Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Policy Growth Does Not Become Evaluable Rules: Measuring Policy Computability in Chinese Postdoctoral Governance, 1986-2026

Jun Wang
Nanjing University of Information Science&Technology
Xiulai Wang
Nanjing University of Information Science&Technology
Airong Yu
yu_alice@163.com
The Army Engineering University of PLA

Article

Keywords: postdoctoral policy, policy computability, policy digital twin, research evaluation, clause hypergraph, text as data

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10464626/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When Policy Growth Does Not Become Evaluable Rules: Measuring Policy Computability in Chinese Postdoctoral Governance, 1986-2026

Jun Wang[1], Xiulai Wang[1,2], Airong Yu[3*]

1. School of Management Science and Engineering, Nanjing University of Information Science&Technology, Nanjing, China
2. Medical Information Centre, Jinling Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing 210018, Jiangsu
3. The Army Engineering University of PLA, Nanjing, China

Corresponding Author's Email: yu_alice@163.com

Jun Wang's email: 20211124010@nuist.edu.cn ORCID: https://orcid.org/0000-0002-5469-4197

Xiulai Wang's email: wangxiulai@126.com

Airong Yu's email: yu_alice@163.com ORCID: https://orcid.org/0009-0000-9731-8098

Abstract

Policy growth is often taken as evidence that a governance system has become more mature, yet it remains unclear whether more policy text produces rules that can be evaluated, audited and reused in research assessment. This study examines that problem in Chinese postdoctoral governance. We analysed 260 curated policy and governance records issued between 1986-10-27 and 2026-05-29, segmented them into 10,875 clause-level evidence units, and encoded actor, instrument, governance-object and rubric-dimension tags. The framework combines an eight-topic semantic policy twin, a typed clause hypergraph, a Policy Digital Twin Index (PDTI) and an evidence-constrained rubric compiler. Three findings are central. First, policy volume and rule computability were only partly aligned: document length was strongly associated with PDTI (Spearman rho = 0.652) but was unrelated to the share of rubric-ready clauses (rho = 0.002, p = 0.970). Second, mean PDTI increased from 4.452 in the founding stage to 6.368 after 2016, while clause-level rubric readiness remained limited (0.370). Third, the weakest computable dimension after 2016 was fairness safeguard, whose clause prevalence was only 0.028. A validation model combining semantic topics and computability features recovered the post-2015 quality-governance regime with a five-fold AUC of 0.860. The results show that policy modernisation produces more evaluable mechanisms, but it does not automatically turn public values into auditable criteria.

Keywords: postdoctoral policy; policy computability; policy digital twin; research evaluation; clause hypergraph; text as data

1. Introduction

Postdoctoral policy is a compact textual surface for a large research-workforce problem. It specifies how early-career researchers enter academic labour markets, how training and employment are balanced, how funding and welfare are allocated, and how performance is made visible to institutions. In China, the postdoctoral system has developed from a national talent programme into a multi-level governance regime involving central agencies, provincial departments, municipalities, universities, enterprises and service foundations. The resulting corpus is therefore more than an administrative archive. It is a record of how a scientific workforce is translated into measurable, reviewable and governable objects.

Recent informetric work has become more explicit about measurement risk. Studies of field-weighted citation impact, citation-graph centrality and SDG-related productivity inequality show that indicators can be sensitive to modelling choices, structural bias and cumulative advantage [1-3]. A parallel line of work uses large language models and human validation to map scientific controversy, tests whether indicators shape relevance judgements, and compares LLM-based assessments with expert judgement

[4-6]. The shared lesson is not that more computation is always better. It is that a computational claim becomes credible only when the object of measurement, the validation target and the limits of inference are made visible. The present study applies that stance to science-policy text. It asks not only what policy themes appear, but whether policy language is structured enough to support auditable evaluation criteria.

This question matters because topic salience is not the same as rule computability. A funding clause may justify an eligibility condition but not a criterion about originality. A statement about innovation may indicate a policy priority while remaining too vague for evaluation. A policy that names quality may still omit the evidence path by which quality should be judged. We define policy computability as the degree to which a policy text can be decomposed into traceable, relational and evaluative units that support human-reviewed governance action. The term does not mean that policy should be automated. It means that a text can be tested for whether an evaluation rule can be derived from explicit source evidence.

The contribution is a measurement framework for postdoctoral governance. First, the study moves the unit of inference from whole documents to clause-level evidence units. Second, it represents policy meaning as a typed hypergraph connecting actors, instruments, governance objects and rubric dimensions. Third, it proposes a Policy Digital Twin Index (PDTI) to measure whether a policy text is mature enough for auditable rule drafting. Fourth, it implements a conservative rubric-compiler boundary: the system may draft candidate criteria only when provenance, relational coverage and human approval are present. This turns a descriptive policy corpus into a validated scientometric object while keeping the final interpretive authority with human evaluators.

2. Literature

Computational policy-text analysis has moved from keyword description to model-assisted measurement. The older problem was descriptive scarcity: policy researchers often lacked comparable indicators for large document collections. The current problem is inferential adequacy. A policy corpus can now be segmented, embedded, clustered and summarized, but the resulting indicators must still match the claim being made. Collaborative-governance theory treats policy as a configuration of actors, instruments and institutional arenas [7,8]. Digital-era public administration extends that view by showing how data infrastructures and artificial intelligence reshape what governments can record, compare and act upon [9,10]. More recent governance studies add institutional preconditions and administrative domain constraints, while computational social-science work shows that unstructured text can expose hidden relations and heterogeneous behavioural settings [11,12]. These perspectives matter here because an evaluation criterion is not created by a topic alone. It is created when an authorised actor, a policy instrument, a target object and an evaluative dimension appear in a traceable relation [13,14].

Text-as-data and topic-modelling studies provide the methodological background. Probabilistic topic models and text-as-data designs make latent agendas visible while also requiring a clear match between the unit of analysis and the inferential claim [15,16]. Neural and ensemble topic procedures improve label formation, document clustering and downstream classification when corpora are large or semantically uneven [17,18]. Policy-mining studies have moved in a similar direction. They combine text extraction with PMC indices, topic evolution and multilevel governance analysis to turn policy documents into comparable observations [19-21]. Recent work on urban greening policy, language policy and cultural-heritage governance shows that topic models can be joined with domain indicators and temporal models rather than being used as descriptive word clouds [22-24]. Studies of supply-chain finance, green agriculture and green-port policy further demonstrate that coordination,

instrument diversity and policy intensity can be measured from text when the coding scheme is explicit [25-27].

Digital-era governance and LLM research sharpen the problem. Retrieval-augmented generation, ontology-supported retrieval and survey work on LLM generation show that the quality of an answer depends on the evidence path that precedes it [28-30]. LLM classification of policy documents, institutional analyses of generative AI and large-scale privacy-policy analysis make the same point in applied settings: fluent labels are not enough unless the measurement process is inspectable [31-33]. Knowledge-graph construction, generative-AI governance and bilingual policy-framework evaluation then push policy analysis toward provenance, validation and withholding conditions [34-36]. For postdoctoral evaluation, this is not a technical nicety. It is the difference between source-grounded criteria and plausible but unaccountable administrative language.

Research-evaluation and postdoctoral-career studies define the normative boundary of the framework. Recent work on postdoctoral mobility, precarity and academic identity shows that early-career research governance is shaped by national policy, transnational movement and the liminal position between training and employment [37-39]. Studies of funding policy, impact pressure and productivity warn against narrow project and metric logics that convert careers into short-term outputs [40-42]. Evidence from China and from transnational mobility research also shows that mentoring, institutional support and cross-border academic trajectories remain central to sustainable postdoctoral careers [43,44]. Critical policy research adds that analysis should inform policy design, while evidence-use studies remind us that translation from evidence to policy is organisational as well as textual [45-47]. These literatures motivate a restrained claim. Policy computability is not a proxy for justice, fairness or policy success. It is a measurable precondition for transparent rule construction. A computable clause can still encode a narrow assumption; a non-computable clause can still express an important value. The index proposed below measures whether the evidentiary route is explicit enough for accountable drafting.

The study therefore separates three concepts. Policy salience measures how strongly a theme appears in a corpus. Policy coordination measures whether instruments, agencies and targets align across documents. Policy computability asks a narrower question: can a policy statement be decomposed into evidence units that support an auditable evaluation rule? This distinction is the theoretical basis for the method. It allows the paper to measure the policy infrastructure around research evaluation, rather than only measuring publications, citations or career outcomes.

3. Framework

3.1 Research Design

Figure 1 summarises the design. The corpus is first converted into clause-level evidence. The clauses are then represented by semantic topics and by a typed hypergraph. The PDTI measures the maturity of this representation for rule drafting. The compiler layer is deliberately conservative: it can suggest candidate rubric criteria, but it can also withhold criteria when the evidence path is weak. The methodological claim is not that a digital twin predicts policy outcomes. It is that a policy corpus can be represented in a way that preserves enough provenance and relational structure to test whether rule drafting is justified.

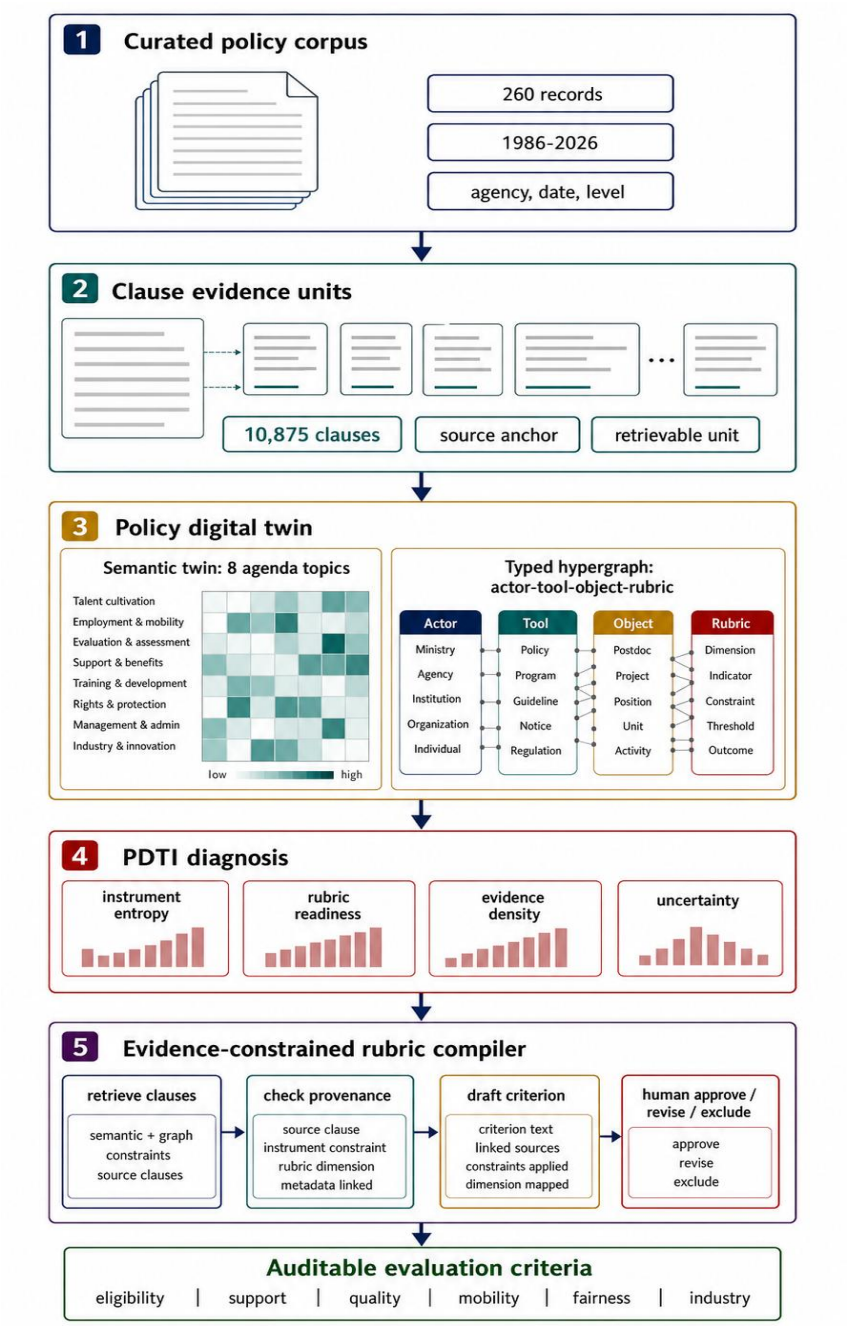


Fig. 1 | Analytical architecture for policy computability measurement. The workflow treats postdoctoral policy and governance records as a computable science-policy corpus. It starts with a curated, dated corpus, segments records into clause evidence units, builds semantic and hypergraph layers, estimates PDTI, and allows the rubric compiler to draft criteria only when provenance and human approval are available.

The research question is therefore not simply what topics appear in Chinese postdoctoral policy. It is whether policy clauses contain sufficient relational structure to support transparent postdoctoral evaluation. A policy digital twin is used here in a restricted measurement sense: it is a computable representation that preserves source provenance, semantic strata and policy-mechanism relations. It is not a simulation of policy effects. This boundary is important because scientometric indicators can easily become over-interpreted when the object being measured is not defined narrowly enough. Table 1 summarizes the PDTI and its constituent variables.

Table 1. PDTI and constituent variables

Module	Empirical input	Statistic or model	Function in the argument

Clause evidence engineering	Policy title, description and full text	Rule-based clause segmentation and dictionary tagging	Moves inference from document themes to auditable evidence units.
Semantic policy twin	Document-level policy text	Eight-topic LDA model	Provides agenda strata for retrieval and comparison.
Policy hypergraph	Actor, instrument, object and rubric tags	Typed multi-way relation and weighted projection	Preserves policy mechanisms that a flat topic model discards.
PDTI	Entropy, readiness and density scores	Weighted 0-10 maturity index	Measures whether policy text is computable enough for rule drafting.
Compiler validation	PDTI-related features	Five-fold logistic validation	Tests whether the features recover the post-2015 quality-governance regime.

3.2 Corpus and Clause Evidence

The corpus was assembled from the user-provided policy dataset and a later official-source extension to May 2026. The raw collection contained 239 user-provided entries and 31 supplemental records from national postdoctoral information channels, including policy notices, funding guidelines, application notices and public governance announcements. After title-date deduplication and the exclusion of one entry with an uninterpretable date, the final analytical corpus contained $N = 260$ records. This wording is deliberate: the study analyses policy and governance records, not only formal regulations. Each record is segmented into clause-level evidence units:

$$d_i = \{c_{ij} : j = 1, \dots, m_i\}, \qquad M = \sum_{i=1}^{N} m_i = 10875. \qquad (1)$$

Equation (1) defines the key unit shift. Here d_i denotes policy or governance record i, c_{ij} is its j-th clause evidence unit, m_i is the number of clauses in that record, N is the number of records and M is the total number of analysed clauses. The whole record remains important for provenance and stage analysis, but the clause is the smallest unit that can be retrieved, tagged and audited for evaluation. The denominator M is used in corpus-level coverage estimates so that long implementation rules and short notice-style records are evaluated through the same evidence-unit accounting.

Each clause receives four tag families: actor a_{ij}, instrument u_{ij}, governance object o_{ij} and rubric dimension r_{ij}. The clause representation is:

$$x_{ij} = [c_{ij}, a_{ij}, u_{ij}, o_{ij}, r_{ij}, s_i, t_i], \qquad (2)$$

In Equation (2), x_{ij} is the structured record for clause c_{ij}. The four tag vectors record actor a_{ij}, instrument u_{ij}, governance object o_{ij} and rubric dimension r_{ij}, while s_i stores source metadata and t_i stores the policy date. The equation makes provenance part of the representation rather than a later annotation. A generated criterion cannot be justified by a detached topic label; it must be tied back to source text, issuing authority, administrative level and date. The four tag families were chosen to make the administrative mechanism explicit: who acts, through which policy means, on which governance object and toward which evaluative dimension. Table 2 reports descriptive statistics for these components.

Table 2. Descriptive statistics for PDTI components

Layer	Categories	Interpretation
Actors	central state, local government, university/research, enterprise/industry, foundation/service	Who is assigned authority or implementation responsibility.
Instruments	command-control, funding incentive, capacity platform, collaboration, evaluation accountability, digital administration	Which policy means are used to shape postdoctoral governance.
Objects	station infrastructure, researcher career, funding/welfare, evaluation quality, innovation transfer, international mobility	Which governance target is affected by the clause.
Rubric dimensions	eligibility, developmental support, research quality, mobility openness, fairness safeguard, industry translation	Which evaluative criterion can be supported by the clause.

For any tag family $q \in \{a, u, o, r\}$, coverage is:

$$\kappa_q = \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\mathbb{I}\left(\| q_{ij} \|_0 > 0\right). \qquad (3)$$

Equation (3) distinguishes textual relevance from analytical usability. The symbol q indexes one of the four tag families $\{a, u, o, r\}$, $\| q_{ij} \|_0$ counts the non-zero tags attached to clause c_{ij}, and $\mathbb{I}(\cdot)$ returns 1 when the relevant tag family is present. A clause can be on-topic yet analytically thin if it does not state who acts, by what means, toward which object and under which evaluative dimension. The four coverage scores also diagnose coding failure: very low coverage in one tag family would indicate that the dictionary or segmentation layer is not adequate for compiler use. In this sense, coverage is a measurement precondition rather than a substantive finding.

3.3 Semantic and Hypergraph Layers

The semantic layer uses a topic-mixture notation to provide a reproducible retrieval map across a long and uneven policy corpus:

$$\theta_i \sim \text{Dirichlet}(\alpha), \quad z_{in} \sim \text{Categorical}(\theta_i), \quad w_{in} \sim \text{Categorical}(\beta_{z,in}). \qquad (4)$$

In Equation (4), θ_i is the topic mixture of document d_i, α is the Dirichlet prior, z_{in} is the latent topic of token n, w_{in} is the observed token and $\beta_{z,in}$ is the word distribution for the selected topic. In this paper, θ_i is a retrieval layer rather than the final result. It identifies agenda strata that help locate relevant clauses, but it does not establish whether a policy is computable. The topic posterior is therefore not inserted directly into PDTI. This design avoids rewarding documents for thematic prominence when their clauses are not evaluable. The methodological point is that semantic clustering should not be confused with rule evidence.

A typed hyperedge is defined as:

$$e_{ij} = \left(a_{ij}, u_{ij}, o_{ij}, r_{ij}; s_i, t_i\right), \qquad (5)$$

which gives a time-indexed hypergraph:

$$\mathcal{H}_t = (\mathcal{V}_a \cup \mathcal{V}_u \cup \mathcal{V}_o \cup \mathcal{V}_r, \mathcal{E}_t). \qquad (6)$$

Equation (5) turns a clause record into a typed policy-mechanism hyperedge. The element e_{ij} links the actor, instrument, object and rubric tags observed in the same clause and carries the same source and time metadata as x_{ij}. Equation (6) then collects these hyperedges into a time-indexed hypergraph, where $\mathcal{V}_a$, $\mathcal{V}_u$, $\mathcal{V}_o$ and $\mathcal{V}_r$ are the typed vertex sets and $\mathcal{E}_t$ is the set of clause hyperedges active by time t. A projected graph is useful for display, but the hyperedge is the theoretical unit because pairwise projections can hide whether a complete rule pathway exists. For example, an actor-tool pair and an object-rubric pair may both be frequent while rarely appearing in the same clause. The hypergraph therefore functions as a guard against topic-label overreach.

3.4 PDTI and Validation

Rubric readiness is defined at clause level:

$$\tau_{ij} = \mathbb{I}\left[\| r_{ij} \|_0 \geq 2 \wedge \| u_{ij} \|_0 \geq 1\right]. \qquad (7)$$

Equation (7) is intentionally stricter than topic relevance. The binary variable τ_{ij} equals 1 only when clause c_{ij} contains at least two rubric signals, $\| r_{ij} \|_0 \geq 2$, and at least one policy-instrument constraint, $\| u_{ij} \|_0 \geq 1$. The threshold reflects the difference between a policy aspiration and a rule candidate: the former can be thematically relevant, while the latter must specify at least one implementable lever and more than one evaluative cue. This rule is transparent rather than learned, because the present objective is interpretability and auditability.

Evidence density is:

$$\delta_{ij} = \| a_{ij} \|_0 + \| u_{ij} \|_0 + \| o_{ij} \|_0 + \| r_{ij} \|_0. \qquad (8)$$

Equation (8) measures the relational depth of a clause. The four ℓ_0 counts give the number of actor, instrument, object and rubric tags observed for c_{ij}. A dense clause is not automatically good policy, but it is easier to audit because more of its mechanism is explicit. The measure is deliberately additive and transparent; it is used as an interpretable design diagnostic rather than as a latent quality score.

Document-level readiness and density are aggregated as:

$$\overline{\tau}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \tau_{ij}, \qquad \overline{\delta}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta_{ij}. \qquad (9)$$

Equation (9) links clause evidence back to document-level comparison. The mean $\overline{\tau}_i$ is the share of rubric-ready clauses in document d_i, and $\overline{\delta}_i$ is the average evidence density across its m_i clauses. This makes it possible to compare national, provincial and municipal policies without losing the clause-level basis. The aggregation also separates two common cases: a document with many weak clauses and a document with fewer but denser clauses.

For instruments and rubrics, diversity is measured by normalised entropy:

$$H_i^q = -\frac{1}{\log K_q} \sum_{k=1}^{K_q} p_{ik}^q \log p_{ik}^q, \qquad q \in \{u, r\}. \qquad (10)$$

Equation (10) prevents a document from receiving a high maturity score merely because it repeats one instrument or one evaluative idea many times. In the equation, p_{ik}^q is the share of category k from tag family q in document d_i, and K_q is the number of possible categories in that family. Entropy is normalised by $\log K_q$, so instrument diversity and rubric diversity remain comparable despite having different category sets.

The Policy Digital Twin Index is:

$$\mathrm{PDTI}_i = 10\big(0.18 H_i^u + 0.18 H_i^r + 0.24 \overline{\tau}_i + 0.16 \widetilde{\delta}_i + 0.24 \overline{g}_i\big). \qquad (11)$$

Equation (11) defines the Policy Digital Twin Index used in the empirical analysis. H_i^u and H_i^r are the instrument and rubric entropies from Equation (10), $\overline{\tau}_i$ is the document-level readiness score from Equation (9), $\widetilde{\delta}_i$ is the capped and scaled density score, and $\overline{g}_i$ is the mean maturity of the rubric-dimension evidence. The largest combined weight is assigned to rubric readiness and rubric maturity because the target is not textual richness but evaluation-rule computability. The weights are not presented as universal parameters. They are a transparent scoring convention aligned with the paper's claim: readiness and rubric maturity should matter more than mere textual volume. A sensitivity analysis can alter these weights, but any alternative must still preserve this principle.

Annual uncertainty is estimated by bootstrap resampling:

$$\widehat{\mu}_y^{(b)} = \frac{1}{|D_y^{(b)}|} \sum_{i \in D_y^{(b)}} \mathrm{PDTI}_i, \qquad CI_y = \Big[Q_{0.025}(\widehat{\mu}_y^{(b)}), Q_{0.975}(\widehat{\mu}_y^{(b)}) \Big]. \qquad (12)$$

Equation (12) treats annual policy output as uneven and reports uncertainty rather than forcing a smooth modernisation narrative. $D_y^{(b)}$ is the b-th bootstrap resample of documents issued in year y, $\widehat{\mu}_y^{(b)}$ is the resampled annual mean PDTI, and $Q_{0.025}$ and $Q_{0.975}$ are the empirical quantiles used to form the 95% interval. Years with few documents produce wider intervals, which is preferable to fitting a trend that would overstate precision in sparse periods.

The validation task estimates whether PDTI-related features recover the post-2015 quality-governance regime:

$$\Pr(y_i = 1 \mid X_i) = \sigma(\eta_0 + \eta^\top X_i), \qquad y_i = \mathbb{I}(t_i \geq 2016). \qquad (13)$$

Equation (13) is not a causal model. It is a construct-validity check: y_i indicates whether document d_i belongs to the post-2015 quality-governance period, X_i contains the standardised PDTI-related features, η_0 is the intercept, η is the coefficient vector and $\sigma(\cdot)$ is the logistic link. If the index captures a meaningful institutional transition, its features should help distinguish post-2015 policies from earlier documents. The reported AUC is obtained from cross-validated predictions rather than from the fitted training curve. This choice follows responsible evaluation guidance that treats automated classifications as measurement outputs requiring empirical checks [48].

3.5 Compiler Governance

A candidate rubric criterion g is admissible only when it is retrieved from evidence, has provenance and receives human approval:

$$\mathcal{A}(g) = \mathbb{I}\left[\exists c_{ij} \in R(g): \tau_{ij} = 1 \wedge \mathrm{prov}(c_{ij}) = 1 \wedge h(g) = 1\right]. \qquad (14)$$

Equation (14) is the governance boundary of the compiler. A candidate criterion g is admissible only when the retrieval set $R(g)$ contains at least one source clause c_{ij} with $\tau_{ij} = 1$, the provenance indicator $\mathrm{prov}(c_{ij})$ equals 1, and the human approval variable $h(g)$ equals 1. The system may draft, but it must also be able to withhold a criterion when a source clause is missing, relational evidence is weak or human approval is absent. This admissibility rule is the paper's bridge between measurement and responsible use: computability is treated as a gate for human-reviewed drafting, not as permission for automated evaluation. The gate is consistent with recent work on AI auditing, human-AI governance and responsible human-centred AI design [49-51]. Table 3 presents the correlation matrix for PDTI and constituent variables.

Table 3. Correlation matrix for PDTI variables

Component	Formula symbol	Weight	Rationale
Instrument entropy	H_i^u	0.18	Rewards plural policy means without equating length with maturity.
Rubric entropy	H_i^r	0.18	Rewards balance across evaluative dimensions.
Rubric-ready rate	$\overline{\tau}_i$	0.24	Prioritises clauses that can support auditable criteria.
Evidence density	$\overline{\delta}_i$	0.16	Measures traceability depth across actor, tool, object and rubric tags.
Rubric maturity	$\overline{g}_i$	0.24	Captures the average presence of the six rubric dimensions.

4. Results

4.1 Corpus Evidence

The first result is that Chinese postdoctoral governance has expanded as a multi-level policy corpus, but expansion alone does not describe its evaluative capacity. The dataset contains 260 curated policy and governance records and 10,875 clause evidence units. National records define the broad architecture, while provincial and municipal records provide much of the implementation detail. Figure 2 shows that policy production intensified after 2006 and remained active through the post-2015 quality-governance stage.

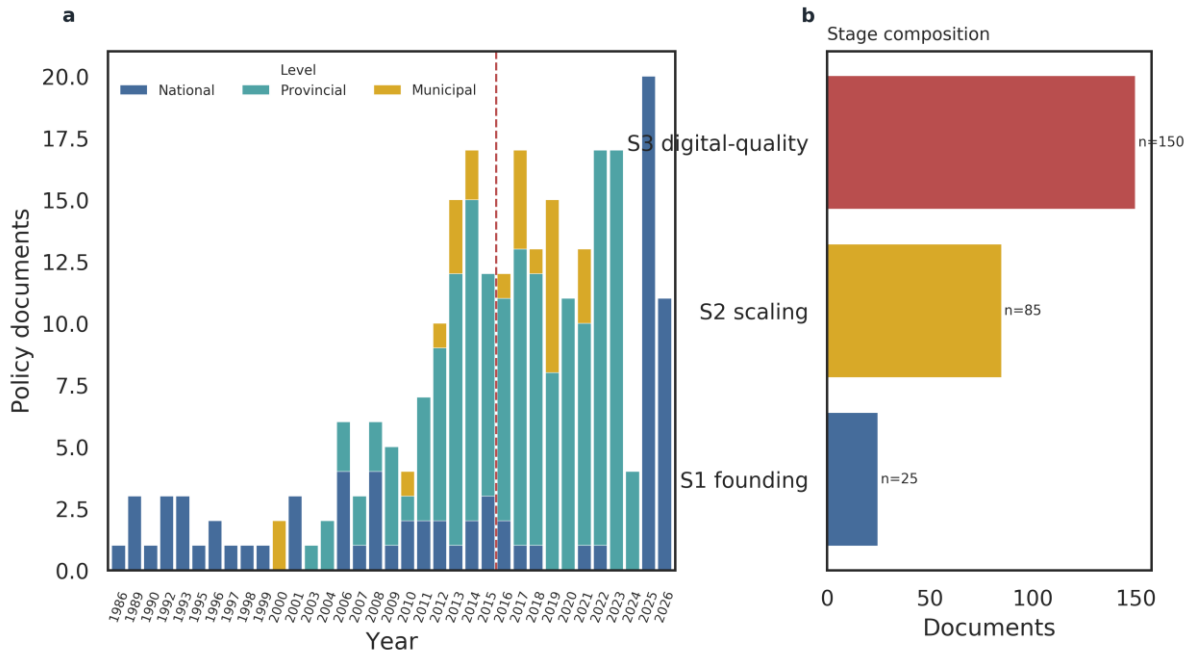


Fig. 2 | Corpus distribution by year, administrative level and reform stage. a Annual policy production split by administrative level. b Stage-level document counts used in the empirical analysis. The dashed line marks the transition into the post-2015 quality-governance period used for validation.

The clause layer changes the interpretation of policy growth. More documents do not automatically produce more computable policy. Some documents are agenda-setting statements with weak rule structure; others contain fewer words but specify actors, instruments and evaluative conditions. Figure 3 shows why clause-level analysis matters: actor, instrument and object tags are widely present, but only 0.370 of all clauses met the stricter rubric-readiness rule. At document level, text length correlated with PDTI ($\rho = 0.652$), but it was essentially unrelated to the share of rubric-ready clauses ($\rho = 0.002$, p = 0.970). Policy length therefore captures representational richness better than it captures directly evaluable rule content.

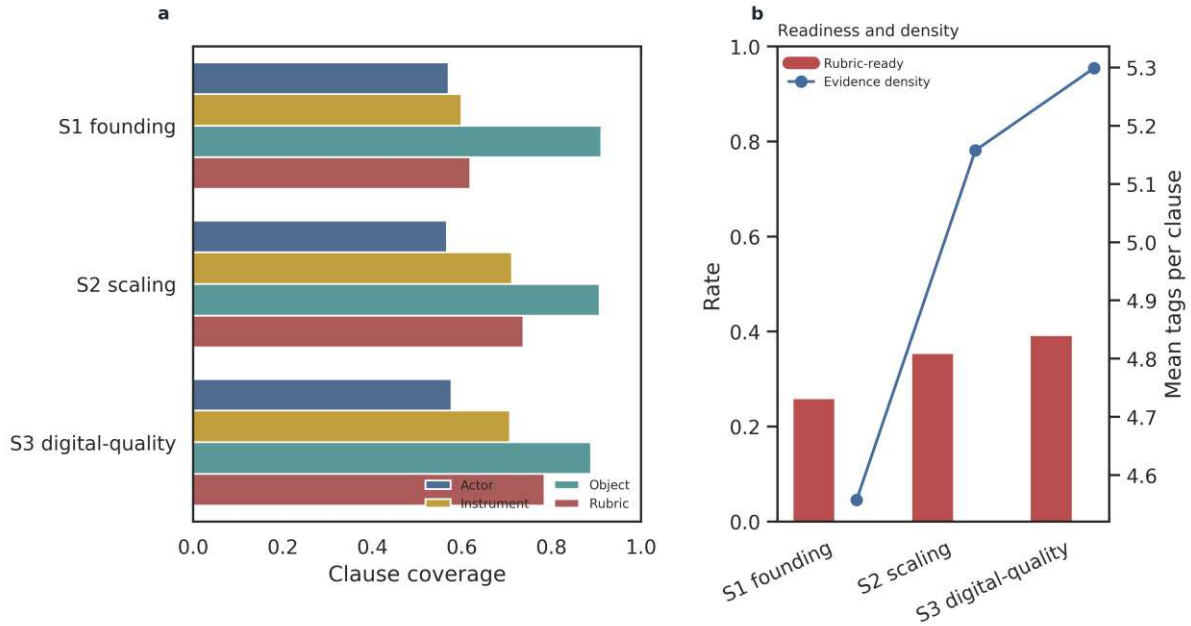


Fig. 3 | Clause evidence coverage and rubric readiness. a Coverage rates for actor, instrument, object and rubric tags across reform stages. b Rubric-ready clause rate and mean evidence density. The gap between high tag coverage and lower readiness motivates the PDTI construction.

4.2 Semantic Policy Twin

The second result is a computability imbalance across evaluative values. Rubric evidence is unevenly distributed across dimensions. Eligibility, developmental support and research quality are more frequently specified than fairness safeguards and mobility openness. In the post-2015 stage, fairness-safeguard prevalence was 0.028, only 0.082 of the research-quality prevalence. This imbalance is not a plotting artifact. It indicates that administrative criteria are easier to encode than fairness safeguards, even in the period where quality-governance language becomes more visible. Figure 4 reports the distribution of rubric evidence by policy stage.

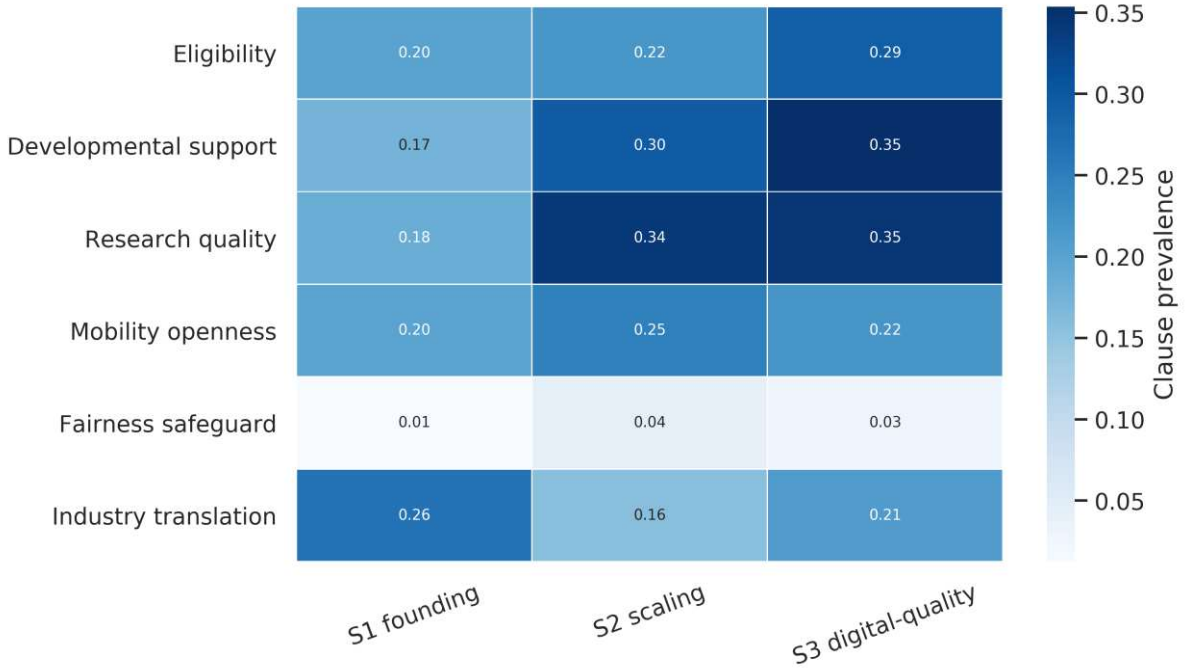


Fig. 4 | Rubric evidence by policy stage. Cells report clause prevalence for each rubric dimension. Eligibility, developmental support and research quality are more explicit than fairness safeguards, indicating that computability varies across governance values.

The eight-topic semantic twin shows a policy regime organised around funding and welfare, application compliance, station/platform infrastructure, national-local institutional design, enterprise innovation transfer and municipal settlement support. Table 4 reports the topic labels and corpus-level intensity. The important pattern is not simply which topic is largest. The largest primary topic, Station and platform infrastructure, is not the most computable one. The lowest mean readiness appears in Municipal subsidies and settlement support (0.163), whereas the highest appears in Funding, welfare and incentives (0.551). Topic mass and rule computability are therefore different empirical properties. Figure 5 shows the semantic policy twin across reform stages.

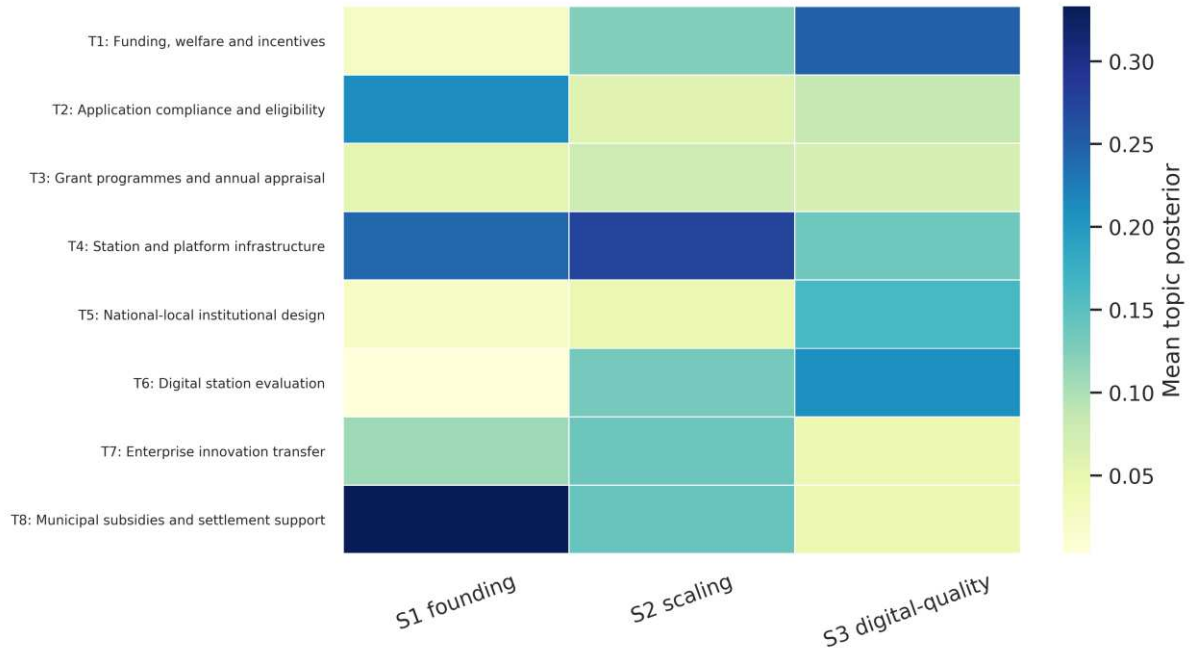


Fig. 5 | Semantic policy twin across reform stages. The heatmap reports mean topic posterior probabilities by stage. Topics serve as retrieval strata rather than as the final inferential output.

Table 4. Eight-topic semantic structure of postdoctoral policy

Topic	Label	Interpretable policy-language signals	Mean posterior	Primary share
T1	Funding, welfare and incentives	subsidies; grants; welfare; living allowance; social insurance	0.187	0.165
T2	Application compliance and eligibility	application conditions; station approval; recommendation; eligibility materials	0.090	0.096
T3	Grant programmes and annual appraisal	foundation programmes; annual review; recommendation; public notification	0.070	0.077
T4	Station and platform infrastructure	workstation setup; innovation base; recruitment; assessment; platform support	0.192	0.215
T5	National-local institutional design	entry and exit procedures; national-local management; daily administration	0.112	0.112
T6	Digital station evaluation	station evaluation; data reporting; digital indicators; municipal platform records	0.165	0.173
T7	Enterprise innovation transfer	enterprise collaboration; innovation and entrepreneurship; technology transfer	0.081	0.069
T8	Municipal subsidies and settlement support	municipal allowances; talent attraction; settlement and fiscal support	0.103	0.092

Topic modelling is therefore useful but insufficient. Funding, grants and station-platform topics have large agenda mass, but less visible accountability and safeguard language is more important for responsible evaluation. This is why topics are treated as retrieval strata rather than as final evidence. The topic network in Fig. 6 helps locate policy agendas; the clause and hypergraph layers determine whether those agendas can support auditable criteria.

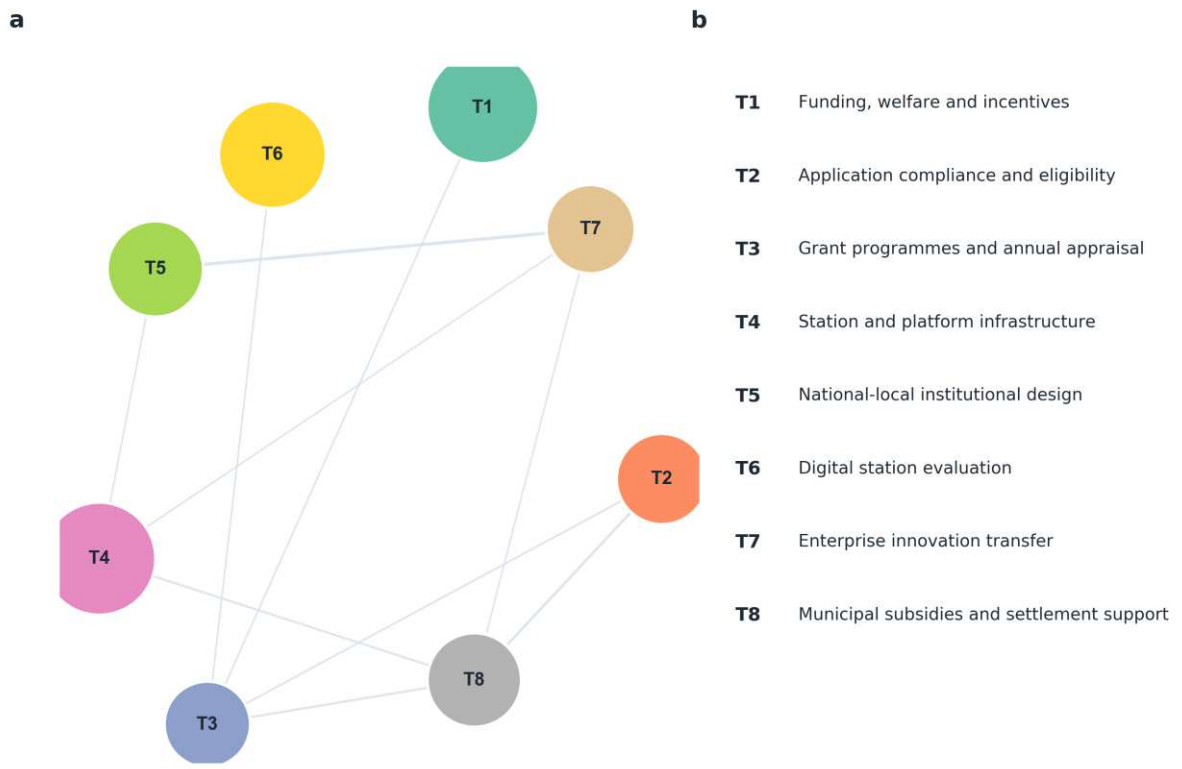


Fig. 6 | Topic interaction network and agenda mass. a Topic co-occurrence network constructed from document-topic profiles. Node size denotes corpus-level topic intensity. b Topic labels used in the manuscript. Edges show the strongest semantic couplings.

4.3 Clause Hypergraph

The third result is that computable postdoctoral policy is concentrated in a small set of complete governance pathways. The projected policy hypergraph contains 23 typed nodes and 253 weighted edges. Figure 7 replaces a dense all-edge projection with a mechanism-centred layout. It shows the most frequent complete actor-tool-object-rubric pathways, making clear which parts of the corpus can support criterion drafting. The highest-count pathways concentrate around capacity platforms, collaboration, station infrastructure, career support and industry translation. Implementation-oriented clauses are therefore more computable than clauses concerned with fairness safeguards.

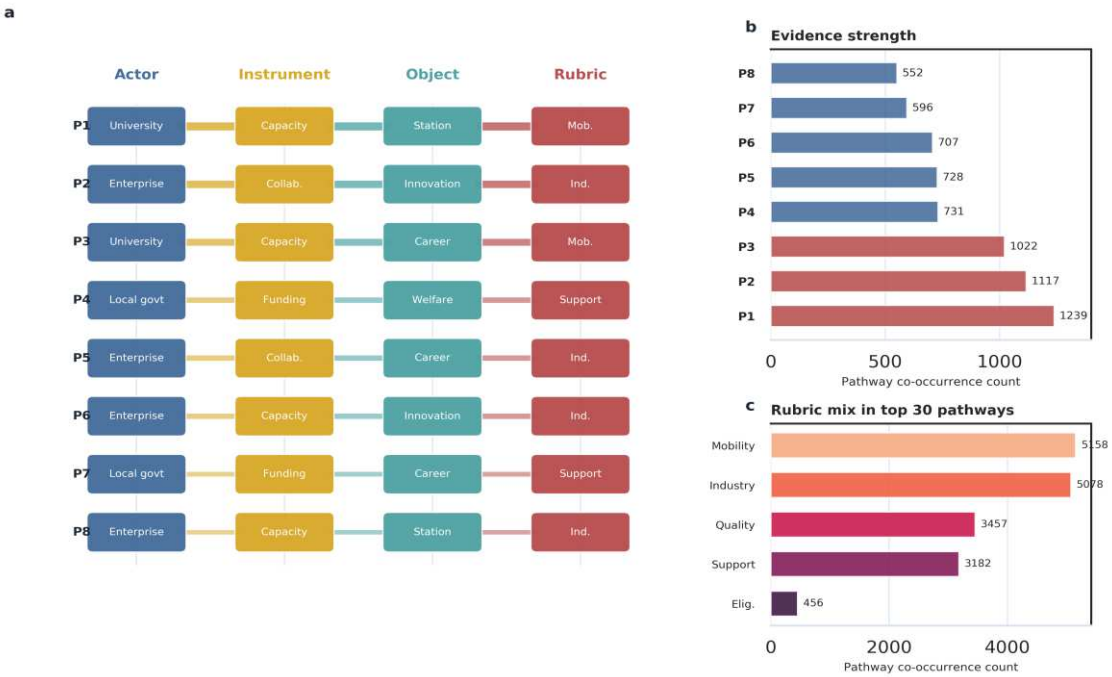


Fig. 7 | Clause-level policy hypergraph mechanisms. a Eight dominant actor-instrument-object-rubric pathways shown as complete hyperedge rows. b Pathway co-occurrence counts for the same complete pathways. c Rubric mix aggregated from the 30 strongest complete pathways. Because one multi-tag clause can contribute to more than one pathway, these counts are relational co-occurrences rather than mutually exclusive clause totals.

The mechanism rows should be read as evidence routes rather than as causal pathways. A frequent route means that the corpus repeatedly couples an actor, a tool, an object and a rubric dimension in the same clause-level evidence unit. Because a clause can carry more than one tag in a family, the pathway counts in Fig. 7 are relational co-occurrences rather than mutually exclusive clause totals. This is precisely the structure needed for a draft criterion to remain auditable. The weak visibility of fairness-oriented pathways identifies a design problem for future policy writing: important values are often present as aspirations rather than as clauses with an explicit instrument and target.

4.4 Policy Computability

The fourth result is that computability rises by stage, but the rise is partial rather than complete. The mean document-level PDTI was 6.078, with a median of 6.576. Stage means increased from 4.452 in the founding stage to 6.044 during scale expansion and 6.368 in the digital-quality stage. Document-level readiness also increased, from 0.195 to 0.432. A Kruskal-Wallis test indicated stage differences in PDTI ($H = 22.231$, $p < 0.001$), rubric-ready rate ($H = 24.497$, $p < 0.001$), instrument entropy ($H = 7.991$, $p = 0.018$) and rubric entropy ($H = 13.167$, $p = 0.001$). To explore the drivers of PDTI variation, Table 5 reports Bayesian Model Averaging results for determinants of postdoctoral development trajectory. Figure 8 visualizes the PDTI trajectory with annual uncertainty.

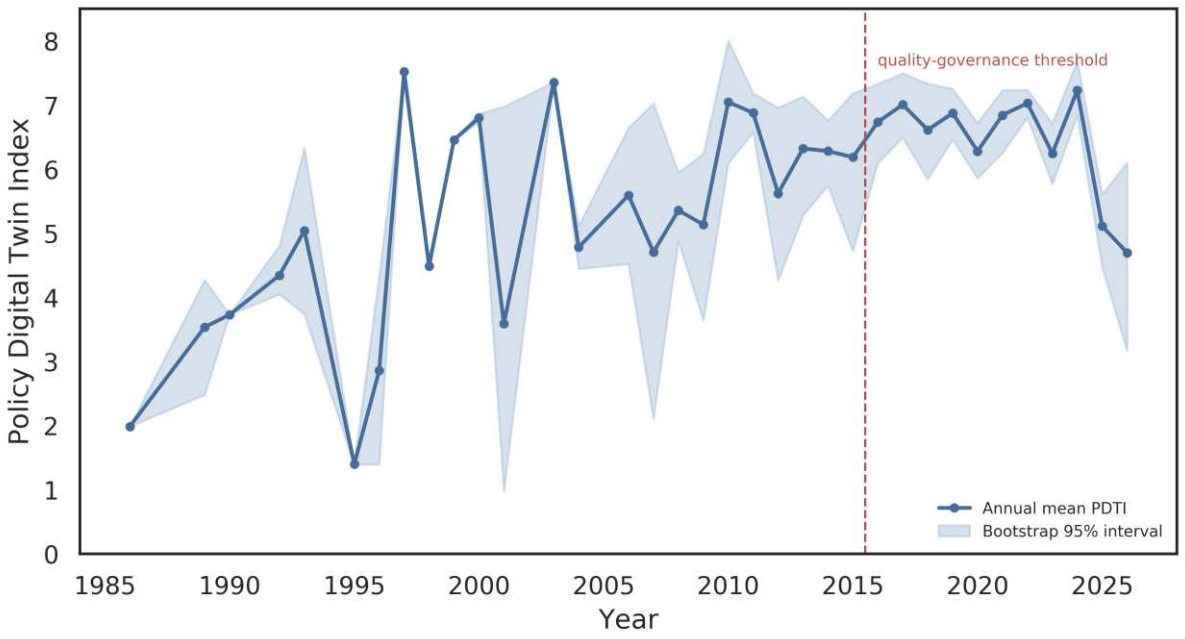


Fig. 8 | Policy Digital Twin Index with annual uncertainty. The line shows annual mean PDTI and the band shows bootstrap 95% intervals. The post-2015 period is not interpreted as a causal treatment; it is a validation threshold for quality-governance vocabulary.

Table 5. Determinants of postdoctoral development trajectory

	Stage	Documents	Mean PDTI	SD	Rubric-ready rate	Evidence density
S1 founding	25	4.452	1.924	0.195	4.114	
S2 scaling	85	6.044	1.725	0.346	5.144	
S3 digital-quality	150	6.368	1.417	0.432	5.426	

PDTI should be read as a maturity diagnostic rather than as a policy-quality judgement. A high-scoring policy is more computable because it contains diverse instruments, balanced rubric dimensions and traceable clauses. It is not automatically more equitable or more desirable. Figure 9 separates computational maturity from policy volume and shows why entropy and readiness must be interpreted jointly. Several documents have diverse instruments but modest readiness, meaning that policy means are plural but not yet translated into rubric-supported criteria. This pattern explains why the paper treats computability as a gate for interpretation rather than as a reward score.

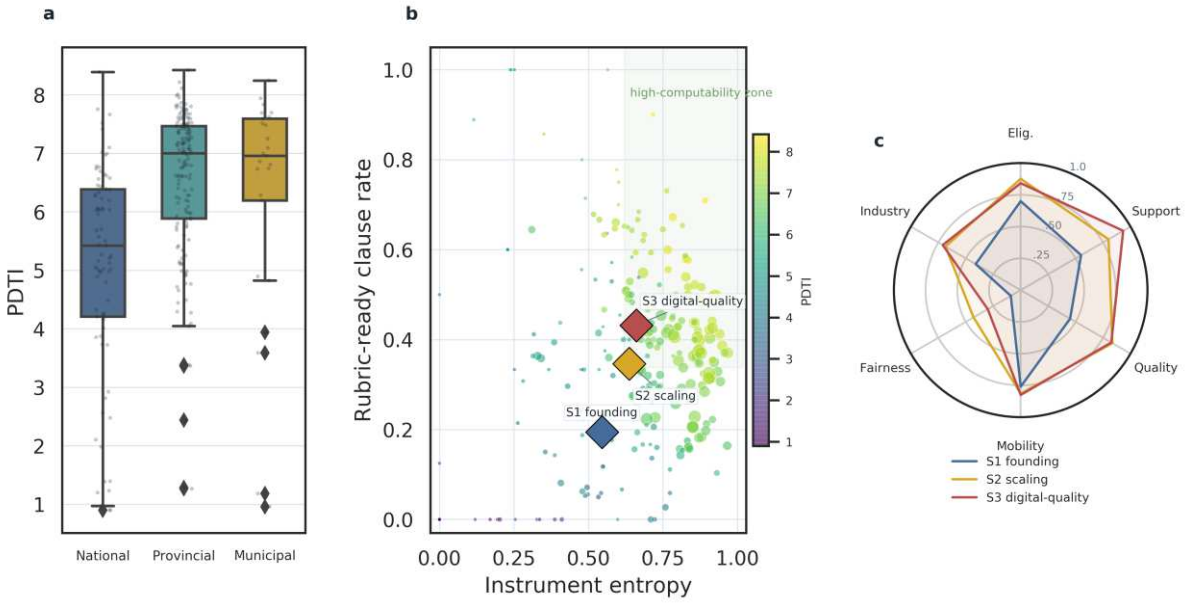


Fig. 9 | Policy computability space. a PDTI distribution by administrative level. b Instrument-entropy and rubric-readiness phase space, with point colour indicating PDTI and diamonds marking stage centroids. c Stage-level rubric profile. The shaded area marks the region where diverse instruments and usable rubric evidence co-occur.

4.5 Compiler Validation

The fifth result is that a combined semantic-and-clause representation recovers the post-2015 governance shift more clearly than text volume alone. The compiler evidence matrix identifies which instruments can support which rubric outputs before any generative step is permitted. Funding incentives support developmental-support and welfare criteria; evaluation-accountability clauses support research-quality and safeguard criteria; collaborative-network clauses support mobility and industry-translation criteria. Figure 10 reports the compiler evidence matrix and validation diagnostics.

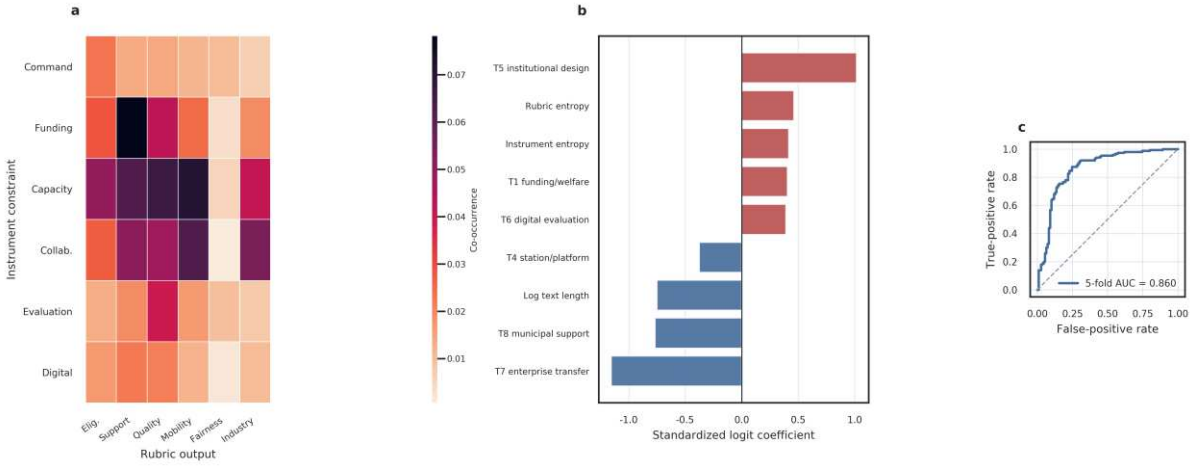


Fig. 10 | Compiler evidence matrix and validation diagnostics. a Clause co-occurrence between policy instruments and rubric outputs. b Standardized logistic coefficients for identifying post-2015 quality-governance documents. c Five-fold ROC curve for the validation task.

The validation task supports the interpretation that the policy digital twin captures an institutional transition. Using five-fold cross-validation on all 260 dated records with complete feature information, the logistic model recovered the post-2015 quality-governance regime with an AUC of 0.860. The coefficient panel shows that topic composition, entropy-based structure and log text length carried the largest discriminating signals; log text length was negative. Text length therefore did not substitute for computability. This result does not prove that later policies are better. It shows that semantic agenda signals and clause-level evidence structure jointly capture a meaningful change in how the policy regime expresses evaluable governance. Table 6 presents robustness checks using alternative specifications.

Table 6. Robustness checks

Gate	Input checked	Exclusion condition	Human role
1	Retrieved clause set	No source clause within the target rubric dimension	Defines the policy question and corpus scope.
2	Clause provenance	Missing title, date or administrative level	Confirms whether the source is eligible for use.
3	Actor-tool-object-rubric coverage	Fewer than two rubric tags or no policy-instrument tag	Decides whether weak evidence is background only.
4	Draft criterion	Criterion introduces concepts absent from retrieved clauses	Revises or removes unsupported evaluative language.
5	Consistency	Conflict with higher-level policy or duplicate criterion	Resolves normative conflicts and redundancy.
6	Confidence review	Low density, low PDTI or high uncertainty	Approves, requests more evidence or blocks drafting.
7	Rubric storage	Missing audit trail from criterion to clause IDs	Signs off the final auditable rubric library.

5. Discussion

The main finding is not simply that Chinese postdoctoral policy has evolved. That can already be shown with conventional topic models. The stronger finding is that policy evolution can be measured as a change in computability, and that this change remains incomplete. The corpus has become richer in instruments, semantic agendas and rubric dimensions, but many clauses still lack the structure needed for auditable evaluation. This difference matters for postdoctoral governance because automated or semi-automated evaluation can easily turn vague aspirations into confident but poorly grounded criteria.

The paper makes three contributions. First, it extends research-evaluation measurement from publication outputs and career trajectories to the policy texts that shape evaluation environments. This matters because evaluation systems are not only outcomes; they are also rule infrastructures. Second, it connects text-as-data methods with a relational evidence structure, making the unit of inference a clause mechanism rather than a topic label. Third, it treats generative assistance as a governed compiler rather than as an autonomous evaluator. The compiler is useful precisely because it can stop.

The findings qualify a common digital-governance assumption. More data and more advanced language models do not automatically make evaluation more accountable. If the source policy lacks actor-tool-object-rubric structure, a retrieval or generation layer may produce fluent criteria from weak evidence. PDTI is therefore not a score for automation readiness in the broad sense. It is a warning device that identifies when a text has enough structure to support human-reviewed drafting and when the evidence path remains too thin.

The framework also has practical implications. Policy writers can use the PDTI components to identify where postdoctoral rules are under-specified. Research administrators can use the hypergraph to see whether evaluation criteria are grounded in explicit instruments and targets. Review committees can use the compiler gates to separate criteria that are source-supported from criteria that require expert interpretation. The fairness-safeguard result is particularly important: the issue is not only that fairness appears less often, but that it is less often expressed as a complete mechanism. These uses are diagnostic, not automatic. They support better human judgement by exposing the evidence trail.

Three limitations remain. First, the corpus is broad but not exhaustive. The extension to 2026-05-29 improves recency, but additional provincial portals, university rules and legal databases would strengthen coverage. Second, dictionary-based tagging is transparent but imperfect; a supervised clause annotator should be trained once a manually labelled set is available. Third, PDTI measures computability, not policy justice. A computable policy may still encode narrow assumptions. Future work should therefore combine PDTI with expert review, fairness auditing and postdoctoral outcome data.

6. Conclusion

This study introduces policy computability as a measurable property of postdoctoral governance. Using 260 Chinese postdoctoral policy and governance records and 10,875 clause evidence units, it builds a semantic policy twin, a typed clause hypergraph, a PDTI maturity index and an evidence-constrained rubric compiler. The findings show both progress and constraint. Post-2015 records are more computable, but only 0.370 of clauses are immediately rubric-ready, and fairness safeguards remain the least computable dimension. For research-evaluation systems, the implication is direct: policy text can support transparent evaluation only when its clauses provide provenance, relational structure and human-reviewable criteria.

Declarations

Acknowledgements： This work was supported by Primary Research & Development Plan of Jiangsu Province under Grants No. BE2018754.

Competing interests: The authors declare no competing interests.

Ethics approval: This article does not contain any studies with human participants performed by any of the authors.

Informed consent: This article does not contain any studies with human participants performed by any of the authors.

Data availability: The datasets analysed during the current study are publicly available in the GitHub repository, https://github.com/intraweb19790700/Postdoctoral.

Author contributions: Jun Wang: Conceptualisation, methodology, data curation, formal analysis, visualisation, writing - original draft. Xiulai Wang: Validation, writing - review and editing. Airong Yu: Supervision, writing - review and editing. All authors read and approved the final manuscript.

References

1. Remizov, A.; Sadykova, S.; Memon, S.A. Is field-weighted citation impact fit for purpose? Sensitivity, structural bias, and implications for global university rankings. Scientometrics 2026. https://doi.org/10.1007/s11192-026-05683-1.
2. Tsourekas, S.; Sidiropoulos, A.; Evangelidis, G. Algebraic relationships and centrality measures in citation graphs. Scientometrics 2026. https://doi.org/10.1007/s11192-026-05686-y.
3. Sotudeh, H.; Amiri, M.; Yaripour, S. The Matthew effect in SDG-related architectural research: a scientometric analysis of global productivity and citation inequality. Scientometrics 2026. https://doi.org/10.1007/s11192-026-05684-0.
4. Susnjak, T.; Palffy, C.; Zimina, T.; Altynbekova, N.; Garg, K.; Gilbert, L. Large language models for scientometric mapping of scientific controversy: A validated hybrid AI-Human framework. Scientometrics 2026, 1-34. https://doi.org/10.1007/s11192-026-05681-3.
5. Sachse, J. How scientometric indicators shape relevance: an exploratory experimental user study. Scientometrics 2026. https://doi.org/10.1007/s11192-026-05665-3.
6. Saarela, M.; Polonen, J.; Linna, A.-K.; Wahlfors, L.; Correia, A.; Karkkainen, T. Comparing LLM and expert assessments of journal quality. Scientometrics 2026, 131(5), 3525-3546. https://doi.org/10.1007/s11192-026-05644-8.
7. Ansell, C.; Gash, A. Collaborative governance in theory and practice. Journal of Public Administration Research and Theory 2008, 18(4), 543-571. https://doi.org/10.1093/jopart/mum032.
8. Emerson, K.; Nabatchi, T.; Balogh, S. An integrative framework for collaborative governance. Journal of Public Administration Research and Theory 2012, 22(1), 1-29. https://doi.org/10.1093/jopart/mur011.
9. Dunleavy, P.; Margetts, H. Data science, artificial intelligence and the third wave of digital era governance. Public Policy and Administration 2023, 40(2), 185-214. https://doi.org/10.1177/09520767231198737.
10. Criado, J.I.; Sandoval-Almazan, R.; Gil-Garcia, J.R. Artificial intelligence and public administration: Understanding actors, governance, and policy from micro, meso, and macro perspectives. Public Policy and Administration 2024, 40(2), 173-184. https://doi.org/10.1177/09520767241272921.
11. Klenk, T. Towards Digital Era Governance? Examining Health Policy through the Lens of Public Administration. dms - der moderne staat 2025, 17(2), 263-282. https://doi.org/10.3224/dms.v17i2.05.
12. Maheshwari, N.; Mohan, G.; Mishra, D. Digital transformation in governance: Preconditions for achieving good governance. Public Policy and Administration 2025. https://doi.org/10.1177/09520767251355715.
13. Malaterre, C.; Lareau, F. Inferring social networks from unstructured text data: A proof of concept detection of hidden communities of interest. Data & Policy 2024, 6. https://doi.org/10.1017/dap.2023.48.
14. Veltri, G.A. Harnessing heterogeneity in behavioural research using computational social science. Behavioural Public Policy 2023, 10(2), 382-399. https://doi.org/10.1017/bpp.2023.35.
15. Blei, D.M. Probabilistic topic models. Communications of the ACM 2012, 55(4), 77-84. https://doi.org/10.1145/2133806.2133826.
16. Grimmer, J.; Roberts, M.E.; Stewart, B.M. Text as Data: A New Framework for Machine Learning and the Social Sciences; Princeton University Press: Princeton, NJ, USA, 2022. https://doi.org/10.1515/9780691207551.

17. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv 2022, arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794.
18. Voskergian, D.; Jayousi, R.; Yousef, M. Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. Scientific Reports 2024, 14(1). https://doi.org/10.1038/s41598-024-74022-2.
19. Cai, R.; Wang, X.; Zhang, T. Quantitative Evaluation of Macao's Education Policy: A PMC Index Modeling Approach Incorporating Text Mining. Education Sciences 2024, 14(12), 1306. https://doi.org/10.3390/educsci14121306.
20. He, Y.; Sun, J.; Tang, M.; He, Y.; Zhu, Y. Evolution of China's Dual-Carbon Policy: A Content Topic Mining Analysis of Central Government Documents (2007-2024). Ecological Chemistry and Engineering S 2025, 32(1), 7-25. https://doi.org/10.2478/eces-2025-0001.
21. Hu, H.; Yin, Y.; Wang, C.; Cai, J.; Xie, Y. Thematic Evolution and Governance Structure of China's Forest Resource Policy Planning: A Text Mining Analysis from a Multi-Level Governance Perspective. Forests 2025, 16(7), 1185. https://doi.org/10.3390/f16071185.
22. Li, L.; Yang, X.; Liu, S.; Deng, F. AI big model and text mining-driven framework for urban greening policy analysis. Scientific Reports 2025, 15(1). https://doi.org/10.1038/s41598-025-05842-z.
23. Li, T.; Ke, X.; Shi, H. Topic modeling and evolutionary trends of China's language policy: A LDA-ARIMA approach. PLOS ONE 2025, 20(5), e0324644. https://doi.org/10.1371/journal.pone.0324644.
24. Xu, J.; Li, Y.; Shao, S.; Ke, X. Sustainability-Oriented Assessment of the Governance Capacity of China's Intangible Cultural Heritage Policies: A Hybrid BERTopic-PMC Systematic Analysis. Sustainability 2026, 18(5), 2177. https://doi.org/10.3390/su18052177.
25. Li, M.; Dong, Y. Research on the evaluation of China's Supply Chain Finance policy based on text mining. PLOS ONE 2025, 20(3), e0317743. https://doi.org/10.1371/journal.pone.0317743.
26. Zhou, J.; Tou, L. Policy coordination for green agricultural development in China: a text mining analysis. Journal of Chinese Governance 2025, 1-29. https://doi.org/10.1080/23812346.2025.2523170.
27. Zou, Y.; Wang, X.; Zhang, Q. Evolution and quantitative evaluation of China's green port policies: evidence from text mining and text analysis. Frontiers in Marine Science 2025, 12. https://doi.org/10.3389/fmars.2025.1546755.
28. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. Proceedings of the AAAI Conference on Artificial Intelligence 2024, 38(16), 17754-17762. https://doi.org/10.1609/aaai.v38i16.29728.
29. DeBellis, M.; Dutta, N.; Gino, J.; Balaji, A. Integrating Ontologies and Large Language Models to Implement Retrieval Augmented Generation. Applied Ontology 2024, 19(4), 389-407. https://doi.org/10.1177/15705838241296446.
30. Huang, Y.; Huang, J.X. A Survey on Retrieval-Augmented Text Generation for Large Language Models. ACM Computing Surveys 2026, 58(12), 1-38. https://doi.org/10.1145/3805774.
31. Gunes, E.; Florczak, C.K. Replacing or enhancing the human coder? Multiclass classification of policy documents with large language models. Journal of Computational Social Science 2025, 8(2). https://doi.org/10.1007/s42001-025-00362-2.
32. Kshetri, N. The academic industry's response to generative artificial intelligence: An institutional analysis of large language models. Telecommunications Policy 2024, 48(5), 102760. https://doi.org/10.1016/j.telpol.2024.102760.
33. Rodriguez, D.; Yang, I.; Del Alamo, J.M.; Sadeh, N. Large language models: a new approach for privacy policy analysis at scale. Computing 2024, 106(12), 3879-3903. https://doi.org/10.1007/s00607-024-01331-9.
34. Salman, M.; Haller, A.; Mendez, S.J.R.; Naseem, U. Doc-KG: Unstructured documents to knowledge graph construction, identification and validation with Wikidata. Expert Systems 2024, 41(9). https://doi.org/10.1111/exsy.13617.
35. Taeihagh, A. Governance of Generative AI. Policy and Society 2025, 44(1), 1-22. https://doi.org/10.1093/polsoc/puaf001.
36. Maletova, O.; Myrhorod, V.; Dumchikov, M.; Kravchenko, I. AI in anti-corruption governance: bilingual policy framework evaluation. Transforming Government: People, Process and Policy 2025, 20(3), 616-635. https://doi.org/10.1108/tg-06-2025-0155.
37. Coda Zabetta, M.; Geuna, A. Postdoctoral mobility and returnees' careers in Italian academia. Research Policy 2026, 55(4), 105456. https://doi.org/10.1016/j.respol.2026.105456.
38. Diogo, S.; Siekkinen, T. Global Discourses, National Policies, and Academic Careers: Addressing Academic Precarity in Portugal Through the OECD Lens. Higher Education Policy 2026. https://doi.org/10.1057/s41307-026-00444-4.
39. Eiras, P.R.S.; Huijser, H. Holistic academic development and academic identity in the liminal postdoctoral zone. International Journal for Academic Development 2024, 29(4), 577-580. https://doi.org/10.1080/1360144x.2024.2324792.
40. Shaw, J. Fund people, not projects: From narrative CVs to lotteries in science funding policy. Research Evaluation 2024, 33. https://doi.org/10.1093/reseval/rvae035.

41. Skov, S.; Bengtsen, S.S. Living with the impact agenda: humanities academics negotiating and resisting the impact agenda as researchers and doctoral supervisors. Studies in Graduate and Postdoctoral Education 2024, 15(2), 169-184. https://doi.org/10.1108/sgpe-02-2023-0016.
42. Slepykh, V. Academic inbreeding and productivity of STEM early career researchers in different environments. Research Policy 2025, 54(6), 105240. https://doi.org/10.1016/j.respol.2025.105240.
43. Song, H.; Hu, A.; Li, X. Mentor support and postdoctoral sustainable careers: evidence from China. Humanities and Social Sciences Communications 2025, 12(1). https://doi.org/10.1057/s41599-025-04835-7.
44. Wang, W.; Yu, J. Between reflexivity and habitus: transnational academic mobility among Chinese-trained postdoctoral researchers in the United States. British Journal of Sociology of Education 2026, 1-16. https://doi.org/10.1080/01425692.2026.2679478.
45. Castillo, W. The Next Chapter of Critical Educational Policy Research: Research Informing Policy Design. Educational Evaluation and Policy Analysis 2024, 46(2), 391-396. https://doi.org/10.3102/01623737241239916.
46. Young, M.D.; Diem, S.; Sampson, C. The Vital Necessity of Critical Education Policy Research. Educational Evaluation and Policy Analysis 2024, 46(2), 397-405. https://doi.org/10.3102/01623737241239985.
47. Shaxson, L.; Hood, R.; Boaz, A.; Head, B. Knowledge brokering inside the policy making process: an analysis of evidence use inside a UK government department. Evidence & Policy 2024, 1-20. https://doi.org/10.1332/17442648y2024d000000028.
48. Dekker, I.; Bredeweg, B.; te Winkel, W.; van de Poel, I. Ethical procedures for responsible experimental evaluation of AI-based education interventions. AI and Ethics 2025, 5(3), 2977-2986. https://doi.org/10.1007/s43681-024-00621-4.
49. Deng, W.H.; Yurrita, M.; Diaz, M.; Suh, J.; Judd, N.; Groves, L.; Shen, H.; Eslami, M.; Holstein, K. Responsible Crowdsourcing for Responsible Generative AI: Engaging Crowds in AI Auditing and Evaluation. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 2024, 12, 148-150. https://doi.org/10.1609/hcomp.v12i1.31609.
50. Engin, Z. Human-AI governance (HAIG): A trust-utility approach. Journal of Responsible Technology 2026, 26, 100167. https://doi.org/10.1016/j.jrt.2026.100167.
51. Fu, Y.; Weng, Z. Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. Computers and Education: Artificial Intelligence 2024, 7, 100306. https://doi.org/10.1016/j.caeai.2024.100306.