológicas para una enseñanza accesible. Revisión Sistemática [Artificial Intelligence and Inclusive Education: Technological Solutions for Accessible Teaching. A Systematic Review]. *Digital Education Review*, (47), pp.62–77. Available from: https://doi.org/10.1344/der.2025.47.62-77.

[32] Wen, W., Zhou, L., Zhang, M. and Hu, D., 2023. Urban/Rural Disparities in Access to Elite Higher Education: The Case of Tsinghua University. *International Journal of Chinese Education*, 12(2), pp.1–12. Available from: https://doi.org/10.1177/2212585X231189338.

[33] Wu, W., 2025. China's Endeavors to Promote Educational Equity through Technology Use. *Science Insights Education Frontiers*, 27(2), pp.4533–4547. Available from: https://doi.org/10.15354/sief.25.re505.

[34] Xue, E., Li, J. and Li, X., 2021. Sustainable Development of Education in Rural Areas for Rural Revitalization in China: A Comprehensive Policy Circle Analysis. *Sustainability*, 13(23), p.13101. Available from: https://doi.org/10.3390/su132313101.

A. Complete search strings and dates

All database searches were executed on 15 October 2025, covering publications from 2010 to the search date, over titles, abstracts, and keywords. Boolean operators (AND/OR/NOT) and wildcards (*) were used as supported by each platform. For CNKI, Simplified Chinese equivalents of every term were used (subject-field search); for the other databases, searches were run in English. The Google Scholar search was restricted with title-focused operators and source exclusions to control noise. Table 5 reports the strings and per-database result counts; the counts sum to the 4,713 records reported in section 3.1, prior to deduplication.

Table 5: Search strategy by database (all searches run 2025-10-15).

Database	Search string	Date	Results
PubMed	("artificial intelligence" OR "AI" OR "generative AI") AND ("teacher performance evaluation" OR "bonus allocation" OR "educational equity") AND ("urban-rural inequality" OR "algorithmic bias" OR "rural revitalization") AND ("SDG4" OR "inclusive education")	2025-10-15	856
Scopus	TITLE-ABS-KEY(("AI empowerment" OR "explainable AI" OR "generative AI") AND ("teacher assessment" OR "performance model" OR "fairness pathways") AND ("urban-rural gap" OR "inequality" OR "ethical risks")) AND PUBYEAR > 2009 AND PUBYEAR < 2026	2025-10-15	1423
Web of Science	TS=("AI applications*" OR "artificial intelligence in education*") AND TS=("teacher performance" OR "bonus equity" OR "assessment bias") AND TS=("urban-rural educational inequality" OR "rural attrition" OR "algorithmic fairness") AND PY=2010–2025	2025-10-15	1120
CNKI	Subject-field (SU) search using Simplified Chinese equivalents of: ("artificial intelligence" OR "AI") AND ("teacher performance evaluation" OR "bonus allocation") AND ("urban-rural inequality" OR "educational equity" OR "rural revitalisation") NOT ("non-education")	2025-10-15	987
Google Scholar	allintitle: "AI education fairness urban-rural" OR "GenAI teacher performance China" (blog and forum domains excluded)	2025-10-15	327

B. Data extraction codebook

Extraction from each included study was performed independently by two coders using the codebook in table 6, with disagreements resolved by discussion and, where needed, third-party arbitration. Thematic coding was non-exclusive. Average extraction time was 30–45 minutes per study; extraction was managed in spreadsheet form. Urban–rural relevance was determined by full-text search for the corresponding terms in English and Chinese; studies without such content were coded "not applicable" on that variable.

Table 6: Data extraction codebook.

Variable	Description	Extraction rule / example
Basic information		
Study ID	Unique sequential identifier	Assigned 1–30
Authors	First author and collaborators	From title page (e.g., Crompton & Burke)
Year	Publication year	From record metadata (e.g., 2023)
Title	Full title	Copied verbatim
Journal/source	Publishing venue	From record metadata
DOI/URL	Persistent identifier	From record; "none" if absent
Methods information		
Study type	Qualitative / quantitative / mixed / review	Classified from the methods section
Sample size	Participants or studies included	From methods or results; "not reported" if absent
Time range	Years covered by the study	From methods
Geographic focus	Country or region	From abstract or introduction (e.g., China; global)
Findings information		
Main themes	AI applications; fairness pathways; urban–rural inequality (multi-select)	Independent dual coding, consensus by discussion
Quantitative estimates	Any reported effect estimates (e.g., d, r, OR)	Recorded verbatim with context; "none" if absent
Key findings	Brief summary of AI role, bias, ethics content	From abstract or conclusion, max. 100 words
Urban–rural relevance	Whether urban–rural disparity is addressed; if so, how	Full-text term search (English and Chinese)
Quality appraisal		
Quality rating	High / moderate / low confidence	AMSTAR-2 domains for reviews; analogous rigour criteria for primary studies (see appendix E)

C. Meta-analytic effect size methodology

This appendix documents the effect-size machinery the review design envisaged for its quantitative subset, so that a future reproducible synthesis can apply it directly.

For a two-group comparison with means M_1, M_2, standard deviations SD_1, SD_2, and group sizes n_1, n_2, the standardized mean difference is

$$d = \frac{M_1 - M_2}{SD_{\text{pooled}}}, \qquad SD_{\text{pooled}} = \sqrt{\frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{n_1+n_2-2}},$$

with approximate variance

$$\text{Var}(d) \approx \frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}.$$

Studies are combined by inverse-variance weighting, $w_i = 1/\text{Var}(d_i)$; under a fixed-effect model the pooled estimate is $\bar{d} = \sum w_i d_i / \sum w_i$ with $SE = 1/\sqrt{\sum w_i}$, and heterogeneity is assessed via $Q = \sum w_i (d_i - \bar{d})^2$ and $I^2 = \max\{0, (Q - df)/Q\} \times 100\%$, switching to a random-effects model (adding the between-study variance τ^2) when heterogeneity is material. Sensitivity analysis excludes influential studies; small-study effects are assessed using funnel plots and Egger's regression. In R, the `metafor` call is `rma(yi = effect, vi = var, method = "REML")`.

D. Simulation details

This appendix specifies the assumption-driven simulation referred to in sections 4.3 and 5.1. Its purpose is to make the 4DG model's intended mechanism explicit and, in principle, quantifiable, not to predict outcomes.

The model is a two-parameter decomposition. Let b denote the share of the urban–rural bonus gap attributable to correctable evaluation bias (as opposed to genuine differences in measured duties, seniority structure, or local funding), and let s denote the proportion of that bias component that 4DG-style safeguards (context-adjusted weighting, bias audits, data-integrity checks) succeed in removing. The relative reduction in the overall gap is then simply

$$R = b \times s.$$

The simulation assumes $b \in [0.30, 0.40]$ – i.e., evaluation bias accounts for 30–40% of the gap, an assumption motivated by, but not derived from, the qualitative evidence that urban-normed criteria systematically under-rate rural work (section 4.2) – and $s \in [0.50, 0.65]$, i.e., safeguards correct one half to two thirds of that component, a deliberately conservative range reflecting that no bias-mitigation program is complete. Sampling (b, s) uniformly over these ranges yields R between 0.15 and 0.26, with central scenarios around 0.20; table 7 shows representative points. This is the basis for the "roughly 15–25%" illustration quoted in the main text.

Table 7
Representative simulation scenarios (all inputs assumed).

Scenario	Bias share of gap b	Correction share s	Gap reduction $R = b \times s$
Conservative	0.30	0.50	15%
Central	0.35	0.57	20%
Optimistic	0.40	0.65	26%

Three caveats bound the exercise. The decomposition is linear and ignores interactions (e.g., between bias correction and participation-driven behaviour change); the parameter ranges are judgments, not estimates; and the gap concept refers specifically to the bonus gap, not the full compensation or resource gap. The empirical program of section 5.3 is designed to replace both parameters with measured quantities.

E. AMSTAR-2 quality appraisal

AMSTAR-2 [as applied in educational AI reviews by, e.g., 5] appraises reviews on sixteen domains, of which seven are critical: protocol registration (item 2), search adequacy (4), exclusion justification (7), risk-of-bias assessment (9), meta-analytic appropriateness (11), interpretation of bias (13), and publication-bias assessment (15). Confidence is rated high (no or one non-critical weakness), moderate (more than one non-critical weakness), low (one critical flaw), or critically low (more than one critical flaw). In the aggregate appraisal performed for this review, most included studies were judged to be of moderate confidence – the most common weakness across the corpus being the absence of pre-registered protocols, which is typical of the AIEd literature – and a small number of large systematic reviews were judged high-confidence. This aggregate characterisation guided the narrative weighting of evidence in sections 4 and 5 (higher-confidence reviews anchor field-level claims; single-context studies are cited as illustrative).