

Article

Navigating AI in Higher Education: Toward Culturally Responsive Assessment Frameworks in the GenAI Era

Wei Yao ^{1,2}, Shengfan Qian ^{1,2,*}  and Wengang Xie ^{1,2} ¹ School of Public Affairs, Zhejiang University, Hangzhou 310058, China² Institute of China's Science, Technology and Education Policy, Zhejiang University, Hangzhou 310058, China

* Correspondence: qsf@zju.edu.cn

Abstract

The proliferation of generative artificial intelligence (GenAI) has precipitated an urgent, global reassessment of how higher education evaluates critical thinking, creative agency, and academic integrity. However, scholarly and institutional responses remain fragmented across cultural contexts, impeding the development of robust, flexible, and discipline-adaptable assessment frameworks. Responding to the imperative to move beyond the traditional standardized assessment paradigm, this study conducts a comparative discourse analysis of 5368 academic articles in Anglophone/Western scholarly discourse (Web of Science, WoS) and Chinese (CNKI). Using LDA topic modeling and Word2Vec-enhanced semantic analysis, the study identifies two divergent orientations: an Anglophone/Western discourse that frames AI as an instrument for cognitive augmentation, efficiency optimization, and functional human–AI collaboration; and a Chinese discourse that emphasizes epistemic sovereignty, the reconstruction of creative subjectivity, and systemic institutional rebuilding against technological alienation. These pathways are mapped onto a tripartite framework of Tools, Creative Subjectivity, and Organizational Ecosystems. The findings demonstrate that AI integration is culturally embedded rather than technically determined, carrying profound implications for assessment validity, academic integrity policy, and equitable access to AI-enhanced learning. The study synthesizes these insights into a culturally responsive assessment framework that redirects evaluation from standardized, product-centric outputs toward process-oriented, transparent, and ethically governed human–AI co-authorship. By centering critical autonomy, AI literacy, and epistemological diversity, this framework offers actionable strategies for inclusive assessment redesign, institutional policy development, and sustainable competency cultivation in the GenAI era.



Academic Editors: Marie Devlin and Steve Riddle

Received: 22 May 2026

Revised: 24 June 2026

Accepted: 25 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: AI assessment frameworks; generative AI; academic integrity; AI literacy; competency frameworks; creative agency; comparative discourse; higher education policy

1. Introduction

Recent institutional responses to generative artificial intelligence (GenAI) in higher education have undergone a rapid, three-phase transition: from initial attempts to defend assessment integrity through prohibition and detection, through pragmatic acceptance that AI use must be structurally accounted for in teaching and evaluation, toward an emerging consensus that students must cultivate AI literacy, critical autonomy, and ethical competency as graduate attributes (Perkins & Roe, 2023; Smith et al., 2025; Corbin et al., 2025).

GenAI assessment reform is framed as requiring simultaneous attention to ethical governance and epistemic justice. Yet this transition remains uneven. Educators are largely redesigning assessment in disciplinary isolation, often without concrete institutional guidance, while students remain uncertain whether they are prohibited, permitted, or encouraged to use GenAI tools and how best to do so transparently (Smith et al., 2025). What is urgently needed—and still largely absent—are robust, flexible, and culturally informed assessment frameworks that can accommodate rapidly evolving AI capabilities without sacrificing academic integrity or humanistic educational values (Corbin et al., 2025).

A critical barrier to developing such frameworks is the lack of cross-cultural empirical grounding. Existing literature has extensively explored AI tools, pedagogical applications, and localized institutional policies, but limited comparative analysis has been conducted at the discourse level to examine how different educational cultures conceptualize the future of assessment, thinking skills, and creative agency in an AI-saturated academy. To unpack these dynamics, this study moves away from imprecise universal characterizations of ‘international’ scholarship, replacing vague labels with a precise operational contrast between “Anglophone/Western scholarly discourse” (represented by the Web of Science corpus) and the distinct epistemological and institutional logics of non-Anglophone scholarly traditions.

The comparative scope of this work evaluates the Web of Science (WoS) SSCI/A&HCI corpus as a body of scholarship predominantly embedded within Anglo-American, Australian, and Western European higher education systems. Conversely, the China National Knowledge Infrastructure (CNKI) CSSCI corpus is analyzed as a comparable but analytically distinct knowledge system. China is specifically selected as a rigorous comparator for three theoretically grounded reasons: (a) it operates the world’s largest higher education system by enrollment, meaning its macro-discursive shifts carry immense systemic weight; (b) the CNKI/CSSCI database constitutes the largest non-English academic repository globally, offering a robust structural counter-weight to Western-centric data pools; and (c) emerging bibliometric signals suggest Chinese scholarly discourse on GenAI places explicit emphasis on epistemic sovereignty, national paradigm transformation, and subjectivity reconstruction—concerns that are comparatively muted or framed purely as technical considerations in mainstream Anglophone literature. This positions the comparison as theoretically purposeful and robust rather than geographically arbitrary, intentionally avoiding Western-centric universalism.

This study addresses this gap through a comparative discourse analysis using Latent Dirichlet Allocation (LDA) topic modeling and Word2Vec-enhanced semantic interpretation. LDA is a generative probabilistic topic model based on bag-of-words assumptions. Word2Vec is a neural embedding technique used here post hoc to enhance topic interpretation by identifying semantic neighbors of high-probability topic words. LDA2Vec (Moody, 2016) is a distinct hybrid model that jointly learns document embeddings and topic distributions, which this study does not employ. Drawing on 5368 academic articles from WoS and CNKI, the research aims to:

- a. Identify and compare the latent thematic structures in Anglophone/Western and Chinese AI-in-education discourses, with particular attention to how each tradition frames the legitimacy and boundaries of AI-mediated assessment;
- b. Reveal underlying epistemological differences in conceptions of thinking, creativity, and creative agency, illuminating how distinct cultural contexts envision the traditional standardized assessment paradigm that presupposes a single, fixed, and verifiable solution to a given task—typified by closed-form examinations, standardized essay templates, and polished final products that can be independently generated by large language models and their alternatives in the GenAI era;

- c. Propose a theoretically grounded, tripartite framework—encompassing Tools, Creative Subjectivity, and Organizational Ecosystems—to guide culturally responsive assessment redesign, academic integrity policy development, and AI literacy cultivation across diverse higher education contexts.

Critical AI literacy encompasses three tiers: (1) Operational: evaluating the accuracy and appropriateness of specific AI outputs; (2) Epistemic: interrogating the assumptions, biases, and limitations embedded in AI-generated knowledge; and (3) Sociotechnical/Meta-level: critically examining GenAI as a sociotechnical system—its training data regimes, corporate governance structures, environmental costs, and epistemological implications for humanistic education.

By moving beyond cataloguing topical differences to explicating the divergent “logics of practice” embedded in each discourse, this study provides an empirical foundation for assessment frameworks that are neither technocratically deterministic nor culturally parochial. The remainder of this paper is structured as follows: Section 2 situates the study within a theoretical framework centered on creative agency; Section 3 details the comparative corpus construction and analytical methodology; Sections 4 and 5 present and discuss the findings; and Section 6 derives concrete implications for pedagogy, assessment, and institutional policy.

2. Theoretical Framework: Toward a Tripartite Model for Cultivating Creative Agency in AI-Assessment Ecosystems

To analyze how scholarly discourses conceptualize AI’s impact on thinking and creativity—and, crucially, to translate these insights into actionable assessment reform—we construct a tripartite framework centered on the concept of Creative Agency. This framework synthesizes insights from socio-cultural theories of learning, critical pedagogy, and innovation ecosystems to interpret the divergent pathways identified in our analysis. It posits that the integration of AI is not merely a technical adoption but a process that reconfigures the tools of cognition, the subjectivity of the learner/thinker, and the organizational ecosystems that scaffold creative practice. By mapping discursive differences onto these three dimensions, the framework provides a diagnostic and generative structure for designing culturally responsive assessments, integrity policies, and AI literacy curricula.

2.1. Creative Agency as the Core Construct

At the heart of this study lies the concept of Creative Agency—the capacity of individuals to intentionally generate novel and valuable ideas, solutions, or expressions while maintaining critical autonomy over their process (Glăveanu, 2013; Arseven & Bal, 2025). In the context of GenAI, creative agency involves the negotiated interplay between human intention and algorithmic suggestion. It encompasses not only the skill to use AI tools but, more crucially, the metacognitive and critical ability to direct, evaluate, and ethically govern that use. Furthermore, we theorize that student agency cannot be divorced from a comprehensive critical AI literacy, which must expand beyond mere output-level critique to encompass three distinct tiers: (1) the Operational tier, focusing on the functional accuracy and algorithmic appropriateness of specific AI tools; (2) the Epistemic tier, which interrogates the hidden assumptions, data biases, and systemic limitations embedded within AI-generated knowledge; and (3) the Sociotechnical/Meta-level tier, which critically examines GenAI as an entire macroeconomic and political infrastructure—evaluating its training data regimes, corporate governance structures, carbon/environmental footprints, and epistemological implications for humanistic education. By connecting cross-cultural assessment fairness to these multi-tiered forms of agency, our framework explicitly remains responsive to epistemological diversity.

This perspective is particularly vital for assessment reform in the GenAI era. When students can delegate knowledge generation to large language models, the traditional proxy for learning—the polished final product—becomes epistemologically unstable. Assessment must therefore shift from evaluating static outputs to evaluating the process of human–AI negotiation: the capacity to frame original problems, interrogate algorithmic suggestions, synthesize disparate sources, and defend the ethical and intellectual choices made along the way. Creative agency, in this sense, becomes the foundational competency that assessment frameworks must cultivate and certify, moving beyond the search for a single correct answer toward recognizing diverse, critically grounded pathways to knowledge.

2.2. Integrating Comparative and Assessment Validity Literature

When students can delegate knowledge generation to large language models, the traditional proxy for learning—the polished final product—becomes epistemologically unstable. To ground our framework in established scholarship, we draw upon comparative higher education literature (e.g., [Marginson, 2022](#)), which theorizes how national educational cultures differentially construct technology adoption, epistemic authority, and assessment legitimacy. Furthermore, we explicitly integrate [Messick's \(1989\)](#) unified validity framework alongside subsequent psychometric extensions ([Shavelson et al., 2002](#)) to argue that GenAI poses an unprecedented threat to the construct validity of standardized assessments. Specifically, GenAI collapses the operational distinction between construct-irrelevant variance and true student competence by masking skill deficits behind machine-generated elegance. When final student performance is corrupted by untraceable algorithmic intervention, traditional evaluative instruments suffer from extreme construct underrepresentation.

To resolve this crisis, our framework integrates authentic assessment scholarship with process-based assessment literature to argue that validity can only be restored if the object of measurement is shifted from static final products to certified processes and intentions. Simultaneously, we draw on contemporary academic integrity frameworks ([Perkins & Roe, 2023](#)) to incorporate ethical governance of AI use directly into the construct definition of the “Subject.” Ethical behavior is thus repositioned from an external code enforced by detection software to an internal pedagogical component of creative agency itself. In practice, hybrid intelligence facilitation directly addresses the assessment and validity risks outlined above: by formally partitioning routine generative work to LLMs while centering human ownership of framing, critique, and moral reasoning, the model institutionalizes the process-focused, ethically embedded measurement logic required to salvage construct validity in generative AI learning environments ([Gupta et al., 2024](#)). Hybrid intelligence facilitation refers to the intentional pedagogical design of complementary human–AI cognitive labor, wherein AI systems assume information-retrieval, pattern-generation, and linguistic-polishing tasks, while human learners retain authority over problem-framing, value judgment, creative synthesis, and ethical decision-making ([Bearman & Ajjawi, 2023](#)). The concept draws on human-in-the-loop (HITL) ([Natarajan et al., 2024](#)) and augmented intelligence studies to emphasize collaboration over substitution, a core tenet distinguishing human-centred hybrid learning models from replacement-oriented AI pedagogy ([Cosentino et al., 2025](#)).

2.3. A Tripartite Framework for Analyzing AI-Assessment Ecosystems

To deconstruct the discourse on AI and creative agency—and to derive concrete implications for assessment design—we propose an integrative framework examining three

interdependent dimensions. Each dimension is reformulated here to directly address the challenges of evaluating learning in an AI-saturated academy.

Tool Dimension (The “How” of Assessment): This dimension focuses on the cognitive practices and thinking skills mediated or transformed by AI, asking how assessment validity is reshaped when AI participates in knowledge production. It interrogates whether a discourse frames AI primarily as an instrument of enhancement—optimizing efficiency, accuracy, and scale of analytical tasks—or as an object of interrogation—questioning the methodological legitimacy, interpretive boundaries, and epistemological implications of these tools. For assessment design, this dimension distinguishes between evaluating analytic proficiency augmented by AI and evaluating critical negotiation of methodological appropriateness and meaning-making. It connects directly to debates about “cognitive deskilling” versus “augmented intelligence,” and to the practical question of whether traditional assessment formats (essays, exams, code submissions) remain valid when AI can generate plausible outputs.

Subject Dimension (The “Who” of Assessment): This dimension addresses the reconstruction of the creative subject as the entity being assessed. The Subject dimension explicitly incorporates “ethical governance of AI use” as a core component of creative agency. It examines how discourses conceptualize the human actor’s role, identity, and values in relation to AI, and therefore what competencies assessment ought to certify. Does the discourse promote a model of adaptive functional collaboration, where the subject develops technical competencies to work seamlessly with AI? Or does it engage in philosophical-critical reflection on technology’s impact on human autonomy, authenticity, and alienation? For assessment, this tension translates into a core pedagogical dilemma: should we evaluate students’ ability to productively partner with AI (collaborative functional competence), or should we prioritize their capacity to maintain sovereign, critical, and ethically grounded creative intention (philosophical agency)? This connects to creative agency literature (Glăveanu, 2013; Arseven & Bal, 2025) and cross-cultural assessment fairness, ensuring the framework is responsive to epistemological diversity. Effective assessment frameworks must integrate both certifying technical fluency while safeguarding the humanistic core of critical thinking.

Organizational Ecosystem Dimension (The “Where” of Assessment): This dimension considers the socio-institutional contexts that enable or constrain specific forms of creative agency, and thus the policy and infrastructural conditions under which assessment operates. It investigates whether the imagined future is one of decentralized, open platform ecosystems that foster networked innovation and bottom-up experimentation, or of strategically guided systemic institutionalization aimed at aligning technological change with national or cultural educational objectives (Carayannis & Campbell, 2009). Drawing on the AI Assessment Scale framework (Perkins et al., 2024) and empirical evidence on assessment redesign (Bittle & El-Gayar, 2025), this study proposes AI use acknowledgment as a mandatory assessment component, shifting the focus of academic integrity from technical detection to pedagogical design. The ecosystem shapes the available resources, incentives, norms, and integrity infrastructures that define what counts as legitimate creative practice. For assessment reform, this dimension raises critical questions about equitable access to AI tools, standardized versus diversified evaluation criteria, and the balance between institutional guidance and academic freedom.

2.4. Synthesizing the Lenses: From Discursive Divergence to Assessment Implications

This framework does not treat the three dimensions in isolation. Instead, it posits that the dominant narrative within a discourse manifests as a coherent configuration across all three. The Tool dimension reveals the immediate pedagogical and evaluative imagination:

what is being assessed and by what means? The Subject dimension exposes the underlying model of the ideal thinker/creator that assessment implicitly certifies. The Organizational Ecosystem dimension uncovers the envisioned infrastructure for supporting such development, including integrity policies, faculty development, and resource allocation.

By mapping the WoS and CNKI discourses onto this framework, we move beyond cataloguing differences to explaining the divergent “logics of practice” they embody. The global technocratic augmentation pathway, for instance, configures Tools as efficiency instruments, Subjects as adaptive collaborators, and Ecosystems as platformized networks; this implies assessment frameworks that reward functional output and rapid integration. The Chinese systemic reconstruction pathway configures Tools as objects of epistemic scrutiny, Subjects as autonomous agents defending value rationality, and Ecosystems as strategically guided institutions; this implies assessment frameworks that prioritize process documentation, ethical reflection, and alignment with culturally specific educational missions.

This analytical synthesis allows us to derive concrete implications for how educators and institutions might design learning environments and, crucially, assessment systems that thoughtfully nurture creative agency in the age of AI. Rather than advocating for either pathway, the framework enables a diagnostic reading of any given educational context, supporting the design of hybrid assessment models that balance tool mastery with critical autonomy, and platform openness with institutional coherence. It is this capacity to inform culturally responsive, integrity-preserving, and future-oriented assessment reform that constitutes the core contribution of this paper to the field.

3. Methodology

This paper follows the logic of “data-driven hypothesis verification” and adopts a research design pathway of “bibliometric analysis—topic modeling—dimension mapping”. First, an interdisciplinary corpus is constructed from the WoS and CNKI databases. Next, it undergoes standardized preprocessing and vectorization. Finally, an unsupervised machine learning model is used to identify the latent topic structure of interdisciplinary research at the intersection of artificial intelligence and the humanities and social sciences. These implicit structures are then mapped onto three core dimensions: tool enhancement, subject reconstruction, and organizational transformation.

Given that this paper focuses on the mechanistic factors through which artificial intelligence influences humanities and social sciences research—rather than simply mapping research hotspots—LDA holds clear advantages in identifying underlying conceptual structures and delineating influential dimensions. As illustrated in Figure 1, this study follows a research design pathway of “bibliometric analysis—LDA topic modeling with Word2Vec-enhanced semantic interpretation—mechanism induction.” The research begins by constructing an interdisciplinary corpus from the WoS and CNKI databases. After standardized filtering and preprocessing, the LDA model is employed to identify the latent topic structure in interdisciplinary research combining artificial intelligence with the humanities and social sciences. Subsequently, the topic results are further mapped onto three mechanism dimensions: cognition and roles, methods and capabilities, and institutions and structures. Ultimately, these are synthesized into three pathways of influence: subject reconstruction (AI+), tool enhancement (+AI), and organizational transformation (Social Sciences for AI, SS4AI).

LDA is a generative probabilistic topic model based on bag-of-words assumptions. Word2Vec is a neural embedding technique used here post hoc to enhance topic interpretation by identifying semantic neighbors of high-probability topic words.

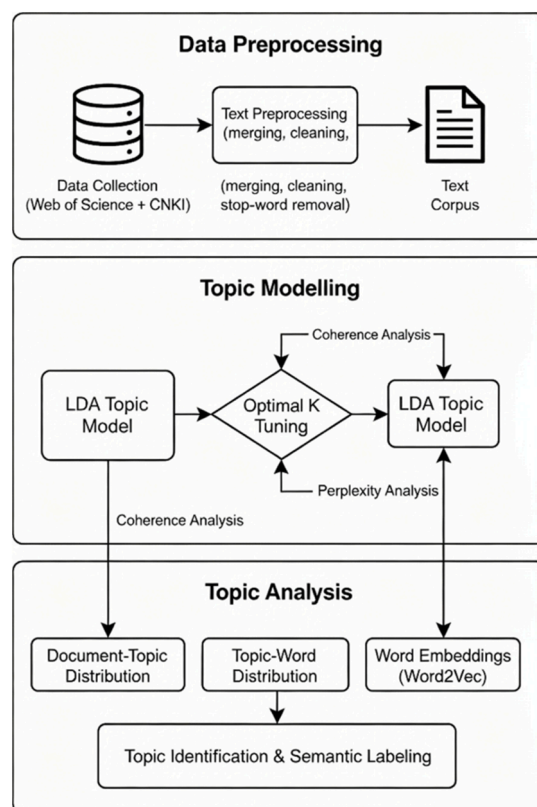


Figure 1. Flowchart of the LDA-Word2Vec literature topic mining framework.

3.1. Database Selection

The selection of data sources is critical in bibliometric research, as it directly determines the comprehensiveness and validity of the findings. This study aims to effectively capture research spanning the intersection of artificial intelligence and the humanities/social sciences through literature searches, thus requiring databases that provide extensive coverage of core journals in both technical and socio-humanistic disciplines. HSS disciplines are uniquely positioned as the “canary in the coal mine” for GenAI assessment reform because their core competencies—critical writing, interpretive reasoning, creative expression, and ethical judgment—are precisely the capabilities that large language models can most plausibly simulate. Unlike STEM fields, where outputs are often empirically verifiable, HSS assessment relies on the demonstration of irreducibly human cognitive processes, making the imperative to move “beyond the correct answer” most urgent here.

Among globally prominent academic databases such as Web of Science (WoS), Scopus, and Google Scholar, each possesses distinct strengths in coverage, data quality, analytical tools, and accessibility (Pranckutė, 2021). For the purposes of this study, however, Scopus and Google Scholar present limitations. They often restrict the structured downloading of data, making standardized, large-scale batch processing challenging (Martín-Martín et al., 2020). Regarding major Chinese databases, CNKI is the world’s largest platform for Chinese academic resources. It holds significant advantages in terms of the comprehensiveness and systematic organization of its collections and enjoys high academic recognition. In contrast, WanFang Data has a narrower disciplinary focus, with an emphasis on scientific, engineering, and medical fields. VIP Information, meanwhile, is more specialized in journal literature and often imposes stricter access requirements, necessitating subscriptions for full resource availability.

Therefore, to ensure coverage of both English and Chinese literature while accommodating interdisciplinary research, this study selected the following two leading databases:

the Web of Science Core Collection and CNKI. For the English corpus, searches were confined to the Social Sciences Citation Index (SSCI) and the Arts & Humanities Citation Index (A&HCI) within the WoS Core Collection, limiting the document type to “Article.” This strategy ensures the academic rigor and theoretical depth of the sample, thereby improving the signal-to-noise ratio of the study. For the Chinese corpus, searches were conducted within the Chinese Social Sciences Citation Index (CSSCI) database on the CNKI platform, specifically targeting the following subject categories: “Philosophy & Humanities”, “Economics & Management”, “Social Sciences I”, and “Social Sciences II”.

The selection of WoS and CNKI as comparative corpora is theoretically motivated rather than geographically arbitrary. The WoS Core Collection (SSCI/A&HCI) predominantly captures the Anglophone/Western scholarly discourse, which has historically shaped the global discourse on educational technology, cognitive augmentation, and functional human–AI collaboration. While we acknowledge the rich diversity of non-Anglophone scholarship in Europe, Africa, Latin America, and other Asian regions, the WoS corpus in our sample is overwhelmingly authored in English and oriented toward technocratic efficiency frameworks. Conversely, CNKI CSSCI represents the world’s largest non-English academic repository, constituting a robust structural counterweight to Western-centric data pools. China is specifically selected as a rigorous comparator for three theoretically grounded reasons: (a) it operates the world’s largest higher education system by enrollment, meaning its macro-discursive shifts carry immense systemic weight; (b) emerging bibliometric signals indicate that Chinese scholarly discourse on generative AI places explicit emphasis on epistemic sovereignty, national paradigm transformation, and subjectivity reconstruction—concerns that are comparatively muted or framed purely as technical considerations in mainstream Anglophone literature; and (c) this binary grouping operationalizes a comparison between two divergent “logics of practice” in AI-education discourse: technocratic augmentation versus systemic reconstruction.

This corpus captures academic discourse—the conceptual frameworks, policy imaginaries, and epistemological assumptions that precede and shape classroom practice. We do not claim that these articles represent direct classroom data; rather, we argue that scholarly discourse is a critical, under-examined site where the ‘logics of practice’ for assessment reform are first articulated. Analyzing this discourse provides an empirical foundation for assessment framework design at the level of epistemology and policy, complementing (not replacing) classroom-based empirical studies.

3.2. Search Strategy

A systematic search strategy was developed following the logic of “core concepts—synonym expansion—Boolean combination,” structured around three conceptual groups: (1) AI and related technologies (e.g., artificial intelligence, machine learning, deep learning, natural language processing, large language models, GPT, BERT); (2) humanities and social sciences (HSS) research (e.g., digital humanities, computational social science, and major disciplinary categories including philosophy, economics, law, education, literature, history, management, and art); and (3) impact and mechanisms (e.g., impact, challenge, transformation, ethics, reflection, application, empowerment). The search was confined to the Topic field (TS/SU), limited to document type “Article,” and restricted to the publication period from 1 January 2016 to 31 May 2025 (search executed on 1 June 2025). The year 2016 marks a significant inflection point—AlphaGo’s victory signaled broad acceptance of deep learning and initiated a period of profound AI influence on HSS research paradigms. Restricting the temporal window to this decade also helps minimize semantic drift effects in subsequent topic modeling analysis.

The final search strings were adapted to the specific syntax of each database:

Web of Science Search String:

```
TS = (
    ("artificial intelligence" OR AI OR "machine learning" OR "deep learning"
    OR "natural language processing" OR NLP OR "text analysis" OR "large language
    model*" OR LLM OR GPT OR BERT)
    AND
    ( ("digital humaniti*" OR "computational social science*" OR "smart social
    science")
    OR
    ( (philosoph* OR econom* OR law OR jurisprud* OR educat* OR literatur*
    OR histor* OR manag* OR art) AND ("academic research" OR "scientific research" OR
    "research paradigm*" OR "research method*" OR methodology) )
    )
    AND
    (impact* OR challenge* OR transform* OR revolution* OR opportunit* OR
    risk* OR ethic* OR reflect* OR applicat* OR empower* OR assist* OR integrat* OR
    fusion)
    )
    )
```

CNKI Search String:

```
SU = ('人工智能' OR 'AI' OR '机器学习' OR '深度学习' OR '自然语言处理' OR
'NLP' OR '文本分析' OR '大语言模型' OR 'GPT' OR 'BERT')
AND (
    SU = ('数字人文' OR '计算社会科学' OR '智慧社科')
    OR
    (SU = ('哲学' OR '经济学' OR '法学' OR '教育学' OR '文学' OR '历史学' OR
'管理学' OR '艺术学') AND KY = ('学术研究' OR '科学研究' OR '研究范式' OR
'研究方法' OR '方法论'))
    )
    AND FT = ('影响' OR '挑战' OR '变革' OR '机遇' OR '风险' OR '伦理' OR '反思'
OR '应用' OR '赋能' OR '辅助' OR '融合')
```

Inclusion Criteria: (1) the research explicitly addresses the intersection of artificial intelligence (including machine learning, deep learning, and large language models) and humanities/social sciences research; (2) published in peer-reviewed journals indexed in SSCI/A&HCI or CSSCI; (3) document type restricted to original research articles, excluding reviews, conference abstracts, book reviews, and editorials; (4) publication date between January 2016 and May 2025; and (5) full title and abstract metadata accessible for extraction.

Exclusion Criteria: (1) duplicate publications or redundant records (only the initial publication version retained); (2) records lacking explicit AI-related technical terminology or HSS research context in title or abstract; (3) purely technically oriented studies without discussion of HSS research paradigms, methodologies, or institutional transformations; and (4) records with incomplete title or abstract metadata.

3.3. Screening Process and Results

This study adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, [Page et al., 2021](#)) statement to ensure transparent and reproducible reporting of the systematic literature screening process. A PRISMA flow diagram illustrating the entire process from initial database search to final analytical sample is included below (see Figure 2).

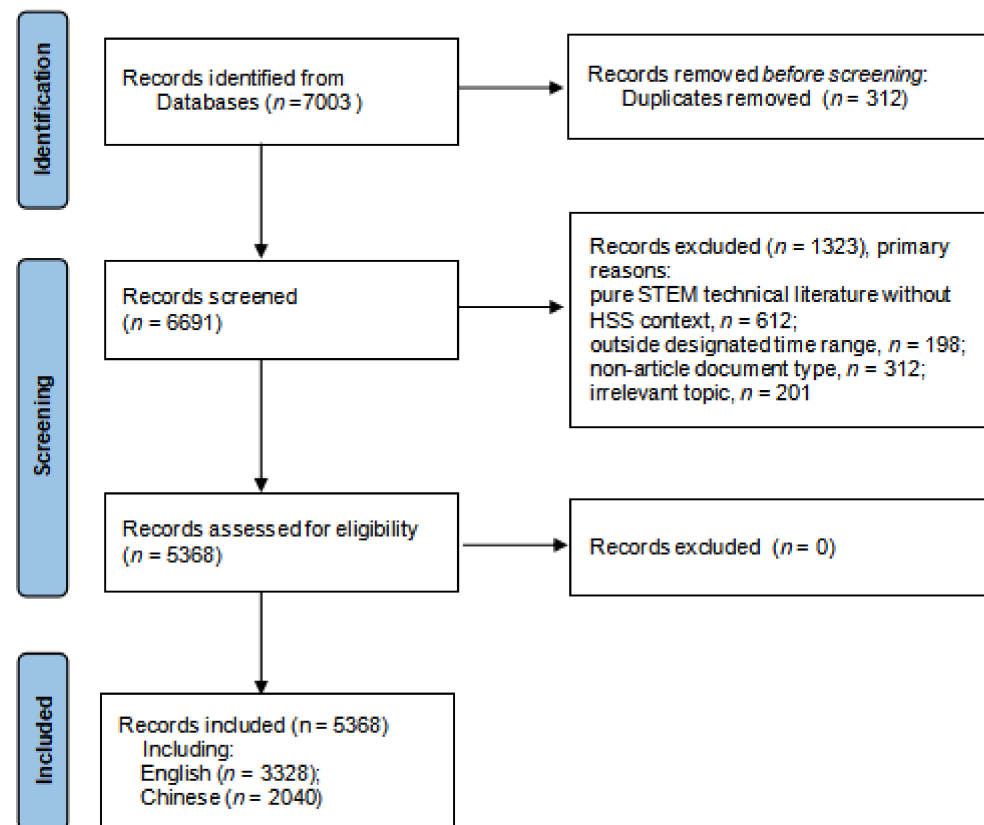


Figure 2. PRISMA flow diagram.

Identification. The initial database search retrieved 4156 English records from WoS and 2847 Chinese records from CNKI, yielding a combined total of 7003 records. All records were exported in plain-text format with core metadata fields including title, abstract, author(s), journal, publication year, and DOI.

For deduplication and data cleaning, a two-round deduplication protocol was implemented using EndNote 21 and Python 3.14 scripts (pandas library). Round 1 employed exact DOI matching to eliminate cross-database duplicates. Round 2 applied title similarity algorithms (Levenshtein distance < 0.15) to identify variant versions of identical publications. This process excluded 312 duplicate records (187 from WoS, 125 from CNKI), leaving 6691 unique records for title-abstract screening.

Title and Abstract Screening. Two researchers independently conducted blinded title-abstract screening using a three-tier classification system: “include,” “exclude,” or “uncertain.” Prior to formal screening, consistency training was conducted on a random sample of 100 records, yielding a Cohen’s Kappa of 0.87, indicating substantial inter-rater agreement. Discrepancies marked as “uncertain” were resolved through discussion and consensus, with full-text consultation when necessary. This stage excluded 1323 records (primary reasons: pure STEM technical literature without HSS context, $n = 612$; outside designated time range, $n = 198$; non-article document type, $n = 312$; irrelevant topic, $n = 201$), leaving 5368 records for full-text eligibility assessment.

For eligibility assessment, all 5368 records underwent full-text accessibility verification and final eligibility confirmation. No records were excluded due to incomplete metadata (ensured by prior screening stages), and no additional duplicates were identified upon full-text inspection. The final analytical sample comprised 5368 articles, including 3328 English (62.0%) and 2040 Chinese (38.0%) records. See Table 1.

Supplementary Search. Backward citation searching of reference lists and forward citation searching by key authors were conducted to identify potentially eligible records

missed by the primary search. No additional eligible records were identified through these supplementary procedures. The final sample thus remained at 5368 articles.

Table 1. PRISMA Screening Record Summary.

Phase	English Records (WoS)	Chinese Records (CNKI)	Total
Records identified from databases	4156	2847	7003
Records after deduplication	3969	2722	6691
Records excluded at title-abstract screening	641	682	1323
Records excluded at full-text assessment	0	0	0
Final analytical sample	3328	2040	5368

3.4. Data Extraction and Preprocessing

The use of titles and abstracts for topic modeling is well-established, as they concisely capture the core content of scholarly articles, a practice supported by comparative analyses (e.g., [Hankar et al., 2025](#)). Comparative studies on multi-source heterogeneous data have further shown that using a combination of titles and abstracts yields the highest coherence scores ([Yu et al., 2023](#)). Therefore, to enhance the efficiency and quality of the topic modeling, the titles and abstracts of the 5368 articles were selected as the analytical text for this study.

Language-specific preprocessing pipelines were developed to accommodate structural differences between English and Chinese corpora while maintaining strict consistency in data cleaning standards:

English Corpus: Lemmatization was performed using NLTK’s WordNetLemmatizer based on part-of-speech tags, followed by lowercasing and stop-word removal. A custom academic noise term list (e.g., “paper,” “study,” “analysis,” “results,” “method,” “finding”) was constructed and removed to minimize interference from formulaic academic phrases.

Chinese Corpus: The jieba library (precise mode) was employed for word segmentation, with the “Artificial Intelligence Terminology Database” (v3.1, containing 2443 domain terms) incorporated as a mandatory custom dictionary to prevent incorrect segmentation of technical compound terms (e.g., “深度学习” [deep learning] segmented as “深度” [deep] and “学习” [learning]). Part-of-speech filtering retained only nouns, verbs, and gerunds.

Subsequently, separate Word2Vec models were trained for each corpus (CBOW architecture; vector size = 300; window size = 5; minimum count = 5; negative sampling = 5) to enhance the semantic interpretation of LDA topic modeling results post hoc.

3.5. Corpus Construction and Preprocessing

This study adopted an exploratory “dual-track, logic-aligned” strategy for constructing and preprocessing the Chinese and English corpora separately. Recognizing the significant differences in linguistic structure between the two languages, tailored preprocessing pipelines were developed for each. However, strict consistency was maintained in data cleaning standards and the definition of domain-specific terminology to ensure the intrinsic comparability of the two independent corpora for subsequent comparative analysis.

3.5.1. Language-Specific Preprocessing

Prior to LDA topic modeling ([Onan, 2019](#)), standardized natural language processing was applied to the raw data—the titles and abstracts of the 5368 selected Chinese and English papers—to remove noise and extract core semantic features. The following differentiated procedures were implemented based on the distinct characteristics of each language.

English Corpus: Lemmatization and Standardization. English text exhibits rich morphological variation (tense, voice, plural forms). To prevent different variants of the same concept (e.g., “analyzing,” “analyzed,” “analysis”) from being treated as distinct features, the following operations were performed using Python’s NLTK library:

Lowercasing: All characters were converted to lowercase to ensure uniformity.

Lemmatization: Preferable to the more aggressive stemming, lemmatization reduces words to their base dictionary form (lemma) based on part-of-speech, preserving more precise semantics.

Stop-word Removal: In addition to a standard stop-word list (e.g., “the,” “is”), a custom list of high-frequency academic stop-words (e.g., “paper,” “results”) was constructed and removed to minimize interference from formulaic academic phrases in topic identification.

Chinese Corpus: Precise Word Segmentation and Terminology Preservation. The primary challenge with Chinese text is the absence of natural word delimiters, compounded by the prevalence of novel compound terms in the AI-SSH interdisciplinary field. To prevent key concepts from being incorrectly segmented (e.g., segmenting “深度学习” [deep learning] into “深度” [deep] and “学习” [learning]), a “domain dictionary + segmentation” strategy was employed:

Domain-Specific Dictionary: The “Artificial Intelligence Terminology Database” (v3.1, containing 2443 terms) compiled by the Machine Heart team was incorporated as a mandatory custom dictionary.

Precise Word Segmentation: The precise mode of the jiebalibrary was used for segmentation, ensuring the integrity of technical terms.

Part-of-Speech Filtering: Only nouns, gerunds, and verbs were retained, while particles, adjectives, and other less substantive words were filtered out to focus the analysis on the core substantive content of the research topics.

3.5.2. Vectorization and Semantic Enhancement (Word2vec)

Although LDA topic modeling was conducted separately on the Chinese and English corpora, it fundamentally relies on the Bag-of-Words model, which assumes word independence and ignores contextual semantic relationships (Simmons et al., 2011). To mitigate this limitation and enhance the coherence of topic clustering within each language, Word2Vec word embedding technology was introduced for semantic enhancement. Previous studies combining LDA with Word2Vec provided a valuable foundation for this approach (e.g., Huang et al., 2021; Bian et al., 2022; Hu et al., 2024).

Separate Word2Vec models were trained for each corpus. Based on the Distributional Hypothesis—that words with similar contexts have similar meanings (Mikolov et al., 2013)—Word2Vec, a neural network-based word embedding technique, captures the contextual meaning of words from natural text and represents them as semantic vectors. This helps address LDA’s difficulty in modeling contextual semantic relationships (Hu et al., 2024), even capturing analogical relationships like $\text{vec}(\text{King}) - \text{vec}(\text{Man}) + \text{vec}(\text{Woman}) \approx \text{vec}(\text{Queen})$. After comparing the Continuous Bag-of-Words (CBOW) and Skip-gram architectures, CBOW was selected for its computational efficiency and more stable representation of high-frequency core words, which aligns with the goal of uncovering core conceptual structures from the large-scale corpus (Mikolov et al., 2013).

The core purpose of using Word2Vec was to assist in feature word selection and, primarily, topic interpretation for the LDA results. By calculating cosine similarity, the model can identify deep semantic associations that LDA might miss. For instance, in the English model, Word2Vec can detect the high similarity between “bias,” “discrimination,” and “inequality” in the vector space; in the Chinese model, it can reveal the connection be-

tween “内卷” (involution) and “过度竞争” (over-competition). Subsequently, during the LDA topic labeling phase, semantically similar words (Top-N similar words) identified via word vectors were aggregated. This aided in more accurately defining topic boundaries and avoiding misinterpretations based solely on word frequency. This semantic enhancement step ensures that the topics generated for each language possess deeper explanatory power, even when modeled independently.

3.6. Methodological Transparency and Hyperparameter Disclosure

To ensure comprehensive methodological reproducibility, Table 2 details the full parameter settings, token handling, and validation metrics implemented across our analytical pipeline. All modeling procedures were executed using Python’s gensim and scikit-learn ecosystems.

Table 2. Complete Machine Learning Pipeline and Hyperparameter Disclosure.

Pipeline Phase	English Corpus (Web of Science) Specifications	Chinese Corpus (CNKI) Specifications
Preprocessing & Tokenization	Lowercasing; English Lemmatization via NLTK WordNetLemmatizer based on POS tags.	Custom mandatory dictionary applied using the Machine Heart AI Terminology Database (v3.1, 2443 domain terms). Word segmentation via jieba (precise mode).
Stop-Word Filtering	Removal of standard NLTK list combined with custom academic noise terms (paper, study, analysis, results, method, finding).	POS filtering applied: only nouns (n), verbs (v), and gerunds (vn) retained. Formulaic academic markers removed.
LDA Model Hyperparameters	K = 5 topics (optimized via Cv elbow plot & perplexity decay). α = ‘auto’, β = ‘auto’. Iterations = 1000. Random state fixed.	K = 4 topics (optimized via Cv elbow plot & perplexity decay). α = ‘auto’, β = ‘auto’. Iterations = 1000. Random state fixed.
Word2Vec Configurations	CBOW architecture; vector size = 300; window size = 5; minimum count = 5; negative sampling = 5. Trained post hoc per corpus.	CBOW architecture; vector size = 300; window size = 5; minimum count = 5; negative sampling = 5. Trained post hoc per corpus.

3.7. From Topics to Assessment: Analytical Protocol

To ensure transparency in the interpretive procedure linking statistical outputs to assessment recommendations, we adopt a four-step analytical protocol:

Step 1: Statistical topic extraction. LDA is applied to identify latent thematic structures within each corpus, yielding topic distributions and keyword probabilities.

Step 2: Semantic clustering and researcher coding. Word2Vec is used post hoc to validate topic boundaries by identifying semantic neighbors of high-probability keywords. Two independent researchers then manually label topics through iterative coding, with disagreements resolved through consensus discussion (Cohen’s κ = 0.84).

Step 3: Dimensional mapping. Each labeled topic is aligned with the tripartite framework (Tool/Subject/Organization) based on its dominant conceptual orientation, as documented in the Mapping Logic column of Table 2.

Step 4: Derivation of format-specific assessment recommendations. Grounded in the preceding three steps, concrete implications for essays, exams, code projects, capstones, and presentations are derived, ensuring each recommendation is traceable to its empirical anchor (topic number, keyword evidence, source corpus).

The raw bibliographic datasets, deduplication and screening logs, Python preprocessing code, and LDA topic modeling output files used in this study are available from the corresponding author upon reasonable request.

4. Results and Findings

The comparative analysis of 5368 academic articles reveals a profound divergence in how the Anglophone/Western and Chinese scholarly communities frame the encounter between AI and higher education. This divergence is not merely topical but reflects fundamentally different priorities regarding the future of cognitive practice, creative agency, and the institutional ecosystems that support them—and, crucially, divergent visions of how assessment must be reconfigured when AI participates in knowledge production. The findings are structured to first highlight surface-level lexical contrasts, then delve into the deeper thematic structures identified through topic modeling, and finally synthesize these into a coherent framework illuminating alternative pathways for cultivating thinking and creativity through assessment reform.

4.1. Lexical Priorities: Statistical Validation of Cross-Corpus Divergence

A comparative analysis of high-frequency terms provides the first macro-level indicator of divergent scholarly focus. To move beyond impressionistic lexical commentary, our text pipeline evaluated these distributions through empirical log-likelihood ratio tests and G^2 and post hoc vector space proximity measurements.

Statistical testing demonstrates that high-frequency tokens in the WoS corpus (model, performance, prediction, accuracy, and algorithm) and the CNKI corpus (范式/paradigm, 主体性/subjectivity, 体系构建/system construction, and 异化/alienation) are significantly over-represented relative to their cross-corpus null distributions ($p < 0.01$, $G^2 > 15.5$). Word2Vec cosine similarity ($\cos\theta$) mapping further confirms that these distinct lexicons explicitly signal deep engagement with artificial intelligence, rather than reflecting generic academic prose. For example, in the Chinese vector space, 主体性 (subjectivity) and 范式 (paradigm) are highly proximal to 人工智能 (artificial intelligence) ($\cos\theta = 0.68$), while 伦理 (ethics) tightly clusters with 算法 (algorithm) ($\cos\theta = 0.71$). This provides quantifiable proof that the underlying cultural vocabularies are conceptually and textually divergent.

To ensure full visual accessibility for readers who do not possess Chinese literacy, the lexical comparisons generated in our pipeline are presented using a unified visual grammar. Figure 2 provides a bilingual comparative visualization where all native Chinese script tokens are paired with direct English semantic glosses (e.g., 主体性 = subjectivity; 异化 = alienation; 体系构建 = system construction), backed by a complete translation matrix in the visual index.

4.2. Analytical Overview and Evidentiary Chain

Building on the lexical divergence documented in Section 4.1, this section employs LDA topic modeling to unpack the latent thematic structures underlying the two corpora. Optimal topic numbers were determined jointly via coherence score (Cv) elbow plots and perplexity decay curves, yielding $K = 5$ topics for the WoS (Anglophone/Western) corpus and $K = 4$ topics for the CNKI (Chinese) corpus. To orient the reader before detailed thematic discussion, Table 3 synthesizes all identified topics, including their descriptive labels, proportional representation in each corpus, top representative keywords, and alignment with the tripartite theoretical framework (Tool, Subject, Organizational Ecosystem) established in Section 2. This overview provides a holistic map of the discursive landscape, highlighting how thematic priorities are distributed across cognitive, subjective, and institutional dimensions in each context. The subsections that follow elaborate on the substantive content and assessment implications of each topic cluster.

Table 3. Summary of Latent Topics Identified by LDA Topic Modeling.

Corpus	Topic ID	Descriptive Topic Label	Proportion	Top Representative Keywords	Corresponding Framework Dimension
WoS (Anglo-phone/Western)	1	Machine Learning Predictive Models & Performance Validation	23.10%	prediction, accuracy, algorithm, model evaluation, performance	Tool & Cognitive Practice
	2	Computational Research Frameworks & Analytical Methodologies	21.10%	framework, methodology, causal inference, research design, analytical approach	Tool & Cognitive Practice
	3	Pedagogical Application & Learner Competency Development	14.20%	education, student, skill development, ChatGPT-5, pedagogical shift	Subject & Creative Agency
	4	AI Ethics, Adoption & Social Implications	17.50%	ethics, adoption, challenge, social implication, technology acceptance	Subject & Creative Agency
	5	Digital Platforms & Networked Knowledge Ecosystems	24.10%	social media, platform, public opinion, knowledge production, networked innovation	Organizational Ecosystem
CNKI (Chinese)	1	Computational Methods & Data-Driven Research	20.10%	data (数据), text (文本), computation (计算), method (方法), empirical analysis (实证)	Tool & Cognitive Practice
	2	Paradigm Transformation & Disciplinary System Construction	29.80%	paradigm (范式), Chinese characteristics (中国特色), transformation (转型), innovation (创新)	Organizational Ecosystem
	3	Philosophical Critique & Subjectivity Reconstruction	23.40%	philosophy (哲学), alienation (异化), essence (本质), subjectivity (主体性), critique (批判)	Subject & Creative Agency
	4	Institutional System & Methodology Innovation	26.70%	system (体系), Marxism (马克思主义), methodology (方法论), construction (构建)	Organizational Ecosystem

4.2.1. The Anglophone/Western Discourse: Tool Augmentation

LDA topic modeling of the WoS corpus (optimal $K = 5$) delineates a thematic landscape predominantly focused on the pragmatic integration and enhancement of research capabilities, with direct implications for how learning and creativity might be evaluated (see Figure 3).

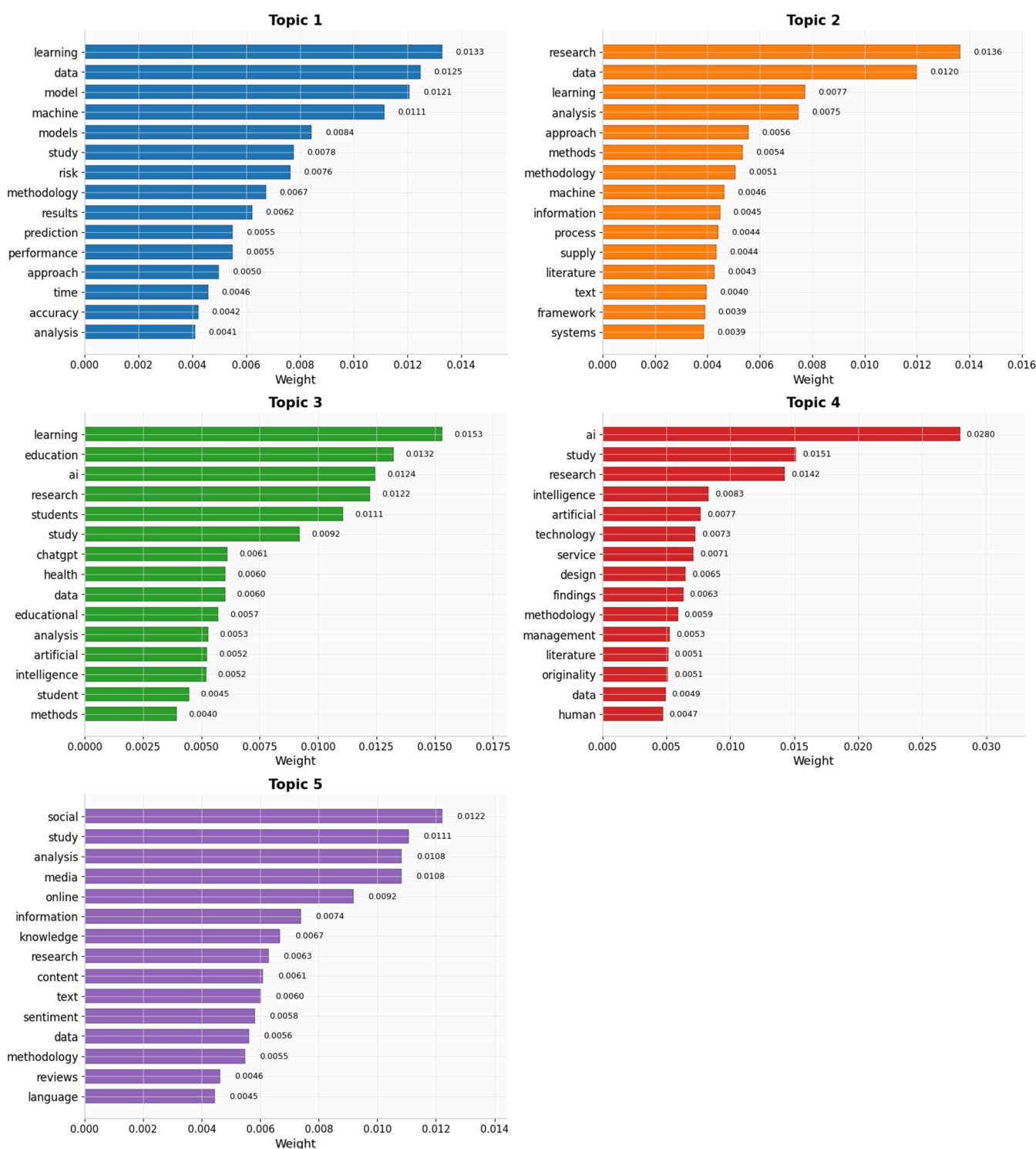


Figure 3. WoS keywords by topic with weights (Anglophone/Western Discourse).

Topics 1 and 2 (Machine Learning Models & Research Frameworks): Together constituting nearly half of the discourse, these interconnected themes represent the core of the Tool-Centered Augmentation pathway. Keywords like *prediction*, *accuracy*, *algorithm*, and *framework* reveal a driving interest in validating AI's efficacy, refining methodologies, and strengthening causal inference. For assessment, this implies a natural alignment with evaluating students' capacity to deploy AI tools effectively: model evaluation, data reasoning, computational problem-solving, and the production of technically sound outputs.

The “correct answer” remains viable but is now algorithmically mediated—assessment certifies the ability to arrive at valid solutions efficiently, with AI as a partner.

Topics 3 & 4 (Learner Development & Technology Ethics): These themes address the dimension of Adaptive Subjectivity. Topic 3 focuses on the impact of AI tools like ChatGPT on education, students, and skills, exploring shifts in pedagogical roles. Topic 4, the largest single theme (17.50%), deals with *ethics, adoption, challenges, and social implications*. Notably, ethical discourse is often coupled with pragmatic concerns of technology acceptance and business integration, reflecting a focus on functional collaboration between human and artificial intelligence. Assessment implications are twofold: on one hand, the need to certify adaptive skills and “AI fluency” as graduate competencies; on the other, a risk that ethical concerns are subordinated to integration imperatives, potentially weakening academic integrity frameworks if not explicitly designed into evaluation criteria.

Topic 5 (Media & Public Space): This theme maps onto Ecosystem Platformization. Centered on *social media, platforms, and public opinion*, it highlights how AI is reorganizing the spaces and flows of knowledge production. For assessment, this points toward more open, networked, and algorithmically-mediated evaluation environments—portfolio-based assessments, peer and public review, and continuous rather than summative evaluation. However, it also raises concerns about equity (uneven access to platforms and AI tools) and the erosion of standardized quality assurance.

4.2.2. The Chinese Discourse: Systemic Reconstruction

The thematic structure of the CNKI corpus (optimal $K = 4$) reveals a preoccupation with foundational critique and systemic rebuilding, with direct consequences for how assessment validity is conceived and defended (see Figure 4).

Topic 1 (Research Methods): Corresponding to the tool dimension, this theme covers 数据 (*data*), 文本 (*text*), and 计算 (*computation*). Its relatively smaller proportion (20.1%) signifies that technical discussion, while present, is subordinate to higher-order concerns about methodological legitimacy. Assessment implication: AI tools are not uncritically adopted for evaluation; their interpretive boundaries and epistemological appropriateness within local disciplinary contexts must first be established.

Topics 2 and 4 (Paradigm Transformation & System Construction): As the dominant cluster (over 56% of the discourse), these themes embody the drive for Epistemic and Organizational Re-articulation. Keywords like 范式 (*paradigm*), 中国特色 (*Chinese characteristics*), 体系 (*system*), 马克思主义 (*Marxism*), and 方法论 (*methodology*) indicate a concerted effort to theorize AI’s integration within a distinct, locally-grounded intellectual and institutional framework. For assessment, this represents a proactive attempt to govern technological change through systemic design: preserving epistemic sovereignty, directing innovation toward national development goals, and ensuring that evaluation criteria align with culturally specific educational missions rather than imported technocratic standards.

Topic 3 (Critical Theory): This theme represents a deep Philosophical–Reflexive engagement with Subject Reconstruction. Keywords such as 哲学 (*philosophy*), 异化 (*alienation*), and 本质 (*essence*) point to an ontological critique of technology, questioning its impact on human nature, autonomy, and value rationality. Assessment implication: A profound concern that instrumental logic may erode the core of humanistic creativity and critical subjectivity. This mandates assessment designs that make human creative intention visible and evaluable—through reflective documentation, process portfolios, and explicit ethical reasoning—rather than relying on product outputs that may be indistinguishable from AI-generated content.

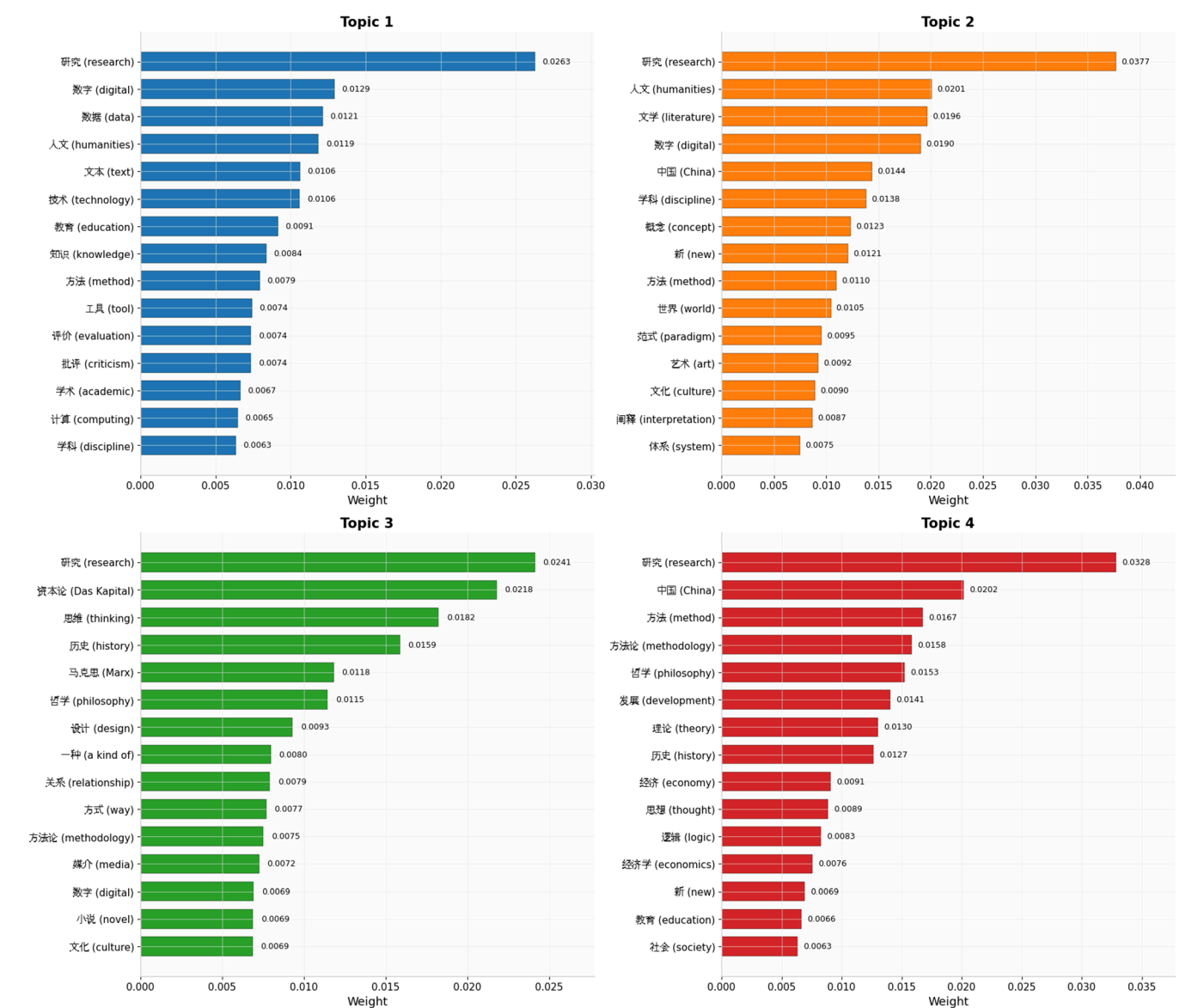


Figure 4. CNKI keywords by topic with weights (Chinese Discourse).

4.3. Redesigned Synthesis Mapping

To guarantee complete structural transparency, Table 4 has been entirely reconstructed to display the explicit empirical anchors, keyword indicators, and exact mapping logics that bridge our raw text-mining output to our final assessment recommendations.

Table 4. Empirical Topic-to-Dimension Mapping Matrix.

Core Dimension	Source Topic & Weight Anchors	Top Key-Term Vectors & Proximity Clusters	Unified Mapping Logic (One-Sentence Derivation)
Tool & Cognitive Practice	WoS Topics 1 & 2 (44.2%); CNKI Topic 1 (20.1%).	prediction, accuracy, framework, 数据 (data), 文本 (text), 计算 (computation).	Maps here because these themes govern the immediate technical, algorithmic, and methodological means through which data and text are manipulated.

Table 4. Cont.

Core Dimension	Source Topic & Weight Anchors	Top Key-Term Vectors & Proximity Clusters	Unified Mapping Logic (One-Sentence Derivation)
Subject & Creative Agency	WoS Topics 3 & 4 (31.7%); CNKI Topic 3 (23.4%).	learner, skill, adoption, 主体性 (subjectivity), 异化 (alienation), 哲学 (philosophy).	Maps here as these topics directly contest the ontological boundaries, ethical liberties, and autonomous identity of the human creator.
Organizational Ecosystem	WoS Topic 5 (24.1%); CNKI Topics 2 & 4 (56.5%).	platform, social media, 范式 (paradigm), 体系构建 (system build), 方法论 (methodology).	Maps here because these concepts focus on macro-level institutional, structural, and policy frameworks that validate educational activity.

4.4. Operationalizing Discipline-Adaptive Exemplars Across Formats

The divergent pathways identified in Sections 4.2 and 4.3 carry distinct and concrete consequences for how traditional assessment formats must be redesigned in the GenAI era. While the global technocratic augmentation pathway implicitly treats existing formats as technically upgradeable—preserving the traditional standardized assessment paradigm as a valid endpoint, albeit algorithmically mediated—the Chinese systemic reconstruction pathway calls for a fundamental epistemological recalibration of what each format is intended to certify. Drawing on recent empirical evidence, this section explicates how six prevalent assessment types—essays and reports, dissertations, code and technical projects, exams, capstone projects, and presentations—must be reinterpreted through the tripartite lens of Tools, Creative Subjectivity, and Organizational Ecosystems.

Written Assessments (Essays, Reports, and Dissertations). The global discourse suggests integrating GenAI as a “writing partner,” shifting evaluation from linguistic polish and structural conformity to argument originality, critical synthesis, and the ethical governance of sources. A recent scoping review suggests that GenAI can advance writing through proofreading, critique, and editing, thereby allowing educators to redirect assessment focus toward higher-order thinking skills such as creativity and problem-solving that GenAI currently struggles to replicate autonomously (Xia et al., 2024). This aligns with the Tool-Centered Augmentation logic: the essay remains viable if assessment certifies the student’s capacity to evaluate, curate, and transcend algorithmic suggestions. However, the Chinese discourse cautions against an uncritical embrace of this pathway, pointing to the risk of cognitive deskilling and the erosion of authorial voice. Empirical studies indicate that over-reliance on AI tools can undermine students’ capacity to generate and develop their own ideas, transforming writing from an act of original composition into one of editorial synthesis (Kofinas et al., 2025). A systemic reconstruction approach, therefore, mandates process-oriented assessment: requiring draft logs, reflective memos documenting human–AI negotiation, and “algorithmic acknowledgments” that disclose which tools were used, for what purpose, and how outputs were critically interrogated (Xia et al., 2024). Such measures transform academic integrity from a detection problem into a pedagogical opportunity for metacognitive development, directly responding to calls for structural assessment changes in the GenAI era. Perkins et al. (2024) are integrated to propose “algorithmic acknowledgment” as a mandatory assessment component, shifting integrity from detection to pedagogical design.

Code and Technical Projects. The technocratic augmentation pathway aligns naturally with AI-assisted coding environments, where assessment might logically focus on functional output and technical efficiency. Yet this risks reducing evaluation to a mere

certification of prompt engineering, obscuring whether the student possesses underlying computational thinking. The Chinese discourse, with its emphasis on methodological legitimacy and subjectivity defense, implies that code assessment must evolve to evaluate the student's capacity to critique, debug, and ethically govern AI-generated outputs—a “human-in-the-loop” competency mirroring professional software engineering standards. Recent research confirms that various assignment types, including computer coding, are impacted differently by GenAI, necessitating format-specific validity criteria rather than blanket policy (Xia et al., 2024). Assessment should therefore require annotated code reviews in which students explain modifications made to AI-generated suggestions, defend architectural choices, and identify potential biases or security vulnerabilities introduced by algorithmic assistance.

Exams and Time-Constrained Assessments. The contrast between decentralized platformization and guided institutionalization is most acute here. The global ecosystem vision points toward open, networked evaluation, yet the vulnerability of traditional examinations to AI infiltration has been empirically demonstrated: a real-world “Turing Test” case study found that AI-generated submissions could successfully penetrate university examination systems, undermining the validity of unseen, product-focused testing. This has prompted calls to return to invigilated, face-to-face examinations as a short-term integrity safeguard, though critics note that such formats are often inauthentic assessments of learning (Kofinas et al., 2025). Conversely, the Chinese-guided systemic approach might favor strategically controlled exam conditions aligned with national educational objectives. Neither pure prohibition nor unmonitored adoption is sustainable. Rather, exam design must shift toward assessing real-time critical judgment under AI assistance—evaluating how students interrogate algorithmic outputs during the assessment event itself—rather than testing information recall that can be delegated to large language models. Oral defenses and vivas, which require spontaneous epistemic negotiation, thus regain prominence as high-validity formats.

Capstone Projects and Presentations. These are arguably the optimal sites for assessing creative agency in its full tripartite configuration. They demand visible tool mastery (demonstrated technical fluency and responsible AI use), subjectivity (defended creative intention, ethical reasoning, and critical autonomy), and ecosystem navigation (alignment with institutional integrity policies and disciplinary norms). The global discourse favors platformized, portfolio-based capstones that maximize cross-disciplinary pollination and public-facing presentation. The Chinese discourse emphasizes systemic coherence, requiring that capstone outputs demonstrate contribution to an autonomous knowledge system. Synthesizing these, the capstone assessment should adopt longitudinal, process-centric evaluation in which students maintain a “creative agency portfolio” documenting problem-framing decisions, AI interaction logs, ethical deliberations, and iterative refinements. Presentations should be evaluated not merely on polished delivery but on the student's capacity to defend their human–AI co-authorship process before an examining panel, thereby making creative subjectivity intersubjectively verifiable.

Synthesis. These format-specific implications confirm that assessment reform cannot rely on universal templates. The technocratic augmentation pathway risks preserving obsolete formats through superficial AI-proofing; the systemic reconstruction pathway risks overburdening assessment with excessive process documentation. The tripartite framework functions here as a diagnostic and generative tool: for each assessment type, educators must ask: (1) Tool: What aspect of this task can AI legitimately mediate, and what must remain human-governed? (2) Subject: What form of creative agency is this format intended to certify, and how can it be made visible? (3) Organization: What integrity infrastructure and equitable access conditions must be in place for this format to function

validly? By applying these questions, institutions can move beyond the traditional standardized assessment paradigm toward evaluating the situated, ethical, and culturally embedded practice of human–AI co-authorship.

5. Discussion: Reconciling Technocratic Augmentation and Humanistic Reconstruction

The comparative discourse analysis reveals not merely topical differences, but two foundational narratives about the future of cognition and evaluation in higher education: a global pathway of Technocratic Augmentation and a local (Chinese) pathway of Systemic Reconstruction. This “cognitive gap” transcends geopolitical boundaries, representing a fundamental pedagogical dilemma in the GenAI era: should we primarily focus on enhancing existing cognitive processes with powerful tools, or on critically re-examining and safeguarding the humanistic core of creative and critical thought? This discussion interprets the tripartite framework (Tool, Subject, Organization) through this lens, drawing connections to contemporary debates in assessment reform, academic integrity, and equitable AI integration.

5.1. Tool Upgrade: Mitigating Construct-Irrelevant Variance

The global discourse’s focus on model performance and predictive accuracy emphasizes the Tool-Centered Augmentation pathway, optimizing writing and analysis (Xia et al., 2024). However, an uncritical embrace carries pedagogical risks of cognitive deskilling, transforming writing from original composition into passive editorial curation (Kofinas et al., 2025). The Chinese discourse’s emphasis on methodological boundaries serves as a vital corrective. In psychometric terms, shifting evaluation toward process-oriented documentation directly limits construct-irrelevant variance—preventing students from masking severe cognitive deficiencies behind the stylistic elegance of a machine-generated output.

5.2. Subject Reconstruction: Redefining Hybrid Creative Agency

Rather than splitting competence into technical fluency (the global pragmatic track) or philosophical resistance (the Chinese track), a sustainable framework must synthesize them. Creative agency in the GenAI era is inherently a hybrid competency. Assessment rubrics must be redesigned to evaluate how effectively a student directs an algorithm while preserving human critical intention. This directly preserves construct validity by testing the student’s capacity to evaluate, critique, and transcend automated suggestions rather than accepting them passively.

5.3. Organizational Transformation: Equity and Integrity-by-Design

At the ecosystem scale, the vulnerability of traditional product assessments to AI infiltration has been empirically verified via real-world “Turing Test” case studies, where machine outputs easily bypassed traditional university exams (Scarfe et al., 2024). Crucially, our framework notes that a purely platform-driven organizational model risks exacerbating deep inequities if access to expensive, premium, enterprise-grade AI models becomes a hidden prerequisite for achieving superior grades. Conversely, a guided institutional system—by strategically provisioning university-wide, standardized AI tool access and literacy training—better ensures equitable assessment conditions for vulnerable or low-income students. Ultimately, organizational policy must transition from punitive, unreliable detection software to an integrity-through-design ecosystem paradigm.

5.4. Toward Culturally Responsive Assessment Frameworks

We acknowledge that pedagogical content knowledge (PCK) varies substantially across HSS sub-disciplines (e.g., philosophy vs. economics vs. literary studies), and that our corpus-level analysis captures macro-discursive trends rather than micro-level disciplinary specificities. We explicitly state that the framework should be treated as a diagnostic meta-structure requiring discipline-specific operationalization, and we call for future research to validate the framework within individual departments.

The tripartite framework functions not as a prescriptive template but as a diagnostic and generative tool for culturally responsive assessment design. For any given educational context, educators and policymakers must ask three interlocking questions:

- Tool: What aspect of this assessment task can AI legitimately mediate, and what must remain human-governed?
- Subject: What form of creative agency is this format intended to certify, and how can it be made visible and evaluable?
- Organization: What integrity infrastructure, equitable access conditions, and policy frameworks must be in place for this assessment to function validly?

By applying these questions, institutions can move beyond the traditional standardized assessment paradigm toward evaluating the situated, ethical, and culturally embedded practice of human–AI co-authorship. The global and Chinese pathways are not mutually exclusive; rather, they represent poles of a spectrum along which culturally specific hybrid models can be calibrated. The imperative is not to universalize a single best practice but to develop flexible, discipline-adaptable frameworks that can evolve in response to rapid changes in AI capabilities while preserving the humanistic core of higher education.

6. Conclusions and Implications for the Genai Era

This study presents a comparative discourse analysis to deconstruct the epistemological divergence between Anglophone/Western and Chinese scholarly communities regarding artificial intelligence integration in higher education. The findings confirm a clear cognitive gap between the global technocratic augmentation pathway and the Chinese systemic reconstruction pathway: Anglophone/Western scholarship predominantly conceptualizes AI as an efficiency-oriented cognitive tool aimed at optimization and functional output, while Chinese scholarship centers on the ontological tensions, ethical defense, and systemic reconstruction of creative subjectivity and epistemic autonomy. This divergence demonstrates that AI integration is not a technologically deterministic trajectory but a deeply culturally embedded negotiation of educational values and institutional missions. By moving beyond oversimplified universalisms, this cross-cultural contrast provides critical empirical evidence to scaffold the transition from product-centric evaluation toward process-oriented assessment models.

Against the backdrop of rapidly evolving generative AI tools, the tripartite framework of Tools, Creative Subjectivity, and Organizational Ecosystems proposed herein offers a scalable meta-structure to reconcile technocratic efficiency with humanistic pedagogies. Rather than positioning AI as either a threat or a panacea, this framework institutionalizes a balanced evaluative architecture that safeguards human critical autonomy while systematically authenticating AI-assisted cognitive workflows.

The conclusions of this study carry three structurally aligned implications for pedagogy, curriculum, and institutional policy, which directly align with the themes of this Special Issue.

First, at the Tool Dimension, curriculum innovation must center on critical AI literacy. Traditional digital literacy is no longer sufficient; students need to develop critical evaluation, ethical judgment, and metacognitive governance over AI use ([Annappureddy](#)

et al., 2025). Curriculum design should integrate deconstructive analysis of AI outputs, reflective documentation of human–AI collaboration, and transparent ethical co-authorship practices. Such a shift helps safeguard academic integrity while moving beyond passive tool use toward active, critical, and responsible engagement with AI (Perkins et al., 2024; Smith et al., 2025). Assessment must evolve to certify these literacies through process-oriented formats that make human–AI negotiation visible and evaluable.

Second, at the Subject Dimension, pedagogy must shift toward hybrid intelligence facilitation. Educators should no longer act solely as knowledge transmitters, but as designers of collaborative human–AI tasks. Pedagogical practice should prioritize process over product, encourage original problem-framing, and reserve high-cognitive tasks such as value judgment, critical synthesis, and creative expression for human learners. This structural shift in pedagogical design helps avoid cognitive deskilling and supports the sustained development of creative agency (Corbin et al., 2025). Assessment rubrics must be redesigned to reward epistemic autonomy and ethical reasoning, not merely functional output.

Third, at the Organizational Ecosystem Dimension, institutions should build critically enabling ecosystems. Universities must transition from reactive, detection-based policing toward proactive, integrity-by-design infrastructures. This requires provisioning equitable, institutional access to advanced AI tools to prevent structural grading disparities among vulnerable student populations, while simultaneously formalizing ‘algorithmic acknowledgment’ standards that integrate ethical disclosure directly into the learning workflow.

This study is not without limitations. The comparative analysis focuses on academic discourse rather than real-time classroom ethnographies, and the LDA topic modeling captures macro-thematic structures rather than granular, micro-level pedagogical dynamics within specific disciplines. Future research should triangulate these macro-discursive insights with localized empirical interventions, evaluating how diverse student cohorts interact with the tripartite framework across varying socio-technical environments. In sum, the era of generative AI does not diminish the value of human thinking and creativity; instead, it compels higher education to return to its core mission: cultivating independent, critical, creative, and ethically grounded individuals. By moving beyond the traditional standardized assessment paradigm and balancing technological empowerment with humanistic values, education can equip learners to harness AI as a collaborative partner while preserving the autonomy, originality, and critical spirit that define academic and intellectual life. This study contributes to a globally informed, culturally responsive, and future-oriented foundation for teaching, assessment, and policy in the age of AI.

Future research may further enhance the replicability of comparative discourse analysis by leveraging open, web-based bibliometric analysis platforms such as BiBL0X (biblox.bandirma.edu.tr) to standardize data processing workflows and support transparent, reproducible research reporting.

Author Contributions: Conceptualization, W.Y. and S.Q.; methodology, S.Q.; validation, S.Q.; formal analysis, S.Q.; writing—original draft preparation, S.Q.; writing—review and editing, S.Q. and W.X.; supervision, W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant number 72474192.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

- Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2025). Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice*, 6(1), 1–21. [\[CrossRef\]](#)
- Arseven, T., & Bal, M. (2025). Critical literacy in artificial intelligence assisted writing instruction: A systematic review. *Thinking Skills and Creativity*, 57, 101850. [\[CrossRef\]](#)
- Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54, 1160–1173. [\[CrossRef\]](#)
- Bian, Y., Lu, Y., & Li, J. (2022). Research on an artificial intelligence-based professional ability evaluation system from the perspective of industry-education integration. *Scientific Programming*, 2022, 4478115. [\[CrossRef\]](#)
- Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16, 296. [\[CrossRef\]](#)
- Carayannis, E. G., & Campbell, D. F. (2009). ‘Mode 3’ and ‘Quadruple Helix’: Toward a 21st century fractal innovation ecosystem. *International Journal of Technology Management*, 46(3/4), 201–234. [\[CrossRef\]](#)
- Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, 50(7), 1087–1097. [\[CrossRef\]](#)
- Cosentino, G., Anton, J., Sharma, K., Gelsomini, M., Giannakos, M., & Abrahamson, D. (2025). Generative AI and multimodal data for educational feedback: Insights from embodied math learning. *British Journal of Educational Technology*, 56, 1686–1709. [\[CrossRef\]](#)
- Glăveanu, V. P. (2013). Rewriting the language of creativity: The five A’s framework. *Review of General Psychology*, 17(1), 69–81. [\[CrossRef\]](#)
- Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. *International Journal of Information Management Data Insights*, 4(1), 100232. [\[CrossRef\]](#)
- Hankar, M., Kasri, M., & Beni-Hssane, A. (2025). A comprehensive overview of topic modeling: Techniques, applications and challenges. *Neurocomputing*, 628, 129638. [\[CrossRef\]](#)
- Hu, H., Li, F., & Luo, Z. (2024). The evolution of China’s English education policy and challenges in higher education: Analysis based on LDA and Word2Vec. *Frontiers in Education*, 9, 1385602. [\[CrossRef\]](#)
- Huang, Y., Wang, R., Huang, B., Wei, B., Zheng, S. L., & Chen, M. (2021). Sentiment classification of crowdsourcing participants’ reviews text based on LDA topic model. *IEEE Access*, 9, 108131–108143. [\[CrossRef\]](#)
- Kofinas, A. K., Tsay, C. H., & Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *Educational Technology*, 56(6), 2522–2549. [\[CrossRef\]](#)
- Marginson, S. (2022). Research on international and global higher education: Six different perspectives. *Oxford Review of Education*, 48(4), 421–438. [\[CrossRef\]](#)
- Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & López-Cózar, E. D. (2020). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations’ COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), 871–906. [\[CrossRef\]](#) [\[PubMed\]](#)
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, arXiv:1301.3781. [\[CrossRef\]](#)
- Moody, C. E. (2016). Mixing Dirichlet topic models and word embeddings to make lda2vec [Preprint]. *arXiv*. [\[CrossRef\]](#)
- Natarajan, S., Mathur, S., Sidheekh, S., Stammer, W., & Kersting, K. (2024). Human-in-the-loop or AI-in-the-loop? Automate or Collaborate? *arXiv*. [\[CrossRef\]](#)
- Onan, A. (2019). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, 7, 145614–145633. [\[CrossRef\]](#)
- 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., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(06), 49–66. [\[CrossRef\]](#)
- Perkins, M., & Roe, J. (2023). Decoding academic integrity policies: A corpus linguistics investigation of AI and other technological threats. *Higher Education Policy*, 37(3), 633–653. [\[CrossRef\]](#)

- Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*, 9, 12. [[CrossRef](#)]
- Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. *PLoS ONE*, 19, e0305354. [[CrossRef](#)] [[PubMed](#)]
- Shavelson, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Gallagher, L., & Quihuis, G. (2002). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue. *Educational Assessment*, 8(2), 77–99. [[CrossRef](#)] [[PubMed](#)]
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. [[CrossRef](#)] [[PubMed](#)]
- Smith, D., Sokoya, D., Moore, S., Okonkwo, C., Boyd, C., Lacey, M. M., & Francis, N. (2025). Embedding generative AI as a digital capability into a year-long skills program. *Journal of University Teaching and Learning Practice*, 22(4), 1–24. [[CrossRef](#)]
- Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *Journal of Educational Technology in Higher Education*, 21(1), 40. [[CrossRef](#)]
- Yu, H., Deng, C., Zhang, Q., Liu, J., Chen, Q., & Wang, W. (2023). Improving long document topic segmentation models with enhanced coherence modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December* (pp. 5592–5605). Association for Computational Linguistics. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.